Department of Electrical and Computer Engineering
North South University

Directed Research

STUDENT-ABSENTEEISM PREDICTION AMONG UNIVERSITY STUDENTS

RIFAT AZIM	ID# 1912353642
MAHADI HASAN	ID# 1911812642
TASFIA TAHMID USHNO	ID# 1831099642
RABEYA AKTER JOYA	ID# 1921268042

Faculty Advisor:

Muhammad Shafayat Oshman

Lecturer

Department of Electrical and Computer Engineering

Spring, 2023

APPROVAL

Rifat Azim (ID # 1912353642), Mahadi Hasan (ID # 1911812642) and Rabeya Akter Joya (ID # 1921268042) and Tasfia Tahmid Ushno (ID # 1831099642) from Electrical and Computer Engineering Department of North South University, have worked on the Directed Research Project titled "Student-Absenteeism Prediction among University Students" under the supervision of Professor Muhammad Shafayat Oshman partial fulfillment of the requirement for the degree of Bachelors of Science in Engineering and has been accepted as satisfactory.

Supervisor's Signature

…………………………………….

Muhammad Shafayat Oshman

Lecturer

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

Chairman's Signature

…………………………………….

Dr. Rajesh Palit

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

DECLARATION

This is to declare that this project/directed research is our original work. No part of this work has been submitted elsewhere partially or fully for the award of any other degree or diploma. All project related information will remain confidential and shall not be disclosed without the formal consent of the project supervisor. Relevant previous works presented in this report have been properly acknowledged and cited. The plagiarism policy, as stated by the supervisor, has been maintained.

Students’ names & Signatures

1. Rifat Azim

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

2. Tasfia Tahmid Ushno

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

3. Rabeya Akter Joya

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

4. Mahadi Hasan

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

ACKNOWLEDGEMENTS

The authors would like to express their heartfelt gratitude towards their project and research supervisor, Muhammad Shafayat Oshman, Lecturer, Department of Electrical and Computer Engineering, North South University, Bangladesh, for his invaluable support, precise guidance and advice pertaining to the experiments, research and theoretical studies carried out during the course of the current project and also in the preparation of the current report.

Furthermore, the authors would like to thank the Department of Electrical and Computer Engineering, North South University, Bangladesh for facilitating the research. The authors would also like to thank their loved ones for their countless sacrifices and continual support.

ABSTRACT

Student-Absenteeism Prediction among University Students

Student absenteeism poses a significant challenge in education systems globally, arising from various causes and leading to numerous adverse consequences. Identifying the specific factors contributing to a student's absence during class hours can be complex.

To address this issue, we employed automation and machine learning techniques. A custom dataset was developed with the participation of students from North South University, encompassing multiple crucial features that predominantly trigger absenteeism. This dataset was subsequently fed into several prominent machine learning models, including K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest. Among these models, Random Forest demonstrated the highest accuracy in predicting student absenteeism that is 77%. Furthermore, by utilizing the feature importance metric inherent to the Random Forest model, we identified that physical health issues are the primary cause of absenteeism among the students of North South University.

This approach not only underscores the efficacy of machine learning in addressing educational challenges but also highlights the critical role of health-related factors in student attendance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Background and Motivation

1.1.1 Introduction

Absenteeism, according to Merriam-Webster dictionary means chronic absence [1]. In the context of the school it is the habitual or intentional failure from going to school. It cannot be denied that every now and then, students may miss some school activities and lessons. But it becomes a problem if the student will be away from school for many days.

Student absenteeism refers to the habitual or frequent absence of a student from school without valid reasons [2]. It is a pervasive issue that affects educational institutions worldwide, impacting not only the individual students but also the overall educational system.

The phenomenon of absenteeism among students can be traced back to various sociolect-economic, psychological, and environmental factors. Historically, absenteeism has been linked to truancy and has been addressed primarily through disciplinary actions [3]. However, contemporary research underscores the complexity of the issue, highlighting the need for a more nuanced understanding and multifaceted approach to address it.

