Faculty 4: Computer Science

Institute for Web Science
and Technologies

Predictive Analytics for Early Identification of At-Risk Students

Master's Thesis

in partial fulfillment of the requirements for
the degree of Master of Science (M.Sc.)
in Web and Data Science

submitted by
Nileena Ans Biju (222100485)

First supervisor:	Prof. Dr. Frank Hopfgartner Institute for Web Science and Technologies
Second supervisor:	Dr.-Ing. Stefania Zourlidou Institute for Web Science and Technologies

Koblenz, September 2024

Statement

I hereby certify that this thesis has been composed by me and is based on my own work, that I did not use any further resources than specified – in particular no references unmentioned in the reference section – and that I did not submit this thesis to another examination before. The paper submission is identical to the submitted electronic version.

	Yes	No
I agree to have this thesis published in the library.	☒	□
I agree to have this thesis published on the Web.	☒	□
The thesis text is available under a Creative Commons License (CC BY-SA 4.0).	☒	□
The source code is available under a GNU General Public License (GPLv3).	☒	□
The collected data is available under a Creative Commons License (CC BY-SA 4.0).	☒	□

KOBLENZ, 03.09.2024

(Place, Date) (Signature)

Note

- If you would like us to contact you for the graduation ceremony, please provide your personal E-mail address: nileenaans95@gmail.com
- If you would like us to send you an invite to join the WeST Alumni and Members group on LinkedIn, please provide your LinkedIn ID : Nileena Ans Biju

Zusammenfassung

Im Bildungsbereich spielt die rechtzeitige Erkennung von Studierenden, die weitere Unterstützung benötigen, um in ihren jeweiligen Kursen erfolgreich zu sein, eine entscheidende Rolle bei der Förderung des akademischen Erfolgs und der Vermeidung möglicher Rückschläge. Diese Arbeit soll daher einen Beitrag zu diesem kritischen Bereich leisten, indem sie sich auf die Entwicklung von Vorhersagemodellen für die frühzeitige Erkennung von Risikostudenten auf ihrem akademischen Weg konzentriert. Der primäre Datensatz, der für diese Arbeit verwendet wird, wird von kaggle zur Verfügung gestellt und umfasst verschiedene Studenteninformationen, einschließlich demographischer, sozioökonomischer Faktoren und akademischer Leistungen, die in drei verschiedene Klassen eingeteilt sind.

Die Hauptziele dieser Arbeit sind daher die Lösung des Problems der unausgewogenen Daten, die Erforschung und Bewertung der Leistung verschiedener Klassifizierungsmethoden wie logistische Regression, Entscheidungsbäume, Zufallswälder und Support Vector Machines (SVM) sowie neuronale Netze und die Entwicklung einer umfassenden End-to-End-Verarbeitungspipeline, die die systematischen Schritte des Datenausgleichs, des Modelltrainings und der Bewertung umfasst. Zusätzlich wird die entwickelte Pipeline an zwei weiteren Datensätzen getestet, um ihre Verallgemeinerbarkeit und Robustheit zu bewerten. Diese Forschung zielt darauf ab, ein umfassendes Verständnis für die Herausforderungen unausgewogener Daten zu schaffen und zu zeigen, wie verschiedene Klassifizierungs- und Regressionsmethoden optimal für die Früherkennung von gefährdeten Schülern eingesetzt werden können. Die Ergebnisse sollen Bildungseinrichtungen dabei helfen, ihre Schüler zu unterstützen und den akademischen Erfolg durch rechtzeitige Interventionen zu verbessern.

Die wichtigsten Ergebnisse zeigen die Robustheit der SVM SMOTE-Balancierungs technik über die in dieser Studie verwendeten Datensätze hinweg, wo sie durchweg die besten Ergebnisse erzielte, wenn sie mit verschiedenen Modellen kombiniert wurde, wobei insbesondere der Erfolg der Kombination des Random Forest-Modells mit SVM SMOTE und des Entscheidungsbaummodells mit SVM SMOTE bei der Erzielung bemerkenswerter Genauigkeitsraten hervorgehoben wird. Dies unterstreicht die Anpassungsfähigkeit der eingesetzten Balancierungstechniken, die eine solide Grundlage für prädiktive Interventionen im Bildungsbereich bilden.

Abstract

In the realm of education, the timely identification of students who need further support to succeed in their respective courses, plays a pivotal role in fostering academic success and preventing potential setbacks. This thesis thus aims to contribute to this critical area by focusing on the development of predictive models for the early detection of at-risk students in their academic journey. The primary dataset used for this thesis is provided by kaggle, encompassing diverse student information, including demographic, socio-economic factors, and academic performance categorized into three different classes, presenting an imbalanced nature that poses a significant challenge.

Thus the primary objectives of this thesis are to address the problem of imbalanced data, explore and assess the performance of multiple classification methods such as, logistic regression, decision tress, random forests and support vector machines (SVM), neural networks, and create a comprehensive end-to-end processing pipeline which includes the systematic steps of balancing the data, model training and evaluation. Additionally the developed pipeline is tested on two additional datasets to assess its generalizability and robustness. This research aims to provide a comprehensive understanding of addressing the challenges of imbalanced data and how different classification methods and regression can be optimally applied to early detection of at-risk students. The findings are expected to aid educational institutions in supporting their students and enhancing academic success through timely interventions.

Key findings demonstrates the robustness of SVM SMOTE balancing technique across the datasets used in this study, where it consistently achieved best results when combined with various models, particularly highlighting the success of the combination of Random Forest model with SVM SMOTE, and Decision tree model with SVM SMOTE in achieving notable accuracy rates. This emphasizes the adaptability of the balancing techniques employed, providing a strong foundation for predictive intervention educational settings.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Motivation

Education opens the doors to individual opportunities. Not only that, education also helps in driving societal progress. Hence, it is a critical mission for educational institutions worldwide to ensure that every student succeeds. In this digital world, predictive analytics has emerged as a powerful tool which can transform educational strategies and outcomes. It uses historical data to predict future events, like in situations where the students are at risk of falling behind. This power of predicting is highly valuable since it allows schools to intervene at the right time, providing the support when it is needed the most. There has been many researches that has successfully integrated machine learning into the educational contexts by uncovering the potential of data mining and learning analytics in the educational field [7].

Educational institutions often face a lot of challenges which hinders student success, such as different learning needs, diverse socio-economic backgrounds and also the wide range of personal circumstances impacting the academic performance. While traditional education models struggle often to address these challenges effectively, since they lack the necessary customization for meeting the educational needs, predictive analytics offers a solution by enabling a personalized approach to education, tailoring the learning experiences according to the unique needs of each student based on the insights derived from the data.

In educational realm, the use of data goes beyond simple record keeping. When it is analyzed intelligently, data can reveal underlying patterns and predictors of student success or failure. This insight not only allows the educators to identify at-risk students, but also helps in understanding the factors that leads to the failure of the student. Data helps in creating a clearer picture of what actions can be taken in order to help improve student outcomes, whether it is tracking attendance, engagement or performance. Predictive models can use this data to forecast the effectiveness of various intervention strategies, which will allow educators to choose the most effective approaches for helping students.

The adoption of predictive analytics in education has been slow despite of its potential benefits. One major barrier is the complexity associated with these technologies. Many educators are intimidated by the technical aspects of predictive analytics and are unclear about how to integrate these tools into their teaching practices. This thesis aims to unmask predictive analytics, by providing clear examples and guidance on how educational data can be effectively leveraged. By simplifying these processes, this study aims to make predictive analytics more accessible to educators, which will help them to make informed decisions that can positively impact the academic journey of the students.

This research is driven by the vision of a more inclusive and effective educational system, where help is provided to the students at-risk at the right time and ensures that no student falls through the cracks. By harnessing the power of predictive analytics, educational students can be more adaptive, responsive, and supportive. The goal is to move beyond the reactive educational models to proactive ones which can anticipate and meet the needs of all students. By doing so, the research hopes to contribute to a future where education becomes more equitable, outcomes are improved, and every student has the opportunity to succeed.

1.2. Objectives

This thesis is structured to address the comprehensive goals using the advanced machine learning techniques, with an aim of developing a robust framework for identifying the student at-risk in educational settings. The methodology integrates several core objectives, each of them designed to optimize and validate the predictive capabilities of our models within the educational field.

1.2.1. Development of Predictive Models

This study involves constructing sophisticated machine learning models that can effectively analyze complex educational data. This approach aims to explore and utilize the performance of various machine learning algorithms including logistic regression, decision trees, random forests, SVMs, and neural networks, to develop a nuanced understanding of the factors influencing the student performance and assess how the data features are utilized in predicting at-risk students [31].

1.2.2. Handling Imbalanced Data

It is very crucial to address the imbalance in educational datasets for ensuring the accuracy and fairness of the predictive models. This thesis employs advanced sampling techniques, such as SMOTE, ADASYN, etc. to balance the data and to enhance the ability of models to identify at-risk students without being biased towards the majority class.

1.2.3. Creation of an End-to-End Processing Pipeline

The core of this research is to develop a comprehensive processing pipeline which encapsulates all the aspects of the predictive modeling process, including scaling, balancing, model development and training. This pipeline is not only designed for ensuring robustness and efficiency, but also for replicability. Each step will be documented in detail, which can then be used as a blueprint by other researchers to follow or adapt according to their specific educational contexts.

1.2.4. Model Generalization and Evaluation

This thesis also tests the effectiveness of the developed models and pipelines on other datasets to evaluate their robustness and generalizability. This ensures that the models can be applied to different educational environments with diverse student populations successfully.

1.2.5. Documentation

A significant step that was carried throughout the research process is comprehensive documentation. This ensures that every aspect of the study, from the initial stage to final results are transparent, scalable and replicable.

1.2.6. Empowering Educational Institutions

The ultimate objective of this thesis is to provide actionable, data-driven tools for identifying at-risk students at an early stage to the educators. These tools aim to improve educational outcomes and reduce dropout rates effectively across various educational settings, by enabling timely and effective interventions.

This thesis aims to contribute significantly to the field of educational technology by achieving these objectives, to provide a practical, evidence based approach to enhance student support systems by leveraging predictive analytics. The detailed framework and findings aims to bridge the gap between the theoretical research and practical applications, thus offering new pathways for educational advancement.

1.3. Thesis Organization

- Chapter 1: Introduction - Provides an overview of the study, background and objectives and organization of the thesis.
- Chapter 2: Literature Review - Reviews existing literature related to the at-risk students, imbalanced data, application of machine learning in the educational domain.

- Chapter 3: Theoretical background - Discusses the theoretical frameworks and academic theories that supports the research.
- Chapter 4: Methodology - Describes the research design, data, pre-processing, exploratory data analysis, pipeline creation, model development and training, model evaluation and testing on three datasets.
- Chapter 5: Results - Presents the findings from model evaluation on the datasets.
- Chapter 6: Conclusion - Summarizes the study findings, draws conclusion.
- References: List of all the sources that are cited in the thesis.
- Appendix: Contains detailed supplementary information on the list of attributes in the datasets and their descriptions.

2. Literature Review

2.1. Existing Works

Using predictive analytics in education field is inspired from a spectrum of previous works, each of them adding special element to the larger picture. Baker's work [8] laid the foundation by illustrating the transformative potential of data mining in understanding student behavior and performance throughout their educational journey. Siemens and Baker [40] expanded this narrative by delving into the intersection of learning analytics and educational data mining, emphasizing collaborative approaches for deeper insights.

Recent literature on early identification of at-risk students reflects a dynamic landscape with a pronounced emphasis on developing and comparing predictive models using diverse machine learning algorithms. Tufail, et al.[43] contributed significantly with a comprehensive review, providing an extensive analysis of a number of various machine learning algorithms and the domains in which these algorithms may be applied. They concluded by mentioning that there is no such thing as a universal algorithm that fits all, rather, the efficacy of any given model depends not only on the model, but also on the kinds of data that it uses. Which can be understood by other research works like article by Hussein, et al. [5], where they investigated four machine learning models to predict the performance of students in a specific subject, computer science, along with focusing on the impact of internet usage on their grades. The highest accuracy was said to be achieved by ANN (Artificial Neural network model) model in their research with 77.04 percent.

Whereas, an attempt has been made on the possibility of developing an academic performance prediction model for at-risk students (with low scores) by Adedokun, et al. [2] in their article of Data Mining Technique For Early Detection Of At-Risk Students, where the aim was to assist the students and their parents or guardian to make an informed decision on the change of selected arm of senior secondary school class as early as possible to achieve better academic performance [2], where five machine learning techniques namely, decision trees, random forest, SVM, ANN, naive bayes, were used to construct the models, the attributes used were previous scores in related subjects and present score in the present class. Clearly, random forest achieved an accuracy of 98.2 percent based on the data used, which shows the potential efficacy of random forest as a predictive model for early detection of at-risk secondary school students in their research. This work has paved a way for other researches like the article in 2023 by Balabied, et al. [35] to use random forest algorithm for early detection of academic under performance in open learning environments. The objective of their study was to use random forest classifier model for analyzing anonymized large datasets available from Open University Learning Analytics (OULAD) to identify patterns and relationships among various factors that contribute to student success or failure. The results of their study indicated that the random forest based model provides a powerful tool for identifying students who are prone to failure and guiding them towards success.

In 2022 , Singh, et al. [42] has published an article on the generation of decision tree model to help in enhancing the quality of the primary educational system by evaluating student data to study the main attributes that may affect the student performance in primary classes. This study showed how decision tree model are efficient in generating "if..then" rules that may be useful for taking decisions to improve academic performance of primary school students.

However according to the research by Cardona, et al. [14] at 2019, SVM technique provides a good resource for the prediction of student success in a Midwest community college for students in STEM majors. The model was developed using data of 282 students with 9 variables. The variables included age, gender, degree, and college GPA, among others. The model results showed a good performance with recall rates over 70 percent and testing rates over 78 percent.

Imbalanced data is another challenge that is encountered while dealing with different datasets while building models. Imbalanced data often skews the outcomes of predictive model, which leads to less accurate or biased interpretations of student performance thus not satisfying the needs of the students.

A blog posted in Analytics Vidhya in 2024 [11], discusses several advanced techniques for addressing class imbalance in machine learning datasets, which is a common issue across several fields including education domain as well. Techniques such as SMOTE, Tomek Links and ensemble methods are outlined. These methods helps in enhancing the representativeness of minority classes in training datasets which are crucial for improving the accuracy as well as fairness of the predictive models.

A study by Pratama, et al. [32] shows the effect of imbalanced data problem and works on finding the best resampling method that can be implemented into the process of machine learning. They used resampling techniques such as SMOTE, Borderline SMOTE, SMOTE Tomek and classifiers such as Logistic regression, KNN (K-Nearest Neighbors), CART, SVM (Support Vector Machine), Random forest for the student's performance dataset. Among which they found that SMOTE Tomek was the best pair that gave the highest accuracy of 85.8 percent on a 10-fold cross validation and a geometric mean of 0.89, which topped the scores among the other combinations [32].

In the study by Barros, et al. [9], three data classification techniques were mainly used namely, decision trees, neural networks, and balanced bagging, for predicting student dropout rates, leveraging different data balancing methods like downsampling, SMOTE and ADASYN. It was found that while the traditional metrics such as accuracy, recall and f1 score was not able to effectively detect the errors in the minority class in imbalanced datasets, the G-mean and UAR metrics proved to be reliable. This study also highlighted that the balanced bagging technique emerged as the most effective model, particularly when evaluated with G-mean and UAR metrics, which captured the error of minority class better and avoided the accuracy paradox. This research also underlines the importance of using appropriate metrics and data balancing techniques to enhance model performance for effectively predicting student dropouts [9].

A research by Alija, et al. [36] in 2023 explores the efficacy of several supervised machine learning methods in predicting the student failures, addressing the challenges that are caused by an imbalanced dataset. Synthetic Minority Over-sampling technique (SMOTE) was employed to balance the class representation and also two feature selection methods were tested, one, a wrapper approach using Particle Swam Optimization (SPO) and the second, Information Gain with ranker. This study found that the wrapper approach with SPO paired with SMOTE, enhanced the algorithm's performance significantly, specifically for the random forest algorithm, which showed best results in terms of true positive rates, recall and ROC curve metrics. This research illustrated the potential of combining advanced feature selection methods and SMOTE to improve the predictive accuracy in the educational domain, especially in forecasting student outcomes [36].

The effectiveness of combining advanced sampling techniques with ensemble classifiers to improve the accuracy of predictions of student performance was highlighted in a study by Hassan, et al. [23] in 2020, where 4413 student records from the faculty of Engineering at a Malaysian university during the first semester of 2017/2018 was used, to which various sampling and ensemble classifier techniques were evaluated for their effectiveness in predicting the student performances. This study compared three types of sampling methods, namely, oversampling, undersampling and hybrid methods, analyzing the performance of five types of ensemble classifiers across seven sampling techniques. The findings showed that the hybrid technique combining Random Oversampling (ROS) with AdaBoost outperformed the other methods. Also, the SMOTE ENN technique consistently produced high results when used with ensemble classifiers, showcasing the potential for enhancing predictive models in the educational settings [23].

2.2. Research Gap

Despite the existing literature, in which efficient model varies according to the dataset and scope, a significant research gap persists in defining a framework that addresses challenges such as imbalanced data in the context of students of higher education. This thesis aims to bridge this gap by exploring and evaluating various classification methods to detect at-risk students, focusing on the selection and utilization of data features and thus providing a comprehensive understanding of how different classification methods can be optimally applied to imbalanced data. This thesis uses three datasets, among which two are open datasets, thus, the results can be used as a benchmark for other studies.

2.3. Research Questions

1. How can the problem of imbalanced data be addressed in the context of identification of students at risk at an early stage in higher education.

2. How to explore and assess the performance of various machine learning models, such as logistic regression, decision trees, random forests, and SVM, neural networks, in accurately predicting students at risk of academic failure.

3. How to create an end-to-end processing pipeline for the higher education data that encapsulates the entire process including systematic steps for data preprocessing, model selection, algorithm training, model evaluation, and interpretation of results, documenting each step to create a replicable process which can be utilized by other researchers or practitioners in the field of educational technology.

