UNIVERSITÄT
KOBLENZ · LANDAU
Fachbereich 4: Informatik

# Data Mining im Fußball

## Masterarbeit

zur Erlangung des Grades einer Master of Science (M.Sc.)
im Studiengang Informatik

vorgelegt von

Christoph Maiwald

| | |
|---|---|
| Erstgutachter: | Prof. Dr. Ulrich Furbach<br>Universität Koblenz-Landau |
| Zweitgutachter: | Dipl.-Inform. Markus Maron<br>Universität Koblenz-Landau |

Koblenz, im September 2013

# Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

| | Ja | Nein |
|---|---|---|
| Mit der Einstellung der Arbeit in die Bibliothek bin ich einverstanden. | ☐ | ☐ |
| Der Veröffentlichung dieser Arbeit im Internet stimme ich zu. | ☐ | ☐ |

.....

(Ort, Datum) (Unterschrift)

## Zusammenfassung

Data Mining ist die Anwendung verschiedener Verfahren, um nützliches Wissen automatisch aus einer großen Menge von Daten zu extrahieren. Im Fußball werden seit der Saison 2011/2012 umfangreiche Daten der Spiele der 1. und 2. Bundesliga aufgenommen und gespeichert. Hierbei werden bis zu 2000 Ereignisse pro Spiel aufgenommen.

Es stellt sich die Frage, ob Fußballvereine mithilfe von Data Mining nützliches Wissen aus diesen umfangreichen Daten extrahieren können.

In der vorliegenden Arbeit wird Data Mining auf die Daten der 1. Fußballbundesliga angewendet, um den Wert bzw. die Wichtigkeit einzelner Fußballspieler für ihren Verein zu quantifizieren. Hierzu wird der derzeitige Stand der Forschung sowie die zur Verfügung stehenden Daten beschrieben. Im Weiteren werden die Klassifikation, die Regressionsanalyse sowie das Clustering auf die vorhandenen Daten angewendet. Hierbei wird auf Qualitätsmerkmale von Spielern, wie die Nominierung eines Spielers für die Nationalmannschaft oder die Note, welche Spieler für ihre Leistungen in Spielen erhalten eingegangen. Außerdem werden die Spielweisen der zur Verfügung stehenden Spieler betrachtet und die Möglichkeit der Vorhersage einer Saison mithilfe von Data Mining überprüft. Der Wert einzelner Spieler wird mithilfe der Regressionsanalyse sowie einer Kombination aus Cluster- und Regressionsanalyse ermittelt.

Obwohl nicht in allen Anwendungen ausreichende Ergebnisse erzielt werden können zeigt sich, dass Data Mining sinnvolle Anwendungsmöglichkeiten im Fußball bietet. Der Wert einzelner Spieler kann mithilfe der zwei Ansätze gemessen werden und bietet eine einfache Visualisierung der Wichtigkeit eines Spielers für seinen Verein.

**Schlagwörter**: Datenanalyse, Sport, Klassifikation, Clustering, Regression

## Abstract

The term Data Mining is used to describe applications that can be applied to extract useful information from large datasets. Since the 2011/2012 season of the german soccer league, extensive data from the first and second Bundesliga have been recorded and stored. Up to 2000 events are recorded for each game.

The question arises, whether it is possible to use Data Mining to extract patterns from this extensive data which could be useful to soccer clubs.

In this thesis, Data Mining is applied to the data of the first Bundesliga to measure the value of individual soccer players for their club. For this purpose, the state of the art and the available data are described. Furthermore, classification, regression analysis and clustering are applied to the available data. This thesis focuses on qualitative characteristics of soccer players like the nomination for the national squad or the marks players get for their playing performance. Additionally this thesis considers the playing style of the available players and examines if it is possible to make predictions for upcoming seasons. The value of individual players is determined by using regression analysis and a combination of cluster analysis and regression analysis.

Even though not all applications can achieve sufficient results, this thesis shows that Data Mining has the potential to be applied to soccer data. The value of a player can be measured with the help of the two approaches, allowing simple visualization of the importance of a player for his club.

**Keywords**: Data analysis, Sports, Classification, Clustering, Regression

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Abkürzungsverzeichnis

| | |
|---|---|
| **EM** | Expectation-Maximization |
| **KDD** | Knowledge Discovery in Databases |
| **RMSE** | Root Mean Squared Error |
| **SQL** | Structured Query Language |
| **VDV** | Vereinigung der Vertragsfußballspieler |
| **Weka** | Waikato Environment for Knowledge Analysis |

# 1 Einleitung

Bereits 1982 schrieb der Zukunftsforscher John Naisbitt in seinem Buch „Megatrends": „We are drowning in information and starved for knowledge" [46]. Ein Satz der in der heutigen Zeit noch mehr an Bedeutung gewinnt. In der Epoche des Informationszeitalters fallen immer mehr Daten an. Im Jahr 2010 ist das weltweite Datenvolumen auf 1,2 Zettabyte gestiegen [19, S. 1]. Dies entspricht rund 1.2 Billionen Gigabyte. Laut einer Studie verdoppelt sich dieses Datenvolumen bis zum Jahr 2020 etwa alle zwei Jahre [20, S. 1]. Wir ertrinken in der derzeitigen Datenmenge und stehen vor der Aufgabe aus diesen Daten nützliches Wissen zu extrahieren. Ein Ansatz zum Wissensgewinn aus einer großen Menge von Daten stellt die Anwendung von Data Mining dar. Data Mining ist die Auswahl geeigneter Verfahren und Algorithmen sowie deren Anwendung auf große Datenmengen, um Wissen automatisch aus diesen Daten zu extrahieren.

Seit der Saison 2011/2012 werden die Spieler der Fußballbundesliga mithilfe mehrerer im Stadion angebrachter Kameras verfolgt und ihr Laufverhalten statistisch erfasst. Während eines Spiels werden dem Fernsehzuschauer so beispielsweise die gelaufenen Kilometer eines Spielers präsentiert. Zur Saison 2011/2012 gab es von der DFL eine Ausschreibung zur offiziellen Erhebung aller Spielaktionen, wie Zweikämpfe oder Ballkontakte der 1. und 2. Bundesliga. Seit dieser Saison werden die Daten offiziell von einer externen Firma erfasst und bereitgestellt. In einem Spiel werden dabei bis zu 2000 Ereignisse aufgenommen [29]. Der enorme Anstieg des Datenvolumens hält somit auch im deutschen Fußball seinen Einzug. Statistiken über einzelne Elemente des Fußballs werden jetzt schon den Trainern, Spielern oder Fans von Vereinen präsentiert. So wird etwa der Ballbesitz oder die Anzahl an Ecken bzw. Torschüssen der Gegner eines Spiels gegenübergestellt. Durch eine größere Datenmenge wird in Zukunft auch das Bedürfnis der automatischen Extraktion von Wissen aus diesen Daten gerade von Verantwortlichen der Vereine steigen.

In der Saison 2011/2012 wurden von den Bundesligavereinen insgesamt rund 296 Millionen Euro in Spielertransfers investiert, was einem Anteil von 14,61 % aller Ausgaben entspricht [14]. Damit wird ein großer Teil der Gelder von Vereinen in neue Spieler investiert. Spieler werden von Vereinen verpflichtet, um die Leistung der Mannschaft zu steigern. Je höher die Ablöse und das Gehalt eines Spielers, desto mehr verspricht sich der Verein eine Leistungssteigerung der Mannschaft. Die Höhe der Ablösesumme eines Spielers sowie sein Gehalt spiegeln demnach den Wert des Spielers für einen Verein wieder. Da viele Parteien an Verhandlungen über die Ablösesummen beteiligt sind, kann anhand der Höhe dieser Summen nicht direkt auf den Wert eines Spielers geschlossen werden. Doch wie kann ein solcher Wert eines Spielers gemessen werden? Lässt sich mithilfe umfangreicher Daten von Spielern unter Zuhilfenahme komplexer Datenanalyse, wie dem Data Mining der Wert bzw. die Wichtigkeit eines Spielers für seinen Verein quantifizieren? Dieser Frage soll im Laufe dieser Arbeit nachgegangen werden.

Der Wert eines Spielers für einen Verein kann aus verschiedenen Blickpunkten erfolgen. So spielt unter anderem die Sympathie eines Spielers eine wichtige Rolle. Ein Spieler, der eine gewisse Ausstrahlungskraft besitzt, kann Sponsoren oder Fans für einen Verein motivieren, sich diesem anzuschließen. Somit hat ein Spieler, neben seiner fußballerischen Qualität auch einen wirtschaftlichen Wert für einen Verein. Der wirtschaftliche Wert soll in dieser Arbeit jedoch nicht betrachtet werden. Es soll nur die reine physische Leistung eines Spielers innerhalb den Bundesligaspielen berücksichtigt werden.

In dieser Arbeit soll der Fokus auf den offensichtlichsten Wert eines Spielers für eine

Mannschaft gelegt werden, nämlich den von ihm erspielten Punkten. Das heißt, diese Arbeit orientiert sich daran die Punkte zu erfassen, die ein einzelner Spieler zu der Gesamtpunktzahl seiner Mannschaft innerhalb einer Saison beiträgt. Da der Erfolg einer Mannschaft am Ende der Saison von den erspielten Punkten der Mannschaft abhängt, kann von einer Aussage über die Punkte, die ein einzelner Spieler zu der Gesamtleistung der Mannschaft beiträgt auf seine Wichtigkeit für den entsprechenden Verein geschlossen werden.

Neben dieser Kennzahl, können die Nominierung zum Nationalspieler oder die von externen Beobachtern vergebene Note für die Leistung in einem Spiel weitere Qualitätsmerkmale eines Spielers sein. Außerdem kann entscheidend sein, wie sich die Spielweise eines Spieler in das vorhandene Potential einer Mannschaft einfügt. Innerhalb dieser Arbeit wird auch auf solche Merkmale eines Spielers geachtet.

Um den Wert der Spieler zu messen, liegen dieser Arbeit Daten über zwei Saisons der 1. Fußballbundesliga vor. Unter dem ausgewählten Einsatz von Data Mining Verfahren soll versucht werden mittels mehrerer beispielhafter Anwendungen die Wichtigkeit einzelner Spieler zu messen.

Nach dieser Einleitung wird in Kapitel 2 auf die Grundlagen des Data Minings eingegangen sowie der Prozess zur Wissensfindung in Daten beschrieben.

Einen Einblick in den derzeitige Forschungsstand der komplexen Datenanalyse im Sport und vor allem im Fußball wird in Kapitel 3 gewährt. Hierzu wird zunächst der Stand der Analysen im Fußball sowie der Stand der Forschung in anderen Sportarten gegenübergestellt. Anschließend werden einzelne ausgewählte Artikel zusammengefasst, um dem Leser eine Vorstellung über den derzeitigen Forschungsstand zu geben.

Kapitel 4 stellt die Daten vor, auf denen die Analysen in dieser Arbeit basieren. Hierzu liegen die Daten über die Saison 2010/2011 sowie 2011/2012 vor. Es wird dabei grundlegend auf die Aufnahme von Daten im Fußball und deren Verfügbarkeit eingegangen. Zusätzlich werden die hier zur Verfügung stehenden Daten beschrieben. Zum Abschluss dieses Kapitels wird die Struktur der endgültigen Datenbasis dargestellt.

In Kapitel 5 werden ausgewählte Data Mining Verfahren auf die beschriebenen Daten angewendet. Hierzu wird in Kapitel 5.1 ein Versuch unternommen, Nationalspieler vorherzusagen. Der Fokus liegt hier auf der Entdeckung von Spielern, die Nationalspielerniveau besitzen, jedoch von Nationaltrainern bei der Nominierung des Nationalkaders nicht berücksichtigt werden. Kapitel 5.2 beschäftigt sich mit der Vergabe von Noten für die Leistung von Spielern in einzelnen Spielen. Hier wird Data Mining zur Beschreibung der Beziehungen zwischen den Spielaktionen von Spielern und der erhaltenen Note benutzt. Kapitel 5.3 nutzt Data Mining um die Spieler in homogene Gruppen zu unterteilen und so die Spielweise einzelner Spieler zu beschreiben. Der Einsatz von Data Mining zur Vorhersage einer zukünftigen Saison mittels Data Mining wird in Kapitel 5.4 überprüft.

Kapitel 6 geht gezielt auf die Eingangsfrage ein. Hier werden die erhaltenen Erkenntnisse aus dem vorangegangen Kapitel zur Messung der Qualität einzelner Spieler herangezogen und eine Kombination zweier Data Mining Verfahren verwendet, um den Anteil der erspielten Punkte einzelner Spieler zur Gesamtpunktzahl ihrer Mannschaft zu ermitteln.

Die Arbeit endet mit Kapitel 7. Hier wird dargestellt, welche Ergebnisse hinsichtlich der aufgeworfenen Frage erzielt worden sind und wie diese erreicht wurden. Es wird erörtert, ob die Anwendung von Data Mining Methoden im Fußball den Verantwortlichen von Vereinen im Allgemeinen helfen kann und im Besonderen, ob die Messung der Wichtigkeit einzelner Spieler sinnvoll umgesetzt werden kann.

# 2 Grundlagen des Data Mining

Das folgende Kapitel soll die Grundlagen beschreiben, auf denen die in dieser Arbeit verwendeten Data Mining Verfahren basieren. Hierbei werden zunächst die Begriffe Data Mining, Knowledge Discovery in Databases (KDD) und Machine Learning thematisiert. Anschließend werden einzelne Data Mining Verfahren, wie die Klassifikation, die Regression und das Clustering beschrieben.

Data Mining ist der Kernschritt des in Abschnitt 2.1 beschriebenen KDD Prozesses [41, S. 1]. Als Data Mining wird der automatisierte bzw. semi-automatisierte Prozess bezeichnet, welcher nützliche Muster in Daten findet. Laut Fayyad, Piatetsky-Shapiro und Uthurusamy [16] ist Data Mining die reine Anwendung von Algorithmen, um Muster aus Daten zu extrahieren, während KDD den gesamten Prozess bezeichnet, um nützliches Wissen in Daten zu identifizieren. Die Autoren definieren in dem Artikel KDD als den nicht-trivialen Prozess um valide, neue, potentiell nützliche und verständliche Muster in Daten zu finden [16, S.6]. Häufig wird in der Literatur KDD und Data Mining als Synonym verwendet.

Ein weiterer Begriff, der in diesem Zusammenhang häufig auftritt ist der Term Machine Learning. Machine Learning ist ein Computer System, welches seine Leistung steigert, indem es aus Erfahrungen lernt [44, S. 2]. Dabei basieren die Data Mining Methoden u.a. auf den Konzepten des Machine Learnings [16, S. 12]. Oft werden Data Mining und Machine Learning jedoch als Synonyme verwendet. In dieser Arbeit wird im Folgenden zur Vereinfachung nur der Begriff Data Mining verwendet und der Begriff Machine Learning ausgelassen.

Data Mining wird in dieser Arbeit als Begriff für die Schritte 5 bis 7 des im nachfolgenden Abschnittes vorgestellten KDD-Prozesses verstanden. Dies sind die Schritte, welche die Auswahl der Data Mining Verfahren, die Auswahl der verwendeten Data Mining Methoden bzw. Algorithmen und die Anwendung der Algorithmen auf die Daten beinhalten.

## 2.1 Der KDD-Prozess

Der KDD-Prozess ist der organisierte Prozess um valide, neue, nützliche und verständliche Muster aus großen und komplexen Datensätzen zu finden [41, S. 1]. Er besteht aus neun aufeinanderfolgenden Schritten. Der Prozess ist in Abbildung 1 dargestellt.

Eine detaillierte Beschreibung des KDD-Prozesses und der Schritte geben Maimon und Rokach [41]. Nachfolgend sollen die Schritte des Prozesses kurz beschrieben werden.

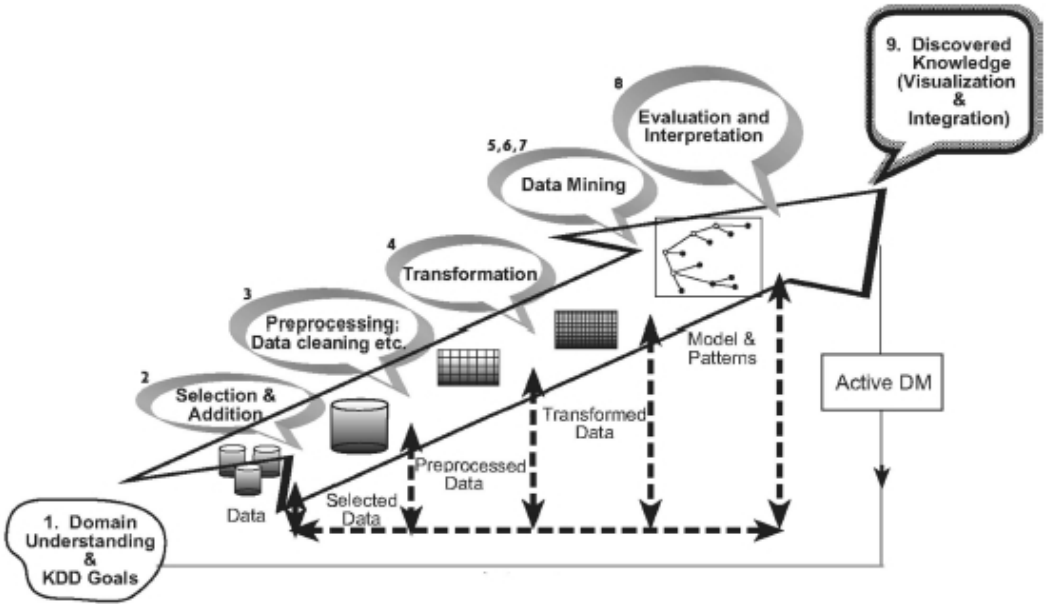
Der Prozess startet mit der Vorbereitung auf die nachfolgenden Schritte. Hier soll die Domäne bzw. der Bereich verstanden werden in dem der Prozess ablaufen soll sowie die Ziele des gesamten Prozesses definiert werden.

Der zweite Schritt beinhaltet die Sammlung der Daten. Hier soll herausgefunden werden, welche Daten zum Erreichen des Ziels benötigt werden und mit welchen zusätzlichen Daten diese Basis erweitert werden kann. Es sollen somit alle Daten gesammelt werden, die für den Prozess wichtig sind.

Schritt 3 beinhaltet die Vorbereitung sowie die Bereinigung der Daten. Das bedeutet, die Daten werden untersucht und zum Beispiel fehlende Werte ergänzt. Sofern Rauschen in der Datenbasis zu finden ist, also Daten welche die vorhandene Basis verfälschen, soll dieses entfernt werden.

Schritt 4 ist zur Datentransformation nötig. Hier werden die Daten für das Data Mining vorbereitet, wie die Auswahl geeigneter Eigenschaften (features) oder die Umwandlung der Attribute, wie die Standardisierung oder Normalisierung.

**Abbildung 1:** Der KDD-Prozess [41, S. 3]




Wie in Abbildung 1 erkennbar ist, beinhalten die Schritte 5, 6 und 7 das eigentliche Data Mining. Entsprechend der definierten Ziele, wie die Vorhersage von Dingen oder die Identifikation unbekannter Strukturen, wird in Schritt 5 das geeignete Data Mining Verfahren ausgewählt. Dabei handelt es sich beispielsweise um die Klassifikation, die Regression oder das Clustering.

In Schritt 6 wird der Data Mining Algorithmus ausgesucht, welcher zum Einsatz kommen soll. Dabei gibt es mehrere spezifische Algorithmen, die eingesetzt werden können. Zum Beispiel kann innerhalb der Klassifikation das neuronale Netzwerk oder Entscheidungsbäume wie der J48graft Algorithmus zum Einsatz kommen. Beim letzten Schritt innerhalb des Data Minings wird der Algorithmus auf die Daten angewendet. Hier wird der Algorithmus meist mehrmals mit verschiedenen Parametereinstellungen durchgeführt bis ein zufriedenstellendes Ergebnis erreicht ist.

Schritt 8 beinhaltet die Evaluation der erkannten Muster im Hinblick auf die definierten Ziele. Es wird erörtert, ob die Ergebnisse nützlich und nachvollziehbar sind. Um den Prozess abzuschließen wird im letzten Schritt das erhaltene Wissen benutzt, um das definierte Ziel zu erreichen. Beispielsweise kann das Wissen für Handlungen in der Zukunft bereitgestellt werden.

Innerhalb der Schritte kann jederzeit zu einem vorangegangenen Schritt zurückgesprungen werden. Dies ist dann nötig, wenn beispielsweise zusätzliche Daten zur Lösung des Problems notwendig sind oder eine veränderte Transformation für die Anwendung nötig ist.

Nach dem KDD-Prozess orientieren sich die Anwendungen zur Wissensfindung, die in dieser Arbeit vorgestellt werden. Dabei wird die Sammlung der Daten in Kapitel 4 beschrieben. In den Abschnitten des Kapitels 5 wird jeweils auf die Problemstellungen eingegangen, die Datenvorbereitung und -transformation beschrieben, das geeignete Data Mining Verfahren ausgewählt sowie geeignete Algorithmen selektiert und angewendet. Zum Abschluss jeden Abschnitts werden die Ergebnisse evaluiert und das Potenzial für einen Einsatz in der Realität eingeschätzt.

## 2.2 Data Mining Verfahren

Nachdem Data Mining im Allgemeinen und der KDD-Prozess im Speziellen beschrieben wurden, sollen nun die in dieser Arbeit verwendeten Data Mining Verfahren vorgestellt werden.

Insgesamt unterscheidet man beim Data Mining zwei Typen von Anwendungen. Einerseits ist dies die Anwendung, welche sich auf die Vorhersage von Werten konzentriert und andererseits die Anwendung, welche auf die Beschreibung von unbekannten Mustern in den Daten fokussiert ist. Eine zweite Einteilung der Anwendungen, die mit dieser Einteilung eng verwandt ist, ist die Einteilung in supervised und unsupervised learning. Beim supervised learning handelt es sich um die Methode, die Beziehungen zwischen einer oder mehreren Eingangsgrößen und einem Zielwert identifiziert [41, S. 7]. Der Zielwert ist dabei bekannt, weshalb die Anwendung als supervised bezeichnet wird. Oft wird mithilfe der Beschreibung der Beziehungen versucht Zielwerte von Daten deren Eingangsgrößen bekannt sind aber die Ausgangsgröße unbekannt ist zu ermitteln. Data Mining kann hier jedoch auch zur Identifikation von Mustern innerhalb der Daten hergenommen werden.

Beim unsupervised learning ist ein solcher Zielwert unbekannt. Hier geht es ausschließlich um das Auffinden von unbekanntem Wissen in den Daten. Ziel ist es innerhalb der Eingangsgrößen nützliche Strukturen bzw. Muster aufzudecken.

Die in den nächsten Abschnitten beschriebene Klassifikation und Regression gehören zur Klasse des supervised learnings, während die Clusteranalyse aus Abschnitt 2.2.2 unter das unsupervised learning fällt.

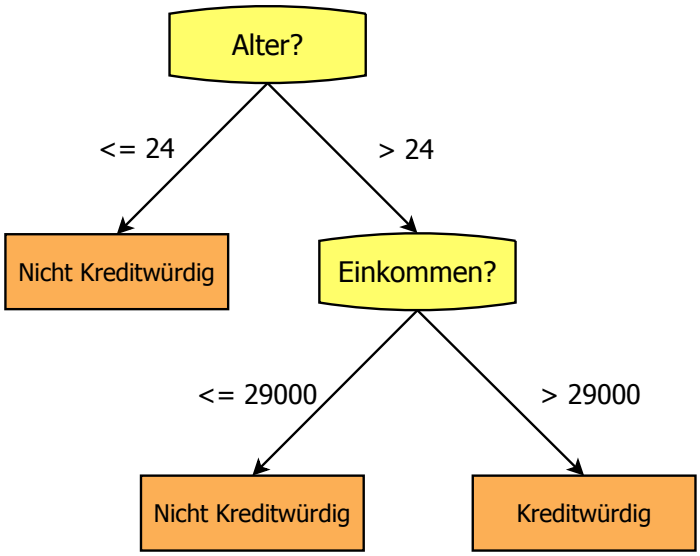
### 2.2.1 Klassifikation und Regression

Die Klassifikation und die Regressionsanalyse gehören zur Klasse der supervised Data Mining Verfahren. Hier wird anhand einer oder mehrerer unabhängiger Variablen (Eingangsgrößen) versucht die abhängige Variable (Ausgangsgröße) zu beschreiben. Die Einganggrößen sowie die Ausgangsgröße sind in diesem Fall bekannt. Beim supervised learning ist es typisch, dass ein Trainingsdatensatz zur Verfügung steht mit dem eine Beschreibung gefunden werden soll, welche imstande ist den Zielwert von ungesehenen Daten vorherzusagen [42, S. 150]. Ein Trainingsdatensatz besteht dabei aus einer Menge von Tupeln, hier auch Instanzen genannt, welche aus mehreren Attributen und ihrem Wertbereich bestehen und zusätzlich einen Zielwert besitzen. Formal ausgedruckt ist dies die Menge von Tupeln $B(A \cup y)$ mit den $n$ Attributen $A = \{a_1, .., a_i, .., a_n\}$ und dem Zielwert $y$ [42, S. 150].

#### 2.2.1.1 Klassifikation

Bei der Klassifikation wird der Zielwert $y$ auch als Klasse $c$ bezeichnet. Die Klassifikation lernt eine Funktion, welche eine Dateninstanz einer oder mehrerer vordefinierten Klassen zuordnet [16, S. 13]. Ein prominentes Beispiel für die Klassifikation kommt aus dem Bankenwesen. Ziel hierbei ist es, Personen anhand ihrer Eigenschaften, wie beispielsweise dem Alter und dem Einkommen in eine der Klassen {Kreditwürdig, Nicht Kreditwürdig} einzuordnen. Anhand der gesammelten Daten aus der Vergangenheit kann erlernt werden, ab welchem Alter und welchem Einkommen eine Person als kreditwürdig eingestuft wird. Ein Auszug aus einem erlernten Modell kann beispielsweise die Regel sein: WENN Alter $>$ 24 UND Einkommen $>$ 29.000 DANN „Kreditwürdig" SONST „Nicht Kreditwürdig". Anhand dieser Regel können zukünftige Kunden aufgrund ihres Alters und Einkommens in eine der zwei Klassen eingeteilt werden.

**Abbildung 2:** Beispiel eines Entscheidungsbaums




Innerhalb der Klassifikation gibt es mehrere Typen, die zur Anwendung kommen können. Bei dem gegebenen Bankenbeispiel handelt es sich um die regelbasierte Klassifikation. In dieser Arbeit kommen Algorithmen aus den folgenden Klassifikationsarten zum Einsatz:

- Entscheidungsbäume
- Regelbasierte Klassifikation
- Neuronale Netzwerke
- Naive-Bayes und Support Vektor Klassifikation

Nachfolgend werden die einzelnen Arten grob beschrieben. In dieser Arbeit soll nicht auf die detaillierte Betrachtung der Algorithmen eingegangen werden. Im Buch von Witten, Frank und Hall [67] sind mehr Informationen über die Algorithmen zu finden. In dem Buch liegt der Fokus auf den Algorithmen, welche in der Data Mining Software Weka (Waikato Environment for Knowledge Analysis) implementiert sind. Diese Software kommt in der vorliegenden Arbeit zum Einsatz.

**Entscheidungsbäume**

Bei Entscheidungsbäumen wird für die Klassifikation ein gerichteter Baum aufgespannt. Die Knoten eines Entscheidungsbaumes beinhalten jeweils einen Test auf den Wert eines Attributs des Datensatzes. Die Blätter des Baumes beinhalten die Klassen, in welche die behandelten Dateninstanzen eingeteilt werden. Abbildung 2 zeigt einen aufgestellten Baum aus dem erwähnten Beispiel der Kreditwürdigkeit von Bankkunden. Jede ungesehene Dateninstanz folgt anhand der Werte der Attribute einem Pfad bis zu einem Blatt. Das Blatt, bei dem die Instanz landet, ist die entsprechende kalkulierte Klasse für diese Instanz. Beispielsweise wird eine 27-jährige Person, die ein Einkommen von 45000 Euro hat, als „Kreditwürdig" klassifiziert. Da die Person älter als 24 ist, läuft sie in den rechten Pfad des Baumes. Bei dem Knoten „Einkommen?" nimmt sie wiederum den rechten Pfad, da das Einkommen höher als 29000 Euro ist. Somit landet die Instanz bei dem Blatt „Kreditwürdig". Die Person wird somit laut dem Entscheidungsbaum als „Kreditwürdig" klassifiziert.

Es gibt eine Vielzahl von Algorithmen die einen Entscheidungsbaum aufstellen. Bei der hier verwendeten Software Weka in der Version 3.6 sind insgesamt 16 Entscheidungsbaum-Algorithmen implementiert.

**Regelbasierte Klassifikation**

Nah verwandt zu den Entscheidungsbäumen sind die Algorithmen der regelbasierten Klassifikation. Regeln sind lediglich eine andere Darstellung für Entscheidungsbäume. Bei den regelbasierten Klassifikationsmethoden werden Regeln meist in Form von „WENN ... DANN" Regeln erstellt. Ein Beispiel für eine Regel ist:

$$\begin{aligned}&(Alter > 24)\ and\ (Einkommen > 29000) => Klasse = Kreditwürdig\\ &=> Klasse = Nicht\ Kreditwürdig\end{aligned} \tag{1}$$

Weka stellt insgesamt 11 verschiedene Implementierungen der regelbasierten Klassifikation bereit.

**Neuronale Netzwerke**

Die Implementierung neuronaler Netze in der Informatik orientiert sich an der Arbeitsweise der Gehirne von Menschen bzw. Tieren. Neuronale Netze sind informationsverarbeitende Systeme, dessen Neuronen sich Informationen über gerichtete Verbindungen zusenden [68, S. 23]. Verbundende Neuronen kommunizieren über den Grad ihrer Aktivierung. Es werden somit keine komplexen Datenstrukturen über die Verbindungen gesendet.

Neurone werden auch als Zellen, Units, Knoten oder Einheiten bezeichnet. Es gibt drei verschiedene Arten von Neuronen [51, S. 17]. Dies sind die Input-Units, welche von der Außenwelt Signale erhalten, die Hidden-Units die sich zwischen den anderen beiden Units befinden und die Ouput-Units, welche Signale an die Außenwelt ausgeben. Die Verbindungen der Units erfolgen durch gewichtete Kanten. Ein positives Gewicht bedeutet, dass ein Neuron auf das andere einen erregenden Einfluss besitzt. Ein negatives Gewicht signalisiert einen hemmenden Einfluss und ein Gewicht von Null übt keine Wirkung aus [51, S.18].

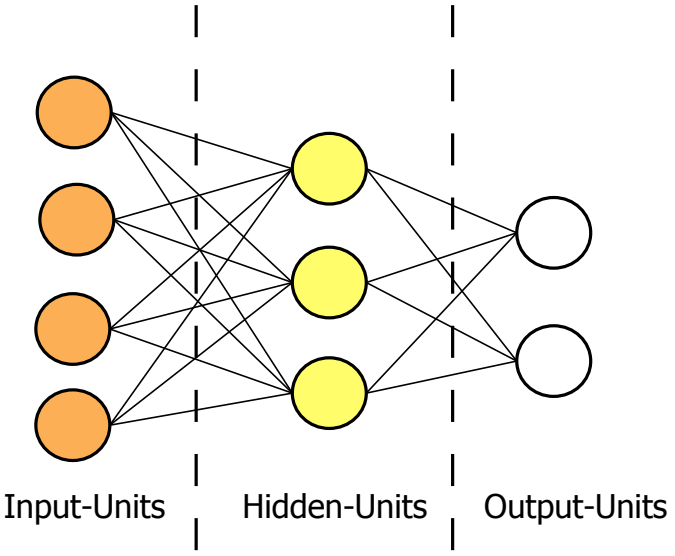
In der Lernphase werden dem neuronalen Netz die Trainingsdaten zur Verfügung gestellt. Mithilfe dieser Trainingsdaten werden die Gewichte der Verbindungen angepasst. Die Units selbst nehmen die einzelnen Inputwerte an, bilden daraus einen Netzinput, ordnen dem Netzinput ein Aktivitätslevel zu und erzeugen daraus einen Output [51, S.19].

Eine erweiterte Einführung in die neuronalen Netze geben Rey und Wender [51]. In Weka repräsentiert der MultilayerPereptron Algorithmus ein neuronales Netzwerk. Die MulitayerPerceptron Methode kann sowohl zur Klassifikation als auch zur Regression (Abschnitt 2.2.1.2) benutzt werden.

**Naive-Bayes und Support Vektor Klassifikation**

Die Naive-Bayes und Support Vektor Klassifikation sollen zur Vollständigkeit kurz erwähnt werden. Naives-Bayes ist eine Bayes-basierte Klassifikationsmethode. Diese Art der Methoden gehört zu den Probabilistisch Graphischen Modellen, welche eine Kombination aus der Graphentheorie und der Wahrscheinlichkeitstheorie bilden [55, S. 193f]. Die Naive-Bayes Methode basiert auf den Bayes' Regeln und nimmt naiverweise an, dass die Attribute voneinander unabhängig sind [67, S.93]. Obwohl diese Annahme in der Realität selten wahr ist, kann die Naive-Bayes Klassifikation oft gute Resultate erzielen. Weitere Informationen zur Bayes-Klassifikation geben Sebastiani, Abad und Ramoni [55].

**Abbildung 3:** Schema eines neuronalen Netzes in Anlehnung an [51, S.17]




Weka stellt 13 Bayes-basierte Klassifkationsmethoden bereit. In dieser Arbeit kommt lediglich die Implementierung des Naives-Bayes Algorithmus zum Einsatz.

Die in Weka implementierte Support Vektor Klassifikation basiert auf dem „sequential minimal optimization algorithm" von John C. Platt. Eine Beschreibung des Algorithmus gibt Platt [50]. Bei der Support Vektor Klassifikation wird jedes Objekt in einem Vektorraum repräsentiert. Ziel des Support Vektor Klassifizierers ist es eine Hyperbene zu finden, die diese Objekte in unterschiedliche Klassen trennt. Eine allgemeine Beschreibung der Support Vektor Klassifizierer liefert Shmilovici [56].

#### 2.2.1.2 Regression

Ziel der Regressionsanalyse ist es eine Funktion zu lernen, welche die Eingangsgrößen auf einen realen Wert abbilden. Die Regression soll die Beziehung zwischen einer oder mehreren unabhängigen Variablen und einer abhängigen Variable beschreiben.

In dieser Arbeit wird unter anderem die lineare Regression genutzt, dessen Funktion die Form

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ...\beta_n * X_n \tag{2}$$

mit dem abhängigen Zielwert $Y$, den unabhängigen Eingangsgrößen $X_1, X_2, ...X_n$ und den zu ermittelnden Koeffizienten $\beta_0, \beta_1, ...\beta_n$ hat. Durch die einfache Darstellung bietet sich die lineare Regression neben der Prognose des Zielwertes auch zur Beschreibung der Beziehungen zwischen mehreren Eingangsgrößen und einer Ausgangsgröße an.

Neben der linearen Regression kommt außerdem die von Weka implementierte SMOreg-Methode zum Einsatz. Dies ist die Implementierung der Support Vektor Maschine für die Regression. Ausführlicher beschreiben die SMOreg Methode Smola und Bernhard [57].

#### 2.2.1.3 Qualität der Vorhersage messen

Innerhalb des supervised Data Mining wird typischerweise zuerst anhand der verfügbaren Daten eine Modell erlernt, welches anschließend auf ungesehene Daten angewendet wird. Um die Qualität der Vorhersage eines erlernten Modells abzuschätzen, wird der verfügbare Datensatz in zwei Teile geteilt. Dies ist einerseits ein

Trainingsdatensatz, auf dem ein Algorithmus angewendet wird und ein Modell erlernt wird. Andererseits wird ein Testdatensatz erstellt, mit dem abgeschätzt werden soll, wie gut das Modell ungesehene Daten prognostiziert. Dies geschieht indem das erlernte Modell auf den Testdatensatz angewendet wird und die vorhergesagten Zielwerte bzw. Klassen mit den tatsächlichen Werten der Testmenge verglichen werden. So wird eine Fehlerrate für das erlernte Modell berechnet.

Bei der Unterteilung in Trainings- und Testdatensatz gibt es mehrere Möglichkeiten. Wenn genügend Daten zur Verfügung stehen kann ein bestimmter Prozentsatz der zur Verfügung stehenden Daten zum Training und ein bestimmter Prozentsatz zum Testen verwendet werden. Dabei ist der Trainingsdatensatz in der Regel größer als der Testdatensatz. Eine weitere Möglichkeit, die vor allem dann zum Einsatz kommt wenn nur wenige Daten zur Verfügung stehen, ist die sogenannte Methode der Crossvalidation. Hier wird der Datensatz in mehrere gleichgroße Teile, auch folds genannt, geteilt. Jede Teilmenge wird dabei einmal zum Testen genutzt und die restlichen Teile dienen dem Training des Modells. Dies wird für jede Teilmenge wiederholt. Legt man beispielsweise fest, dass zehn Teilmengen gebildet werden sollen, so wird der Datensatz in zehn Teile geteilt. Anschließend wird jeder Teil einmal als Testdatensatz genutzt und die restlichen Teile jeweils zum Training. Für jeden Testdatensatz wird der Fehler der Vorhersage berechnet. Die Fehlerraten der zehn Durchläufe werden anschließend gemittelt um eine Einschätzung der Qualität des erlernten Modells zu ermitteln [67, S. 153]. Das finale Modell wird mit dem gesamten Datensatz erlernt.

### 2.2.2 Clusteranalyse

Die Clusteranalyse gehört zum unsupervised Data Mining. Der Zielwert ist im Gegensatz zur Klassifikation oder Regression unbekannt. Dies bedeutet, dass bei der Clusteranalyse die Beschreibung der Daten im Vordergrund steht. Ziel der Clusteranalyse ist es die Objekte bzw. Instanzen in homogene Gruppen zu teilen. Dabei sollen ähnliche Objekte den gleichen Gruppen bzw. Clustern zugeordnet werden und unähnliche Objekte in verschiedene Cluster eingeteilt werden. Bei der Clusteranalyse ist man an der Aufteilung der Objekte interessiert und damit an den identifizierten Gruppen.

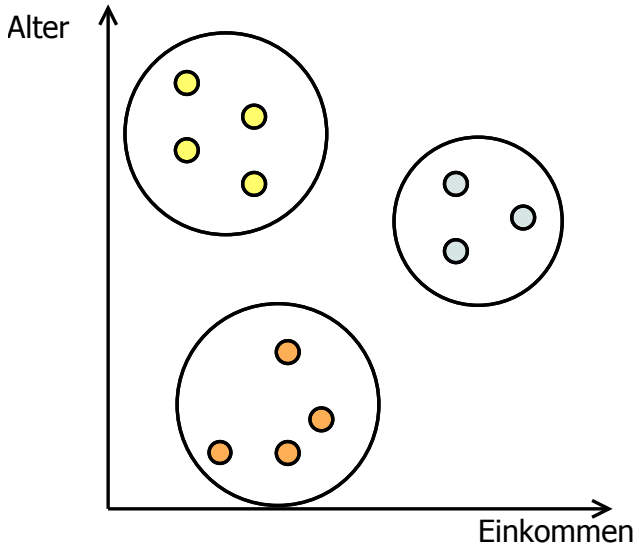
Stellt man sich die Objekte in einen Vektorraum vor, so nutzen viele Clustering Methoden die Distanz der Objekte um ähnliche bzw. unähnliche Objekte zu identifizieren. So sind Objekte mit einer geringen Distanz zueinander ähnlich. Je weiter die Objekte auseinander liegen, desto unähnlicher sind sie sich. Abbildung 4 zeigt dieses Beispiel in einem zweidimensionalen Raum. Hier werden drei Cluster identifiziert, die sich in ihren Eigenschaften ähneln. Dabei können je nach Methode die Objekte nur zu einem oder zu mehreren Clustern gehören.

Die Clusteranalyse wird oft im Marketing zur Marktsegmentierung eingesetzt. Ziel ist es beispielsweise mehrere Kundensegmente durch die Analyse des Kaufverhaltens der Kunden zu identifizieren, um so Gruppen von Kunden anhand ihrer Vorlieben gezielt anzusprechen.

Es gibt eine Vielzahl von Algorithmen für das Clustering, die sich in ihren Eigenschaften unterscheiden. So müssen bei manchen Methoden die Anzahl an Clustern, in denen die Objekte eingeteilt werden sollen, im Vorhinein angegeben werden. Der Clusteralgorithmus der in dieser Arbeit verwendet wird ist der EM-Algorithmus (Expectation-Maximization-Algorithmus), bei dem die Clusteranzahl nicht vordefiniert sein muss. Der EM-Algorithmus ist eine Implementierung des probabilistischen Clusterings.

Neben dem EM-Algorithmus hat Weka zusätzliche 15 Clusteralgorithmen implementiert. Weitere Beschreibungen zur Clusteranalyse geben Maimon und Rokach [40] sowie Witten, Frank und Hall [67].

**Abbildung 4:** Beispiel Clusteranalyse




### 2.2.3 Weitere Data Mining Verfahren

Weitere Data Mining Verfahren beinhalten unter anderem die Assoziationsanalyse oder die soziale Netzwerkanalyse, welche in dieser Arbeit nicht zum Einsatz kommen. Bei der Assoziationsanalyse wird nach Zusammenhängen in den Daten gesucht. Beispielsweise kann anhand einer Warenkorbanalyse ermittelt werden, welche Artikel oft zusammen in einem Warenkorb auftauchen. Mithilfe dieses Wissens lassen sich Kaufempfehlungen für Kunden kreieren oder die Platzierung von Artikeln im Handel beeinflussen.

Die soziale Netzwerkanalyse findet Anwendung in der Analyse von Beziehungen von Individuen. Beispielsweise können soziale Netzwerke analysiert werden, um Individuen zu identifizieren, welche eine hohe Autorität in einem Netzwerk innehaben. Ein verwandtes Feld ist die Link Analyse, bei denen die ein- und ausgehenden Links einer Webseite analysiert werden. So kann beispielsweise die Linkstruktur des Webs abgebildet werden. Eine Art der Link Analyse nutzt der Suchalgorithmus von Google. Der Algorithmus nutzt die Linkstruktur des Webs um die Resultatliste von Suchanfragen zu ordnen.

## 2.3 Verwendete Software

Für die Anwendung von Data Mining wird in dieser Arbeit die Software „Weka 3: Data Mining Software in Java" benutzt. Weka ist eine open source Software und bietet verschiedene Data Mining Algorithmen zur Anwendung an. Weka stellt Werkzeuge zur Datenvorbereitung, Klassifikation, Regression, Clusteranalyse, Assoziationsanalyse und zur Visualisierung bereit. Die Software Webseite ist unter `cs.waikato.ac.nz/ml/weka/` zu finden.

Weka steht für „Waikato Environment for Knowledge Analysis" und wurde von der Universität von Waikato in der Programmiersprache Java entwickelt. Die Software kann einerseits mithilfe einer grafischen Benutzeroberfläche benutzt werden oder durch eigenen Java Code aufgerufen werden. In dieser Arbeit wird ausschließlich mit der grafischen Benutzeroberfläche gearbeitet. Die hier verwendete Version von Weka ist die Version 3.6.8.

Neben der Data Mining Software wird in dieser Arbeit das open source relationale Datenbankverwaltungssysteme MySQL zur Datenhaltung genutzt. Weka bietet

die Möglichkeit durch eine Schnittstelle auf die Datenbank zuzugreifen. Mithilfe des Oracle SQL Developers in der Version 3.2.10.09 wird die Datenbank verwaltet. Mittels dieser Software werden u.a. Tabellen angelegt, Daten restrukturiert und eigene einfache Analysen durchgeführt. Die Sammlung der Daten aus dem Internet wird mit dem HTTrack Website Copier in der Version 3.46 und der Programmiersprache Python (Version 3.2) verwirklicht. Nähere Informationen hierzu finden sich in Kapitel 4.

# 3 Stand der Forschung

In diesem Kapitel wird der Stand der Forschung in der Analyse von Sportdaten dargestellt. Bei der Suche nach geeigneter Literatur fällt auf, dass im deutschsprachigen Raum wenige bis keine Veröffentlichungen über die Analyse von Sportdaten, besonders unter der Berücksichtigung des Bereichs Data Mining vorhanden sind. Die Datenanalyse innerhalb des Sports ist vor allem in Amerika weit verbreitet. Ein Grund dafür ist das Aufkommen einer neuen Ära der Statistiken in der von Amerikanern begehrten Sportart Baseball. Statt die reinen Zahlen, wie die Anzahl geschlagener oder gefangener Bälle zu analysieren, wurden diese Statistiken von dem amerikanischen Historiker und Statistiker Bill James hinterfragt und neue Analysen erfunden. Diese „Revolution" ist unter dem Namen Sabermetrics bekannt und werden von Schumaker, Solieman und Chen [54, S. 36] als der Wandel von traditionellen Statistiken hin zum Wissensmanagement (knowledge management) beschrieben. Bill James hat die reine Aufstellung der Zahlen in Frage gestellt und neue Maßzahlen erfunden, welche die Spieler objektiver bewerten sollten. Statt nur Teilaspekte, wie die Trefferquote zu betrachten, wird bei Sabermetrics versucht Maßzahlen zu erfinden, welche die Gesamtleistung eines Spielers bewerten.

Nachdem Bill James 1977 die ersten Analysen veröffentlicht hat, haben die Sportfans die Ideen aufgenommen und eigene Maßzahlen zu Sabermetrics beigesteuert. Vorerst haben die Maßzahlen nur wenig Anwendung bei Vereinen gefunden bis der Manager des Proficlubs Oakland A's Billy Bean im Jahr 2002 die Maßzahlen von Sabermetrics genutzt hat, um neue Spieler für seine Mannschaft auszuwählen. Ergebnis war, dass der Club für seine Verhältnisse sehr großen Erfolg durch die Adaption der Maßzahlen erfahren hat. Inspiriert von diesem Erfolg, konnten die Boston Red Sox mit der Hilfe von Sabermetrics ihre Mannschaft sogar so stark verstärken, dass sie 2004 und 2007 die Meisterschaft gewannen. [54, S. 36]

Ebenso wie bei der Sabermetrics Revolution, startete 1980 der Statistiker Dean Oliver eine Revolution in der Sportart Basketball. Dean Oliver stellte ebenfalls die alten Messwerte zur Beurteilung der Leistung von Spielern in Frage und erfand neue Maßzahlen. Mithilfe der Beratung von Dean Oliver konnten die Seattle Supersonics 2005 die Basketball Division der USA gewinnen [54, S. 36].

Ein Grund, warum diese beiden Sportarten und andere amerikanische Sportarten, wie z.B. Football einen großen Analysehintergrund haben, liegt in der besseren Möglichkeit die Spiele dieser Sportarten zu quantifizieren. Beim Baseball beispielsweise werden die Spielaktionen hintereinander ausgeführt und können gut in einzelne Teile zerlegt werden. Dadurch lassen sich die Aktionen dieser Sportart leichter manuell aufnehmen. Beim Basketball ist ein Spiel nicht in seine Einzelteile zerlegbar. Jedoch sind aufgrund der vielen Ereignisse im Basketball, die aufgenommen Daten sehr gehaltvoll. Im Basketball werden in einem Spiel bis zu 100 Punkte pro Mannschaft erzielt, Spieler werfen sehr häufig auf den Korb oder die Mannschaftsaufstellung wechselt ständig während des Spiels. Somit sind hier mehr Daten manuell aufnehmbar, als es im Fußball der Fall ist.

Im Gegensatz dazu ist ein Fußballspiel ereignisarm. Mithilfe neuer technologischer Möglichkeiten und der Professionalisierung der Datenaufnahme, kann der Fußball jedoch mit der Aufnahme von Werten wie beispielsweise der Laufleistung inzwischen auch eine große Anzahl an Daten aufbereiten. Die Grundlage für eine statistische Revolution im Fußball ist inzwischen gelegt.

Ein Ansatz für die Einbindung von Fußballfans und Analytikern in die Analyse von Fußballdaten wurde von dem englischen Premier League Club Manchester City verfolgt. Diese haben in Zusammenarbeit mit der Firma Opta, welche Sportdaten pro-

fessionell aufnimmt und verkauft, die Daten der Premier League Saison 2011/2012 frei im Internet zur Verfügung gestellt. Die Hoffnung war, dass die Nutzer der Daten neue Maßzahlen oder Analyseansätze veröffentlichen, die für Vereine oder Fans hilfreich sein können. Dabei wurden zahlreiche Analysen der Daten in verschiedenen Blogs über Fußball und deren Taktiken veröffentlicht.

Innerhalb des Fußballs gibt es inzwischen verschiedene Indexe, die Fußballspielern anhand der aufgenommenen Daten, wie Tore, Vorlagen oder der Zweikampfquote einen Wert zuteilen. Solche Indexe sollen die Qualität der Spieler messen. Es gibt dabei unter anderem den vom früheren englischen und jetzigen russischen Nationaltrainers Fabio Capello erstellten Capello Index [1], das Castrol EDGE Ranking [2] oder der Opta Index der auch von der Webseite `bundesliga.de` genutzt wird [5]. Die genaue Berechnung dieser Indexe ist nicht bekannt. Beim Capello Index handelt es sich beispielsweise um einen per Hand erstellten Index. Solche Indexe sind kritisch zu betrachten, da den subjektiv als wichtig empfundenen Ereignissen in einem Spiel, wie beispielsweise den geschossenen Toren die höchste Aufmerksamkeit bei der Berechnung solcher Indexe beigemessen werden und andere Spielaktionen nahezu ignoriert werden.

Im Weiteren werden verschiedene ausgewählte Veröffentlichungen vorgestellt, um einen Einblick in den Stand der Forschung zu gewähren. Dabei wird besonderer Fokus auf Data Mining Verfahren gelegt. Außerdem werden Analysen berücksichtigt, die spielspezifische Attribute von Spielern bzw. Mannschaften in ihre Analysen aufnehmen. Dabei ist dies keine vollständige Liste der veröffentlichten Artikel. Schumaker, Solieman und Chen [54] stellen einen ausführlicheren Einblick in den derzeitigen Stand der komplexen Datenanalyse im Sport zur Verfügung.

Einige Anwendungen der einfachen und komplexen Datenanalyse im Fußball werden im folgenden Abschnitt 3.1 beschrieben, die nicht mit den Analysen der vorliegenden Arbeit in Verbindung stehen, aber zur Vollständigkeit erwähnt werden. In den vorgestellten Forschungen wird nicht auf die spielspezifischen Attribute von Spielern eingegangen.

In den weiteren Abschnitten werden Veröffentlichungen vorgestellt, die Methoden des Data Mining nutzen, um Daten im Sport zu analysieren. Dabei nutzt der Artikel aus Abschnitt 3.2 die Regressionsanalyse, um die Erfolgsfaktoren einer Saison bzw. Liga zu ermitteln sowie die Varianzanalyse um die Top-Teams von den restlichen Teams zu unterscheiden. Abschnitt 3.3 beschäftigt sich mit einem Artikel, der die Clusteranalyse nutzt, um Eishockeyspieler in Kategorien einzuteilen. Der Artikel aus Abschnitt 3.4 beschreibt einen Ansatz, der Trainern bei der Entscheidung für die perfekten Zeitpunkte bei Ein- und Auswechslungen unterstützen soll. Dazu nutzt der Autor einen Entscheidungsbaum. Neuronale Netzwerke kommen in den Artikeln aus Abschnitt 3.5 zum Einsatz. Hier sollen die Ausgänge von Basketballspielen mithilfe der spielspezifischen Attribute der Mannschaften prognostiziert werden. Der Artikel aus Abschnitt 3.6 kombiniert die Cluster- und Regressionsanalyse, um den Wert einzelner Eishockeyspieler zu messen.

## 3.1 Anwendungen der Datenanalyse im Fußball

Eine vielversprechende Anwendung von Data Mining im Sport findet bei dem italienischen Serie A Verein AC Mailand statt. Hier werden neuronale Netzwerke in der medizinischen Abteilung des Vereins eingesetzt [43]. Wie Kuper [36] beschreibt wurde aufgrund einer Verletzung eines für 30 Millionen Euro eingekauften Spielers von

[1] `www.capelloindex.com`

[2] `www.castrolfootball.com/rankings`

AC Mailand, das sogenannte Milan Lab eingeführt, welches mithilfe modernster Methoden versucht, die Verletzungen von Spielern zu vermeiden. Idee ist es mögliche Verletzungen vorherzusagen, um so die Spieler vor dem Eintritt einer Schädigung zu warnen. Laut dem Artikel von Flinders [18] konnte das eingesetzte System im Jahr 2002 in über 70 % der Fälle Verletzungen aus der Vergangenheit vorhersagen. Kuper [36] beschreibt, dass alleine anhand von Sprunganalysen, die medizinische Abteilung 70 % der Verletzungen prognostizieren kann. Mithilfe mehrerer Daten konnten die nichttraumatischen Verletzungen, also Verletzungen ohne Außeneinwirkung, im Jahr 2008 zu den Jahren zuvor um 90 % reduziert werden.

Auch bei Neuverpflichtungen werden die Mitarbeiter des Milan Labs konsultiert. So studieren sie die Bewegungsabläufe eines Spielers, um vorherzusagen wie verletzungsanfällig er ist. Damit kann eine Entscheidung zum Einkauf des Spielers maßgeblich beeinflusst werden. Inzwischen werden die Analysen auch zur Verbesserung der physischen Verfassung der Sportler genutzt. Dabei nutzen die Fitnesstrainer des Vereins die Daten, um die Spieler gezielt zu verbessern. Eine genaue Beschreibung der Analysen ist nicht veröffentlicht.

Bei der WM 2006 in Deutschland zog ein Zettel die Aufmerksamkeit der Fußballfans auf sich. Dieser wurde dem deutschen Torhüter Jens Lehmann vor dem Elfmeterschießen gegen Argentinien im Viertelfinale ausgehändigt. Darauf sollen die Elfmeterschützen der Argentinier und ihre bevorzugte Ecke, in welche sie den Elfmeter schießen, aufgelistet gewesen sein [58]. In dem Buch von Kuper und Szymanski [37], welches sich mit Statistiken aus dem Bereich Fußball beschäftigt, gibt es ein Kapitel welches sich ausschließlich mit der Statistik von Elfmeterschießen beschäftigt. In dem Kapitel „The Economist's Fear of the Penalty Kick" wird das Elfmeterschießen aus Sicht der Spieltheorie erörtert. Es wird hinterfragt, ob eine Handlungsempfehlung für Schützen oder Torhüter gegeben werden kann. Die Ergebnisse, die in dem Kapitel vorgestellt werden zeigen, dass die Analyse von Elfmeterschießen sinnvoll angewendet werden kann. Beispielsweise wird in dem Kapitel das Elfmeterschießen des Champions League Finales 2008 zwischen dem FC Chelsea und Manchester United untersucht und interessante Erkenntnisse gewonnen. Eine Analyse mithilfe von Data Mining Verfahren könnte an diese Analysen angeschlossen werden.

Im Gegensatz zu den Artikeln aus Abschnitt 3.5, welche die spielspezifischen Attribute von Spielern zur Vorhersage nutzen, verwenden Rotshtein, Posner und Rakityanskaya [52], Awerbuch [1] oder Dyte und Clarke [15] nur die reinen Ergebnisse und andere leicht zur Verfügung stehenden Daten zur Prognose von Fußballspielen.