1.1.2 Understanding student absenteeism is critical for several reasons:

Academic Performance: Regular attendance is directly correlated with academic success. Students who frequently miss school are more likely to fall behind in their studies, leading to lower grades and higher dropout rates.

Long-term Outcomes: Chronic absenteeism can have long-lasting effects on students' future prospects, including reduced chances of higher education and limited career opportunities. It can also contribute to a cycle of poverty and social exclusion.

School Funding and Resources: In many educational systems, school funding is tied to student attendance. High rates of absenteeism can lead to reduced financial resources for schools, impacting the quality of education provided.

Community and Social Impact: Absenteeism is often linked to broader social issues such as poverty, family instability, and community disorganization. Addressing absenteeism can have positive ripple effects on the community by promoting social cohesion and stability.

Early Identification of Problems: Regular monitoring of attendance can help in the early identification of students who may be at risk of academic failure, social problems, or other difficulties, allowing for timely intervention.

1.1.3 Motivation for Research

The motivation for researching student absenteeism is driven by the need to develop effective strategies to combat this issue and promote a more inclusive and supportive educational environment. Key motivations include:

Identifying Underlying Causes: To develop effective interventions, it is essential to understand the root causes of absenteeism. These may include personal factors (e.g., health issues, mental health, family responsibilities), school-related factors (e.g., school climate, teacher-student relationships), and broader sociolect-economic factors (e.g., economic hardship, community safety).

Developing Targeted Interventions: Research can inform the design and implementation of targeted interventions such as mentoring programs, counseling services, and engagement initiatives that address the specific needs of at-risk students.

Enhancing Educational Policies: Insights from absenteeism research can guide the development of educational policies that promote attendance and reduce barriers to regular school attendance.

Improving Educational Equity: By addressing the factors contributing to absenteeism, schools can work towards greater educational equity, ensuring that all students have the opportunity to succeed regardless of their background or circumstances.

Contributing to Educational Research: Research on student absenteeism contributes to the broader field of educational research, providing valuable data and insights that can inform future studies and educational practices.

1.2 Purpose and Goal of the Project

1.2.1 Purpose of the Project

The purpose of this research project is to leverage machine learning (ML) techniques to analyze, predict, and reduce student absenteeism. Student absenteeism is a complex issue influenced by a multitude of factors, including personal, sociolect-economic, and environmental variables [4]. Traditional methods of addressing absenteeism often fall short due to their reactive nature and inability to account for the nuanced interplay of these factors. This project aims to employ data-driven methodologies to proactively identify students at risk of chronic absenteeism and uncover the underlying patterns and predictors. By doing so, the project seeks to enable educators and policymakers to implement timely and effective interventions that enhance student attendance and academic performance.

1.2.2 Goals of the Project

Data Collection and Preparation:

Goal: To compile a comprehensive datasets from various sources such as student records, academic performance metrics, sociolect-economic background information, and school environment details.

Objective: Ensure the data is clean, accurate, and representative of the student population to build a robust and reliable machine learning model.

Exploratory Data Analysis (EDA):

Goal: To conduct detailed exploratory data analysis to understand the distribution and relationships within the data.

Objective: Identify key features and potential correlations that could inform the development of predictive models for absenteeism.

Feature Engineering:

Goal: To create meaningful features from the raw data that can enhance the predictive power of the ML models.

Objective: Transform and engineer features that capture the essence of factors contributing to student absenteeism, ensuring the model has high explanatory and predictive capability.

Model Development:

Goal: To develop and train various machine learning models to predict student absenteeism based on the identified features.

Objective: Evaluate multiple algorithms (e.g., logistic regression, decision trees, random forests, gradient boosting, neural networks) to determine the most effective model for accurate predictions.