3. Theoretical Background

This section provides an overview of theoretical frameworks and academic theories that helped this research. This sets the foundation of understanding the significance and application of predictive analytics in education.

3.1. Educational Data Mining (EDM)

Educational Data Mining (EDM) is a field that explores the application of data mining techniques to educational data. Education data mining can be further characterised as an exciting area of research that extracts patterns from databases of educational data that can be used for comprehending, improving academic performance and assessing student's learning process [19]. Which can be further elaborated by stating that EDM focuses on developing methods to better understand student's learning processes, predict educational outcomes, and identify at-risk students. Knowledge discover and data mining uses several different classification methods and techniques, each having its advantages and disadvantages [34]. The key objectives of EDM are to transform raw educational data into actionable insights and to support decision making processes in educational institutions.

EDM leverages theories such as constructive learning theory, which propose that the learners constructs their own understanding and knowledge of the world through experiences and reflecting on those experiences [1]. This theoretical perspective supports the analysis of how students interact with educational content, enabling the creation of personalized learning experiences. Additionally cognitive load theory, which addresses the limitations of working memory, which helps in designing interventions that minimizes cognitive overload and enhances learning efficiency [17].

Educational Data Mining can be further enriched by incorporating sophisticated analytical frameworks and emerging theories in educational psychology and machine learning that has both established. The recent advancements focus on multimodal data analysis [27], that significantly enriches the insights that are derived from the educational settings. For instance, researchers are now exploring the simultaneous use of textual, audio, behavioral and video data for gaining a more comprehensive understanding of student engagement and patterns of interactions within the digital learning environments.

Moreover, the integration of network analysis techniques has opened new pathways that helps in understanding social learning dynamics [20]. EDM can uncover important insights into the community aspects of learning by analysing the relationships and interactions between learners within the online forums and collaborative platforms, which are critical for cognitive and social development. This analytical aspect not only helps in enhancing individual learning outcomes but also fosters a more collaborative and supportive environment for learning.

These advancements highlights the transformative potential of EDM in not just in optimizing educational outcomes, but also in revolutionizing educational practices. By making use of the detailed insights provided by these complex data analyses, educators and policymakers will be able to tailor more effective educational strategies to meet the diverse learner needs and contexts. This advancements in EDM not only promises to just enhance the academic performance, but also in equipping learners with the skills that are necessary to thrive in an increasingly digital and interconnected world.

3.2. Predictive Analytics

Predictive analytics involves using statistical techniques and machine learning algorithms for analyzing current and historical data to make predictions about future outcomes. In the educational context, predictive analytics is used for identifying students who are at risk of failing or dropping out, thus enabling timely interventions. Predictive analytics enables the customization of learning paths for individual students through data-driven insights [6].

The theoretical foundation of predictive analytics includes regression analysis, classification algorithms, and ensemble methods.

3.2.1. Regression Analysis

Regression analysis is a key statistical method used to model the relationship between a dependent variable with one or more independent variables. Any application of regression analysis must distinguish between the roles of the two quantitative variables. One, which we wish to predict or we believe which is being influenced is called the dependent variable, or response or outcome variable [18]. The other that we use as the basis of our prediction, or that we believe is causing some change is called the independent, explanatory or predictor variable [18]. The dependent variable is traditionally labeled as 'y' and the independent variable is labeled as 'x' [18]. This statistical method helps in understanding how the changes in independent variable impacts student outcomes, which thus allows educators to identify factors that significantly influence academic performance and do the needful.

There are different types of regression analysis, including linear regression, polynomial regression and logistic regression [39]. Linear regression models the relationship between two variables by fitting a linear equation to the observed data, which helps in predicting the value of the dependent variable based on the value of the independent variable. Multiple regression extends this concept by using two or more dependent variables to predict the dependent variable. Polynomial regression is a variant of this multiple regression. On the other hand, logistic regression is used when the dependent variable is binary, estimating the probability that a given input point belongs to a specific class and is particularly used for classification problems in education [39].

3.2.2. Classification Algorithms

Classification algorithms are a fundamental component of predictive analytics, especially in the field of education where it is used to categorize students based on their risk levels. Classification is a type of supervised learning, where the model is trained on a labeled dataset. The primary aim of classification is to assign input data into predefined classes or categorize based on extracted features from the data [30].

Classification algorithms analyze features such as demographic information, academic history, and socio-economic status to predict whether a student is at risk of academic failure. By doing this, educators can prioritize interventions for the students in need. These algorithms work by learning patterns from a labeled training dataset and using this knowledge to classify new, unseen data. Common classification algorithms include logistic regression, decision trees and support vector machines (SVMs) etc. In summary, classification algorithms are a powerful tool in predictive analytics for the education domain. They help in the identification of at-risk students by analyzing various features and predicts the risk levels of students which helps the educators to allocate resources more effectively and thus improve educational outcomes.

3.2.3. Ensemble Methods

Ensemble methods are advanced techniques used in predictive analytics to improve the accuracy and robustness of models by combining the predictions from multiple individual models. The main concept of ensemble method is that a group of weak learners can come together, forming a strong learned, thereby enhancing the predictive performance compared to any single model. Ensemble methods divide the training datasets into subsets and for those, independent learning models are constructed and then combined to form the right hypothesis as illustrated in the Figure 3.1. [4].

There are several ensemble methods, including bagging, boosting, stacking and random forests [41]. Bagging, or Bootstrap Aggregating involves training multiple model versions on different training subsets of data and averaging their predictions to reduce variance and prevent overfitting. Random forest is an extension of

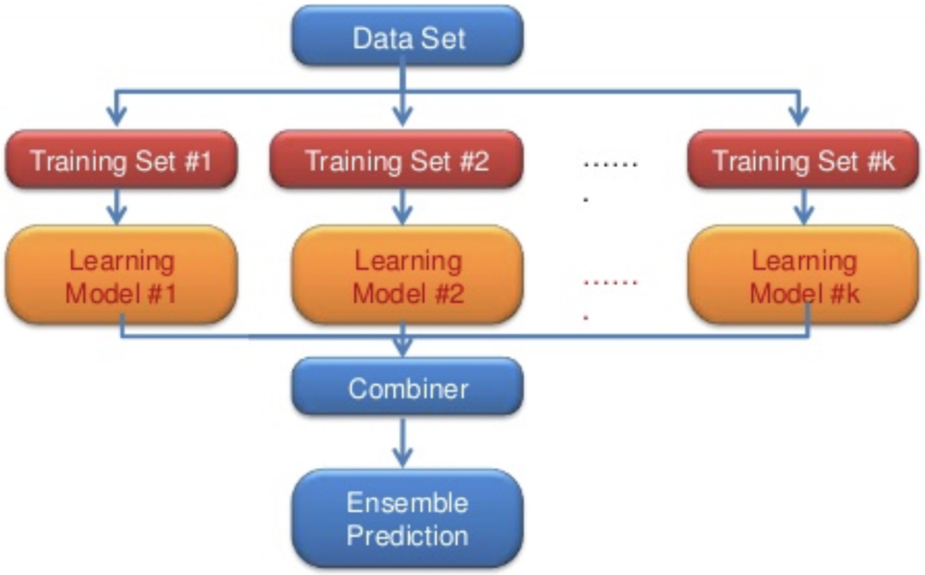

Figure 3.1.: Ensemble Algorithm [4]

decision trees using bagging, building multiple trees on various data and feature subsets, aggregating their outputs to enhance prediction stability and accuracy by capturing a broader range of data patterns [41].

Boosting focuses on sequentially correcting errors of previous models to improve accuracy, thus giving more weight to the misclassified instances of data and reducing bias. Techniques like AdaBoost and Gradient Boosting Machines(GBM), iteratively enhances weak learners to build strong learners by addressing earlier errors. Stacking is another ensemble method which involves training multiple different types of models and then using another model which is called a meta-learner, to combine their predictions. This method makes use of the strengths of various models to achieve better overall performance, stacking is flexible and can be used with a variety of base models, making it a powerful technique for predictive analytics in education [41].

Ensemble methods are particularly valuable in educational predictive learning because they can handle the complexities and nuances of educational data, leading to more accurate and reliable predictions of student outcomes. By combining the strengths of multiple models, ensemble methods help the educators to identify at-risk students better and allocate the needed resources effectively to support their academic success.

3.3. Machine Learning Algorithms

Machine learning algorithms are the backbone of predictive analytics in education that offers various methods to analyze and interpret educational data. The following subsections covers five primary machine learning algorithms which are used in this study: decision trees, logistic regression, random forest, support vector machines(SVMs) and neural networks.

3.3.1. Decision Trees

Decision trees are a type of supervised learning algorithm which is used for classification as well as regression tasks. They operate by splitting the data into subsets based on the value of the input features which creates a tree-like model of decisions. Each internal node in a tree represents a test on an attribute, the outcome of the test is represented by each branch, and leaf nodes represents class label or continuous value. Due to their simplicity and interpretability, decision trees are particularly useful in the educational settings. This algorithm allows educators to understand the decision-making process easily and also to identify the key factors that contributes to the student success or failure. However decision trees can suffer from overfitting, especially when they become too complex and captures the noise in the data [29]. The preferred measures to evaluate the model performance of the classification tree are gini index and entropy. The equation is [4] :

$$\text{GINI} = \sum_{m=1}^{k} p_{ik}(1 - p_{ik}) \tag{3.1}$$

$$\text{Entropy} = -\sum_{m=1}^{k} p_{ik} \log p_{ik} \tag{3.2}$$

Figure 3.2 shows an example of a basic decision tree structure. The top node is known as the root node. The root node starts the decision making process, dividing the data based on specific features to optimize the predictions of outcomes. Below are the decision nodes, which splits the data further, leading to leaf nodes that are the final decision outputs based on the input conditions. Additionally, the diagram highlights a sub-tree, which is a smaller division within the tree that is representing a sequence of decisions taken from the root to the leaf, showing the recursive nature of the decision tree algorithm.

.

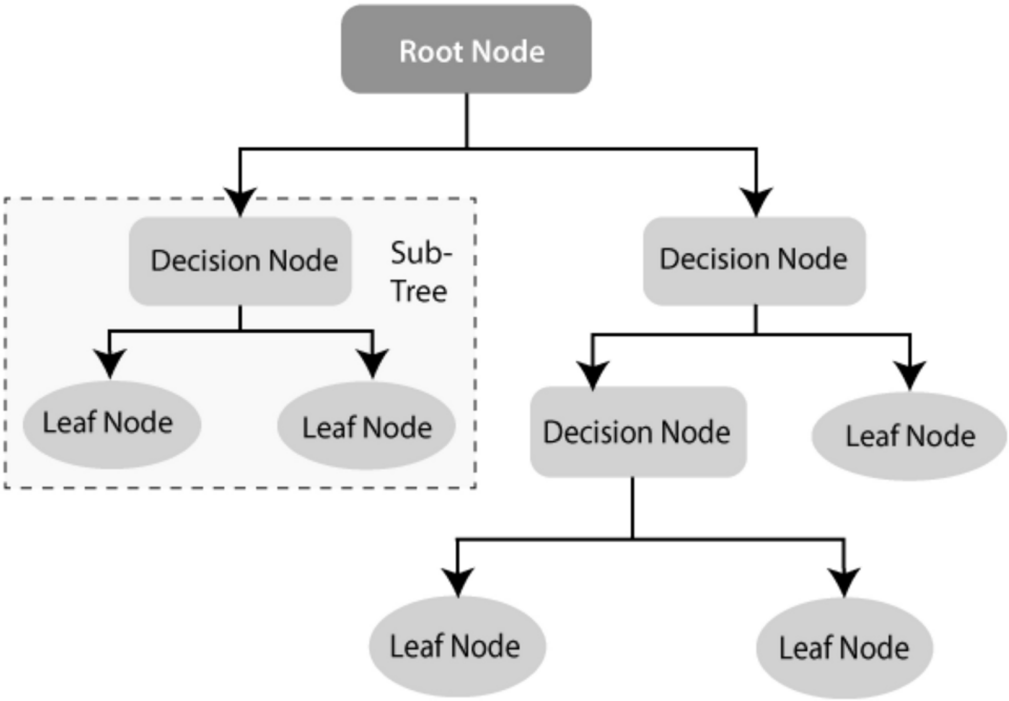

Figure 3.2.: Example of Decision Tree Structure [37]

3.3.2. Logistic Regression

Logistic regression is a statistical model which is commonly used for binary classification problems. It uses a logistic function to model the probability of a given input belonging to a specific class. In the field of education, logistic regression can predict whether a student will pass or fail based on various predictor variables such as grades, attendance, socio-economic factors etc. This model estimates the odds of an event occuring as a function of independent variables, thus providing insights into the relative importance of different predictors in determining student outcomes. Logistic regression is valued for several reasons such as, its simplicity, ease of implementation, and interpretability which makes it a popular choice for predictive analytics in education [26].

The logistic function is represented as [4] :

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.3}$$

The output can be defined as y_{pred} such that [4]:

$$y_{\text{pred}} = \begin{cases} 0 & \text{if } \hat{p} < 0.5 \\ 1 & \text{if } \hat{p} > 0.5 \end{cases}$$

Figure 3.3 shows the logistic regression applied to values ranging from -20 to 20,

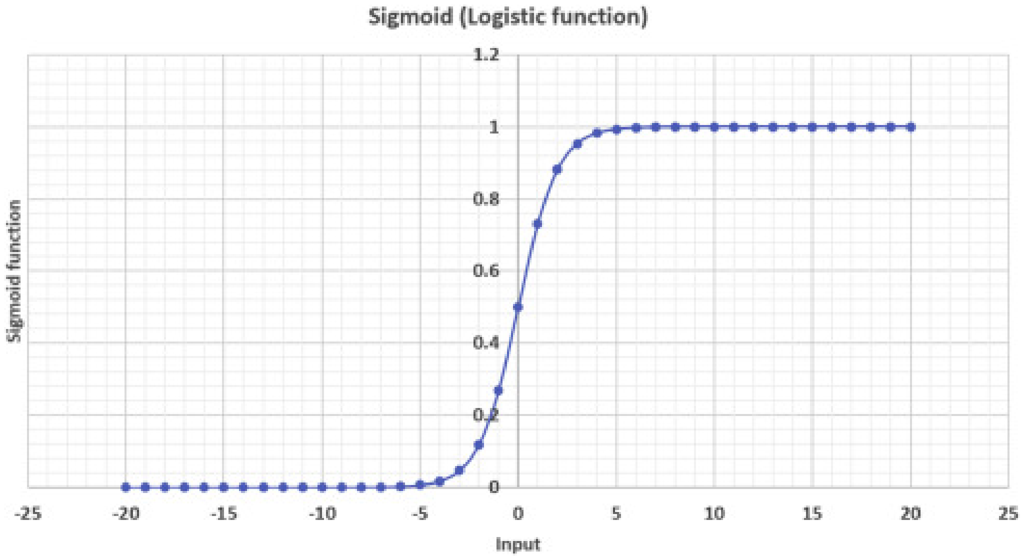

Figure 3.3.: Logistic Regression Applied To a Range of -20 to 20 [10]

3.3.3. Random Forest

Random forest is an ensemble learning method that constructs a multitude of decision tress during training, and aggregates their predictions. Each tree in a random forest is built using a random subset of data and random subset of features that helps in ensuring that the trees are uncorrelated. The final prediction of the random forest is the mode of the classes for the classification tasks or the mean prediction for the regression tasks. This method solved the decision tree's tendency to overfit to the training data by averaging multiple trees, thus improving generalization and robustness. Random forest algorithms are particularly effective in handling large datasets that has high dimensionality and can capture complex interactions between the variables, which makes them a powerful tool in educational predictive analytics [25].

The main concept of random forest is to build as much decision trees possible using multiple data samples by using the majority vote of each group for categorization and also the average if regression is performed [4]. This is illustrated in Figure 3.4.

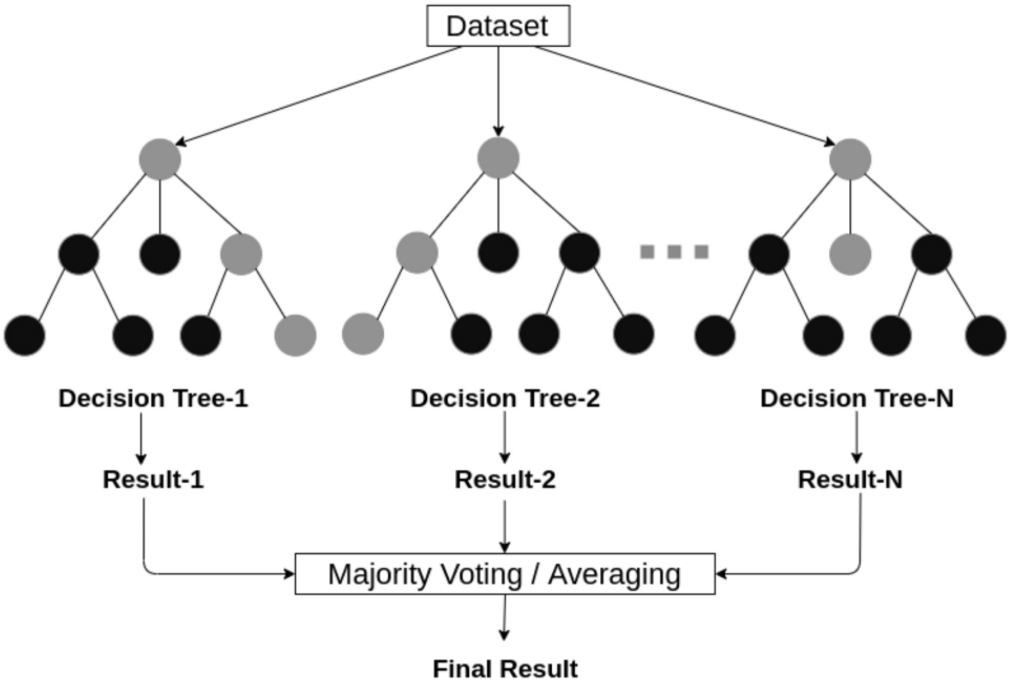

Figure 3.4.: Example of Random Forest Structure Considering Multiple Decision Trees [37]

The mean importance feature calculated from all the trees in the random forest is represented as [4] :

$$F_i = \frac{\sum_{j=1}^{n} f_{ij}}{n} \tag{3.4}$$

where F_i is the mean feature importance for all the trees and f_{ij} represents the feature importance of i in jth tree and n is the number of trees in the forest [4].