Rotshtein, Posner und Rakityanskaya [52] nutzen zur Vorhersage eines Spiels die Ergebnisse der letzten fünf Spiele einer Mannschaft und des Gegners sowie die Resultate der letzten zwei Spiele, bei denen die Mannschaften gegeneinander angetreten sind. Als Zielwert soll das Vorhersagemodell, welches auf der Fuzzylogik basiert, die Höhe des Sieges oder der Niederlage ausgeben. Dabei bezeichnen die Autoren einen hohen Sieg bzw. Niederlage als ein Spiel, bei dem die Tordifferenz mindestens drei Tore beträgt. Ein weniger hoher Sieg oder Niederlage liegt vor, wenn die Tordifferenz beider Mannschaften zwischen eins und zwei liegt. Außerdem soll das Modell in der Lage sein ein Unentschieden vorherzusagen.

Es werden ein genetischer Algorithmus sowie ein neuronales Netzwerk zur Lösung des aufgestellten nicht-linearen Optimierungsproblems genutzt [52, S. 623]. Als Trainingsdatensatz liegen 1056 Spiele der Jahre 1994 bis 2011 aus der finnischen Liga vor. Wie gut die Spiele vorhergesagt werden können, wird mithilfe von Spielen aus den Jahren 1991 bis 1993 ermittelt. Dies sind 350 Spiele.

Laut Rotshtein, Posner und Rakityanskaya werden mithilfe des neuronalen Netzwerks 304 Spiele richtig vorhergesagt. Dies bedeutet eine Klassifikationsrate von 87 %,

was eine sehr hohe Trefferquote darstellt. Damit kann dieses Modell bessere Werte erzielen als die Artikel aus Abschnitt 3.5 über die Sportart Basketball. Ob eine solch hohe Klassifikationsrate tatsächlich zu erreichen ist, können weitere Forschung mit dem vorgestellten Ansatz offenlegen.

Die Autoren sehen den Einsatz für richtige Vorhersagen jedoch kritisch, da das Modell weder Verletzungen von Spielern noch die Anzahl an Spielern auf dem Feld und auf der Ersatzbank einschließt. Außerdem werden die Objektivität des Schiedsrichters und die Wetterbedingungen in der Analyse nicht betrachtet. [52, S. 628]

Es gibt weitere Analysen im Fußball, die einfache Daten, wie das reine Ergebnis oder die Anzahl an geschossenen Tore betrachten. So versuchen Van Calster, Tim und Van Huffel [60] mithilfe des Anteils von Unentschieden gespielten Spielen, geschossenen und erhaltenen Toren pro Spiel, Anzahl aller Tore pro Spiel und erhaltene Punkte pro Spiel zu erklären, wie torlose Unentschieden im Fußball zustande kommen. Dabei ist offensichtlich, dass die Anzahl aller Tore pro Spiel mit den torlosen Unentschieden stark verknüpft sind. Außerdem stehen die erhaltenen Punkte pro Spiel mit den Unentschieden in Verbindung.

Weitere Analysen von Sportdaten beschreiben Schumaker, Solieman und Chen [54].

## 3.2 Differenzierung der Spielweisen von Fußballmannschaften

Oberstone [47, 48] nutzt die lineare Regression zur Beschreibung der Zusammenhänge zwischen verschiedenen spielspezifischen Attributen und der erspielten Punktzahl einer Mannschaft. Daneben wird die Varianzanalyse (ANOVA von analysis of variance) genutzt, um Unterschiede in den Spielweisen von Mannschaften und Ligen zu erforschen.

Oberstone [48] beschäftigt sich mit der Analyse der englischen Premier League Saison 2007/2008. Einerseits versucht der Autor mithilfe der Regressionsanalyse die Attribute herauszufinden, welche maßgeblich zum Erfolg eines Teams in der Saison beigetragen haben. Andererseits wird mithilfe der Varianzanalyse versucht, die Unterschiede der Spielweisen von erfolgreichen Vereinen zu den Spielweisen der anderen Mannschaften zu ermitteln.

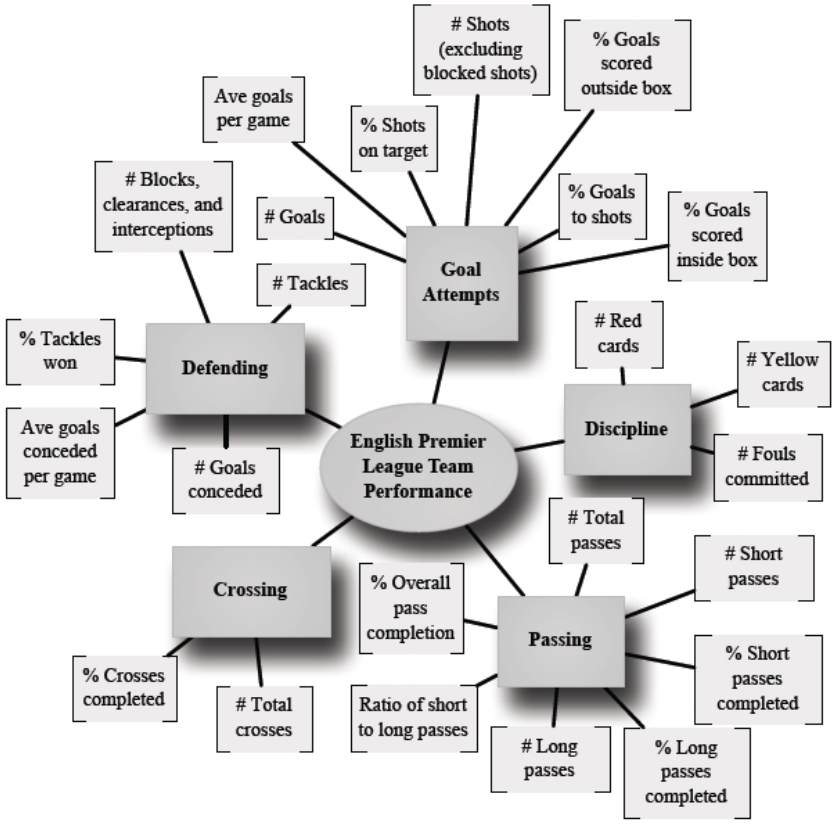
Für die Analysen liegen dem Autor die Mannschaftsdaten der Saison 2007/2008 zur Verfügung. Dabei handelt es sich um die in Abbildung 5 dargestellten Attribute, welche von der Firma Opta bereitgestellt werden.

Die Daten werden dabei vom Autor Oberstone in fünf verschieden Kategorien eingeteilt, nämlich in Torversuche (Goal Attempts), Passspiel (Passing), Flanken (Crossing), und Verteidigung (Defending). Insgesamt besteht der Datensatz aus 24 Attributen, wobei sieben dieser Attribute redundante Informationen speichern und vom Autor aus der Analyse ausgeschlossen werden. Anschließend werden von diesen 17 Attributen, sukzessive Attribute entfernt bis der Datensatz für die Regression ausschließlich aus statistisch signifikanten Attributen besteht. Dadurch werden die 17 Attribute auf folgende sechs Attribute reduziert [48, S. 7]:

- $X_5$ : Erfolgreiche Torschussquote in Prozent (% goals to shots)
- $X_7$ : Tore außerhalb des Strafraums in Prozent (% goals scored outside of box)
- $X_{11}$ : Verhältnis kurzer zu langer Pässe (ratio of short/long passes)
- $X_{15}$ : Anzahl an Flanken (total crosses)
- $X_{18}$ : Durchschnittliche Anzahl an zugelassenen Toren pro Spiel (average goals conceded per game)
- $X_{23}$ : Anzahl gelber Karten (yellow cards)

Mithilfe dieser Werte als Eingangsgrößen und dem Zielwert, der erspielten Punktzahl der beobachteten Mannschaft am Ende der Saison 2007/2008, wird in dem Artikel eine Regressionsanalyse durchgeführt, um die Beziehungen zwischen den Werten

**Abbildung 5:** Opta Attribute [48, S. 6]




und der Punktzahl zu ermitteln. Die resultierende Formel der Regression lautet [48, S.10]:

$$\begin{aligned} Y =& 30,999 + 99,231X_5 + 80,159X_7 + 4,471X_{11} \\ &+ 0,029X_{15} - 31,708X_{18} - 0,161X_{23} \end{aligned} \tag{3}$$

Anhand der Koeffizienten der unabhängigen Variablen, kann eine Aussage über die Erfolgsfaktoren einer Mannschaft getroffen werden.

Die Ergebnisse zeigen, dass eine gute Torquote, viele erzielte Treffer von außerhalb des Strafraums, ein hoher Anteil kurzer zu langer Pässe und viele Flanken die Punktzahl einer Mannschaft positiv beeinflussen. Zusätzlich sollten Mannschaften versuchen möglichst wenige Tore zuzulassen. Ebenfalls sollten sie eine faire Spielweise bevorzugen, da so gelbe Karten verhindert werden können.

In einem weiteren Artikel führt Oberstone [47] die gleichen Regressionsanalysen für die Saison 2008/2009 in den Ligen der englischen Premier League, der spanischen La Liga und der italienischen Serie A durch. Dabei wird für jede Liga eine eigene Regression durchgeführt. Die Ergebnisse zeigen, dass die Erfolgsfaktoren in den einzelnen Ligen variieren. Außerdem unterscheidet sich die erlernte Formel aus der Premier League Saison 2008/2009 wesentlich zu der oben genannten Formel aus der Saison 2007/2008.

Zur Analyse der Unterschiede in den Spielweisen erfolgreicher Mannschaften zu den restlichen Teams der Liga, nutzt der Autor die Varianzanalyse. Die Anwendung dieser Analyse für die vorliegenden Daten ist kritisch zu betrachten, da die Varianzanalyse eine Varianzhomogenität und eine Normalverteilung voraussetzt [53, S. 381]. Ob dies in den Daten des Artikels der Fall ist, wird vom Autor nicht erwähnt. In den Daten der hier vorliegenden zwei Bundesligasaisons ist eine Normalverteilung nicht

gegeben.

Um die Varianzanalyse anzuwenden, werden alle 24 Attribute des Datensatzes genutzt. Als Top-Teams definiert der Autor die vier bestplatzierten Mannschaften, als schlechteste Teams die vier letzten Teams der Tabelle und die restlichen zwölf Mannschaften als mittelmäßige Teams. Mithilfe der Varianzanalyse versucht der Autor herauszufinden, ob sich die Werte der verschiedenen Attribute der drei Gruppen signifikant unterscheiden.

Die Analyse findet heraus, dass sich die Mannschaften in 13 Attributen statistisch signifikant unterscheiden. Die detaillierten Ergebnisse stellt Oberstone [48, S. 16] dar. Hier zeigt sich vor allem, dass die Top-Teams:

- mehr Torchancen kreieren
- öfter und präziser Passen
- weniger lange Pässe favorisieren
- länger im Ballbesitz sind
- mehr erfolgreiche Zweikämpfe führen
- weniger Fouls begehen.

In dem zweiten Artikel von Oberstone, vergleicht der Autor mithilfe derselben Methode die Unterschiede in den drei Ligen aus England, Spanien und Italien. Dabei ergeben sich Unterschiede u.a. in der Anzahl an Torschüssen, in der Torquote, in der Passgenauigkeit, im Zweikampfverhalten, bei den Fouls und in der Anzahl an gelben und roten Karten. Eine Aufbereitung der Ergebnisse gibt Oberstone [47, S. 11f].

Die Ergebnisse der beiden Artikel unter Zuhilfenahme der gleichen Methoden geben eine interessante Beschreibung der Spielweisen der einzelnen Mannschaften und Ligen wieder. Inwiefern die dargestellten Erkenntnisse Vereine helfen können bleibt vom Autor unbeantwortet. Bei den ermittelten Formeln handelt es sich um Formeln, die sehr angepasst (overfitted) an die gegebenen Daten sind. Das bedeutet, dass die Formeln lediglich die vorgestellten Saisons sehr gut beschreiben, andere Saisons jedoch nur gering repräsentieren. Zu einer Vorhersage sind die Formeln demnach ungeeignet. Ob sich Handlungsempfehlungen für Vereine aus den extrahierten Mustern generieren lassen ist demnach kritisch zu betrachten.

Die Resultate bei der Unterscheidung von Top-Teams und den Ligen mithilfe der Varianzanalyse zeigen, wie sich die Spielweisen der besten Mannschaften von denen der restlichen Mannschaften abheben. Dadurch können Vereine ihre Spielweise kritisch untersuchen und gegebenenfalls ihren Spielstil in die Richtung erfolgreicher Mannschaften justieren. Unter Zuhilfenahme weiterer Saisons und Attribute könnte eine solche Art von Analyse weitere und detaillierte Ergebnisse hervorbringen. In Kapitel 5.3 wird ein Versuch der Differenzierung der Bundesligavereine unter Zuhilfenahme eines weiteren Data Mining Verfahrens unternommen. Hier wird die Clusteranalyse benutzt, um die Spielweisen der Teams zu unterscheiden und eine Kategorisierung vorzunehmen. Die Ergebnisse des Clusterings werden mit den Tabellenplatzierungen der Mannschaften in Beziehung gesetzt, um zu testen, ob sich die Spielweisen erfolgreicher Teams ähneln und sich zu den restlichen Teams abgrenzen.

### 3.3 Clusteranalyse zur Kategorisierung von Spielern

Vincent und Eastman [61] nutzen die Clusteranalyse, um Eishockeyspieler der National Hockey League (NHL) anhand ihrer Spielweisen zu kategorisieren. Die Idee hinter der Analyse ist es die Kategorien, welche von Fans und Kommentatoren anhand ihrer subjektiven Wahrnehmung definiert werden, mithilfe des Clusterings von Spielern zu hinterfragen. Beispielsweise teilt die Öffentlichkeit die Spieler unter anderem in die Kategorien „Grinder" und „Enforcer" ein. Dabei weisen „Grinder" eher defensive Qualitäten auf. Spieler der Kategorie „Enforcer" halten die Gegner mit Körpereinsatz

von den eigenen Angreifern fern, damit diese frei zum Torabschluss kommen können [61, S. 1]. Ob sich eine solche Einteilung mit den objektiv erhobenen Daten deckt, wird von Vincent und Eastman [61] überprüft.

Den Autoren stehen dabei insgesamt Daten von 625 Spielern zur Verfügung. Dabei werden die folgenden Attribute behandelt [61, S. 3ff]:

- Punkte pro Spiel
- Strafminuten pro Spiel
- Plus-Minus Statistik pro Spiel
- Gewicht

Die Autoren betrachten nicht nur die „pro Spiel" Statistik, sondern führen auch eine Analyse für die Attribute „pro Minute" aus. Die Plus-Minus Statistik ist die Differenz von geschossenen und erhaltenen Toren der Mannschaft eines Spielers, während der entsprechende Spieler gespielt hat. Tore und Gegentore während Strafminuten werden nicht mitgerechnet. Das Gewicht wird vom Autor hinzugefügt, da die physischen Eigenschaften beim Eishockey eine wichtige Rolle spielen, um Gegenspieler zu schwächen und das eigene Spiel positiv zu beeinflussen. Außerdem werden nur Spieler betrachtet, die mindestens zehn Spiele in der NHL gespielt haben und für welche Informationen über das Gehalt zur Verfügung stehen.

Um die Clusteranalyse durchzuführen nutzen die Autoren die K-Means Methode. Als Ähnlichkeitsmaß wird die euklidische Distanz gewählt. Da bei K-Means die Anzahl an Cluster voreingestellt werden muss, wird mithilfe des Calinski-Harabsz pseudo-F Indexes, welcher den Anteil der Quadratsummen zwischen den Clustern und die Quadratsummen innerhalb der Cluster berechnet, die optimale Clusteranzahl ausgewählt. Dabei soll der Wert des pseudo Indexes maximiert werden. Außerdem werden die Werte der Spieler standardisiert, damit kein Attribut beim Clustering dominiert. Zusätzlich werden Extremwerte aus den Daten gelöscht. [61, S. 3ff] Es werden jeweils separate Analysen für die offensiven und defensiven Spieler durchgeführt.

Innerhalb der Clusteranalyse für die Offensive, werden jeweils drei Cluster identifiziert. Die Cluster werden von den Autoren als „Scorer", „Enforcer" und „Grinder" kategorisiert. Bei der „pro Spiel" Analyse zeigt sich in den Mittelwerten der einzelnen Cluster, dass die Spieler der Kategorie „Scorer" am meisten Tore schießen und die höchste Plus-Minus Statistik aufweisen. Am aggressivsten agieren die Spieler der Kategorie „Enforcer", welche auch die Spieler mit dem höchsten Gewicht sind. Die dritte Kategorie sind die „Grinder", welche am leichtesten sind, wenige Strafminuten erhalten und mittlere Werte in der Anzahl an Toren und in der Plus-Minus Statistik aufweisen. Weitere Analysen zeigen, dass die „Scorer" im Schnitt die höchste Einsatzzeit haben, die „Grinder" am zweitmeisten und die „Enforcer" am wenigsten. Die „Scorer" verdienen mit 3,9 Millionen Dollar pro Jahr wesentlich mehr als die „Grinder" und „Enforcer", welche im Schnitt 1 Millionen Dollar pro Jahr verdienen. Die Unterschiede zur Analyse des „pro Minute" Datensatzes sind sehr gering und werden im Folgenden nicht weiter betrachtet.

Die Clusteranalyse der defensiven Spieler ermittelt zwei Cluster. Einerseits sind dies die „Scorer", welche viele Tore schießen und eine höhere Plus-Minus Statistik besitzen. Andererseits wird die Spielerkategorie „Agressors" identifiziert. Die Spieler dieses Clusters besitzen eine höhere Anzahl an Strafminuten und sind schwerer. Auch hier sind die beiden Analysen der zwei Datensätze sehr ähnlich. Ein „Scorer" der Defensive verdient mit 2,2 Millionen Dollar pro Jahr fast doppelt so viel wie ein „Agressor".

Die Analyse des Artikels zeigt, dass Spieler in der Offensive in drei Kategorien und in der Defensive in zwei Kategorien eingeteilt werden. Dabei zeigt sich, dass diese Kategorien mit den Kategorien, die subjektiv Wahrgenommen werden im Allgemei-

nen übereinstimmen. Da die Autoren die Spielerkategorien mit dem Gehalt in Beziehung setzen, kann die Analyse Verantwortlichen von Vereinen als Überprüfung des Gehaltsgefüge der eigenen Mannschaft dienen.

Kritisch zu betrachten ist, dass nur wenige Attribute für die Analyse zur Verfügung stehen. So können detaillierte Informationen, zu genaueren Spielerkategorien führen. In Kapitel 5.3 dieser Arbeit wird eine Clusteranalyse mit den Spielern der Bundesliga durchgeführt. Dabei sind mehr Attribute vorhanden, als dies in diesem Artikel der Fall ist. Die Analyse zeigt eine höhere Differenzierung der Spielweisen. Solche Analysen können sehr gut zum Scouting genutzt werden, da es beim Scouting meist von Interesse ist Spieler zu kaufen, welche abgegebene Spieler optimal ersetzen. Durch die Clusteranalyse lassen sich leicht Spieler aus gleichen Spielerkategorien identifizieren. Ein weiterer Artikel, der die Clusteranalyse nutzt ist in Abschnitt 3.6 beschrieben.

## 3.4 Der beste Zeitpunkt für eine Einwechslung

Ein interessanter Ansatz zur Anwendung von Data Mining im Fußball bietet Myers [45]. In diesem Artikel werden zwar keine spielspezifischen Attribute von Spielern bzw. Mannschaften behandelt, da hier jedoch die Entscheidungsbäume als Data Mining Methode eingesetzt werden, soll der Artikel zur Vollständigkeit trotzdem beschrieben werden. Unter Verwendung der Klassifikation wird in dieser Analyse eine Strategie ermittelt, welche den Trainern helfen kann die Zeitpunkte zu erfahren, wann eine Ein- bzw. Auswechslung eines Spielers sinnvoll ist.

Innerhalb der Regeln des Weltfußballverbands FIFA (Fédération Internationale de Football Association) sind bei einem Spiel maximal drei Wechsel pro Mannschaft erlaubt. Eine erneute Einwechslung eines bereits ausgewechselten Spielers ist dagegen ausgeschlossen. Durch diese Regeln kann ein Trainer durch passende Wechsel nur beschränkt auf den Spielverlauf einwirken. Ziel des Autors ist es eine effektive Wechselstrategie zu entwickeln, welche die Wahrscheinlichkeit erhöht das Spiel positiv zu beeinflussen [45, S. 1].

Dem Autor Myers stehen für die Analyse 485 Spiele zu Verfügung, bei denen die Tordifferenz vor und nach jedem Spielerwechsel gemessen wurden. Außerdem liegt der Fakt vor, ob eine Mannschaft zu Hause spielte oder auswärts. Die aufgenommenen Spiele stammen aus der englischen Premier League (155 Instanzen), der spanischen Liga La Liga (158) und der Serie A aus Italien (172).

Bevor der Autor mit der Anwendung der Klassifikation beginnt, beschreibt er das Verhalten der Ein- und Auswechslungen der Trainer in den zur Verfügung stehenden Daten. So ist es üblich, dass die Trainer in über 80 % der Fälle alle drei Wechselmöglichkeiten nutzen und in keinem der 485 Spiele ein Trainer gar keinen Wechsel vornimmt. Der erste Wechsel wird im Schnitt in der 56. Minuten vorgenommen, der zweite Wechsel in der 70. Minute und die letzte Einwechslung passiert im Mittel in der 80. Minute. [45, S. 4]

Weitere Analysen der Daten zeigen, dass sich die Zeitpunkte der Einwechslungen unterscheiden, sofern eine Mannschaft führt, zurückliegt oder der Spielstand unentschieden ist. So nehmen die Trainer bei einem Unentschieden oder einer bevorstehende Niederlage die Wechsel früher vor. Sofern der Verein vorne liegt, vertraut der Trainer dem Team und lässt die Spieler länger spielen. [45, S. 7]

Zusätzlich untersucht der Autor, ob sich die Wechselzeitpunkte unterscheiden, sofern eine Mannschaft zu Hause spielt oder nicht. Der Autor kann hier jedoch keine statistische Signifikanz bezüglich der Zeitpunkte feststellen.

Innerhalb der Ligen vollziehen die Trainer der italienischen Liga ihre erste Einwechslung früher als die spanische und englische Liga. In der zweiten und dritten Einwechslung liegt keine statistische Signifikanz für eine Unterscheidung vor.

**Abbildung 6:** Wechselstrategie von Myers in Anlehnung an [45, S. 11]

- Wenn Mannschaft zurückliegt:
  - Wechsle zum ersten Mal vor der 58. Minute
  - Wechsle zum zweiten Mal vor der 73. Minute
  - Wechsle zum dritten Mal vor der 79. Minute
- Sonst:
  - Wechsle nach belieben

Um eine statistische Signifikanz festzustellen, nutzt der Autor, wie auch in dem Artikel aus Abschnitt 3.2, die Varianzanalyse und zusätzlich den Zweistichproben-t-Test. Bei der Varianzanalyse gelten allgemein die Varianzhomogenität und die Normalverteilung des Stichprobenumfangs als Voraussetzung [53, S. 381]. Der Zweistichproben-t-Test setzt ebenfalls die Normalverteilung voraus. Eine Erklärung, ob diese Voraussetzungen gegeben sind, bleibt der Autor schuldig. Ob eine Anwendung der Varianzanalyse oder dem t-Test sinnvoll ist bleibt somit fraglich.

Bei der Anwendung von Data Mining wählt der Autor die Entscheidungsbäume als geeignete Klassifikationsmethode. Ein Entscheidungsbaum ist leicht darstellbar und da es Ziel ist, den Trainern eine Strategie vorzulegen, welche sie einfach nutzen können, wird diese Methode in dem Artikel gewählt. Als Trainingsdaten werden die 485 Spiele eingesetzt. In Abbildung 6 ist die Strategie in einer Regel dargestellt, welche von der Klassifikation erlernt wird.

Zusätzlich zu der entstandenen Regel gibt der Autor an, dass Mannschaften, welche vor dem ersten kritischen Punkt in Minute 58 hinten lagen und sich an die Regel hielten in 41 % der Fälle ihre Lage verbesserten. Sofern ein Trainer nicht vor der 58. Minuten gewechselt hat, konnten in 18 % der Spiele das Ergebnis verbessert werden. Zu 30 % konnten Mannschaften ihre Situation verbessern, wenn sie vor der 73. Minuten hinten lagen und ihre zweite Einwechslung vor diesem Punkt getätigt haben. Sofern sie dies nicht taten, trat eine Verbesserung in nur 6 % der Spiele ein. Sobald eine Mannschaft bis zur 79. Minute hinten lag und der Trainer bis zu diesem Zeitpunkt einen Spieler ausgewechselt hat, brachte dies in 24 % der Fälle eine Verbesserung. 7 % der Spiele konnten gedreht werden, falls die Trainer sich nicht an diese Regel hielten.

Für Mannschaften die vorne liegen oder bei Spielen bei denen es Unentschieden steht, können keine Empfehlungen für die Wechsel ermittelt werden. Ebenfalls gilt die Regel nicht für Wechsel, die auf eine Verletzung folgen oder wenn Spieler eine rote Karten erhalten. [45, S.10]

Um die Regel zu validieren, nutzt der Autor weitere 1283 Spiele aus verschiedenen Ligen, u.a. aus der Bundesligasaison 2009/2010 und der Weltmeisterschaft 2010, bei denen die Regel angewendet werden können. Die Instanzen aus den Trainingsdaten sind ebenfalls im Validierungsdatensatz enthalten.

Die Validierung zeigt, dass in 34,29 % der Fälle die Trainer nach der Regel gewechselt haben. In diesen 440 Spielen war die Regel in 42,27 % der Fälle erfolgreich. Von den 843 Spielen bei denen die Regel nicht befolgt wurde, konnten nur 20,52 % der Vereine ihr Ergebnis verbessern. Somit kann ausgesagt werden, dass bei einer Befolgung der Regel die Chance auf eine Ergebnisverbesserung fast verdoppelt werden kann. [45, S. 13]

Der vorgestellte Artikel bietet einen interessanten Ansatz zur Anwendung von Data Mining im Fußball. Hierbei sollen die Trainer in ihren Entscheidungen unterstützt

werden, zu welchem Zeitpunkt ein Wechsel stattfinden soll. Wie der Autor in seinem Fazit schreibt, ist es schwer eine exakte Regel zu ermitteln, da ein Spielausgang von vielen verschiedenen Faktoren abhängt. Der Autor erwähnt ebenfalls, dass zukünftige Forschungen in diesem Bereich mehrere Faktoren in Betracht ziehen sollten. Beispielsweise fehlt in der Analyse welche Art von Spieler, wie Abwehrspieler oder Stürmer, eingewechselt werden soll. Außerdem wird die Qualität der Einwechselspieler nicht betrachtet. Weitere Analysen können auf dieser Analyse aufbauen und sie verfeinern.

### 3.5 Vorhersage von Spielausgängen

Viele Forschungen untersuchen die Möglichkeit mithilfe von Data Mining Sportereignisse vorauszusagen. Beispielsweise versuchen Rotshtein, Posner und Rakityanskaya [52] mithilfe der Ergebnissen aus der Vergangenheit Spielausgänge im Fußball zu prognostizieren. Chen et al. [10] wenden neuronale Netzwerke zur Vorhersage von Hunderennen an. Lyle [39] sowie West und Lamsal [66] beschäftigen sich mit der Vorhersage von Prognosen im Baseball bzw. Football. Im Basketball werden von Beckler, Wang und Papamicheal [2], Loeffelholz, Bednar und Bauer [38] sowie Ivankovic et al. [32] versucht die Spielausgänge vorherzusagen. Einen Überblick über den Einsatz von Data Mining zur Vorhersage im Basketball ist in der Masterthesis von Cao [6, S. 33ff] zu finden. Im Weiteren wird die Anwendung von neuronalen Netzwerken für die Vorhersage von Spielausgängen vorgestellt, wie sie von Ivankovic et al. [32] beschrieben werden.

Ivankovic et al. [32] nutzen bei der Anwendung der neuronalen Netze die Daten der serbischen Basketballliga der Saisons 2005/2006 bis 2009/2010. Damit stehen den Autoren Daten über fünf Saisons zur Verfügung, was insgesamt 890 Spielen entspricht. Als Attribute stehen unter anderem die Vorlagen (Assists), die offensiven und defensiven Rebounds (das Fangen von abgeprallten Bällen), die blockierten Würfe oder die Fouls zur Verfügung. Zusätzlich stehen detaillierte Informationen über die Punktewürfe bereit. Beispielsweise steht die Position der zwei und drei Punktewürfe zur Verfügung. Das Spielfeld ist hierbei in sechs Positionen unterteilt. Für diese sechs Zonen sind dabei die Anzahl an Würfen sowie die Anzahl an Treffern gespeichert.

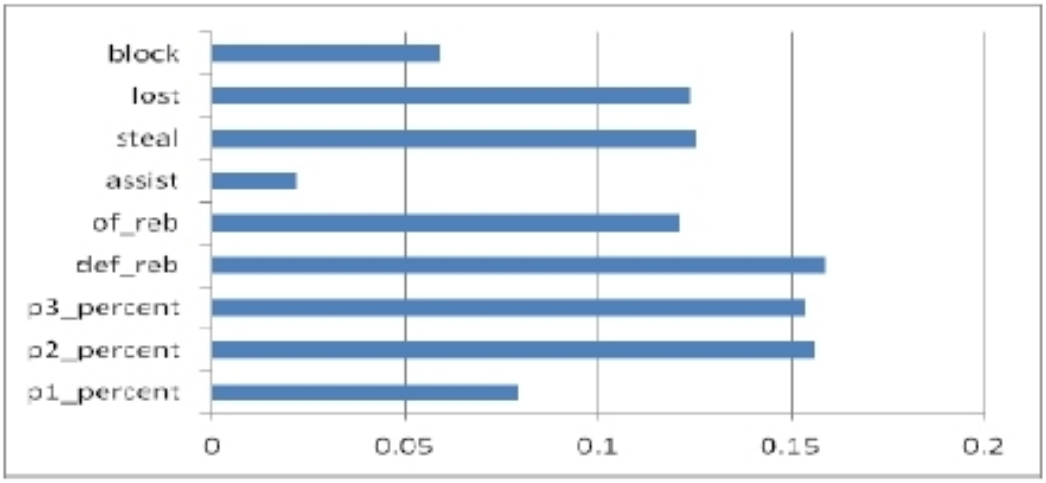
In einer ersten Anwendung des neuronalen Netzes werden nur diese Wurfdaten als Eingangsgrößen für das neuronale Netz genutzt. Dabei wird für jede Zone der Anteil an Treffern zu Würfen pro Team und Spiel berechnet. Als Zielwerte werden die Klassen „Sieg" und „Niederlage" definiert.

Die Autoren teilen die Daten in 75 Prozent Trainingsdaten und 25 Prozent Testdaten. Die Klassifikation erreicht dabei eine Klassifikationsrate von 66,4 % richtig vorhergesagter Spiele [32, S. 176]. Um diese Rate zu verbessern, führen die Autoren eine zweite Klassifikation unter Berücksichtigung aller verfügbaren Attribute durch. Hierbei wird vernachlässigt, in welcher Zone die Würfe abgegeben werden. Es werden lediglich die Trefferquoten der 1-Punkte-, 2-Punkte- und 3-Punktewürfen im Allgemeinen betrachtet. Zusätzlich liegen die Anzahl an offensiven und defensiven Rebounds, Vorlagen, Steals (das Abnehmen des Balls vom Gegner), verlorenen Bällen und blockierten Würfen vor.

In Abbildung 7 sind die Attribute und die Höhe ihres Einflusses auf den Ausgang des Spiels visualisiert. Es zeigt sich, dass die Rebounds in der Defensive, sowie die Trefferquoten der Zwei- und Dreipunktwürfe die wichtigsten Attribute darstellen.

Insgesamt werden von dem neuronalen Netzwerk unter Berücksichtigung des erweiterten Datensatzes 1441 von 1780 Instanzen richtig klassifiziert. Das entspricht einer Klassifikationsrate von 80,96 %. Dies ist eine sehr gute Vorhersage, welche jedoch in dieser Weise zu keiner Prognose genutzt werden kann. Die Klassifikation ist gut darin den Ausgang anhand der Werte vorherzusagen, welche im Spiel erreicht wor-

**Abbildung 7:** Einfluss der Attribute auf den Spielausgang [32, S. 178]




den sind. Vor einem Spiel sind diese jedoch nicht bekannt. Ein noch nicht gespieltes Spiel könnte somit nicht prognostiziert werden. Eine Möglichkeit wäre es, die Annahme zu machen, dass eine Mannschaft im nächsten Spiel so spielt, wie sie in den Partien zuvor gespielt hat. Somit könnte man Durchschnittswerte der letzten Spiele berechnen und darauf eine Klassifikation anwenden. Jedoch wird bei einer solchen Analyse die Stärke des Gegners nicht berücksichtigt.

Durch die Visualisierung der Einflussstärke der einzelnen Attribute, kann man jedoch die wichtigsten Attribute für einen Sieg ermitteln und diese als Information für die eigene Taktik nutzen. Eine andere Möglichkeit wäre es, statt eines neuronalen Netzwerks, einen Entscheidungsbaum als Methode zu nutzen. Mit einem Entscheidungsbaum ist es möglich die Zusammenhänge der Einflussgrößen zu visualisieren. Im Gegensatz dazu sind die Muster, die ein neuronales Netzwerk ermittelt für den Anwender verborgen. Solche Methoden werden deshalb als Black-Box Methoden angesehen.

Loeffelholz, Bednar und Bauer [38] verwenden ebenso ein neuronales Netzwerk zur Vorhersage von Spielausgängen im Basketball. Die Autoren dieses Artikels, nutzen im Gegensatz zum vorherigen Artikel, die Durchschnittswerte von Mannschaften zur Prognose von Spielen. Außerdem setzen sie die Werte der beiden Gegner gegenüber. Somit wird auch die Stärke der gegnerischen Mannschaft betrachtet.

Bei der Analyse stehen folgende Attribute für die Teams zur Verfügung [38, S. 3]:

- Trefferquote der Zweipunktewürfe (Field Goal %)
- Trefferquote der Dreipunktewürfe (Three Point %)
- Trefferquote der Freiwürfe (Free Throw %)
- Offensive Rebounds (Offensive Rebound)
- Defensive Rebounds (Defensive Rebound)
- Vorlagen (Assists)
- Steals
- Blockierte Würfe (Blocks)
- Ballverluste (Turnovers)
- Fouls (Personal Fouls)
- Punkte

Diese Werte sind jeweils für die Heim- und Gastmannschaft gegenübergestellt. Zusätzlich ist aufgenommen, ob eine Mannschaft zuhause (1) oder auswärts (0) spielt. Als Klassen dienen die Werte {1} für einen Heimsieg und {2} für einen Auswärtssieg.

**Abbildung 8:** Beispiel 6 Basketballspiele [38, S. 3]

Percentage Input

| Away | | | | | | | | | | | | Home | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FG | 3P | FT | Oreb | Dreb | Ast | Stl | Blk | TO | PF | PTS | Away | FG | 3P | FT | Oreb | Dreb | Ast | Stl | Blk | TO | PF | PTS | Home | Winner |
| 50.0 | 46.2 | 76.5 | 8 | 32 | 15 | 1 | 4 | 17 | 19 | 97 | 0 | 47.1 | 25.0 | 69.2 | 12 | 28 | 21 | 8 | 4 | 9 | 19 | 106 | 1 | 2 |
| 52.2 | 45.5 | 83.3 | 16 | 40 | 24 | 9 | 7 | 19 | 26 | 117 | 0 | 41.6 | 26.1 | 68.4 | 7 | 30 | 19 | 8 | 9 | 20 | 28 | 96 | 1 | 1 |
| 45.9 | 27.3 | 67.7 | 12 | 37 | 23 | 10 | 5 | 19 | 30 | 95 | 0 | 42.1 | 25.0 | 60.0 | 11 | 26 | 18 | 16 | 3 | 12 | 22 | 93 | 1 | 1 |
| 36.4 | 30.0 | 71.1 | 23 | 33 | 15 | 6 | 6 | 18 | 30 | 110 | 0 | 41.3 | 44.8 | 83.3 | 14 | 42 | 20 | 12 | 6 | 17 | 33 | 119 | 1 | 2 |
| 37.9 | 23.5 | 59.1 | 18 | 30 | 16 | 4 | 4 | 15 | 28 | 83 | 0 | 43.8 | 54.5 | 74.3 | 9 | 33 | 15 | 2 | 9 | 10 | 22 | 102 | 1 | 2 |
| 45.8 | 41.2 | 60.9 | 16 | 31 | 22 | 3 | 6 | 18 | 17 | 97 | 0 | 49.4 | 56.3 | 88.2 | 8 | 27 | 23 | 8 | 6 | 11 | 20 | 106 | 1 | 2 |

In Abbildung 8 sind die Daten für sechs Spiele aufgelistet.

Es liegen die Daten der Saison 2007/2008 vor. Dies entspricht 650 Instanzen. Die Spiele werden so vorausgesagt, wie es auch in einer realen Anwendung eingesetzt werden könnte [38, S. 2]. Wie bereits erwähnt werden zur Prognose die Durchschnittswerte der Mannschaften für die gesamte Saison genutzt.

Statt nur ein neuronales Netzwerk zu nutzen, stellen die Autoren mehrere Arten der neuronalen Netze, wie das Feed Forward Neural Network oder Radial Basis Functions gegenüber. Außerdem werden alle benutzten neuronale Netze mithilfe des Bayesian Belief Networks und des Probalisitc Neural Network fusioniert. Idee ist es, alle Arten der neuronalen Netze zu kombinieren und so eine bessere Klassifizierung zu erzielen [38, S. 8ff]. Um die erlernten Modelle zu testen, nutzen die Autoren insgesamt zehn Trainings- und Testdatensätze. Der erste Datensatz wird in 620 Trainingsdaten und 30 Testdaten geteilt. Dabei sind die Trainingsinstanzen die ersten 620 Spiele und die Testdaten die 30 darauf folgenden Spiele. Die anderen neun Trainings- und Testdatensätze werden wie bei der Cross-validation zufällig gewählt [38, S. 10].

Die Autoren führen die neuronalen Netzwerke auf insgesamt vier verschiedene Datensätze aus. Der erste Datensatz betrachtet alle 22 Attribute. Für den zweiten Datensatz wird die Feature Selection benutzt, welche die Attribute aus dem Datensatz ermittelt, die den größten Einfluss auf den Zielwert haben. Dabei werden die Attribute Ballverluste und Punkte als wichtigste Attribute erkannt. Der dritte Datensatz betrachtet alle offensiven Attribute, wie die Zweipunktetrefferquote, die Dreipunktetrefferquote und die Trefferquote für die Freiwürfe der beiden Teams. Der letzte Datensatz nutzt lediglich die offensiven Attribute der Dreipunktewürfe und Freiwürfe als Einflussgrößen.

Die durchschnittlichen Klassifikationsraten über die zehn Testdatensätze sind in Tabelle 1 dargestellt. Um einen Vergleich von Data Mining zu menschlichen Experten vorzunehmen, sind zusätzlich die Klassifikationsraten der Experten der Zeitschrift USA Today dargestellt. Die Experten der Zeitschrift versuchen vor den jeweiligen Spieltagen der Basketballliga die Spiele vorherzusagen.

Die Ergebnisse zeigen, dass Data Mining mehr Spiele vorhersagen kann, als es die menschlichen Experten tun. Am besten lassen sich die Spieler anhand der Trefferquoten der Dreipunktewürfe und Freiwürfe prognostizieren. Eine Fusionierung aller Netzwerke, führt zu keinen besseren Ergebnissen. Mit 74,33 % durchschnittlicher Klassifikationsrate kann man sagen, dass eine gute Vorhersagbarkeit durch Data Mining erreicht werden kann. Bei der Einteilung der ersten 620 Spiele der Saison als Trainingsdatensatz und die nachfolgenden 30 Spiele als Testdaten, kann sogar eine Klassifikationsrate von 83,33 % erzielt werden [38, S.13f].

Die Resultate dieses Artikels zeigen, dass eine Vorhersage von Spielen sehr gut möglich ist. Unter Zuhilfenahme mehrere Attribute oder mehreren Trainingsdaten

**Tabelle 1:** Durchschnittliche Klassifikatonsraten in Anlehnung an [38, S. 13]

| Netzwerk | Alle Attribute | Feature Selection | Offensive (6 Attr.) | Offensive (4 Attr.) |
|---|---|---|---|---|
| FFNN | 71,67 | 70,67 | 72,67 | 74,33 |
| RBF | 68,67 | 69,00 | 68,00 | 72,00 |
| PNN | 71,33 | 69,00 | 72,33 | 73,34 |
| GRNN | 71,33 | 69,00 | 72,33 | 73,34 |
| PNN Fusion | 71,67 | 70,67 | 72,67 | 72,24 |
| Bayes Fusion | 72,67 | 70,67 | 72,67 | 71,57 |
| Experten | 68,67 | 68,67 | 68,67 | 68,67 |

FFN: Feed Forward Neural Network, RBF: Radial Basis Function
PNN: Probalistic Neural Network, GRNN: Generalized Neural Network

könnten sogar noch bessere Ergebnisse erzielt werden.

### 3.6 Kombination aus Cluster- und Regressionsanalyse

Chan, Cho und Novati [7] sowie Chan und Novati [8], kombinieren das Clustering und die Regressionsanalyse, um den Punkteanteil eines Spielers an der Gesamtpunktzahl einer Mannschaft zu messen. Die Fragestellung, die hier beantwortet wird, ist die gleiche, die auch dieser Arbeit zu Grunde liegt. Der hier vorgestellte Idee zur Quantifizierung der Wichtigkeit eines Spielers, wird in dieser Arbeit in Kapitel 6 aufgenommen und auf die Spieler der Bundesliga angewendet.

Die Artikel orientieren sich an der Clusteranalyse aus dem in Abschnitt 3.3 vorgestellten Artikel, bei dem Eishockeyspieler in verschiedene Kategorien eingeteilt werden. Als erstes clustern die Autoren Chan, Cho und Novati, die ihnen zur Verfügung stehenden Eishockeyspieler anhand ihrer Spielweisen. In einem zweiten Schritt wird mithilfe der Regression analysiert, in welchem Maß die einzelnen Spielercluster zu den erzielten Punkten einer Mannschaft beitragen. Der Einfluss eines Spielers auf die Gesamtpunktzahl, wird in Punkten angegeben. Je höher diese Punktzahl ist, umso wichtiger war ein Spieler für seinen Verein.

Chan, Cho und Novati [7, S. 134] liegen die Daten über die Saisons 2005/2006 bis 2009/2010 vor. Für die Saison 2008/2009 sind dies beispielsweise 582 Offensivspieler, 303 Defensivspieler und 89 Torhüter.

Für die Clusteranalyse werden für jede Position die oberen 75 % an Spielern ausgewählt, welche auf dieser Position am meisten Einsatzzeit hatten. Die unteren 25 % werden aus dem Datensatz gefiltert. Dadurch umgehen die Autoren das Problem, dass Spieler im Datensatz enthalten sind, die nur wenig Einsatzzeit hatten und deren Spielweise durch diese geringen Einsatzminuten nur schlecht abgebildet werden [7, S. 134].

Folgende Attribute sind über die Laufzeit einer Saison gesammelt [7, S. 133f]:

- Offensive Spieler:
  - Tore
  - Vorlagen
  - Plus-Minus Statistik
  - Anzahl an Körpereinsätzen gegenüber dem Gegner (Hits)
  - Blockierte Schüsse
  - Strafminuten

- Defensive Spieler:
  - Punkte (Tore + Vorlagen)
  - Plus-Minus Statistik
  - Anzahl an Körpereinsätzen gegenüber dem Gegner
  - Blockierte Schüsse
  - Strafminuten
- Torhüter:
  - Gehaltene Schüsse in %
  - Gegentorschnitt
  - Siege / Spiele in der Startaufstellung
  - Spiele ohne Gegentor / Spiele in der Startaufstellung

Dabei werden die einzelnen Attribute durch die Anzahl an Spielminuten eines Spielers geteilt. Die Plus-Minus Statistik beschreibt die Differenz von geschossenen und erhaltenen Toren der Mannschaft eines Spielers, während der entsprechende Spieler auf dem Eis stand. Als Siege gelten die Anzahl an Spielen bei denen der Torwart auf dem Eis stand und die Mannschaft das Siegtor geschossen hat. Der Gegentorschnitt wird im Englischen auch als GGA (Goals Against Average) bezeichnet und beschreibt die durchschnittliche Anzahl an Gegentreffern pro Spiel.

Für die Clusteranalyse werden alle Attribute in den gleichen Wertbereich abgebildet. Dazu werden die Werte für Offensiv- und Defensivspieler anhand der Einsatzzeit normalisiert. Zusätzlich werden die Werte standardisiert, indem der Mittelwert abgezogen wird und durch die Standardabweichung dividiert wird. Für die Torhüter werden die Werte durch die Anzahl an Spielen in der Startaufstellung normalisiert und anschließend standardisiert. Die Werte werden für jeweils eine Saison betrachtet. Die Clusteranalyse wird separat für jede Saison ausgeführt. [7, S. 134]

Wie bei der Clusteranalyse von Vincent und Eastman [61] aus Abschnitt 3.3, wird die K-Means Clustermethode benutzt. Da hier die Zahl der zu identifizierenden Cluster im Vorhinein definiert werden muss, nutzen Chan, Cho und Novati den Clasinski-Harabasz pseudo-F Index und den Silhouette Value, um die optimale Clusteranzahl zu bestimmen. Diese Methoden messen die Dichte der Cluster. Beispielsweise berechnet der pseudo-F Index den Anteil der Quadratsummen zwischen den Clustern und die Quadratsummen innerhalb der Cluster. Je höher die beiden Werte der beiden Methoden sind, desto besser repräsentieren die Cluster die Daten. [7, S. 134]

Es zeigt sich, dass die optimale Clusteranzahl für die einzelnen Positionen zwischen zwei und fünf Clustern liegt. Die Autoren entscheiden sich für die Clusteranzahl innerhalb dieses Bereichs, welche im Sinne des Eishockeys die aussagekräftigsten Cluster identifizieren. Für die Offensivspieler sind dies vier Cluster, welche die Spieler als Top-Sturmreihe (Top Line), zweite Sturmreihe (Second Line), defensive Angreifer (Defensive) und körperbetont-spielende Angreifer (Physical) kategorisiert. Von der Top-Sturmreihe zu den körperbetont-spielenden Angreifern verringert sich die Anzahl an Toren, Vorlagen und die Plus-Minus Statistik, während sich die Anzahl an Körpereinsätzen und Strafminuten erhöht. [7, S. 135]

In der Defensive werden ebenfalls vier Cluster identifiziert. Diese bestehen aus offensiven (Offensive), defensiven (Defensive), durchschnittlichen (Average) und körperbetont-spielenden (Physical) Defensivspielern. Dabei erzielen die offensiven Spieler die meisten Punkte. Die defensiv ausgerichteten Spieler erreichen die höchste Plus-Minus Statistik und die höchste Anzahl an blockierten Schüssen. Die körperbetont-spielenden Defensivspieler begehen die meisten Körpereinsätze und erhalten die meisten Strafminuten. Die durchschnittlichen Defensivspieler haben in den meisten Attributen durchschnittliche Werte. [7, S. 135]

Bei den Torhütern liegt die optimale Clusteranzahl bei drei Clustern. Hier gibt es die

Elite-Torhüter (Elite), die durchschnittlichen Torhüter (Average) und die schlechteren Torhüter (Bottom). Dabei erzielen die Elite-Torhüter die besten Werte in allen Attributen und die schlechteren Torhüter die schlechtesten Werte. Die durchschnittlichen Torhüter liegen in der Mitte dieser beiden Spielerkategorien. [7, S. 135f]

Aufbauend auf der Clusteranalyse, nutzen die Autoren die Regressionsanalyse, um den Einfluss der einzelnen Spielertypen auf die erspielten Punkte zu messen. Um diese Analyse durchzuführen, wird für jede Mannschaft ermittelt, wie lange die einzelnen Spielertypen für eine Mannschaft über eine gesamte Saison gespielt haben. Hierzu werden die gespielten Minute eines Spieler durch die Gesamtanzahl an gespielten Minuten der gesamten Saison geteilt. Insgesamt kann ein Spieler 4967 Minuten für eine Mannschaft spielen. Das entspricht den 82 Spielen einer Saison, welche 60 Minuten dauern. Zusätzlich wird diese Zahl mit 47 addiert, welche die Anzahl an Verlängerungsminuten für die Spiele einer gesamten Saison repräsentiert. Ein Spiel dauert demnach im Schnitt 60,6 Minuten. Dies bedeutet, wenn ein Elite-Torwart 1500 Minuten für eine Mannschaft über eine Saison gespielt hat, so werden die 1500 Minuten durch 4967 Minuten geteilt. Damit hätte dieser Torwart für das Cluster der Elite-Torhüter seiner Mannschaft mit einem Anteil von etwa 0,3 beigetragen. [7, S. 137]

Für die Regression werden diese Werte der Cluster einer Mannschaft als die unabhängigen Variablen genommen, während die Anzahl an erspielten Punkten über die Saison die abhängige Variable repräsentiert. Die Autoren führen zwei Regressionsanalysen durch. Eine Regression, welche nur die wichtigsten Cluster betrachtet und eine Regression, welche alle Cluster berücksichtigt. Dabei zeigt sich, dass die Regression unter Berücksichtigung der wichtigsten Cluster als Eingangsgrößen das beste Ergebnis erzielt. Im Schnitt liegt die Prognose 10,2 Punkte neben der tatsächlich erreichten Punktzahl, sofern die Saison 2009/2010 als Testdatensatz dient und die restlichen Saisons als Trainingsdaten. Die finale Formel der Regression wird mithilfe aller Saisons als Trainingsdatensatz ermittelt. In Tabelle 2 sind die Ergebnisse der Regression dargestellt. [7, S. 137]

Es zeigt sich, dass die Elite-Torhüter die meisten Punkte für ihre Mannschaft einspielen. Danach sind die Top-Offensivspieler am wichtigsten. Die durchschnittlichen Torhüter und die zweite Sturmreihe haben etwa den gleichen Einfluss auf die Punktzahl, welcher jedoch geringer ist als die der Elite-Torhüter und der Top-Sturmreihe. Die Defensivspieler haben von allen Clustern den geringsten Einfluss. Die körperbetont-spielenden Angreifer, die durchschnittlichen und körperbetont-spielenden Defensivspieler sowie die schlechteren Torhüter sind nicht statistisch signifikant für die Regression.

Im weiteren Verlauf des Artikels, setzen Chan, Cho und Novati den Einfluss der Spielerkategorien auf die Teampunkte mit dem Gehalt der Spieler in Beziehung. Dabei zeigt sich, dass die defensiv-agierenden Angreifer zwar weniger als die Spieler der besten- und zweiten Sturmreihe verdienen, verglichen mit dem Einfluss auf die Teampunkte jedoch eine gute Investition für Vereine darstellen [7, S. 138f]. Laut der Analyse sind die Top-Angreifer überbezahlt, während die Elite-Torhüter ebenfalls viel verdienen, dies jedoch mit der erbrachten Leistung im Gleichgewicht steht. Insgesamt sind die Defensivspieler überbezahlt, da sie nur wenig Einfluss auf die Punktzahl haben, jedoch etwa gleichviel wie die Offensivspieler verdienen.

Zum Abschluss ihres Artikels, stellen die Autoren ihr entwickeltes „Trade Tool" vor, welches auf den vorgestellten Ergebnissen aufbaut. Dabei werden im Verlauf des Artikels verschiedene Transfers exemplarisch verglichen. Die genaue Analyse der Beispiele werden von Chan, Cho und Novati [7, S. 139-141] beschrieben.

Den Einsatz dieses Werkzeugs betrachten die Autoren als kritisch. So erklären sie, dass das „Trade Tool" zur Evaluation von vergangenen Transfers nützlich ist, jedoch

**Tabelle 2:** Ergebnisse der Kombination in Anlehnung an [7, S. 138]

| Position | Cluster | Teampunkte |
|---|---|---|
| Offensivspieler | Top | 288 |
| | Zweite Sturmreihe | 199 |
| | Defensiv | 187 |
| | Körperbetont-spielend | – |
| Defensivspieler | Offensiv | 93 |
| | Defensiv | 37 |
| | Durchschnitt | – |
| | Körperbetont-spielend | – |
| Torhüter | Elite | 321 |
| | Durchschnitt | 212 |
| | Schlecht | – |

der Einsatz für eine Analyse von zukünftigen Transfers schwierig ist. So kann nicht ermittelt werden wie viele Minuten ein gekaufter Spieler in Zukunft für eine Mannschaft aufläuft, ob sich ein Spieler verletzt oder ob er seine bisherige Form hält [7, S. 141].

Der vorgestellte Ansatz bietet eine Möglichkeit die Wichtigkeit einzelner Spielertypen für die Leistung einer Mannschaft zu ermitteln und diese mit dem Gehalt der Spieler zu vergleichen. So kann evaluiert werden, ob ein Spieler seinen erhaltenen Lohn rechtfertigt. Verantwortliche können so das Gehaltsgefüge ihrer Mannschaft kritisch hinterfragen sowie den Mix ihrer Mannschaft aus verschiedenen Spielertypen überprüfen. Zusätzliche Daten können den Ansatz erweitern.

In dem fortführenden Artikel von Chan und Novati [8] wird ein anderer Ansatz der Clusteranalyse durchgeführt. Hier wird ein Spieler nicht nur einer Kategorie zugeordnet, sondern wird durch eine Kombination der Cluster seiner Position repräsentiert. Beispielsweise wird so ein Offensivspieler, der vorher in die Kategorie der Top-Angreifer eingeteilt wurde, nun als ein Spieler repräsentiert, der zu 40 % wie ein Top-Angreifer, zu 25 % wie ein zweiter Sturmreihenspieler, zu 20 % wie ein defensiver Angreifer und zu 15 % wie ein körperbetont-spielender Angreifer spielt. Durch diese Darstellung umgeht man das Problem in der vorangegangen Analyse, dass gleiche Spielertypen auch den gleichen Wert für eine Mannschaft haben. Durch die detaillierte Betrachtung der Spielweise kann die Quantifizierung der Wichtigkeit eines Spielers genauer betrachtet werden. Ebenfalls werden so keine Spieler aus der Analyse ausgeschlossen. In der vorigen Regressionsanalyse hatten Spieler, die beispielsweise unter die körperbetont-spielenden Angreifer fielen keinen Koeffizienten, da dieses Cluster als nicht signifikant eingestuft wurde [8, S. 1]. Im weiteren Verlauf können die Autoren somit einen individuellen Wert für jeden Spieler der NHL ermitteln.

In beiden Analysen wurde mit den vorgestellten Attributen der Spieler gearbeitet. Mit weiteren Attributen könnte das Clustering noch detailliertere Spielweisen betrachten.

Die Idee aus diesem Artikel wird in dieser Arbeit in Kapitel 6 aufgenommen und mit den hier vorliegenden Daten durchgeführt. Dadurch soll ermittelt werden, in welchem Umfang einzelne Spieler der Bundesliga die Leistung einer Mannschaft beeinflussen.