Model Evaluation and Validation:

Goal: To rigorously assess the performance of the ML models using appropriate metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

Objective: Ensure the models are robust, generalizable, and provide reliable predictions on unseen data, validating their applicability in real-world settings.

Implementation and Monitoring:

Goal: To implement the predictive model and intervention strategies within an educational setting.

Objective: Continuously monitor the impact of the interventions on student attendance, refining the models and strategies based on feedback and outcomes to ensure their effectiveness and sustainability.

Reporting and Dissemination:

Goal: To document the findings, methodologies, and outcomes of the project in a comprehensive research paper.

Objective: Share the insights and recommendations with stakeholders, including educators, policymakers, and the research community, to promote the adoption of data-driven approaches in education.

Specific Outcomes:

Predictive Accuracy: Achieve a high level of predictive accuracy in identifying students at risk of absenteeism, with metrics such as precision and recall meeting or exceeding predefined thresholds.

Intervention Effectiveness: Demonstrate a measurable reduction in absenteeism rates among students who receive targeted interventions based on the ML predictions.

Scalability and Generalization: Develop models and strategies that can be scaled and adapted to various educational contexts and institutions.

Knowledge Contribution: Contribute to the academic and practical understanding of the factors influencing student absenteeism and the potential of machine learning to address educational challenges.

Long-term Vision:

The long-term vision for this project is to establish a sustainable and adaptive framework for using machine learning to improve student attendance and overall educational outcomes. By continuously refining the models and intervention strategies based on ongoing data collection and feedback, the project aims to create a dynamic system that supports student success and well-being across diverse educational environments. This proactive, data-driven approach has the potential to significantly enhance educational practices and policies, ultimately leading to better academic performance and reduced absenteeism rates.

1.3 Organization of the Report

This paper is organized into several comprehensive segments, each addressing a distinct aspect of our project.

In Chapter 1, we begin with an introduction that outlines our motivation and goals, setting the stage for the subsequent discussion and providing the foundational context for the study.

Chapter 2 is dedicated to a thorough literature review. Here, we discuss the various research papers we examined, highlighting the methodologies and findings that informed our approach. This chapter includes an analysis of the strengths and limitations of the reviewed literature, helping us discern which aspects were beneficial to our study and which were less relevant.

In Chapter 3, we delve into the methodologies employed in our project. This section provides a detailed explanation of the techniques and processes used to collect, analyze, and interpret the data. We explain our choice of machine learning models, the rationale behind these choices, and the specific configurations and parameters utilized.

Chapter 4 focuses on the software implementation of our project. We describe the development process, including the tools and technologies used, the structure of our codebase, and the steps taken to ensure the robustness and reliability of our system. This chapter serves as a technical guide to the practical aspects of our project.

Finally, Chapter 5 presents our results and discusses their implications. We include detailed graphs and charts to illustrate our findings, providing a visual representation of the data and the outcomes of our analyses. In this chapter, we also reflect on the limitations of our study and propose potential improvements and directions for future research.

Chapter 2 Research Literature Review

2.1 Existing Research and Limitations

Numerous studies have been conducted to predict employee absenteeism using a variety of machine learning techniques. But very few machine learning project has been done on student absenteeism. In this section, a review of these approaches is presented.

Machine Learning for Attendance Rate Prediction (Educational infrastructure) by **Team Neuralcraft** [5]

The aim of this project was to improve educational infrastructure by predicting school attendance rates using machine learning techniques with a focus on XGBoost Regression. It draws attention to how machine learning algorithms are used in African nations to forecast attendance and enhance academic performance. Key characteristics of the dataset include household ownership of electronics at the district level, access to water and sanitation, school attendance rates, and literacy rates. The relationship between attendance rate and education level is also discussed in the document, along with subnational African countries' access to better water. Lastly, it discussed the model's evaluation using metrics like R2Score, MAE, and MSE, showing that the XGBOOST model successfully increases the accuracy of attendance rate prediction and that the model is prepared for real-world implementation. Enhanced performance of the model.