3.3.4. Support Vector Machine (SVM)

Support Vector Machines (SVMs) are also a type of supervised learning where models are used for classification and regression analysis. They function by finding the hyperplane that separates the classes in the feature space in the best possible way. SVMs are particularly effective in high-dimensional spaces and are known for their ability to handle large sets of features. In educational applications, SVMs are used to classify students based on their risk levels and also predict academic outcomes based on various predictors, like test scores and attendance records. SVMs are known to be powerful for complex datasets but can be intensive computationally and requires careful tuning of parameters to achieve optimal performance [29]. Figure 3.5 shows an example of linear support vector classifier.

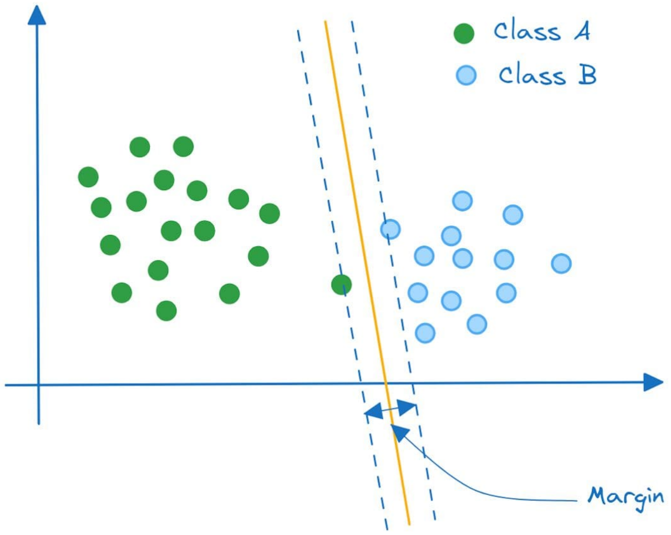

Figure 3.5.: Example of Linear Support Vector Classifier [13]

SVM is known for its robustness and adaptability. It is able to conduct linear or non linear classification and regression tasks and also even identify outliers with sufficiently good accuracy [43]. The core concept of SVM is to put each feature vector in a high-dimensional space and then draw an imaginary high-dimensional line which is known as a hyperplane [43]. The expression of hyperplane is given as [43] :

$$w \cdot x - b = 0 \tag{3.5}$$

$$wx = \sum_{i=1}^{n} w_i x_i \tag{3.6}$$

where w represents the real valued vector, x, the input feature vector, b is a real number, and n represents the number of dimensions of the feature vector [4].

3.3.5. Neural Networks

Neural networks are a class of machine learning algorithms which is inspired by the structure and function of the human brain. They consists of interconnected layers of nodes which are called neurons, which process input data and learn to make predictions through the process called backpropagation. Neural networks are particularly powerful for handling large and complex datasets with numerous features. In education field, neural networks can be used to predict student performance, identify at-risk students, and thus provide personalised recommendations for learning. The ability of neural networks to capture non-linear relationships and interactions among features makes them very useful to work with complex educational data. However neural networks require large amounts of data and computation resources, also the models are challenging to interpret when compared to simpler algorithms. Another advantage of neural networks is that they can adapt to changing input without having to redesign the output criteria and give the best possible result [29].

Every neural network has at least three layers, an input layer, a hidden layer and an output layer. The generalizability of these networks make them well suited for every type of classification problems, with all kinds of data [43]. The neural network can work also for various regression problems as well with higher accuracy than linear regression, decision trees, etc. The simplest form of all neural network is feed forward. The most basic one of feed forward neural network has one input layer, one hidden layer and one output layer. And the processing of the data in this network only occurs in one way [43]. Although, the data can travel through a number of hidden layers. The structure of a neural network is illustrated in Figure 3.6.

The output of the neuron is represented as [43] :

$$z = \sum_{i=1}^{n} (a_i \cdot w_i) + b \tag{3.7}$$

where z is the output of the neuron, n represents the number of neurons in the previous layer, a is the input vector, w is the weight vector, and b is the bias.

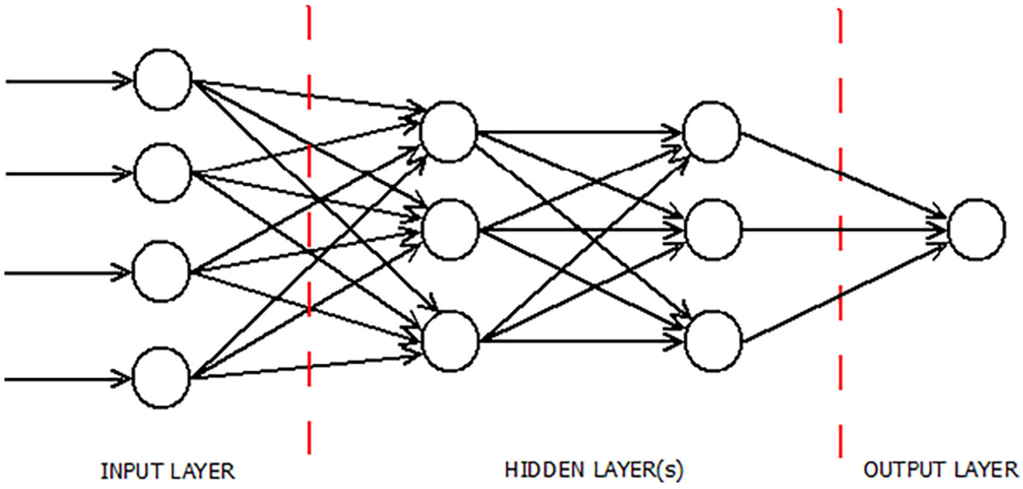

Figure 3.6.: Structure of a Neural Network Model [21]

In summary, each of these machine learning algorithms has its advantages and disadvantages. Decision trees and random forest are easy to interpret and handles non-linear relationships well, but they can overfit without proper tuning. Whereas logistic regression provides clear probabilistic interpretations and it is straightforward to implement. But it may fail to capture the complex patterns in the data. SVMs excel in high-dimensional spaces and are very effective for complex classification tasks but can be also resource-intensive. Neural networks offer powerful modeling capabilities for complex dataset, but requires substantial data and computational resources and can be less interpretable as well. By leveraging these algorithms in education field, predictive analytics can provide valuable insights on the student performance and help the educators in a timely manner to implement better solutions.

3.4. Handling Imbalanced Data

Imbalanced learning is one of the most formidable challenges within the realm of machine learning and data mining. Despite having continuous researches and advancements over the past decades, learning from data with imbalanced classification distribution remains a compelling research area [45]. Imbalanced data refers to data that has non-equal distribution of classes [3]. Handling imbalanced data is a critical aspect of predictive analytics, particularly in educational datasets where instances of at-risk students are often much fewer than those of the instances of successful students. Imbalanced datasets can lead to biased models which performs well on the majority class but performs poorly on the minority class. There are various techniques that have been developed to address the imbalance data issue, to ensure that machine learning models can effectively identify and support at-risk students.

There are many techniques that has been proposed to handle the problem of imbalance dataset. Those techniques are classified mainly into two, data-level and algorithm level. The aim of data-level techniques is to resample the datasets by increasing the frequency of samples or decreasing the frequency of samples in the classes. These techniques are simple and effective in handling imbalanced dataset problem. This resampling is mainly divided into three approaches, oversampling, undersampling and hybrid methods. While the undersampling approach is about eliminating some majority class sample to match the number of samples in the minority class, oversampling approach is the opposite where new minority class samples are synthesized to match the number of samples in the majority class [3]. Whereas the hybrid approach achieves balanced class distribution by combining these two methods. An algorithm-level technique, also known as ensemble-based classifier, focuses on improving the classifiers and not the datasets. The main focus of these techniques is to adapt a special strategy to merge many different classifiers from one original dataset into one classifier, and then aggregate the results of this classification. The algorithm-level approach has been extensively used to handle imbalance dataset problem and there are also several approaches that has been proposed to build the ensemble classifiers [3].

3.4.1. SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE is a popular method for addressing the problem of imbalanced datasets by generating synthetic examples for the minority class. Unlike simply duplicating the existing minority instances, SMOTE creates new synthetic points by interpolating between existing minority instances. This is achieved by selecting pairs of nearest neighbor minority class samples and generating new synthetic samples along the line segments that joins these pairs. This methods introduces more variety into the minority class without just replicating existing data, which helps prevent overfitting [38].

The process of SMOTE involves identifying the k-nearest neighbors first for each instances in the minority class. For each minority class instance, synthetic samples are generated by selecting one or more of its k-nearest neighbors and creates new instances that are randomly located along the line segment joining the instance and its selected neighbors. This results in a more even distribution of the minority class across the feature space, that helps the classifier to better learn the decision boundary between the minority and majority classes. The primary advantage of SMOTE is that it addresses class imbalance by effectively increasing the diversity of the training dataset, which leads to an improved generalization and robustness of the machine learning models.

3.4.2. ADASYN (Adaptive Synthetic Sampling)

ADASYN is an extension of SMOTE that generates synthetic data in a more targeted manner. The key idea behind ADASYN is to adaptively generate more synthetic data for the instances that are in the minority class which are harder to learn. This is done by calculating the density distribution of the minority class samples and generating more synthetic samples for those instances that are in low-density regions or are difficult to classify by the learning algorithms [28].

The process of ADASYN involves several steps. Firstly, the degree of difficulty or difficulty ratio is calculated for each minority class instance based on its k-nearest neighbors. Instances which have more majority class neighbors are considered to be harder to learn. ADASYN then generates synthetic samples in proportion to the level of difficulty, with more synthetic samples being created for the harder to learn instances. This adaptive approach ensures that the classifier pays more attention to the challenging instances, which helps in improving the overall ability to distinguish between the classes. By focusing on difficult instances, ADASYN helps to balance the dataset more effectively and enhance the model's performance on the minority class [28].

3.4.3. SMOTE Tomek

SMOTE combined with Tomek links is a hybrid method introduced first by Batista, et al.(2003) designed to improve the quality of the synthetic data generated by SMOTE and to clean the dataset by removing the borderline instances. Tomek links are pairs of instances where each instance is the nearest neighbor of the other but belongs to different classes. The removal of these Tomek links helps to clarify the decision boundary between classes, thus reducing the overlap and potential noise in the data [44].

The process starts with applying SMOTE to generate synthetic minority class instances. After which Tomek links are identified and removed from the dataset. This not just eliminates noisy or overlapping instances, but also enhances the class separability. By cleaning the dataset in this manner, the combined SMOTE with Tomek approach leads to a more balanced and less noisy training dataset, which thus can improve the performance and the robustness of the machine learning models. This hybrid method is particularly useful for datasets with high overlap between classes, since it helps to refine the decision boundary and reduce misclassification errors [44].

3.4.4. SMOTE ENN

SMOTE ENN is another hybrid technique which combines SMOTE with Edited Nearest Neighbors (ENN) to handle imbalanced datasets. After generating synthetic samples using SMOTE, ENN is applied to clean the dataset by removing the noisy and misclassified instances. ENN works by examining each instances and its k-nearest neighbors, and then removing instances that do not agree with the majority class of their neighbors. Integrating ENN with oversampled data by SMOTE helps in extensive data cleaning which results in a more clear and concise class separation [38].

The integration of ENN with SMOTE provides a two step approach for solving the imbalanced dataset problem. First the dataset is balanced through the synthetic oversampling and then it is cleaned to remove noise. This step ensures that the training dataset is not only balanced, but is also more accurate and representative of the true class distributions. SMOTE ENN is particularly effective in improving the quality of the dataset by addressing the imbalance and noise that helps in better model performance and classification [38].

3.4.5. Borderline SMOTE

Borderline SMOTE is a variant of SMOTE that aims on generating synthetic samples specifically for the minority class instances, which are close to the decision boundary. These instances which are known as borderline instances are more likely to be misclassified. But focusing on these critical points, Borderline SMOTE aims to strengthen the classifier's ability to distinguish between the classes near the decision boundary [22].

The technique involves identifying the borderline instances of the minority class by examining the nearest neighbors. Once that process of identifying the instances are done, synthetic samples are generated in their vicinity. This targeted approach ensures that the synthetic data is generated where it was most needed, which improves the classifier's sensitivity and reduces the likelihood of misclassification. Borderline SMOTE is particularly useful for the classifier's performance enhancement on the instances that are difficult to classify, thereby improving the overall accuracy and robustness.

3.4.6. SVM SMOTE

SVM SMOTE combines the principles of SMOTE with the power of Support Vector Machine (SVM) and focuses on increasing the minority points along the decision boundaries. This technique uses SVM to identify the support vectors of the minority class, which are the critical points lying on the edge of the class boundary. These support vectors are critical because they define the decision boundary between the classes. Then SMOTE is applied to these support vectors to generate synthetic samples which ensures that the synthetic data focuses on the most informative and challenging instances [12].

By aiming the support vectors, SVM SMOTE enhances the classifier's ability to learn the decision boundary more accurately. This method is particularly effective for datasets with complex, non-linear boundaries since it uses the strengths of SVM in identifying crucial instances and the strengths of SMOTE in generating synthetic samples. SVM SMOTE helps in creating an informative and balanced training dataset which leads to improved classification performance and robustness.

In summary, all these advanced techniques for handling imbalanced data are essential for improving the performance of the machine learning models that are for educational predictive analytics. By ensuring a more balanced representation of classes, these methods helps in building robust models which can effectively identify at-risk students and give proper support to them and thereby enhancing educational outcomes.

3.5. Evaluation Metrics

Evaluating the performance of the machine learning models is significant to ensure the effectiveness and reliability of the models, specifically in the context of predictive analytics for educational data. Different metrics provide insights into the various aspects of the performance of the model, which helps in understanding how well the model can predict and where can be improvements made. This section will cover several key evaluation metrics that are commonly used to evaluate the classification and regression models, such as accuracy, precision, recall, F1 score, AUC-ROC, confusion matrix, Mean Squared Error(MSE).

3.5.1. Accuracy

Accuracy is the most straightforward evaluation metric which represents the proportion of correctly classified instances out of the total instances. In general, accuracy is the measure of the ratio of correct predictions over the total number of instances evaluated [24], which is calculated as :

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \tag{3.8}$$

While accuracy is a useful evaluation metric, it can also be misleading when dealing with the imbalanced datasets, where the majority class vastly outnumbers the minority class, a high accuracy score can still be achieved by simply predicting the majority class for all instances. As an example, if in a dataset 90 percent of the students are not at risk and only 10 percent are at risk, then predicting the students who are not at risk would give a 90 percent accuracy but will fail to identify the at-risk students. Therefore, accuracy must be considered along with other metrics to get a comprehensive understanding on the model performance.

3.5.2. Precision

Precision is also known as positive predictive value, which is used for measuring the proportion of true positive predictions out of all positive predictions. In other words, precision is used to measure the positive patterns which are correctly predicted from the total predicted patterns in a positive class [24]. This is an important metric when the cost of false positives are high. Precision is calculated as :

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.9}$$

The high precision of the model indicates that the model has a low false positive rate, which means it is reliable when it predicts a positive class. In educational domain, high precision is very critical when interventions are very time consuming or costly, which ensures that the resources are allocated to the students who needs it the most. For instance, if a model has high precision, it implies that most of the students who are identified as at-risk are really in need of intervention, which is significant when the resources are limited.

3.5.3. Recall

Recall, which is also known as sensitivity or true positive rate, is used for measuring the proportion of actual positives which are correctly identified by the model. Or in simple words, recall is used to measure the fraction of positive patterns which are correctly classified [24] and is particularly significant when the cost of the false negatives is high, such as missing at-risk students. Recall is calculated as :

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.10}$$

High recall of the model indicates that the model effectively identifies most of the positive instances, which ensures that the at-risk students are not overlooked and receives the needed support. For example, in educational domain, a high recall means that the model was successful in identifying most students who are actually at-risk, which thereby allows timely interventions.

3.5.4. F1 Score

The F1 score is the harmonic mean of precision and recall [24], thus providing a single metric which balances both concerns. This metric is particularly useful when there is a necessity to find an optimal balance between precision and recall. The F1 score is calculated as :

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.11}$$

The F1 score is beneficial for situations where both the false positives and false negatives carries significant costs, like as in the educational context with predictive analytics where both misidentifying and missing at-risk students can have serious consequences. High F1 score indicates that there is good balance between precision and recall, which means the model is effective at identifying positive cases without having too many false positives.

3.5.5. AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

The AUC-ROC curve is a performance measurement used in classification problems at different threshold settings. Unlike the threshold and probability metrics, the AUC value indicates the classifier's overall ranking performance [24]. The ROC curve is a graphical representation of the true positive rate or recall against the false positive rate. The AUC represents the degree of separability the model has achieved, which indicates how well it distinguishes between the classes. An AUC of value 1 implies the model is perfect, while an AUC of 0.5 suggests that the model has no discrimination capability, similar to random guessing.

For a two class problem, the AUC value can be calculated as [24] :

$$\text{AUC} = \frac{S_p - n_p\left(n_n + 1\right)/2}{n_p n_n} \tag{3.12}$$

Where, S_p is the sum of all positive examples that are ranked, and n_p and n_n represents the number of positive and negative examples respectively [24]. The AUC was theoretically and empirically proven to be better than the accuracy metric in evaluating the performance of the classifier as well as in discriminating an optimal solution during the classification training [24].

A high value for AUC implies that the model performs well across various threshold levels, thus providing a comprehensive measure of performance. This metric is specifically used for model comparison and for selecting the best one for a given task. In educational domain, a high value for AUC would imply that the model can discriminate effectively between at-risk students to those not at risk, which can lead to more accurate interventions.