# 4 Datengrundlage

Bis vor wenigen Jahren wurden nur die offensichtlichsten Daten eines Fußballspiels erhoben. So sind beispielsweise die Aufstellung der Mannschaften, die Vorlagengeber eines Tores oder die Torschützen über Jahre hinweg aufgenommen worden und im Internet einsehbar. Erst seit wenigen Jahren werden gehaltvollere Daten aufgenommen und Spielern, Trainern oder Fans präsentiert. Dabei sind z.B. auf der offiziellen Seite der Bundesliga `bundesliga.de` seit der Saison 2009/2010 umfangreichere Statistiken über die Teams einzusehen, wie beispielsweise die Anzahl an Ballkontakten der gesamten Mannschaft in einem Spiel. In den Spielstatistiken ab der Saison 2011/2012 sind zusätzliche Statistiken, wie das Zweikampfverhalten, das Passspiel oder die Laufdistanz für einzelne Spieler bereitgestellt. Ab wann genau die erweiterte Datenaufnahme begonnen hat ist nicht bekannt. Opta, eines der führende Unternehmen, welches Daten im Sport und vor allem im Fußball aufnimmt und bereitstellt, hat 2006 die Datensammlung um die Aufnahme der x/y Koordinaten des Spielfelds erweitert. Außerdem werden seit diesem Jahr die Daten von jeweils zwei Spielbeobachtern und einem Aufseher gesammelt [49]. In Deutschland sind die Stadien der ersten und zweiten Bundesliga seit der Saison 2011/12 mit dem Tracking-System Vis.Track, einem Bildbearbeitungs-System ausgestattet [31]. Dieses ermöglicht Spiele aufzuzeichnen und teilweise automatisiert auszuwerten.

In anderen Sportarten wie Baseball, Basketball, Eishockey oder Football werden schon länger detaillierte Daten aufgenommen und ausgewertet. Gerade im Baseball, wo durch den Statistiker Bill James im Jahr 1977 eine Revolution in der Statistik von Baseball losgetreten wurde, gibt es große, frei zugängliche Datenbasen zu finden. Beispielsweise sind auf `seanlahman.com/baseball-archive/statistics/` Schlag- und Wurfstatistiken seit 1871 gesammelt. Wie man in der Readme-Datei der Datenbank sieht, sind hier eine Vielzahl von spielspezifischen Attributen für die einzelnen Spieler und Spiele gesammelt. Beim Basketball findet man unter anderem auf der Seite `databasebasketball.com` umfangreiche Statistiken von Spielen der NBA seit der Saison 1946/47. Für Football stehen beispielsweise unter `pro-football-reference.com` Daten bis zurück zum Jahr 1920 zur Verfügung.

Der große Vorteil der Datenbanken dieser Sportarten ist, dass die Daten frei zum Download verfügbar sind. Beim Fußball ist hier ein Defizit zu erkennen. So kann man Webseiten finden, welche beispielsweise alle Ergebnisse der Spiele oder die Torschützenlisten der Saisons zur Verfügung stellen. Es finden sich jedoch keine detaillierten Informationen, wie beispielsweise über die Zweikampfquoten oder über das Laufverhalten von Mannschaften oder Spielern. Einen Anfang zur Änderung dieses Umstandes wurde in der Saison 2011/2012 gemacht. Detaillierte Daten für diese Saison der Premier League wurden vom Verein Manchester City und Opta frei zur Verfügung gestellt. Derzeit existiert die Aktion nicht mehr, es konnte sich über einen bestimmten Zeitraum jedoch jede Person die umfangreichen Daten dieser Saison herunterladen und eigene Analysen kreieren. Inwiefern in Zukunft Daten frei zur Verfügung gestellt werden ist derzeit nicht absehbar.

Die Daten, welche in dieser Arbeit zum Einsatz kommen werden hauptsächlich von der IMPIRE AG zur Verfügung gestellt. Die IMPIRE AG ist ein Dienstleister von Fußballdaten sowie den zugehörigen Services und Technologien [30]. Die Firma wurde 1988 gegründet und hat ihren Hauptsitz in Ismaning bei München [27]. Ab der Saison 2011/2012 ist die IMPIRE AG zur Erhebung der offiziellen Spieldaten der 1. sowie 2. Bundesliga berechtigt. Die Entscheidung zur offiziellen Vergabe der Datenerhebung durch die DFL wurde aufgrund des Ziels zur Vereinheitlichung und Steigerung der Datenqualität der Wettbewerbe des Ligaverbandes getroffen [4]. Die IMPIRE AG hält

zusätzlich die Rechte für die weltweit exklusive Vermarktung der Daten für Medien und Sportwetten [28].

Die Erhebung der Daten erfolgt bei der IMPIRE AG in zwei Teilen. Einerseits gibt es die Spielbeobachtung (Scouting) und andererseits das Tracking. Das Scouting nimmt die Aktionen auf dem Spielfeld auf. Das Tracking speichert die physischen Daten, wie etwa die Laufleistung, mithilfe von zwei im Stadion installierten Kameras [59]. Laut der IMPIRE AG erheben insgesamt sieben Personen die Daten eines Spiels. Dabei sind vier dieser Personen die Spielbeobachter. Die restlichen drei Personen dienen als Trackingoperatoren für das Spiel. Zusätzlich werden die Daten im Nachgang überarbeitet und verfeinert [29]. Die IMPIRE AG nimmt pro Spiel über 2000 Ereignisse auf und hat seit 1992 jedes Bundesligaspiel per Videoanalyse ausgewertet. Die aufgenommenen Ereignisse beinhalten z.B. die Pässe, Zweikämpfe, Fouls, gelaufenen Kilometer oder die Anzahl an Sprints [29].

Für die hier vorliegende Arbeit wurden von der IMPIRE AG die Daten für die Saison 2010/2011 sowie 2011/2012 zur Verfügung gestellt. Zur Vereinfachung wird im Folgenden die Saison 2010/2011 als Saison 2010 bezeichnet sowie die Saison 2011/2012 als Saison 2011. Dabei konnten insgesamt sechs Ereignisse bzw. Attribute ausgewählt werden, welche die IMPIRE AG pro Spieler freigibt. Dabei handelt es sich um die folgenden spielspezifischen Attribute:

- Gewonnene Zweikämpfe in %
- Anzahl an Ballkontakten
- Anzahl an Pässen
- Erfolgreiche Passquote in %
- Anzahl an Torschüssen
- Anzahl an Torschussvorlagen

Diese Ereignisse sind für jeden Spieler in den Spielen gespeichert, an denen er teilgenommen hat. Zusätzlich sind in der von der IMPIRE AG zur Verfügung gestellten Excel Tabelle die Saison, der Spieltag, das Datum des Spiels, die Heimmannschaft, die Gastmannschaft und der Verein sowie der Name des Spielers gelistet. Tabelle 3 zeigt die Daten für das Spiel zwischen dem 1. FC Köln und dem SV Werder Bremen des 19. Spieltags der Saison 2010, wie sie von der IMPIRE AG bereitgestellt werden.

Mithilfe dieser Daten sind weitere Attribute eines Spielers berechenbar. So kann die **„Prozentzahl der verlorenen Zweikämpfe"** mithilfe der „gewonnenen Zweikämpfe in Prozent" ausgerechnet werden, indem man die Zahl von 100 abzieht. Die **„fehlerhafte Passquote in %"**, die **„Anzahl an erfolgreichen Pässen"** sowie die **„Anzahl an fehlerhaften Pässen"** lassen sich mithilfe der zwei Attribute „Anzahl an Pässen" und „erfolgreiche Passquote in %" ausrechnen. Die „fehlerhafte Passquote in %" ergibt sich indem man von 100 die „erfolgreiche Passquote in %" abzieht. Die „Anzahl an Pässen" multipliziert mit der „Erfolgreichen Passquote in %" sowie geteilt durch 100 ergibt die „Anzahl an erfolgreichen Pässen". Die „fehlerhaften Pässe" ergeben sich wiederum aus der „Anzahl an Pässen" minus der „Anzahl an erfolgreichen Pässen".

Als weitere Quelle für Spielerdaten im Fußball dient in dieser Arbeit das Internet. Dabei gibt es mehrere Webseiten, welche die Ergebnisse von jedem Spiel der Bundesliga in einem Spielbericht aufzeichnen. Zusätzlich zu dem reinen Ergebnis des Spiels, werden hier leicht aufzunehmende Ereignisse gelistet. So sind bei den meisten Seiten die Torschützen, Vorlagengeber und auch das Erhalten einer gelben Karte aufgezeichnet.

In der vorliegenden Arbeit wird die Seite `fussballdaten.de` als weitere Quelle genutzt, um die Daten der IMPIRE AG anzureichern. Auf der Seite `fussballdaten.de` sind vielfältige Informationen zu den Bundesligen, zum DFB-Pokal, zu Länderspielen und auch zu ausländischen Ligen zu finden. In erster Linie

**Tabelle 3:** Beispiel IMPIRE AG Daten

| Saison | Spieltag | Datum | Heim | Gast | Verein | Spieler | ZK | BK | P | PQ | TS | TV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Marko Arnautovic | 15 | 18 | 10 | 80 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Denni Avdic | 30 | 13 | 3 | 66,67 | 2 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Philipp Bargfrede | 47,62 | 72 | 38 | 81,58 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Torsten Frings | 48 | 72 | 41 | 78,05 | 0 | 1 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Clemens Fritz | 52,63 | 85 | 41 | 75,61 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Aaron Hunt | 38,46 | 50 | 32 | 87,5 | 1 | 2 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Felix Kroos | 35,71 | 22 | 15 | 80 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Marko Marin | 47,83 | 26 | 16 | 87,5 | 0 | 1 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Per Mertesacker | 69,57 | 69 | 47 | 89,36 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Petri Pasanen | 62,5 | 77 | 42 | 100 | 1 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Claudio Pizarro | 34,38 | 41 | 23 | 82,61 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Mikael Silvestre | 48,28 | 96 | 58 | 75,86 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Sandro Wagner | 25 | 5 | 2 | 100 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Bremen | Tim Wiese | 100 | 39 | 15 | 53,33 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Andrezinho | 68 | 63 | 12 | 66,67 | 1 | 2 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Christian Clemens | 46,67 | 37 | 13 | 92,31 | 1 | 1 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Fabrice Ehret | 70,83 | 68 | 24 | 66,67 | 0 | 2 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Christian Eichner | 72,73 | 29 | 6 | 83,33 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Pedro Geromel | 61,11 | 28 | 3 | 100 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Martin Lanig | 48,72 | 56 | 21 | 90,48 | 2 | 1 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Adam Matuschyk | 65,38 | 39 | 17 | 82,35 | 2 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Kevin McKenna | 0 | 2 | 0 |  | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Milivoje Novakovic | 42,86 | 35 | 11 | 90,91 | 2 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Slawomir Peszko | 43,24 | 40 | 17 | 94,12 | 1 | 2 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Lukas Podolski | 54,55 | 58 | 22 | 68,18 | 2 | 4 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Michael Rensing |  | 21 | 5 | 60 | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Reinhold Yabo | 0 | 0 | 0 |  | 0 | 0 |
| 2010 | 19 | 22.01.2011 | Köln | Bremen | Köln | Taner Yalcin | 0 | 2 | 1 | 100 | 1 | 0 |

ZK = % Zweikampf gewonnen, BK = Ballkontakte, P = Pässe, PQ = % Passquote, TS = Torschüsse , TV = Torschussvorlagen

**Abbildung 9:** Spielstatistik fussballdaten.de

DIE SPIELSTATISTIK 1. FC KÖLN - SV WERDER BREMEN

# 1. FC KÖLN - SV WERDER BREMEN

## 3:0 (2:0)

Bundesliga 2010/2011, 19. Spieltag
22.01.2011, 18:30 Uhr
RheinEnergieStadion (Köln), 45.100 Zuschauer
Schiedsrichter: Peter Sippel (München)

**TORE**

| | | |
|---|---|---|
| 1:0 | Lukas Podolski | 6. (Linksschuss, Ehret) |
| 2:0 | Adam Matuschyk | 33. (Rechtsschuss, Andrezinho) |
| 3:0 | Lukas Podolski | 84. (Rechtsschuss, Peszko) |

| AUFSTELLUNG FC KÖLN | | AUFSTELLUNG BREMEN | |
|---|---|---|---|
| Michael Rensing | (3,0) | Tim Wiese | (3,5) |
| Andrezinho | (3,0) | Clemens Fritz | (5,0) |
| Pedro Geromel | (2,5) | Per Mertesacker | (5,0) |
| Fabrice Ehret | (2,5) | Petri Pasanen | (5,0) |
| Christian Eichner | (3,0) | Mikaël Silvestre | (4,5) |
| Martin Lanig | (3,0) | Aaron Hunt | (5,5) |
| Adam Matuschyk | (2,0) | Torsten Frings | (5,5) |
| Christian Clemens | (2,5) | Philipp Bargfrede | (5,0) |
| Slawomir Peszko | (2,5) | Felix Kroos | (5,0) |
| Lukas Podolski | (1,5) | Marko Arnautovic | (5,5) |
| Milivoje Novakovic | (3,0) | Claudio Pizarro | (5,0) |

| WECHSEL FC KÖLN | | WECHSEL BREMEN | |
|---|---|---|---|
| Kevin McKenna | (–) | Marko Marin | (4,5) |
| für Slawomir Peszko | (85.) | für Felix Kroos | (39.) |
| Taner Yalcin | (–) | Denni Avdic | (5,0) |
| für Lukas Podolski | (88.) | für Marko Arnautovic | (56.) |
| Reinhold Yabo | (–) | Sandro Wagner | (–) |
| für Christian Clemens | (88.) | für Aaron Hunt | (75.) |

| KARTEN FC KÖLN | KARTEN BREMEN |
|---|---|
| Gelb für Pedro Geromel | Gelb für Marko Arnautovic |
| Gelb für Martin Lanig | Gelb für Tim Wiese |
| Gelb für Adam Matuschyk | |

| TRAINER FC KÖLN | TRAINER BREMEN |
|---|---|
| Frank Schaefer | Thomas Schaaf |

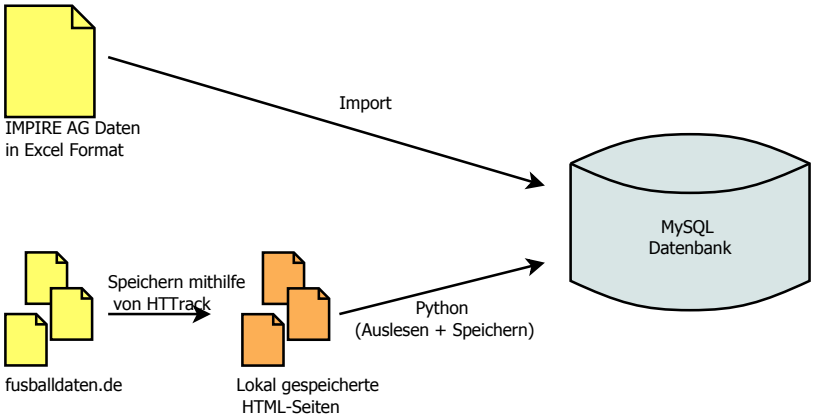
sind die Spielberichte der beiden Saisons 2010 und 2011 der 1. Bundesliga für diese Arbeit von Bedeutung. Zusätzlich werden die Spielerstatistiken der Seite genutzt, um weitere Informationen, wie die Spielerposition aber auch die Anzahl an Länderspielen eines Spielers zu ermitteln.

Bei `fussballdaten.de` wird zu jedem Spiel ein kurzer Bericht geschrieben, welcher den Spielverlauf des Spiels zusammenfasst. Außerdem gibt es eine Statistik zu dem jeweiligen Spiel. Hier sind Informationen gespeichert, wie die Paarung des Spiels, das Ergebnis, der Spieltag, das Datum, die Aufstellungen, die Trainer, die Torschützen, die Vorbereiter der Tore, die Ein- und Auswechslungen sowie die Karten, die einzelne Spieler erhalten haben. Abbildung 9 zeigt ein Beispiel einer solchen Statistik für das Spiel des 1. FC Kölns gegen den SV Werder Bremen.

Neben den Spielberichten werden Daten aus den Statistiken der einzelnen Spieler von `fussballdaten.de` ausgelesen. In der Spielerstatistik sind Informationen, wie das Geburtsdatum, die Nationalität, die Position des Spielers sowie seine Größe und sein Gewicht aufgelistet. Zusätzlich sind hier die Anzahl an Vereinsspielen aufgeteilt auf die einzelnen Saisons sowie die Spiele als Nationalspieler in der jeweiligen Saison gespeichert.

Um die entsprechenden Daten aus den HTML-Seiten von `fussballdaten.de` auszulesen werden die Spielberichte sowie die Spielerstatistiken in einem ersten

**Abbildung 10:** ETL-Prozess




Schritt mittels der Software HTTrack automatisiert heruntergeladen und lokal gespeichert. HTTrack ist ein frei verfügbarer offline Browser, welcher im Stande ist HTML Seiten, Bilder und andere Dateien von einem Server herunterzuladen und diese lokal abzuspeichern [25]. Dabei kann HTTRack, je nach Einstellung zu Links verzweigen und die entsprechenden Seiten herunterladen oder nur spezielle Seiten oder Dateien speichern. Mithilfe der Software ist es somit möglich, alle Spielerberichte der beiden Saisons sowie die Spielerstatistiken der aufgestellten Spieler auf dem lokalen System abzuspeichern. Anschließend können die HTML-Seiten entsprechend ausgewertet werden sowie deren Daten in einer geeigneten Form abgelegt werden.

Mithilfe der Programmiersprache Python werden die abgelegten HTML-Seiten ausgewertet. Python unter Einsatz der Bibliothek „lxml" ermöglicht es XML und HTML Seiten weiterzuverarbeiten. Dadurch können die benötigten Informationen aus den HTML-Seiten ausgelesen, verarbeitet und gespeichert werden. Nähere Informationen zu Python finden sich unter `python.org`. Informationen zur Bibliothek lxml unter `lxml.de`. Die Daten werden mithilfe von Python unter der Unterstützung von MySQLdb (`sourceforge.net/projects/mysql-python/`) in einer MySQL Datenbank gespeichert. Eine Datenbank vereinfacht die Speicherung, Restrukturierung und Auswertung der Daten. Zudem kann der einfache Zugriff auf die Daten aus der hier benutzten Data Mining Software Weka automatisch geschehen. Entsprechend werden die Daten aus der Excel-Datei der IMPIRE AG in eine geeignete Tabelle der MySQL Datenbank importiert. Das Extrahieren, Transformieren und Speichern wird auch der ETL-Prozess genannt. ETL steht hierbei für Extract, Transform und Load. Der ETL-Prozess wird in Abbildung 10 nochmals bildlich dargestellt.

Mithilfe von Python werden die folgenden Daten aus den Spielberichten gelesen und mit den Daten von der IMPIRE AG angereichert:

- Anzahl an Toren
- Anzahl an Vorlagen
- Anzahl an gelben Karten
- Minute der Einwechslung
- Minute der Auswechslung
- Ergebnis

Anhand dieser Werte und den Ereignissen der IMPIRE AG werden zusätzlich noch die Werte

- Vergebene Torschüsse = Torschüsse - Tore
- Erfolgreiche Torschussquote = Tore / Torschüsse
- Fehlerhafte Torschussquote = 100 - Erfolgreiche Torschussquote
- Pass pro Ballkontakt = Pässe / Ballkontakte
- Erfolgreicher Pass pro Ballkontakt = Erfolgreiche Pässe / Ballkontakte
- Torschussvorlage pro Pass = Torschussvorlage / Pässe
- Gespielte Minuten = Anzahl an gespielten Minuten berechnet durch die Minuten der Ein- und Auswechslung

berechnet. Das „Attribut Pass Pro Ballkontakt" gibt dabei den Wert zurück, der aussagt wie schnell ein Spieler den Pass weiterspielt, sobald er in Ballbesitz kommt. „Erfolgreicher Pass pro Ballkontakt" gibt wieder, inwiefern dieser Pass gleichzeitig noch erfolgreich war. Das Attribut „Torschussvorlage pro Pass" soll darstellen, wie hoch der Anteil aller Pässe ist, die direkt zu einer Torchance führen und somit zielführender sind. Die anderen Werte gelten als Anreicherung der schon aufgenommen Ereignisse. Alle bisher aufgelistete Daten werden in der Tabelle **Fakten** der MySQL-Datenbank gespeichert. Diese Tabelle enthält die folgenden spielspezifischen Attribute:

- Zweikampf
- VerlZweikampf
- Ballkontakte
- ErfPaesse
- FehlPaesse
- Paesse
- ErfPassquote
- FehlPassquote
- Torschuesse
- VergTorschuesse
- ErfTorschussquote
- FehlTorschussquote
- Torschussvorlagen
- Vorlage
- Tore
- Gelb
- PassProBallkontakt
- ErfPassProBallkontakt
- TorschussvorlageProPass

Zusätzlich werden aus den Spielerstatistiken weitere Daten ausgelesen, welche für spätere Analysen zur Verfügung stehen sollen. Dabei handelt es sich einerseits um die Spielerposition, die entweder den Wert Torwart, Abwehr, Mittelfeld oder Angriff annehmen kann. Die Spielerposition wird in den weiteren Analysen dieser Arbeit eine wichtige Rollen spielen, da Analysen für die Position oft separat durchgeführt werden. Zusätzlich werden noch das Geburtsdatum, das Alter, das Gewicht und die Größe der Spieler aufgenommen. Diese Werte werden in der Tabelle **Spielereigenschaften** gespeichert.

Die Statistik aus Tabelle 4 gibt die maximalen Werte (Max), minimalen Werte (Min), Mittelwerte (MW) und die Standardabweichung (SD für Standard Deviation) der aufgelisteten Attribute der beiden Saisons wieder. Dabei werden die Torhüter ausgelassen, da die aufgelisteten Attribute keine Torhüter-spezifischen Daten darstellen. Die Qualität eines Torhüters wird anhand anderer Daten gemessen, wie etwa die gehaltenen Torschüsse oder abgefangenen Flanken. Ein Torhüter hat in der Regel wenige

**Tabelle 4:** Statistik Fakten

| Wert | Min | Max | MW | SD |
|---|---|---|---|---|
| Zweikampf | 0 | 100 | 49,0432 | 19,4007 |
| VerlZweikampf | 0 | 100 | 50,9568 | 19,4007 |
| Ballkontakte | 0 | 145 | 44,0655 | 24,3196 |
| ErfPaesse | 0 | 131 | 22,3373 | 16,1939 |
| FehlPaesse | 0 | 27 | 5,6135 | 4,0366 |
| Paesse | 0 | 134 | 27,9508 | 18,4587 |
| ErfPassquote | 0 | 100 | 76,0820 | 18,3166 |
| FehlPassquote | 0 | 100 | 23,9180 | 18,3166 |
| Torschuesse | 0 | 10 | 1,0275 | 1,3217 |
| VergTorschuesse | 0 | 10 | 0,9177 | 1,2156 |
| ErfTorschussquote | 0 | 100 | 5,4104 | 19,0713 |
| FehlTorschussquote | 0 | 100 | 94,5896 | 19,0713 |
| Torschussvorlagen | 0 | 15 | 0,9872 | 1,3175 |
| Vorlage | 0 | 3 | 0,0913 | 0,3169 |
| Tore | 0 | 4 | 0,1098 | 0,3605 |
| Gelb | 0 | 1 | 0,0799 | 0,2711 |
| PassProBallkontakt | 0 | 1 | 0,6043 | 0,1731 |
| ErfPassProBallkontakt | 0 | 1 | 0,4732 | 0,1742 |
| TorschussvorlageProPass | 0 | 2 | 0,0454 | 0,0864 |
| Gespielt | 1 | 90 | 70,4096 | 29,4826 |

Ballkontakte oder schießt keine Tore. Um eine Verfälschung der Statistik zu umgehen werden Torhüter demnach aus der Statistik ausgelassen. Die Tabelle **Fakten** beinhaltet insgesamt 16933 Instanzen, wobei 46 dieser Instanzen Eigentore bezeichnen und keinem Spieler zugeordnet sind. Insgesamt wurde somit ein Datensatz mit 16887 Fakten von der IMPIRE AG zur Verfügung gestellt und anschließend mit Daten aus dem Internet angereichert. Ohne Torhüter sind dies 15646 Instanzen.

Anhand der von `fussballdaten.de` gelesenen Ergebnisse der Spiele wird zusätzlich die Tabelle **Abschlusstabelle** berechnet. Bei einem Sieg erhält die siegreiche Mannschaft drei Punkte, bei einem Unentschieden erhalten beide Vereine einen Punkt. Für eine Niederlage werden keine Punkte vergeben. Diese erspielten Punkte werden für die jeweiligen Hin- und Rückrunden der Saison 2010 sowie der Saison 2011 pro Verein zusammengezählt und gespeichert. Die Hinrunde behandelt die Spieltage 1 bis 17 und die Rückrunde die Spieltage 18 bis 34. Zusätzlich wird anhand der Punkte der Rang der Mannschaft berechnet. Rang bedeutet hier, in welchem Tabellenteil die betrachtete Mannschaft gelandet ist. Dabei wird die Tabelle für die Hin- und Rückrunde in drei Teile geteilt. Die ersten fünf Mannschaften sind die erfolgreicheren Mannschaften. Sie qualifizieren sich für die internationalen Wettbewerbe, wie die Europa League und die Champions League. Sie nehmen in der Tabelle den Wert {2} an. Die letzten drei der insgesamt 18 Mannschaften einer Saison gelten als die Absteiger und nehmen den Rang {0} an. Die Zahl {1} gilt für die Mannschaften, welche zwischen diesen zwei Regionen landen. Bei Punktgleichheit zum fünften bzw. 16. Platz wird der entsprechende Rang auf {2} bzw. {0} gesetzt. Die Hin- und Rückrunden werden isoliert betrachtet. Das heißt, die Punkte der Rückrunde werden nicht mit den Punkten der Hinrunde summiert, sondern separat behandelt.

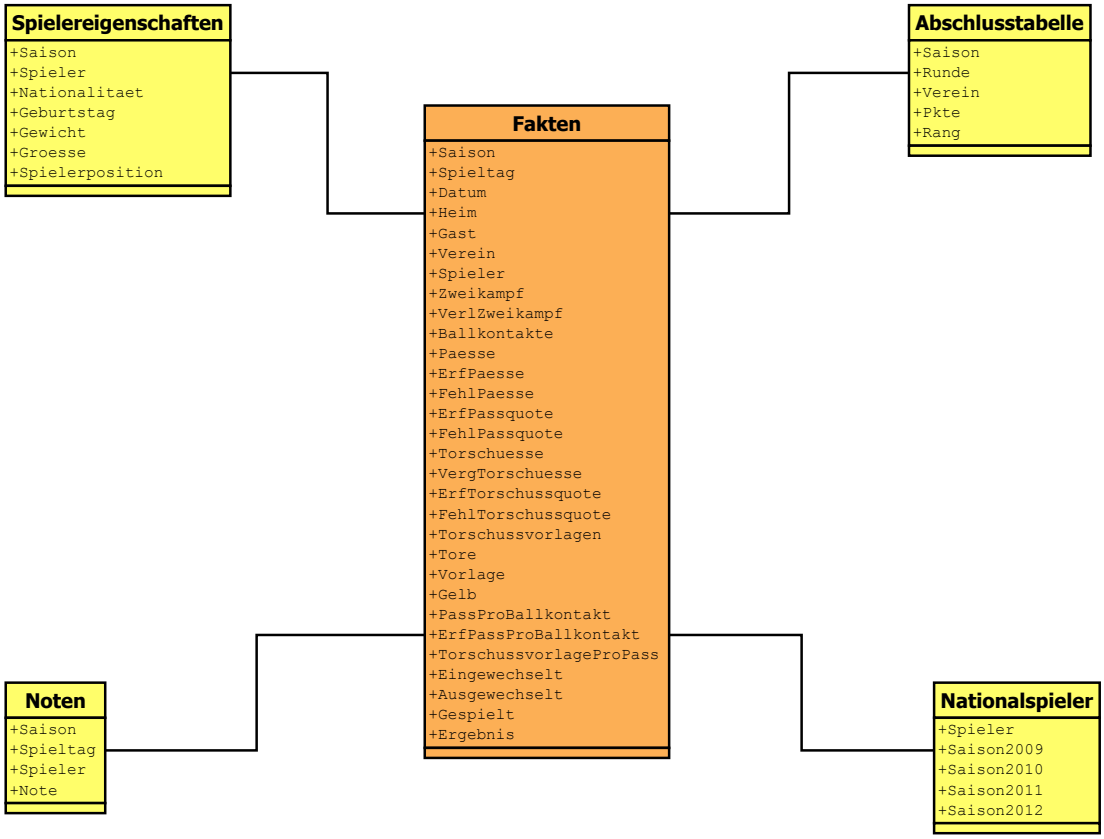
Des Weiteren sind im Spielbericht von `fussballdaten.de` die Noten eines Spielers gespeichert, welche er von der Redaktion des Portals für seine Leistung im Spiel erhalten hat. Die Notenvergabe ist näher in Kapitel 5.2 beschrieben, in dem die Noten

**Abbildung 11:** Nationalspieler fussballdaten.de

ALS NATIONALSPIELER

| SAISON | VEREIN | POSITION | SPIELE | TORE | VORL. | GELB | G/R | ROT |
|---|---|---|---|---|---|---|---|---|
| 2010 | Deutschland | Abwehr | 3 | 0 | 1 | 1 | 0 | 0 |
| 2011 | Deutschland | Abwehr | 6 | 0 | 0 | 0 | 0 | 0 |
| 2012 | Deutschland | Abwehr | 1 | 0 | 0 | 0 | 0 | 0 |
| 2013 | Deutschland | Abwehr | 2 | 0 | 0 | 0 | 0 | 0 |
| Summe Länderspiele | | | 12 | 0 | 1 | 1 | 0 | 0 |

**Abbildung 12:** Datenmodell




auch zur Anwendung beim Data Mining kommen. Die Noten der Spieler werden in der Tabelle **Noten** gespeichert. Hier ist die Saison, der Spieltag, der Spieler und die Note für das entsprechende Spiel hinterlegt.

Wie bereits erwähnt, sind neben den Spielereigenschaften auch die Anzahl an Nationalspielen in den einzelnen Jahren in dieser Arbeit von Bedeutung. Diese Information ist für die Data Mining Anwendung in Kapitel 5.1 nötig. In Abbildung 11 sind die Spiele als Nationalspieler von Dennis Aogo, wie sie von `fussballdaten.de` angezeigt werden, abgebildet. Für die hier vorgestellten Analysen wird die Anzahl an Nationalspielen eines Spielers der Jahre 2009 bis 2012 ausgelesen und in der Tabelle **Nationalspieler** gespeichert. Sofern ein Spieler in einem Jahr kein Spiel für sein Land gespielt hat, nimmt die entsprechende Zelle in der Tabelle die Zahl 0 an.

In Abbildung 12 sind die erstellten Tabellen und ihre Attribute dargestellt. Dabei sind die Tabellen im sogenannten Star-Schema angeordnet. Die Darstellung macht es möglich die Daten für die folgenden Analysen einfach zu verknüpfen und neu zu strukturieren.

# 5 Anwendungen des Data Mining im Fußball

In den folgenden Abschnitten werden die vorgestellten Verfahren, wie die Klassifikation, Regression und Clusteranalyse exemplarisch auf die vorhandenen Daten angewendet. Dazu wird in Abschnitt 5.1 die Klassifikation genutzt, um Nationalspieler anhand ihrer Spielweisen zur klassifizieren. Die aufgezeigte Anwendung soll dazu genutzt werden, gute Spieler im Fußball zu identifizieren.

Abschnitt 5.2 nutzt die Regressionsanalyse, um die Beziehungen zwischen der Leistung eines Spielers und der von Redakteuren vergebenen Note zu ermitteln. Die Ergebnisse sollen klären, welche Spielaktionen zu einer besseren Beurteilung der Leistung eines Spielers führen.

Die Clusteranalyse wird in Abschnitt 5.3 angewendet, um einerseits Mannschaften in verschiedene Gruppen zu unterteilen und andererseits um Spieler zu gruppieren. Dabei soll die Einteilung der Mannschaften klären, inwiefern sich die besten Mannschaften von den restlichen Teams differenzieren. Die Identifikation von Spielerkategorien soll dazu dienen, ähnliche und unähnliche Spieler zu ermitteln.

Abschnitt 5.4 untersucht die Möglichkeit der Vorhersage einer gesamten Saison. Hier werden die Regression und die Klassifikation für eine solche Prognose angewendet.

## 5.1 Wer wird der nächste Nationalspieler?

Nationalspieler gelten als die besten Spieler ihres Landes und repräsentieren damit die Vorstellung eines guten bzw. besseren Spielers auf der ihm zugewiesenen Position. In den folgenden Abschnitten wird untersucht wie dieses Wissen zur Findung von gut spielenden Spielern mithilfe von Data Mining unterstützt werden kann. Zunächst wird näher auf die Problemstellung eingegangen sowie die Vorgehensweise in diesem Kapitel erläutert.

Die Spieler einer Fußballmannschaft sind eine der wichtigsten Komponenten eines Fußballvereins. Von ihnen hängt maßgeblich ab, wie ein Verein in der anstehenden Saison abschneidet. Gleichzeitig finanzieren sich Vereine unter anderem durch Spielerverkäufe, wobei bessere Spieler höhere Ablösesummen generieren als leistungsschwache Spieler. Eine wichtige Aufgabe für Vereine ist es somit Spieler zu entdecken, die den Verein sportlich sowie wirtschaftlich verbessern.

Für die Spielersuche bzw. den Spielersucher werden meist die englischen Begriffe Scouting bzw. Scout eingesetzt. Diese werden auch in der vorliegenden Arbeit genutzt. Nach einem Artikel der Badischen Zeitung [35] investieren die Vereine der Bundesliga siebenstellige Beträge in das Scouting von Spielern. Laut dem Autor des Artikels hat der Verein Bayer 04 Leverkusen acht hauptamtliche Scouts angestellt.

Als „Scouting 2.0" bezeichnet der FC Bayern München seine Spielersuche mithilfe von Datenbanken [23]. Der Chefscout des FC Bayern München beschreibt in dem Interview von Heindl [23], inwiefern Datenbanken die Spielersuche unterstützen. Dabei stellt er heraus, dass mithilfe der gespeicherten Daten eines Spielers Informationen wie Name, Nationalität, Größe oder die Torgefährlichkeit abgerufen werden. Der Artikel „Digitales Scouting" von Spiegel–Online [3] beschreibt ebenfalls den Einsatz von Datenbanken beim Scouting von Spielern. Der Artikel stellt dabei heraus, dass es Scouting-Systeme erlauben, Wunschprofile mit bis zu 30 Kriterien zu erstellen. Diese Profile können dann mit den in der Datenbank gespeicherten Spielern abgeglichen und ausgegeben werden.

Wie die vorliegenden Artikel zeigen, setzen professionelle Fußballvereine Datenbanken im Scouting ein. Inwiefern Data Mining bei Vereinen genutzt wird ist nicht bekannt. Jedoch bietet Data Mining die Möglichkeit das Scouting zu unterstützen. Ein

geeignetes Verfahren ist die Klassifikation, die im Folgenden zum Einsatz kommt. Wie Kapitel 2.2.1.1 beschreibt, dient die Klassifikation dazu, Objekten bestimmten Klassen zuzuordnen, wobei die Klassen vordefiniert sind. Die Objekte bei der Anwendung der Klassifikation zum Scouting sind die Spieler. Diese Objekte werden in zwei Klassen {0,1} unterteilt. Dabei beschreibt die Klasse {1} „gute" Spieler und die Klasse {0} schwächere Spieler.

Es stellt sich die Frage, wie sich gute bzw. schwache Spieler charakterisieren lassen. Ein Ansatz, der hier verwendet wird, ist es die Nationalspieler einzelner Länder als gute Spieler anzusehen. Nationalspieler repräsentieren die besten Spieler eines Landes und werden von den Trainern der Nationalmannschaften zu Freundschafts- bzw. Qualifikationsspielen und Turnieren wie die Europa- und Weltmeisterschaft eingeladen. Die Anzahl an nominierten Spielern für solche Spiele oder Turniere ist begrenzt. So dürfen für die kommende Weltmeisterschaft 2014 in Brasilien insgesamt 23 Spieler endgültig gemeldet werden [17, S. 23]. Wegen dieser Begrenzung werden viele Spieler bei der Nominierung nicht berücksichtigt und genau dies macht sich das Scouting mittels Data Mining in dieser Arbeit zur Hilfe. Die Idee ist es, die Spieler mittels Klassifikation festzustellen, welche von den Trainern nicht berücksichtigt werden, es aufgrund ihre Leistung jedoch verdient hätten. Somit ist man an den fälschlich als Nationalspieler klassifizierten Spielern interessiert, da diese laut Data Mining Methoden in die Klasse der guten Spieler fallen, jedoch als solche nach außen nicht erkennbar sind - nämlich nicht nominiert sind.

In den nächsten Abschnitten wird beschrieben wie das genannte Problem gelöst werden kann. Dabei wird zunächst die Datengrundlage dargelegt und die Daten entsprechend transformiert. Anschließend werden mehrere Data Mining Algorithmen ausgeführt und die Ergebnisse dargestellt. Zuletzt erfolgen der Einsatz und die Evaluierung der angewendeten Klassifikation.

### 5.1.1 Datengrundlage

Aus dem in Kapitel 4 beschriebenen Datenmodell sind zur Lösung des Data Mining Problems die Tabellen **Fakten**, **Spielereigenschaften** und **Nationalspieler** nötig. Diese Tabellen werden zu einer Tabelle **SpielerProSaison** verknüpft, auf der die in diesem Kapitel durchgeführten Analysen basieren.

Aus der Tabelle **Fakten**, welche die Daten von jedem Spiel beinhalten, werden die einzelnen Daten für jeden Spieler aggregiert. Um die einzelnen Werte vergleichbar zu machen, werden die Daten durch die gespielten Minuten geteilt. Beispielsweise hat ein Spieler mit höherer Einsatzzeit typischerweise mehr Ballkontakte als ein Spieler, der kürzere Zeit spielt. Um dies zu umgehen, werden bei der Aggregation die Werte, wie die Ballkontakte oder die Tore eines Spielers für die Saison summiert und durch die Gesamtanzahl an gespielten Minuten der Saison geteilt. Dadurch erhält man den Wert pro gespielte Minute. Ein Beispiel für die Ballkontakte ist in Abbildung 13, als Pseudocode in Structured Query Language (SQL) ausgedrückt, zu sehen. Gesondert betrachtet werden die Werte, welche in Prozentzahlen gespeichert sind, wie das Zweikampfverhalten oder die Passquote. Um diese bei der Aggregation wieder in Prozent abzubilden, werden diese Werte eines Spielers in einem Spiel mit den gespielten Minuten des Spielers in dem Spiel multipliziert, dann summiert und anschließend durch die gespielten Minuten der ganzen Saison geteilt. Dadurch werden außerdem die Werte der Spiele höher gewichtet, in denen ein Spieler eine längere Einsatzzeit hatte im Gegensatz zu den Spielen in denen er kürzer gespielt hat. Abbildung 14 zeigt den Pseudocode für die gewonnene Zweikampfquote.

Die Tabelle **Spielereigenschaften** beinhaltet die Eigenschaften eines Spielers, wie seine Nationalität, sein Gewicht oder die Position auf der er spielt. Für das spätere

**Abbildung 13:** Pseudocode Ballkontakte

```
SELECT
Saison, Spieler, Summe(Ballkontakte) / Summe(gespielte Minuten
    )
FROM fakten
GROUP BY Saison, Spieler
```

**Abbildung 14:** Pseudocode Zweikampf

```
SELECT
Saison, Spieler, Summe(Zweikampf * gespielte Minuten) / Summe(
    gespielte Minuten)
FROM fakten
GROUP BY Saison, Spieler
```

Data Mining Verfahren ist die Spielerposition, wie Abwehr, Mittelfeld oder Angriff eines Spielers von besonderem Interesse. Für jede Position wird ein eigenes Data Mining Modell erlernt, da z.B. die Anforderungen an einen Abwehrspieler andere sind, als die Anforderungen an einen Angreifer.

In der Tabelle **Nationalspieler** sind die Anzahl an Spielen eines Spielers in der zugehörigen Nationalmannschaft der Jahre 2009, 2010, 2011 und 2012 gespeichert. Für die Tabelle **SpielerProSaison** werden mit dieser Anzahl zwei Werte berechnet. So wird anhand der Anzahl an Nationalspielen in der vorangegangenen Saison und der aktuell betrachteten Saison die Spalte **nm1** berechnet. Ist die Anzahl an Spielen dieser zwei Saisons größer als eins, so gilt der Spieler als bereits für die Nationalmannschaft angetreten. Im Gegensatz dazu bedeutet die Spalte **nm** das ein Spieler in der nachfolgenden Saison an Nationalspielen teilgenommen hat. Als Nationalspieler gilt er dann, wenn er in der nächsten Saison mehr als ein Nationalspiel bestritten hat. Für beide Spalten wird der Wert {1} für positiv (z.B. war bereits Nationalspieler) und {0} für negativ (z.B. hat kein Nationalspiel bestritten) definiert. Ein Beispiel SQL-Statement für die Saison 2010 ist in Abbildung 15 zu sehen.

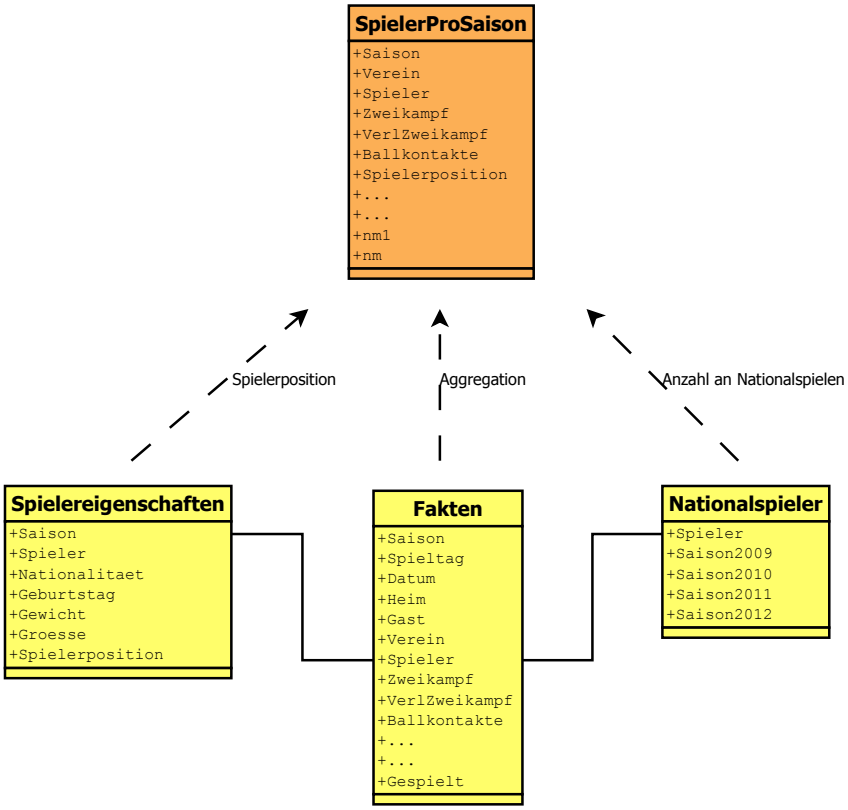
In Abbildung 16 ist die Zusammenführung der Daten dargestellt. Im Anhang ist das gesamte SQL-Statement in Abbildung 23 für die Tabelle zu sehen.

Erste Einblicke in die Daten sollen kleinere Analysen gewähren. Die Analysen werden ohne Torhüter erfolgen, da die Daten von Torhütern die Datengrundlage

**Abbildung 15:** SQL Nationalspieler

```
SELECT
fakten.Saison, fakten.Spieler,
(case when (nationalspieler.saison2009 + nationalspieler.
    saison2010) > 1 then 1 else 0 end) AS nm1,
(case when (nationalspieler.saison2011) > 1 then 1 else 0 end)
     AS nm
FROM fakten, nationalspieler
WHERE fakten.Spieler = nationalspieler.Spieler AND fakten.
    saison = 2010
GROUP BY fakten.Saison, fakten.Spieler
```

**Abbildung 16:** Datengrundlage SpielerProSaison




verfälschen. Torhüter haben spezielle Aufgaben innerhalb des Spiels, wie Bälle halten oder Flanken abfangen. Torhüter werden nicht daran gemessen wie oft sie auf das Tor schießen oder wie viele Ballkontakte sie innerhalb des Spiels haben. Die Torhüterspezifischen Daten liegen nicht vor und Torhüter werden deswegen in der nachfolgenden Analyse sowie dem nachfolgendem Data Mining Verfahren nicht berücksichtigt.

In Tabelle 5 sind für die einzelnen Spielerattribute jeweils die maximalen Werte (Max), minimalen Werte (Min), Mittelwerte (MW) und die Standardabweichung (SD für Standard Deviation) dargestellt.

In Tabelle 6 ist eine Statistik der Nationalspieler für die einzelnen Positionen beschrieben. Insgesamt spielten 192 von 870 Feldspielern in den Saisons 2010 und 2011 in der Nationalmannschaft ihres Landes. Dabei wurden 59 Feldspieler zum ersten mal für die Nationalmannschaft berufen (Debütanten).

Die Statistik über die Anzahl an Nationalspielern zeigt bereits, dass Nationalspieler in der Spielermenge unterrepräsentiert sind. Diese Erkenntnis ist für das Data Mining Verfahren wichtig und wird im folgenden Abschnitt 5.1.2 näher erklärt.

Innerhalb der nachfolgenden Analysen wird zudem eine Klassifikation auf die beschriebenen Daten mit reduzierter Attributmenge angewendet. Dies dient der Dimensionsreduktion und wird als „Feature Selection" bezeichnet. Die Feature Selection dient der Identifikation von aussagekräftigen Attributen und dem Eliminieren von nicht relevanten und redundanten Attribute innerhalb des Datensatzes [11, S. 94]. Durch die Anwendung der Feature Selection wird in diesem Kapitel überprüft, ob mit einer reduzierten Dimensionalität eine höhere Vorhersagegenauigkeit erreicht werden kann. Dabei wird die Attributselektion automatisch von der von Weka zur Verfügung gestellten Methode „CfsSubsetEval" ausgeführt. Hier werden solche Attribute bevorzugt, die mit dem Zielwert stark korrelieren, jedoch untereinander eine geringe Korrelation aufweisen [22, S. 69].

**Tabelle 5:** Statistik SpielerProSaison

| Wert | Min | Max | MW | SD |
|---|---|---|---|---|
| Zweikampf in % | 0,00 | 100 | 48,8642 | 12,7792 |
| VerlZweikampf in % | 0,00 | 100 | 51,1358 | 12,7792 |
| Ballkontakte | 0,00 | 2,00 | 0,6186 | 0,1709 |
| ErfPaesse | 0,00 | 1,25 | 0,3060 | 0,1344 |
| FehlPaesse | 0,00 | 1,00 | 0,0809 | 0,0473 |
| Paesse | 0,00 | 1,3750 | 0,3869 | 0,1444 |
| ErfPassquote in % | 0,00 | 100 | 76,4308 | 11,6682 |
| FehlPassquote in % | 0,00 | 100 | 23,5692 | 11,6682 |
| Torschuesse | 0,00 | 0,2500 | 0,0154 | 0,0167 |
| VergTorschuesse | 0,00 | 0,250 | 0,0141 | 0,0159 |
| ErfTorschussquote in % | 0,00 | 100 | 4,7408 | 7,355 |
| FehlTorschussquote in % | 0,00 | 100 | 95,2592 | 7,355 |
| Torschussvorlagen | 0,00 | 0,2500 | 0,0146 | 0,0171 |
| Vorlage | 0,00 | 0,0227 | 0,0012 | 0,0019 |
| Tore | 0,00 | 0,0200 | 0,0013 | 0,0022 |
| Gelb | 0,00 | 0,0141 | 0,0011 | 0,0015 |
| PassProBallkontakt | 0,00 | 1,00 | 0,6048 | 0,108 |
| ErfPassProBallkontakt | 0,00 | 0,8750 | 0,4719 | 0,1168 |
| TorschussvorlageProPass | 0,00 | 1,00 | 0,0452 | 0,0556 |
| Gespielt in Min. | 1 | 3060 | 1266,5299 | 923,2746 |
| Gewicht | 60 | 96 | 75,6414 | 6,4649 |
| Groesse | 166 | 198 | 182,4195 | 6,2297 |
| Alter | 17 | 37 | 24,8563 | 4,1698 |

**Tabelle 6:** Statistik Nationalspieler

| Position | Anzahl | Nationalspieler | Debütanten |
|---|---|---|---|
| Abwehr | 301 | 56 | 14 |
| Mittelfeld | 377 | 76 | 29 |
| Angriff | 192 | 60 | 16 |
| Gesamt | 870 | 192 | 59 |

### 5.1.2 Klassifikation von Nationalspielern

Das beschriebene Problem, Spieler den vordefinierten Klassen {1} und {0} zuzuordnen, kann mithilfe der Klassifikation gelöst werden. Im weiteren Verlauf des Kapitels werden verschiedene Klassifikationsmethoden durchgeführt und bewertet. Zuerst wird mit einem Datensatz gearbeitet, der nur die Attribute enthält, die von der Spielweise des Spielers abhängen. Die physischen Werte eines Spielers werden außer Acht gelassen und es wird der Fakt ignoriert, dass ein Spieler bereits Nationalspieler war. Der Grund dafür ist, dass man bei dieser Anwendung an gut spielenden Spielern interessiert ist und Attribute wie das Alter, Gewicht oder ob er bereits Nationalspieler war, für einen Verein bei einer Neuverpflichtung in erster Linie nicht interessant sind. Somit sollen nur spielspezifische Werte wie das Zweikampfverhalten, Anzahl an Ballkontakten etc., das Data Mining Verfahren beeinflussen.

In einer weiteren Anwendung wird ein zweiter Datensatz eingesetzt. Bei diesem Datensatz wird zusätzlich das Attribut aufgenommen, ob ein Spieler bereits vorher für die Nationalmannschaft angetreten ist.

Wie bereits erwähnt, wird für jede Spielerposition ein eigenes Data Mining Modell erlernt. Die Anforderungen an einen Abwehrspieler sind im Fußball andere als die eines Mittelfeldspielers oder Angreifers.

Das grobe Vorgehen in diesem Abschnitt wird wie folgt aussehen:

- Durchführen der Klassifikation mit spielspezifischen Werten
    1. Durchführung mehrerer Testfälle
    2. Anwendung der ausgewählten Algorithmen und Darstellung der Ergebnisse
- Durchführung der Klassifikation unter Einbeziehung des Attributs **nm1**
    1. Durchführung mehrerer Testfälle
    2. Anwendung der ausgewählten Algorithmen und Darstellung der Ergebnisse
- Einsatz und Evaluierung der Klassifikation

Der erste Datensatz beinhaltet die spielspezifischen Werte, welche aus folgenden Attributen der Tabelle **SpielerProSaison** bestehen und im Folgenden als Datensatz 1 bezeichnet wird:

- Zweikampf
- VerlZweikampf
- Ballkontakte
- ErfPaesse
- FehlPaesse
- Paesse
- ErfPassquote
- FehlPassquote
- Torschuesse
- VergTorschuesse
- ErfTorschussquote
- FehlTorschussquote

- Torschussvorlagen
- Vorlage
- Tore
- Gelb
- PassProBallkontakt
- ErfPassProBallkontakt
- TorschussvorlageProPass

Hinzu kommt die zu klassifizierende Klasse **nm**.

**Datensatz 1: Durchführung mehrerer Testfälle**

Zur Durchführung der Data Mining Methoden wird die in Kapitel 2.3 beschriebene Data Mining Software Weka benutzt. Die ersten Testfälle werden mit dem Datensatz 1 durchgeführt. Für die einzelne Position werden separate Analysen ausgeführt. Als Testmethode wird die Cross-validation Methode mit zehn Teilmengen gewählt.

Führt man erste Data Mining Methoden mit den Datensätzen durch, werden bei Methoden, wie den Entscheidungsbäumen (J48graft) oder den regelbasierten Methoden (JRip) für z.B. die Position Abwehr keine Bäume bzw. keine Regeln aufgebaut und alle Spieler der Klasse $\{0\}$ zugeordnet. Dieses Verhalten erklärt sich dadurch, dass in dem vorliegenden Fall ein unbalanciertes Klassenverhältnis zugrunde liegt. Von 301 Abwehrspielern sind lediglich 56 Spieler Nationalspieler. Die Methoden wählen den naiven Ansatz und klassifizieren jeden Spieler in die überrepräsentierte Klasse und erhalten damit die Korrektklassifikationsrate von ca. 81 %. Um Probleme in unbalancierten Datensätzen zu umgehen gibt es mehrere Methoden die von Chavla [9] näher vorgestellt werden. Bei der folgenden Anwendung wird das Resampling der Datensätze zum Einsatz kommen, da bei diesem zusätzlich die verhältnismäßig kleinen Datensätze künstlich erweitert werden. Das Resampling kann eine einheitliche Klassenverteilung auf zwei Arten erzielen. Einerseits gibt es das Undersampling, bei dem zufällig ausgewählte Instanzen der überrepräsentierten Klasse aus dem Datensatz gelöscht werden. Die andere Methode kopiert Instanzen aus der unterrepräsentierten Klasse und fügt diese dem Datensatz erneut hinzu. Dies nennt man Oversampling. Mithilfe von Weka lässt sich das Oversampling automatisch ausführen. Das Resampling von Weka wird so eingestellt, dass die Datensätze verdoppelt werden und die Klassenverteilung gleichzeitig ausgeglichen wird.

Der Nachteil bei der Verwendung von Resampling ist, dass ein Cross-validation Test nun nicht mehr verwendet werden kann. Jede Teilmenge die zum Testen verwendet wird kann kopierte Werte aus der Trainingsmenge enthalten. Die Ergebnisse der Cross-validation Methode wären damit verfälscht. Um trotzdem einige Tests durchzuführen werden aus dem ursprünglichen Datensatz die Spieler der Mannschaften des FC Bayern München, des 1. FSV Mainz 05 und des FC Schalke 04 herausgenommen. Auf den reduzierten Datensatz wird das Resampling angewendet. Zum Testen werden die ungesehenen Spieler der drei Mannschaften genutzt.

Zur Bewertung der Klassifikation wird die Wahrheitsmatrix zur Hilfe genommen, welche die vorhergesagten Klassenzuordnungen und die tatsächlichen Klassenzuordnungen gegenüberstellt. Eine solche Matrix ist in Tabelle 7 zu sehen.

Es soll eine möglichst hohe Klassifikationsrate erzielt werden, also möglichst viele Spieler richtig eingeordnet werden. Gleichzeitig ist es aber auch von besonderem Interesse die Nationalspieler richtig zu klassifizieren, da diese als die guten Spieler gelten, welche für diese Anwendung von Bedeutung sind. Somit gilt neben der Klassifikationsrate die sogenannten Precision (Genauigkeit) als wichtigste Maßzahl. Die Precision wird berechnet in dem man die positiv erkannten Instanzen durch die An-

**Tabelle 7:** Wahrheitsmatrix

| | Nicht Nationalspieler | Nationalspieler |
|---|---|---|
| Klassifiziert als Nicht-Nationalspieler | Richtig Negativ | Falsch Positiv |
| Klassifiziert als National-spieler | Falsch Negativ | Richtig Positiv |

zahl tatsächlich positiver Instanzen teilt. Die Berechnung ist in Formel 4 dargestellt.

$$Precision = \frac{RichtigPositiv}{FalschPositiv + RichtigPositiv} \tag{4}$$

Es werden im Laufe der Tests mehrere Klassifikationsmethoden aus den verschiedenen Bereichen wie den Entscheidungsbäumen, den regelbasierten Klassifikationsmethoden, der Bayes Klassifikation und komplexere Algorithmen, wie neuronale Netze ausprobiert. Zusätzlich werden diese Methoden parametrisiert. So werden die Methoden in mehreren Testläufen unter verschiedenen Einstellungen angewendet, bis die Klassifikationen die besten Ergebnisse erzielen.