Predicting and analysis student absenteeism, using ml model, L.mulki,A.rista _ A study conducted at the Faculty of Information Technology [6] analyzed the factors contributing to student absenteeism using machine learning algorithms. Data from 500 students were collected, focusing on family, demographic, social, university, and personal aspects. Findings indicated that factors such as family problems, low income, divorced parents, and dissatisfaction with the field of study and university leaders were linked to higher rates of absenteeism. The study recommends universities provide support services for students facing family issues, enhance campus safety, and improve student engagement and motivation. However, the results may not be applicable to other universities or regions. The algorithm which are used in this project is, bayes net, naive bayes, logistic regression, hoeffding tree, j48, random forest, random tree, REPTTree. To predict the accuracy of the model all algorithm provide a good result which is 97% only in case of naive bayes the result is 83%.

Samir Qaisar Ajmi develop a prediction model “Predicting Absenteeism at Work Using Machine Learning Algorithms” [7]. This concept provides a solution to the issue of employee absenteeism in the workplace. The four machine learning algorithms employed in this study they are logistic regression algorithm, the decision tree algorithm, the neural network algorithm, and the support vector machine algorithm. This article uses a database with 20 attributes that was constructed from absence records from July 2007 to July 2010 at a Brazilian courier company. The accuracy metric and ROC index measure have been used to compare the models' performances. With an AUC of 0.834, the decision tree model has the best accuracy at 83.33%, while the support vector machine has the lowest accuracy at 68.47% with an AUC of 0.760.

Another research has been done by Akash saxena to predict Employee absenteeism prediction. They discussed the methodology used in analyzing employee absenteeism data, covering topics such as missing value analysis, feature selection, feature scaling, and model selection. And they also includes code snippets in Python that demonstrate data preprocessing and model evaluation techniques. The conclusion highlights insights into the causes of absenteeism and projected losses if the trend continues. The code performs various tasks, including K-nearest neighbor imputation, correlation analysis, scaling, and implementing prediction algorithms such as KNN, linear regression, decision tree, random forest, and naive Bayes. They used dataset that has 21 features and 742 data records. As the model is regression model prediction, they used MSE/RMSE and MAE to predict the accuracy of the model and KNN and random forest provided a better accurate result in comparison with the decision tree.

Wahid & Zaman have employed four machine learning algorithms to forecast the period of job absenteeism. The dataset they used was gathered from a Brazilian courier service. Tree Ensemble, Gradient Boosted Tree, Random Forest, and Decision Tree techniques have all been applied. Gradient Boosted Tree, with 82% accuracy, was the best model.

Incorporating a Machine Learning Model into a Web-Based Administrative Decision Support Tool for Predicting Workplace Absenteeism by **Gopal Nath** "et al.” [8]

They talked about integrating machine learning into an online platform that forecasts employee absenteeism. The study trained machine learning models using data from a Brazilian courier company, and it discovered that the Support Vector Machine (SVM) and Multinomial Logistic

Regression (MLR) models had the best accuracy. A web-based interactive application was created to increase the tool's accessibility, enabling managers to forecast absenteeism without any prior experience with machine learning. With the use of this tool, managers may make more informed decisions to lower absenteeism and its detrimental effects on productivity. The report also makes reference to earlier investigations on the causes of absenteeism, which included the application of data mining and predictive analytics methods.

Chapter 3 Methodology

3.1 System Design

Predicting absenteeism among student or employees is a complicated task, given that there are many variables and reasons which causes a person's absence in working days. However, in our project, we tried to simplify each and every problems into some categories and then unified the data by going from person to person.

In our prediction system, we've used these categories to predict the outcome, which is our target variable, Absent per Week. The basic idea was to see whether the absence among the student caused by the number of problems they are facing, which we calculated and documented using the categories we mentioned earlier.