3.5.6. Confusion Matrix

A confusion matrix is used for providing a detailed breakdown on the performance of the model by showing the actual versus predicted classifications. It includes four key metrics which are:

1. **True Positives (TP) :** The data point in the confusion matrix is said to be true positive, when a positive outcome is predicted and it matches the actual value [46].
2. **True Negatives (TN) :** The data point in the confusion matrix is said to be true negative, when a negative outcome is predicted and the actual value is also the same [46].
3. **False Positives (FP) :** The data point in the confusion matrix is said to be false positive, when a positive outcome is predicted but the actual value is negative. This scenario is also known as Type 1 Error [46].
4. **False Negatives (FN) :** The data point in the confusion matrix is said to be false negative, when a negative outcome is predicted but in actual the outcome is positive. This scenario is also known as Type 2 Error [46].

The confusion matrix helps in visualizing the types of error that are made by the model, which helps in gaining more insights into the areas where the models needs improvement. Table 3.1 shows the confusion matrix for the binary classification.

The confusion matrix is specifically useful for understanding the performance of the model in a detailed manner, especially for imbalanced datasets. This helps in the calculation of various metric such as accuracy, precision, recall and F1 score, thus providing a comprehensive view of the model performance. By leveraging confusion matrix and analyzing it, educators can understand the correct balance between the at-risk students that are identified and the instances where the model misclassifies the students, thereby helping to further refine and improve the model.

Class designation		Actual class	
		True (1)	False (0)
Predicted class	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table 3.1.: Confusion Matrix for the Binary Classification Problem [46].

3.5.7. Mean Squared Error (MSE)

Mean Squared Error is a common evaluation metric for the regression tasks, which measures the average of the squares of the errors, that is the average squared difference between the estimated values and the actual value. In simple words, MSE measures the difference between the predicted solutions and desired solutions [24]. MSE is calculated as :

$$\text{MSE} = \frac{1}{n}\sum_{j=1}^{n}(P_j - A_j)^2 \tag{3.13}$$

where P_j is the predicted value of instance j, A_j is the real target value of instance j, and n is the total number of instances [24].

MSE gives large errors a higher weight, which makes it specifically sensitive to outliers. Whereas, a lower MSE implies a better fit of the model to the data. In educational predictive analytics, MSE can be used for evaluating models that can predict continuous outcomes, like grades of the students or scores. By analyzing the average squared error, MSE provides a clear measure of the accuracy of the model in predicting these outcomes. Models that are generally preferred are those that has lower MSE, since they indicate more precise predictions, which is critical for making informed decisions and educational interventions.

4. Methodology

This chapter focuses on the methodologies employed to develop, evaluate and validate the machine learning models that are aimed for identifying at-risk students within a higher education context.

4.1. Python As The Programming Foundation

Python serves as the foundational programming language for this study. Python is known for its versatility and an extensive library, which makes it exceptionally well-suited for handling complex data-driven tasks in educational research.

4.1.1. Data Handling Capabilities

- **Data Extraction and Storage**: Python excels in interfacing with several different data formats. Its libraries like Pandas are crucial for efficient data manipulations, which helps in facilitating the handling of large datasets typical in educational settings.
- **Data Pre-processing**: Pandas and NumPy, which are the libraries of python are significant in cleaning, transforming and normalizing data to ensure that the high standards required for accurate modeling are met. This pre-processing step is very crucial in maintaining the integrity of data and the readiness for analysis.

4.1.2. Data Visualization

- **Matplotlib and Plotly**: These libraries are used to create visualizations in detail to provide more insight on the dataset, that helps in the exploratory data analysis phase. Effective visualization helps in presenting distributions and underlying patterns in the data which are crucial for initial explorations of the data and evaluation of the model in depth.

4.1.3. Scikit-learn for Machine Learning Model Development

- **Comprehensive Use of Scikit-learn in Modeling**: Scikit-learn library is extensively used during the entire phase of modeling process, since its robust suite of tools and algorithms supports several machine learning tasks such as classification, regression and clustering. This library is essential in developing predictive models for identifying at-risk students.
- **Integration of Pipeline**: Scikit-learn library helps to build pipelines, thus enhancing the modeling process. Pipelines are used to streamline the workflow, from data pre-processing step to model fitting and evaluation. This integration of pipeline helps in ensuring consistency and efficiency in executing every step of the model development process.
- **Cross-validation and Evaluation Metrics**: Scikit-learn also has the functionality of performing cross-validation which helps in ensuring the reliability and generalizability of the model. Additionally, the effectiveness of the models can be assessed by employing the comprehensive metrics offered by this library, such as accuracy, precision, recall and F1-score.

By leveraging Python and the vast libraries it offers, this study aims to establish a solid framework for tackling the challenges in predicting educational outcomes. This approach not only helps in boosting the precision of the predictive models, but also helps in ensuring the reproducibility and scalability of the analysis, thus promoting a broader application and also facilitating future researches in the field.

4.2. Methodology - Steps

The methodology followed for this thesis is shown in Figure 4.1 and the steps in this workflow are explained in the following subsections.

4.2.1. Understanding the Requirements and Research (Literature Review)

The initial phase of this study involved a comprehensive process of defining the project's goals and understanding the existing landscape of predictive analytics in the field of education. This stage was crucial in setting a clear direction for this research and for ensuring that the methods developed were all well informed.

The study started by identifying the primary objective, which is to develop predictive models capable of identifying at-risk students at an early stage in their academic journey. An extensive review of existing literature was conducted to achieve this. The review was done by analyzing published research papers and articles which

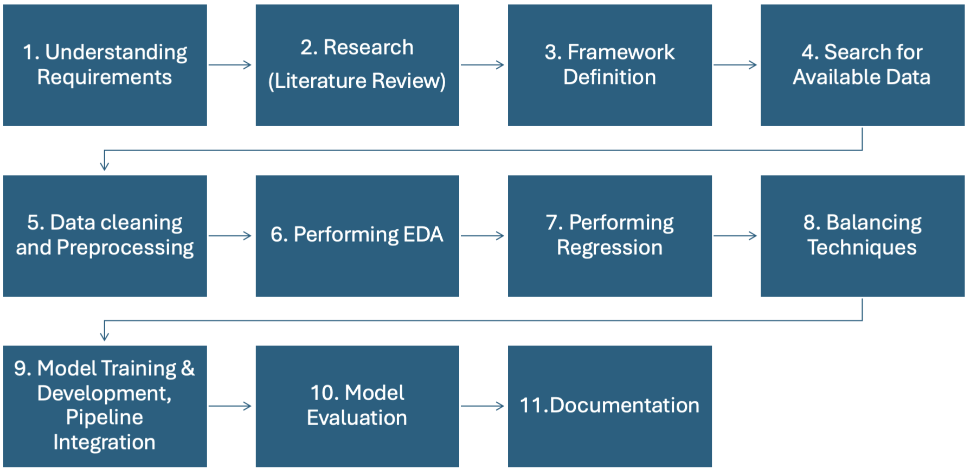
1. Understanding Requirements
2. Research (Literature Review)
3. Framework Definition
4. Search for Available Data
5. Data cleaning and Preprocessing
6. Performing EDA
7. Performing Regression
8. Balancing Techniques
9. Model Training & Development, Pipeline Integration
10. Model Evaluation
11.Documentation

Figure 4.1.: Workflow

have explored similar concepts within the domain of data analytics. The aim was to not only gather existing knowledge, but also to identify the gaps where the thesis can contribute new insights.

As the research progressed, the literature review remained to be an ongoing activity, that continuously informed and refined the research approach. This iterative process ensured that each phase of this thesis was built on a solid foundation of existing knowledge by remaining adaptive to new findings and techniques which were discovered along the way. This dynamic approach has helped in developing a structured methodology for this study, which led to a more accurate and reliable model development.

This foundational phase was very crucial for providing with the necessary insights for effectively tackling the complex challenge of predicting academic risk. By understanding the requirement and the state of the research in predictive analytics in educational domain thoroughly, this thesis was certain to develop solution which are both innovative and directly responsive to the educational institution needs.

The detailed and reflective approach to understanding the requirements and doing an extensive literature review has laid a strong groundwork for the next phases of the thesis, focused on creating effective predictive models for using in the educational realm.

4.2.2. Framework Definition

This phase of the thesis functions as a blueprint for the entire research project. It outlines both the theoretical and operational structure within which the predictive models were developed and tested. This stage is crucial because it creates the guidelines and standards which are required to ensure that the research is methodologically sound and is aligned with academic and practical objectives.

The theoretical groundwork of this thesis is rooted in the principles of data science, educational psychology and predictive analytics. This approach helps in a comprehensive understanding of the factors which influence the student performance and outcomes. The review of existing literature in the previous phase plays a significant role in providing the theoretical justifications for the chosen methodologies and technologies. Thus, this framework helps in guiding the selection of variables, model development, interpretation of results, and ensures that the research remains focused and relevant to the field of education.

The thesis is structured around a series of steps that are designed to effectively develop and validate predictive models. First one is data collection, where types and source of required data are outlined, ensuring that they contain a broad spectrum of variables which can impact student success. Then comes data pre-processing, where the selected datasets are subjected to cleaning, to handle missing values and other transformations if necessary to prepare the dataset for analysis. This step is critical for ensuring model accuracy and performance. Then comes data analysis step, where the cleaned dataset are subjected to analysis to uncover patterns, anomalies, relationships and trends. Exploratory Data Analysis (EDA) is very important to understand the underlying structure of the data that will help in making choice of modeling techniques and other strategies. Several plots are employed to gain deep insights into the data.

A key component in the development of the framework of this thesis is the integration of an end-to-end-processing pipeline, to ensure the consistency, efficiency and reproducibility of the research. This pipeline incorporates all critical stages such as scaling the data which ensures that all the features contribute equally to the performance of the model, thus preventing any single feature from dominating due to scale differences, applying balancing techniques, to solve the imbalanced data problem to enhance the model's ability to generalize from the training data and improve its predictive accuracy on minority class, model development and evaluation, where various machine learning models are developed and evaluated. Each model is configured and optimized within the pipeline to identify best the at-risk students. The last and crucial component of the framework is documentation and standardization throughout the research process, in which every step is thoroughly documented. This ensures transparency and reproducibility and facilitates the potential adaptation of these methodologies to other educational settings.

By defining this framework carefully, this thesis lays a solid foundation for a systematic exploration of predictive analytics in the educational domain. This framework intends to provide a clear road map for achieving the research objectives, along with ensuring that the results are robust, reliable and is relevant for the educators to provide timely interventions for improving student outcomes. This structured approach is not only helping to enhance the quality of the research, but also contributes significantly to the educational field by offering a replicable model for future studies.

4.2.3. Datasets

The datasets which are selected for this thesis are integral for developing robust predictive models for identifying students who are at-risk across various educational contexts. These datasets are sourced from reputable platforms, each of them providing unique insights into the dynamics of student performance and academic success.

Dataset 1

The primary dataset of this thesis, is taken from Kaggle, which is provided by Zenodo [33]. The dataset "Predict student dropout and academic success" was created from a higher education institution in 2021 related to students who were enrolled in different undergraduate degrees and these data were acquired from several disjoint databases [33]. The data provides a holistic view of the student's academic journey. The data includes essential information known at the time of a student's enrollment, such as demographics, socioeconomic factors, marital status etc. In addition to that, it tracks the performance of the students across the first and second semesters offering valuable insights on their academic progress and outcomes. The dataset consists of 4,424 records, each with 35 attributes. All features of this dataset are listed in the Appendix. These attributes cover a wide range of data points such as demographic details, academic paths and performance metrics.

This dataset is specifically designed for analyzing and predicting the student performance metrics on semester basis. This detailed tracking makes it a powerful tool for identifying potential dropouts along with other academic challenges at an early stage in the student's university life. Rigorous data cleaning and pre-processing were already undertaken in the available data to ensure the quality and consistency of the dataset. This includes standardizing the data formats, handling missing values, and also correcting any inconsistencies for preparing the dataset for effective and reliable analysis.

Dataset 2

This dataset from the UCI Machine Learning Repository provides an extensive overview of academic and personal factors affecting the student performance in two Portuguese schools [15]. The dataset includes a wide range of variables, such as student demographics, family background, social habits and school related features alongside academic grades. The dataset includes 649 records, each detailed with 33 different attributes. The list of attributes is given in the Appendix. This comprehensive collection helps in the robust analysis of the factors that influence academic success and risks.

The diversity of this dataset makes it invaluable for testing the applicability of the models across different educational environments and cultural contexts, for ensuring that the models are versatile and effective in diverse educational settings. The UCI dataset is a reputed source of educational data which provides a broad overview of student academic performance, making it ideal for comparative and generalizability studies.

Dataset 3

The Module Uni Assessment Dataset is provided by the second supervisor of this thesis from a Computer Science semester module, containing detailed academic assessment data and is not an open dataset and hence not available to the public. This dataset includes granular details about student assessments and outcomes specific to the curriculum of the university. The dataset contains 109 records, each with 12 attributes. The attributes includes, the id of the students, assignment grades of 10 assignments with values ranging from 0 to 100 and the final grade, which is also in the range of 0 to 100.

This data is instrumental for developing predictive models which are tailored according to the academic structure and challenges of the university, which helps in personalized student support and intervention strategies. The data is handled with utmost confidentiality, and is used under strict ethical guidelines, ensuring compliance with data protection regulations and respecting the privacy of students.

These datasets together provide a comprehensive foundation for the thesis, which allows an in-depth exploration of the factors that influence academic success and risks across various educational settings. Each dataset has been made available after undergoing rigorous pre-processing to ensure that the data used for analysis is of highest quality, thus enhancing the predictive power and reliability of the developed models. This approach ensures that the research is grounded in accurate data, thus providing reliable insights which can be helpful for educators to make informed educational strategies and interventions.

4.2.4. Data Handling - Data Cleaning and Pre-processing

Data cleaning and pre-processing is the backbone of any data-driven analysis, particularly in the context of research, where the accuracy and reliability of the results are of greatest importance. This thesis uses datasets that have already undergone comprehensive data cleaning and pre-processing and only basic categorical encoding was done in this phase of research after getting the datasets.

Data cleaning and pre-processing is very crucial for ensuring quality and consistency of the data. The primary aim of these initial steps is to ensure that the datasets are not prone to the common issues which can skew the analysis or predictive modeling. These issues include data format inconsistencies, missing values, outliers, which can significantly affect the outcomes of the statistical models.

Another crucial step in this phase is to standardize the data formats and integrating datasets from multiple sources. This is mainly done to ensure the seamless handling of the data and prevents issues that arise from dataset discrepancies, especially when datasets come from different sources with various collection methodologies.

Handling missing data is another step which is addressed through imputation techniques that are tailored to the nature of the missing values and type of the data. Mean or median imputation methods are often used for continuous variables, while for categorical data, mode imputation method or more sophisticated predictive imputation methods might be used depending on the overall dataset.

Z-score or IQR (Interquartile Range) techniques are used to identify outliers. Each outlier is evaluated to understand whether it represents a true anomaly or an error while entering the data. This technique involves adjusting the values or removing the outlier records together, based on their impact on the overall dataset.

Advanced pre-processing techniques includes feature engineering, normalization and scaling which are crucial for preparing the data for predictive modeling. Feature engineering involves creating new variable from existing dataset for capturing the underlying patterns better and to provide the models with information which are not readily apparent in the raw data, like transforming raw test scores into categorical performance levels. Normalization and scaling techniques are used to adjust the range and distribution of the variable scales to ensure that there would be no single attribute that unduly influences the outcome of the model. While normalization technique typically scales data to a range of 0 to 1, standardization transforms data to have a mean of 0 and a standard deviation of 1.

All the datasets that we use in this thesis had already undergone extensive data pre-processing before being available to the public for using it for further research. The data types and formats were aligned after being collected from multiple educational databases. For the data that is not open to public, pre-processing focused on anonymizing student identifiers and standardizing grading scales across different modules to create a uniform dataset for analysis. Other pre-processing efforts involved in the preparation of data were to correct data entry errors and normalizing grades to account for the different scoring systems which was collected from different sources.

By having the data cleaning and pre-processing phase in the methodology, this thesis focuses on ensuring that the data used in the subsequent analyses is of the highest quality. This significant approach is not only helpful to enhance the accuracy of the predictive models, but also helps in ensuring the finding of this study are robust, reliable and are capable for supporting the educators to make significant interventions.

4.2.5. Performing Exploratory Data Analysis (EDA)

Another crucial step in data-driven projects is Exploratory Data Analysis (EDA), especially in the context of educational research where understanding the underlying factors that influence student outcomes is of utmost importance. EDA helps to uncover insights that not only helps in forming the modeling strategy but also help in guiding critical educational interventions. For this thesis study, EDA is conducted separately for each of the three datasets.

Data 1 - Predict Student Dropout and Academic Success

Target Variable Distribution Figure 4.2 is a count plot that provides a visual representation of the frequency of each class in the target variable, highlighting any imbalances. It is very important to understand the distribution of the target for selecting appropriate balancing and modelling techniques, especially if the dataset shows a significant skew towards one class.

Distribution of 1st Semester Grade by Target Class Figure 4.3 shows a box plot used to examine the distribution of '1st semester grades' by different target classes. This visualization helps us to observe the variation in the academic performance across categories, which thus helps in potentially identifying disparities or achievements that correlate with the target classification.

Distribution of Age at Enrollment by Target Class Figure 4.4 shows a box plot used to visualize the spread and tendency of 'Age at Enrollment' across the target classes. This visualization helps in identifying whether certain age groups are more prevalent in specific target class, possibly suggesting age related trends or targeted information.