Es zeigt sich, dass die folgenden Methoden die besten Ergebnisse erreichen:

- Entscheidungsbäume
  - J48graft
  - RandomForest
- Regelbasierte Klassifikation
  - JRip
- Neuronales Netzwerk
  - MutlilayerPerceptron

Die Naives-Bayes Klassifikation ergibt in den Tests eine Klassifikationsrate von nur 40 % und wird somit für die weitere Verwendung als ungeeignet eingestuft. Auch andere von Weka zur Verfügung gestellte Algorithmen können in den Tests nicht überzeugen.

Ein zweiter Testlauf wird mit der Feature Selection der Daten vorgenommen. Bei der Feature Selection werden nur die wichtigsten Attribute bei dem Aufbau des Modells berücksichtigt. Auch bei diesem Testlauf werden dieselben Algorithmen benutzt, jedoch mit verschiedenen Einstellungen erneut getestet. So kann man vergleichen, ob eine Feature Selection zu einer Verbesserung der Klassifikation führt. Die Ergebnisse beider Tests sind in Tabelle 8 dargestellt.

Wie man sieht, führt der Einsatz der Feature Selection zu anderen Ergebnissen. Teilweise schneiden die Klassifizierer mit Feature Selection besser ab, ein klarer Trend ist jedoch nicht zu erkennen. Auf der Position Angriff wird jede Methode positiv von der Feature Selection beeinflusst, sowohl in der Klassifikationsrate als auch bei der Precision. Die besten Resultate erreicht hier die MultilayerPerceptron Methode mit einer Klassifikationsrate von 80,77 % und einer Precision von 0,923. Im Bereich Abwehr hat die JRip Methode ohne Feature Selection den besten Wert mit 88,57 %. Im Mittelfeld erzielen der Entscheidungsbaum J48graft ohne Feature Selection sowie die JRip Methode mit Feature Selection mit 82,05 % und 0,833 die besten Klassifikationsraten und die besten Precisionwerte. Insgesamt kann man anhand der Tests erkennen, dass sich die Spieler mit Bestwerten von über 80 % Klassifikationsrate in allen Positionen voraussichtlich zufriedenstellend zuordnen lassen.

**Tabelle 8:** Datensatz 1: Ergebnisse Tests

| Methode | Position | Rate (Precision) | Rate (Precision)* |
|---|---|---|---|
| J48graft | Abwehr | 77,14 % (0,727) | 82,86 % (0,818) |
| | Mittelfeld | 82,05 % (0,833) | 66,67 % (0,417) |
| | Angriff | 61,54 % (0,692) | 76,92 % (0,923) |
| RandomForest | Abwehr | 82,86 % (0,636) | 80,00 % (0,636) |
| | Mittelfeld | 82,05 % (0,583) | 71,79 % (0,583) |
| | Angriff | 73,08 % (0,769) | 76,92 % (0,769) |
| JRip | Abwehr | 88,57 % (0,727) | 80,00 % (0,727) |
| | Mittelfeld | 79,49 % (0,833) | 82,05 % (0,833) |
| | Angriff | 69,23 % (0,692) | 73,08 % (0,846) |
| MultilayerPerceptron | Abwehr | 85,71 % (0,909) | 71,43 % (0,818) |
| | Mittelfeld | 74,36 % (0,667) | 71,79 % (0,667) |
| | Angriff | 76,92 % (0,769) | 80,77 % (0,923) |

* Mit Feature Selection

**Datensatz 1: Anwendung und Ergebnisse**

Ziel der Anwendung ist es herauszufinden, welche Spieler vom Data Mining als Talente bzw. gute Spieler erkannt werden. Aufgrund der Erkenntnisse aus den Testfällen werden die verschiedenen Data Mining Algorithmen parametrisiert. Es werden nicht die erlernten Modelle aus den Testfällen benutzt, sondern für einen Durchlauf werden jeweils neue Modelle erlernt und auf ungesehene Daten angewendet. Dies wird wie folgt aussehen: Ein Durchlauf besteht aus zwei Datensätzen, einem Trainingsdatensatz und einem Datensatz auf dem das erlernte Modell angewendet wird. Der Trainingsdatensatz beinhaltet die Spieler der 20 Mannschaften aus zwei Saisons, abgesehen von genau einer Mannschaft. Dies sind insgesamt 19 Mannschaften. Mit diesem Trainingsdatensatz werden jeweils die ausgesuchten Data Mining Modelle erlernt. Angewendet werden die erlernten Modelle auf die Spieler der ausgelassenen Mannschaft. Pro Durchlauf wird immer eine vorher noch nicht ausgelassene Mannschaft ausgelassen. Es gibt somit insgesamt 20 Durchläufe für jede Methode. Zum Abschluss wird für jede Methode die in den Durchläufen klassifizierten Spieler zusammengetragen und die Ergebnisse dargestellt.

Bei dieser Vorgehensweise werden für jede Position und Methode insgesamt 20 Modelle erlernt, die im Endeffekt verschieden aussehen. Es wird hier also nicht klassisch ein bestehendes Modell erlernt, sondern mehrere Modelle. Diese gesonderte Anwendung wird deswegen ausgesucht, weil insgesamt wenige Spieler pro Position für die zwei Saisons vorliegen. Sofern mehrere Spieler zur Verfügung stehen, könnte man auch ein Modell erlernen und dies auf ungesehene Daten anwenden, um zu sehen, wie sich die Klassifikation auswirkt.

Wie in Tabelle 9 zu sehen ist, sind die Ergebnisse der Anwendung schlechter als die der Testfälle. Keine der Methoden erreicht die Klassifikationsrate der naiven Ansätze von 81,4 % für die Abwehr, 79,8 % für das Mittelfeld und 68,8 % für den Angriff. In der Abwehr erlangt der RandomForest Algorithmus ohne Feature Selection die höchste Klassifikationsrate mit 71,8 %, jedoch ist der Wert der Precision mit 0,07 zu schlecht als das die Ergebnisse für den realen Einsatz in Frage kommen. Die meisten Nationalspieler der Abwehr klassifiziert die MultilayerPerceptron Methode ohne Feature Selection mit 22 von 56 Nationalspielern richtig. Die Klassifikationsrate ist mit 61,1 % zwar nicht die Beste aber nur geringfügig schlechter als die der anderen Methoden. Im Mittelfeld überzeugen die RandomForest Methode sowie die Multi-

**Tabelle 9:** Datensatz 1: Ergebnisse

| Methode | Position | Rate (Precision) | Rate (Precision)* |
|---|---|---|---|
| J48graft | Abwehr | 61,8 % (0,232) | 64,4 % (0,25) |
| | Mittelfeld | 65,5 % (0,395) | 66,6 % (0,382) |
| | Angriff | 58,5 % (0,483) | 58,5 % (0,417) |
| RandomForest | Abwehr | 71,8 % (0,07) | 69,8 % (0,143) |
| | Mittelfeld | 76,1 % (0,263) | 74,8 % (0,25) |
| | Angriff | 67,4 % (0,433) | 65,8 % (0,433) |
| JRip | Abwehr | 63,8 % (0,296) | 61,8 % (0,232) |
| | Mittelfeld | 66,8 % (0,316) | 65,2 % (0,368) |
| | Angriff | 57,5 % (0,383) | 61,7 % (0,567) |
| MultilayerPerceptron | Abwehr | 61,1 % (0,393) | 56,8 % (0,339) |
| | Mittelfeld | 69,2 % (0,421) | 67,4 % (0,513) |
| | Angriff | 62,2 % (0,4) | 64,2 % (0,517) |

* Mit Feature Selection

layerPerceptron Methode beide ohne Feature Selection. Die RandomForest Methode hat eine hohe Klassifikationsrate (76,1 %) aber eine geringere Precision (0,263). Die MultilayerPerceptron Methode hat mit 0,421 den besseren Precisionwert aber eine schlechtere Klassifikationsrate (69,2 %). Im Angriff kann die MultilayerPerceptron Methode mit Feature Selection bei der Kombination von Klassifikationsrate (64,2 %) und Precision (0,517) überzeugen. Mehr als die Hälfte der Nationalspieler werden hier richtig erkannt.

**Datensatz 2: Durchführung mehrerer Testfälle**

Betrachtet man die Ergebnisse aus dem Datensatz 1, sind diese schlechter als die naiven Ansätze und mit einer richtigen Vorhersage von weniger als der Hälfte an Nationalspielern als schlecht zu bewerten. Ein Versuch die Werte zu erhöhen ist es weitere Attribute in den Datensatz aufzunehmen. Der Datensatz 2 beinhaltet zusätzlich zu Datensatz 1 den Fakt, ob ein Spieler in der Vergangenheit bereits für die Nationalmannschaft angetreten ist. Wie man im Folgenden sehen wird ist gerade dieser Fakt in der Mehrheit der Fälle besonders wichtig. Ein Versuch der Erklärung wird im letzten Abschnitt dieses Kapitels unternommen. Die physischen Werte, wie das Alter oder das Gewicht werden weiterhin nicht berücksichtigt. Zum Testen wird der gleiche Trainings- und Testdatensatz wie im vorigen Testfall benutzt, nur mit dem zusätzlichen Attribut **nm1**. Es werden die gleichen Methoden verwendet um den Einfluss des Attributs zu vergleichen. Die Ergebnisse sind in Tabelle 10 dargestellt.

Im Gegensatz zum Datensatz 1 werden die Ergebnisse in den meisten Fällen verbessert. Beispielsweise erzielt die RandomForest Methode in allen Positionen eine bessere Klassifikationsrate. Bei den anderen Methoden sind die Ergebnisse unterschiedlich. Achtet man auf die maximalen Klassifikationsraten übertrifft der neue Datensatz den Datensatz 1. In allen Postionen ist eine maximale Rate von über 80 % erreichbar. Bei der Wahl der Feature Selection ist keine generelle Verbesserung bzw. Verschlechterung zu erkennen.

**Tabelle 10:** Datensatz 2: Ergebnisse Tests

| Methode | Position | Rate (Precision) | Rate (Precision)* |
|---|---|---|---|
| J48graft | Abwehr | 88,57 % (0,909) | 82,86 % (0,727) |
| | Mittelfeld | 61,54 % (0,417) | 66,67 % (0,667) |
| | Angriff | 76,92 % (0,846) | 76,92 % (0,846) |
| RandomForest | Abwehr | 85,71 % (0,636) | 82,86 % (0,636) |
| | Mittelfeld | 82,05 % (0,583) | 84,62 % (0,583) |
| | Angriff | 80,77 % (0,846) | 80,77 % (0,846) |
| JRip | Abwehr | 85,71 % (0,636) | 85,71 % (0,818) |
| | Mittelfeld | 69,23 % (0,583) | 79,49 % (0,75) |
| | Angriff | 76,92 % (0,769) | 73,08 % (0,615) |
| MultilayerPerceptron | Abwehr | 88,57 % (0,909) | 77,14 % (0,636) |
| | Mittelfeld | 87,18 % (0,833) | 71,79 % (0,75) |
| | Angriff | 73,08 % (0,692) | 69,23 % (0,769) |

* Mit Feature Selection

**Datensatz 2: Anwendung und Ergebnisse**

Wie auch beim Datensatz 1, werden beim Datensatz 2 insgesamt 20 Durchläufe pro Methode absolviert, wobei in jedem Durchlauf eine Mannschaft zum Lernen ausgelassen wird und anschließend die Spieler dieser Mannschaft klassifiziert werden. Wie man in Tabelle 11 sieht, wird mithilfe des zusätzlichen Attributs in fast allen Methoden die Klassifikationsraten gesteigert. Die Precision wird ebenfalls erhöht bzw. erreicht mindestens die gleichen Werte wie die des Datensatzes 1. In vielen Fällen wird über die Hälfte der Nationalspieler richtig klassifiziert. In der Abwehr schneiden die RandomForest und MultilayerPerceptron Methode, beide unter Verwendung der Feature Selection am besten ab. Sie erreichen eine Klassifikationsrate von über 80 % und Precisionwerte von über 0,5. Im Mittelfeld erreicht die RandomForest Methode ohne Feature Selection die beste Rate (78,0 %), hat jedoch eine schwachen Precisionwert von 0,361. Die MultilayerPerceptron Methode mit Feature Selection hat die schwächere Klassifikationsrate von ca. 68 % aber eine Precision von 0,553. Im Angriff erreichen die RandomForest und die MultilayerPerceptron Methode ohne Feature Selection mit Klassifikationsraten von 69,9 % bzw. 67,9 % und Precisionwerte von über 0,5 die besten Resultate.

Datensatz 2 erzielt mit dem zusätzlichen Attribut bessere Werte. Vor allem die Precision wird durch das zusätzliche Attribut **nm1** positiv beeinflusst. Für den Einsatz im Scouting muss man entscheiden welche Ergebnisse man letztendlich einsetzt. Im folgenden Abschnitt werden die klassifizierten Spieler des Datensatzes 2 für das Scouting genutzt, da diese Methoden wesentlich besser die Nationalspieler erkennen. Es werden hierzu die Methoden mit den besten Ergebnissen herangezogen und kombiniert, um verlässlichere Aussagen zu treffen.

**Tabelle 11:** Datensatz 2: Ergebnisse

| Methoden | Position | Rate (Precision) | Rate (Precision)* |
|---|---|---|---|
| J48graft | Abwehr | 77,4 % (0,643) | 79,7 % (0,536) |
| | Mittelfeld | 69,2 % (0,421) | 67,6 % (0,368) |
| | Angriff | 61,7 % (0,57) | 63,7 % (0,583) |
| RandomForest | Abwehr | 81,4 % (0,25) | 83,4 % (0,536) |
| | Mittelfeld | 78,0 % (0,316) | 77,0 % (0,303) |
| | Angriff | 69,9 % (0,5) | 65,3 % (0,517) |
| JRip | Abwehr | 73,4 % (0,446) | 79,4 % (0,643) |
| | Mittelfeld | 66,8 % (0,434) | 68,2 % (0,316) |
| | Angriff | 57,5 % (0,533) | 65,3 % (0,55) |
| MultilayerPerceptron | Abwehr | 72,1 % (0,536) | 80,1 % (0,679) |
| | Mittelfeld | 68,4 % (0,421) | 68,4 % (0,553) |
| | Angriff | 67,9 % (0,517) | 64,2 % (0,667) |

* Mit Feature Selection

### 5.1.3 Einsatz und Evaluierung

Der Einsatz der Methoden zur Talententdeckung soll nun zeigen welche Spieler als Nationalspieler klassifiziert werden. Im Folgenden werden nur die Spieler angezeigt, die laut der Annahme vorher noch nie für die Nationalmannschaft angetreten sind, also den Wert $\{0\}$ für das Attribut **nm1** annehmen. Diese Annahme besagt, dass ein Spieler noch kein Nationalspieler war, sofern er in der aktuell betrachteten Saison sowie der vorangegangen Saison weniger als zwei Spiele für die Nationalmannschaft bestritten hat. Diese Spieler sind also in den Jahren zuvor nicht als beste Spieler aufgefallen. Außerdem werden nur Spieler dargestellt, die in der betrachteten Saison keine Nationalspieler sind und damit im Attribut **nm** den Wert $\{0\}$ besitzen.

#### Abwehr

In der Abwehr erreichen die RandomForest und die MultilayerPerceptron Methode mit Feature Selection die besten Werte hinsichtlich der Klassifikationsrate und der Precision. Die RandomForest Methode ordnet sechs Spieler in die Klasse der Nationalspieler ein, die bisher noch nicht in der Nationalmannschaft angetreten sind. Bei der MultilayerPerceptron Methode sind dies neun Spieler. Es gibt dabei keinen Spieler, der von beiden Methoden positiv klassifiziert wird. Vereinigt man beide Spielermengen gibt es also insgesamt 15 Spieler, welche von den Methoden als Talente in der Abwehr für eine Mannschaft in Frage kommen. Dabei ist zu beachten, dass ein Spieler je einmal pro Saison in dieser Menge vorkommen kann. In Tabelle 12 sind diese Spieler dargestellt.

Der im Jahr 2013 auffälligste dieser Spieler ist Dante, der 2010 beim Verein Borussia Mönchengladbach spielte. Dante wechselte zum Beginn der Saison 2012\2013 zum Rekordmeister FC Bayern München und wurde im Januar 2013 für die brasilianische Nationalmannschaft berufen. Seine Leistung in 2010 wurde als positiv eingestuft. Seine Leistung in der Saison darauf jedoch nicht. Mithilfe von Data Mining konnte man trotzdem frühzeitig erkennen, welche Qualität der Spieler besitzt. Christian Pander wird als einziger Spieler in beiden Saisons als potenzieller Nationalspieler erkannt. Christian Pander war im Jahr 2007 bereits Nationalspieler Deutschlands, wurde aufgrund vieler Verletzungen jedoch nicht mehr nominiert. Ebenfalls vor 2009 für die jeweiligen Nationalmannschaften nominiert waren u.a. Mikale Silvestre, Alexandre Vasoski, Andreas Hinkel oder Christian Schulz. Juri Judt, Lukas Schmitz und

**Tabelle 12:** Klassifizierte Spieler in der Abwehr

| Saison | Verein | Spieler |
|---|---|---|
| 2010 | Borussia Mönchengladbach | Anderson |
| 2010 | FC St. Pauli | Fabio Morena |
| 2010 | 1. FC Nürnberg | Timothy Chandler |
| 2011 | SC Freiburg | Andreas Hinkel |
| 2011 | 1. FC Nürnberg | Juri Judt |
| 2011 | SV Werder Bremen | Lukas Schmitz |
| 2010 | Eintracht Frankfurt | Aleksandar Vasoski |
| 2010 | FC Schalke 04 | Christian Pander |
| 2010 | Hannover 96 | Christian Schulz |
| 2010 | 1. FC Köln | Christopher Schorch |
| 2010 | Borussia Mönchengladbach | Dante |
| 2010 | Eintracht Frankfurt | Maik Franz |
| 2011 | Hannover 96 | Christian Pander |
| 2011 | SV Werden Bremen | Mikael Silvestre |
| 2011 | 1. FC Kaiserslautern | Willi Orban |

Christopher Schorch waren bereits Spieler in den Junioren-Nationalmannschaften, konnten sich bisher aber nicht merklich in den Mittelpunkt der Bundesliga spielen. Timothy Chandler wurde mehrmals für die Nationalmannschaft der USA nominiert, entschied sich aber für die deutsche Nationalmannschaft zu spielen, für die er bisher jedoch nicht nominiert wurde.

**Mittelfeld**

Im Mittelfeld erreicht die RandomForest Methode ohne Feature Selection die höchste Klassifikationsrate, aber einen vergleichsweise niedrigen Precisionwert. Die MultilayerPerceptron Methode mit Feature Selection schneidet hinsichtlich der Precision am besten ab, hat aber eine niedrigere Klassifikationsrate.

Nationalspieler, die vorher noch nicht für die Nationalmannschaft nominiert wurden, von den Methoden jedoch als solche klassifiziert sind, sind bei der RandomForest Methode neun Spieler. Bei dem MultilayerPerceptron Algorithmus sind es 42 Spieler. Von beiden Methoden werden sechs gleiche Spieler der positiven Klasse zugeordnet, welche in Tabelle 13 zu sehen sind. Ivo Illicevic ist dabei nach der Saison 2010 für die kroatische Nationalmannschaft aufgelaufen. Christian Gentner, David Jarolim und Dawdah Bah hatten bereits Einsätze für ihre Nationalmannschaften. Pierre des Wit war bereits für die Junioren-Nationalmannschaft der unter 21-jährigen nominiert, spielte nach der Saison 2010 für Bayer 04 Leverkusen wurde aber wegen einer schweren Verletzung wenig eingesetzt und spielt inzwischen wieder beim 1. FC Kaiserslautern in der 2. Liga. Raffael stand im Jahr 2013 als Leihspieler beim Champions League Teilnehmer FC Schalke 04 unter Vertrag und wechselte im Sommer zum Erstligisten Borussia Mönchengladbach.

**Angriff**

Im Angriff schneiden die RandomForest sowie die MultilayerPerceptron Methode ohne Feature Selection am besten ab. Die beiden Methoden klassifizieren jeweils 12 bzw. 14 Spieler als Nationalspieler, die vorher nicht für ihre Nationalmannschaft angetreten sind. Fünf Spieler überschneiden sich bei beiden Klassifizierern, welche in Tabelle 14

**Tabelle 13:** Klassifizierte Spieler im Mittelfeld

| Saison | Verein | Spieler |
|---|---|---|
| 2010 | 1. FC Kaiserslautern | Ivo Ilicevic |
| 2011 | VfB Stuttgart | Christian Gentner |
| 2011 | Hamburger SV | David Jarolim |
| 2011 | FC Augsburg | Dawda Bah |
| 2011 | 1. FC Kaiserslautern | Pierre De Wit |
| 2011 | Hertha BSC | Raffael |

**Tabelle 14:** Klassifizierte Spieler im Angriff

| Saison | Verein | Spieler |
|---|---|---|
| 2010 | 1899 Hoffenheim | Denis Thomalla |
| 2010 | FC St. Pauli | Gerald Asamoah |
| 2011 | Hertha BSC | Adrian Ramos |
| 2011 | VfL Wolfsburg | Patrick Helmes |
| 2011 | FC Schalke 04 | Raul |

aufgelistet sind.

Gerald Asamoah und Raul sind hierbei Spieler, die früher bereits für ihre Nationalmannschaft gespielt haben, aber mit über 34 Jahren inzwischen zu alt für diese sind. Denis Thomalla spielt in der Junioren-Nationalmannschaft der unter 19-jährigen für Deutschland, und ist unter den Spielern mit seinen jungen 18 Jahren sicherlich das interessanteste Talent. Adrian Ramos spielt inzwischen für Ghana in der Nationalmannschaft. Patrick Helmes trat bereits für die deutsche Nationalmannschaft an, wurde aber von vielen Verletzungen in seiner Karriere gestoppt.

**Einschätzung der Ergebnisse**

Die Ergebnisse zeigen, dass Spieler mit höherem Niveau entdeckt werden konnten. Viele der Spieler sind bereits früher für Nationalmannschaften angetreten bzw. wurden nach den Saisons für die Nationalmannschaft nominiert. Einige der Spieler konnten sich jedoch bisher nicht auf höherem Niveau beweisen.

Die Methoden erreichen in etwa die gleichen Klassifikationsraten wie die naiven Ansätze, was die Mindestanforderung an geeignete Data Mining Methoden sein sollte. Jedoch sind die Algorithmen, angewendet auf die vorliegenden Daten, nicht robust genug um zuverlässige Ergebnisse zu liefern. Analysiert man die in den Durchläufen aufgestellten Modelle untereinander, so sieht man, dass sowohl bei der Feature Selection als auch bei der Baum bzw. Regelaufstellung unterschiedliche Werte bevorzugt werden. Daraus lässt sich ableiten, dass die Daten nicht eindeutig auf Nationalspieler schließen lassen. Würden die Daten eindeutige Strukturen erkennen lassen, würden die einzelnen erlernten Modelle ein ähnliches Schema aufweisen. Die Daten würden die Modelle beeinflussen.

Kritisch zu betrachten sind ebenfalls die geringen Werte der Precision. Die Modelle sind nicht stark genug die tatsächlichen Nationalspieler zu erkennen. Wäre die Precision hoch genug könnte man davon ausgehen, dass die Methoden zuverlässig gute Spieler klassifizieren. Gleichzeitig hätte man eine höhere Gewissheit, dass die falsch als Nationalspieler klassifizierten Spieler ebenso die Charakteristika von guten Spielern aufweisen.

Wie in den Ergebnissen erkennbar ist, führt der Datensatz 2 mit dem zusätzlichen Attribut **nm1** zu besseren Ergebnissen. Dies liegt daran, dass bereits nominierte Na-

tionalspieler von Trainern bevorzugt werden. Die Nationaltrainer sind daran interessiert eine eingespielte Mannschaft für die Turniere auflaufen zu lassen. Das macht die hier vorgestellte Anwendung schwieriger, da nicht immer die spielspezifischen Werte entscheidend für eine Nominierung sind. Anhand dieser Werte lässt sich somit ein schlechteres Ergebnis erwarten. Eine Möglichkeit wäre es nur Spieler zu betrachten, welche für die Nationalmannschaft debütieren. Diese schaffen es offensichtlich lediglich anhand ihrer Leistung die Trainer zu überzeugen. Die hier zur Verfügung stehenden Daten beinhalten leider zu wenige Debütanten um diese Option zu testen.

Die Attribute der Datensätze sind nur ein Ausschnitt der Daten, die aufgenommen werden können. Mithilfe weiterer Attribute kann versucht werden, die Methoden positiv zu beeinflussen. Auch der Einfluss zusätzlicher nicht spielspezifischer Daten könnten die Modelle beeinflussen, wie die Größe oder das Gewicht. In dieser Arbeit wird der Fokus jedoch auf die spielspezifischen Daten gelegt. Interessant wäre es zu sehen, wie sich eine höhere Anzahl an Daten auf die Anwendung auswirkt.

Zudem wird hier nur ein Ansatz vorgestellt, um gute Spieler zu entdecken. Statt Nationalspieler als gute Spieler anzusehen, ist es auch möglich, zahlreiche andere Annahmen zu treffen. So können z.B. die Anzahl an gewonnenen Titeln oder Spielen aber auch Nominierungen für die Elf des Tages die Klasse von guten Spielern beeinflussen. Auch Trainer oder Scouts könnten individuell gute Spieler der Liga klassifizieren. Die daraufhin erlernten Modelle könnten auf Spieler anderer Ligen angewendet werden.

Die Liste an zugeordneten Spielern kann man zur weiteren Verwendung im Scouting nutzen. So können Scouts die positiv klassifizierten Spieler als vorgefilterte Liste nehmen, um gewisse Spieler näher zu beobachten. Da aufgrund der Entfernung zu ausländischen Ligen dort nur geringfügig Spieler beobachtet werden können, kann eine Vorauswahl anhand der gespeicherten Daten mithilfe der Klassfizierern vorgenommen werden. So können Scouts sich auf Spiele dieser Spieler konzentrieren. Zudem können Spieler, die aufgrund der subjektiven Beurteilung von Scouts als ungeeignet eingestuft werden, erkannt und erneut betrachtet werden. Eine Einstufung von Spielern anhand ihrer persönlichen Werte muss jedoch zuletzt immer eine Person vornehmen. Schlussendlich spielt auch der Charakter einzelner Spieler eine Rolle bei Transfers von Spielern.

Die Anwendung lässt sich zudem auch in die entgegengesetzte Richtung nutzen. So können Verantwortliche die Spieler des eigenen Vereins klassifizieren, um zu sehen, welcher Spieler auf welchem Niveau spielt. Dieses Wissen kann auf Vertragsverhandlungen Einfluss nehmen.

Abschließend kann man sagen, dass eine solche Klassifikation im realen Einsatz Anwendung finden kann, sofern die Ergebnisse zuverlässige Aussagen tätigen lassen. Mithilfe der Klassifikation können so Spieler mit höherer spielerischer Qualität identifiziert werden.

## 5.2 Die Notenvergabe im Fußball

Wie in Kapitel 2.2.1.2 erläutert, versucht die Regressionsanalyse den Zusammenhang zwischen einer abhängigen Variablen und einer oder mehreren unabhängigen Variablen zu beschreiben. Eine Möglichkeit ist es, die Regressionsanalyse für Vorhersagen zu nutzen. Dabei wendet man ein erlerntes Regressionsmodell auf neue, ungesehene Daten an, um Prognosen zu erstellen. Der Ansatz, welcher in diesem Kapitel verwendet wird, dient jedoch nicht zur Aufstellung eines solchen Vorhersagemodells. In den nachfolgenden Abschnitten wird die Regressionsanalyse dazu genutzt, um die Stärke des Zusammenhangs von unabhängigen Variablen und einer abhängigen Variablen aufzudecken. Dazu wird im Folgenden versucht die wichtigsten Werte zu ermitteln, welche die Note eines Spielers in einem Spiel beeinflussen. Zunächst wird näher auf

die Notenvergabe eingegangen. In weiteren Abschnitten wird Weka dazu genutzt einzelne Regressionsmodelle aufzustellen, welche im Anschluss interpretiert werden. Dazu werden ebenfalls die benötigten Daten vorgestellt. Zum Abschluss werden die Ergebnisse evaluiert und ein Ansatz vorgestellt, der zur Verbesserung der Analysen führen soll.

Noten werden von vielen Zeitschriften bzw. Internetportalen vergeben. Meist sind diese Noten an sogenannte Fußball-Managerspiele gekoppelt, wie z.B. das der Zeitschrift Kicker oder das Managerspiel der Seite `comunio.de`. In diesen Spielen können die Teilnehmer Spieler aus den Bundesliga-Mannschaften virtuell kaufen und erhalten für diese pro Spieltag eine bestimmte Anzahl an Punkten. Die Punkte werden einerseits für Tore oder gelbe Karten vergeben, andererseits aber auch für die von Redakteuren zugewiesenen Noten von Spielern [12].

Bei dem Online-Managerspieler `comunio.de` werden die Noten eines Spiels von den Redakteuren der Seite `sportal.de` vergeben. Bei `sportal.de` vergibt pro Spiel ein Redakteur die Noten, welche anschließend mit einem Notenkoordinator diskutiert werden. Laut `sportal.de` fließen die spielspezifischen Daten, wie die Tore oder die Zweikampfbilanz eines Spielers in die Note ein. Schlussendlich bleibt die Beurteilung der Leistung jedoch eine subjektive Einschätzung des Redakteurs [21]. Das Managerspiel des Kicker Magazins basiert auf einem ähnlichen Prinzip. Hier benoten zwei Redakteure die Spieler eines Spiels [34]. Auch bei der Bild Zeitung bewerten die Redakteure die Spieler aufgrund ihrer subjektiven Meinung. Es soll jedoch die komplette Leistung in die Note einfließen, wie Tore, Vorlagen sowie gewonnene und verlorene Zweikämpfe eines Spielers [33].

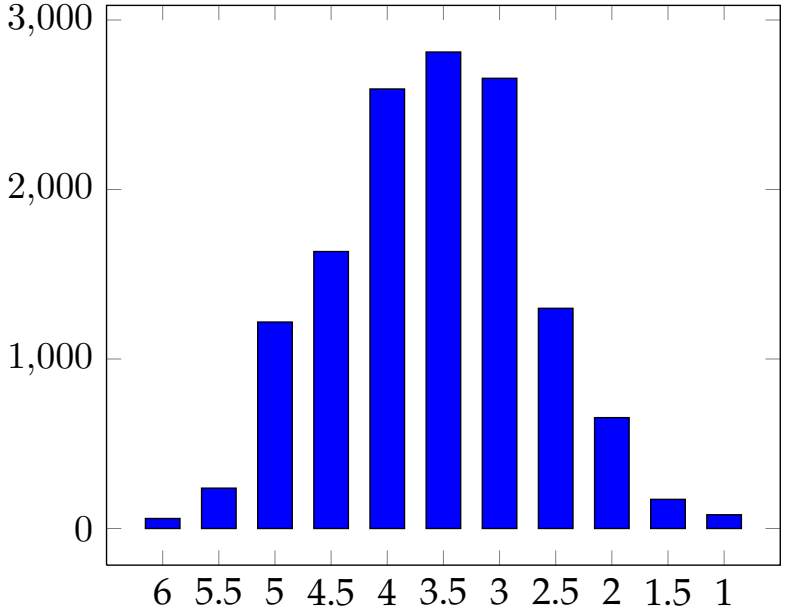
Dass die Noten nicht nur von Managerspielen genutzt werden, glaubt der Kicker-Chefredakteur Klaus Smentek. Laut ihm ist der Stellenwert der Noten bei Spielern als auch bei Fans hoch [34]. Nach einem Artikel von Spiegel-Online spielen die Noten eine wichtige Rolle bei den Spielern [33]. Der Artikel stellt zudem heraus, dass viele Spieler die Notenvergabe nicht nachvollziehen können und sich oftmals falsch bewertet fühlen. Bei den interviewten Spielern herrscht die Meinung vor, dass bei der Notenvergabe die Sympathie des Spielers oder der Kooperationswille mit Redakteuren eine wichtige Rolle einnehmen.

Trotz der Aussagen, dass ausschließlich die Leistung in die Notenvergabe einfließt ist unklar wie sich die Noten tatsächlich zusammensetzen. Mithilfe von Data Mining kann versucht werden, Zusammenhänge zwischen den Daten und der Note zu enträtseln. Die beschriebene Regressionsanalyse wird im weiteren Verlauf als Verfahren genutzt, um diese Abhängigkeiten aufzudecken.

Geht man von der Annahme aus, dass ausschließlich die Leistung die Note bestimmt, kann das Wissen über die Stärke der Abhängigkeiten nützlich sein. So können Spieler anhand der herausgefundenen wichtigen Attribute neu beurteilt werden. Hängt eine gute Leistung beispielsweise stark vom Zweikampfverhalten ab, so können Spieler gezielt nach dieser Eigenschaft ausgewählt werden.

In den USA wurden unter dem Begriff Sabermetrics neue Analysen im Baseball erstellt. Dabei sollen die Sabermetrics Statistiken Spieler objektiver analysieren als veraltete Statistiken [13, S. 1]. Die in diesem Kapitel gefundenen Zusammenhänge können ebenfalls zur objektiveren Bewertung der Spieler beitragen. Mithilfe der bedeutendsten Faktoren kann die Leistung von Spielern neu eingeschätzt werden. Beispielsweise kann ein Angreifer nicht nur an geschossenen Toren gemessen werden, sondern anhand der Kombination der als wichtig eingestuften Attribute.

**Abbildung 17:** Verteilung der Noten




### 5.2.1 Datengrundlage

Die Daten, die zur Aufstellung der Regressionsanalyse benötigt werden sind in der Tabelle **Fakten** sowie der Tabelle **Noten** gespeichert. Zusätzlich ist die Spielerposition aus der Tabelle **Spielereigenschaften** nötig. Das gesamte Datenmodell dieser Arbeit ist in Kapitel 4 beschrieben.

Die Tabelle **Fakten** beinhaltet die Daten eines jeden Spiels. Hierzu sind für jeden Spieler seine spielspezifischen Daten für das gegebene Spiel gespeichert.

Die Noten der Spieler sind in der Tabelle **Noten** hinterlegt. Die Noten sind aus der `fussballdaten.de` Webseite gelesen worden. Die Notenvergabe orientiert sich am Schulnotensystem, d.h. die beste zu erreichende Note ist eine 1 und die schlechteste eine 6. Die Abstufung erfolgt in 0,5 Schritten. Wie die Notenvergabe bei `fussballdaten.de` erfolgt ist nicht näher bekannt. Es ist davon auszugehen, dass die Noten ähnlich zu den beschriebenen Beispielen vergeben werden.

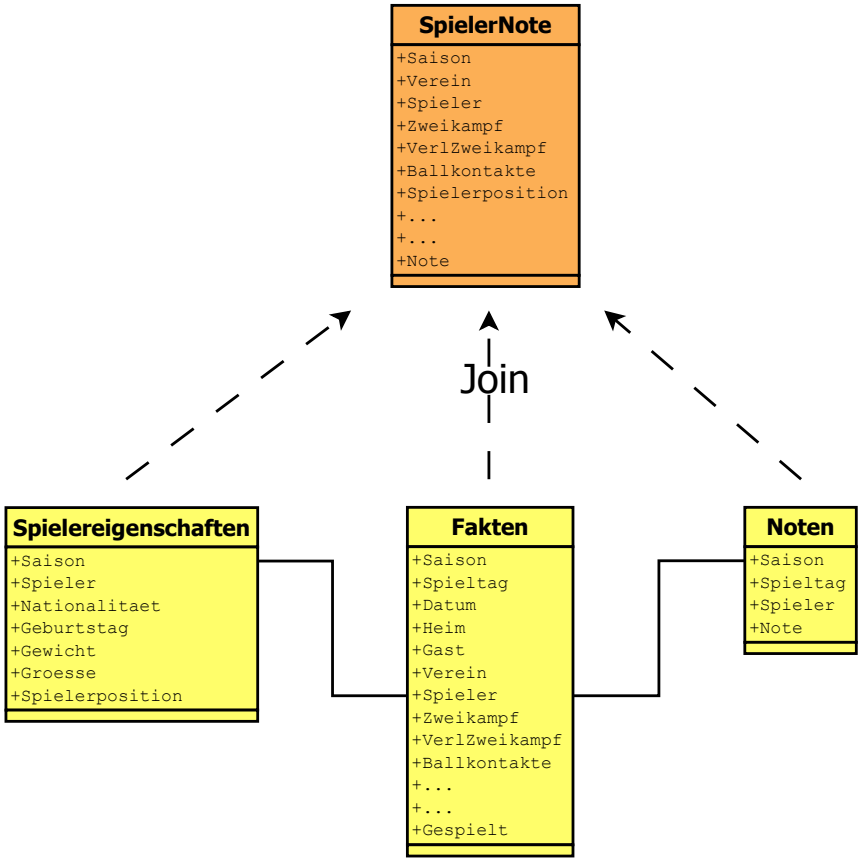
Insgesamt sind 13415 Noten in der Tabelle **Noten** gespeichert. Es stehen weniger Noten zur Verfügung als tatsächlich Spieler gespielt haben. Dies liegt daran, dass ein Spieler eine bestimmte Zeit an Minuten auf dem Feld gestanden haben muss, bevor eine Note vergeben wird. Bei `kicker.de` muss ein Spieler mindestens 30 Minuten am Spiel teilgenommen haben, um eine Note zu erhalten [33]. In der vorhandenen Tabelle von `fussballdaten.de` ist die geringste Einsatzzeit 16 Minuten bei der ein Spieler noch benotet wurde.

In Abbildung 17 ist die Verteilung der Noten der Saisons 2010 und 2011 zu sehen. Die Note 1 wurde 81-mal vergeben. Die schlechteste Note 6 haben insgesamt 59 Spieler erhalten. Wie man sieht, ist die Note 3,5 mit 2811-mal die am häufigsten vergebene Note.

Zusätzlich ist die Spielerposition aus der Tabelle **Spielereigenschaften** nötig, da verschiedene Regressionsmodelle für die einzelne Position aufgestellt werden. Dies wird gemacht, da z.B. ein Abwehrspieler anhand anderer Eigenschaften gemessen wird, als ein Mittelfeldspieler oder Angreifer.

Um mit den Daten zu arbeiten, werden die drei Tabellen verbunden. In Abbildung 18 sind die beiden Tabellen und ihre Zusammenführung abgebildet. Das zugehörige

**Abbildung 18:** Datengrundlage Noten




SQL-Statement ist im Anhang in Abbildung 24 zu sehen.

Bei der Analyse der Daten erkennt man, dass eine weitere Transformation nötig ist. Die einzelnen Attribute besitzen verschiedene Wertebereiche. So liegen die gewonnenen Zweikämpfe zwischen 0 % und 100 %. Im Gegensatz dazu ist die maximale Anzahl an Ballkontakten nach oben hin offen. Die Anzahl an gelben Karten kann sogar nur Werte zwischen null und zwei annehmen (zwei gelbe Karten bedeuten eine Gelb-Rote Karte und der Spieler wird für das laufende Spiel gesperrt). Mithilfe einer Normalisierung können die Werte in einen Wertebereich abgebildet werden. Dadurch sind die Werte untereinander vergleichbar. Liegt ein gleicher Wertebereich vor, lassen sich Aussagen wie „Variable X ist 3-mal so stark gewichtet wie Variable Y" treffen.

Weka bietet an, den Datensatz automatisch zu normalisieren. Hier werden alle Werte in den Bereich zwischen null und eins abgebildet. Die Note als Zielwert wird nicht normalisiert.

### 5.2.2 Aufstellung des Regressionsmodells

Im Folgenden werden Regressionsanalysen für die einzelnen Spielerpositionen Abwehr, Mittelfeld und Angriff durchgeführt. Die Torhüter werden ausgelassen, da Torhüter-spezifische Daten für diese Arbeit nicht vorliegen.

Weka bietet mehrere Möglichkeiten an, um Regressionsanalysen durchzuführen. Diese reichen von der einfachen linearen Regression bis zur Regression Tree Methode M5P. Auch neuronale Netzte sind in der Lage numerische Werte zu ermitteln. Diese werden im Weiteren jedoch nicht angewendet, da neuronale Netze in erster Linie Modelle aufstellen, welche lediglich eine Zielvariable prognostizieren, statt die Zu-

**Tabelle 15:** Ergebnisse Note

| Methode | Position | RMSE |
|---|---|---|
| Lineare Regression | Abwehr | 0,7559 |
| | Mittelfeld | 0,696 |
| | Angriff | 0,6102 |
| SMOreg | Abwehr | 0,7598 |
| | Mittelfeld | 0,6973 |
| | Angriff | 0,611 |
| M5P | Abwehr | 0,7559 |
| | Mittelfeld | 0,6965 |
| | Angriff | 0,6062 |

sammenhänge aufzudecken. Da in dieser Anwendung die Beschreibung der Zusammensetzung der Zielvariable wichtig ist, werden solche sogenannten „Black Box" Methoden ausgelassen.

Zur Beurteilung der Methoden wird die in Kapitel 2.2.1.3 beschriebene Methode der Cross-validation mit zehn Teilmengen genutzt. Die Qualität der Methoden wird mittels der Wurzel aus der mittleren quadratischen Abweichung, im Englischen Root Mean Squared Error (RMSE) genannt, ermittelt. Diese Maßzahl bestimmt die Abweichung zwischen den vorhergesagten Werten und den tatsächlich beobachteten Werten. Der zugehörige Ausdruck ist in Formel 5 abgebildet, wobei $p_i$ den vorhergesagten Wert und $y_i$ den beobachteten Wert für die Instanz *i* bezeichnet.

$$RMSE = \sqrt{\frac{(p_1 - y_1)^2 + ... + (p_n - y_n)^2}{n}} \tag{5}$$

In dieser Anwendung werden mehrere Regressionsmethoden eingesetzt. Dies sind die folgenden Methoden:

- Lineare Regression
- SMOreg
- M5P Tree

Die ausgewählten Methoden werden mit unterschiedlichen Einstellungen getestet und nur die besten Ergebnisse dargestellt.

Wie die Ergebnisse in Tabelle 15 zeigen, sind die Abweichungen in den Methoden sehr gering. Eine Feature Selection, bei der nur die wichtigsten Attribute für den Zielwert betrachtet werden, verbessert lediglich bei dem M5P Tree den RMSE. Die Verbesserung ist jedoch so gering, dass sie ignoriert werden kann. In der Tabelle sind nur die Ergebnisse ohne die Feature Selection dargestellt.

In der Abwehr und im Mittelfeld führt die lineare Regression zu dem geringsten RMSE. Im Angriff schneidet der M5P Tree um 0,004 besser ab als die lineare Regression. Insgesamt kann festgestellt werden, dass sich die Resultate der Regressionen nur gering unterscheiden.

Die höher bzw. weniger hoch gewichteten Variablen sind in allen drei Modellen sehr ähnlich. Zur Vereinfachung werden im Folgenden nur die Formeln der linearen Regression untersucht. Dabei ist zu beachten, dass die lineare Regression von Weka kollineare Attribute beim Erlernen entfernt. Somit enthalten die folgenden Formeln nicht alle Attribute die der Datensatz enthält.

**Tabelle 16:** Attribute für die Abwehr gegenübergestellt

| Positiv | Negativ |
|---|---|
| -1,4596 * Tore | 1,2581 * erfPassProBallkontakt |
| -1,1973 * Torschussvorlagen | 0,7835 * fehlPaesse |
| -1,1696 * Vorlage | 0,3608 * verlZweikampf |
| -1,1195 * erfPaesse | 0,3598 * TorschussvorlageProPass |
| -0,8796 * Zweikampf | 0,1158 * Gelb |
| -0,5282 * PassProBallkontakt | 0,0897 * fehlPassquote |
| -0,3425 * vergTorschuesse | 0,0797 * fehlTorschussquote |
| -0,0859 * erfTorschussquote | 0,0536 * erfPassquote |
| -0,0753 * Ballkontakte | |

**Abwehr**

Formel 6 zeigt die aufgestellte Formel der linearen Regression für die Abwehr. In Tabelle 16 sind die Attribute nach dem positiven (verringert die Note) bzw. negativen Einfluss (erhöht die Note) getrennt dargestellt sowie nach der Höhe des Einflusses geordnet.

$$\begin{aligned} note = & -0,8796 * Zweikampf + 0,3608 * verlZweikampf \\ & - 0,0753 * Ballkontakte - 1,1195 * erfPaesse \\ & + 0,0536 * erfPassquote + 0,7835 * fehlPaesse \\ & + 0,0897 * fehlPassquote - 0,0859 * erfTorschussquote \\ & - 0,3425 * vergTorschuesse + 0,0797 * fehlTorschussquote \\ & - 1,1973 * Torschussvorlagen - 1,4596 * Tore \\ & - 1,1696 * Vorlage + 0,1158 * Gelb \\ & - 0,5282 * PassProBallkontakt + 1,2581 * erfPassProBallkontakt \\ & + 0,3598 * TorschussvorlageProPass + 3,8115 \end{aligned} \tag{6}$$

Wie man sieht, sind die Attribute Tore (-1,4596), Torschussvorlagen (-1,1973), Vorlagen (-1,1696) und die erfolgreichen Pässe (-1,1195) die wichtigsten Attribute, welche die Note positiv beeinflussen. Etwa um die Hälfte weniger bedeutend als die Tore, ist das erfolgreiche Zweikampfverhalten mit -0,8796 und die Pässe pro Ballkontakt mit -0,5282. Die Anzahl an vergebenen Torschüssen beeinflusst die Note positiv mit -0,3425, während die erfolgreiche Torschussquote und die Anzahl an Ballkontakten mit etwa -0,08 nur minimal positiven Einfluss auf die Note haben.

Negativ belasten die Note vor allem die erfolgreichen Pässe pro Ballkontakt mit 1,2581. Die Fehlpässe (0,7835) und die Prozent an verlorenen Zweikämpfen (0,3608) beeinflussen die Note ebenfalls negativ, aber weniger stark. Während die Torschussvorlagen pro Pass (0,3598) und die Anzahl an gelben Karten (0,1158) die Note noch leicht verbessern, fallen die fehlerhafte Pass- und Torschussquote mit 0,0897 bzw. 0,0797 sowie die erfolgreiche Passquote mit 0,0536 nur gering ins Gewicht.

**Tabelle 17:** Attribute für das Mittelfeld gegenübergestellt

| Positiv | Negativ |
|---|---|
| -2,6718 * Tore | 0,7089 * fehlPaesse |
| -1,8942 * Vorlage | 0,4679 * erfPaesse |
| -1,732 * Ballkontakte | 0,4287 * verlZweikampf |
| -0,7382 * vergTorschuesse | 0,3522 * erfPassProBallkontakt |
| -0,6939 * PassProBallkontakt | 0,0734 * fehlTorschussquote |
| -0,4971 * Torschussvorlagen | 0,0628 * fehlPassquote |
| -0,4428 * TorschussvorlageProPass | 0,0582 * Gelb |
| -0,3893 * Zweikampf | 0,0485 * erfPassquote |
| -0,1222 * erfTorschussquote | |

**Mittelfeld**

In Formel 7 und in Tabelle 17 sind die Ergebnisse der linearen Regression für das Mittelfeld dargestellt.

$$\begin{aligned} note = & -0,3893 * Zweikampf + 0,4287 * verlZweikampf \\ & - 1,732 * Ballkontakte + 0,4679 * erfPaesse \\ & + 0,0485 * erfPassquote + 0,7089 * fehlPaesse \\ & + 0,0628 * fehlPassquote - 0,1222 * erfTorschussquote \\ & - 0,7382 * vergTorschuesse + 0,0734 * fehlTorschussquote \\ & - 0,4971 * Torschussvorlagen - 2,6718 * Tore \\ & - 1,8942 * Vorlage + 0,0582 * Gelb \\ & - 0,6939 * PassProBallkontakt + 0,3522 * erfPassProBallkontakt \\ & - 0,4428 * TorschussvorlageProPass + 4,3956 \end{aligned} \tag{7}$$

Im hohen Maße wird die Note positiv von den Toren mit -2,6718 beeinflusst. Etwas schwächer gehen die Vorlagen (-1,8942) und die Anzahl an Ballkontakten (-1,732) positiv in die Note ein. Die vergebenen Torschüsse und die Pässe pro Ballkontakt beeinflussen mit etwa -0,7 die Note mehr als 3-mal weniger positiv als die Anzahl an Toren. Die Torschussvorlagen, die gewonnene Zweikampfquote sowie die Torschussvorlagen pro Pass wirken ebenfalls positiv auf die Note ein, mit etwa -0,4 im Vergleich aber gering. Mit -0,1222 beeinflusst die erfolgreiche Torschussquote die Note noch geringer.

Am negativsten wirkt die Anzahl an fehlerhaften Pässen auf die Note ein, beeinflusst mit 0,7089 die Note jedoch 3-mal weniger als die Tore. Die erfolgreichen Pässe, die verlorene Zweikampfquote und die erfolgreichen Pässe pro Ballkontakt belasten die Note mit 0,4679 bis 0,3522 zusätzlich negativ. Ebenfalls negativen Einfluss besitzen die fehlerhafte Torschussquote und die Quote an fehlerhaften Pässen mit etwa 0,07. Die gelben Karten und die erfolgreiche Passquote haben mit etwa 0,05 nur geringe Einwirkung auf die Zielvariable.

**Angriff**

Die Ergebnisse aus Formel 8 und Tabelle 18 zeigen, dass die Anzahl an geschossenen Toren mit etwa -3,4 den größten positiven Einfluss auf die Note des Angriffs haben. Mit etwa -2 ist die Anzahl an Vorlagen ebenfalls sehr wichtig für die Zusammensetzung der Note. Mit rund -1 wirkt die Anzahl an Torschüssen ebenfalls positiv auf die

**Tabelle 18:** Attribute für den Angriff gegenübergestellt

| Positiv | Negativ |
|---|---|
| - 3,4207 * Tore | 0,278 * fehlTorschussquote |
| - 1,9409 * Vorlage | 0,2151 * fehlPassquote |
| - 1,0935 * Torschuesse | 0,2123 * verlZweikampf |
| - 0,7666 * Torschussvorlagen | 0,1152 * erfPassProBallkontakt |
| - 0,7645 * Ballkontakte | 0,1134 * vergTorschuesse |
| - 0,2473 * erfTorschussquote | 0,0972 * PassProBallkontakt |
| - 0,2202 * Zweikampf | 0,0216 * Gelb |
| - 0,1663 * erfPaesse | |
| - 0,15 * fehlPaesse | |
| - 0,1087 * TorschussvorlageProPass | |
| - 0,0933 * erfPassquote | |

Note ein. Die Torschussvorlagen und Ballkontakte sind mit etwa -0,76 fast 5-mal weniger wichtig als die Tore. Die erfolgreiche Torschussquote (-0,2473) und die gewonnenen Zweikämpfe (-0,2202) haben geringen Einfluss auf die Note. Leicht positiven Einfluss, mit Koeffizienten zwischen -0,17 und -0,09 haben die Anzahl an erfolgreichen und fehlerhaften Pässen, die Torschussvorlagen pro Pass und die erfolgreiche Passquote.

$$
\begin{aligned}
note = &-0,2202 * Zweikampf + 0,2123 * verlZweikampf \\
&- 0,7645 * Ballkontakte - 0,1663 * erfPaesse \\
&- 0,0933 * erfPassquote - 0,15 * fehlPaesse \\
&+ 0,2151 * fehlPassquote - 1,0935 * Torschuesse \\
&- 0,2473 * erfTorschussquote + 0,1134 * vergTorschuesse \\
&+ 0,278 * fehlTorschussquote - 0,7666 * Torschussvorlagen \\
&- 3,4207 * Tore - 1,9409 * Vorlage \\
&+ 0,0216 * Gelb + 0,0972 * PassProBallkontakt \\
&+ 0,1152 * erfPassProBallkontakt \\
&- 0,1087 * TorschussvorlageProPass \\
&+ 4,4604
\end{aligned}
\tag{8}
$$

Die Attribute der negativen Seite haben im Vergleich geringen Einfluss. Mit 0,278 hat die fehlerhafte Torschussquote noch den größten Einfluss, zeigt aber dass sie über 12-mal weniger wichtig ist als die Anzahl an Toren. Mit etwas über 0,2 ist die fehlerhafte Passquote sowie die Quote der verlorenen Zweikämpfe weniger bedeutend für die Zusammensetzung der Note. Die erfolgreichen Pässe pro Ballkontakt, die vergebenen Torschüsse und die Pässe pro Ballkontakt wirken mit rund 0,1 ebenfalls negativ auf die Note ein. Die Anzahl an gelben Karten (0,0216) haben nur gering negativen Einfluss.

**Zusammenfassung der Ergebnisse**

Das mit Abstand wichtigste Attribut für den **Angriff** ist die Anzahl an Toren. Ein Umstand der zu erwarten ist, da Stürmer in der öffentlichen Wahrnehmung fast ausschließlich an den Toren gemessen werden. Danach ist es von Bedeutung viele Vorlagen zu geben und somit nicht selbst das Tor zu erzielen, sondern den Mitspieler in

Szene zu setzen. Zudem sind die Anzahl an Torschüssen und die Torschussvorlagen wichtige Attribute. Dies zeigt, dass es für Stürmer am Wichtigsten ist in den offensiven Bereichen gute Werte zu erlangen. Sie müssen viele Tore bzw. Torversuche tätigen und viele Tore bzw. Torschüsse auflegen. Des Weiteren ist es wichtig genügend Ballkontakte zu erhalten. Das bedeutet, dass ein auffälliger Stürmer mehr Chancen hat eine gute Note zu erhalten als ein unauffälliger. Positiv beeinflusst wird die Note im Sturm außerdem durch die erfolgreiche Torschussquote und das positive Zweikampfverhalten, jedoch wesentlich geringer als die anderen Werte.

Auf der negativen Seite wird die Note nur gering beeinflusst. Dabei ist die fehlerhafte Torschussquote ein schlechtes Merkmal, also eine negative Eigenschaft für die Offensive. Auch ein negatives Zweikampfverhalten und eine negative Passbilanz beeinflussen die Note negativ. Das bedeutet, dass Stürmer insgesamt ein gutes Zweikampfverhalten besitzen sollten, genauso wie ein sicheres Passspiel.

Im **Mittelfeld** findet sich die Anzahl an Ballkontakten unter den drei bedeutendsten Attributen wieder. Dies heißt, dass die Leistung eines Mittelfeldspielers dann besser angesehen wird, wenn er viel am Spiel teilnimmt und viele Ballkontakte hat. Am wichtigsten ist es jedoch viele Tore zu schießen und viele Tore vorzulegen, also offensiv gute Werte zu erreichen. Die Note wird ebenfalls anhand der Anzahl an vergebenen Torschüssen verbessert. Dies ist interessant, da es eigentlich ein negativ einzuschätzender Wert ist. Es scheint jedoch so, dass Redakteure darauf achten, dass ein Mittelfeldspieler oft aufs Tor schießt, unbeachtet dessen ob er trifft oder nicht. Für einen Mittelfeldspieler ist es zudem wichtig den Ball schnell weiterzuleiten. Dies ist daran zu erkennen, dass ein guter Wert bei den Pässen pro Ballkontakt die Note positiv beeinflusst. Viele Torschussvorlagen sind außerdem bedeutend. Mittelfeldspieler sollen die Stürmer anspielen, und je öfter ein Pass beim Stürmer ankommt und so die gegnerische Abwehr überwunden wird, desto besser wird er bewertet. Ein gutes Zweikampfverhalten hat zudem gering positiven Einfluss. Die lineare Regression zeigt, dass Mittelfeldspieler eher an ihren offensiven Eigenschaften gemessen werden. Der positive Einfluss des Zweikampfverhaltens deutet darauf hin, dass gute defensive Eigenschaften ebenfalls gewürdigt werden.