Here is a simplified version of the block diagram of our system design:

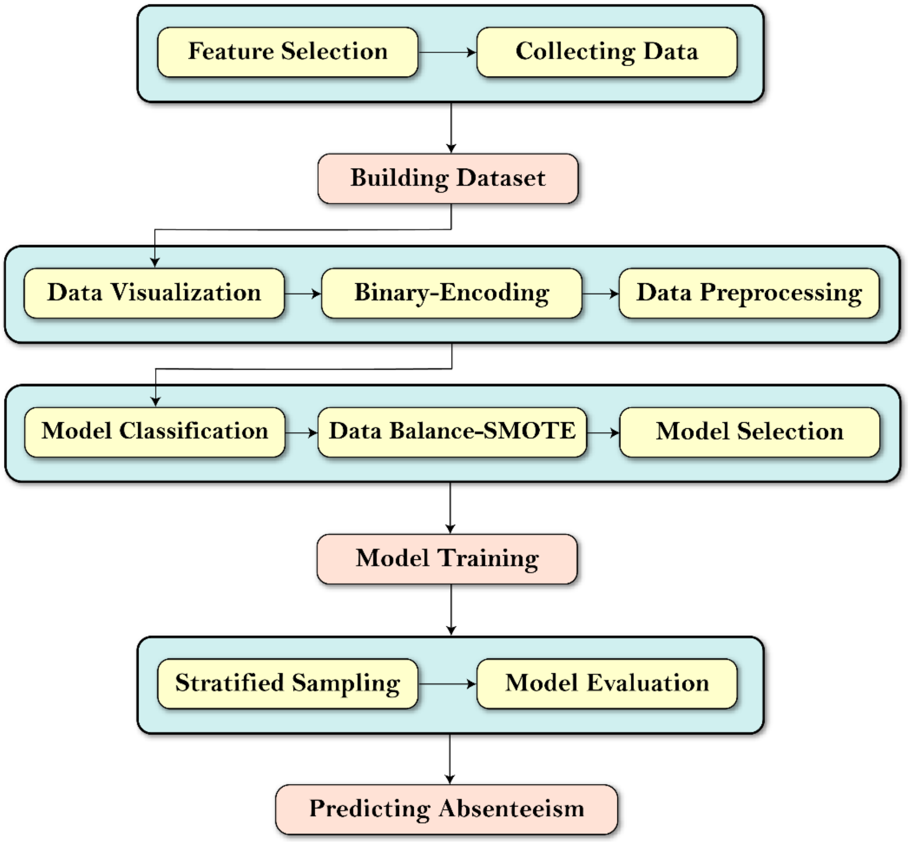


Figure 1: System Design

3.2 Software Components

It is a general understanding that in a machine learning project, tons of software components are used. From handling the dataset to preprocessing and model training, we use various components and libraries for our project. The components that we used however, are sequentially and briefly explained below.

3.2.1 Dataset

Given the lack of suitable online datasets for predicting university student absenteeism, we have resolved to create our own dataset tailored to this objective. After extensive brainstorming and discussion sessions, we have opted to design a binary classification dataset. This approach, while less complex than deep learning models, can yield results of comparable precision.

The primary variable of interest in our dataset is the "Absent per Month" rate, which serves as our target variable. We have identified several features that may influence student absenteeism, encompassing a broad range of potential causes. These features include:

i. Physical Health Issues (PHI): Encompasses all health-related problems, from common illnesses such as colds and fevers to more severe health conditions and accidental injuries.

ii. Mental Health Issues (MHI): Addresses the challenges faced by students struggling with personal mental health issues, which can significantly impact their attendance.

iii. Financial Issues (FI): Recognizes the role of financial difficulties in hindering a student’s ability to attend classes regularly.

iv. Learning Difficulties (LD): Considers various learning disabilities and difficulties that may demotivate students from attending classes.

v. Social Communication Issues (SCI): Includes problems related to social interactions and communication, which can lead to absenteeism.

vi. University Environment (UE): Reflects the impact of the university's environment on student motivation and comfort. A supportive and welcoming environment may encourage attendance, while a negative environment can have the opposite effect.

vii. Transport Issues (TI): Highlights the role of transportation in student absenteeism. The distance between a student’s home and the university, as well as the reliability and convenience of the transportation system, can significantly influence attendance rates.