Distribution of Marital Status by Target Class Figure 4.5 is a bar chart used to identify trends and patterns that are related to marital status that might affect student outcomes. Observing the outcomes based on marital status can offer insights into the social and personal dynamics that can affect the students. For instance, married students might have different challenges and support systems when compared to single students, which help in their academic studies. Or it can also be as,

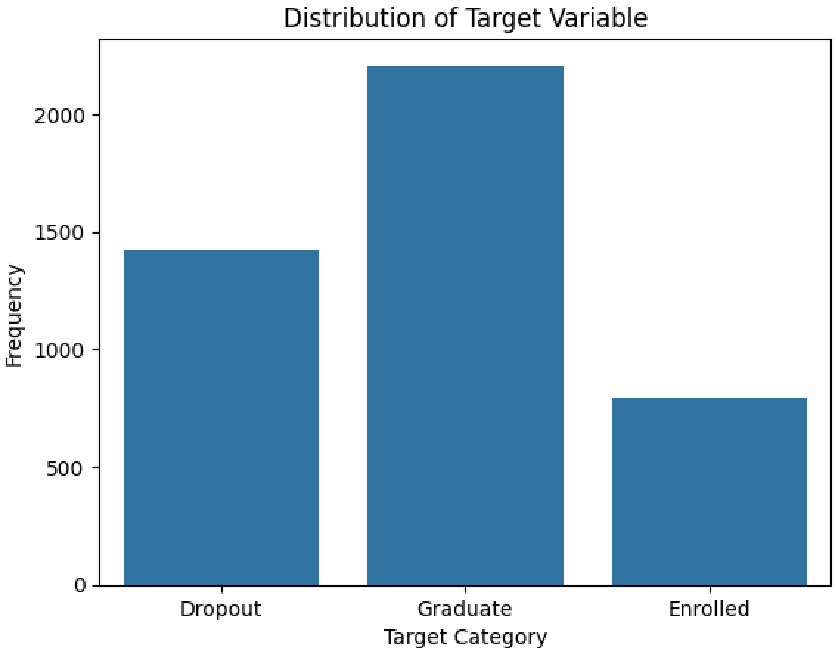

Figure 4.2.: Dataset 1: Target Distribution

the married students have more responsibilities than single students which can impact their outcomes negatively. Here in this graph, 1 represents single, 2- married, 3-widower, 4-divorced, 5-facto union, and 6-legally separated.

Distribution of Gender by Target Class Figure 4.6 is a bar chart used to assess how gender distribution varies across different target class, and thus can reveal gender related disparities or advantages in educational contexts. This visualization helps in understanding any gender biases in the academic performance, retention rates or any other educational metrics. This can be very critical in developing gender sensitive policies and support mechanisms that are aimed to provide fair educational opportunities. Here in the figure, 1 represents male and 0 represents female.

Correlation Matrix of Features When the target variable is encoded numerically, it is possible to determine how it correlates with other numerical variables. The figure 4.7 shows a correlation heatmap that offers insights into potential predictors and their relationships with the target. It also helps in identifying attributes that have a stronger linear relationship with the target variable. This analysis is important for feature selection during model building, by focusing on variables with significant correlations.

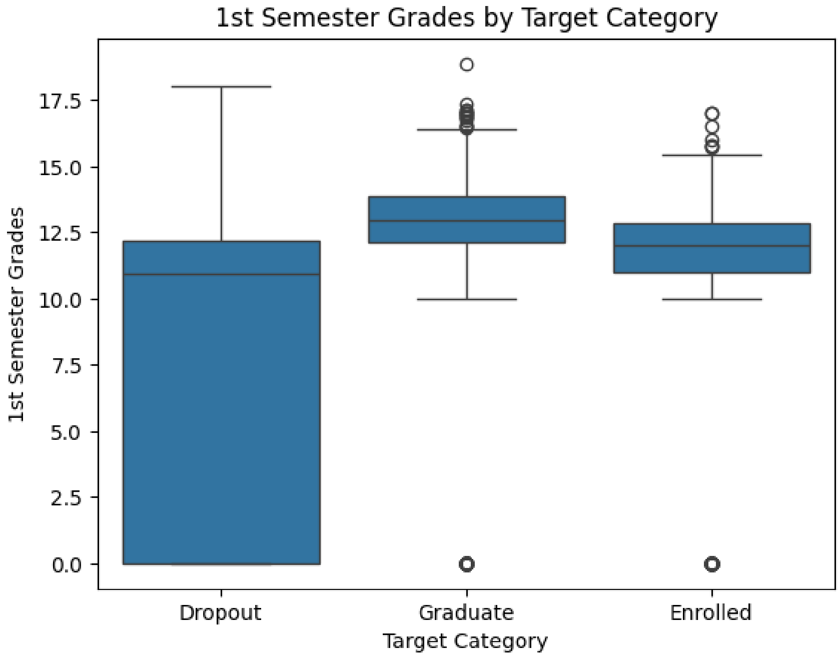

Figure 4.3.: Dataset 1: 1st Semester Grades by Target Category

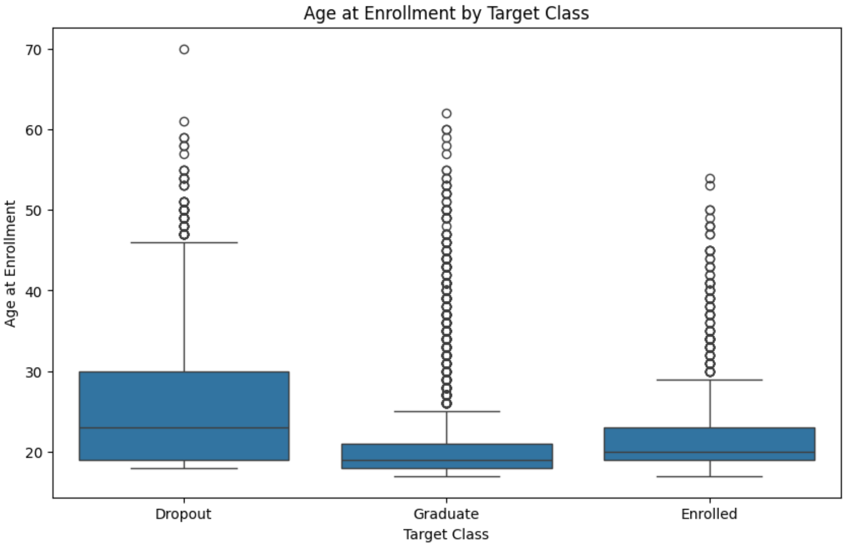

Figure 4.4.: Dataset 1: Age at Enrollment by Target Category

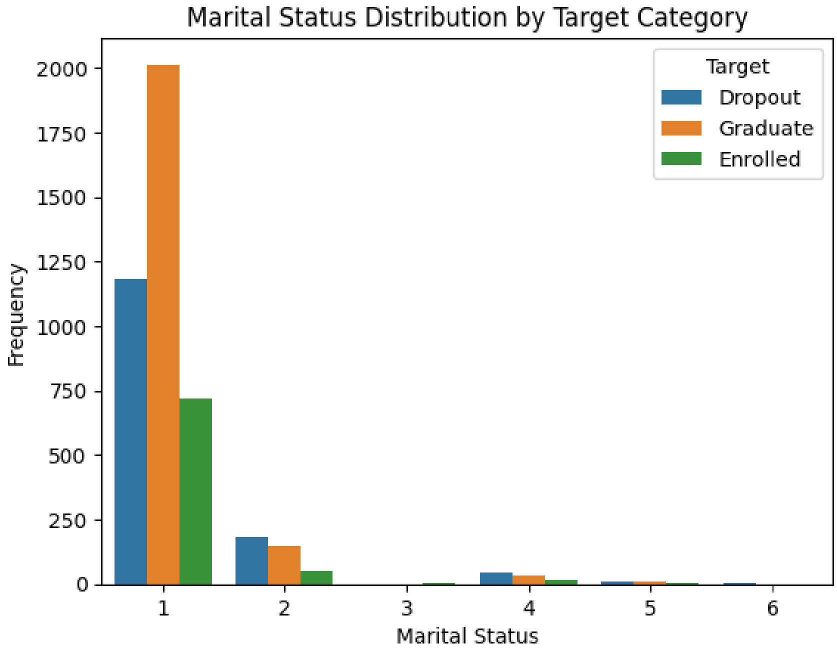

Figure 4.5.: Dataset 1: Distribution of Marital Status by Target Category

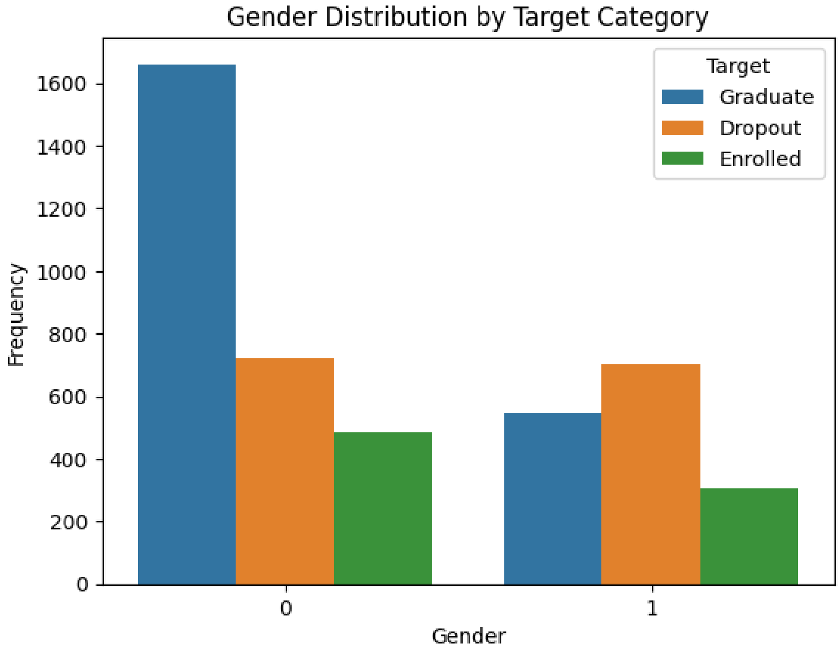

Figure 4.6.: Dataset 1: Distribution of Gender by Target Category

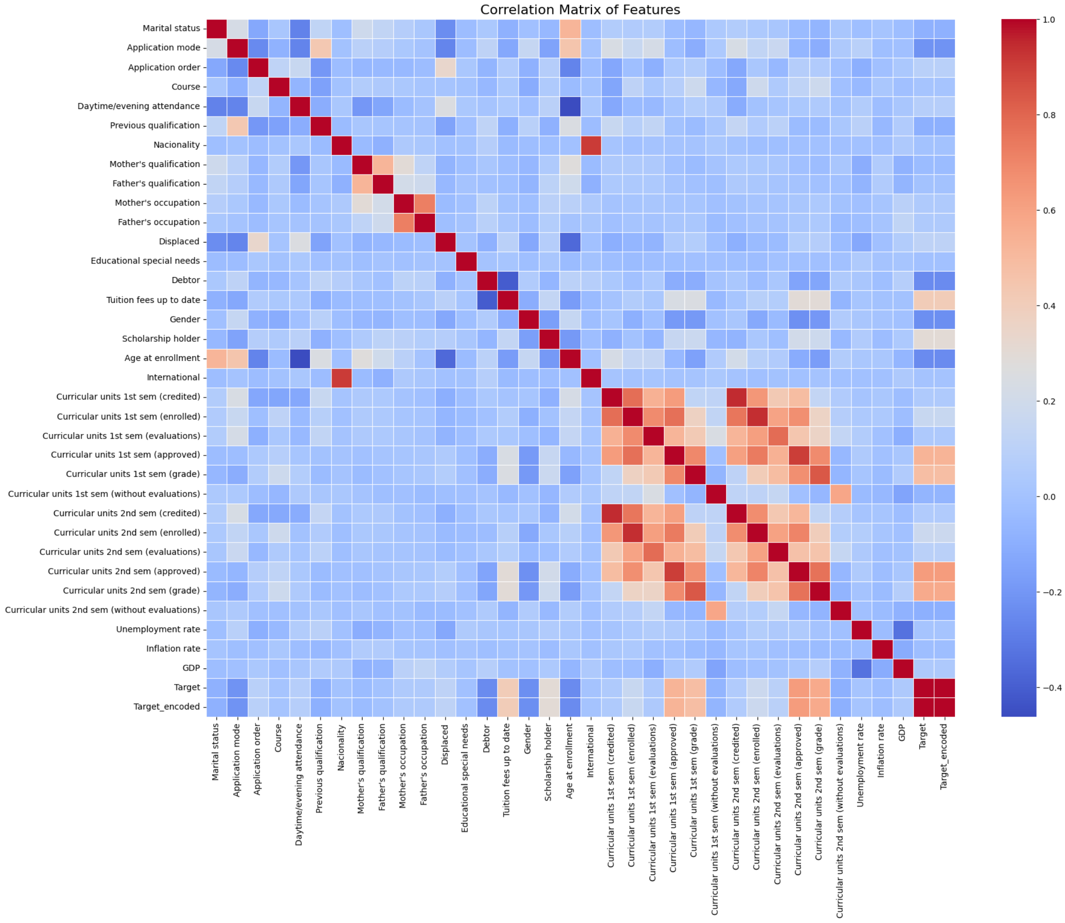

Figure 4.7.: Dataset 1: Correlation Matrix of Features

Dataset 2 - UCI Student Performance

The target variable of UCI student Performance data is created by assessing the G3 variable. If the G3 value is greater than or equal to 10, then it is considered pass or else fail [16].

Target variable Distribution Figure 4.8 shows the distribution of target variable, that is, pass or fail using a count plot, highlighting potential imbalance that could impact modeling.

Distribution of Absences by Target The box plot shown in Figure 4.9 shows the distribution of absences by the target category which helps in identifying the impact of absences in passing or failure of the student.

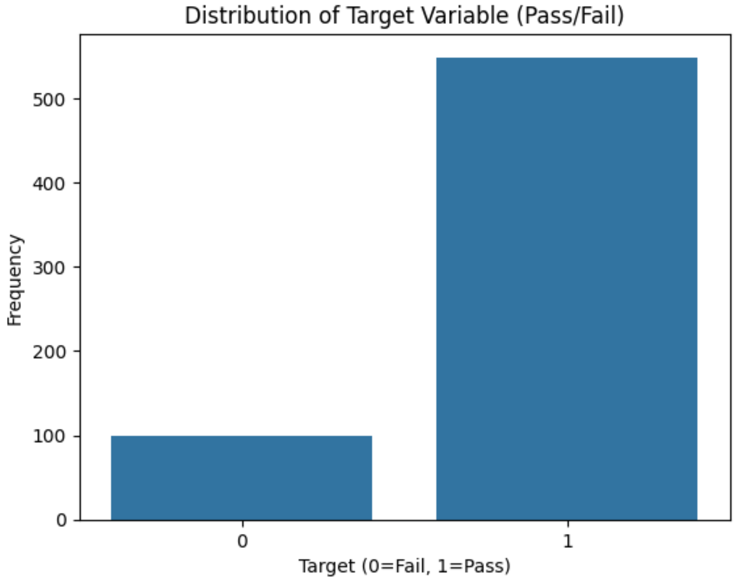

Figure 4.8.: Dataset 2: Distribution of Target Variable

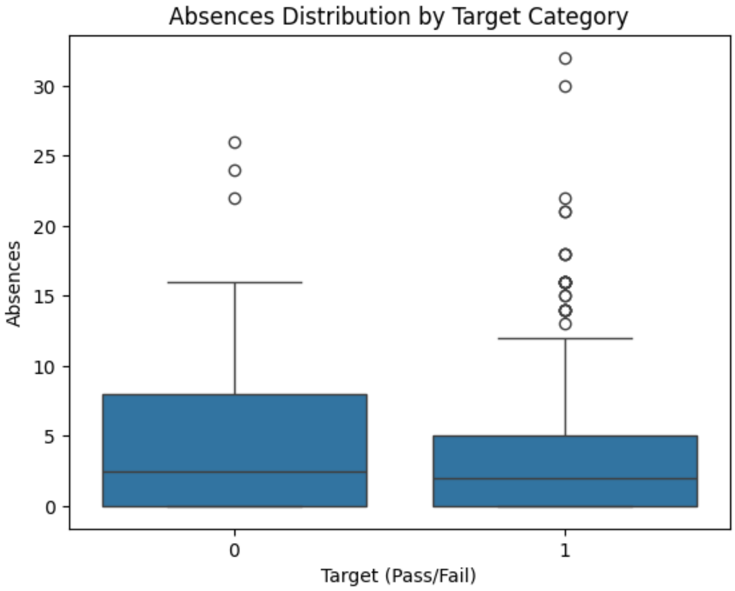

Figure 4.9.: Dataset 2: Distribution of Absences by Target Class

Study Time Categorized by Target Figure 4.10 shows a histogram that visualizes the impact of study time in the passing or failure of students.

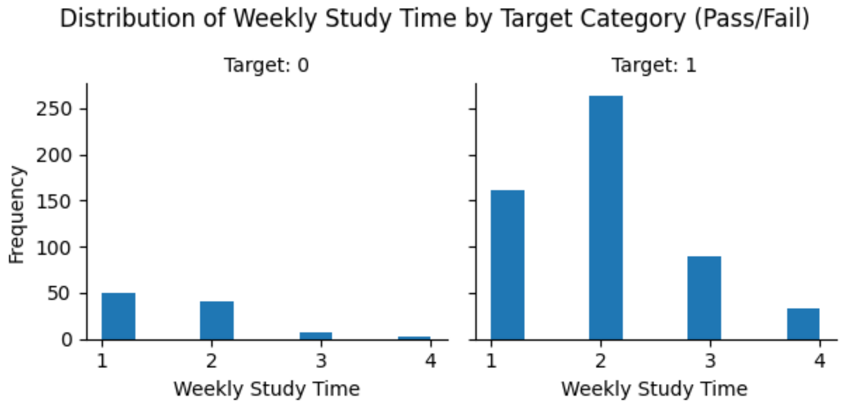

Figure 4.10.: Dataset 2: Distribution of Weekly Study Time by Target Class

Distribution of Gender by Target Figure 4.11 shows a count plot that visualizes gender distribution across the target class. This analysis will help in understanding any gender bias and ensuring timely and fair interventions for the students.