Negativ beeinflusst die Note eines Mittelfeldspielers die Anzahl an Fehlpässen. Dies war zu erwarten, da Mittelfeldspieler die Schnittstelle zwischen Abwehr und Angriff sind und dafür sorgen müssen, dass die Bälle erfolgreich weitergeleitet werden. Wird der Ball oft zum Gegner gepasst, bewerten Redakteure dies negativ. Interessant ist, dass die Anzahl an erfolgreichen Pässen und erfolgreichen Pässen pro Ballkontakt ebenfalls negativ einwirken. Das steht auch im Widerspruch mit dem positiven Attribut der Pässe pro Ballkontakt, da dies aussagt, dass viele Pässe scheinbar positiv bewertet werden. Eher zu erklären ist, dass die Quote an verlorenen Zweikämpfen negativ auf die Note einwirkt, da diese Quote offensichtlich ein unerwünschtes Verhalten für Mittelfeldspieler ist.

In der **Abwehr** fällt auf, dass die offensiven Qualitäten die wichtigsten sind. Die Anzahl an Toren, Torschussvorlagen und Vorlagen sind die wichtigsten Attribute. Im Gegensatz zum Mittelfeld bewirken die erfolgreichen Pässe, dass der Spieler besser bewertet wird. Gerade für Abwehrspieler ist es wichtig, dass Pässe in der Nähe des eigenen Tores den eigenen Mitspieler erreichen, da ein Fehlpass schnell zu einem Gegentor führen kann. Dies zeigt auch, dass die Anzahl an Fehlpässen negativen Einfluss auf die Note hat. Im Gegensatz zu den anderen Positionen hat das Zweikampfverhalten größeren positiven Einfluss. Dies war zu erwarten, da Abwehrspieler hauptsächlich eingesetzt werden, um Aktionen der Gegenspieler zu verhindern. Des Weiteren wird positiv bewertet, wenn ein Abwehrspieler viele Pässe pro Ballkontakt spielt, was bedeutet, dass der Ball schnell weitergespielt wird um so schnell Angriffe zu kreieren.

Der Zusammenhang der Ballkontakte mit der Note ist nur sehr gering. Ein Abwehrspieler muss somit nicht zwangsweise viel in das Spiel integriert sein, um eine gute Note zu erhalten. Noch geringen Einfluss haben die vergebenen Torschüsse, was bedeutet, dass Aktionen in der Offensive auch positiv bewertet werden. Erfolgreiche Pässe pro Ballkontakt haben einen hohen negativen Einfluss. Eine Erklärung dafür fehlt leider, genauso warum viele Torschussvorlagen pro Pass schlecht bewertet werden. Dass verlorene Zweikämpfe und die Anzahl gelber Karten geringen negativen Einfluss haben, ist dagegen offensichtlich.

In allen drei Spielerpositionen sind die Anzahl an Toren sowie die Anzahl an Vorlagen unter den drei wichtigsten Attributen. Für alle Spieler gilt somit, dass ihre Leistung als besser angesehen wird, sobald sie an einem Tor direkt bzw. indirekt beteiligt sind. Dies ist darauf zurückzuführen, dass Tore im Fußball die wichtigsten Ereignisse sind und somit am meisten gewürdigt werden. Interessant ist diese Tatsache für Managerspiele, da sich Tore und Vorlagen doppelt auf die Gesamtpunktzahl eines Spielers auswirken. Erstens wird eine bessere Note erzielt und zweitens werden die Tore und Vorlagen mit extra Punkten im Spiel gewürdigt. Neben den Attributen Tore und Vorlagen sind in der Abwehr zusätzlich die erfolgreichen Pässe und das Zweikampfverhalten wichtig, also ein sicheres Passspiel und gutes Abwehrverhalten. Im Mittelfeld liegt der Fokus, neben den Toren und Vorlagen, auf dem Attribut Ballkontakte. Das heißt, dass ein Spieler, welcher das Spiel durch seine Präsenz steuert eher gewürdigt wird, als ein unauffälliger Spieler, der sich wenig ins Spiel einschaltet. Im Angriff sind die offensiven Attribute, wie die Anzahl an Toren, Torschüssen oder Vorlagen wesentlich bedeutender als bei den anderen zwei Positionen. Stürmer sind dafür zuständig, Tore zu erzielen bzw. vorzubereiten. Die Abwehr hat den kleinsten Koeffizienten bei den Toren. Im Mittelfeld ist dieser größer, während er im Sturm den höchsten Wert annimmt. Je eher ein Spieler für die Offensive zuständig ist, umso mehr wird er an den Toren gemessen.

### 5.2.3 Implikationen für den realen Einsatz

Insgesamt zeigen die Ergebnisse aus Tabelle 15, dass die Note im Schnitt zwischen 0,6 und 0,75 falsch prognostiziert wird. Ein Fehler von über einer halben Note ist als kritisch zu betrachten. Die Interpretation der gefundenen Zusammenhänge zeigt jedoch, dass sich einige Attribute in den Formeln stark unterscheiden. Diese sind auch meist nachvollziehbar und stimmen mit der öffentlichen Wahrnehmung überein. Das die erfolgreichen Pässe pro Ballkontakt stark negativ in die Note der Abwehrspieler eingehen ist dabei ein Zusammenhang, der nicht offensichtlich ist. Welche Verbindung es hier zur Note gibt ist unersichtlich.

Beispielhaft soll nun vorgestellt werden, wie das gewonnene Wissen zur neuen Beurteilung von Spielern beitragen kann. Dafür werden die aggregierten Werte der Spieler, die mindestens ein Drittel der Saison 2011 gespielt haben, in die Formeln eingetragen. Diese sind in der Tabelle **SpielerProSaison** gespeichert, welche als Grundlage für die Anwendung von Data Mining im Kapitel 5.1 dient. Dort sind die Werte der Saison 2011 von jedem Spieler summiert und durch Anzahl an gespielten Minuten der Saison geteilt. Die Daten sind nicht normalisiert, sodass die Noten nicht im Bereich von 0 bis 6 zu erwarten sind. Da man aber nur an den niedrigsten Zahlen interessiert ist und diese nicht in einen Kontext gesetzt werden müssen, wird dieser Umstand ignoriert. Eine niedrigere Zahl bedeutet einen besseren Wert als eine Hohe. In Tabelle 19 sind die fünf Spieler mit der niedrigsten Note der Saison 2011 für jede Spielerposition aufgelistet. Acht dieser 15 Spieler wurden in der Saison 2011 von der zugehörigen Multilayer-Perceptron Methode in Kapitel 5.1 als Nationalspieler vorhergesagt. Die Ergebnisse

**Tabelle 19:** Spieler mit den besten Noten der Saison 2011

| Abwehr | Mittelfeld | Angriff |
|---|---|---|
| D.van Buyten (-35,94) | S. Reinartz (7,96) | K. J. Huntelaar (17,79) |
| G. Sankoh (-35,78) | M. Lanig (8,74) | Raul (21,11) |
| Naldo (-35,56) | P. Niemeyer (8,76) | P. Helmes (21,12) |
| S. Rajkovic (-35,30) | W. Kvist (9,35) | S. Kießling (21,31) |
| P. Wollscheid (-32,68) | S. Bender (10,04) | S. Okazaki (21,75) |

sind schwer zu interpretieren. Zum Beispiel schneidet Philipp Lahm, der Kapitän der deutschen Nationalmannschaft, mit einer Note von -15,13 im Vergleich schlecht ab. Im Mittelfeld hat der Spieler der Saison 2011 Marco Reus, ein sehr hohen Wert von 22,22. Somit ist schwer zu sagen, ob dieses Experiment tatsächlich ein guter Messwert für die Qualität von Spielern ist. Trotzdem können die Aufschlüsse für ähnliche Bewertungen genutzt werden.

Wie erwähnt liegen die Vorhersagen im Schnitt über eine halbe Note neben der tatsächlichen Note. Bisher sind nur die spielspezifischen Werte als unabhängige Variablen in die lineare Regression eingegangen. Ein Ansatz zur Verbesserung ist es, weitere Attribute in den Datensatz aufzunehmen. Die Tests zeigen, dass die Feature Selection die Qualität der Methoden nicht verbessert. Dies ist ein Anzeichen dafür, dass viele Attribute zu besseren Ergebnissen führen als wenige. Wie bereits angemerkt, gibt es noch viele weitere Attribute, welche im Fußball aufgenommen werden können. Beispielsweise die gelaufenen Kilometer oder die Anzahl an Sprints pro Spiel. Diese Werte könnten die Regressionen wesentlich beeinflussen und so zu besseren Prognosen führen. Neben diesen spielspezifischen Werten ist es außerdem möglich dem Datensatz Attribute hinzuzufügen, die nicht unbedingt mit der Spielweise des Spielers zusammenhängen aber trotzdem Einfluss auf die Note haben können. Ein Attribut, welches hierbei heraus sticht ist die Differenz von geschossenen Toren zu erhaltenen Toren. Die Idee ist, je höher eine Mannschaft gewinnt umso mehr werden die Leistungen der Spieler der erfolgreichen Mannschaft positiv wahrgenommen. Ein hoher Sieg bedeutet, dass die siegreiche Mannschaft wesentlich besser gespielt haben muss. Gleichzeitig bedeutet es, dass die Leistungen der Verlierer unzureichend waren.

Nimmt man das Attribut Differenz in die verschiedenen Datensätze auf, so werden wesentlich bessere Ergebnisse erzielt. So erreicht die lineare Regression mit dem zusätzlichen Attribut in der **Abwehr** einen RMSE von 0,5984 statt vorher 0,7559. Das Attribut Differenz ist mit -3,3868 das mit Abstand wichtigste Attribut, gefolgt von der Anzahl an Toren mit -1,1697. Die Differenz ist somit fast 3-mal höher gewichtet als die Tore.

Im **Mittelfeld** erreicht man mit dem erweiterten Datensatz einen RMSE von 0,573 (vorher 0,696). Auch hier hat die Differenz den niedrigsten Koeffizienten mit -2,9729. Zum Vergleich hat die Anzahl an Toren den Koeffizienten -1,828.

Der **Angriff** erreicht ohne das Attribut einen RMSE von 0,6102. Mit der Differenz kann dieser auf 0,5363 gesenkt werden. Die Differenz hat hier nicht den höchsten Stellenwert mit -2,3887, sondern die Anzahl an Toren mit -2,688.

Die Ergebnisse zeigen, dass vor allem die Abwehr davon profitiert wesentlich weniger Tore zu erhalten als zu schießen. Im Angriff ist es wichtiger Tore zu schießen, als eine hohe Differenz zu erreichen. Der RMSE konnte in allen Fällen verbessert werden. Zusätzliche Tests haben gezeigt, dass ein Heimvorteil kaum Auswirkung auf die Note hat. Auch ob das Spiel gewonnen wurde oder nicht, ist für die Zusammensetzung der Note nicht entscheidend.

Um das Kapitel abzuschließen, kann man sagen, dass die lineare Regression gehol-

fen hat, die Zusammenhänge zwischen den Attributen und der Note aufzudecken. Auch wenn die Qualität der Ergebnisse nicht vollständig überzeugt, so ist das aufgedeckte Wissen durchaus nutzbar. Wie sich zeigt, vergeben Redakteure die Note hauptsächlich anhand der offensiven Eigenschaften der Spieler. Die spielentscheidenden Aktionen wie Tore und Vorlagen sind dabei die zwei wichtigsten spielspezifischen Attribute, welche die Vergabe der Note bestimmen. Den größten Einfluss hat die Tordifferenz in einem Spiel. Gewinnt eine Mannschaft hoch, so werden die Spieler auch sehr gut bewertet.

Wie beschrieben, können die Ergebnisse für eine neue Beurteilung der Spieler genutzt werden. Konkrete Möglichkeiten können weitere Untersuchungen ergeben. Insgesamt bieten Regressionen eine einfache Darstellung der Beziehungen zwischen mehrere unabhängiger Variablen und einer abhängigen Variable. Regressionen können zudem zur Vorhersage genutzt werden. Damit bietet die Regressionsanalyse vielfältige Möglichkeiten um im Fußball angewendet zu werden.

## 5.3 Einteilung von Teams und Spielern in homogene Gruppen

Der beschriebene Artikel aus Kapitel 3.3 zeigt, dass die Clusteranalyse sinnvolle Anwendung im Bereich des Sports findet. Wie in Kapitel 2.2.2 beschrieben, fällt die Clusteranalyse unter die Kategorie des unsupervised learnings. Im Gegensatz zur Klassifikation oder Regression ist dabei der Zielwert unbekannt. Das Clustering kommt dann zur Anwendung, wenn die Entdeckung von unbekannten Mustern im Vordergrund der Analyse steht. Die Clusteranalyse dient dazu Objekten Klassen zuzuordnen. Hierbei sind die Klassen, auch Cluster oder Gruppen genannt, jedoch nicht vordefiniert und müssen von den Methoden erst identifiziert werden. Die Clusteranalyse versucht, Objekte die ähnlich sind den gleichen Gruppen zuzuordnen und unähnliche Objekte in unterschiedliche Gruppen einzuteilen. Ziel ist es, die Objekte in homogene Gruppen zu unterteilen.

Obwohl die verwandten Arbeiten keine Anwendung im Fußball betrachten, zeigen die Ergebnisse, dass es möglich ist Spieler verschiedener Sportarten anhand ihrer spielspezifischen Werte in Gruppen zu unterteilen, um so Ähnlichkeiten in den Spielweisen aufzudecken. Bisher ist nicht bekannt, inwiefern Vereine die Clusteranalyse im Profifußball anwenden. Dass Data Mining jedoch Möglichkeiten in diesem Bereich bietet, sollen die folgenden Abschnitte zeigen.

Vereine stehen am Saisonende oder zur Winterpause vor dem Problem, dass Spieler ihren Verein verlassen und diese adäquat ersetzt werden müssen. So musste zum Beispiel der Verein Borussia Dortmund nach der Meistersaison 2010/2011 ihren defensiven Mittelfeldspieler Nuri Sahin ersetzen. Mit Ilkay Gündogan vom 1. FC Nürnberg wurde die Lücke erfolgreich geschlossen und Borussia Dortmund konnte die Meisterschaft verteidigen. Nicht so gut reagiert hat der Verein Borussia Mönchengladbach nach dem Abgang des Spielers Marco Reus nach der Saison 2011/2012. Borussia Mönchengladbach hat zwar versucht den Verlust des offensiv agierenden Spielers mit neuen Angreifern zu kompensieren, musste jedoch erkennen, dass dies nicht zufriedenstellend geglückt ist. So ärgerte sich der Trainer Lucien Favre nach dem aus in der Qualifikation zur Champions League mit den Worten: „... Wir haben zu ähnliche Stürmer, die bleiben alle nur im Zentrum. Die Mischung ist nicht gut." [24]. Ein Umstand der mithilfe der Clusteranalyse hätte umgangen werden können. Schließlich ist hier das Ziel andersartige Spieler zu trennen und ähnliche Spieler zusammenzufassen.

In dem Artikel aus Kapitel 3.2 wird mithilfe statistischer Methoden und der linearen Regression versucht die Top-Mannschaften von den durchschnittlichen Mannschaften abzugrenzen. Ziel ist es zu erklären, was besser spielende Teams anders machen als ihre Konkurrenten. In dieser Arbeit soll die Clusteranalyse benutzt werden, um solche

Unterschiede aufzudecken. So wird im weiteren Verlauf ebenfalls versucht, Teams in Gruppen zu unterteilen, um zu identifizieren worin sich Top-Teams ähneln und von anderen Teams abgrenzen.

Im folgenden Abschnitt 5.3.1 wird die Clusteranalyse auf die Mannschaften angewendet. Zuerst werden die Daten für das Data Mining näher erklärt. Daran anschließend wird der EM-Algorithmus angewendet. Abschließend werden die Ergebnisse dargestellt und interpretiert.

Im Abschnitt 5.3.2 wird auf die Spieler eingegangen. Dabei werden für jede Spielerposition wie Abwehr, Mittelfeld und Angriff eigene Clusteranalysen durchgeführt. Auch hier werden zunächst die Daten dargestellt. Zum Abschluss findet eine Darstellung der Ergebnisse und Interpretation dieser statt. Um das Kapitel über die Clusteranalyse abzuschließen, werden die Implikationen für den realen Einsatz eingeschätzt.

### 5.3.1 Clusteranalyse von Mannschaften

In den folgenden Abschnitten wird untersucht, inwiefern sich einzelne Mannschaften ähneln bzw. unterscheiden. Dabei wird ein besonderer Fokus auf die Unterteilung von sehr erfolgreichen Teams und weniger erfolgreichen Teams gelegt. Es gilt herauszufinden, ob es eine klare Abgrenzung zwischen den Teams gibt, die im oberen Drittel der Tabellenhälfte landen zu denen, die sich in der unteren Region der Tabelle einfinden.

#### 5.3.1.1 Datengrundlage

Innerhalb des in dieser Arbeit vorliegenden Datensatzes sind die Daten für die Saison 2010 und 2011 gespeichert. Sobald ein Spieler an einem Spiel teilgenommen hat, wurden die entsprechenden Daten für ihn aufgenommen. Diese reichen von der Anzahl an geschossenen Toren bis zur Quote der gewonnenen Zweikämpfe. Für die folgende Anwendung werden die Werte der gesamten Mannschaft betrachtet. Dies bedeutet, dass im Folgenden die Werte der Spieler eines Vereins zusammengefasst werden. Dadurch wird eine spielspezifische Beschreibung für jede Mannschaft in der Bundesliga erreicht.

Die Daten werden für diese Anwendung pro Halbsaison aggregiert. Dies bedeutet, dass für die Clusteranalyse insgesamt 72 Dateninstanzen zur Verfügung stehen. Diese Zahl setzt sich folgendermaßen zusammen: Pro Saison nehmen 18 Mannschaften an der 1. Bundesliga teil. In jeder Saison werden insgesamt 34 Spiele gespielt. Diese 34 Spiele werden in zwei Hälften ausgespielt, nämlich in einer Hinrunde und einer Rückrunde. Damit gibt es pro Saison genau zwei Runden. Da die Daten von zwei Saisons zur Verfügung stehen, gibt es insgesamt 18 Mannschaften * 2 Runden * 2 Saisons = 72 Instanzen.

Die Basis für die Aggregation bildet die Tabelle **Fakten**. Für jedes Spiel sind hier spielspezifische Daten, wie die Tore, das Zweikampfverhalten oder die Vorlagen eines jeden Spielers der an dem betrachteten Spiel teilgenommen hat, hinterlegt. Die komplette Liste der Attribute, die in der Tabelle **Fakten** gespeichert sind findet sich in Kapitel 4 über die Datengrundlage dieser Arbeit.

Da in der folgenden Anwendung nur die summierten Werte der Vereine von Bedeutung sind, werden die Werte aller Spieler eines Vereins zusammengefasst. Man erhält so die spielspezifischen Werte für eine Mannschaft. Bei den Werten, welche in Prozent angegeben sind, wie die erfolgreiche oder fehlerhafte Passquote werden die summierten Werte durch die Anzahl an Instanzen geteilt. Die berechnete Zahl wird so wiederum in Prozent abgebildet. Bei den Werten bei denen die Mengen gespeichert sind, wie die Anzahl an Torschüssen oder die Ballkontakte, werden die Werte für die

**Abbildung 19:** Pseudocode Aggregation von Vereinen

```
SELECT
Saison, 'Hinrunde' AS Runde, Verein, Spieler,
Summe(Tore),
Summe(ErfPassquote) / Count(*)
FROM fakten WHERE Saison = 2010 AND Spieltag > 0 AND Spieltag
   < 18
GROUP BY Saison, Verein
```

jeweilige Runde lediglich summiert. Hier wird nicht durch die gespielten Minuten geteilt. Dies ist nicht erforderlich, da jede Mannschaft die gleiche Anzahl an Minuten pro Runde spielt. In dieser Arbeit wird angenommen das jedes Spiel 90 Minuten dauert und es keine Nachspielzeit gibt. Da die Nachspielzeit nur ein geringer Teil der gespielten Minuten sind, kann man diese zusätzliche Zeit ignorieren.

Abbildung 19 zeigt einen Beispielbefehl für die Tore und die erfolgreiche Passquote der Hinrunde der Saison 2010 in Pseudocode. Abbildung 25 aus dem Anhang zeigt den gesamten SQL-Ausdruck der Aggregation.

Der Fokus dieser Analyse liegt auf der Unterscheidung von erfolgreichen Mannschaften einer Liga zu den weniger erfolgreichen Teams. Um die identifizierten Cluster zu überprüfen, wird im Anschluss der Clusteranalyse die Mannschaften der Cluster mit dem Tabellenrang der jeweiligen Runde verglichen. In Tabelle **Abschlusstabelle** sind für jede Saison und Runde die jeweiligen Punkte eines Vereins gespeichert, welche dieser für die Runde erreicht hat. Anhand der Punkte ist der „Rang" der Vereine berechnet. Rang bedeutet hier die Tabellenregion, in der eine Mannschaft für die betrachtete Runde gelandet ist. Die Regionen sind in drei Teile unterteilt. Die ersten fünf Mannschaften aus der Tabelle sind die Top-Mannschaften und nehmen den Wert {2} für die Spalte *Rang* an. Sofern es Mannschaften gibt, die anhand der gleichen Punktzahl zusammen Platz 5 einnehmen, fallen diese gemeinsam unter die Top-Mannschaften. Als schlechteste Mannschaften gelten die Vereine, welche die Plätze in den Abstiegsregionen belegen. Dies sind die letzten drei Vereine der Tabelle und besitzen den Wert {0}. Auch hier nimmt eine Mannschaft den Wert {0} an, sofern sie die gleiche Punktzahl besitzt wie die Mannschaft, die auf Platz 16 steht. Die restlichen Mannschaften fallen unter die Kategorie der mittelmäßigen Mannschaften und besitzen den Wert {1}. Jede Runde wird separat betrachtet. Beispielweise ist der 1. FSV Mainz 05 in der Hinrunde der Saison 2010 in Region {2} gelandet. Zusammengenommen mit der Rückrunde erreichte der 1. FSV Mainz 05 auch am Ende der Saison 2010 einen der fünf besten Plätze der Tabelle. Betrachtet man die Rückrunde isoliert, erreichte der 1. FSV Mainz 05 in dieser Runde jedoch nur die mittlere Tabellenregion. Separat gesehen wurde also zuerst der Wert {2} für die Hinrunde und der Wert {1} für die Rückrunde angenommen.

Mithilfe der aggregierten Daten der Vereine wird im Folgenden eine Clusteranalyse durchgeführt und mittels den Werten aus der Abschlusstabelle interpretiert.

#### 5.3.1.2 Durchführung der Clusteranalyse

In diesem Abschnitt wird die Clusteranalyse durchgeführt. Die Clusteranalyse ist näher in Kapitel 2.2.2 beschrieben.

Die Data Mining Software Weka bietet mehrere Algorithmen zur Clusteranalyse an, die in dieser Anwendung zum Einsatz kommen könnten. Im Folgenden wird der

**Tabelle 20:** Matrix Cluster Mannschaften

| | | Clusterzuordnung | | |
|---|---|---|---|---|
| | | 0 | 1 | 2 |
| Rang | 0 | 5 | 6 | 1 |
| | 1 | 19 | 17 | 3 |
| | 2 | 4 | 1 | 16 |

EM-Algorithmus für die Clusteranalyse genutzt. Der Vorteil dieser Methode ist, dass die Anzahl an Clustern durch den Benutzer nicht angegeben werden muss, wie dies beispielsweise für die K-Means Methode der Fall ist. Obwohl man daran interessiert ist, die Daten in drei Teile zu segmentieren (Top-Mannschaften, mittlere Region, Absteiger), ist es nicht sicher, ob mehr bzw. weniger Cluster die Daten besser beschreiben. Deswegen wird die automatische Identifizierung der Anzahl an Clustern des EM-Algorithmus zur Hilfe genommen. Die Methode wird in mehreren Iterationen parametrisiert bis das bestmögliche Ergebnis gefunden wird. Dabei gilt je höher der log-likelihood Wert der gefunden Cluster ist, desto besser ist die gefundene Lösung. Da der Algorithmus abhängig von der initialen Konfiguration arbeitet, werden mehre Durchläufe mit veränderten initialen Konfigurationen durchgeführt.

Die Anwendung der Clusteranalyse führt dazu, dass die Mannschaften in drei Cluster aufgeteilt werden. In Tabelle 20 sieht man die Clusterzuordnungen und den Rang in einer Matrix gegenübergestellt. Wie man sieht, sind Cluster 0 und Cluster 1 hauptsächlich eine Mischung der schlechtesten drei Mannschaften sowie der mittleren Mannschaften. Festzuhalten ist, dass nur eine Mannschaft in Cluster 1 eingeteilt wurde, die es unter die besten fünf einer Runde schafften. Aus Cluster 0 schafften dies immerhin vier Vereine. Cluster 2 besteht zu 80 % aus Top-5-Mannschaften. Nur eine Mannschaft innerhalb dieses Clusters ist in der Abstiegsregion gelandet. Der Spielstil des Clusters 2 ist somit der erfolgreichste und Mannschaften erhöhen ihre Chancen wesentlich, wenn sie ähnlich zu diesen Mannschaften spielen.

In Tabelle 21 sind die Mittelwerte jeden Attributs der einzelnen Cluster dargestellt. Für eine bessere Überschaubarkeit der Ergebnisse wurden die Zellen der Tabelle in eine Farbe aus dem Bereich Grün bis Rot eingefärbt. Der maximale Wert einer Zeile wird in Grün eingefärbt und der Minimale in Rot. Die Zellen der Werte zwischen diesen beiden Extremwerten nehmen je nach Höhe des entsprechenden Mittelwerts eine Farbe zwischen Grün und Rot an. Dabei repräsentieren Werte nahe dem maximalen Wertes eine grünere Farbe. Je niedriger ein Wert wird, desto eher geht die Farbe der entsprechenden Zelle von Grün zu Gelb und schließlich in Rot über.

Es ist zu erkennen, dass bei den Quoten, wie dem Zweikampfverhalten, erfolgreiche- und fehlerhafte Passquote und die erfolgreiche- bzw. fehlerhafte Torschussquote keine neuen Informationen gewonnen werden. Wird z.B. im erfolgreichen Zweikampfverhalten der größte Mittelwert aller Cluster angenommen, wird gleichzeitig in der Quote der verlorenen Zweikämpfe der niedrigste Mittelwert der Cluster erreicht. Gleiches gilt für die Torschuss- und Passquote.

Wie in der Tabelle zu erkennen ist, weisen die Mannschaften aus Cluster 2 und damit die erfolgreichen Mannschaften, in der Mehrheit der Attribute die höchsten Mittelwerte auf. Im negativ zu bewertenden Attribut, der Anzahl an Fehlpässen haben die Mannschaften aus Cluster 2 den zweithöchsten Mittelwert. Dies heißt, dass viele Pässe zum Gegner gingen. Da aber diese Teams insgesamt mehr Pässe spielen, ist dieser Wert verständlich. Anhand der erfolgreichen Passquote kann man sehen, dass die Vereine aus Cluster 2 trotzdem das sicherste Passspiel aufweisen. Gleiches

**Tabelle 21:** Cluster Mannschaften

| Wert | 0 | 1 | 2 |
|---|---|---|---|
| Zweikampf | 48,486 | 48,539 | 50,448 |
| VerlZweikampf | 51,514 | 51,461 | 49,552 |
| Ballkontakte | 8796,212 | 9512,138 | 10743,88 |
| ErfPaesse | 3997,823 | 4870,046 | 6033,423 |
| FehlPaesse | 1247,505 | 1184,319 | 1224,107 |
| Paesse | 5245,329 | 6054,365 | 7257,53 |
| ErfPassquote | 72,727 | 77,329 | 79,339 |
| FehlPassquote | 27,273 | 22,672 | 20,661 |
| Torschuesse | 213,555 | 209,295 | 253,79 |
| VergTorschuesse | 191,803 | 190,461 | 220,901 |
| ErfTorschussquote | 5,279 | 4,245 | 6,994 |
| FehlTorschussquote | 94,721 | 95,755 | 93,007 |
| Torschussvorlagen | 204,517 | 200,3 | 245,739 |
| Vorlage | 18,076 | 15,151 | 27,952 |
| Tore | 21,752 | 18,834 | 32,889 |
| Gelb | 16,835 | 18,902 | 16,236 |
| PassProBallkontakt | 0,573 | 0,609 | 0,643 |
| ErfPassProBallkontakt | 0,429 | 0,482 | 0,525 |
| TorschussvorlageProPass | 0,05 | 0,042 | 0,044 |

gilt für die vergebenen Torschüsse. Die Vereine aus Cluster 2 schießen öfter aufs Tor und verschießen demnach auch wesentlich öfter. Den niedrigsten Wert erhalten die Mannschaften in der Anzahl an gelben Karten. Daraus lässt sich erkennen, dass die Top-Mannschaften weniger grobe Fouls begehen als die anderen Mannschaften. In den Torschussvorlagen pro Pass nehmen die Teams aus Cluster 2 den zweitbesten Wert an. Es wird nicht versucht mit jedem Pass eine Torchance zu kreieren, sondern das Spiel wird behutsam aufgebaut.

Die Ergebnisse zeigen, dass sich die Mannschaften aus Cluster 2 klar gegenüber den anderen Teams abgrenzen und in der Mehrheit der Werte die besten Ergebnisse aufweisen. Eine Empfehlung für den Spielstil kann anhand der Ergebnisse nicht ausgesprochen werden, da Vereine natürlicherweise versuchen in all diesen Attributen die besten Werte zu erspielen.

Vergleicht man die Mannschaften aus Cluster 0 sowie Cluster 1 miteinander, fällt auf, dass die Unterschiede gering ausfallen. Auffällig ist, dass die Teams aus Cluster 0 vor allem in den offensiven Attributen, wie die Anzahl an Toren, Torschussvorlagen, Vorlagen oder Torschüssen bessere Werte als die Mannschaften aus Cluster 1 erzielen aber in den Attributen, wie dem Zweikampfverhalten, Anzahl an Ballkontakten oder dem Passspiel schlechter abschneiden. Die Mannschaften teilen sich hier also in zwei Kategorien. Cluster 0 beschreibt die offensiv stärkeren Mannschaften, welche aber das Passspiel vernachlässigen oder weniger Ballkontakte haben. Im Gegensatz dazu können die Mannschaften aus Cluster 1 bei diesen Attributen bessere Werte erzielen, haben dafür aber in der Offensive ihre Schwächen. Um also in der oberen Region mitspielen zu wollen, müssen Defensive und Offensive gleich erfolgreich verbunden werden.

Ein weiterer Ansatz zur Analyse der Mannschaften ist es die Mannschaftsteile Abwehr, Mittelfeld und Angriff einzeln zu betrachten. Damit ist gemeint, die Spieler der einzelnen Positionen einer Mannschaft zusammenzufassen und separat zu analysieren. Beispielsweise fasst man alle Abwehrspieler einer Mannschaft zusammen und

**Tabelle 22:** Matrix Cluster Abwehr

| | | Clusterzuordnung | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| Rang | 0 | 0 | 0 | 3 | 5 | 4 |
| | 1 | 6 | 1 | 14 | 12 | 7 |
| | 2 | 5 | 7 | 3 | 3 | 2 |

führt darauf eine Clusteranalyse aus. So können Unterschiede in den einzelnen Positionen der Mannschaften identifiziert werden.

Für diese Analysen werden die Daten wie bei der Analyse der gesamten Mannschaft aggregiert, jedoch auf die Spieler der jeweiligen Position wie Abwehr, Mittelfeld oder Angriff gefiltert. So werden nur die Spieler der einzelnen Positionen der Vereine zusammengefasst. Auf diese Daten werden die Clusteranalysen ausgeführt. Hierbei ist zu beachten, dass sich die Anzahl an gespielten Minuten zwischen den Mannschaften unterscheiden. Dies liegt an dem Fakt, dass verschiedene Spielsysteme gespielt werden. Beispielsweise bevorzugen es manche Mannschaften mit zwei Angreifern zu starten, während andere Mannschaften über 90 Minuten nur mit einem Stürmer spielen. Die Daten müssen hier also zusätzlich durch die gespielten Minuten geteilt werden, sodass die Werte vergleichbar sind. Dies gilt nicht für die Attribute bei denen Prozentzahlen gespeichert sind. Diese bleiben vergleichbar, da sie durch die Anzahl an Instanzen geteilt werden. Da Weka innerhalb der Clusteranalyse mit maximal vier Nachkommastellen rechnet und durch die Division sehr kleine Werte angenommen werden, werden die Werte pro Spiel berechnet. Dies wird erreicht indem man die Werte mit 90 multipliziert, da ein Spiel regulär 90 Minuten dauert.

Die Matrix aus Tabelle 22 beschreibt die Verteilung der Mannschaften auf die Cluster der Position **Abwehr**. Tabelle 23 zeigt die eingefärbten Mittelwerte der Clusteranalyse. Cluster 0 und Cluster 1 sind dabei die Cluster, welche in keiner der beobachteten Runden Mannschaften enthalten, die unter den letzten drei der Tabelle landeten. Die anderen Cluster dagegen sind sehr gemischt. Wie die Tabelle zeigt, haben die Vereine aus Cluster 1 gute bzw. beste Werte in den Attributen Zweikampf sowie Ballkontakte und stechen besonders im Passspiel heraus. Auffällig in diesem Cluster ist jedoch, dass in den offensiven Attributen, wie Torschüsse, Torschussvorlagen, Vorlagen und auch bei der Anzahl an Toren eher geringe Mittelwerte erzielt werden. Sieben der acht Mannschaften aus diesem Cluster sind unter den besten fünf Mannschaften gelandet. Eine Mannschaft kann somit ihre Chance erhöhen, wenn die Abwehr weniger an offensiven Aktionen teilnimmt, dafür aber viele Ballkontakte und ein gutes Passspiel aufweist. Auch im Attribut Pass pro Ballkontakt wird ein hoher Wert erzielt.

Das andere Cluster, welches erfolgreichere Mannschaften beschreibt, ist Cluster 0. Hier weisen die Mannschaften in allen Attributen gute Werte auf. Besonders in den offensiven Attributen, wie Torschüsse oder Torschussvorlagen. Grundsätzlich ist also nicht zu sagen, dass sich offensive Qualitäten negativ auswirken.

Die übrigen Cluster beinhalten keine interessanten, neuen Erkenntnisse und in vielen Attributen sind die Unterschiede gering. Die Vereine aus Cluster 3 schießen pro Spiel die meisten Tore. In diesem Cluster sind mit fünf Mannschaften die meisten „Absteiger" eingeteilt. Das bedeutet, dass es nicht besser ist torgefährliche Abwehrspieler zu verpflichten. Ein Indiz dafür, dass sich Abwehrspieler auf den ihren angedachten Positionen fokussieren sollten, statt die Aufgaben der offensiven Spieler zu übernehmen.

Das Clustering im **Mittelfeld** identifiziert insgesamt vier Cluster. Wie die Matrix

**Tabelle 23:** Tabelle Cluster Abwehr

| Wert | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Zweikampf | 59,983 | 59,575 | 57,501 | 58,883 | 56,078 |
| VerlZweikampf | 40,017 | 40,425 | 42,499 | 41,117 | 43,922 |
| Ballkontakte | 64,802 | 76,122 | 60,749 | 59,122 | 52,077 |
| ErfPaesse | 34,509 | 49,404 | 31,773 | 28,104 | 22,348 |
| FehlPaesse | 7,095 | 6,511 | 6,995 | 7,618 | 7,053 |
| Paesse | 41,604 | 55,916 | 38,768 | 35,722 | 29,4 |
| ErfPassquote | 80,21 | 85,43 | 79,174 | 75,541 | 72,325 |
| FehlPassquote | 19,79 | 14,57 | 20,826 | 24,459 | 27,675 |
| Torschuesse | 0,73 | 0,451 | 0,467 | 0,598 | 0,47 |
| VergTorschuesse | 0,676 | 0,408 | 0,455 | 0,538 | 0,44 |
| ErfTorschussquote | 3,652 | 3,25 | 0,713 | 4,287 | 1,924 |
| FehlTorschussquote | 96,348 | 96,75 | 99,287 | 95,713 | 98,076 |
| Torschussvorlagen | 0,897 | 0,579 | 0,689 | 0,667 | 0,556 |
| Vorlage | 0,101 | 0,038 | 0,054 | 0,052 | 0,052 |
| Tore | 0,054 | 0,044 | 0,011 | 0,06 | 0,031 |
| Gelb | 0,122 | 0,102 | 0,104 | 0,121 | 0,1 |
| PassProBallkontakt | 0,623 | 0,706 | 0,616 | 0,586 | 0,548 |
| ErfPassProBallkontakt | 0,511 | 0,619 | 0,499 | 0,455 | 0,408 |
| TorschussvorlageProPass | 0,024 | 0,011 | 0,019 | 0,02 | 0,024 |

**Tabelle 24:** Matrix Cluster Mittelfeld

| | | Clusterzuordnung | | | |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 |
| Rang | 0 | 0 | 6 | 2 | 4 |
| | 1 | 6 | 11 | 2 | 21 |
| | 2 | 8 | 3 | 6 | 3 |

aus Tabelle 24 zeigt, beinhaltet Cluster 0 keinen „Absteiger". Cluster 2 ist ein weiteres Cluster, welches Top-Teams beschreibt. Hier sind sechs von zehn Teams unter den ersten fünf Mannschaften gelandet. Die eingefärbten Ergebnisse aus Tabelle 25 zeigen, dass die Mannschaften aus beiden Cluster in allen Attributen gute Werte erzielen. Die Teams aus Cluster 0 haben die besten Zweikampfwerte, viele Ballkontakte und ein sicheres Passspiel. In den offensiven Attributen, wie die Anzahl an Torschüssen, Vorlagen oder Toren werden gute, aber nicht beste Werte erreicht. Die Mannschaften aus Cluster 2 haben dagegen in diesen Attributen die höchsten Werte und in den anderen Attributen gute, jedoch keine Bestwerte. Cluster 1 beinhaltet die meisten „Absteiger". Die Mannschaften dieses Clusters schießen mehr Tore als die des Clusters 3, haben jedoch die wenigsten Ballkontakte, die schlechtesten Zweikampfwerte und das schlechteste Passspiel. Die offensiven Qualitäten sind somit nicht unbedingt die wichtigsten Eigenschaften. Ein sicheres Passspiel, viele Ballkontakte und gutes Zweikampfverhalten erhöhen die Chancen nicht unter den letzten drei in der Tabelle zu landen.

Im Gegensatz zur Abwehr und zum Mittelfeld, wird in der Clusteranalyse des **Angriffs** ein Cluster (Cluster 0) identifiziert, welches keine Mannschaft enthält, die unter den fünf besten Mannschaften gelandet ist. Die erfolgreicheren Mannschaften beschreibt das Cluster 1 mit acht von zwölf Vereinen unter den fünf besten Mannschaften. Die Matrix der Clusterverteilung ist in Tabelle 26 zu sehen.

Tabelle 27 zeigt, dass die Teams aus Cluster 0 die schlechteste erfolgreiche Passquote

**Tabelle 25:** Tabelle Cluster Mittelfeld

| Wert | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Zweikampf | 49,7298 | 45,9896 | 47,6022 | 46,2712 |
| VerlZweikampf | 50,2702 | 54,0104 | 52,3978 | 53,7288 |
| Ballkontakte | 69,0091 | 51,6541 | 61,3779 | 57,9687 |
| ErfPaesse | 40,9031 | 24,2105 | 32,6002 | 30,6484 |
| FehlPaesse | 8,0582 | 7,6673 | 8,2399 | 7,6119 |
| Paesse | 48,9613 | 31,8777 | 40,8401 | 38,2602 |
| ErfPassquote | 81,344 | 73,3116 | 77,4565 | 77,465 |
| FehlPassquote | 18,656 | 26,6884 | 22,5435 | 22,535 |
| Torschuesse | 1,5378 | 1,4554 | 1,7702 | 1,2624 |
| VergTorschuesse | 1,373 | 1,3404 | 1,5725 | 1,183 |
| ErfTorschussquote | 6,4235 | 4,7928 | 7,6063 | 3,1813 |
| FehlTorschussquote | 93,5765 | 95,2072 | 92,3937 | 96,8187 |
| Torschussvorlagen | 1,7394 | 1,5807 | 2,0785 | 1,4816 |
| Vorlage | 0,1915 | 0,1343 | 0,1961 | 0,1215 |
| Tore | 0,1647 | 0,115 | 0,1977 | 0,0794 |
| Gelb | 0,1111 | 0,109 | 0,0667 | 0,1103 |
| PassProBallkontakt | 0,6901 | 0,5941 | 0,6428 | 0,6372 |
| ErfPassProBallkontakt | 0,5714 | 0,4468 | 0,5084 | 0,5054 |
| TorschussvorlageProPass | 0,0415 | 0,0587 | 0,0591 | 0,0427 |

**Tabelle 26:** Matrix Cluster Angriff

| | | Clusterzuordnung | | | |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 |
| Rang | 0 | 7 | 1 | 2 | 2 |
| | 1 | 13 | 3 | 12 | 12 |
| | 2 | 0 | 8 | 6 | 6 |

und dazu schlechte Werte in der Offensive aufweisen. Die erfolgreiche Torschussquote ist sehr gering und es werden wenige Torschüsse pro Spiel abgegeben sowie die wenigsten Tore geschossen. Will eine Mannschaft demnach erfolgreich spielen, müssen gewisse offensive Qualitäten bei den Angreifern gegeben sein.

Die Teams aus Cluster 1 haben dagegen die besten offensiven Werte. Die Mannschaften aus dieser Gruppe geben die meisten Torschüsse ab und schießen die meisten Tore. Auffällig ist, dass diese Mannschaften im Angriff weniger Zweikämpfe gewinnen, die wenigsten Ballkontakte haben, die wenigsten Pässe spielen und eine niedrige erfolgreiche Passquote aufweisen. Angreifer sollten sich somit auf offensive Aktionen beschränken, statt sich viel ins Spiel einzubinden und viele Pässe zu spielen. Auch die defensiven Qualitäten wie das Zweikampfverhalten sind nicht entscheidend. Die Angriffsreihen dieser Teams, legen zudem wenige Torschüsse auf, was bedeuten kann, dass die Stürmer eher selbst auf Tor schießen, als weiter zu passen. Außerdem sollten Angreifer eine gewisse Aggressivität aufweisen, da die Mannschaften dieses Clusters die meisten gelben Karten erhalten.

Statt die Mannschaftsteile einzeln zu betrachten, kann man als Erweiterung des vorgestellten Ansatzes die beschriebenen 19 Attribute jeder Position in einen Datensatz übertragen. Das bedeutet, dass der Datensatz alle oben genannten Attribute für jede Position beinhaltet und demnach 57 spielspezifische Attribute enthält. Führt man die Clusteranalyse darauf aus, erhält man drei identifizierte Cluster. Wie bei der Analy-

**Tabelle 27:** Tabelle Cluster Angriff

| Wert | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Zweikampf | 37,6471 | 37,9666 | 38,5387 | 40,0588 |
| VerlZweikampf | 62,3529 | 62,0334 | 61,4613 | 59,9412 |
| Ballkontakte | 38,2463 | 36,6995 | 42,4826 | 43,9781 |
| ErfPaesse | 14,2652 | 13,5706 | 19,8533 | 18,3551 |
| FehlPaesse | 5,5935 | 5,0109 | 5,8096 | 6,4753 |
| Paesse | 19,8587 | 18,5814 | 25,6629 | 24,8304 |
| ErfPassquote | 67,5155 | 70.4339 | 74,3755 | 72,945 |
| FehlPassquote | 32,4845 | 29,5661 | 25,6245 | 27,055 |
| Torschuesse | 2,2684 | 3,4781 | 2,2657 | 2,958 |
| VergTorschuesse | 2,0341 | 2,7999 | 1,8451 | 2,5997 |
| ErfTorschussquote | 6,5534 | 14,7603 | 13,8134 | 9,037 |
| FehlTorschussquote | 93,4466 | 85,2397 | 86,1866 | 90,963 |
| Torschussvorlagen | 1,5286 | 1,4873 | 1,4911 | 1,7353 |
| Vorlage | 0,1289 | 0,1664 | 0,1585 | 0,162 |
| Tore | 0,2343 | 0,6782 | 0,4206 | 0,3583 |
| Gelb | 0,0799 | 0,1079 | 0,0732 | 0,0772 |
| PassProBallkontakt | 0,501 | 0,4933 | 0,5796 | 0,5552 |
| ErfPassProBallkontakt | 0,3545 | 0,3596 | 0,4429 | 0,409 |
| TorschussvorlageProPass | 0,0766 | 0,0919 | 0,0663 | 0,0781 |

se der gesamten Mannschaften gibt es zwei Cluster, welche sehr gemischt sind und ein Cluster, welches die Top-Mannschaften beschreibt. Dieses Cluster beinhaltet 17 Mannschaften, wovon 14 in der oberen und die restlichen drei in der mittleren Region landeten. Dabei sind keine neuen Muster erkennbar. Dieses Cluster übertrifft in der Mehrheit der Attribute die anderen Cluster. Die Abwehr weist sowohl in der Offensive, als auch in der Defensive die besten Werte auf. Gleiches gilt für das Mittelfeld. Im Angriff hat dieses Cluster die besten Werte in der Offensive, nämlich den Attributen Torschüsse, Torschussvorlagen, erfolgreiche Torschussquote, Vorlagen und Tore. In den Attributen Ballkontakte, Pässe und erfolgreiche Passquote werden nicht die besten Werte erreicht. Dies bedeutet, dass im Angriff die offensiven Qualitäten die wichtigsten sind. Da hier sonst keine besonderen neuen Informationen erkennbar sind, wird auf die Darstellung der Ergebnisse verzichtet.

### 5.3.2 Clusteranalyse von Spielern

Spieler zu finden, welche sich in ihrer Spielweise ähneln, bietet im Fußball besonders bei Spielereinkäufen gute Anwendungsmöglichkeiten. Verlässt ein Spieler eine Mannschaft, liegt es an den Verantwortlichen des Vereins für adäquaten Ersatz zu sorgen. Ein Spieler der sehr ähnlich zu dem abgegebene Spieler spielt, kann diesen optimal ersetzen ohne das Mannschaftsgefüge zu ändern. In den folgenden Abschnitten wird eine solche Anwendung mithilfe der Clusteranalyse durchgeführt. Zuerst wird auf die Datengrundlage eingegangen und anschließend die Analyse durchgeführt.

#### 5.3.2.1 Datengrundlage

Im Folgenden werden für jeden Spieler seine spielspezifischen Werte für jede Halbsaison aggregiert und durch die gespielten Minuten geteilt. Die Division ist deswegen wichtig, da so eine Vergleichbarkeit der Spieler erreicht wird. Beispielsweise ist es

**Abbildung 20:** Pseudocode Spielerattribute

```
SELECT
Saison, 'Hinrunde' AS Runde, Verein, Spieler,
(Summe(Tore) / Summe(gespielte Minuten))*90,
Summe(ErfPassquote * gespielte Minuten) / Summe(gespielte
    Minuten),
FROM fakten WHERE Saison = 2010 AND Spieltag > 0 AND Spieltag
    < 18
GROUP BY Saison, Spieler
```

nicht hilfreich, Angreifer nur an der Gesamtanzahl an Toren zu messen. Ein Spieler A mit fünf Treffern ist nicht gleich besser einzuschätzen als ein Spieler B mit nur einem Tor, sofern Spieler A insgesamt zehn Spiele für die Tore gebraucht hat. Spieler B hingegen nur ein Spiel gespielt hat. Für eine bessere Vergleichbarkeit muss man die Tore im Vergleich zu der Einsatzzeit setzten. So hat Spieler A pro Spiel 0,5 Tore geschossen. Stürmer B hingegen ein Tor pro Spiel. Spieler B hätte somit eine bessere Torquote als Spieler A.

Die Werte der Spieler werden aus der Tabelle **Fakten** aggregiert, die näher in Kapitel 4 beschrieben ist. In dieser Tabelle sind für jedes Spiel der beiden Saisons die spielspezifischen Werte der Spieler gespeichert. Diese Werte werden für jeden Spieler für die Hinrunde und Rückrunde für beide Saisons aggregiert. Torhüter werden im Folgenden ausgelassen, da Torhüter-spezifische-Daten für diese Arbeit nicht zur Verfügung stehen. Die Spielerposition wird aus der Tabelle **Spielereigenschaften** gelesen. Innerhalb der Anwendung werden für die Positionen Abwehr, Mittelfeld und Angriff eigene Clusteranalysen durchgeführt.

Für Werte bei denen die Anzahl gespeichert ist, wie die geschossenen Tore oder die Anzahl an gelben Karten, werden die Werte für die jeweilige Runde summiert und durch die gespielten Minuten der betrachteten Runde geteilt. Dadurch erhält man den zugehörigen Wert pro Minute. Zusätzlich wird dieser Wert mit 90 multipliziert. Da ein Spiel regulär 90 Minuten dauert, erreichen man so den entsprechenden Wert pro Spiel. Dies wird aus dem Grund getan, da Weka nur mit vier Nachkommastellen in der Clusteranalyse arbeitet und die Werte hier sehr gering sein können. Durch die Multiplikation wird dieses Problem umgangen.

Die Werte, welche in Prozent gespeichert sind, werden mit der Anzahl an gespielten Minuten pro Spiel multipliziert und anschließend durch die Gesamtanzahl an gespielten Minuten für die jeweilige Runde geteilt. Dadurch erhält man wieder einen Wert in Prozent. Diese Prozentwerte sind zusätzlich anhand der gespielten Minuten eines Spiels gewichtet. Werte von Spielen in denen ein Spieler länger gespielt hat sind höher gewichtet. Abbildung 20 zeigt ein Beispiel in Pseudocode für die Tore und die erfolgreiche Passquote für die Hinrunde der Saison 2010.

Durch die Aggregation erhält man insgesamt 1523 Instanzen. Dabei handelt es sich um die Spieler aus den Positionen Abwehr, Mittelfeld und Angriff über die Hin- und Rückrunden der Saisons 2010 und 2011. Das bedeutet, dass Spieler bis zu viermal in diesen Instanzen vorkommen, sofern sie beide Saisons gespielt haben.

In der Analyse der Datengrundlage zeigt sich, dass der Datensatz Spieler beinhaltet, die sehr wenige Minuten in einer Runde gespielt haben. Zum Beispiel gibt es 43 Instanzen, die in einer Runde insgesamt weniger als zehn Minuten gespielt haben. Es wird angenommen, dass Spieler die weniger als ein Drittel einer Runde gespielt haben, nur schlecht durch die Daten beschrieben werden. Durch die kurze Einsatzzeit repräsentieren die Daten nicht den wirklichen Spielstil des Spielers. Diese Spieler

werden aus dem Datensatz aussortiert. In einer Runde werden insgesamt 17 Spiele gespielt, die in der Regel 90 Minuten dauern. Damit kann ein Spieler maximal 1530 Minuten pro Runde spielen. Nachspielzeiten werden in dieser Arbeit nicht beachtet. Es werden alle Spieler aussortiert, die weniger als 510 Minuten in einer Runde gespielt haben. Dadurch erhält man insgesamt 927 Instanzen mit denen die Clusteranalyse durchgeführt wird. Dabei handelt es sich um 349 Abwehrspieler, 401 Mittelfeldspieler und 177 Angreifer. Das zugehörige SQL Statement ist im Anhang in Abbildung 26 zu sehen.

#### 5.3.2.2 Durchführung der Clusteranalyse

Wie bei der Clusteranalyse der Mannschaften werden auch die Spieler mithilfe der EM-Methode segmentiert. Es werden verschiedene Durchläufe durchgeführt und das beste Ergebnis ausgewählt. Für die einzelnen Positionen Abwehr, Mittelfeld und Angriff werden separate Analysen durchgeführt.

**Abwehr**

In der öffentlichen Meinung gibt es insgesamt zwei verschiedene Abwehrspieler. Dies sind einerseits die Innenverteidiger, welche im Zentrum der Abwehr spielen. Sie sind oft großgewachsen, um Vorteile in den Kopfballduellen ausspielen zu können. Sie sind außerdem für den Spielaufbau zuständig. Bei ihnen sind ein gutes Zweikampfverhalten und ein gutes Passspiel zu erwarten. Die andere Art der Verteidiger sind die Außenverteidiger, welche auf den Außenbahnen der Verteidigung spielen und eine wesentliche Rolle in den offensiven Aktionen der Mannschaft einnehmen. So sind sie für die Flanken von außen auf die eigenen Stürmer zuständig. Im Gegensatz zum Innenverteidiger sind sie eher klein und verfügen über einer hohe Schnelligkeit. Bei ihnen erwartet man eine höhere Anzahl an Vorlagen.

Die Clusteranalyse identifiziert insgesamt fünf unterschiedliche Arten der Abwehrspieler. In Tabelle 28 sind die Mittelwerte der einzelnen Cluster aufgelistet. Die Farben innerhalb der Tabelle, dienen der besseren Übersicht der einzelnen Werte. Die Farbskala reicht dabei von Grün bis Rot. Der höchste Wert einer Zeile nimmt die Farbe Grün an, während der niedrigste Wert der Zeile in Rot eingefärbt wird. Je niedriger ein Wert ist, umso eher färbt sich die entsprechende Zelle Rot. Durch diese Visualisierung lässt sich leichter erkennen, in welchen Attributen die einzelnen Cluster gute bzw. schlechtere Werte annehmen.

Analysiert man die Ergebnisse, so ähneln sich die Spieler aus Cluster 1 und Cluster 4 in gewissen Attributen. In den Zweikampfwerten, der Anzahl an Ballkontakten sowie dem Passspiel haben die Spieler dieser beiden Cluster hohe Mittelwerte. Die Spieler aus Cluster 1 unterscheiden sich jedoch in den beiden offensiven Attributen der erfolgreichen Torschussquote sowie den Toren. Dort haben diese Verteidiger die besten Werte. Die Spieler aus Cluster 4 erreichen dagegen die schlechtesten Werte in diesen Bereichen.

Ähnlich verhalten sich die Spieler aus Cluster 0 und Cluster 2. Sie sind in vielen Attributen gleichartig. Sie haben die schlechtesten Zweikampfwerte und ein schwaches Passspiel. In den Attributen Torschussvorlagen und Vorlagen sind diese Spieler die stärksten. Wobei die Spieler aus Cluster 0 gegenüber denen aus Cluster 2 leicht bessere Werte aufweisen. Die Verteidiger aus Cluster 0 sind zudem torgefährlicher als die Spieler aus Cluster 2.

Die Spieler aus Cluster 3 können lediglich durch gutes Zweikampfverhalten und einer guten erfolgreichen Passquote überzeugen.