These categories were chosen because they collectively encompass the primary reasons for student absenteeism. Physical health issues cover a wide array of health-related reasons for absence. Mental health issues are also critical, as students facing personal struggles may be less inclined to attend classes. Financial issues, learning difficulties, and social communication problems can further demotivate students. The university environment plays a pivotal role in either mitigating or exacerbating these issues, while transport problems can be a significant logistical barrier to regular attendance.

After determining the features for our study, we advanced to the data collection phase, targeting a sample from the student population of North South University. Our initial goal was to gather data from at least 2,000 students, aiming for a robust dataset that would enhance the reliability of our analysis. However, we exceeded our expectations and managed to collect responses from 2,500 students, significantly enriching our dataset.

To collect the data, we designed a survey comprising seven yes-or-no questions, each aligned with one of our identified features: Physical Health Issues, Mental Health Issues, Financial Issues, Learning Difficulties, Social Communication Issues, University Environment, and Transport Issues. These questions were aimed at capturing the factors that potentially contribute to student absenteeism.

We approached the students individually and explained the purpose of our project, ensuring transparency and seeking their voluntary participation. With their consent, we recorded their responses along with their ID numbers, which would be used exclusively for the research purposes outlined. This approach not only ensured the ethical handling of participant information but also helped in maintaining the accuracy and traceability of the data.

The collected data was then meticulously organized and stored, ready for the preprocessing and analysis stages. This comprehensive data collection effort laid a solid foundation for our study, allowing us to move forward with a dataset that is both large and representative of the student population at North South University.

By incorporating these features into our dataset, we aim to provide a comprehensive framework for analyzing and predicting absenteeism rates among university students. This approach will facilitate the development of effective interventions to address and reduce absenteeism.

After training our models, we've found out that Physical Health Issues mostly caused in the increase of absenteeism rate among the students [9]. Here is the plot that showed us that data:

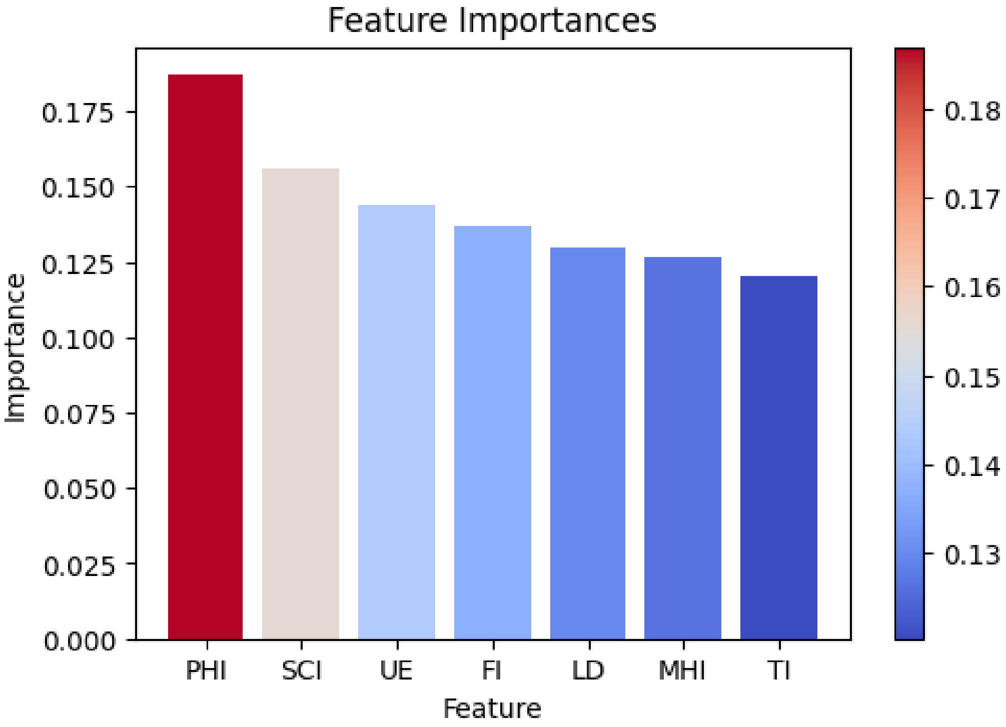


Figure 2: Important features that cause absenteeism.

3.2.2 Data Preprocessing

The initial step in our data preprocessing involved converting the dataset from its original format into a more suitable format for analysis. The dataset was initially in an XLSX file, which was converted into a CSV file for easier handling and processing.

Feature Selection and Binary Encoding: From the original dataset containing 10 columns, we selected the relevant features that correspond to our predefined categories: Physical Health Issues (PHI), Mental Health Issues (MHI), Financial Issues (FI), Learning Difficulties (LD), Social Communication Issues (SCI), University Environment (UE), and Transport Issues (TI). These features were then subjected to binary encoding. Each feature was encoded as follows: 'y' (yes) was converted to 1, and 'n' (no) was converted to 0. This binary encoding simplifies the data, making it suitable for machine learning algorithms that require numerical input.

Handling Missing Values: After encoding, we checked the dataset for any null values. Identifying and handling missing values is crucial as they can significantly impact the performance of machine learning models. Depending on the amount and pattern of missing data, appropriate strategies such as imputation or deletion would be employed to address these gaps.

SMOTE (Synthetic Minority Over-sampling Technique): Given the potential for imbalanced data, particularly with binary target variables, we applied SMOTE. SMOTE is used to generate synthetic samples for the minority class, thereby balancing the dataset [10]. This technique helps prevent models from being biased towards the majority class and improves the model's ability to generalize. We had 10 minority classes, and 3 majority classes. Here we've attached a pie chart that shows our classes before and after SMOTE.

Normalization: Normalization was performed to ensure that all features contribute equally to the model [11]. This step scales the data such that each feature has a mean of zero and a standard deviation of one. Normalization is essential for algorithms sensitive to the scale of data, such as distance-based methods.

Stratified Sampling: To ensure that our training and testing datasets were representative of the overall population, we employed stratified sampling. Stratified sampling involves dividing the data into homogeneous subgroups (strata) and then sampling from each subgroup proportionally [12]. This technique preserves the distribution of the target variable across the train and test sets, providing a more accurate evaluation of the model's performance.

Through these preprocessing steps, we have prepared a clean and balanced dataset, ready for the subsequent stages of model training and evaluation. These steps are critical in ensuring that the machine learning model performs effectively and yields reliable predictions.

3.2.3 Machine Learning Models

For this project, we've used various models which are Logistic Regression, K-Nearest Neighbor, Support Vector Machine (SVM), Random Forest, and Decision Tree. Among these models, Random Forest outperformed the other models, attaining 77% accuracy score.

1. Logistic Regression: Logistic regression is utilized for binary classification, employing the sigmoid function, which takes independent variables as input and produces a probability value between 0 and 1. For example, given two classes, Class 0 and Class 1, if the logistic function's value for an input exceeds 0.5 (the threshold value), the input is assigned to Class 1; otherwise, it is assigned to Class 0. It is termed regression because it extends linear regression, though its primary application is in classification problems. [13]