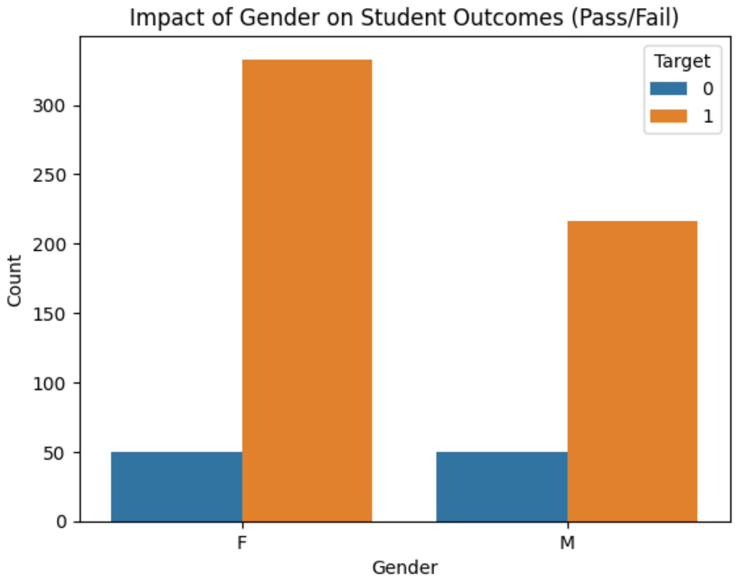

Figure 4.11.: Dataset 2: Distribution of Gender by Target Class

Correlation Matrix of Numeric Features Figure 4.12 shows the correlation matrix of the numerical attributes which will help in understanding which features are more critical to be used in the model and thus help in efficient model development and training.

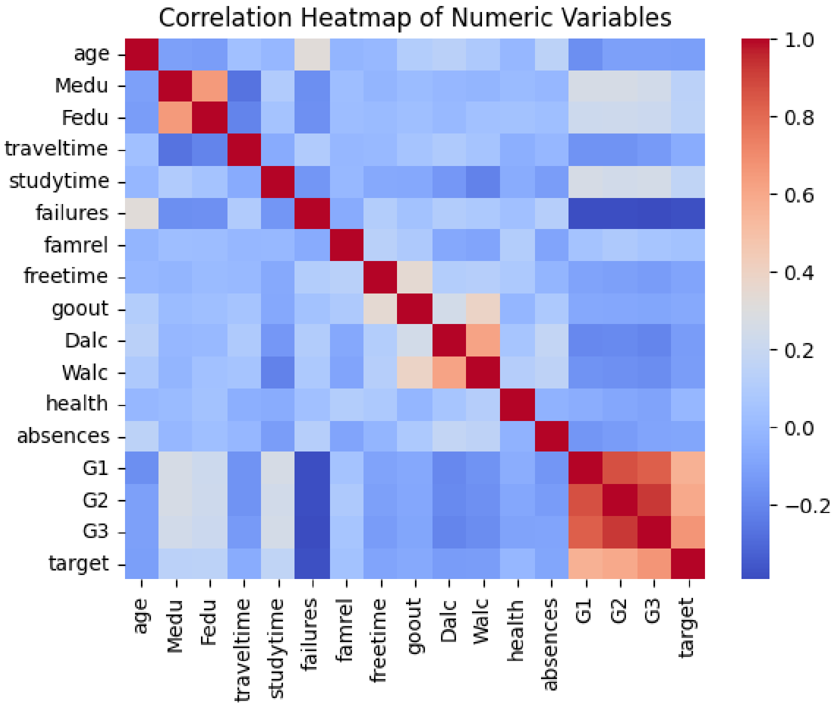

Figure 4.12.: Dataset 2: Correlation Matrix of Numerical Features

Dataset 3 - Module UNI Assessment

The analysis documented for this data is limited as it is not an open dataset.

Distribution of Target Variable The target variable distribution of this anonymized data is shown in Figure 4.13, which shows the very evident imbalance issue of the data. For this data, the threshold of passing is considered to be 50 percent and if the score of a student is 0, then it is considered as the student has not given the exam and hence those rows are removed from the dataset. The target variable is set by considering this passing threshold. Pass value is represented by 1 and fail is represented by 0.

Comparative Analysis of Pass/Fail Groups Figure 4.14 shows a comparative analysis of the target class with a bar chart that visualizes average scores for the assignments for the pass and fail class.

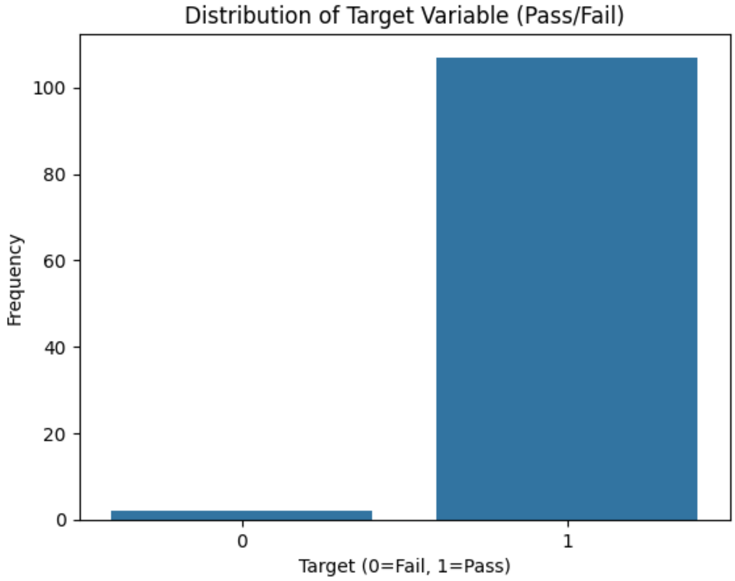

Figure 4.13.: Dataset 3: Distribution of Target Variable (Pass/Fail)

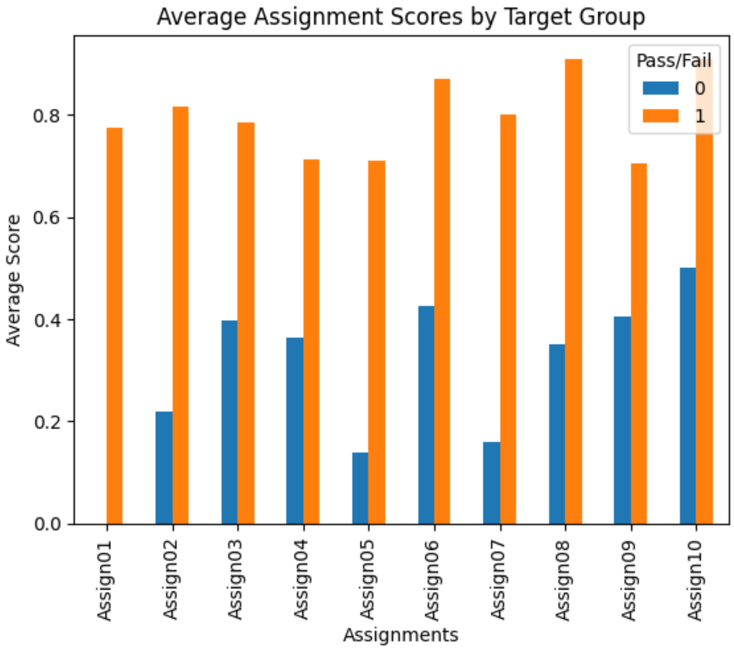

Figure 4.14.: Dataset 3: Average Assignment Scores by Target Variable (Pass/Fail)

4.2.6. Performing Regression

This thesis focuses on employing linear regression analysis for exploring the relationships between various predictor variables and a continuous outcome within the educational datasets. The aim is to quantify how well can these predictors estimate the student performance, usually measured by final grades or scores.

Linear regression can be defined as a fundamental statistical and machine learning technique that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. Linear regression is widely used because of its simplicity, which makes it a standard approach for regression tasks. The linear relationship between the variables is expressed in the form of a linear equation as :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \tag{4.1}$$

Where Y represents the dependent variable, $X_1, X_2, \ldots, X_n$ represents independent variables, β_0 is the y-intercept, $\beta_1, \beta_2, \ldots, \beta_n$ represents the coefficients, and error term is represented by ϵ.

While doing the model implementation for each of the datasets, data was split into training and testing sets to evaluate the performance of the model. The Linear-Regression class from scikit-learn is used for fitting the model on the training data. The performance of the model is evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics.

MSE measures the average of the squares of the errors, or in simple words, it is the average squared difference between the estimated values and the actual value. R^2 indicates the goodness of fit and therefore is defined as the measure of how well the unseen data are likely to be predicted by the model, by explaining the percentage of the response variable variation which is explained by a linear model. MSE provides a quantifiable measure of how much the predictions of the model deviates from the actual data points. If the value of MSE is lower, it indicates that the model fits the data more accurately. Whereas a higher R^2 value closer to 1 implies a model that explains a large portion of the variance in exam scores based on the features.

This phase of the thesis provides a comprehensive approach to using linear regression for interpreting educational data, which helps in contributing valuable insights into the factors that impacts student outcomes significantly. This analysis not only helps in enhancing academic understanding, but also in offering practical recommendations based on the quantitative evidence.

4.2.7. Balancing Techniques

Balancing techniques are essential to address the challenge of imbalanced datasets in predictive modeling, especially within the context of machine learning. Imbalanced datasets can lead to biased models which can overly favor the majority class and compromise the predictive accuracy concerning the minority class. For a more equitable representation of the datasets, balancing techniques are applied to adjust the dataset for enhancing the model performance and fairness.

The common balancing techniques are mentioned below:

Oversampling techniques This technique is used to increase the size of the minority class by duplicating the samples or generating synthetic samples. Techniques like SMOTE, ADASYN, BorderlineSMOTE are widely used for creating synthetic samples based on feature space similarities.

Undersampling Techniques This technique is applied to reduce the size of the majority class for balancing the dataset. Random undersampling and cluster centroids are some of the methods that involves removing samples from the majority class to equalize the distribution of the class.

Hybrid Techniques This technique is the combination of both oversampling and undersampling, aiming to balance the dataset by adding synthetic instances of the minority class and removing the borderline majority instances or noisy data. Some of the examples of this type of balancing are SMOTE ENN and SMOTE Tomek.

Various balancing techniques were strategically applied to the datasets in this research for evaluating their impact on model performance. For the data 'Predict Student Dropout and Academic Success' and for 'UCI Student Performance' data, balancing techniques were applied after splitting the data into training and testing datasets. This step ensured that the model is trained on a balanced representation of classes and then validated on unseen testing set that mirrors real world class distribution. Techniques like SMOTE, ADASYN, along with their hybrid forms were integrated into a machine learning pipeline using the imblearn library in python. This paved a way to dynamically balance the datasets during model training.

For the dataset 'Module UNI Assessment', a different approach was adopted where the balancing technique was applied before the data was split into training and testing datasets. Since the class distribution of this dataset was known to be extremely skewed as seen in the analysis section, and the number of samples is relatively small, this approach ensured that both the training and testing datasets are balanced. The models were then trained on the resampled training dataset and evaluated on the resampled testing dataset. This ensured consistency in the class representation across all stages of model development and evaluation.

The balancing technique implemented using imblearn.pipeline ensured that the balancing occurred with the cross validation loops during the model training step, and maintained the integrity of the validation process. This setup prevents the data from leakage and ensures that the results are reliable and indicates how well the models perform on unseen data.

The careful implementation of these balancing techniques in this study has allowed for a thorough investigation into their effect on the robustness and accuracy of the model. Balancing the datasets prior to the model training helps in mitigating the bias towards the majority class, thereby enhancing the predictive performance across all classes. This step in this research ensures that the findings are not just statistically significant but also practically relevant in the real world educational domain.

4.2.8. Model Training and Development, Pipeline Integration

A critical phase in the data science and machine learning lifecycle is model training and development. In this phase theoretical models are translated into practical applications. This phase includes steps like selecting appropriate algorithms, configuring them with the right parameters, training them on datasets to learn from patterns. The process often requires adjustments based on model performance metric until the optimal setup is achieved and hence is iterative.

Integration of pipeline in machine learning provides a structured approach for automating the flow of data through various stages of processing such as data cleaning, feature selection and normalization and applying machine learning methods. This systematic approach not only ensures that all the steps are reproducible and scalable, but also helps in minimizing errors and inconsistencies which can arise when the steps are implemented separately.

In this thesis, a comprehensive machine learning pipeline was implemented using the scikit-learn and imblearn libraries of python to streamline the process from data pre-processing to model evaluation. This integration is crucial to maintain the integrity of the data through the model training process and for ensuring that the performance metrics are measured accurately.

Before training, the data was pre-processed to ensure it is clean and suitable for modeling. This included handling missing values, encoding categorical variables. These pre-processing steps were already done earlier based on each dataset's cleaning needs. Thus only scaling the numerical features using StandardScaler from the pre-processing step was included in the pipeline to normalize the data distribution, which is a generic step that is needed for all the datasets.

Also, given the imbalance in the class distribution of the datasets, various resampling methods were employed for ensuring fair representation of all classes. Balancing techniques such as SMOTE, ADASYN, SMOTE TOMEK, SMOTE ENN, Borderline SMOTE, SVM SMOTE were applied dynamically during the model training phase within the pipeline. A variety of models were explored, such as decision trees, random forests, logistic regression, support vector machines and neural networks. Each models were selected based on its suitability for handling the specific characteristics of the dataset and for addressing research questions.

The imblearn.pipeline was used to ensure that the sampling methods and the model training steps were conducted in a controlled sequence, for preventing the data leakage between the train and test datasets. Additionally the models were evaluated using cross validation techniques for assessing the performance robustly. Metrics such as accuracy, precision, recall, F1 score and confusion matrix were computed for analyzing the effectiveness of the model in classifying the target variable correctly.

The integration of these machine learning pipeline has enhanced the efficiency and effectiveness of the research significantly, which allows for a systematic evaluation of different models and techniques. This pipeline framework has helped in ensuring that each step from data pre-processing to final evaluation was reproducible and robust, making it easy for a thorough investigation into the predictive capabilities of the models developed. This structured approach not only helped in streamlining the model development process but also in providing insights that are crucial in making informed decisions in educational settings.

4.2.9. Model Evaluation

Model evaluation is an important stage in the machine learning pipeline where the model performance is assessed to determine the effectiveness of the model in making predictions. This stage utilizes several statistical measures and tests that helps in evaluating how well the predictions of the model matches the actual data.

Overview of Model Evaluation Techniques

Model evaluation is not limited to just checking accuracy, it goes beyond that involving a series of metrics that provide insights into different aspects of the model performance.

Accuracy This metric measures the overall correctness of the model in predicting the target variable.

Precision and Recall While precision measures the correctness achieved in positive prediction, recall measures the ability of the model in identifying the relevant cases within a dataset.

F1 Score It is the harmonic mean of precision and recall, thus providing a balance between the two metrics, which is especially useful in the case of imbalanced datasets.

Confusion Matrix It is a table used for describing the performance of a classification model on a set of test data for which the true values are known.

In this study, the evaluation of machine learning models was systematically approached by integrating a variety of metrics, that reflects a comprehensive assessment of the performance of each model. Along with these, cross-validation is extensively used to ensure the robustness of the model evaluation. K-Fold cross-validation methods were applied for guaging the effectiveness of the model across different subsets of the dataset, thus mitigating any overfitting and providing a more generalized performance estimate. Models were assessed using metrics such as accuracy, precision, recall, f1 score and the confusion matrix. All these metrics were calculated by using the scikit-learn library of python, providing extensive support for model evaluation. The analysis of confusion matrix played a critical role in understanding the types of errors made by the models. This includes the analysis of true positives, false positives, true negatives and false negatives, which helped in the fine tuning of the models further.

The application of these evaluation metrics enabled a holistic assessment of the models, by revealing strengths and weaknesses that could be used for addressing in the future iterations of the model training. By documenting and analyzing the performance of each model, this thesis not only advances academic knowledge, but also serves as a valuable resource for the practical implementation in the field of education. This comprehensive evaluation process helped in ensuring the reliability and effectiveness of the deployed models, allowing them to be capable of making accurate predictions that can impact educational interventions and strategies significantly.

4.2.10. Documentation

In any research project, particularly in data science and machine learning projects, documentation is crucial for several reasons. It provides a detailed account of the methodologies used, the decision making processes, the results obtained, and the interpretations and conclusions drawn from the datasets. Effective documentation not only ensures the transparency of the research, but also its reproducibility, easiness to review for validation and is helpful for the future researchers who may want to build upon the work.

Importance of Thorough Documentation

Reproducibility Proper documentation allows other researchers to replicate the work under similar conditions for verifying the findings and applying the methods to new datasets.

Transparency Detailed records helps in maintaining the integrity of the research process, allowing for a clear understanding of how the conclusions were drawn.

Continuity Comprehensive documentation helps in maintaining continuity in ongoing projects, as different team members may engage with the project at various stages.

Documentation Strategies in This Thesis

Code Documentation All the coding activities were conducted using Visual Studio Code, which offered an efficient environment for developing, testing and managing the project files. Extensive inline comments were maintained throughout the coding for describing the functionalities, requirements and usage of each code segment.

Methodology Description Detailed documentation of this thesis covered entire methodology workflow, employed from requirement understanding to model evaluation which explains the rationale behind each methodological choice including data selection, data pre-processing, model selection etc. This section also discusses the reasoning for selecting different machine learning models and the balancing techniques which helped in providing a clear linkage to the research questions.

Results and Analysis Documentation All the findings from the outputs of the model were meticulously documented, by incorporating charts, graphs and other analyses to substantiate the interpretations and conclusions. This documentation not only highlights the quantitative outcomes but also contextualized these results within the broader scope of the research objectives.

Technical and User Documentation Technical documentation is aimed at developers and future researches, all tailored around the use of visual studio code. User documentation is designed for end-users like educational administrators, by translating complex technical details into practical insights and actionable recommendations, thus ensuring the applicability of the research in real world settings.

Project Documentation Tools and Formats Jupyter notebooks were extensively used for their ability to combine executable code, visual data representations and rich text in a single document. Latex editor was utilized to create a well-formatted documentation that is easy to navigate and interpret, supporting the project's detailed narrative.

The meticulous documentation approach adapted in this thesis enhances the reliability, utility and scalability of the research. By outlining each step of the research clearly from conceptualization to execution, this documentation ensures that the research can serve as a foundational model for future studies which can drive in further advancements in the application of machine learning in the context of education. This dedication to provide a detailed documentation is pivotal in extending the impact of the research in academic circles, along with making it a valuable resource for educators for practical implementation and policy development.