**Tabelle 28:** Cluster Abwehrspieler

| Wert | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Zweikampf | 55,835 | 63,285 | 56,193 | 62,527 | 61.733 |
| VerlZweikampf | 44,165 | 36,715 | 43,807 | 37,473 | 38,267 |
| Ballkontakte | 62,303 | 61,023 | 60,151 | 50,673 | 63,428 |
| ErfPaesse | 26,402 | 35,455 | 25,301 | 26,567 | 37,391 |
| FehlPaesse | 8,435 | 6,404 | 8,444 | 5,641 | 6,321 |
| Paesse | 34,837 | 41,858 | 33,744 | 32,208 | 43,711 |
| ErfPassquote | 74,297 | 83,516 | 73,734 | 80,921 | 84,31 |
| FehlPassquote | 25,703 | 16,484 | 26,266 | 19,079 | 15,69 |
| Torschuesse | 0,77 | 0,663 | 0,498 | 0,51 | 0,431 |
| VergTorschuesse | 0,676 | 0,531 | 0,498 | 0,439 | 0,431 |
| ErfTorschussquote | 6,171 | 10,41 | 0 | 5,292 | 0 |
| FehlTorschussquote | 93,829 | 89,59 | 100 | 94,709 | 100 |
| Torschussvorlagen | 1,466 | 0,357 | 0,96 | 0,234 | 0,448 |
| Vorlage | 0,132 | 0,032 | 0,084 | 0,013 | 0,037 |
| Tore | 0,093 | 0,132 | 0 | 0,071 | 0 |
| Gelb | 0,1 | 0,104 | 0,109 | 0,107 | 0,119 |
| PassProBallkontakt | 0,552 | 0,666 | 0,551 | 0,614 | 0,668 |
| ErfPassProBallkontakt | 0,415 | 0,559 | 0,409 | 0,5 | 0,567 |
| TorschussvorlageProPass | 0,045 | 0,009 | 0,03 | 0,007 | 0,01 |

Betrachtet man die Werte, so würde man die Spieler aus Cluster 1 und Cluster 4 in den Bereich Innenverteidung einteilen. Dies kommt daher, dass diese Spieler ein gutes Zweikampfverhalten aufweisen und im Passspiel hohe Werte enthalten, was gut für den Spielaufbau ist. Dagegen haben die Verteidiger aus Cluster 0 und Cluster 2 gute Werte in den Vorlagen, was auf Außenverteidiger schließen lässt. Die Spieler aus Cluster 3 weisen keinen besonderen Spielstil auf.

Stellt man die Spieler der beiden Cluster gegenüber, so sind die Spieler aus Cluster 1 und Cluster 4 im Durchschnitt 188 cm groß. Dagegen sind die Spieler aus Cluster 0 und Cluster 2 im Schnitt 182 cm groß. Sie sind außerdem im Schnitt 5 kg schwerer als die Spieler aus Cluster 0 und 2. Dies lässt wiederum auf die Unterteilung in Innen- bzw. Außenverteidiger schließen. Wie erwähnt sind Innenverteidiger meist großgewachsen aufgrund der Vorteile bei Kopfballduellen. Außenverteidiger sind eher klein und schnell.

Auch die nähere Analyse der Spieler bestätigt diese Einteilung. So sind Spieler, die jeweils in allen vier Runden in das Cluster 0 und 2 eingeteilt werden, die beiden Außenverteidiger von Borussia Dortmund Marcel Schmelzer und Lukasz Piszczek, der Linksverteidiger Felix Bastians und der Schalker Außenverteidiger Christian Fuchs. Jedoch ist nicht jeder Spieler dieser Cluster typischer Außenverteidiger. So fällt unter anderem Khalid Boulahrouz, typischer Innenverteidiger, 3-mal in diese Kategorie.

Cluster 1 und Cluster 4 haben dagegen typische Innenverteidiger in ihren Reihen. So werden unter anderem die Innenverteidiger Serdar Tasci, Rodnei, Holger Badstuber und Mats Hummels in allen vier Runden in diese Gruppen eingeteilt. Ein typischer Außenverteidiger wie Philipp Lahm ist jedoch ebenfalls 4-mal in diesen Clustern zu finden. Mats Hummels, Heiko Westermann und Daniel van Buyten wurden dreimal in die torgefährlichen Innenverteidiger aus Cluster 1 eingeteilt.

Cluster 3 beinhaltet in der Mehrheit Spieler, die nur in einer Runde unter dieses Cluster fallen. Hier sind eher Innenverteidiger zu finden. Benedikt Höwedes wurde für zwei Runden hier eingeteilt, spielte in den anderen zwei Runden jedoch wie ein

**Tabelle 29:** Zuordnung der Cluster der Abwehrspieler

| Cluster | Anzahl Spieler | Typische Spieler |
|---|---|---|
| 0 | 32 | C. Fuchs, C. Pander, L. Piszczek |
| 1 | 51 | M. Hummels, H. Westermann, D. v. Buyten |
| 2 | 68 | F. Bastians, R. Zabavnik, M. Schmelzer |
| 3 | 31 | N. Noveski, K. Haggui, B. Höwedes |
| 4 | 167 | L. Schmitz, H. Badstuber, D. Schwaab |

Spieler aus Cluster 4.

Tabelle 29 zeigt eine Statistik über die zugeordneten Spieler der einzelnen Cluster. Hierbei sind die Anzahl an Spielern in den einzelnen Clustern aufgelistet sowie eine Auswahl an Spielern aufgezählt, die unter dieses Cluster fallen.

**Mittelfeld**

Subjektiv werden Mittelfeldspieler zahlreich unterteilt. So redet man z.B. von sogenannten 6er, 8er oder dem Spielmacher, die Nummer 10. Die Beschreibung dieser ist schwer und detailreich. Eine leichtere Segmentierung der Mittelfeldspieler ist es, sie in defensive und offensive Spieler einzuteilen. Wobei die offensiven Spieler in Attributen, wie Vorlagen, Torschüsse und Tore hohe Werte erzielen. Bei defensiven Spielern ist eine gute Zweikampfquote zu erwarten. Defensive Mittelfeldspieler sind in der Regel für die Kontrolle des Spiels zuständig. So sind viele Ballkontakte und gutes Passspiel zu erwarten. Eine weitere Gruppe der Mittelfeldspieler sind die „Außen". Sie sind wie die Außenverteidiger für viele Flanken zuständig. Hier sind hohe Werte in den Torschussvorlagen bzw. Vorlagen zu erwarten.

Die Durchführung der Clusteranalyse teilt die Spieler des Mittelfelds in fünf Cluster ein. Die Ergebnisse des Clusterings sind in Tabelle 30 zu sehen. Wie auch in der Abwehr gibt es hier zwei Paare von Clustern, die sich ähneln. Ein weiteres Cluster beinhaltet Mittelfeldspieler, die insgesamt eher schlechte Werte aufweisen.

Das Cluster 0 der Analyse, liefert Spieler die sehr durchschnittliche bis schlechte Werte in fast allen Attributen aufweisen. Sie weisen lediglich ein gutes Zweikampfverhalten auf, fallen sonst jedoch nicht positiv auf.

Die Spieler aus Cluster 1 und Cluster 4 sind Spieler, die sich in ihrer Spielweise ähnlich sind. Dabei können die Spieler aus Cluster 1 als die besseren Mittelfeldspieler der beiden Gruppen angesehen werden. Die Spieler aus diesem Cluster haben in den offensiven Attributen, wie Torschüsse, erfolgreiche Torschussquote, Torschussvorlagen, Vorlagen und Tore hohe Mittelwerte. In den Attributen Zweikampf, Ballkontakte und erfolgreiche Passquote sind die Werte vergleichsweise schlecht. Im Gegensatz dazu haben die Spieler aus Cluster 4 in diesen Attributen die schlechtesten Werte aller Mittelfeldspieler. In den offensiven Attributen weisen die Spieler aus Cluster 4 gute Werte auf. Die Spieler dieser beiden Cluster erhalten eher wenige gelbe Karten pro Spiel.

Im Gegensatz dazu haben die Spieler aus Cluster 2 und Cluster 3 gute Werte in den Attributen Zweikampf und Ballkontakte sowie gute Werte im Passspiel. Die Spieler aus Cluster 2 besitzen in den offensiven Attributen die schlechtesten Mittelwerte. Die Spieler aus Cluster 3 erreichen hier gering bessere Mittelwerte. Insgesamt beschreiben die Spieler aus Cluster 3 die besseren Mittelfeldspieler der beiden Cluster. Die Mittelwerte der Spieler aus Cluster 3 sind in der Mehrheit besser als die des Clusters 2. Beide Cluster beschreiben jedoch einen ähnlichen Spielstil.

Die Verteilung der Spieler auf die Cluster stellt Tabelle 31 dar. Im Vergleich beschrei-

**Tabelle 30:** Cluster Mittelfeldspieler

| Wert | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Zweikampf | 50,778 | 43,25 | 50,917 | 52,228 | 43,449 |
| VerlZweikampf | 49,222 | 56,75 | 49,083 | 47,772 | 56,551 |
| Ballkontakte | 54,08 | 55,237 | 59,907 | 68,901 | 49,854 |
| ErfPaesse | 25,768 | 26,609 | 33,327 | 41,502 | 20,876 |
| FehlPaesse | 8,738 | 7,592 | 7,71 | 8,217 | 6,819 |
| Paesse | 34,506 | 34,202 | 41.037 | 49,719 | 27,694 |
| ErfPassquote | 74,117 | 77,088 | 79,906 | 82,698 | 74,107 |
| FehlPassquote | 25,883 | 22,912 | 20,094 | 17,302 | 25,893 |
| Torschuesse | 1,308 | 2,225 | 0,925 | 1,419 | 1,914 |
| VergTorschuesse | 1,191 | 1,89 | 0,925 | 1,249 | 1,81 |
| ErfTorschussquote | 5,643 | 14,191 | 0 | 8,743 | 3,257 |
| FehlTorschussquote | 94,358 | 85,809 | 100 | 91,257 | 96,743 |
| Torschussvorlagen | 1,274 | 2,302 | 1,191 | 1,476 | 2,441 |
| Vorlage | 0,082 | 0,264 | 0,099 | 0,136 | 0,245 |
| Tore | 0,117 | 0,335 | 0 | 0,17 | 0,104 |
| Gelb | 0,09 | 0,096 | 0,112 | 0,117 | 0,092 |
| PassProBallkontakt | 0,633 | 0,612 | 0,674 | 0,714 | 0,546 |
| ErfPassProBallkontakt | 0,471 | 0,474 | 0,542 | 0,592 | 0,408 |
| TorschussvorlageProPass | 0,039 | 0,071 | 0,032 | 0,031 | 0,094 |

**Tabelle 31:** Zuordnung der Cluster der Mittelfeldspieler

| Cluster | Anzahl Spieler | Typische Spieler |
|---|---|---|
| 0 | 55 | M. Caligiuri, B. Vukcevic, S. Pinto |
| 1 | 69 | T. Müller, M. Reus, F. Ribery |
| 2 | 137 | M. Schmiedebach, P. Bargfrede, Josue |
| 3 | 85 | L. Bender, S. Rolfes, B. Schweinsteiger |
| 4 | 55 | M. Marin, C. Clemens, C. Tiffert |

ben Cluster 1 und Cluster 4 die offensiven Mittelfeldspieler. Diese bestehen bei genauerer Analyse aus den Außenspielern, wie Franck Ribery, Marco Reus oder Thomas Müller aber auch den Spielmachern aus dem Zentrum wie Mario Götze oder Shinji Kagawa. Das Cluster 1 die besseren Offensivspieler beschreibt, zeigt sich auch daran, dass sich in diesem Cluster etwa 45 % an Nationalspielern befinden. In Cluster 4 sind dagegen nur 11 % an Nationalspielern vorhanden.

In Cluster 2 und Cluster 3 sind wie erwartet überwiegend defensive Spieler zu finden. So befinden sich Bastian Schweinsteiger, Simon Rolfes oder Lars Bender in jeder Runde in einem dieser Cluster. Cluster 3 beinhaltet ca. 33 % Nationalspieler, Cluster 2 etwa 23 %. Dies deckt sich mit der Clusterbeschreibung. Die Spieler aus Cluster 3 haben überwiegend bessere Mittelwerte als die Spieler aus Cluster 2.

### Angriff

Im Angriff spricht man meist von den klassischen Mittelstürmern, welche oft das Tor treffen. Außerdem gibt es die Außenstürmern, welche auf den Außenbahnen für Flanken oder Vorlagen zuständig sind.

Tabelle 32 zeigt das eingefärbte Ergebnis der Clusteranalyse für den Angriff. Wie zu sehen ist, teilt die EM-Methode die Angreifer in sechs verschiedene Cluster ein. Die Verteilung der Spieler auf die Cluster ist in Tabelle 33 zu sehen.

Cluster 0 beinhaltet Spieler, die viele Torschüsse abgeben, die eine sehr gute erfolgreiche Torschussquote haben und die mit Abstand die meisten Tore schießen. Sie können als die klassischen „Goalgetter" kategorisiert werden. Diese Gruppe hat in den Bereichen Zweikampf, Ballkontakte und Passspiel durchschnittliche Werte. Diese Spieler erhalten auch die meisten gelben Karten.

Die Spieler aus Cluster 1 heben sich durch die meisten Vorlagen ab. Die erfolgreiche Torschussquote ist hoch, es werden insgesamt jedoch wenige Torschüsse abgegeben. Dieses Cluster beinhaltet die wenigsten Spieler mit zehn Angreifern.

Die Angreifer aus Cluster 2 sowie Cluster 4 sind sich relativ ähnlich. Hier finden sich Spieler wieder, welche viele Torschussvorlagen und Vorlagen abgeben. Zudem haben die Spieler aus diesen beiden Clustern die besten Zweikampfwerte und meisten Ballkontakte. Auch in der Anzahl an Pässen werden die höchsten Mittelwerte angenommen. In der erfolgreichen Passquote haben die Spieler aus Cluster 4 den höchsten Wert. Die Angreifer aus Cluster 2 dagegen den schlechtesten. Die Spieler aus Cluster 4 haben ebenfalls die meisten Pässe pro Ballkontakt.

Die Stürmer aus Cluster 3 und Cluster 5 fallen auf, weil sie viele Torschüsse abgeben. Die Spieler aus Cluster 5 weisen dagegen eine schlechte erfolgreiche Torschussquote aus. Die Spieler aus Cluster 3 haben eine bessere Torschussquote und treffen das Tor am zweithäufigsten aller Angreifer. In den sonstigen Attributen sind keine großen Auffälligkeiten zu erkennen, wobei die Angreifer aus Cluster 5 noch eine gute erfolgreiche Passquote erspielen.

Cluster 0 beinhaltet die Spieler, welche außergewöhnlich häufig treffen. Dies sind z.B. Mario Gomez, Papiss Cisse, Stefan Kießling oder Vedad Ibisevic. Diese Spieler spielen in ihren Mannschaften auch hauptsächlich die Rolle des Mittelstürmers. Bis auf wenige Ausnahmen spielen die Spieler nicht ausschließlich in jeder Runde in diesem Cluster. Sofern die Spieler nicht unter das Cluster 0 fallen, so teilen sie sich in die Cluster 3 bzw. Cluster 5 auf. Man kann sagen, dass das Cluster 0 die Ausreißer der Mittelstürmer beinhaltet, die in der betrachtenden Runde außergewöhnlich oft treffen. Papiss Cisse ist ein Spieler, der in all seinen Runden in dieses Cluster eingeteilt wird.

Die Angreifer Mario Mandzukic, Mohamadou Idrissou oder Martin Harnik sind in allen Runden in Cluster 2 eingeteilt. Dies sind auch eher Spieler, die im Zentrum spielen. Jedoch sind diese Spieler eher mitspielende Stürmer mit guten Zweikampfwerten.

Cluster 4 beinhaltet in der Mehrheit Spieler, die als sogenannte „Außen" bezeichnet werden. Darunter fallen z.B. Jefferson Farfan, Jan Schlaudraff oder Ryan Babel. Zudem finden sich hier Mittelstürmer wie Raul oder Claudio Pizarro wieder. Diese beiden Spieler gelten als technisch versierte Spieler. Das gute Passspiel sowie die zahlreichen Ballkontakte beweisen diese Einschätzung. Diese Spieler legen eher Tore vor, als sie selbst zu erzielen.

Cluster 1 beinhaltet nur zehn Spieler. Mohammed Abdellaoue wird in allen vier Runden in dieses Cluster eingeteilt. Keiner der zehn Spieler ist ein klassischer Außenspieler, sondern eher zentral spielende Angreifer.

**Zusammenfassung der Clusteranalyse der Spieler**

Das Clustering der Spieler teilt diese in mehrere Kategorien auf, welche sich im Grunde mit der öffentlichen Wahrnehmung der Spielstile decken. Schaut man sich den Wechsel von Nuri Sahin von Borussia Dortmund zu Real Madrid nach der Saison 2010 an, so wurde dieser mit Ilkay Gündögan vom 1. FC Nürnberg ersetzt. Die Clusteranalyse hätte Ilkay Gündogan nicht als perfekten Ersatz vorgeschlagen. So ist Nuri Sahin in Cluster 3 eingeteilt und wird damit als ein eher defensiver Mittelfeldspieler mit

**Tabelle 32:** Cluster Angreifer

| Wert | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Zweikampf | 38,85 | 31,36 | 43,9 | 38,68 | 42,75 | 39,28 |
| VerlZweikampf | 61,15 | 68,64 | 56,1 | 61,32 | 57,25 | 60,72 |
| Ballkontakte | 39,21 | 24,08 | 44,26 | 35,33 | 52,43 | 42,13 |
| ErfPaesse | 16,0 | 8,3 | 17,3 | 11,48 | 27,85 | 18,49 |
| FehlPaesse | 5,5 | 2,92 | 8,05 | 5,25 | 6,77 | 5,15 |
| Paesse | 21,51 | 11,22 | 25,35 | 16,73 | 34,62 | 23,64 |
| ErfPassquote | 74,14 | 73,88 | 68,05 | 68,8 | 79,8 | 78,12 |
| FehlPassquote | 25,86 | 26,12 | 31,95 | 31,2 | 20,2 | 21,88 |
| Torschuesse | 2,88 | 2,0 | 2,39 | 2,72 | 2,05 | 2,8 |
| VergTorschuesse | 2,2 | 1,63 | 2,12 | 2,33 | 1,77 | 2,54 |
| ErfTorschussquote | 24,39 | 16,25 | 8,95 | 12,46 | 9,9 | 7,92 |
| FehlTorschussquote | 75,61 | 83,75 | 91,05 | 87,54 | 90,1 | 92,08 |
| Torschussvorlagen | 1,61 | 1,03 | 1,79 | 1,46 | 1,74 | 1,51 |
| Vorlage | 0,17 | 0,18 | 0,18 | 0,11 | 0,18 | 0,13 |
| Tore | 0,69 | 0,36 | 0,27 | 0,39 | 0,28 | 0,26 |
| Gelb | 0,12 | 0,05 | 0,07 | 0,06 | 0,07 | 0,1 |
| PassProBallkontakt | 0,54 | 0,46 | 0,57 | 0,47 | 0,66 | 0,56 |
| ErfPassProBallkontakt | 0,4 | 0,34 | 0,39 | 0,32 | 0,52 | 0,44 |
| TorschussvorlageProPass | 0,08 | 0,09 | 0,08 | 0,09 | 0,05 | 0,07 |

**Tabelle 33:** Zuordnung der Cluster der Angreifer

| Cluster | Anzahl Spieler | Typische Spieler |
|---|---|---|
| 0 | 37 | P. Cisse, M. Gomez, S. Kießling |
| 1 | 10 | M. Abdellaoue, E. Hoffer, C. Marica |
| 2 | 26 | M. Mandzukic, Cacau, I. De Camargo |
| 3 | 38 | C. Eigler, S. Allagui, M. Petric |
| 4 | 29 | Raul, J. Farfan, J. Schlaudraff |
| 5 | 37 | M. Arnautovic, S. Okazaki, L. Podolski |

durchschnittlicher Offensivkraft charakterisiert. Ilkay Gündogan ist 2010 als ein Spieler aus Cluster 1 beschrieben. Dieses Cluster beinhaltet Spieler, welche ihre Stärken in der Offensive haben. Somit läge laut der Analyse kein ähnlicher Spielstil vor. Dass Ilkay Gündögan trotzdem als Nuri Sahin Ersatz verpflichtet wurde, zeigt die Clusterzuordnung der Saison 2011. In beiden Runden wird Ilkay Gündögan in Cluster 3 eingeteilt und ersetzte somit Nuri Sahin, der ebenfalls dem Spielstil von Cluster 3 entspricht.

Schaut man sich dagegen den Wechsel von Marco Reus nach der Saison 2011 von Borussia Mönchengladbach zu Borussia Dortmund an, so wurde kein direkter Ersatz eingekauft. Statt einen neuen offensiven Mittelfeldspieler zu kaufen, wurden mit Luuk de Jong oder Peniel Mlapa Angreifer eingekauft, um so die Offensive zu stärken. Das Internetportal von T-Online vermutete bereits im Januar 2012, dass auch kein neuer Spieler als Ersatz für Marco Reus nötig ist, sondern das Eigengewächs Patrick Herrmann die Lücke füllen kann [63]. Eine Vermutung die sich mit der Clusteranalyse deckt. Marco Reus spielt in allen Runden laut der Beschreibung des Clusters 1. Patrick Hermann ist ebenfalls in allen Runden in dieses Cluster eingeteilt. Somit war es für Borussia Mönchengladbach nicht nötig den Spieler direkt zu ersetzen, da bereits ein ähnlicher Spieler bei dem Verein spielte.

### 5.3.3 Implikationen für den realen Einsatz

Der exemplarische Einsatz der Clusteranalyse auf die vorhandenen Daten der zwei Saisons zeigt, dass das Clustering sinnvolle Möglichkeiten zur Anwendung im Profifußball bietet. Die identifizierten Cluster für die einzelnen Teile der Mannschaften aber auch für die gesamten Mannschaften zeigen, dass sich Top-Teams in bestimmten Bereichen ähneln und sich von schlechteren Teams abgrenzen. Das Wissen über den Spielstil erfolgreicher Mannschaften, kann zur Kaderplanung genutzt werden oder das Training bzw. Spielsystem einer Mannschaft beeinflussen. In den Beispielen für die Mannschaften sind zwar viele der gewonnenen Erkenntnisse offensichtlich, jedoch liefert die Clusteranalyse auch Resultate, die nicht direkt offenkundig sind. So sind die offensiven Qualitäten in der Abwehr nicht entscheidend für den Erfolg einer Mannschaft, sondern können eher vernachlässigt werden. Die Mittelfeldreihen sollten überwiegend ein sicheres Passspiel aufweisen und viele Ballkontakte haben. Auch hier ist die Offensive nicht unbedingt entscheidend. Mannschaften die erfolgreich spielen wollen brauchen zudem keine spielstarken Stürmer, sondern welche die viele Torschüsse abgeben und viele Tore erzielen.

Weitere Erkenntnisse könnten eine größere Menge an spielspezifischen Attributen liefern. Hier könnten Antworten darauf zu finden sein, inwiefern die Fitness der Spieler eine entscheidende Rolle spielt. Sind Mannschaften die mehr Kilometer in einem Spiel zurücklegen oder mehr Sprints laufen erfolgreicher? Genauso kann untersucht werden inwiefern Angriffe über die linke bzw. rechte Seite den Erfolg beeinflussen. Durch mehrere Attribute könnte der Spielstil der einzelnen Mannschaften noch detaillierter beschrieben werden. Führen z.B. Flankenläufe zu mehr Punkten oder sollte über das Zentrum gespielt werden? Die Erkenntnisse könnten genutzt werden um den eigenen Spielstil in bestimmter Weise anzupassen.

Auch die Beschreibung des Spielstils der Gegner kann einem Verein Vorteile verschaffen. So wählen gewisse Trainer, wie Thomas Tuchel vom 1. FSV Mainz 05 oder Jürgen Klopp von Borussia Dortmund, je nach Gegner ihre Spieler und Taktik aus [62]. Die Beschreibung anhand der Cluster können zur Spielvorbereitung genutzt werden. Statt Runden können hier auch einzelne Spiele geclustert werden. Sucht man z.B. die perfekte Taktik gegen eine bestimmte Mannschaft, kann man die letzten Spiele des Gegners clustern um zu sehen, wie die gegnerischen Mannschaften gegen diesen Ver-

ein aufgetreten sind. Man kann so versuchen zu identifizieren gegen welche Spielstile der Gegner Schwächen offenbart bzw. welche Spielstile dem Gegner liegen. Führen z.B. Angriffe über eine gewisse Seite zu mehr Torchancen, als Angriffe über die entgegengesetzte Seite oder über das Zentrum?

Innerhalb der Clusteranalyse der Spieler liefern die Beschreibungen der Cluster vor allem Nutzen für das Scouting von Vereinen. Die Einteilung in Gruppen hilft Spieler zu finden, welche im Spielstil ähnlich zu anderen Spielern sind. So kann gezielt nach adäquatem Ersatz von Spielern gesucht werden. Auch eine heterogene Mischung einer Mannschaft kann hier erzielt werden. So kann der Kader aus Spielern aus verschiedenen Clustern gebildet werden. Die Kategorisierung kann zudem helfen, Spieler aus ausländischen Ligen vorzufiltern. Da die Spielweise dieser Spieler nicht so bekannt ist, wie es bei Spielern der heimischen Liga ist, kann bereits eine Einteilung in Gruppen vor dem eigentlichen Scouting stattfinden. So müssen nur die Spieler näher beobachtet werden, welche in die gewünschte Gruppe eingeteilt wurden. Auch hier würden mehrere Attribute detailreichere Analysen erlauben. Auch physische Attribute, wie die Größe, das Alter oder das Gewicht der Spieler könnten mit einbezogen werden

Insgesamt eignet sich die Clusteranalyse sehr gut um eine Segmentierung der Mannschaften bzw. Spieler durchzuführen. Die spielspezifischen Daten sind für eine solche Analyse gut geeignet.

## 5.4 Vorhersage der nächsten Saison

In den folgenden Abschnitten wird versucht die nächste Saison vorherzusagen. Dabei wird in einem ersten Schritt die Regressionsanalyse genutzt, um die Punkte der einzelnen Vereine einer zukünftigen Saison zu prognostizieren. In einem weiteren Schritt wird mithilfe der Klassifikation versucht den Rang der Vereine vorherzusagen. Um die Realisierbarkeit solcher Vorhersagen zu testen wird die Saison 2011 prognostiziert und die Resultate mit den tatsächlichen Werten der Saison überprüft.

Die folgenden Prognosen innerhalb der Regressionsanalysen sind zweigeteilt. Einerseits wird eine Prognose erstellt, um den Punkteverlust zu bestimmen, welcher ausschließlich durch Spielerverkäufe eintritt. Dabei werden die bestehenden Kader der Teams genommen und die Spieler aus den Teams gelöscht, welche den Verein am Ende der Saison 2010 verlassen haben. Andererseits werden die Punkte der neu zusammengestellten Teams vorhergesagt, also die Mannschaften unter Berücksichtigung aller Zu- und Abgänge. So ist auch erkennbar, welche Auswirkungen die Zukäufe von Spielern auf die Leistung einer Mannschaft haben.

Bei der Klassifikation wird geprüft, inwiefern die Methoden dieser Anwendung den Rang der Hin- und Rückrunde der Saison 2011 prognostizieren können. Hierbei werden ausschließlich die Teams inklusive aller Zu- und Abgänge betrachtet. Auch hier wird geprüft wie sich die tatsächlichen Ränge der Saison 2011 mit denen der Prognose unterscheiden.

Eine Prognose der nächsten Saison kann für Vereine eine große Hilfe darstellen. Durch eine akkurate Vorhersage kann geprüft werden, inwiefern sich Verkäufe von Spielern auf die Saisonleistung auswirken und ob sie durch die Spielereinkäufe aufgefangen werden können bzw. ob eine Leistungssteigerung zu erwarten ist. Ein Vorhersagemodell kann so als Frühwarnsystem gelten, sofern die Transfers ein Ungleichgewicht aufweisen. Es kann anhand des vorhergesagten Punkte- bzw. Rangverlustes geprüft werden, ob die abgegebenen Leistungsträger durch geplante Spielerzukäufe adäquat ersetzt werden.

Ob ein Spiel gewonnen wird hängt von vielen Faktoren ab. Somit ist eine Vorhersage schwer zu treffen. Jedes Jahr gibt es etliche Experten die aufgrund der Entwicklungen auf dem Transfermarkt versuchen den Meister oder die Absteiger der nächsten

Saison festzulegen. So hat 1899 Hoffenheim vor der Saison 2012 mehrere neue Spieler verpflichtet. Als Saisonziel wurde ein Platz unter den besten sechs Mannschaften ausgesprochen [26]. Das Ergebnis nach mehreren Trainerwechseln war, dass 1899 Hoffenheim am Ende der Saison 2012 auf Platz 16 landete. Das Saisonziel wurde also um zehn Plätze verfehlt.

Jede Saison ist von solchen Überraschungen gespickt, die nur wenige Experten vorhersagen. Ob eine Prognose mithilfe von Data Mining möglich ist, soll in den folgenden Abschnitten geklärt werden. Dabei wird zunächst auf die Datengrundlage eingegangen. Anschließend werden die Regressionsanalyse und die Klassifikation auf die Daten angewendet. Eine Einschätzung der Ergebnisse wird im letzten Abschnitt vorgenommen.

### 5.4.1 Datengrundlage

Für die folgenden Analysen werden insgesamt drei Datensätze aufgebaut. Wie auch bei der Clusteranalyse der Mannschaften, wird in diesen Datensätzen die Gesamtleistung der Vereine betrachtet. Es werden alle Spieler eines Vereins zu einem Datensatz aggregiert. So wird eine spielspezifische Beschreibung jeder Mannschaft erreicht.

Bei dem ersten Datensatz handelt es sich um den Trainingsdatensatz. Hier werden die Vereine jeweils für die Hin- und Rückrunde der beiden Saisons 2010 und 2011 aggregiert. Dadurch erhält man einen Trainingsdatensatz von 72 Mannschaften. Der Trainingsdatensatz wird dazu genutzt ein Data Mining Modell aufzubauen, auf dem neue Daten angewendet werden.

Bei den neuen Datensätzen wird ein Datensatz die Mannschaften der Saison 2010 enthalten, jedoch ohne die Spieler, welche den Verein zur Winterpause bzw. Saisonende verlassen haben. Die Idee ist es zu zeigen, wie sich die Wechsel der Spieler auf den Rang bzw. die erspielten Punkte auswirken.

Ein zweiter Datensatz, der auf die erlernten Modelle angewendet wird, berücksichtigt die Mannschaften mit allen Zu- und Abgängen. Dieser Datensatz betrachtet ebenfalls nur die Saison 2010. Spieler, die zu dem betrachteten Verein hinzugekommen sind und in der Saison 2010 bei einem anderen Verein innerhalb der Liga gespielt haben, können dem Datensatz mit den Werten aus 2010 hinzugefügt werden. Spieler die aus einer anderen Liga zu dem Verein gewechselt sind und für die keine Werte in der Saison 2010 vorliegen, werden mit den Daten aus 2011 zum Verein hinzugefügt. Normalerweise sind diese Daten aus der Zukunft nicht vorhanden, um jedoch den Einsatz der Prognosen zu testen, werden diese in dieser Arbeit trotzdem hinzugefügt. Ziel ist es somit die Saison 2011 mit allen Zu- und Abgängen eines jeden Vereins zu prognostizieren. Anhand der Ergebnisse kann eingeschätzt werden, inwiefern sich eine Saison vorhersagen lässt.

Der Trainingsdatensatz wird anhand der Werte aus der Tabelle **Fakten** aufgebaut. Hier sind die Daten aus jedem Spiel gespeichert. Eine genaue Beschreibung der Daten findet sich in Kapitel 4. Sobald ein Spieler an einem Spiel teilgenommen hat, werden Daten, wie die Anzahl an Torschüssen oder Pässen aufgenommen und gespeichert.

Die Attribute, welche für jeden Spieler in jedem Spiel in der Tabelle **Fakten** zur Verfügung stehen, werden für die einzelnen Mannschaft aggregiert. Gleichzeitig werden die Attribute anhand der gespielten Minuten gewichtet. So wird erreicht, dass Spieler mit mehr Spielminuten die Daten der Mannschaft höher beeinflussen, als Spieler die nur wenig Einsatzzeit haben. Die Idee dahinter ist, dass Spieler mit mehr Spielminuten den Erfolg einer Mannschaft maßgeblich beeinflussen und Spieler mit wenig Einsatzzeit nur geringen Einfluss auf die Leistung haben. Dies wird erzielt indem jeder Wert eines Spielers mit der Einsatzzeit des Spielers in dem Spiel multipliziert wird. Dieses Produkt wird dann für die jeweiligen Runden von allen Spielern eines

**Abbildung 21:** Pseudocode Aggregation von Vereinen mit Gewicht

```
SELECT
Saison, 'Hinrunde' AS Runde, Verein,
Summe(Zweikampf*gespielt)/Summe(gespielt),
Summe(Tore*gespielt)/Summe(gespielt)
FROM fakten WHERE Saison = 2010 AND Spieltag > 0 AND Spieltag
   < 18
GROUP BY Saison, Verein
```

Vereins summiert und durch die gespielten Minuten der gesamten Mannschaft in dieser Runde geteilt. Abbildung 21 zeigt ein Beispiel für die gewonnenen Zweikämpfe und Anzahl an Toren für die Hinrunde der Saison 2010. Das gesamte SQL-Statement ist in Abbildung 27 im Anhang zu finden.

Der Datensatz ohne die Spieler, welche den Verein verlassen haben, wird auf dieselbe Weise aggregiert. Für die Aggregation wird eine Kopie der Tabelle **Fakten** erstellt. Dabei wird die Hin- sowie Rückrunde der Saison 2010 mit der Hin- und Rückrunde der Saison 2011 verglichen. Sobald ein Spieler eines Vereins in der folgenden Saison nicht mehr in diesem Verein gespielt hat, wird dieser aus der Faktentabelle gelöscht. Auf diese Tabelle der Fakten ohne die Spieler wird dann das SQL Statement aus Abbildung 27 angewendet.

Für den Datensatz mit den neu zusammengestellten Kadern wird ebenfalls die Tabelle **Fakten** kopiert. Hierbei werden die Spieler aus der Saison 2010 mit den Spielern der Saison 2011 verbunden. Sofern ein Spieler in der Saison 2011 nicht mehr in der Bundesliga gespielt hat, wird er bei diesem Verbund (Join) nicht weiter berücksichtigt. Durch den Join erhält man alle Spieler der Saison 2011, welche auch in der Saison 2010 gespielt haben. Für diese Spieler werden die Daten aus der Saison 2010 genommen. Der Verein des Spielers wird jedoch auf den neuen Verein der Saison 2011 gesetzt. Dadurch werden alle Wechsel innerhalb der Liga berücksichtigt. Zusätzlich müssen die Spieler, für die keine Daten in der Saison 2010 vorhanden sind berücksichtigt werden. Um diese Spieler der Faktentabelle hinzuzufügen, werden sie und ihre Daten aus der Saison 2011 ausgewählt und mit der Kopie der Tabelle **Fakten** vereinigt. Abbildung 22 zeigt, wie die Kopie der Tabelle **Fakten** aufgebaut wird. Die Daten werden wie in dem SQL-Statement aus Abbildung 27 aggregiert, jedoch auf die Kopie der Faktentabelle angewendet.

Es werden somit drei Datensätze kreiert, welche die gleiche Struktur aufweisen. Einen, der die Saisons 2010 und 2011 mit den tatsächlichen aufgetretenen Werten beschreibt und als Trainingsdatensatz dient. Einen Weiteren, der die Saison 2010 beschreibt, jedoch ohne die Spieler, die den Verein verlassen haben. Der dritte Datensatz berücksichtigt alle Zu- und Abgänge der Vereine. Hier werden die Daten aus der Saison 2010 benutzt, sofern sie vorhanden sind. In dem Fall, in dem ein Spieler aus einer anderen Liga gewechselt ist, werden die Daten der Saison 2011 benutzt.

Bei den neuen Datensätzen werden außerdem Spieler als Abgänge betrachtet, sofern die Spieler in der Saison 2011 verletzt waren und kein Spiel bestritten haben. Als Neuzugänge werden Spieler behandelt, welche in der Saison 2010 verletzt bzw. nicht berücksichtigt wurden, aber in der Saison 2011 für den Verein angetreten sind.

In der Tabelle **Abschlusstabelle** sind die Zielvariablen gespeichert, welche für die Regressionsanalyse und Klassifikation nötig sind. Innerhalb der Regressionsanalyse wird mit den Punkten gearbeitet. Die Tabelle **Abschlusstabelle** enthält die Punkte für die Hin- und Rückrunden der Saison 2010 sowie der Saison 2011.

Innerhalb der Durchführung der Klassifikation wird mit dem Rang der Vereine ge-

**Abbildung 22:** Kopie der Tabelle Fakten mit Transfers

```
SELECT
f.Saison, f.Spieltag, f.Datum, f.Heim, f.Gast, s.Verein, f.
    Spieler, f.Zweikampf, f.verlZweikampf, f.Ballkontakte, f.
    Paesse,
..., f.gespielt
FROM
(SELECT DISTINCT verein, spieler FROM fakten WHERE saison =
    2011) s
JOIN
(select * from fakten where saison = 2010) f
ON f.spieler = s.spieler
UNION
SELECT * FROM fakten WHERE saison = 2011 AND spieler NOT IN (
    SELECT spieler FROM fakten WHERE saison = 2010)
```

arbeitet. Die ersten fünf Plätze beschreiben dabei den Rang {2}, die Plätze 6 bis 15 den Rang {1} und die letzten drei Plätze den Rang {0}. Sofern Mannschaften punktgleich auf Platz 5 bzw. Platz 16 stehen, nehmen sie den Rang {2} bzw. {0} an.

In beiden Fällen werden die Runden separat betrachtet. Die Punkte der Hinrunde haben keinen Einfluss auf die Punkte der Rückrunde.

### 5.4.2 Durchführung der Prognose

Im Folgenden wird versucht mithilfe der vorgestellten Datensätze, die Auswirkungen von Spielerein- bzw. verkäufen auf die Leistung einer Mannschaft zu quantifizieren. Dabei wird in einem ersten Schritt die Regressionsanalyse genutzt. Mithilfe der Regressionsanalyse soll herausgefunden werden, inwiefern sich die Transfers auf den Punktestand von Vereinen nach einer Runde auswirken.

In Abschnitt 5.4.2.2 wird mithilfe der Klassifikation versucht den Rang der neu zusammengestellten Mannschaften zu ermitteln. So wird versucht zu erklären, ob die Ein- und Verkäufe von Spielern den Rang erhöhen bzw. verringern.

#### 5.4.2.1 Punkte

Wie erwähnt wird mithilfe der Regressionsanalyse und des Trainingsdatensatzes ein Regressionsmodell erlernt, auf dem die neu zusammengestellten Daten angewendet werden. Man nimmt dazu die 72 Instanzen aus der Aggregation der Mannschaften. Als Zielvariable gelten die entsprechenden Punkte der einzelnen Mannschaften. Um die Qualität der Regressionsanalyse zu testen, wird die Cross-validation Testmethode mit zehn Teilmengen genutzt. Die Qualität wird mithilfe des RMSEs gemessen. Dieser wurde bereits in Kapitel 5.2 beschrieben. Der RMSE misst die durchschnittlichen Abweichung der prognostizierten Punkten von den tatsächlich erzielten Punkten.

Die besten Ergebnisse bei der Durchführung mehrerer Regressionsmethoden unter verschiedenen Einstellungen bietet die lineare Regression und die SMOreg Methode. Bei der linearen Regression wurde dabei die Option „kollineare Attribute entfernen" ausgewählt. Die SMOreg Methode kann ohne die Attribute, welche keine neuen Informationen darstellen, wie verlorene Zweikämpfe und fehlerhafte Passquote einen besseren RMSE erreichen. Das Eliminieren des Attributs fehlerhafte Passquote verschlechterte den RMSE wiederum und wird somit beibehalten.

**Tabelle 34:** RSME der Regressionsanalyse für die Prognose

| Methode | RMSE |
|---|---|
| Lineare Regression | 5,1997 |
| SMOreg | 5,1515 |

Die Ergebnisse sind in Tabelle 34 dargestellt. Wie man sieht, liegt die Differenz der vorhergesagten Punkte im Schnitt etwa fünf Punkte neben den tatsächlichen Punkten.

Im nächsten Schritt wird der Datensatz der Mannschaften ohne die abgegebenen Spieler auf die erlernten Modelle angewendet. In Tabelle 35 sind die vorhergesagten Punkte ohne diese Spieler aufgelistet. Zusätzlich ist die Differenz zum vorhergesagten Wert aus dem Training des Modells abgebildet. Dadurch erhält man eine Darstellung wie sich die Vereine ohne die Spieler in der Saison laut den Regressionen entwickelt hätten. Nutzt man beispielsweise den Trainingsdatensatz wiederum als Testdatensatz, hat Borussia Dortmund in der Hinrunde im Training bei der linearen Regression bereits einen Fehler von -9,7 Punkten zu den tatsächlich erspielten Punkten. Ob sich eine Mannschaft laut Methode verschlechtert, wird anhand des Unterschieds zum Trainingswert bestimmt.

Wie man sieht, sind die Tendenzen der beiden Regressionen in 25 der 32 Fällen gleich. Obwohl man davon ausgehen würde, dass sich jede Mannschaft durch die Abgabe von Spielern verschlechtert, gibt es Mannschaften die sich laut dem Modell verbessern würden. Dies ist beispielsweise beim FC Bayern München der Fall. Die Verbesserung kommt dadurch zustande, dass die Spieler die mehr zum Erfolg beigetragen haben nun stärker gewichtet werden als vorher. Der FC Bayern München hätte ohne die Spieler, die nach der Saison den Verein verlassen haben laut den Methoden zwischen 3 und 7 Punkten mehr in der Hinrunde der Saison 2011 erreicht.

Laut der linearen Regression hat der 1. FSV Mainz 05 in der Hinrunde am meisten an Qualität verloren. Auch laut der SMOreg Methode hätte der 1. FSV Mainz 05 an Punkten verloren. Bei beiden Regressionen liefert die Prognose einen Unterschied zwischen 5 und 10 Punkten für die Hinrunde. In der Rückrunde hätte sich der 1. FSV Mainz 05 laut linearer Regression leicht verbessert und laut SMOreg Methode um etwa 2 Punkte verschlechtert. Analysiert man den 1. FSV Mainz 05 so sieht man, dass viele der Leistungsträger den Verein verlassen haben. So wechselten zum Abschluss der Saison 2010 die heutigen Nationalspieler Andre Schürrle (Bayer 04 Leverkusen), Christian Fuchs und Lewis Holtby (beide FC Schalke 04). Außerdem zog sich Adam Szalai einen Kreuzbandriss zu und fiel ebenfalls für die Hinrunde der Saison 2011 aus. Diese Spieler haben in der Hinrunde jeweils über 1000 Minuten für den 1. FSV Mainz 05 gespielt.

Betrachtet man die prognostizierte Punktzahl der gesamten Saison, hätten laut der linearen Regression Borussia Dortmund (-12,68 Punkte), der SV Werder Bremen (-10,12) und Bayer 04 Leverkusen (-8,73) neben dem 1. FSV Mainz 05 (-9,56) am meisten an Qualität durch Spielerverkäufe verloren. Danach folgen der VfL Wolfsburg, der Hamburger SV und der FC Schalke 04 die zwischen -6 und -3 Punkte verloren hätten. Für den SC Freiburg, den 1. FC Nürnberg und Hannover 96 prognostiziert die lineare Regression kaum Unterschiede in der Punktzahl. Bei den Vereinen aus Bayern, Mönchengladbach, Stuttgart, Hoffenheim, Köln und Kaiserslautern ist laut linearer Regression über die gesamte Saison sogar eine Qualitätssteigerung durch die Abgänge zu erwarten. Besonders beim 1. FC Kaiserslautern ist diese Vorhersage überraschend, da der 1. FC Kaiserslautern u.a. seinen besten Stürmer Srdan Lakic abgegeben hat. Nach dem siebten Platz aus der Saison 2010 ist der 1. FC Kaiserslautern in der nachfolgenden Saison als letzter Verein der Tabelle abgestiegen. Laut Regression wäre dies

**Tabelle 35:** Ergebnisse der Prognose der Punkte ohne Spieler

| Runde | Verein | LinReg | SMOReg |
|---|---|---|---|
| Hin | SV Werder Bremen | 18,71 (-3,45) | 23,09 (-3,15) |
| Hin | Borussia Dortmund | 25,59 (-7,71) | 29,64 (-3,25) |
| Hin | FC Bayern München | 41,82 (6,82) | 40,61 (4,82) |
| Hin | SC Freiburg | 25,19 (1,36) | 25,91 (2,60) |
| Hin | Hamburger SV | 20,18 (-6,34) | 29,26 (0,34) |
| Hin | Hannover 96 | 26,32 (0,76) | 21,44 (0,66) |
| Hin | 1899 Hoffenheim | 31,83 (3,48) | 30,78 (1,22) |
| Hin | 1. FC Kaiserslautern | 26,67 (2,41) | 18,16 (-4,91) |
| Hin | 1. FC Köln | 21,73 (2,75) | 20,49 (1,89) |
| Hin | Bayer 04 Leverkusen | 23,57 (-3,30) | 22,43 (-6,44) |
| Hin | 1. FSV Mainz 05 | 14,73 (-10,67) | 17,16 (-5,55) |
| Hin | Borussia Mönchengladbach | 22,64 (1,49) | 24,17 (2,85) |
| Hin | 1. FC Nürnberg | 19,76 (-1,65) | 16,83 (-5,17) |
| Hin | FC Schalke 04 | 16,73 (-4,92) | 18,01 (-2,84) |
| Hin | VfB Stuttgart | 23,65 (1,96) | 29,02 (4,68) |
| Hin | VfL Wolfsburg | 17,63 (-5,69) | 10,78 (-10,43) |
| Rück | SV Werder Bremen | 17,01 (-6,67) | 19,47 (-4,39) |
| Rück | Borussia Dortmund | 23,70 (-4,97) | 27,88 (-1,80) |
| Rück | FC Bayern München | 45,33 (4,75) | 38,81 (1,98) |
| Rück | SC Freiburg | 13,45 (-2,37) | 16,18 (0,18) |
| Rück | Hamburger SV | 20,44 (2,05) | 22,04 (3,65) |
| Rück | Hannover 96 | 27,94 (0,16) | 26,52 (0,13) |
| Rück | 1899 Hoffenheim | 22,25 (3,03) | 23,61 (0,35) |
| Rück | 1. FC Kaiserslautern | 27,37 (6,55) | 19,99 (-1,87) |
| Rück | 1. FC Köln | 29,42 (4,21) | 26,34 (3,19) |
| Rück | Bayer 04 Leverkusen | 24,28 (-5,43) | 31,83 (-2,10) |
| Rück | 1. FSV Mainz 05 | 25,76 (1,11) | 22,32 (-2,49) |
| Rück | Borussia Mönchengladbach | 29,54 (2,23) | 28,31 (2,31) |
| Rück | 1. FC Nürnberg | 26,24 (1,55) | 29,04 (3,94) |
| Rück | FC Schalke 04 | 13,95 (1,35) | 15,04 (0,97) |
| Rück | VfB Stuttgart | 29,48 (3,60) | 28,52 (-0,81) |
| Rück | VfL Wolfsburg | 20,75 (-0,01) | 22,30 (2,66) |

* Werte in Klammern: Differenz zum vorhergesagten Wert aus dem Training

in der Prognose nicht offensichtlich gewesen.

Bei der SMOreg Methode sind bei den Vereinen aus Leverkusen, Mainz, Wolfsburg und Bremen mit etwa -8 Punkten die größten Punkteverluste zu erwarten. Dahinter folgen der 1. FC Kaiserslautern und Borussia Dortmund mit Punkteverlusten von -6,78 bzw. -5,05 Punkten. Laut der SMOreg Methode haben die Vereine aus Schalke, Nürnberg, Hannover und Hoffenheim mit Punkteverlust- bzw. gewinn von -2 bis +2 einen geringen Punkteunterschied zu erwarten. Der SC Freiburg, VfB Stuttgart, der Hamburger SV, der 1. FC Köln, Borussia Mönchengladbach und der FC Bayern München können durch die Spielerverkäufe ihre Leistung um 2 bis 7 Punkte steigern.

Betrachtet man den Meister der Saison 2010 Borussia Dortmund wurden nur sehr wenige Spieler abgeben. Insgesamt wurden fünf Spieler verkauft, wobei drei dieser Spieler in den jeweiligen Runden unter 30 Minuten gespielt haben und ein Spieler nur in der Rückrunde zum Einsatz kam und dabei weniger als 100 Minuten gespielt hat. Somit beeinflussten diese vier Spieler nur wenig die Mannschaftsleistung von Borussia Dortmund in der Saison 2010. Der größte Verlust von Borussia Dortmund war der Abgang des Spielers Nuri Sahin, welcher für festgeschriebene 10 Millionen Euro zu Real Madrid wechselte. Dieser spielte große Teile der Hin- und Rückrunde für den Verein. Laut der linearen Regression machten die Verkäufe etwa -12 Punkte aus. Bei der SMOreg Methode etwa -5 Punkte. Laut beider Regressionen wäre durch die Verkäufe also eine Verschlechterung der Leistung eingetreten, welche die Verantwortlichen von Borussia Dortmund hätten auffangen müssen. Anhand der folgenden Prognose mit den Spielereinkäufen, soll nun geklärt werden ob dies laut Data Mining vorhergesagt werden würde. In der Realität konnte Borussia Dortmund u.a. durch den Einkauf von Ilkay Gündögan aus Nürnberg der Verlust von Nuri Sahin kompensiert werden und die Meisterschaft verteidigt werden.

Die Daten, welche alle Wechsel berücksichtigen werden nun auf die erlernten Modelle angewendet. Hierbei handelt es sich also um eine vollständige Prognose der nächsten Saison. Wie beschrieben, beinhalten die Daten die Spieler des neu zusammengestellten Kaders jeder Mannschaft der Saison 2010. Es handelt sich dabei um die Daten der Spieler aus der Saison 2010, sofern diese in dieser Saison gespielt haben. Ist dies nicht der Fall, werden die Daten aus der Saison 2011 hergenommen.

In Tabelle 36 und 37 sind die Ergebnisse beider Methoden dargestellt. Die Mannschaften sind nach den Punkten der Prognose absteigend geordnet. Zusätzlich sind die Punkte der Saison 2010 aufgelistet, die Punkte aus dem Training der Methoden und die Punkte, welche die Vereine tatsächlich in der Saison 2011 erspielt haben. Es handelt sich dabei um 16 Vereine. Die zwei Ab- und Aufsteiger sind nicht berücksichtigt, da von ihnen nur Daten von einer Saison vorliegen.

Die Ergebnisse zeigen, dass die hier vorgestellten Regressionen zur Prognose der nächsten Saison nur bedingt geeignet sind. Der RMSE liegt bei der linearen Regression bei über 14,74 Punkten und bei der SMOreg Methode bei 12,23 Punkten. Auch die prognostizierten Tabellen weichen wesentlich von der tatsächlichen Abschlusstabelle ab.

Analysiert man die Prognosen einzelner Mannschaften separat, so finden sich einzelne Erkenntnisse wieder, die für die Vereine interessant sein können. Aus beiden Regressionen ist zu erkennen, dass sich der FC Bayern München nach der Saison verstärkt hat. In der Realität konnte der FC Bayern München ebenfalls mehr Punkte erreichen. Laut den Methoden wäre der FC Bayern München Meister geworden, dies ist jedoch nicht eingetreten. Auch bei Bayer 04 Leverkusen sind beide Prognosen richtig. Bayer 04 Leverkusen holte insgesamt über 10 Punkte weniger als in der Saison zuvor. Beide Regressionen sagen einen hohen Punkteverlust von Saison 2010 zur Saison 2011 für Bayer 04 Leverkusen voraus. Laut beider Methoden gewinnt Borus-

sia Mönchengladbach an Qualität durch die Spielerwechsel. Dies ist auch tatsächlich eingetreten. Wobei der Unterschied in der Realität höher war, als es die Regressionen prognostizieren. Gleiches gilt für den FC Schalke 04. Sehr nah an der wirklichen Punktevergabe sind die Methoden beim 1. FSV Mainz 05. Mainz konnte in der Saison 2011 insgesamt 39 Punkte erspielen. Laut beider Regressionen liegt die Prognose für den 1. FSV Mainz 05 zwischen 35 und 38 Punkten. Beide Methoden konnten den erheblichen Qualitätsverlust durch die Abgabe der wichtigen Stammkräfte vorhersagen.

Vergleicht man die Prognose unter Berücksichtigung aller Spielerwechsel mit der Analyse der Punktevergabe ohne die Einkäufe von neuen Spielern, so konnten sich bei der linearen Regression der SV Werder Bremen und der FC Schalke 04 am stärksten verbessern. Durch Spielerzukäufe konnten die Leistung um etwa 11 bis 13 Punkten verbessert werden. Die Verbesserung beim 1. FC Köln und beim 1. FC Nürnberg liegt bei 1 bis 3 Punkten. Bei Hannover 96, dem SC Freiburg, Borussia Dortmund, Borussia Mönchengladbach, Bayer 04 Leverkusen und dem 1. FC Kaiserslautern ist durch Spielereinkäufe kein großer Unterschied zur Punktevergabe ohne Einkäufe zu erkennen. Die Vereine aus Mainz, Hamburg, Wolfsburg, München, Stuttgart und Hoffenheim verlieren durch die Spielereinkäufe sogar an Qualität von -6 bis -2 Punkten

Bei der SMOreg Methode können sich ebenfalls der FC Schalke 04, der SV Werder Bremen sowie der VfL Wolfsburg mit einer Differenz von 6 bis 9 Punkten durch Spielerzukäufe am stärksten verbessern. Der 1. FC Nürnberg, der 1. FC Köln, VfB Stuttgart, Bayer 04 Leverkusen, 1899 Hoffenheim und der 1. FC Kaiserslautern verbessern sich um 1 bis 2 Punkte durch die erfolgreichen Einkäufe. Kaum Unterschiede sind bei den Vereinen aus Freiburg, Köln, Hannover, Hamburg, Mönchengladbach, Mainz und Dortmund zu erkennen. Der 1. FC Nürnberg und der FC Bayern München verschlechtern sich sogar um etwa -2 Punkte.

Die Beispiele zeigen, dass die Prognosen in manchen Fällen richtig liegen und so zur Einschätzung der Leistung einiger Vereine in der nächsten Saison hergenommen werden können. Gerade Vereine wie der 1. FSV Mainz 05 oder Bayer 04 Leverkusen wären durch die Prognose frühzeitig gewarnt worden, dass in der nächsten Saison ein Leistungsabfall zu erwarten ist. Beim FC Schalke 04 und dem SV Werder Bremen ist es so, dass die Spielereinkäufe die Spielverkäufe kompensieren und die Leistung der Mannschaft sogar verbessern. Laut den Analysen konnten die Verantwortlichen von Borussia Dortmund den Verlust des Spielers Nuri Sahin nicht kompensieren, in der Realität ist ihnen dies jedoch gelungen.