5. Results

This section presents a detailed analysis of the results obtained from applying regression and classification models to the three different educational datasets. Each part of this section describes the regression and classification techniques applied to each dataset along with the evaluation of the models including various performance metrics and the implications of these findings.

5.1. Dataset 1 - Predict Student Dropout and Academic Success

5.1.1. Classification Results

The evaluation of classification models on the 'Predict Student Dropout and Academic Success' dataset involved a systematic application of various machine learning models which were combined with different data balancing techniques. The aim was to identify the most effective model and balancing technique combination for predicting the student outcome. Each model, balancing technique combination were also subjected to cross-validation using a 5-fold KFold strategy for estimating the performance stability of the model across different data splits. The models were then trained on the full training set and evaluated on the testing set.

The results of accuracy with cross validation and on testing datasets are shown in Table 5.5. The highest accuracy was shown by Random Forest Model combined with SVM SMOTE and also random forest combined with SMOTE with an accuracy of 76.38 percent. The high cross-validation mean accuracy demonstrated by random forest with SVM SMOTE shows robustness, suggesting that it generalizes well across different subsets of the data. Additionally, neural network and SVM models showed consistently good performance across various balancing techniques, indicating their effectiveness in handling imbalanced data in this context.

In terms of precision, recall and f1 score, random forest consistently showed high performance across all metrics with various balancing techniques. It was particularly effective at predicting graduates (Class 2), which showed a strong balance between precision and recall. Across most models, Class 2 that represents graduates, achieved higher precision and f1 scores often exceeding 0.80. This implies the models effectiveness in identifying the students who are likely to graduate which is crucial for resource allocation and program planning. But class 1 representing enrolled students consistently showed lower precision and recall, which presents a challenge in modeling, possibly requiring more distinctive features or advanced model tuning to improve the identification accuracy.

The confusion matrix and classification report for Random Forest with SVM SMOTE is shown in Table 5.1 and 5.2 respectively.

	Predicted Class 0	**Predicted Class 1**	**Predicted Class 2**
Actual Class 0	246	26	44
Actual Class 1	35	60	56
Actual Class 2	17	31	370

Table 5.1.: Dataset 1: Confusion Matrix of Random Forest with SVM SMOTE

Class	**Precision**	**Recall**	**F1-score**
0	0.83	0.78	0.80
1	0.51	0.40	0.45
2	0.79	0.89	0.83

Table 5.2.: Dataset 1: Classification Report for Random Forest with SVM SMOTE

And the confusion matrix and classification report for Random Forest with SMOTE is shown in Table 5.3 and 5.4 respectively.

	Predicted Class 0	**Predicted Class 1**	**Predicted Class 2**
Actual Class 0	236	37	43
Actual Class 1	28	73	50
Actual Class 2	14	37	367

Table 5.3.: Dataset 1: Confusion Matrix of Random Forest with SMOTE

Class	**Precision**	**Recall**	**F1-score**
0	0.85	0.75	0.79
1	0.50	0.48	0.49
2	0.80	0.88	0.84

Table 5.4.: Dataset 1: Classification Report for Random Forest with SMOTE

Balancing Technique	Model	CV Mean Accuracy	Test Accuracy	Macro Average F1 Score	Weighted Average F1 Score
SMOTE	Decision Tree	0.672777	0.670056	0.60	0.67
SMOTE	Random Forest	0.765463	0.763842	0.71	0.76
SMOTE	Logistic Regression	0.746259	0.737853	0.68	0.75
SMOTE	SVM	0.741170	0.753672	0.71	0.76
SMOTE	Neural Network	0.725908	0.690395	0.62	0.69
ADASYN	Decision Tree	0.687484	0.662147	0.61	0.67
ADASYN	Random Forest	0.766595	0.762712	0.70	0.76
ADASYN	Logistic Regression	0.747105	0.742373	0.69	0.75
ADASYN	SVM	0.742012	0.748023	0.71	0.76
ADASYN	Neural Network	0.717719	0.703955	0.63	0.70
SMOTE_Tomek	Decision Tree	0.686067	0.661017	0.58	0.66
SMOTE_Tomek	Random Forest	0.769985	0.753672	0.69	0.75
SMOTE_Tomek	Logistic Regression	0.744846	0.737853	0.68	0.75
SMOTE_Tomek	SVM	0.740040	0.744633	0.70	0.75
SMOTE_Tomek	Neural Network	0.722523	0.726554	0.66	0.73
SMOTE_ENN	Decision Tree	0.655265	0.663277	0.62	0.68
SMOTE_ENN	Random Forest	0.727322	0.733333	0.70	0.75
SMOTE_ENN	Logistic Regression	0.693132	0.692655	0.67	0.72
SMOTE_ENN	SVM	0.681829	0.680226	0.65	0.71
SMOTE_ENN	Neural Network	0.685503	0.699435	0.66	0.71
BorderlineSMOTE	Decision Tree	0.682953	0.662147	0.60	0.67
BorderlineSMOTE	Random Forest	0.769422	0.755932	0.68	0.75
BorderlineSMOTE	Logistic Regression	0.747669	0.740113	0.68	0.75
BorderlineSMOTE	SVM	0.740884	0.748023	0.70	0.75
BorderlineSMOTE	Neural Network	0.715456	0.712994	0.65	0.71
SVMSMOTE	Decision Tree	0.670805	0.674576	0.62	0.68
SVMSMOTE	Random Forest	0.772247	0.763842	0.69	0.76
SVMSMOTE	Logistic Regression	0.752753	0.731073	0.67	0.73
SVMSMOTE	SVM	0.751058	0.757062	0.70	0.76
SVMSMOTE	Neural Network	0.719975	0.710734	0.64	0.71

Table 5.5.: Dataset 1: Classification Results

5.2. Dataset 2 - UCI Student Performance

5.2.1. Classification Results

In the analysis of the UCI student performance dataset, pre-processing was a crucial step due to the presence of categorical variables. The `get_dummies` function from pandas was used to handle these variables effectively. This function converted categorical variables into dummy or indicator variables. Initially the dataset had 33 columns, but after applying the function, this number increased to 42 columns. This increase in column number is due to the transformation of categorical columns into multiple binary columns, one for each category. This enhances the ability of the model to leverage the categorical data for predictions.

Like Dataset 1, Dataset 2 - the UCI student performance dataset was also subjected to validation using a 5-fold cross-validation approach for each model and balancing technique combination, for estimating its stability and performance across different subsets of the data. The models were then also evaluated on testing datasets as well to measure their predictive accuracy and other evaluation metrics such as precision, recall, f1 score to understand their performance in distinguishing between the binary classes 0 and 1, which represents fail and pass respectively.

Table 5.6 shows the results of testing accuracy and cross-validation accuracy on the each combination of model and balancing techniques. For the UCI student performance data, the combination of decision tree with SVM SMOTE turned to be more efficient with the highest accuracy of approximately 93.07 percentage. This result suggests that SVM SMOTE effectively addresses class imbalances by generating synthetic samples in a manner that suits the decision boundaries learned by SVM. This enhances the ability of decision tree to classify minority classes more accurately, leading to an improved overall performance on the test data. This accuracy indicates that the decision tree model with SVM SMOTE could be particularly useful in scenarios where precise binary classification of student performance is critical, like identifying students who needs special attention to succeed, thus allowing educators to apply targeted strategies to support the students.

Whereas the cross-validation mean accuracy is highest for the combination of Random Forest with Borderline SMOTE with a mean accuracy of 94.22 percent. The inherent strength of random forest in managing overfitting, combined with the effectiveness of BorderlineSMOTE in handling overlapping class distributions and noisy data, likely contributed to this robust performance. The high cross validation score indicates that this combination not only performs well on average across several subsets of data, but also shows consistent performance, which highlights the reliability and stability of the model. This is very important in the educational domain where the model needs to perform well across different schools or several demographic groups.

Balancing Technique	**Model**	**CV Mean Accuracy**	**Test Accuracy**	**Macro Average F1 Score**	**Weighted Average F1 Score**
SMOTE	Decision Tree	0.890105	0.892308	0.75	0.90
SMOTE	Random Forest	0.936426	0.915385	0.80	0.92
SMOTE	Logistic Regression	0.892065	0.876923	0.75	0.89
SMOTE	SVM	0.876662	0.892308	0.74	0.89
SMOTE	Neural Network	0.899776	0.915385	0.79	0.91
ADASYN	Decision Tree	0.897872	0.907692	0.77	0.91
ADASYN	Random Forest	0.934485	0.907692	0.79	0.91
ADASYN	Logistic Regression	0.890123	0.861538	0.73	0.88
ADASYN	SVM	0.874720	0.900000	0.75	0.90
ADASYN	Neural Network	0.886296	0.907692	0.77	0.91
SMOTE_Tomek	Decision Tree	0.890105	0.892308	0.75	0.90
SMOTE_Tomek	Random Forest	0.936426	0.915385	0.80	0.92
SMOTE_Tomek	Logistic Regression	0.892065	0.876923	0.75	0.89
SMOTE_Tomek	SVM	0.876662	0.892308	0.74	0.89
SMOTE_Tomek	Neural Network	0.899776	0.915385	0.79	0.91
SMOTE_ENN	Decision Tree	0.874757	0.869231	0.75	0.88
SMOTE_ENN	Random Forest	0.896004	0.884615	0.78	0.90
SMOTE_ENN	Logistic Regression	0.817009	0.846154	0.74	0.87
SMOTE_ENN	SVM	0.818895	0.869231	0.75	0.88
SMOTE_ENN	Neural Network	0.791934	0.853846	0.74	0.87
BorderlineSMOTE	Decision Tree	0.895836	0.900000	0.76	0.90
BorderlineSMOTE	Random Forest	0.942177	0.915385	0.80	0.92
BorderlineSMOTE	Logistic Regression	0.890161	0.884615	0.76	0.89
BorderlineSMOTE	SVM	0.878603	0.915385	0.79	0.91
BorderlineSMOTE	Neural Network	0.890161	0.915385	0.79	0.91
SVMSMOTE	Decision Tree	0.899757	0.930769	0.83	0.93
SVMSMOTE	Random Forest	0.936408	0.900000	0.77	0.90
SVMSMOTE	Logistic Regression	0.890161	0.884615	0.76	0.89
SVMSMOTE	SVM	0.886296	0.900000	0.77	0.90
SVMSMOTE	Neural Network	0.888219	0.915385	0.79	0.91

Table 5.6.: Dataset 2: Classification Results

Across most models, Class 1 which represented pass outcome, achieved a higher precision, recall and f1 score with precision and f1 score always exceeding 0.90. This implies the effectiveness of the model in identifying the students who are likely to pass which is really important for resource allocation and program planning for the students who are prone to fail. This binary classification tended to gain more accuracy for the models than the accuracy gained by models in the multi classification problem with Dataset 1.

The confusion matrix and classification report for Decision tree with SVM SMOTE is shown in Table 5.7 and 5.8 respectively.

	Predicted Class 0	**Predicted Class 1**
Actual Class 0	10	5
Actual Class 1	4	111

Table 5.7.: Dataset 2: Confusion Matrix of Decision Tree with SVM SMOTE

Class	**Precision**	**Recall**	**F1-score**
0	0.71	0.67	0.69
1	0.96	0.97	0.96

Table 5.8.: Dataset 2: Classification Report for Decision Tree with SVM SMOTE

5.3. Dataset 3 - Module UNI Assessment

5.3.1. Regression Results

The regression analysis conducted on the 'Module UNI Assessment' dataset highlights several key findings. The primary goal of this analysis was to predict the final exam grade based on the grades of ten assignments. The dataset was already cleansed to exclude students who did not attend the exam.

Graphical Representation of Regression Results Figure 5.1 shows a scatter plot visualizing the actual vs predicted values, where the concentration of points along the diagonal suggests that the predictions of the model are generally aligned with the actual grades, but with some variance. Points that are significantly above or below the line would indicate over estimations or under estimations by the model.

Many of the data points lie close to the red dashed line, which is the line of prediction. This proximity indicates that for many students, the predictions of the model were quite close to their actual exam grades. Thus suggesting that for a considerable

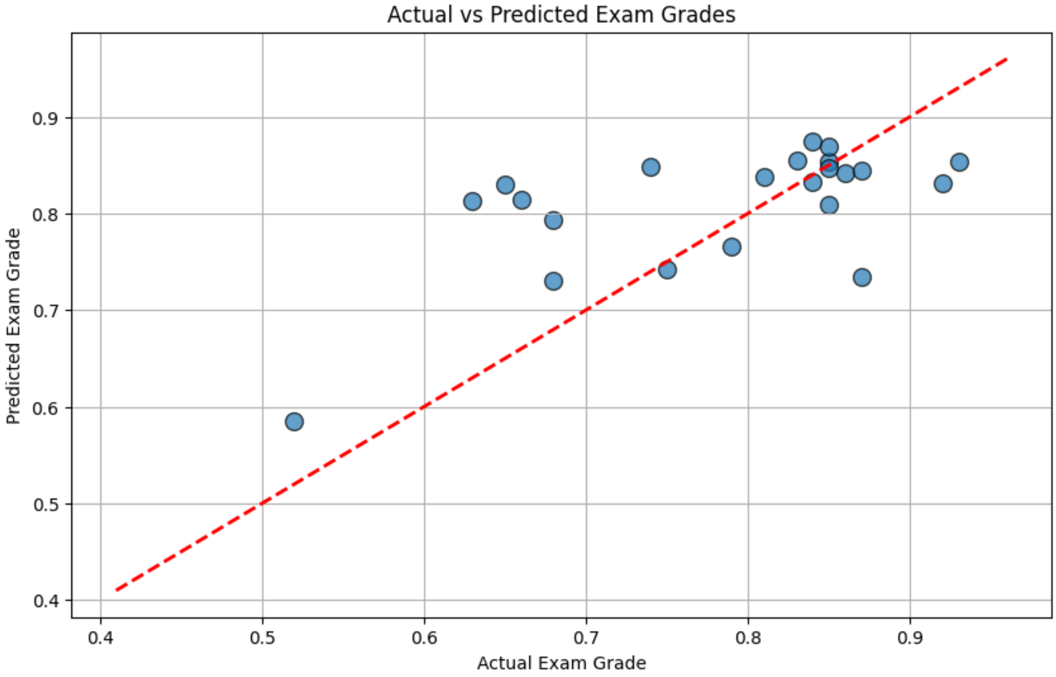

Figure 5.1.: Regression on Dataset 3: Actual vs Predicted

portion of the data, the model effectively captured the underlying pattern between the features used and the exam outcomes.

Metric	**Value**
Mean Squared Error (MSE)	0.007256201268553489
R-squared (R²)	0.3302436420886228

Table 5.9.: Regression Metrics for Dataset 3

The MSE of the model is relatively low, which suggests that the predictions of the model are on average close to the actual exam grades. However, the presence of points that deviate from the prediction line also highlights the limitations of the model in capturing all the complexities of the data.

The value of R-squared is approximately 0.3302, which is not very high, but is not negligible. It indicates that about a third of the variability in exam grades can be explained through the model. And also this could be seen as moderate performance, suggesting that, while some predictive ability is present, there might be room for improvement. Such as, having additional features or different features might help in improving the explanatory power of the model. Features such as student attendance, participation in class etc. could have significant impact on the exam grades and might be able to improve the model if included.

To summarize, considering the MSE AND R² values together, while the predictions are relatively close to actual grades, the ability of the model to explain the variance fully is not very high.

Means Squared Error vs. Number of Features The graph shown in Figure 5.2 illustrates the relationship between the number of assignments and the MSE of a regression model when applied to the Module UNI assessment data. This chart shows a significant trend in reduction of error when the number of features considered increases, particularly noticeable from 1 to 2 assignments where the MSE drops sharply and then with the most marked improvement occurring at the ninth feature.

This analysis suggests that for predictive modeling, particularly in the context of education or similar settings, where assignments are used as predictors, it is important to identify the optimal number of features that contribute to the performance of the model without leading to overfitting or unnecessary complexity. This can help in efficiently utilizing the data for making accurate predictions while maintaining the simplicity and computational efficiency of the model.

These analyses can be valuable for educational administrators and curriculum planners to understand the predictive power of the features related to the students and give planning and support to the students in need by assessing these features.

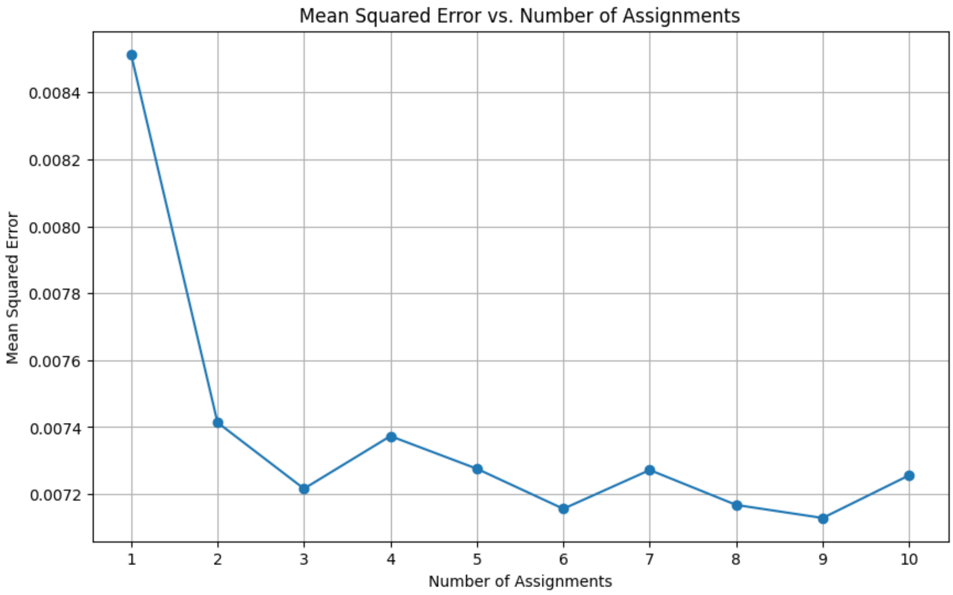

Figure 5.2.: Dataset 3: Mean Squared Error vs Number of Features

5.3.2. Classification Results

The primary goal of classification for this dataset was to predict whether the students would pass or fail on the basis of their exam scores, where pass is defined as a score above 0.5 and fail is when the score is under 0.5. The dataset was already cleansed to exclude students who did not attend the exam. The high difference between the pass and fail class was already demonstrated in the analysis section in Figure 4.13.