**Tabelle 36:** Lineare Regression: Prognose der Punkte für die Saison 2011

| Verein | 2010 | Training | Prognose | 2011 |
|---|---|---|---|---|
| FC Bayern München | 65 | 75,58 | 83,39 | 73 |
| Hannover 96 | 60 | 53,34 | 54,73 | 48 |
| 1. FC Köln | 44 | 44,19 | 54,00 | 30 |
| 1. FC Kaiserslautern | 46 | 45,08 | 53,31 | 23 |
| Borussia Mönchengladbach | 36 | 48,46 | 51,68 | 60 |
| SV Werder Bremen | 41 | 45,84 | 49,14 | 42 |
| Borussia Dortmund | 75 | 61,97 | 48,99 | 81 |
| 1899 Hoffenheim | 43 | 47,57 | 47,99 | 41 |
| VfB Stuttgart | 42 | 47,57 | 47,91 | 53 |
| Bayer 04 Leverkusen | 68 | 56,58 | 47,30 | 54 |
| 1. FC Nürnberg | 47 | 46,10 | 47,21 | 42 |
| FC Schalke 04 | 40 | 34,25 | 42,56 | 64 |
| SC Freiburg | 44 | 39,65 | 39,02 | 40 |
| 1. FSV Mainz 05 | 58 | 50,05 | 38,38 | 39 |
| Hamburger SV | 45 | 44,91 | 38,13 | 36 |
| VfL Wolfsburg | 38 | 44,08 | 35,55 | 44 |

**Tabelle 37:** SMOreg: Prognose der Punkte für die Saison 2011

| Verein | 2010 | Training | Prognose | 2011 |
|---|---|---|---|---|
| FC Bayern München | 65 | 72,62 | 77,44 | 73 |
| VfB Stuttgart | 42 | 53,67 | 59,85 | 53 |
| Borussia Dortmund | 75 | 62,57 | 56,66 | 81 |
| Bayer 04 Leverkusen | 68 | 62,80 | 56,21 | 54 |
| 1899 Hoffenheim | 43 | 52,82 | 55,67 | 41 |
| Borussia Mönchengladbach | 36 | 47,32 | 52,13 | 60 |
| Hamburger SV | 45 | 47,31 | 51,18 | 36 |
| SV Werder Bremen | 41 | 50,10 | 49,40 | 42 |
| Hannover 96 | 60 | 47,17 | 48,01 | 48 |
| 1. FC Köln | 44 | 41,75 | 47,03 | 30 |
| 1. FC Nürnberg | 47 | 47,10 | 43,97 | 42 |
| VfL Wolfsburg | 38 | 40,85 | 42,53 | 44 |
| SC Freiburg | 44 | 39,31 | 42,29 | 40 |
| FC Schalke 04 | 40 | 34,92 | 39,55 | 64 |
| 1. FC Kaiserslautern | 46 | 44,93 | 39,37 | 23 |
| 1. FSV Mainz 05 | 58 | 47,52 | 38,91 | 39 |

**Tabelle 38:** Klassifikationsraten für das Training der Prognose

| Methode | Klassifikationsrate |
|---|---|
| MultiLayer Perceptron | 77,78 % |
| Naives Bayes | 76,39 % |
| Random Forest | 76,39 % |
| SMO | 79,17 % |

**Tabelle 39:** Klassifikationsraten der Prognose

| Methode | Klassifikationsrate |
|---|---|
| MultiLayer Perceptron | 59,38 % |
| Naives Bayes | 65,63 % |
| Random Forest | 62,5 % |
| SMO | 62,5 % |

#### 5.4.2.2 Rang

In diesem Abschnitt wird die Saison 2011 mithilfe der Klassifikation prognostiziert. In Kapitel 2.2.1.1 wird die Klassifikation beschrieben und in Kapitel 5.1 beispielhaft angewendet. Es werden bei dieser Prognose die Daten der Vereine mit allen Spielerverund einkäufen genutzt.

Zu Beginn der Klassifizierung werden mehrere Modelle anhand des Trainingsdatensatzes erlernt. Bei der Lernphase wird mit mehreren Methoden und unterschiedlicher Parametrisierung gearbeitet. Als Testmethode wird die Cross-validation Methode mit zehn Teilmengen gewählt. Die besten Ergebnisse können die Klassifizierer aus Tabelle 38 erzielen. Die Klassifikationsraten von über 76 % deuten auf eine gute Vorhersagbarkeit durch Data Mining hin.

Anschließend werden die neuen Daten auf die erlernten Modelle angewendet, um so eine Vorhersage zu generieren. Da man die Prognose mit den tatsächlichen Tabellenplatzierungen der Mannschaften aus der Saison 2011 testen kann, sind in Tabelle 39 die Klassifikationsraten der Anwendung aufgelistet.

Wie zu sehen ist, liegen die Klassifikationsraten der Methoden nah beieinander. Die Naives-Bayes Methode prognostiziert 21 der 32 Instanzen richtig, während die RandomForest Methode und die SMO Methode 20 Instanzen richtig einordnen. Die MultilayerPerceptron Methode ordnet 19 Vereine richtig ein. In Tabelle 40 sind die Prognosen der vier Klassifizierer dargestellt. Insgesamt werden 18 Instanzen von allen Methoden in die gleiche Klasse zugeordnet.

Die abschließende Prognose der Hin- und Rückrunde ist in Tabelle 41 dargestellt. Zur Generierung dieser Prognose werden alle Methoden miteinander kombiniert. Ein Verein wird der Klasse zugeordnet, in welche die Mehrheit der Methoden den entsprechenden Verein klassifiziert haben. Sofern ein Verein in einer Klasse nicht überrepräsentiert ist, wird die zugeordnete Klasse der SMO Methode genommen, da dieser Klassifizierer im Test am besten abschneidet.

Die kombinierte Prognose klassifiziert 22 von 32 Instanzen richtig und hat somit eine Klassifikationsrate von 68,75 %. Damit schneidet die Klassifikation besser ab als der naive Ansatz. Beim naiven Ansatz würde man alle Vereine in die Klasse {1} einordnen. Dadurch würde dieser Ansatz 18 der 32 Mannschaften richtig einordnen, was einer Klassifikationsrate von 56,25 % entspricht.

Bei Borussia Dortmund und dem FC Bayern München liegt die Prognose richtig. Beide Mannschaften haben sich durch die Wechsel nicht verschlechtert. Beim FC Schalke 04 und dem SV Werder Bremen wird der Aufstieg aus der mittleren Region

**Tabelle 40:** Prognosen der Klassifikation aller Methoden

| Runde | Verein | MP | NB | RF | SMO |
|---|---|---|---|---|---|
| Hin | FC Bayern München | 2 | 2 | 2 | 2 |
| Hin | Borussia Dortmund | 2 | 2 | 2 | 2 |
| Hin | FC Schalke 04 | 1 | 1 | 2 | 2 |
| Hin | Borussia Mönchengladbach | 2 | 2 | 2 | 2 |
| Hin | SV Werder Bremen | 2 | 2 | 1 | 0 |
| Hin | Bayer 04 Leverkusen | 1 | 1 | 1 | 1 |
| Hin | Hannover 96 | 2 | 1 | 2 | 1 |
| Hin | 1899 Hoffenheim | 2 | 1 | 1 | 2 |
| Hin | VfB Stuttgart | 2 | 2 | 0 | 2 |
| Hin | 1. FC Köln | 1 | 1 | 1 | 1 |
| Hin | VfL Wolfsburg | 1 | 1 | 1 | 1 |
| Hin | Hamburger SV | 1 | 2 | 2 | 1 |
| Hin | 1. FC Nürnberg | 0 | 0 | 1 | 1 |
| Hin | 1. FSV Mainz 05 | 1 | 1 | 1 | 1 |
| Hin | 1. FC Kaiserslautern | 1 | 0 | 1 | 1 |
| Hin | SC Freiburg | 1 | 1 | 1 | 1 |
| Rück | Borussia Dortmund | 2 | 2 | 2 | 2 |
| Rück | FC Bayern München | 2 | 2 | 2 | 2 |
| Rück | VfB Stuttgart | 2 | 2 | 2 | 2 |
| Rück | FC Schalke 04 | 1 | 1 | 1 | 1 |
| Rück | Bayer 04 Leverkusen | 2 | 2 | 2 | 2 |
| Rück | SC Freiburg | 0 | 0 | 0 | 0 |
| Rück | Borussia Mönchengladbach | 1 | 1 | 1 | 1 |
| Rück | Hannover 96 | 2 | 0 | 2 | 2 |
| Rück | VfL Wolfsburg | 1 | 0 | 1 | 1 |
| Rück | 1. FC Nürnberg | 1 | 1 | 1 | 1 |
| Rück | 1. FSV Mainz 05 | 1 | 1 | 1 | 1 |
| Rück | 1899 Hoffenheim | 1 | 1 | 1 | 0 |
| Rück | Hamburger SV | 1 | 1 | 1 | 1 |
| Rück | SV Werder Bremen | 2 | 0 | 0 | 0 |
| Rück | 1. FC Köln | 2 | 1 | 2 | 2 |
| Rück | 1. FC Kaiserslautern | 1 | 0 | 1 | 1 |

MP: MulitlayerPerceptron Methode
NB: Naive-Bayes Methode
RF: RandomForest Methode
SMO: SMO Methode

**Tabelle 41:** Prognose der Klassifikation

| Runde | Verein | Prognose | Saison2011 |
|---|---|---|---|
| Hin | FC Bayern München | 2 | 2 |
| Hin | Borussia Dortmund | 2 | 2 |
| Hin | FC Schalke 04 | 2 | 2 |
| Hin | Borussia Mönchengladbach | 2 | 2 |
| Hin | SV Werder Bremen | 2 | 2 |
| Hin | Bayer 04 Leverkusen | 1 | 1 |
| Hin | Hannover 96 | 1 | 1 |
| Hin | 1899 Hoffenheim | 2 | 1 |
| Hin | VfB Stuttgart | 2 | 1 |
| Hin | 1. FC Köln | 1 | 1 |
| Hin | VfL Wolfsburg | 1 | 1 |
| Hin | Hamburger SV | 1 | 1 |
| Hin | 1. FC Nürnberg | 1 | 1 |
| Hin | 1. FSV Mainz 05 | 1 | 1 |
| Hin | 1. FC Kaiserslautern | 1 | 0 |
| Hin | SC Freiburg | 1 | 0 |
| Rück | Borussia Dortmund | 2 | 2 |
| Rück | FC Bayern München | 2 | 2 |
| Rück | VfB Stuttgart | 2 | 2 |
| Rück | FC Schalke 04 | 1 | 2 |
| Rück | Bayer 04 Leverkusen | 2 | 2 |
| Rück | SC Freiburg | 0 | 1 |
| Rück | Borussia Mönchengladbach | 1 | 1 |
| Rück | Hannover 96 | 2 | 1 |
| Rück | VfL Wolfsburg | 1 | 1 |
| Rück | 1. FC Nürnberg | 1 | 1 |
| Rück | 1. FSV Mainz 05 | 1 | 1 |
| Rück | 1899 Hoffenheim | 1 | 1 |
| Rück | Hamburger SV | 1 | 1 |
| Rück | SV Werder Bremen | 0 | 1 |
| Rück | 1. FC Köln | 2 | 0 |
| Rück | 1. FC Kaiserslautern | 1 | 0 |

in die Top-5 der Hinrunde erkannt. In der Rückrunde werden jedoch beide Vereine schlechter als der tatsächliche Wert eingestuft. Laut der Prognose wäre der SV Werder Bremen sogar auf einem Abstiegsplatz gelandet. Im Gegensatz zur Regressionsanalyse aus dem vorherigen Kapitel, kann die Klassifikation den großen Sprung von Borussia Mönchengladbach von einem Abstiegsplatz in der Hinrunde 2010 auf einen Platz unter den besten Fünf erkennen. Der Rang von Borussia Mönchengladbach in der Rückrunde wird ebenfalls richtig klassifiziert. Bei Bayer 04 Leverkusen liegt die Klassifikation in beiden Runden richtig. In der Hinrunde wird ein Leistungsabfall von der Top-Region zu den mittleren Tabellenplätzen vorhergesagt. Die Vorhersage für den 1. FSV Mainz 05 stimmt ebenfalls. Auch hier werden die Spielverkäufe durch die Einkäufe nicht kompensiert. Laut Prognose wäre der 1. FSV Mainz 05 in der Hinrunde aus den Top-5 in die mittlere Region gefallen. Im Gegensatz dazu hätte die Klassifikation den 1. FC Kaiserslautern nicht als Absteiger vorhergesagt. Der 1. FC Kaiserslautern wird in beiden Runden auf einen der Plätze zwischen sechs und 15 klassifiziert.

Insgesamt werden die Absteiger nicht richtig ermittelt. Die tatsächlichen Absteiger werden nicht erkannt und mit dem SV Werder Bremen und dem SC Freiburg zwei Mannschaften als Absteiger klassifiziert, welche auf einem mittleren Tabellenplatz landeten. Wobei der SV Werder Bremen in der Rückrunde nur um zwei Punkte die Abstiegsregion vermied.

Die Ergebnisse der Klassifikation sind zufriedenstellender als die der Regressionsanalyse. Mit dem Nachteil bei der Klassifikation, dass nur die drei Regionen betrachtet werden und keine vollständige Tabelle für die Zukunft erstellt werden kann.

### 5.4.3 Einschätzung der Ergebnisse

Die Prognose der Punkte liefert unzureichende Ergebnisse verglichen mit dem tatsächlichen Ausgang der Saison. Die Vorhersage mit der Klassifikation kann dagegen gering bessere Ergebnisse erzielen.

Für die schlechten Ergebnisse kann es mehrere Gründe geben. Einen großen Einfluss haben die Daten auf die Ergebnisse der Verfahren. In dieser Arbeit wurde ein Ansatz für die Aggregation der Mannschaften vorgestellt, welcher eine Gewichtung vornimmt. Es sind hier weitere Ansätze denkbar, die einerseits eine verschiedenartige Gewichtung vornehmen oder generell die Mannschaften anders repräsentieren. Zusätzlich zu der hier vorgestellten Repräsentation, wurde außerdem versucht die Teams in ihre Mannschaftsteile aufzusplitten, um so einen höheren Detailgrad zu erreichen. Dabei wurden für jede Mannschaft die vorgestellten Attribute für jede Position wie Abwehr, Mittelfeld und Angriff zusammengefasst. Statt eines Zweikampfwerts für die gesamte Mannschaft, enthält so beispielsweise jedes Team die Zweikampfwerte für die Abwehr, für das Mittelfeld und für den Angriff separat im Datensatz. Die Ergebnisse bei der linearen Regression und bei der Klassifikation schneiden jedoch wesentlich schlechter ab, als die Betrachtung der gesamten Mannschaft. Somit wird auf die Darstellung der Ergebnisse im Weiteren verzichtet.

Zusätzlich sind in dem hier zur Verfügung stehenden Datensatz nur einige der Attribute enthalten, welche pro Spieler und Spiel aufgenommen werden können. So fehlen beispielsweise Attribute, wie die Laufleistung, Anzahl an Angriffen im vorderen Drittel oder die genaue Position bei Torschüssen. Solche Informationen könnten die Ergebnisse verbessern. Außerdem fehlen die Daten der Torhüter, welche bei einem Fußballspiel eine bedeutende Rolle spielen. So kann ein Torwart den Ausgang eines Spiels mit seinen Paraden wesentlich beeinflussen. Auch Daten über die Trainer oder das Zuschaueraufkommen können Informationen darstellen, welche Einfluss auf die Ergebnisse haben.

Es liegen zudem Daten für lediglich zwei Saisons vor. Dadurch entstehen in diesem Beispiel insgesamt nur 72 Instanzen aus denen ein Data Mining Verfahren lernen kann. Dies sind sehr wenige Daten um unentdeckte Muster zu erkennen. Eine Analyse über wesentlich mehr Saisons könnten die Ergebnisse positiv beeinflussen.

Ob eine akkurate Prognose mit erweitertem Datensatz möglich ist, ist jedoch weiterhin fraglich. Wie erwähnt, ist es schwer möglich eine Saison vorherzusagen. Beispielsweise können Ereignisse, wie Spielerverletzungen oder Trainerwechsel die Leistung einer Mannschaft wesentlich mindern bzw. steigern. Trotzdem sind die hier aufgezeigten Anwendungen für Verantwortliche eines Vereins ein Ansatz zur Analyse ihrer Spielertransfers.

In dieser Arbeit kann keine Prognose für die Saison 2012 abgegeben werden. Da viele Spieler aus anderen Ligen in die Bundesliga gewechselt sind und für diese keine Daten vorliegen, sind die neu zusammengestellten Teams nur unzureichend abbildbar. Eine Prognose für die Saison 2012 wäre somit nicht hilfreich.

# 6 Die Wichtigkeit eines Spielers messen

Die vorhergehenden Kapitel über die einzelnen Data Mining Verfahren, wie der Klassifikation, Regression sowie der Clusteranalyse sollten exemplarisch aufzeigen, wie Data Mining im Fußball angewendet werden kann. Dabei ist bisher nicht darauf eingegangen worden, inwiefern spezielle Spieler, zum Erfolg einer Mannschaft beitragen. In den folgenden Abschnitten, soll mithilfe der Regression separat sowie einer Kombination aus Clusteranalyse und linearer Regression dieser Frage nachgegangen werden.

Spieler werden an ihrem Marktwert gemessen. Das heißt, der Betrag den ein Spieler bei einem potenziellen Verkauf erwirtschaftet, spiegelt seinen Wert für den Verein wieder. Für Nuri Sahin erwirtschaftete Borussia Dortmund nach der Saison eine festgeschriebene Ablösesumme von 10 Millionen Euro [64]. Mario Götze wechselt im Sommer 2013 für festgeschriebene 37 Millionen zum FC Bayern München [65]. Inwiefern spiegelt jedoch die Höhe der Ablösesumme den tatsächlichen Wert eines Spielers wieder? Da mehrere Parteien an den Verhandlungen über die Ablöse beteiligt sind und diese versuchen das Beste für sich aus der Situation zu verhandeln, kann man nicht direkt von der Ablösesumme auf die Wichtigkeit des Spielers schließen.

Entscheidend für einen Verein ist es zu wissen, inwiefern der Spieler zum Erfolg der Mannschaft beiträgt. Da die Punkteausbeute die Leistung einer Mannschaft widerspiegelt, ist man daran interessiert mit welchem Anteil ein Spieler an den Punkten einer Saison beteiligt ist. Im Folgenden soll genau dieser Anteil mithilfe von Data Mining versucht werden zu berechnen. Dazu soll am Ende diesen Kapitels die Frage beantworten werden, ob man messen kann, wie sehr sich der Verlust eines Spielers auf die Leistung der Mannschaft auswirkt.

Abschnitt 6.1 nimmt die beiden erlernten Modelle der Regressionsanalyse aus Kapitel 5.4.2.1, um zu Messen, inwiefern einzelne Spieler zur Punkteausbeute der Mannschaften beitragen. Der gleiche Ansatz wurde in dem Kapitel schon mithilfe des Datensatzes verfolgt, welcher die Mannschaften ohne die verkauften Spieler betrachtet. In diesem Kapitel wird die gleiche Anwendung für einzelne Spieler durchgeführt.

In Abschnitt 6.2 wird ein Ansatz zur Messung der Wichtigkeit eines Spielers vorgestellt, welcher die Clusteranalyse aus Kapitel 5.3 mit der Regressionsanalyse kombiniert. Dabei wird die Idee aus dem in Kapitel 3.6 vorgestellten Artikel von Chan, Cho und Novati [7] aufgegriffen. Hier wird für jede Mannschaft berechnet, wie viele Minuten alle Spieler der einzelnen Cluster für die betrachtete Mannschaft in einer Runde gespielt haben. Mithilfe dieser Repräsentation wird ein Regressionsmodell erlernt, welches Aufschluss über die Wichtigkeit einzelner Spieler und ihrer Spielweise geben soll.

## 6.1 Verwendung der Regressionsanalyse

In Kapitel 5.4.2.1 wurden mithilfe der linearen Regression sowie der SMOreg Methode zwei Regressionsmodelle erlernt, welche die Punkte einer Saison vorhersagen. In dem Kapitel wurden die Modelle genutzt, um eine zukünftige Saison vorherzusagen. Eine weitere Möglichkeit ist es, die erlernten Formeln zu nehmen und für einzelnen Mannschaften zu ermitteln, wie sie ohne bestimmte Spieler in der Saison abgeschnitten hätten. Dadurch ist es möglich die Wichtigkeit eines Spielers in erspielten Punkten zu messen.

Die Formeln der beiden Regressionen sind in Formel 9 für die lineare Regression und Formel 10 für die SMOreg Methode dargestellt. Da die spielspezifischen Werte, wie die Zweikampfquote oder die Anzahl an Ballkontakten in verschiedenen Wertebereichen liegen, kann anhand der Koeffizienten keine Aussage getätigt werden, welche

Werte die Punkte am ehesten beeinflussen. Auf eine Normalisierung der Werte wurde verzichtet, da es so möglich ist, die erlernten Formeln auf neue Daten anzuwenden.

$$
\begin{aligned}
pkte = \\
& 0,0572 * zweikampf - 0,0572 * verlZweikampf \\
& + 1,9089 * fehlPaesse + 0,9236 * Paesse \\
& + 0,3213 * erfPassquote - 0,3215 * fehlPassquote \\
& - 25,3893 * vergTorschuesse + 0,4447 * erfTorschussquote \\
& - 0,4438 * fehlTorschussquote + 23,2081 * Torschussvorlagen \\
& + 32,2395 * Vorlage + 55,1416 * Tore \\
& - 16,9865 * Gelb - 129,4646 * PassProBallkontakt \\
& - 40,7453 * erfPassProBallkontakt \\
& - 167,6621 * TorschussvorlageProPass \\
& + 99,3397
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
pkte = \\
& 0,0203 * zweikampf + 1,5055 * Ballkontakte \\
& - 0,477 * erfPaesse - 0,0271 * fehlPaesse \\
& - 0,5042 * Paesse - 0,0488 * erfPassquote \\
& + 0,187 * Torschuesse - 0,0653 * vergTorschuesse \\
& + 1,2205 * erfTorschussquote - 1,2205 * fehlTorschussquote \\
& + 0,4069 * Torschussvorlagen + 0,3476 * Vorlage \\
& + 0,2523 * Tore - 0,4332 * Gelb \\
& + 0,0053 * PassProBallkontakt + 0,0001 * erfPassProBallkontakt \\
& - 0,0097 * TorschussvorlageProPass \\
& + 84,7672
\end{aligned}
\tag{10}
$$

Die Formeln lassen sich auf die zusammengefassten Werte der Vereine anwenden. Das heißt, für die Zweikampfquote werden alle Quoten der Spieler des gesamten Vereins für die jeweilige Runde summiert. Es wird hier zusätzlich eine Gewichtung vorgenommen, sodass die Werte der Spieler die mehr Einsatzzeit hatten höher bewertet werden als die, welche nur wenige Minuten gespielt haben. Dies wird erreicht indem man jeden Wert mit der Einsatzzeit des Spielers multipliziert, diese Produkte jeweils addiert und durch die gespielten Minuten der Mannschaft in der jeweiligen Runde teilt. Die Werte, wie die Anzahl an Ballkontakten oder die Torschussvorlagen kommen aus der Tabelle **Fakten**. Hier sind die Aktionen eines jeden Spielers in einem Spiel gespeichert. Diese Daten liegen für die beiden Saisons 2010 und 2011 vor.

Um die Wichtigkeit eines Spielers zu ermitteln werden im Folgenden anhand der Formeln die Punkte mit dem betrachteten Spieler errechnet und von diesen Punkten werden die prognostizierten Punkte ohne den Spieler abgezogen. Dies wird erreicht indem man die Aggregation für den jeweiligen Verein mit den Originaldaten aus der Tabelle **Fakten** durchführt. Diese Werte können dann in die Formel eingesetzt werden. Um die Punkte ohne den Spieler zu ermitteln, wird die gleiche Aggregation durchgeführt, jedoch wird der betrachtete Spieler aus den Originaldaten herausgefiltert und so seine Daten ignoriert. Damit werden zwei Punkteprognose erreicht. Einmal eine Prognose mit dem ursprünglichen Kader und einmal ohne den Spieler dessen Wert ermittelt werden soll.

**Tabelle 42:** Spielercluster Statistik

| Position | Cluster | Anzahl | Beschreibung |
|---|---|---|---|
| Abwehr | 0 | 32 | Top Außenverteidiger |
| | 1 | 51 | Torgefährliche Innenverteidiger |
| | 2 | 68 | Außenverteidiger |
| | 3 | 31 | Zweikampfstarke Verteidiger |
| | 4 | 167 | Innenverteidiger |
| Mittelfeld | 0 | 55 | Durchschnittliche Spieler |
| | 1 | 69 | Top offensives Mittelfeld |
| | 2 | 137 | Defensives Mittelfeld |
| | 3 | 85 | Top defensives Mittelfeld |
| | 4 | 55 | Offensives Mittelfeld |
| Angriff | 0 | 37 | Top Mittelstürmer |
| | 1 | 10 | Top Vorlagengeber |
| | 2 | 26 | Mitspielende Stürmer |
| | 3 | 38 | Mittelstürmer |
| | 4 | 29 | Außenstürmer |
| | 5 | 37 | Mittelstürmer |

Bevor die Werte für die einzelnen Spieler ermittelt werden, wird im folgenden Abschnitt ein weiterer Ansatz für die Messung der Wichtigkeit eines Spielers vorgestellt. Beide Möglichkeiten kommen in Abschnitt 6.3 exemplarisch zur Anwendung.

## 6.2 Kombination aus Clustering und Regression

Chan, Cho und Novati [7] kombinieren zur Messung der Wichtigkeit von Eishockeyspielern die Cluster- und Regressionsanalyse. In diesem Artikel werden die Spieler der National Hockey League anhand ihrer Position, wie Torhüter, Defensive und Offensive geclustert. Anschließend wird jede Mannschaft anhand dieser Clusterzuordnungen repräsentiert. Für jede Mannschaft wird berechnet, wie viele Minuten einer Saison die Spieler einer Mannschaft im Durchschnitt in den identifizierten Clustern gespielt haben. Mithilfe dieser Repräsentation der Mannschaft wird dann eine Regressionsfunktion erlernt, an der man erkennen kann mit welchem Gewicht die einzelnen Cluster zu den erzielten Punkten am Saisonende beitragen.

In dieser Arbeit wurde in Kapitel 5.3 die Bundesligaspieler der beiden Saisons 2010 und 2011 in homogene Gruppen eingeteilt. Dabei wurden die Spieler der Positionen Abwehr, Mittelfeld und Angriff, welche mindestens ein Drittel einer Runde gespielt haben geclustert.

Die Ergebnisse ergeben in der Abwehr und Mittelfeld jeweils fünf identifizierte Cluster. Im Angriff werden sechs Cluster identifiziert. Tabelle 42 zeigt die Anzahl an Spielern der einzelnen Cluster sowie eine Beschreibung der einzelnen Cluster.

Für die Regressionsanalyse berechnet man pro Mannschaft die Anzahl an gespielten Minuten der Spieler des Vereins für die einzelnen Cluster. Dies wird für jede Runde berechnet. Hat beispielsweise ein Verein in der Hinrunde der Saison 2010 zwei Abwehrspieler aus Cluster 2 je 100 Minuten eingesetzt, so besitzt der entsprechende Verein in der Zelle für das Cluster 2 die Summe der Einsatzzeit der zwei Spieler, nämlich 200 Minuten. Die Anzahl an gespielten Minuten wird aus der Tabelle **Fakten** aggregiert. Das Cluster und die Position des Spielers kommt aus der Tabelle **Clusteredspieler**, welche die Ergebnisse aus der Clusteranalyse dieser Arbeit beinhaltet. Das zugehörige SQL Statement ist im Anhang in Abbildung 28 zu finden.

Zusätzlich werden die Punkte als Zielwert für die Regressionsanalyse benötigt. Diese sind in der Tabelle **Abschlusstabelle** gespeichert. Die Vereine werden mit den entsprechenden Daten aus dieser Tabelle verknüpft.

Für die Analyse liegen 72 Dateninstanzen vor. Dabei handelt es sich um die Daten der Hin- und Rückrunden der beiden Saisons 2010 und 2011. Die Zielwerte sind die Punkte. Als unabhängige Variablen gelten die Summen der gespielten Minuten in den verschiedenen Clustern.

Im Folgenden wird nur die lineare Regression zur Analyse eingesetzt. Die Tests mit der Cross-validation Methode bei zehn Teilmengen zeigen bei den verschiedenen Regressionsmethoden nur geringe Abweichungen im RMSE.

Die lineare Regression, welche kollineare Attribute entfernt, erreicht einen RMSE von 6,2584. Formel 11 zeigt die zugehörige Formel der Analyse.

$$
\begin{aligned}
pkte = \\
&+ 0,0025 * Abwehr0 + 0,0027 * Abwehr1 \\
&+ 0,0017 * Abwehr2 + 0,0030 * Abwehr3 \\
&+ 0,0010 * Abwehr4 \\
&+ 0,0023 * Mittelfeld0 + 0,0064 * Mittelfeld1 \\
&+ 0,0023 * Mittelfeld2 + 0,0037 * Mittelfeld3 \\
&+ 0,0022 * Mittelfeld4 \\
&+ 0,0070 * Angriff0 + 0,0025 * Angriff1 \\
&+ 0,0009 * Angriff2 + 0,0046 * Angriff3 \\
&+ 0,0036 * Angriff4 + 0,0024 * Angriff5 \\
&- 15,0567
\end{aligned}
\tag{11}
$$

Die Gleichung stellt die Wichtigkeit der einzelnen Cluster dar. Je höher der Koeffizient des Attributs, desto wichtiger ist die unabhängige Variable. Wie man sieht hat keines der Cluster einen negativen Einfluss auf die Anzahl der Punkte.

In der **Abwehr** liegen die Koeffizienten nah beieinander. Keines der Cluster sticht in der Wichtigkeit heraus. Cluster 3 hat den höchsten Koeffizienten mit 0,0030. Cluster 0 und 1 mit 0,0025 und 0,0027 sind nur gering weniger wichtig. Cluster 2 fällt mit 0,0017 im Vergleich etwas ab.

Im **Mittelfeld** ist Cluster 1 das wichtigste Cluster. Es hat einen Koeffizienten von 0,0064. Cluster 3 ist mit 0,0037 wesentlich weniger wichtig, beeinflusst die Punkte jedoch mehr als das wichtigste Cluster der Abwehr. Cluster 0 (0,0023), Cluster 2 (0,0023) und Cluster 4 (0,0022) haben geringeren Einfluss.

Cluster 0 ist das herausragende Cluster im **Angriff**. Mit einem Koeffizienten von 0,0070 ist es die wichtigste Variable der gesamten Formel. Cluster 3 mit 0,0046 liegt in der Wichtigkeit dahinter. Das Cluster 4 beeinflusst die Punkte mit 0,0036. Cluster 1 und Cluster 4 haben mit 0,0025 und 0,0024 noch geringen Einfluss, während Cluster 2 mit 0,0009 den geringsten Koeffizienten der Formel aufweist.

Nach der Größe des Koeffizienten geordnet ergibt sich folgende Reihenfolge:

- Angriff0 (0,0070)
- Mittelfeld1 (0,0064)
- Angriff3 (0,0046)
- Mittelfeld3 (0,0037)
- Angriff4 (0,0036)
- Abwehr3 (0,0030)
- Abwehr1 (0,0027)
- Abwehr0 (0,0025)

- Angriff1 (0,0025)
- Angriff5 (0,0024)
- Mittelfeld0 (0,0023)
- Mittelfeld2 (0,0023)
- Mittelfeld4 (0,0022)
- Abwehr2 (0,0017)
- Abwehr4 (0,0010)
- Angriff2 (0,0009)

Vergleicht man die Wichtigkeit der Cluster mit den identifizierten Clustern und ihren Eigenschaften so erkennt man, dass sich die vorgenommene Einstufung der Cluster in der Formel widerspiegelt.

Das wichtigste Cluster laut der Regressionsanalyse stellt Cluster 0 des Angriffs dar. Dieses Cluster beschreibt die Top-Mittelstürmer der Bundesliga. Dies sind die Spieler, welche mit Abstand die meisten Tore schießen und die meisten Torschüsse abgeben. Im Mittelfeld erspielt Cluster 1 die meisten Punkte. Hierbei handelt es sich um die besseren offensiven Mittelfeldspieler. Die Spieler dieses Clusters geben die meisten Vorlagen und schießen die meisten Tore aller Mittelfeldspieler. An dritter Stelle steht das Cluster 3 des Angriffs, welches überwiegend Mittelstürmer beinhaltet. Diese Spieler haben die zweitbeste Torquote der Cluster des Angriffs. Im Cluster 3 des Mittelfelds, welches in der Wichtigkeit dahinter folgt, ist die Top-Defensive der Mittelfeldspieler zu finden. Dies sind die zweikampfstärksten Mittelfeldspieler mit den meisten Ballkontakten und den zweitmeisten Toren aller Mittelfeldspieler. Kurz dahinter folgt Cluster 4 des Angriffs. Dies sind überwiegend die Außenstürmer mit vielen Ballkontakten, bester Passquote und den meisten Vorlagen. Dahinter liegen die Abwehrspieler aus Cluster 3 und Cluster 1. Cluster 3 beinhaltet Spieler, die gute Zweikampfwerte besitzen aber sonst keine besonderen Merkmale aufweisen. In Cluster 1 sind überwiegend Innenverteidiger eingeteilt, welche verhältnismäßig viele Tore schießen, gut passen und ein gutes Zweikampfverhalten aufweisen. Danach folgen die Top-Außenverteidiger in der Reihenfolge. Die restlichen Cluster der einzelnen Position weisen keine besonderen Merkmale auf.

Mithilfe der Koeffizienten lässt sich nun die Wichtigkeit einzelner Spieler für die Saison bewerten. Im Folgenden Abschnitt soll die Wichtigkeit einzelner Spieler evaluiert werden. Dies geschieht mit den beiden vorgestellten Methoden.

## 6.3 Wie wichtig ist ein Spieler?

Wie in Abschnitt 6.1 beschrieben, lassen sich durch die erlernten Modelle aus Kapitel 5.4, die Punkte vergleichen, die laut den Regressionsanalysen tatsächlich erspielt wurden und die Punkte, die ohne bestimmte Spieler erspielt worden wären. Die Differenz dieser Punkte gibt die Punkte wieder, welche der Spieler zur Leistung der Mannschaft beigetragen hat.

Bei dem Ansatz, welcher die Cluster und Regressionsanalyse kombiniert, werden die Koeffizienten des Cluster in dem ein Spieler in der betrachteten Runde gespielt hat mit der Einsatzzeit des Spielers in der jeweiligen Runde multipliziert. Das Produkt ergibt die erspielten Punkte für die jeweilige Runde, welche der Spieler laut der Analyse zur Gesamtpunktzahl beigetragen hat. In den Tabellen 43 und 44 sind die zehn Spieler der Saison 2010 und 2011, welche laut Kombination aus Clustering und Regression für die meisten Punkte in der gesamten Saison zuständig waren. Diese bestehen ausschließlich aus Angreifern und Mittelfeldspielern.

Um die ermittelten Punkte im realen Einsatz zu testen, können einige vergangene Transfers der Bundesliga als Beispiel hergenommen werden. Ein gutes Beispiel liefern dabei die Transfers von Borussia Dortmund in der Saison 2010. Wie bereits in dieser

**Tabelle 43:** Die Zehn besten Punktelieferanten 2010

| Saison | Spieler | Position | Punkte |
|---|---|---|---|
| 2010 | Papiss Cisse | Angriff (C0) | 19,628 |
| 2010 | Srdjan Lakic | Angriff (C0) | 18,655 |
| 2010 | Marco Reus | Mittelfeld (C1) | 18,2848 |
| 2010 | Thomas Müller | Mittelfeld (C1) | 17,8944 |
| 2010 | Didier Konan Ya | Angriff (C0) | 17,227 |
| 2010 | Kevin Großkreutz | Mittelfeld (C1) | 16,6592 |
| 2010 | Mario Götze | Mittelfeld (C1) | 16,2624 |
| 2010 | Milivoje Novakovic | Angriff (C0) | 13,5766 |
| 2010 | Mario Gomez | Angriff (C3) | 13,5318 |
| 2010 | Theofanis Gekas | Angriff (C0) | 13,0155 |

**Tabelle 44:** Die Zehn besten Punktelieferanten 2011

| Saison | Spieler | Position | Punkte |
|---|---|---|---|
| 2011 | Klaas Jan Huntelaar | Angriff (C0) | 19,621 |
| 2011 | Marco Reus | Mittelfeld (C1) | 17,7792 |
| 2011 | Thomas Müller | Mittelfeld (C1) | 17,28 |
| 2011 | Robert Lewandowski | Angriff (C0) | 16,7048 |
| 2011 | Shinji Kagawa | Mittelfeld (C1) | 15,4688 |
| 2011 | Juan Arango | Mittelfeld (C1) | 15,3346 |
| 2011 | Kevin Großkreutz | Mittelfeld (C1) | 14,944 |
| 2011 | Mario Gomez | Angriff (C0) | 14,5662 |
| 2011 | Vedad Ibisevic | Angriff (C0) | 14,098 |
| 2011 | Claudio Pizarro | Angriff (C0) | 13,377 |

Arbeit erwähnt, war Nuri Sahin der Einzige der abgegebenen Spieler, welcher Stammspieler in der Mannschaft war. Er wurde zudem zum Spieler der Saison gewählt. Nuri Sahin hat laut der Kombination insgesamt 9,6052 Punkte für Borussia Dortmund in der Saison erspielt. Als Alternativ wurde Ilkay Gündögan vom 1. FC Nürnberg verpflichtet. Ilkay Gündögan hat in der Saison 2010 für den 1. FC Nürnberg insgesamt 8,5632 Punkte erspielt, wobei Ilkay Gündögan große Teile der Rückrunde verletzungsbedingt pausieren musste. Somit konnte Nuri Sahin durch den neuen Spieler im Sinne der Punkte sehr gut ersetzt werden.

Bei den Regressionsanalysen hat Nuri Sahin laut der linearen Regression 13,51 und laut SMOreg 5,85 Punkte für die Saison 2010 erspielt. Die Wichtigkeit des Spielers wird anhand aller drei Analysen aufgedeckt. Ilkay Gündögan war laut linearer Regression dagegen nur an 0,08 Punkten beteiligt. Laut SMOreg Methode wären ohne Ilkay Gündögan sogar 0,05 Punkte mehr erspielt worden. Hier ist also ein klarer Unterschied in den Analysen zu erkennen.

In der vorherigen Prognose aus Kapitel 5.4 wurde der Qualitätsunterschied vom 1. FC Kaiserslautern in der Saison 2010 durch die Spielerverkäufe nicht erkannt. Untersucht man die abgegebenen Stammkräfte vom 1. FC Kaiserslautern in den Regressionsanalysen separat, so hat Srdjan Lakic laut der linearen Regression die Saisonleistung mit -2,65 verlorenen Punkten negativ beeinflusst. Die SMOreg Methode schätzt die Wichtigkeit des Spielers auf 5,90 Punkten in der Saison 2010 ein. Eine weitere Stammkraft, nämlich Jan Moravek hätte laut den beiden Regressionen etwa 1 bis 2 Punkte zur Saisonleistung beigesteuert. Laut der Kombination aus Clusteranalyse und Regression war Srdjan Lakic wesentlich am Erfolg der Mannschaft in der Saison 2010 beteiligt. Er hat alleine 18,655 Punkte erspielt. Jan Moravek beispielsweise zusätzliche 5,9134 Punkte.

Beim 1. FSV Mainz 05 wird der Verlust in der Saison 2010 der drei Stammkräfte Andre Schürrle (10,561 Punkte), Lewis Holtby (8,6402) und Christian Fuchs (5,6502) ebenfalls durch Kombination aus Cluster- und Regressionsanalyse deutlich. Die lineare Regression sowie die SMOreg Methode haben in Kapitel 5.4 den Qualitätsunterschied vom 1. FSV Mainz 05 zur Vorsaison sehr gut prognostiziert. In der separaten Analyse hat Andre Schürrle seinem Verein dem 1. FSV Mainz 05 nach der linearen Regression -1,44 Punkte gekostet, Lewis Holtby 2,33 Punkte erspielt und Christian Fuchs 7,41 Punkte gewonnen. Die SMOreg Methode schätzt Andre Schürrle mit 4,69 erspielten Punkten positiv, Lewis Holtby mit -0,21 Punkten negativ und Christian Fuchs mit 3,45 Punkten positiv ein.

Analysiert man den berühmtesten Transfer der Saison 2011, den von Marco Reus von Borussia Mönchengladbach zu Borussia Dortmund, so ergibt die Kombination der Analysen, dass Marco Reus 2011 17,78 Punkte für Borussia Mönchengladbach erspielt hat. Die beiden Regressionsanalysen werten den Spieler auf etwa 5 bis 6 erspielten Punkte für die Saison 2011. Es wurden einige neue Spieler für Marco Reus transferiert, es konnte sich jedoch keiner dieser Spieler in der Saison 2012 wesentlich in den Vordergrund spielen. Das Ergebnis für Borussia Mönchengladbach war, dass insgesamt 13 Punkte weniger gewonnen wurden als in der Saison 2011 mit Marco Reus. Der Transfer des wichtigen Spielers konnte nicht aufgefangen werden. Die Analysen liegen somit sehr nahe am Punkteverlust der durch den Transfer entstanden ist.

Borussia Dortmund dagegen konnte 2012 auch mit Marco Reus, der in den beiden Saisons davor über 17 Punkte für Borussia Mönchengladbach erspielt hat, nicht mehr Punkte erreichen als in der Saison 2011. Dies kann daran liegen, dass mit Shinji Kagawa laut Clusteranalyse in Kombination mit der Regressionsanalyse ein Spieler abgegeben wurde der 15,4688 Punkte in der Saison 2011 erspielte. Also ein wesentlicher Verlust für die Mannschaft. Die beiden Regressionen bewerten den Spieler mit etwa 3

erspielten Punkten.

Anhand der Beispiele scheint es, als würde die Kombination aus Clustering und Regression die Spieler, welche subjektiv betrachtet die Leistung ihrer Mannschaften positiv beeinflusst haben, näher an dieser subjektiven Betrachtung zu bewerten. Man kann sagen, dass Stürmer Srdjan Lakic mit insgesamt 16 erzielten Toren in der Saison 2010 wesentlichen Anteil am Erfolg vom 1. FC Kaiserslautern hatte. Die Regressionsmethoden bewerten ihn eher schlecht. Die Kombination der Analysen bewertet den Spieler dagegen mit 18 Punkten sehr positiv. Gleiches gilt auch für Ilkay Gündogan der aufgrund seiner Leistung in der Saison 2010 im August 2011 zum Nationalspieler berufen wurde. Die beiden Regressionsanalysen bewerten Ilkay Gündögan als einen Spieler, der die Punkte weder positiv noch negativ beeinflusst hat. Bei der Kombination der Cluster- und Regressionsanalyse wird der Spieler sehr gut bewertet, was er auch in der nachfolgenden Saison bei Borussia Dortmund unter Beweis gestellt hat.

Anschließend an die beiden vorgestellten Ansätze zur Messung der Wichtigkeit eines Spielers, sollen diese Ergebnisse mit den Resultaten aus den Anwendungen aus Kapitel 5 für eine Auswahl an Spielern aufgeführt werden. Dazu werden die zehn Feldspieler genommen, welche von den Mitgliedern der Vereinigung der Vertragsfußballspieler (VDV) für die Saison 2010 sowie 2011 in die Bundesliga-Mannschaft der Saisons gewählt worden sind. Bei dieser Wahl können Mitglieder der VDV inklusive aller Spieler der Bundesliga, 2. Bundesliga, 3. Liga und der Regionalligen, die ihrer Meinung nach besten Spieler der Saison wählen. In Tabelle 45 und Tabelle 46 sind die zehn besten Feldspieler für die Saison 2010 bzw. 2011 aufgelistet. Zusätzlich ist dargestellt, ob der Spieler als Nationalspieler der beiden Klassifikationsmethoden aus Kapitel 5.1 klassifiziert wurde, wie seine Note aus der in Kapitel 5.2 identifizierten linearen Regression für die Saison aussieht, sowie die Punktzahl, die er in diesem Kapitel innerhalb der linearen Regression, der SMOreg Methode sowie bei der Kombination aus Clustering und Regression für seine Mannschaft erspielt hat.

Es zeigt sich, dass 2010 sechs bzw. fünf der zehn Feldspieler als Nationalspieler vom neuronalen Netzwerk bzw. der RandomForest Methode erkannt werden. Bei den Noten zeigt sich, dass sieben der zehn Spieler eine Note unter dem Durchschnitt erhalten. Bei beiden Regressionsmethoden sind die Mittelfeldspieler sowie die Angreifer überdurchschnittlich wichtig für ihren Verein. In der Abwehr sind die Ergebnisse unterschiedlich. Philipp Lahm ist der einzige Abwehrspieler in 2010 der laut beider Regressionen über dem Durchschnitt gepunktet hat. Innerhalb der Kombination von Clusteranalyse und Regression haben alle der zehn Feldspieler einen höheren Einfluss auf die Punkte als der Durchschnitt.

In der Saison 2011 werden sechs, der als bestgewählten Feldspielern als Nationalspieler von beiden Klassifizierern eingeordnet. Beide Stürmer haben eine bessere Note als der Durchschnitt. Im Mittelfeld und in der Abwehr gibt es jeweils zwei Spieler, die im Mittel besser abschneiden. Bei der linearen Regression können sechs Spieler besser als der Durchschnitt punkten. Bei der SMOreg Methode sind dies fünf Spieler. Innerhalb der Kombination punkten sieben der zehn Feldspieler im Schnitt besser als die restlichen Spieler.

Die zusammengefassten Resultate zeigen sehr unterschiedliche Ergebnisse. Geht man davon aus, dass die aufgelisteten zehn Spieler die tatsächlich besten Spieler der Saison sind, ist nicht von jeder Methode auf diese Tatsache zu schließen. In Tabelle 47 und 48 sind die Spieler der Saison 2010 und Saison 2010 aufgelistet, welche überdurchschnittliche Ergebnisse auf ihrer Position in den vorgestellten Methoden erreichen und von den beiden Klassifizierern aus Kapitel 5.1 als Nationalspieler eingeordnet werden.

Abschließend ist zu erwähnen, dass eine Beurteilung von zukünftigen Trans-

**Tabelle 45:** Zusammenfassung der Mannschaft der Saison 2010

| Position | Verein | Spieler | NM_MP | NM_RF | Note | LinReg | SMOreg | Kombi |
|---|---|---|---|---|---|---|---|---|
| Abwehr | FC Bayern München | Philipp Lahm | 1 | 1 | -19,09 | 4,13 | 2,88 | 5,66 |
| | Borussia Dortmund | Neven Subotic | 1 | 0 | -29,40 | -3,75 | -1,49 | 5,81 |
| | Borussia Dortmund | Mats Hummels | 1 | 1 | -27,47 | -2,18 | 1,89 | 7,65 |
| | Borussia Dortmund | Marcel Schmelzer | 0 | 0 | -17,28 | 1,78 | -0,47 | 5,20 |
| Durchschnitt | | | | | -20,36 | 0,60 | -0,09 | 3,50 |
| Mittelfed | Borussia Dortmund | Mario Götze | 0 | 0 | 22,05 | 11,02 | 3,55 | 16,26 |
| | Bayer 04 Leverkusen | Arturo Vidal | 1 | 1 | 8,4 | 11,61 | 8,39 | 10,03 |
| | Borussia Dortmund | Nuri Sahin | 1 | 1 | 10,21 | 13,51 | 5,85 | 9,61 |
| | FC Bayern München | Arjen Robben | 0 | 0 | 14,25 | 4,17 | 2,79 | 7,25 |
| Durchschnitt | | | | | 17,83 | 0,38 | 0,01 | 5,52 |
| Angriff | FC Bayern München | Mario Gomez | 1 | 1 | 18,54 | 2,09 | 4,46 | 13,53 |
| | SC Freiburg | Papiss Cisse | 0 | 0 | 22,68 | 2,32 | 10,21 | 19,63 |
| Durchschnitt | | | | | 28,00 | -1,07 | 1,07 | 6,66 |

NM_MP: Spieler wurde als Nationalspieler durch die MultilayerPercepton Methode klassifiziert
NM_RF: Spieler wurde als Nationalspieler durch die RandomForest Methode klassifiziert
Note: Note die durch die lineare Regression berechnet wird (Je niedriger desto besser)
LinReg: Erspielte Punkte anhand der linearen Regression berechnet
SMOreg: Erspielte Punkte anhand der SMOreg Methode berechnet
Kombi: Erspielte Punkte anhand der Kombination aus Clustering und linearer Regression berechnet

**Tabelle 46:** Zusammenfassung der Mannschaft der Saison 2011

| Position | Verein | Spieler | NM_MP | NM_RF | Note | LinReg | SMOreg | Kombi |
|---|---|---|---|---|---|---|---|---|
| Abwehr | Borussia Mönchengladbach | Dante | 0 | 0 | -26,61 | -2,70 | -1,84 | 2,97 |
| | FC Bayern München | Philipp Lahm | 1 | 1 | -15,13 | 2,52 | -0,36 | 2,75 |
| | Borussia Dortmund | Mats Hummels | 1 | 0 | -30,15 | -1,30 | -1,39 | 5,38 |
| | Borussia Dortmund | Lukasz Piszczek | 0 | 0 | -20,11 | 7,31 | 2,82 | 7,00 |
| Durchschnitt | | | | | -20,36 | 0,74 | 0,06 | 3,23 |
| Mittelfed | Borussia Mönchengladbach | Marco Reus | 1 | 1 | 22,22 | 5,57 | 6,57 | 17,78 |
| | FC Bayern München | Franck Ribery | 1 | 1 | 16,55 | 8,94 | 5,56 | 11,93 |
| | FC Bayern München | David Alaba | 1 | 1 | 14,338 | -2,23 | -2,29 | 4,29 |
| | Borussia Dortmund | Shinji Kagawa | 0 | 1 | 27,34 | 3,00 | 2,26 | 15,45 |
| Durchschnitt | | | | | 17,71 | 0,40 | 0,18 | 6,13 |
| Angriff | FC Schalke 04 | Klaas Jan Huntelaar | 0 | 0 | 17,79 | 4,95 | 8,70 | 19,62 |
| | FC Bayern München | Mario Gomez | 1 | 1 | 22,91 | -1,26 | 0,53 | 14,57 |
| Durchschnitt | | | | | 27,35 | -0,74 | 1,03 | 6,37 |

NM_MP: Spieler wurde als Nationalspieler durch die MultilayerPercepton Methode klassifiziert
NM_RF: Spieler wurde als Nationalspieler durch die RandomForest Methode klassifiziert
Note: Note die durch die lineare Regression berechnet wird (Je niedriger desto besser)
LinReg: Erspielte Punkte anhand der linearen Regression berechnet
SMOreg: Erspielte Punkte anhand der SMOreg Methode berechnet
Kombi: Erspielte Punkte anhand der Kombination aus Clustering und linearer Regression berechnet

**Tabelle 47:** Überdurchschnittliche Spieler der Saison 2010

| Position | Verein | Spieler | NM_MP | NM_RF | Note | LinReg | SMOreg | Kombi |
|---|---|---|---|---|---|---|---|---|
| Abwehr | VfL Wolfsburg | Sascha Riether | 1 | 1 | -21,21 | 2,71 | 0,85 | 4,17 |
| Mittelfeld | Bayer 04 Leverkusen | Arturo Vidal | 1 | 1 | 8,42 | 11,61 | 8,39 | 10,03 |
| Mittelfeld | 1. FC Kaiserslautern | Jan Moravek | 1 | 1 | 16,76 | 0,79 | 2,02 | 5,91 |
| Mittelfeld | Borussia Dortmund | Nuri Sahin | 1 | 1 | 10,21 | 13,51 | 5,85 | 9,61 |
| Mittelfeld | Bayer 04 Leverkusen | Renato Augusto | 1 | 1 | 15,62 | 1,42 | 2,31 | 10,80 |
| Mittelfeld | Bayer 04 Leverkusen | Simon Rolfes | 1 | 1 | 10,83 | 0,40 | 0,02 | 7,03 |
| Mittelfeld | FC Bayern München | Thomas Müller | 1 | 1 | 15,83 | 5,61 | 2,16 | 17,89 |
| Mittelfeld | VfB Stuttgart | Zdravko Kuzmanovic | 1 | 1 | 14,52 | 2,56 | 4,21 | 8,87 |
| Angriff | 1. FC St. Pauli | Gerald Asamoah | 1 | 1 | 27,48 | 1,79 | 1,36 | 9,97 |
| Angriff | 1. FC Köln | Lukas Podolski | 1 | 1 | 25,12 | 4,11 | 5,13 | 13,00 |
| Angriff | FC Bayern München | Mario Gomez | 1 | 1 | 18,54 | 2,09 | 4,46 | 13,53 |
| Angriff | VfB Stuttgart | Martin Harnik | 1 | 1 | 27,44 | 0,48 | 1,34 | 8,98 |
| Angriff | Bayer 04 Leverkusen | Stefan Kießling | 1 | 1 | 23,74 | -0,56 | 1,53 | 7,26 |

NM_MP: Spieler wurde als Nationalspieler durch die MultilayerPercepton Methode klassifiziert
NM_RF: Spieler wurde als Nationalspieler durch die RandomForest Methode klassifiziert
Note: Note die durch die lineare Regression berechnet wird (Je niedriger desto besser)
LinReg: Erspielte Punkte anhand der linearen Regression berechnet
SMOreg: Erspielte Punkte anhand der SMOreg Methode berechnet
Kombi: Erspielte Punkte anhand der Kombination aus Clustering und linearer Regression berechnet

**Tabelle 48:** Überdurchschnittliche Spieler der Saison 2011

| Position | Verein | Spieler | NM_MP | NM_RF | Note | LinReg | SMOreg | Kombi |
|---|---|---|---|---|---|---|---|---|
| Abwehr | Hamburger SV | Heiko Westermann | 1 | 1 | -28,60 | 0,88 | 2,46 | 5,57 |
| Abwehr | Hannover 96 | Karim Haggui | 1 | 1 | -27,87 | 1,41 | 2,43 | 6,78 |
| Mittelfeld | FC Bayern München | Franck Ribery | 1 | 1 | 16,55 | 8,94 | 5,56 | 11,93 |
| Mittelfeld | Bayer 04 Leverkusen | Lars Bender | 1 | 1 | 10,04 | 2,39 | 0,91 | 8,82 |
| Mittelfeld | 1899 Hoffenheim | Sejad Salihovic | 1 | 1 | 16,46 | 3,71 | 4,97 | 9,33 |
| Angriff | 1. FC Köln | Lukas Podolski | 1 | 1 | 24,59 | 3,68 | 5,68 | 12,42 |
| Angriff | FC Schalke 04 | Raul | 1 | 1 | 21,11 | -0,12 | 4,11 | 10,08 |
| Angriff | Bayer 04 Leverkusen | Stefan Kießling | 1 | 1 | 21,31 | 2,79 | 4,87 | 13,00 |

NM_MP: Spieler wurde als Nationalspieler durch die MultilayerPercepton Methode klassifiziert
NM_RF: Spieler wurde als Nationalspieler durch die RandomForest Methode klassifiziert
Note: Note die durch die lineare Regression berechnet wird (Je niedriger desto besser)
LinReg: Erspielte Punkte anhand der linearen Regression berechnet
SMOreg: Erspielte Punkte anhand der SMOreg Methode berechnet
Kombi: Erspielte Punkte anhand der Kombination aus Clustering und linearer Regression berechnet

fers mithilfe der beiden Ansätze nur bedingt zum Einsatz kommen kann. Da in der Zukunft viele Unwägbarkeiten vorherrschen, wie die Verletzung des Spielers, überraschender Formverlust durch äußerliche Einflüsse oder andere Dinge, die auf die Leistung eines Spielers einwirken, kann nicht von der erspielten Punktzahl auf zukünftige Leistungen geschlossen werden. Eine Auswahl an Spielern, wie in Tabelle 47 und Tabelle 48 kann jedoch helfen, eine Vorauswahl von Spielern zu ermitteln und diese mithilfe menschlicher Scouts weiter zu beobachten.