Unlike Dataset 1 and Dataset 2, for the classification of this dataset, the data was split into training and testing dataset after applying the balancing techniques, since the minority class was very low for splitting the dataset, with a possibility that the training dataset would not consist of any minority class to train with. Thus ensured splitting of data after balancing for training with both classes for a better prediction. The classification results for the student dataset using various balancing methods and models have shown remarkable accuracy rates across multiple configurations as shown in Table 5.10.

Random Forest showed a test accuracy of 1.0 with SMOTE, ADASYN, SMOTE Tomek, SMOTE ENN, AND SVM SMOTE balancing techniques, highlighting the robustness of random forest model in handling balanced datasets. SVM model achieved a test accuracy of 1.0 with the combination of every balancing technique used in this thesis, demonstrating its efficiency in various balanced scenarios. Every classification model when combined with SVM SMOTE balancing technique has achieved a test accuracy 1.0 indicating the effective balancing of imbalanced data by SVM SMOTE technique. Additionally decision tree and logistic regression model showed a test accuracy of 1.0 when combined with the balancing technique SMOTE ENN.

This dataset, dataset 3, was comparatively much smaller than dataset 1 and 2. But every model with a combination of different balancing technique has gained an accuracy of more than 0.95 for this dataset. These results underscore the effectiveness of integrating advanced balancing techniques with robust classification models to improve the predictive accuracy in educational settings.

Along with accuracy, the other evaluation metrics like precision and f1 score, had a perfect score for both the classes, pass and fail for this dataset, which ensures that when the data is properly balanced and handled, models will be very efficient in predicting the student outcomes. Such insights are crucial in developing early interventions and tailored student support programs.

Balancing Technique	Model	CV Mean Accuracy	Test Accuracy	Macro Average F1 Score	Weighted Average F1 Score
SMOTE	Decision Tree	0.976975	0.976744	0.98	0.98
SMOTE	Random Forest	0.994286	1.000000	1.00	1.00
SMOTE	Logistic Regression	0.988571	0.976744	0.98	0.98
SMOTE	SVM	0.994286	1.000000	1.00	1.00
SMOTE	Neural Network	0.994286	0.976744	0.98	0.98
ADASYN	Decision Tree	0.982353	0.976744	0.98	0.98
ADASYN	Random Forest	1.000000	1.000000	1.00	1.00
ADASYN	Logistic Regression	0.988235	0.976744	0.98	0.98
ADASYN	SVM	0.994118	1.000000	1.00	1.00
ADASYN	Neural Network	0.994118	0.976744	0.98	0.98
SMOTE_Tomek	Decision Tree	0.976975	0.976744	0.98	0.98
SMOTE_Tomek	Random Forest	0.994286	1.000000	1.00	1.00
SMOTE_Tomek	Logistic Regression	0.988571	0.976744	0.98	0.98
SMOTE_Tomek	SVM	0.994286	1.000000	1.00	1.00
SMOTE_Tomek	Neural Network	0.994286	0.976744	0.98	0.98
SMOTE_ENN	Decision Tree	0.982175	1.000000	1.00	1.00
SMOTE_ENN	Random Forest	1.000000	1.000000	1.00	1.00
SMOTE_ENN	Logistic Regression	1.000000	1.000000	1.00	1.00
SMOTE_ENN	SVM	1.000000	1.000000	1.00	1.00
SMOTE_ENN	Neural Network	1.000000	0.953846	1.00	1.00
BorderlineSMOTE	Decision Tree	0.976975	0.976744	0.98	0.98
BorderlineSMOTE	Random Forest	1.000000	0.953846	1.00	1.00
BorderlineSMOTE	Logistic Regression	0.988571	0.976744	0.98	0.98
BorderlineSMOTE	SVM	0.994286	1.000000	1.00	1.00
BorderlineSMOTE	Neural Network	0.994286	0.976744	0.98	0.98
SVMSMOTE	Decision Tree	0.970085	1.000000	1.00	1.00
SVMSMOTE	Random Forest	1.000000	1.000000	1.00	1.00
SVMSMOTE	Logistic Regression	0.977208	1.000000	1.00	1.00
SVMSMOTE	SVM	0.992593	1.000000	1.00	1.00
SVMSMOTE	Neural Network	0.984615	1.000000	1.00	1.00

Table 5.10.: Dataset 3: Classification Results

6. Conclusion

This thesis has successfully addressed the pressing need within the educational domain to identify and support the students at-risk of academic failure through predictive modeling. With a focus on overcoming the challenges presented by the imbalanced datasets, this research explores a range of machine learning algorithms combined with various balancing techniques to enhance the prediction and intervention processes within academic environments.

The effectiveness of the developed models combined with balancing techniques were affirmed through their application to two other datasets ('UCI Student Performance', 'Module UNI Assessment') in addition to the primary dataset ('Predict Student Dropout and Academic Success'), which confirmed their generalizability and the effectiveness of the integrated processing pipeline in different educational contexts.

The methodology adapted for Dataset 3 ('Module UNI Assessment') was particularly innovative based on the distribution of the data, where data splitting post balancing ensured the inclusion of minority classes in the training process. This approach guaranteed that the models were well-trained and reflective of the features, enhancing their accuracy across varied data splits.

In this study, the SVM SMOTE balancing technique consistently yielded superior results across every datasets used in this thesis when paired with various models. Specifically the combination of random forest model with SVM SMOTE emerged to be the most efficient in dataset 1, achieving an accuracy of 76.38 percent. In dataset 2, the decision tree model paired with SVM SMOTE gained an accuracy of 93.07 percent making it most efficient in predicting the outcome of student success. Where the dataset 3 sets a benchmark with a perfect accuracy score of 100 percent for all models integrated with SVM SMOTE, while maintaining high accuracy levels above 95 percent for other model, balancing technique combinations as well. These findings underlines the adaptability and robustness of SVM SMOTE balancing technique across different educational datasets, highlighting its potential as a key tool for predictive modeling for imbalanced datasets in educational settings.

The utilization of resampling techniques, especially SVM SMOTE has demonstrably enhanced model performance across the educational datasets used in this thesis. By addressing the imbalance in the training data, these techniques allowed for a more accurate and fair prediction of students at-risk. The improved modeling outcomes are evident in the substantial increases in the accuracy and F1 scores, confirming the effectiveness of resampling in tackling the skewness in the dataset.

The insights gained from this research are aimed to offer significant benefits to the educational institutions. By facilitating early identification of students at-risk, the models enable timely and targeted interventions, potentially improving educational outcomes through customized support measures. The research provides a clear guide on how to harness technology to foster educational success. By making the predictive tools actionable and accessible, this study helps in empowering the educators to implement data-driven strategies effectively.

This thesis not only contributes to the educational domain by refining predictive model development and data handling strategies, but also sets a precedent for the future research in the educational technology. It creates a pathway for ongoing improvements and innovations in student support systems, by bridging the gap between theoretical approaches and practical applications, thus ultimately leading to a more fair and effective educational environment.

This research can further be expanded by exploring alternative resampling strategies that might further refine the robustness of the model, particularly against data drift in different educational environments. Additionally, provisions can be made to integrate real-time data acquisition and updating models dynamically that could make the system even more responsive to changes in the patterns of student behavior. Also, experimenting with emerging machine learning algorithms along with feature selection techniques may also provide deeper insights and improved predictive performance.

Bibliography

[1] Constructivist learning theory. *Constructivism, Learning Theories, ELM Learning*, 2024.

[2] A. Adedokun, J. Akinjobi, T. Omolara, and K. Oladapo. Data mining technique for early detection of at-risk students. 2019.

[3] A. Alashoor and S. Abdulla. Examining techniques to solving imbalanced datasets in educational data mining systems. 21:205–213, 06 2022.

[4] A. Alnuaimi and T. Albaldawi. An overview of machine learning classification techniques. *BIO Web of Conferences*, 97, 2024.

[5] H. Altabrawee, O. Ali, and S. Qaisar. Predicting students' performance using machine learning techniques. *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences*, 27, 2019.

[6] J. A. Bacus and R. Cascaro. Impact of predictive learning analytics in higher education: A systematic literature review. *2024 13th International Conference on Educational and Information Technology (ICEIT)*, 2024.

[7] R. Baker. Educational data mining anf learning analytics. 2014.

[8] R. S. J. Baker. Data mining for education. *International encyclopedia of education*, 2010.

[9] T. M. Barros, P. A. Souza Neto, I. Silva, and L. A. Guedes. Predictive models for imbalanced data: A school dropout perspective. *Education Sciences*, 9(4), 2019.

[10] H. Belyadi and A. Haghighat. Chapter 5 - supervised learning. pages 169–295, 2021.

[11] G. blog. 10 techniques to solve imbalanced classes in machine learning. *Analytics Vidhya*, 2024.

[12] A. Bordia. Handling imbalanced data by oversampling with smote and its variants. *Analytics Vidhya,Medium*, 2022.

[13] B. P. C. A gentle introduction to support vector machines. *KDnuggets*, 2023.

[14] T. A. Cardona and E. a. Cudney. Predicting student retention using support vector machines. *Procedia Manufacturing*, 2019.

[15] P. Cortez. Student Performance. UCI Machine Learning Repository, 2014.

[16] P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. 2008.

[17] T. de Jong. Cognitive load theory, educational research, and instructional design: some food for thought. 2010.

[18] W. J. W. Donna L. Mohr and R. J. Freund. Statistical methods (fourth edition). *Science Direct*, 2023.

[19] N. Eleyan, M. Al Akasheh, E. F. Malik, and O. Hujran. Predicting student performance using educational data mining. *022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2022.

[20] Figueira. Communication and resource usage analysis in online environments: An integrated social network analysis and data mining perspective. pages 1027–1032, 04 2017.

[21] P. Giudici, A. Gramegna, and E. Raffinetti. Machine learning classification model comparison. *Socio-Economic Planning Sciences*, 87, 2023.

[22] D. S. Goswami. Class imbalance, smote, borderline smote, adasyn. *Towards Data Science,Medium*, 2020.

[23] H. Hassan, N. B. Ahmad, and S. Anuar. Improved students,Äô performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining. *Journal of Physics: Conference Series*, 1529(5):052041, may 2020.

[24] M. Hossin and M. Sulaiman. A review on evaluation metrics for data classification evaluations. *nternational Journal of Data Mining Knowledge Management Process (IJDKP)*, 2015.

[25] M. Jiang, X. Huang, D. Liu, and S. Hu. Teaching evaluation index of college students based on random forest. *2023 3rd International Conference on Educational Technology (ICET)*, 2023.

[26] V. Kanade. What is logistic regression? equation, assumptions, types, and best practices. *Artificial Intelligence, Spiceworks*, 2022.

[27] Z. Liu, S. Yang, J. Tang, N. Heffernan, and R. Luckin. Recent advances in multimodal educational data mining in k-12 education. pages 3549–3550, 08 2020.

[28] V. M. Data imbalance: How is adasyn different from smote? *Medium*, 2023.

[29] B. Mahesh. Machine learning algorithms -a review. *International Journal of Science and Research (IJSR)*, 9, 2019.

[30] Melanie. Classification algorithms: Definition and main models. *Data Scientest*, 2023.

[31] E. Nimy, M. Mosia, and C. Chibaya. Identifying at-risk students for early intervention—a probabilistic machine learning approach. *Applied Sciences*, 13, 2023.

[32] I. Pratama, Y. Pristyanto, and P. T. Prasetyaningrum. Imbalanced class handling and classification on educational dataset. 2021.

[33] V. Realinho, J. Machado, L. Baptista, and M. V. Martins. Predict students' dropout and academic success, 2021.

[34] A. A. Saa. Educational data mining students' performance prediction. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 7, 2016.

[35] B. SAA and E. HF. Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. 2023.

[36] A. S. G. A. K. H. Sadri Alija, Edmond Beqiri. Predicting students performance using supervised machine learning based on imbalanced dataset and wrapper feature selection. *Informatica, An international journal of computing and informatics*, 2023.

[37] I. H. Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2, 2021.

[38] S. Satpathy. Smote for imbalanced classification with python. *Analytics Vidhya*, 2024.

[39] P. Sharma. Different types of regression models. *Analytics Vidhya*, 2024.

[40] G. Siemens and R. Baker. Learning analytics and educational data mining: Towards communication and collaboration. *ACM International Conference Proceeding Series*, 2012.

[41] A. Singh. A comprehensive guide to ensemble learning (with python codes). *Analytics Vidhya*, 2024.

[42] M. Singh, M. Vyas, R. Chaudhary, and U. Soni. Decision tree academic performance model for primary school students. 2022.

[43] S. Tufail, H. Riggs, M. Tariq, and A. I. Sarwat. Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. *Electronics*, 2023.

[44] R. A. A. Viadinugroho. Imbalanced classification in python: Smote-tomek links method. *Towards Data Science,Medium*, 2021.

[45] Z. Y. Y. S. . C. L. P. C. Wuxing Chen, Kaixiang Yang. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review 57,*, 2024.

[46] Željko Đ. Vujović. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 2021.

A. Appendix

A.1. Dataset 1 - Attribute List

Attribute	**Description**
Marital Status	The marital status of the student (e.g., single, married).
Application Mode	How the application was submitted to the university (e.g., online, paper-based).
Application Order	The order of preference of the university application.
Course	The specific course or program the student is enrolled in.
Daytime/Evening Attendance	Indicates if classes are attended during daytime or evening schedules.
Previous Qualification	The highest academic qualification obtained before current enrollment.
Nationality	Nationality of the student.
Mother's Qualification	The educational qualification of the student's mother.
Father's Qualification	The educational qualification of the student's father.
Mother's Occupation	The occupation of the student's mother.
Father's Occupation	The occupation of the student's father.
Displaced	Indicates whether the student is displaced due to any reason like conflicts or natural disasters.
Educational Special Needs	Whether the student has special educational needs.
Debtor	Indicates if the student has any outstanding debts.
Tuition Fees Up to Date	Whether the student's tuition fees are paid up to date.
Gender	The gender of the student.
Scholarship Holder	Indicates whether the student is receiving a scholarship.
Age at Enrollment	Age of the student at the time of enrollment.
International	Whether the student is an international or domestic student.
Curricular Units 1st Sem (credited)	Number of curricular units credited to the student in the first semester.
Curricular Units 1st Sem (enrolled)	Number of curricular units the student enrolled in during the first semester.
Curricular Units 1st Sem (evaluations)	Number of evaluations completed by the student for the first semester units.
Curricular Units 1st Sem (approved)	Number of curricular units the student passed during the first semester.
Curricular Units 1st Sem (grade)	Average grade achieved by the student in the first semester units.
Curricular Units 1st Sem (without evaluations)	Number of curricular units without formal evaluations in the first semester.
Curricular Units 2nd Sem (credited)	Number of curricular units credited to the student in the second semester.
Curricular Units 2nd Sem (enrolled)	Number of curricular units the student enrolled in during the second semester.
Curricular Units 2nd Sem (evaluations)	Number of evaluations completed by the student for the second semester units.
Curricular Units 2nd Sem (approved)	Number of curricular units the student passed during the second semester.
Curricular Units 2nd Sem (grade)	Average grade achieved by the student in the second semester units.
Curricular Units 2nd Sem (without evaluations)	Number of curricular units without formal evaluations in the second semester.
Unemployment Rate	The local or national unemployment rate at the time of study.
Inflation Rate	The inflation rate relevant to the student's country at the time of study.
GDP	The Gross Domestic Product of the student's country during the period of study.
Target	Indicates whether the student dropped out or is enrolled or graduated.

Table A.1.: Dataset 1: Description of Dataset Columns for Predicting Student Dropout and Academic Success

A.2. Dataset 2 - Attribute List

Attribute	Description
School	Identifier for the student’s school (GP - Gabriel Pereira or MS - Mousinho da Silveira).
Sex	Student’s sex (female or male).
Age	Age of student (numeric: from 15 to 22).
Address	Student’s home address type (urban or rural).
Famsize	Family size (LE3 - less or equal to 3 or GT3 - greater than 3).
Pstatus	Parent’s cohabitation status (T - living together or A - apart).
Medu	Mother’s education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education).
Fedu	Father’s education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education).
Mjob	Mother’s job (’teacher’, ’health’, ’services’ (e.g., administrative or police), ’at$_{h}ome'or'other'$).
Fjob	Father’s job (’teacher’, ’health’, ’services’ (e.g., administrative or police), ’at$_{h}ome'or'other'$).
Reason	Reason to choose this school (close to ’home’, ’reputation’, ’course preference’ or ’other’).
Guardian	Student’s guardian (’mother’, ’father’ or ’other’).
Traveltime	Home to school travel time (1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour).
Studytime	Weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours).
Failures	Number of past class failures (numeric: n if 1<=n<3, else 4).
Schoolsup	Extra educational support (yes or no).
Famsup	Family educational support (yes or no).
Paid	Extra paid classes within the course subject (Math or Portuguese) (yes or no).
Activities	Extra-curricular activities (yes or no).
Nursery	Attended nursery school (yes or no).
Higher	Wants to take higher education (yes or no).
Internet	Internet access at home (yes or no).
Romantic	In a romantic relationship (yes or no).
Famrel	Quality of family relationships (numeric: from 1 - very bad to 5 - excellent).
Freetime	Free time after school (numeric: from 1 - very low to 5 - very high).
Goout	Going out with friends (numeric: from 1 - very low to 5 - very high).
Dalc	Workday alcohol consumption (numeric: from 1 - very low to 5 - very high).
Walc	Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high).
Health	Current health status (numeric: from 1 - very bad to 5 - very good).
Absences	Number of school absences (numeric: from 0 to 93).
G1	First period grade (numeric: from 0 to 20).
G2	Second period grade (numeric: from 0 to 20).
G3	Final grade (numeric: from 0 to 20).

Table A.2.: Dataset 2: Description of Dataset Columns for UCI Student Performance