Vergangene Transfers können durch die beiden Ansätze jedoch auf ihre Qualität überprüft werden. Die ermittelten Werte können zudem bei Verhandlungen über die Ablösesumme oder Vertragsgesprächen genutzt werden, um die Höhe der Ablöse bzw. des Gehalts zu berechnen. Die Analysen mit erweitertem Datensatz können zudem noch detaillierte und bessere Ergebnisse liefern.

Die Quantifizierung der Wichtigkeit eines Spielers durch die von ihm erspielten Punkte bietet eine leichte Darstellung des Wertes für die Verantwortlichen eines Vereins an. Die beiden vorgestellten Ansätze sind dabei zwei Arten der Analyse die möglich sind. Erweiterte Ansätze sind denkbar. So können einerseits andere Verfahren gewählt und kombiniert werden. Andererseits ist eine veränderte Repräsentation der Mannschaften möglich.

# 7 Die Zukunft des Data Mining im Fußball

Das Ziel der vorliegenden Arbeit bestand darin, mittels komplexer Datenanalyse, nämlich dem Data Mining, den Wert eines Spielers für einen Verein zu quantifizieren. Zu diesem Zweck wurden mehrere Data Mining Verfahren exemplarisch auf die spielspezifischen Daten der Bundesligaspieler der beiden Saisons 2010/2011 und 2011/2012 angewendet. Dies diente der Überprüfung, inwiefern Data Mining im Sport und speziell im Fußball eingesetzt werden kann. Der Fokus lag dabei auf der Bemessung der Wichtigkeit eines Spielers für seinen Verein.

Im Laufe der Arbeit wurden Data Mining Verfahren angewendet, um Nationalspieler zu klassifizieren, die Beziehungen zwischen der Spielweise und der erhaltenen Note eines Spielers zu identifizieren, Spieler in homogene Gruppen einzuordnen und eine zukünftige Saison vorherzusagen. Um die Eingangsfrage zu beantworten wurden zusätzlich zwei Data Mining Verfahren kombiniert.

Die Klassifikation von Nationalspielern diente der Entdeckung von Spielern, welche zwar auf dem Niveau eines Nationalspielers spielen, jedoch von den Nationaltrainern nicht berücksichtigt werden. Es konnten dabei Klassifikationsraten von rund 70 % erreicht werden. Die Precision und damit das Qualitätsmerkmal der richtigen Klassifizierung von Nationalspielern, ist jedoch nicht hoch genug, um in dem gezeigten Beispiel von einer zuverlässigen Klassifizierung von Nationalspielern auszugehen. Die Idee der Durchführung einer solchen Klassifikation bietet jedoch eine realistische Möglichkeit zum Einsatz in der Realität. Die Verwendung mehrerer Daten über zahlreichere Saisons könnten die Ergebnisse des aufgezeigten Ansatzes positiv beeinflussen. Außerdem könnte die Idee aus anderen Blickwinkeln betrachtet werden. In dieser Arbeit werden Nationalspieler als Talente bzw. gute Spieler definiert. Andere Betrachtungen sind hier jedoch möglich. So könnten Spieler einer Liga beispielsweise von Trainern eines Vereins in einer der beiden Kategorien „gut" oder „schlecht" eingeteilt werden. Aufgrund dieser Einteilung könnten dann Modelle erlernt werden und auf ungesehene Daten von Spielern angewendet werden.

Die Identifikation der Beziehungen zwischen den dargestellten Attributen eines Spielers und der Note, welche die Spieler von Redakteuren mehrerer Zeitschriften oder Internetseiten erhalten, sollten klären, welche Aktionen eines Spielers zu einer guten Beurteilung seiner Leistung führen. Die Analysen ergeben, dass vor allem die offensiven Aktionen eines Spielers die Beurteilung positiv beeinflussen. Die Qualität der Ergebnisse ist kritisch zu betrachten. Eine Analyse mit erweiterter Datenbasis könnte zu besseren Resultaten führen. Auch hier besteht die Möglichkeit weitere Forschungen anzuschließen. So können andere Werte als die Note als Zielwert für die Regression herhalten. Insgesamt lässt sich sagen, dass die Regression eine einfache Möglichkeit bietet, die Zusammenhänge von diversen Eingangsgrößen zur Ausgangsgröße zu ermitteln. Somit ist eine Anwendung dieses Verfahrens im Sport in Zukunft sehr realistisch, was auch die Veröffentlichungen im Bereich der komplexen Datenanalyse zeigen.

Die Anwendung der Clusteranalyse in dieser Arbeit beweist ihre Anwendungsmöglichkeit im realen Einsatz. Es können nicht nur einzelne Mannschaften in sinnvolle Kategorien eingeteilt werden, vor allem die Einordnung der Spieler in die verschiedenen Spielertypen kann Verantwortlichen von Vereinen bei der Zusammensetzung der Mannschaft helfen. So können gleichartige Spieler, aber auch unterschiedlich agierende Spieler innerhalb der Liga identifiziert werden. Das Wissen über die Spielweisen einzelner Spieler kann zur Kaderplanung genutzt werden. Dadurch können Spieler im Sinne ihrer Spielweise gleichwertig ersetzt werden bzw. eine möglichst heterogener Kader zusammengestellt werden.

Eine Vorhersage einer Saison mithilfe von Data Mining muss kritisch betrachtet werden. Eine solche Vorhersage, genauso wie die Prognose einzelner Spielausgängen ist immer schwierig, da die Leistung einer Mannschaft von vielen Dingen abhängt. So können beispielsweise Verletzungen von Spielern, die Schiedsrichterleistung oder äußerliche Einflüsse, wie das Zuschaueraufkommen oder die Wetterbedingungen die Ausgänge von Spielen beeinflussen, sind jedoch in einer Data Mining Anwendung schwer abbildbar. Auch die in dieser Arbeit dargestellten Ergebnisse der Prognose sind nicht zufriedenstellend und für einen Einsatz in der Realität nicht brauchbar.

Die Kernfrage dieser Arbeit, wie man den Wert eines Spielers für seinen Verein im Sinne der fußballerischen Leistung misst, wird mithilfe der Regression separat sowie einer Kombination aus Clustering und Regression beantwortet. Dabei werden bei der isolierten Verwendung der Regressionsanalyse einerseits die ermittelten Punkte des Verfahrens mit einem gewissen Spieler errechnet, sowie die Punkte die laut der Analyse ohne den Spieler erreicht worden wären. Die Differenz der beiden Punktevorhersagen entspricht den erspielten Punkten eines einzelnen Spielers, die er laut Regression zur Gesamtpunktzahl des Vereins beiträgt. Diese reine Anwendung der Regression basierte auf den Ergebnissen der Vorhersage, welche im vorangegangenen Kapitel unternommen wurde. Wie erwähnt, sind die Ergebnisse der Vorhersage nicht zufriedenstellend. Demnach muss diese Art der Messung für die Wichtigkeit eines Spieler ebenfalls kritisch betrachtet werden.

Bei der Kombination aus Cluster- und Regressionsanalyse, werden nicht die einzelnen Spieler in den Mittelpunkt gestellt, sondern ihre Art zu spielen. Hier wird mithilfe der Regression gelernt, welche Cluster und damit welche Spielweisen am ehesten die erhaltenen Punkte des Vereins beeinflussen. Dabei wird erkannt, dass gewisse Spielweisen mehr zur Leistung beitragen als andere. Der Ansatz lässt es außerdem zu, die Punkte jeden Spielers zu errechnen, welche er zur Gesamtleistung der Mannschaft beigetragen hat. Anhand der aufzeigten Beispiele, bei denen die ermittelten Punkte gewisser Spieler mit der subjektiven Wahrnehmung der Leistung dieser Spieler verglichen werden, zeigt sich, dass sich der aufgeführte Ansatz zur Bemessung von Spielern eignet. Es muss jedoch gesagt sein, dass nur wenige Daten für die Analysen zur Verfügung stehen. Weitere Untersuchungen mit mehreren Attributen sowie Daten über mehrere Saisons könnten weitere Erkenntnisse offenbaren und die Qualität einer solchen Analyse detaillierter überprüft werden.

Beide Ansätze der Quantifizierung der Wichtigkeit von Spielern zeigen auf, dass eine Analyse mittels Data Mining möglich ist. Durch die begrenzte Anzahl an Daten, können jedoch keine verlässlichen Aussagen über die Qualität der hier durchgeführten Anwendungen gemacht werden. In weiteren Untersuchungen, könnten die Ansätze mit erweiterten Datensätzen überprüft werden. Auch eine Veränderung der Bedingungen innerhalb der Durchführung, wie beispielsweise ein anderer Aufbau der Datensätze oder eine anderer Auswahl von Data Mining Methoden können als weitere Forschungen angedacht werden.

Die Quantifizierung des Wertes durch die erspielten Punkte eines einzelnen Spielers, kann von Verantwortlichen genutzt werden, um vergangene Transfers zu beurteilen und kann als Anhaltspunkt über Ablösesummen oder Gehälter bei Vertragsgesprächen dienen. Dafür kann aufbauend auf den Ergebnissen ein Informationssystem erstellt werden, dass eine einfache Betrachtung der Ergebnisse sowie ein Vergleich von Spielern für Benutzer ermöglicht.

Das Potential zukünftiger Transfers zu Beurteilen ist als kritisch zu betrachten. Es kann nicht darauf geschlossen werden, dass ein Spieler den Wert, den er in einer vergangenen Saison erreicht hat auch in zukünftigen Saisons halten oder verbessern kann. Wie auch bei der Vorhersage von Spielen oder einer ganzen Saison, liegen hier

viele Unwägbarkeiten vor, wie beispielsweise die Verletzung eines Spielers oder andere Einflüsse, welche die Form eines Spielers beeinflussen. Einen Anhaltspunkt über den Wert eines Spielers können die Ergebnisse für die Verantwortlichen eines Vereins trotzdem bieten.

Zusätzlich zur Bemessung der Wichtigkeit der Spieler, können die Ergebnisse aus den aufgezeigten Analysen der Nationalspieler, der Noten sowie der Clusteranalyse betrachtet werden und Verantwortlichen als Entscheidungshilfe dienen.

Wie anhand des Einblickes in den derzeitigen Forschungsstand sowie der in dieser Arbeit durchgeführten Analysen zu sehen ist, gibt es vielfältige Möglichkeiten der Anwendung von Data Mining im Sport und Fußball. Durch die erweiterte und automatisierte Aufnahme von Daten der Bundesliga, wird in Zukunft ein höheres und gehaltvolleres Datenvolumen zur Verfügung stehen und damit einhergehend wird die Nachfrage der manuellen, aber im Besonderen auch die der automatisierten Analysen der Daten steigen. Die Anwendungen von Data Mining in dieser Arbeit fokussieren sich primär auf die Spieler einer Mannschaft. Die automatisierte Analyse bietet jedoch in den vielfältigen Aufgabenbereichen eines Vereins, wie dem Training, der Jugendförderung oder der Spielanalyse sinnvolle Anwendungsmöglichkeiten. In Zukunft kann somit die komplexe Datenanalyse, wie sie Data Mining darstellt, in vielen Bereichen des Fußballs Verwendung finden und Vereinen als Entscheidungshilfe dienen.

# Anhang

**Abbildung 23:** SQL SpielerProSaison

```
SELECT
fakten.saison, fakten.verein, fakten.spieler,
CAST(SUM(fakten.zweikampf*gespielt)/SUM(fakten.gespielt) AS
    DECIMAL(18,9)) AS zweikampf,
CAST(SUM(fakten.verlZweikampf*gespielt)/SUM(fakten.gespielt)
    AS DECIMAL(18,9)) AS verlZweikampf,
CAST(SUM(fakten.Ballkontakte)/SUM(fakten.gespielt) AS DECIMAL
    (18,9)) AS Ballkontakte,
CAST(SUM(fakten.erfPaesse)/SUM(fakten.gespielt) AS DECIMAL
    (18,9)) AS erfPaesse,
CAST(SUM(fakten.fehlPaesse)/SUM(fakten.gespielt) AS DECIMAL
    (18,9)) AS fehlPaesse,
CAST(SUM(fakten.Paesse)/SUM(fakten.gespielt) AS DECIMAL(18,9))
     AS Paesse,
CAST(SUM(fakten.erfPassquote*gespielt)/SUM(fakten.gespielt) AS
     DECIMAL(18,9)) AS erfPassquote,
CAST(SUM(fakten.fehlPassquote*gespielt)/SUM(fakten.gespielt)
    AS DECIMAL(18,9)) AS fehlPassquote,
CAST(SUM(fakten.Torschuesse)/SUM(fakten.gespielt) AS DECIMAL
    (18,9)) AS Torschuesse,
CAST(SUM(fakten.vergTorschuesse)/SUM(fakten.gespielt) AS
    DECIMAL(18,9)) AS vergTorschuesse,
CAST(SUM(fakten.erfTorschussquote*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS erfTorschussquote,
CAST(SUM(fakten.fehlTorschussquote*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS fehlTorschussquote,
CAST(SUM(fakten.Torschussvorlagen)/SUM(fakten.gespielt) AS
    DECIMAL(18,9)) AS Torschussvorlagen,
CAST(SUM(fakten.Vorlage)/SUM(fakten.gespielt) AS DECIMAL(18,9)
    ) AS Vorlage,
CAST(SUM(fakten.Tore)/SUM(fakten.gespielt) AS DECIMAL(18,9))
    AS Tore,
CAST(SUM(fakten.Gelb)/SUM(fakten.gespielt) AS DECIMAL(18,9))
    AS Gelb,
CAST(SUM(fakten.PassProBallkontakt*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS PassProBallkontakt,
CAST(SUM(fakten.erfPassProBallkontakt*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS erfPassProBallkontakt,
CAST(SUM(fakten.TorschussvorlageProPass*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS TorschussvorlageProPass,
(CASE WHEN (nationalspieler.saison2009 + nationalspieler.
    saison2010) > 1 THEN 1 ELSE 0 END) AS nm1,
(CASE WHEN (nationalspieler.saison2011) > 1 THEN 1 ELSE 0 END)
     AS nm,
spielereigenschaften.spielerposition
FROM fakten, spielereigenschaften, nationalspieler
WHERE spielereigenschaften.saison = fakten.saison AND
    spielereigenschaften.spieler = fakten.spieler
```

```
AND nationalspieler.spieler = fakten.spieler AND fakten.saison
    = 2010
GROUP BY fakten.saison, fakten.spieler
UNION
SELECT fakten.saison, fakten.verein, fakten.spieler,
CAST(SUM(fakten.zweikampf*gespielt)/SUM(fakten.gespielt) AS
   DECIMAL(18,9)) AS zweikampf,
CAST(SUM(fakten.verlZweikampf*gespielt)/SUM(fakten.gespielt)
   AS DECIMAL(18,9)) AS verlZweikampf,
CAST(SUM(fakten.Ballkontakte)/SUM(fakten.gespielt) AS DECIMAL
   (18,9)) AS Ballkontakte,
CAST(SUM(fakten.erfPaesse)/SUM(fakten.gespielt) AS DECIMAL
   (18,9)) AS erfPaesse,
CAST(SUM(fakten.fehlPaesse)/SUM(fakten.gespielt) AS DECIMAL
   (18,9)) AS fehlPaesse,
CAST(SUM(fakten.Paesse)/SUM(fakten.gespielt) AS DECIMAL(18,9))
    AS Paesse,
CAST(SUM(fakten.erfPassquote*gespielt)/SUM(fakten.gespielt) AS
    DECIMAL(18,9)) AS erfPassquote,
CAST(SUM(fakten.fehlPassquote*gespielt)/SUM(fakten.gespielt)
   AS DECIMAL(18,9)) AS fehlPassquote,
CAST(SUM(fakten.Torschuesse)/SUM(fakten.gespielt) AS DECIMAL
   (18,9)) AS Torschuesse,
CAST(SUM(fakten.vergTorschuesse)/SUM(fakten.gespielt) AS
   DECIMAL(18,9)) AS vergTorschuesse,
CAST(SUM(fakten.erfTorschussquote*gespielt)/SUM(fakten.
   gespielt) AS DECIMAL(18,9)) AS erfTorschussquote,
CAST(SUM(fakten.fehlTorschussquote*gespielt)/SUM(fakten.
   gespielt) AS DECIMAL(18,9)) AS fehlTorschussquote,
CAST(SUM(fakten.Torschussvorlagen)/SUM(fakten.gespielt) AS
   DECIMAL(18,9)) AS Torschussvorlagen,
CAST(SUM(fakten.Vorlage)/SUM(fakten.gespielt) AS DECIMAL(18,9)
   ) AS Vorlage,
CAST(SUM(fakten.Tore)/SUM(fakten.gespielt) AS DECIMAL(18,9))
   AS Tore,
CAST(SUM(fakten.Gelb)/SUM(fakten.gespielt) AS DECIMAL(18,9))
   AS Gelb,
CAST(SUM(fakten.PassProBallkontakt*gespielt)/SUM(fakten.
   gespielt) AS DECIMAL(18,9)) AS PassProBallkontakt,
CAST(SUM(fakten.erfPassProBallkontakt*gespielt)/SUM(fakten.
   gespielt) AS DECIMAL(18,9)) AS erfPassProBallkontakt,
CAST(SUM(fakten.TorschussvorlageProPass*gespielt)/SUM(fakten.
   gespielt) AS DECIMAL(18,9)) AS TorschussvorlageProPass,
(CASE WHEN (nationalspieler.saison2010 + nationalspieler.
   saison2011) > 1 THEN 1 ELSE 0 END) AS nm1,
(CASE WHEN (nationalspieler.saison2012) > 1 THEN 1 ELSE 0 END)
    AS nm,
spielereigenschaften.spielerposition
FROM fakten, spielereigenschaften, nationalspieler
WHERE spielereigenschaften.saison = fakten.saison AND
   spielereigenschaften.spieler = fakten.spieler
```

```
AND nationalspieler.spieler = fakten.spieler AND fakten.saison
    = 2011
GROUP BY fakten.saison, fakten.spieler
```

**Abbildung 24:** SQL Spielernoten

```
SELECT
fakten.saison, fakten.spieltag, fakten.verein, fakten.spieler,
fakten.Zweikampf,
fakten.verlZweikampf,
fakten.Ballkontakte,
fakten.Paesse,
fakten.erfPaesse,
fakten.erfPassquote,
fakten.fehlPaesse,
fakten.fehlPassquote,
fakten.Torschuesse,
fakten.erfTorschussquote,
fakten.vergTorschuesse,
fakten.fehlTorschussquote,
fakten.Torschussvorlagen,
fakten.Tore,
fakten.Vorlage,
fakten.Gelb,
fakten.PassProBallkontakt,
fakten.erfPassProBallkontakt,
fakten.TorschussvorlageProPass,
spielereigenschaften.spielerposition,
noten.note
FROM
fakten JOIN noten
ON fakten.saison = noten.saison AND fakten.spieltag = noten.
    spieltag AND fakten.spieler = noten.spieler
JOIN spielereigenschaften
ON spielereigenschaften.spieler = fakten.spieler AND
    spielereigenschaften.saison = fakten.saison
```

**Abbildung 25:** SQL Verein pro Runde

```
SELECT fakten.saison, 'Hinrunde' AS Runde, fakten.verein,
CAST(SUM(fakten.zweikampf)/COUNT(*) AS DECIMAL(18,9)) AS
    zweikampf,
CAST(SUM(fakten.verlZweikampf)/COUNT(*) AS DECIMAL(18,9)) AS
    verlZweikampf,
CAST(SUM(fakten.Ballkontakte) AS DECIMAL(18,0)) AS
    Ballkontakte,
CAST(SUM(fakten.erfPaesse) AS DECIMAL(18,0)) AS erfPaesse,
CAST(SUM(fakten.fehlPaesse) AS DECIMAL(18,0)) AS fehlPaesse,
CAST(SUM(fakten.Paesse) AS DECIMAL(18,0)) AS Paesse,
CAST(SUM(fakten.erfPassquote)/COUNT(*) AS DECIMAL(18,9)) AS
    erfPassquote,
CAST(SUM(fakten.fehlPassquote)/COUNT(*) AS DECIMAL(18,9)) AS
```

```
    fehlPassquote,
CAST(SUM(fakten.Torschuesse) AS DECIMAL(18,0)) AS Torschuesse,
CAST(SUM(fakten.vergTorschuesse) AS DECIMAL(18,0)) AS
    vergTorschuesse,
CAST(SUM(fakten.erfTorschussquote)/COUNT(*) AS DECIMAL(18,9))
    AS erfTorschussquote,
CAST(SUM(fakten.fehlTorschussquote)/COUNT(*) AS DECIMAL(18,9))
     AS fehlTorschussquote,
CAST(SUM(fakten.Torschussvorlagen) AS DECIMAL(18,0)) AS
    Torschussvorlagen,
CAST(SUM(fakten.Vorlage) AS DECIMAL(18,0)) AS Vorlage,
CAST(SUM(fakten.Tore) AS DECIMAL(18,0)) AS Tore,
CAST(SUM(fakten.Gelb) AS DECIMAL(18,0)) AS Gelb,
CAST(SUM(fakten.PassProBallkontakt)/COUNT(*) AS DECIMAL(18,9))
     AS PassProBallkontakt,
CAST(SUM(fakten.erfPassProBallkontakt)/COUNT(*) AS DECIMAL
    (18,9)) AS erfPassProBallkontakt,
CAST(SUM(fakten.TorschussvorlageProPass)/COUNT(*) AS DECIMAL
    (18,9)) AS TorschussvorlageProPass
FROM fakten WHERE fakten.spieltag > 0 AND fakten.spieltag < 18
      AND fakten.spieler IN (SELECT spieler FROM
    spielereigenschaften WHERE spielerposition IN ('Abwehr', '
    Mittelfeld','Angriff'))
GROUP BY fakten.saison, fakten.verein
union
SELECT fakten.saison, 'Rueckrunde' AS Runde, fakten.verein,
CAST(SUM(fakten.zweikampf)/COUNT(*) AS DECIMAL(18,9)) AS
    zweikampf,
CAST(SUM(fakten.verlZweikampf)/COUNT(*) AS DECIMAL(18,9)) AS
    verlZweikampf,
CAST(SUM(fakten.Ballkontakte) AS DECIMAL(18,0)) AS
    Ballkontakte,
CAST(SUM(fakten.erfPaesse) AS DECIMAL(18,0)) AS erfPaesse,
CAST(SUM(fakten.fehlPaesse) AS DECIMAL(18,0)) AS fehlPaesse,
CAST(SUM(fakten.Paesse) AS DECIMAL(18,0)) AS Paesse,
CAST(SUM(fakten.erfPassquote)/COUNT(*) AS DECIMAL(18,9)) AS
    erfPassquote,
CAST(SUM(fakten.fehlPassquote)/COUNT(*) AS DECIMAL(18,9)) AS
    fehlPassquote,
CAST(SUM(fakten.Torschuesse) AS DECIMAL(18,0)) AS Torschuesse,
CAST(SUM(fakten.vergTorschuesse) AS DECIMAL(18,0)) AS
    vergTorschuesse,
CAST(SUM(fakten.erfTorschussquote)/COUNT(*) AS DECIMAL(18,9))
    AS erfTorschussquote,
CAST(SUM(fakten.fehlTorschussquote)/COUNT(*) AS DECIMAL(18,9))
     AS fehlTorschussquote,
CAST(SUM(fakten.Torschussvorlagen) AS DECIMAL(18,0)) AS
    Torschussvorlagen,
CAST(SUM(fakten.Vorlage) AS DECIMAL(18,0)) AS Vorlage,
CAST(SUM(fakten.Tore) AS DECIMAL(18,0)) AS Tore,
CAST(SUM(fakten.Gelb) AS DECIMAL(18,0)) AS Gelb,
```

```
CAST(SUM(fakten.PassProBallkontakt)/COUNT(*) AS DECIMAL(18,9))
    AS PassProBallkontakt,
CAST(SUM(fakten.erfPassProBallkontakt)/COUNT(*) AS DECIMAL
    (18,9)) AS erfPassProBallkontakt,
CAST(SUM(fakten.TorschussvorlageProPass)/COUNT(*) AS DECIMAL
    (18,9)) AS TorschussvorlageProPass
FROM fakten WHERE fakten.spieltag > 17 AND fakten.spieltag <
    35 AND fakten.spieler IN (SELECT spieler FROM
    spielereigenschaften WHERE spielerposition IN ('Abwehr', '
    Mittelfeld','Angriff'))
GROUP BY fakten.saison, fakten.verein;
```

**Abbildung 26:** SQL Spieler pro Runde

```
SELECT fakten.saison, 'Hinrunde' AS Runde, fakten.verein,
    fakten.spieler,
CAST(SUM(fakten.zweikampf*gespielt)/SUM(fakten.gespielt) AS
    DECIMAL(18,9)) AS zweikampf,
CAST(SUM(fakten.verlZweikampf*gespielt)/SUM(fakten.gespielt)
    AS DECIMAL(18,9)) AS verlZweikampf,
CAST((SUM(fakten.Ballkontakte)/SUM(fakten.gespielt))*90 AS
    DECIMAL(18,9)) AS Ballkontakte,
CAST((SUM(fakten.erfPaesse)/SUM(fakten.gespielt))*90 AS
    DECIMAL(18,9)) AS erfPaesse,
CAST((SUM(fakten.fehlPaesse)/SUM(fakten.gespielt))*90 AS
    DECIMAL(18,9)) AS fehlPaesse,
CAST((SUM(fakten.Paesse)/SUM(fakten.gespielt))*90 AS DECIMAL
    (18,9)) AS Paesse,
CAST(SUM(fakten.erfPassquote*gespielt)/SUM(fakten.gespielt) AS
    DECIMAL(18,9)) AS erfPassquote,
CAST(SUM(fakten.fehlPassquote*gespielt)/SUM(fakten.gespielt)
    AS DECIMAL(18,9)) AS fehlPassquote,
CAST((SUM(fakten.Torschuesse)/SUM(fakten.gespielt))*90 AS
    DECIMAL(18,9)) AS Torschuesse,
CAST((SUM(fakten.vergTorschuesse)/SUM(fakten.gespielt))*90 AS
    DECIMAL(18,9)) AS vergTorschuesse,
CAST(SUM(fakten.erfTorschussquote*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS erfTorschussquote,
CAST(SUM(fakten.fehlTorschussquote*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS fehlTorschussquote,
CAST((SUM(fakten.Torschussvorlagen)/SUM(fakten.gespielt))*90
    AS DECIMAL(18,9)) AS Torschussvorlagen,
CAST((SUM(fakten.Vorlage)/SUM(fakten.gespielt))*90 AS DECIMAL
    (18,9)) AS Vorlage,
CAST((SUM(fakten.Tore)/SUM(fakten.gespielt))*90 AS DECIMAL
    (18,9)) AS Tore,
CAST((SUM(fakten.Gelb)/SUM(fakten.gespielt))*90 AS DECIMAL
    (18,9)) AS Gelb,
CAST(SUM(fakten.PassProBallkontakt*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS PassProBallkontakt,
CAST(SUM(fakten.erfPassProBallkontakt*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS erfPassProBallkontakt,
```

```
CAST(SUM(fakten.TorschussvorlageProPass*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS TorschussvorlageProPass,
spielereigenschaften.spielerposition
FROM fakten, spielereigenschaften WHERE fakten.saison =
    spielereigenschaften.saison AND fakten.spieler =
    spielereigenschaften.spieler AND fakten.spieltag > 0 AND
    fakten.spieltag < 18  AND fakten.spieler IN (SELECT spieler
     FROM spielereigenschaften WHERE spielerposition IN ('
    Abwehr','Mittelfeld','Angriff'))
GROUP BY fakten.saison, fakten.spieler
HAVING SUM(fakten.gespielt) >= 510
union
SELECT fakten.saison, 'Rueckrunde' AS Runde, fakten.verein,
    fakten.spieler,
CAST(SUM(fakten.zweikampf*gespielt)/SUM(fakten.gespielt) AS
    DECIMAL(18,9)) AS zweikampf,
CAST(SUM(fakten.verlZweikampf*gespielt)/SUM(fakten.gespielt)
    AS DECIMAL(18,9)) AS verlZweikampf,
CAST((SUM(fakten.Ballkontakte)/SUM(fakten.gespielt))*90 AS
    DECIMAL(18,9)) AS Ballkontakte,
CAST((SUM(fakten.erfPaesse)/SUM(fakten.gespielt))*90  AS
    DECIMAL(18,9)) AS erfPaesse,
CAST((SUM(fakten.fehlPaesse)/SUM(fakten.gespielt))*90  AS
    DECIMAL(18,9)) AS fehlPaesse,
CAST((SUM(fakten.Paesse)/SUM(fakten.gespielt))*90  AS DECIMAL
    (18,9)) AS Paesse,
CAST(SUM(fakten.erfPassquote*gespielt)/SUM(fakten.gespielt) AS
     DECIMAL(18,9)) AS erfPassquote,
CAST(SUM(fakten.fehlPassquote*gespielt)/SUM(fakten.gespielt)
    AS DECIMAL(18,9)) AS fehlPassquote,
CAST((SUM(fakten.Torschuesse)/SUM(fakten.gespielt))*90  AS
    DECIMAL(18,9)) AS Torschuesse,
CAST((SUM(fakten.vergTorschuesse)/SUM(fakten.gespielt))*90  AS
     DECIMAL(18,9)) AS vergTorschuesse,
CAST(SUM(fakten.erfTorschussquote*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS erfTorschussquote,
CAST(SUM(fakten.fehlTorschussquote*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS fehlTorschussquote,
CAST((SUM(fakten.Torschussvorlagen)/SUM(fakten.gespielt))*90
    AS DECIMAL(18,9)) AS Torschussvorlagen,
CAST((SUM(fakten.Vorlage)/SUM(fakten.gespielt))*90  AS DECIMAL
    (18,9)) AS Vorlage,
CAST((SUM(fakten.Tore)/SUM(fakten.gespielt))*90  AS DECIMAL
    (18,9)) AS Tore,
CAST((SUM(fakten.Gelb)/SUM(fakten.gespielt))*90  AS DECIMAL
    (18,9)) AS Gelb,
CAST(SUM(fakten.PassProBallkontakt*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS PassProBallkontakt,
CAST(SUM(fakten.erfPassProBallkontakt*gespielt)/SUM(fakten.
    gespielt) AS DECIMAL(18,9)) AS erfPassProBallkontakt,
CAST(SUM(fakten.TorschussvorlageProPass*gespielt)/SUM(fakten.
```

```
    gespielt) AS DECIMAL(18,9)) AS TorschussvorlageProPass,
spielereigenschaften.spielerposition
FROM fakten, spielereigenschaften WHERE fakten.saison =
    spielereigenschaften.saison AND fakten.spieler =
    spielereigenschaften.spieler AND fakten.spieltag > 17 AND
    fakten.spieltag < 35  AND fakten.spieler IN (SELECT spieler
     FROM spielereigenschaften WHERE spielerposition IN ('
    Abwehr','Mittelfeld','Angriff'))
GROUP BY fakten.saison, fakten.spieler
HAVING SUM(fakten.gespielt) >= 510
```

**Abbildung 27:** SQL Verein pro Runde gewichtet

```
SELECT saison, 'Hinrunde' AS Runde, verein,
CAST(SUM(zweikampf*gespielt)/SUM(gespielt) AS DECIMAL(18,9))
    AS zweikampf,
CAST(SUM(verlZweikampf*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS verlZweikampf,
CAST(SUM(Ballkontakte*gespielt)/SUM(gespielt) AS DECIMAL(18,9)
    ) AS Ballkontakte,
CAST(SUM(erfPaesse*gespielt)/SUM(gespielt) AS DECIMAL(18,9))
    AS erfPaesse,
CAST(SUM(fehlPaesse*gespielt)/SUM(gespielt) AS DECIMAL(18,9))
    AS fehlPaesse,
CAST(SUM(Paesse*gespielt)/SUM(gespielt) AS DECIMAL(18,9)) AS
    Paesse,
CAST(SUM(erfPassquote*gespielt)/SUM(gespielt) AS DECIMAL(18,9)
    ) AS erfPassquote,
CAST(SUM(fehlPassquote*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS fehlPassquote,
CAST(SUM(Torschuesse*gespielt)/SUM(gespielt) AS DECIMAL(18,9))
     AS Torschuesse,
CAST(SUM(vergTorschuesse*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS vergTorschuesse,
CAST(SUM(erfTorschussquote*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS erfTorschussquote,
CAST(SUM(fehlTorschussquote*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS fehlTorschussquote,
CAST(SUM(Torschussvorlagen*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS Torschussvorlagen,
CAST(SUM(Vorlage*gespielt)/SUM(gespielt) AS DECIMAL(18,9)) AS
    Vorlage,
CAST(SUM(Tore*gespielt)/SUM(gespielt) AS DECIMAL(18,9)) AS
    Tore,
CAST(SUM(Gelb*gespielt)/SUM(gespielt) AS DECIMAL(18,9)) AS
    Gelb,
CAST(SUM(PassProBallkontakt*gespielt)/SUM(gespielt)as DECIMAL
    (18,9)) AS PassProBallkontakt,
CAST(SUM(erfPassProBallkontakt*gespielt)/SUM(gespielt)as
    DECIMAL(18,9)) AS erfPassProBallkontakt,
CAST(SUM(TorschussvorlageProPass*gespielt)/SUM(gespielt)as
    DECIMAL(18,9)) AS TorschussvorlageProPass
```

```
FROM fakten WHERE spieltag > 0 AND spieltag < 18  AND spieler
    IN (SELECT spieler FROM spielereigenschaften WHERE
    spielerposition IN ('Abwehr', 'Mittelfeld','Angriff'))
GROUP BY saison, verein
UNION
SELECT saison, 'Rueckrunde' AS Runde, verein,
CAST(SUM(zweikampf*gespielt)/SUM(gespielt) AS DECIMAL(18,9))
    AS zweikampf,
CAST(SUM(verlZweikampf*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS verlZweikampf,
CAST(SUM(Ballkontakte*gespielt)/SUM(gespielt) AS DECIMAL(18,9)
    ) AS Ballkontakte,
CAST(SUM(erfPaesse*gespielt)/SUM(gespielt) AS DECIMAL(18,9))
    AS erfPaesse,
CAST(SUM(fehlPaesse*gespielt)/SUM(gespielt) AS DECIMAL(18,9))
    AS fehlPaesse,
CAST(SUM(Paesse*gespielt)/SUM(gespielt) AS DECIMAL(18,9)) AS
    Paesse,
CAST(SUM(erfPassquote*gespielt)/SUM(gespielt) AS DECIMAL(18,9)
    ) AS erfPassquote,
CAST(SUM(fehlPassquote*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS fehlPassquote,
CAST(SUM(Torschuesse*gespielt)/SUM(gespielt) AS DECIMAL(18,9))
     AS Torschuesse,
CAST(SUM(vergTorschuesse*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS vergTorschuesse,
CAST(SUM(erfTorschussquote*gespielt)/SUM(gespielt)as DECIMAL
    (18,9)) AS erfTorschussquote,
CAST(SUM(fehlTorschussquote*gespielt)/SUM(gespielt)as DECIMAL
    (18,9)) AS fehlTorschussquote,
CAST(SUM(Torschussvorlagen*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS Torschussvorlagen,
CAST(SUM(Vorlage*gespielt)/SUM(gespielt) AS DECIMAL(18,9)) AS
    Vorlage,
CAST(SUM(Tore*gespielt)/SUM(gespielt) AS DECIMAL(18,9)) AS
    Tore,
CAST(SUM(Gelb*gespielt)/SUM(gespielt) AS DECIMAL(18,9)) AS
    Gelb,
CAST(SUM(PassProBallkontakt*gespielt)/SUM(gespielt) AS DECIMAL
    (18,9)) AS PassProBallkontakt,
CAST(SUM(erfPassProBallkontakt*gespielt)/SUM(gespielt) AS
    DECIMAL(18,9)) AS erfPassProBallkontakt,
CAST(SUM(TorschussvorlageProPass*gespielt)/SUM(gespielt) AS
    DECIMAL(18,9)) AS TorschussvorlageProPass
FROM fakten WHERE spieltag > 17 AND spieltag < 35  AND spieler
     IN (SELECT spieler FROM spielereigenschaften WHERE
    spielerposition IN ('Abwehr', 'Mittelfeld','Angriff'))
GROUP BY saison, verein;
```

**Abbildung 28:** SQL Clusteredspieler pro Verein

```
SELECT
saison, runde, verein,
SUM((CASE WHEN spielerposition = 'Abwehr' AND cluster = '
    cluster0' THEN gespielt ELSE 0 END)) AS Abwehr0,
SUM((CASE WHEN spielerposition = 'Abwehr' AND cluster = '
    cluster1' THEN gespielt ELSE 0 END)) AS Abwehr1,
SUM((CASE WHEN spielerposition = 'Abwehr' AND cluster = '
    cluster2' THEN gespielt ELSE 0 END)) AS Abwehr2,
SUM((CASE WHEN spielerposition = 'Abwehr' AND cluster = '
    cluster3' THEN gespielt ELSE 0 END)) AS Abwehr3,
SUM((CASE WHEN spielerposition = 'Abwehr' AND cluster = '
    cluster4' THEN gespielt ELSE 0 END)) AS Abwehr4,
SUM((CASE WHEN spielerposition = 'Mittelfeld' AND cluster = '
    cluster0' THEN gespielt ELSE 0 END)) AS Mittelfeld0,
SUM((CASE WHEN spielerposition = 'Mittelfeld' AND cluster = '
    cluster1' THEN gespielt ELSE 0 END)) AS Mittelfeld1,
SUM((CASE WHEN spielerposition = 'Mittelfeld' AND cluster = '
    cluster2' THEN gespielt ELSE 0 END)) AS Mittelfeld2,
SUM((CASE WHEN spielerposition = 'Mittelfeld' AND cluster = '
    cluster3' THEN gespielt ELSE 0 END)) AS Mittelfeld3,
SUM((CASE WHEN spielerposition = 'Mittelfeld' AND cluster = '
    cluster4' THEN gespielt ELSE 0 END)) AS Mittelfeld4,
SUM((CASE WHEN spielerposition = 'Angriff' AND cluster = '
    cluster0' THEN gespielt ELSE 0 END)) AS Angriff0,
SUM((CASE WHEN spielerposition = 'Angriff' AND cluster = '
    cluster1' THEN gespielt ELSE 0 END)) AS Angriff1,
SUM((CASE WHEN spielerposition = 'Angriff' AND cluster = '
    cluster2' THEN gespielt ELSE 0 END)) AS Angriff2,
SUM((CASE WHEN spielerposition = 'Angriff' AND cluster = '
    cluster3' THEN gespielt ELSE 0 END)) AS Angriff3,
SUM((CASE WHEN spielerposition = 'Angriff' AND cluster = '
    cluster4' THEN gespielt ELSE 0 END)) AS Angriff4,
SUM((CASE WHEN spielerposition = 'Angriff' AND cluster = '
    cluster5' THEN gespielt ELSE 0 END)) AS Angriff5
FROM
(SELECT
f.saison, c.runde, f.verein,f.spieler, f.gespielt,  c.
    spielerposition, c.cluster
FROM (SELECT saison, 'Hinrunde' AS runde, verein, spieler, SUM
    (gespielt) AS gespielt FROM fakten
WHERE spieltag > 0 AND spieltag < 18
GROUP BY saison,verein,spieler) f,
clusteredspieler c WHERE c.spieler = f.spieler AND c.saison =
    f.saison AND c.runde = f.runde
UNION
SELECT
f.saison, c.runde, f.verein,f.spieler, f.gespielt, c.
    spielerposition, c.cluster
FROM (SELECT saison, 'Rueckrunde' AS runde, verein, spieler,
```

```
   SUM(gespielt) AS gespielt  FROM fakten
WHERE spieltag > 17 AND spieltag < 35
GROUP BY saison,verein,spieler) f,
clusteredspieler c WHERE c.spieler = f.spieler AND c.saison =
   f.saison AND c.runde = f.runde) a
GROUP BY saison, runde, verein;
```

## Literaturverzeichnis


[1] Wladimir Awerbuch. „Anwendung von Data Mining zu statistischen Auswertungen und Vorhersagen im Sport“. Diplomarbeit. TU Darmstadt, Knowledge Engineering Group, 2009. URL: http://www.ke.tu-darmstadt.de/lehre/arbeiten/diplom/2009/Awerbuch_Wladimir.pdf.

[2] Mathew Beckler, Hingfei Wang und Michael Papamicheal. *NBA Oracle*. Zuletzt besucht am 17.06.2013. URL: http://www.mbeckler.org/coursework/2008-2009/10701_report.pdf.

[3] Christoph Biermann. *Digitales Scouting: Wunderstürmer gesucht*. Zuletzt besucht am 14.03.2013. URL: http://www.spiegel.de/sport/fussball/digitales-scouting-wunderstuermer-gesucht-a-688624.html.

[4] BUNDESLIGA.de. *Ligavorstand vergibt Auftrag zur Erhebung offizieller Spieldaten*. Zuletzt besucht am 03.06.2013. URL: http://www.bundesliga.de/de/liga/news/2010/index.php?f=0000176864.php.

[5] BUNDESLIGA.de. *System = Erfolg: bundesliga.de nutzt Opta Index*. Zuletzt besucht am 23.06.2013. URL: http://www.bundesliga.de/de/liga/news/2007/index.php?f=79974.php.

[6] Chenjie Cao. *Sports Data Mining Technology Used in Basketball Outcome Prediction*. Athens, Gerogia, 2012.

[7] Timothy C. Y. Chan, Justin A. Cho und David C. Novati. „Quantifying the Contribution of NHL Player Types to Team Performance“. In: *Interfaces* 42.2 (März 2012), S. 131–145.

[8] Timothy C. Y. Chan und David C. Novati. „Split personalities of NHL players: Using clustering, projection and regression to measure individual point shares“. In: *MIT Sloan Sports Analytics Conference*. 2012. URL: http://www.sloansportsconference.com/wp-content/uploads/2012/02/59-Chan_Novati_Split-personalities-of-NHL-players.pdf.

[9] Nitesh V. Chavla. „Data Mining For Imbalanced Datasets: An Overview“. In: *Data Mining and Knowledge Discovery Handbook*. Hrsg. von Oded Z. Maimon und Lior Rokach. Berlin, Heidelberg: Springer, 2005, S. 853–863.

[10] Hsinchun Chen, Peter Buntin Rinde, Linlin She, Siunie Sutjahjo, Chris Sommer und Daryl Neely. „Expert Prediction, Symbolic Learning, and Neural Networks: An Experiment on Greyhound Racing“. In: *IEEE Expert: Intelligent Systems and Their Applications* 9.6 (Dez. 1994), S. 21–27.

[11] Barak Chizi und Oded Maimon. „Dimension Reduction and Feature Selection“. In: *Data Mining and Knowledge Discovery Handbook*. Hrsg. von Oded Z. Maimon und Lior Rokach. Berlin, Heidelberg: Springer, 2005, S. 93–109.

[12] COMUNIO.de. *Spielregeln*. Zuletzt besucht am 08.04.2013. URL: http://www.comunio.de/rules.phtml.

[13] Gabriel B. Costa, Michael R. Huber und John T. Saccoman. *Understanding Sabermetrics – An Introduction to the Science of Baseball Statistics*. North Carolina: McFarland, 2008.

[14] DFL. *Report 2012: Die wirtschaftliche Situation im Lizenzfußball*. Zuletzt besucht am 04.07.2013. URL: http://www.bundesliga.de/media/native/autosync/report_2013_dt_72dpi.pdf.

[15] David Dyte und Stephen R. Clarke. „A Ratings Based Poisson Model for World Cup Soccer Simulation“. In: *The Journal of the Operational Research Society* 51.8 (2000), S. 993–998.

[16] Usama M. Fayyad, Gregory Piatetsky-Shapiro und Ramasamy Uthurusamy. „From Data Mining to Knowldge Discovery: An Overview“. In: *Advances in Knowledge Discovery and Data Mining*. Hrsg. von Usama M. Fayyad, Gregory Piatetsky-Shapiro und Ramasamy Uthurusamy. Cambridge: MIT Press, 1996.

[17] FIFA. *Reglement: FIFA Fussball–Weltmeisterschaft Brasilien 2014*. Zuletzt besucht am 18.03.2013. URL: http://de.fifa.com/mm/document/tournament/competition/01/47/38/17/regulationsfwcbrazil2014_de.pdf.

[18] Karl Flinders. *Football injuries are rocket science*. Zuletzt besucht am 23.06.2013. URL: http://www.v3.co.uk/v3-uk/news/1950164/football-injuries-rocket-science.

[19] John F. Gantz und David Reinsel. *The Digital Universe Decade – Are You Ready?* 2010. URL: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf.

[20] John F. Gantz und David Reinsel. *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*. 2012. URL: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

[21] SPORTAL GmbH. *Bundesliga–Noten von sportal.de für Comunio*. Zuletzt besucht am 08.04.2013. URL: http://www.sportal.de/bundesliga-noten-von-sportalde-fuer-comunio-1-20120824232946000000.

[22] Mark A. Hall. „Correlation-based Feature Subset Selection for Machine Learning“. Diss. Hamilton, New Zealand: University of Waikato, 1998.

[23] Nikolaus Heindl. *Scouting 2.0*. Zuletzt besucht am 13.03.2013. URL: http://www.fcbayern.telekom.de/de/aktuell/news/2010/24055.php.

[24] Christian Hornung. *Favre legt sich mit Eberl an: Zoff um 12–Mio–Stürmer de Jong*. Zuletzt besucht am 22.04.2013. URL: http://www.bild.de/sport/fussball/lucien-favre/zoff-mit-eberl-wegen-12-mio-stuermer-de-jong-25796910.bild.html.

[25] HTTRACK.com. *HTTrack Website Copier*. Zuletzt besucht am 03.06.2013. URL: http://www.httrack.com/.

[26] Fabio Huber. *1899 Hoffenheim: Forsche Töne von Markus Babbel*. Zuletzt besucht am 30.05.2013. URL: http://www.echo-online.de/sport/fussball/1bundesliga/buli2012./1899-Hoffenheim-Forsche-Toene-von-Markus-Babbel;art2400,3117714.

[27] IMPIRE. *Die IMPIRE AG*. Zuletzt besucht am 03.06.2013. URL: http://www.bundesliga-datenbank.de/de/aboutus/.

[28] IMPIRE. *Offizielle Spieldaten der Bundesliga*. Zuletzt besucht am 03.06.2013. URL: http://www.bundesliga-datenbank.de/fileadmin/impire/home/Kommunikation_offizielle_Spieldaten_impire.de.pdf.

[29] IMPIRE. *Produkte*. Zuletzt besucht am 03.06.2013. URL: http://www.bundesliga-datenbank.de/de/products/.

[30] IMPIRE. *Willkommen bei IMPIRE*. Zuletzt besucht am 03.06.2013. URL: http://www.bundesliga-datenbank.de/index.php?topic=Unternehmen.

[31] INSPECT-ONLINE.de. *Fußballer unter der Lupe*. Zuletzt besucht am 04.06.2013. URL: http://www.inspect-online.com/topstories/automation/fussballer-unter-der-lupe.

[32] Zdravko Ivankovic, Milos Rackovic, Branko Markoski, Dragica Radosav und Miodrag Ivkovic. „Appliance of Neural Networks in Basketball Scouting“. In: *ACTA POLYTECHNICA HUNGARICA* 7.4 (2010), S. 167–180.

[33] Philipp Köster. *Noten für Fußballer: Setzen, Sechs!* Zuletzt besucht am 08.04.2013. URL: http://www.spiegel.de/sport/fussball/noten-fuer-fussballer-setzen-sechs-a-543832.html.

[34] Ullrich Kroemer. *Kicker-Noten – Die Zeugnisse der Profis*. Zuletzt besucht am 08.04.2013. URL: http://www.news.de/sport/855093240/die-zeugnisse-der-profis/1/.

[35] Rene Kübler. *Wie der SC Freiburg nach Talenten sucht – auf der ganzen Welt*. Zuletzt besucht am 13.03.2013. URL: http://www.badische-zeitung.de/sport/scfreiburg/wie-der-sc-freiburg-nach-talenten-sucht-auf-der-ganzen-welt--40857783.html.

[36] Simon Kuper. *Milan Lab's secret of youth*. Zuletzt besucht am 23.06.2013. URL: http://www.ft.com/intl/cms/s/0/c56b9be6-e6ff-11dc-b5c3-0000779fd2ac.html#axzz2ZOx9LJyX.

[37] Simon Kuper und Stefan Szymanski. *Soccernomics - Why England Loses, Why Spain, Germany, and Brazil Win, and Why the US, Japan, Australia, Turkey-And Even Iraq-Are Destined to Become the Kings of the World's Most Popular Sport*. Nation Books, 2012.

[38] Bernard Loeffelholz, Earl Bednar und Kenneth W Bauer. „Predicting NBA Games Using Neural Networks“. In: *Journal of Quantitative Analysis in Sports* 5.1 (2009), S. 1–17.

[39] Arlo Lyle. *Baseball Prediction Using Ensemble Learning*. Athens, Gerogia, 2007.

[40] Oded Z. Maimon und Lior Rokach. „Clustering Methods“. In: *Data Mining and Knowledge Discovery Handbook*. Hrsg. von Oded Z. Maimon und Lior Rokach. Berlin, Heidelberg: Springer, 2005, S. 321–349.

[41] Oded Maimon und Lior Rokach. „Introduction To Knowledge Discovery In Databases“. In: *Data Mining and Knowledge Discovery Handbook*. Hrsg. von Oded Z. Maimon und Lior Rokach. Berlin, Heidelberg: Springer, 2005, S. 1–13.

[42] Oded Maimon und Lior Rokach. „Introduction to Supervised Methods“. In: *Data Mining and Knowledge Discovery Handbook*. Hrsg. von Oded Z. Maimon und Lior Rokach. Berlin, Heidelberg: Springer, 2005, S. 149–162.

[43] AC Milan. *Milan Lab*. Zuletzt besucht am 23.06.2013. URL: http://www.acmilan.com/en/club/milan_lab.

[44] Thomas M. Mitchell. *Machine Learning*. 1. Aufl. New York, USA: McGraw-Hill, Inc., 1997.

[45] Bret R. Myers. „A Proposed Decision Rule for the Timing of Soccer Substitutions“. In: *Journal of Quantitative Analysis in Sports* 8.1 (2012), S. 1–24.

[46] John Naisbitt. *Megatrends – Ten New Directions Transforming Our Lives*. Warner Books, 1982.

[47] Joel Oberstone. „Comparing Team Performance of the English Premier League, Serie A, and La Liga for the 2008-2009 Season“. In: *Journal of Quantitative Analysis in Sports* 7.1 (2011), S. 1–18.

[48] Joel Oberstone. „Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success“. In: *Journal of Quantitative Analysis in Sports* 5.3 (2009), S. 1–29.

[49] OPTA. *Company history*. Zuletzt besucht am 04.06.2013. URL: http://www.optasports.com/en/about/who-we-are/company-history.aspx.

[50] John C. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. Zuletzt besucht am 09.06.2013. URL: http://research.microsoft.com/en-us/um/people/jplatt/smo-book.pdf.

[51] Günter Daniel Rey und Karl F. Wender. *Neuronale Netze: Eine Einführung in die Grundlagen, Anwendungen und Datenauswertung*. 1. Aufl. Bern, Schweiz: Huber, 2008.

[52] Alexander P. Rotshtein, Morton Posner und Hanna B. Rakityanskaya. „Football Predictions Based on a Fuzzy Model with Genetic and Neural Tuning“. In: *Cybernetics and Systems Analysis* 41.4 (Juli 2005), S. 619–630.

[53] Lothar Sachs. *Angewandte Statistik – Planung und Auswertung – Methoden und Modelle*. 4. Aufl. Berlin: Springer DE, 1974.

[54] Robert P. Schumaker, Osama K. Solieman und Hsinchun Chen. *Sports Data Mining*. 1st. Springer Publishing Company, Incorporated, 2010.

[55] Paola Sebastiani, Maria M. Abad und Marco F. Ramoni. „Bayesian Networks“. In: *Data Mining and Knowledge Discovery Handbook*. Hrsg. von Oded Z. Maimon und Lior Rokach. Berlin, Heidelberg: Springer, 2005, S. 193–226.

[56] Armin Shmilovici. „Support Vector Machines“. In: *Data Mining and Knowledge Discovery Handbook*. Hrsg. von Oded Z. Maimon und Lior Rokach. Berlin, Heidelberg: Springer, 2005, S. 257–273.

[57] Alex J. Smola und Scholkopf Bernhard. *A Tutorial on Support Vector Regression*. Zuletzt besucht am 09.06.2013. URL: http://alex.smola.org/papers/2003/SmoSch03b.pdf.

[58] SPIEGEL.de. *Fußball-WM: Lehmanns Geheimnis gelüftet*. Zuletzt besucht am 24.06.2013. URL: http://www.spiegel.de/sport/fussball/fussball-wm-lehmanns-geheimnis-gelueftet-a-438988.html.

[59] TAGESBLATT.de. *Wie die Firma Impire die Spieldaten in der Fußball-Bundesliga erfasst*. Zuletzt besucht am 03.06.2013. URL: http://www.tagblatt.de/Home/sport/ueberregionaler-sport_artikel,-Wie-die-Firma-Impire-die-Spieldaten-in-der-Fussball-Bundesliga-erfasst-_arid,170716.html.

[60] Ben Van Calster, Smits Tim und Sabine Van Huffel. „The Curse of Scoreless Draws in Soccer: The Relationship with a Team's Offensive, Defensive, and Overall Performance“. In: *Journal of Quantitative Analysis in Sports* 4.1 (2008), S. 1–24.

[61] Claude B. Vincent und Byron Eastman. „Defining the Style of Play in the NHL: An Application of Cluster Analysis“. In: *Journal of Quantitative Analysis in Sports* 5.1 (2009), S. 1–23.

[62] Lars Wallrodt und Udo Muras. *Die Trainer Tuchel und Klopp zeigen, wie es geht*. Zuletzt besucht am 05.05.2013. URL: http://www.welt.de/sport/fussball/bundesliga/1-fsv-mainz/article9886638/Die-Trainer-Tuchel-und-Klopp-zeigen-wie-es-geht.html.

[63] Björn Wannhoff. *Ist Herrmann das neue Super-Fohlen?* Zuletzt besucht am 05.05.2013. URL: http://www.t-online.de/sport/fussball/bundesliga/id_53378844/patrick-herrmann-tritt-in-die-fussstapfen-von-marco-reus.html.

[64] WELT.de. *BVB–Star Nuri Sahin wechselt zu Real Madrid*. Zuletzt besucht am 12.05.2013. URL: http://www.welt.de/sport/fussball/bundesliga/borussia-dortmund/article13360770/BVB-Star-Nuri-Sahin-wechselt-zu-Real-Madrid.html.

[65] WELT.de. *Götze wechselt für 37 Millionen zum FC Bayern*. Zuletzt besucht am 12.05.2013. URL: http://www.welt.de/sport/fussball/article115517510/Goetze-wechselt-fuer-37-Millionen-zum-FC-Bayern.html.

[66] Brady T. West und Madhur Lamsal. „A New Application of Linear Modeling in the Prediction of College Football Bowl Outcomes and the Development of Team Ratings“. In: *Journal of Quantitative Analysis in Sports* 4.3 (2008), S. 1–21.

[67] Ian H. Witten, Eibe Frank und Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques – Practical Machine Learning Tools and Techniques*. 3. Aufl. Amsterdam: Elsevier, 2011.

[68] Andreas Zell. *Simulation neuronaler Netze*. 1. Aufl. Deutschland: Addison-Wesley, 1994.