   Although the Logistic Regression model is Meta for binary classification, however it did not perform as significantly as we expected, because it gave us only 55% accuracy.
2. K-Nearest Neighbors: This is a fundamental yet crucial classification algorithm in machine learning. It falls under the domain of supervised learning and is extensively used in pattern recognition, data mining, and intrusion detection. This algorithm is highly applicable in real-world scenarios because it is non-parametric, meaning it does not make any assumptions about the underlying data distribution (unlike other algorithms such as GMM, which assume a Gaussian distribution). In KNN, we use existing data, known as training data, to classify new data points based on their proximity to the known groups identified by specific attributes [14]. The KNN algorithm gave us 60% testing and overall 63% training accuracy.
3. Support Vector Machine (SVM): This is a supervised machine learning algorithm used for both classification and regression. Although it can be applied to regression problems, it is best suited for classification tasks. The primary objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can effectively separate data points into different classes within the feature space. [15] The equation for the linear hyperplane can be expressed as:

   $$w^t x + b = 0$$

   The SVM model gave us 64% accuracy.

4. Random Forest: The Random Forest algorithm is a highly popular supervised machine learning algorithm used for both classification and regression problems. It operates as a classifier comprising multiple decision trees constructed on various subsets of the given dataset, and it aggregates the results by averaging to enhance predictive accuracy. This algorithm is based on the concept of ensemble learning, which involves combining multiple

classifiers to address complex problems and improve the overall performance of the model. [16]

The Random Forest model has provided the highest accuracy score so far, achieving an impressive 77% accuracy. This success is attributed to the architecture of the Random Forest algorithm. Additionally, the ensemble learning process, which combines multiple classifiers, played a crucial role in achieving this significant improvement in accuracy.

5. Decision Tree: We opted for the Decision Tree model primarily due to its typically faster computation speed compared to Random Forest in many cases. Our objective was to assess whether the Decision Tree could not only reduce computation time but also potentially surpass the performance of the Random Forest model. However, the outcome was not as anticipated. The Decision Tree model achieved only 52% accuracy, highlighting the Random Forest's superiority in our project scenario.

TABLE I. ML MODELS AND CLASSIFICATION REPORT

Model	**Specialty**	**F1-Score**
Logistic Regression	Renowned for binary classification as it uses the sigmoid function.	55%
K-Nearest Neighbor	Extraordinary algorithm to recognize patterns in dataset.	63%
Support Vector Machine	Good for both classification and regression problems.	67%
Random Forest	Similar to SVM, but aggregates results by averaging the classifiers.	77%
Decision Tree	Lighter than Random Forest algorithm, and takes lesser computation time than RF.	52%

3.3 Software Implementation

Building the project from scratch posed significant challenges, particularly in the initial phase of sourcing a suitable dataset. Despite extensive searching, none of the available datasets met our specific requirements, prompting us to create a bespoke dataset using Microsoft Excel. This dataset included seven carefully selected features, with "Absent per Month" (APM) serving as the target variable, quantifying the number of days a student was absent in a given month.

With the dataset finalized, our next step was to conduct a thorough Exploratory Data Analysis (EDA) to gain insights into the data [17]. This process not only involved visualizing the dataset but also delving deeper into the relationships between the selected features and the absenteeism rate among students. We discovered that these features indeed played significant roles in influencing student absenteeism, reinforcing the relevance of our chosen variables.

Moving forward, our preprocessing phase encompassed several critical steps to ensure data quality and readiness for analysis. Initially, we converted the dataset from its original .xlsx format to .csv, enhancing ease of handling and manipulation. Subsequently, employing binary encoding transformed categorical data (expressed as 'yes' and 'no') into numerical equivalents ('1' and '0'), facilitating compatibility with machine learning algorithms.

Following encoding, meticulous checks were conducted to identify and rectify null values, empty columns, and other data imperfections that could potentially distort our analysis or model performance. Addressing these issues was crucial to mitigating risks of underfitting or overfitting during model training and evaluation.

Despite initially suspecting interrelationships among the selected features, our correlation matrix analysis yielded surprising results: minimal correlation existed among them.

Here is our correlation heatmap which was plotted using an existing machine learning library “matplotlib” from pyplot:

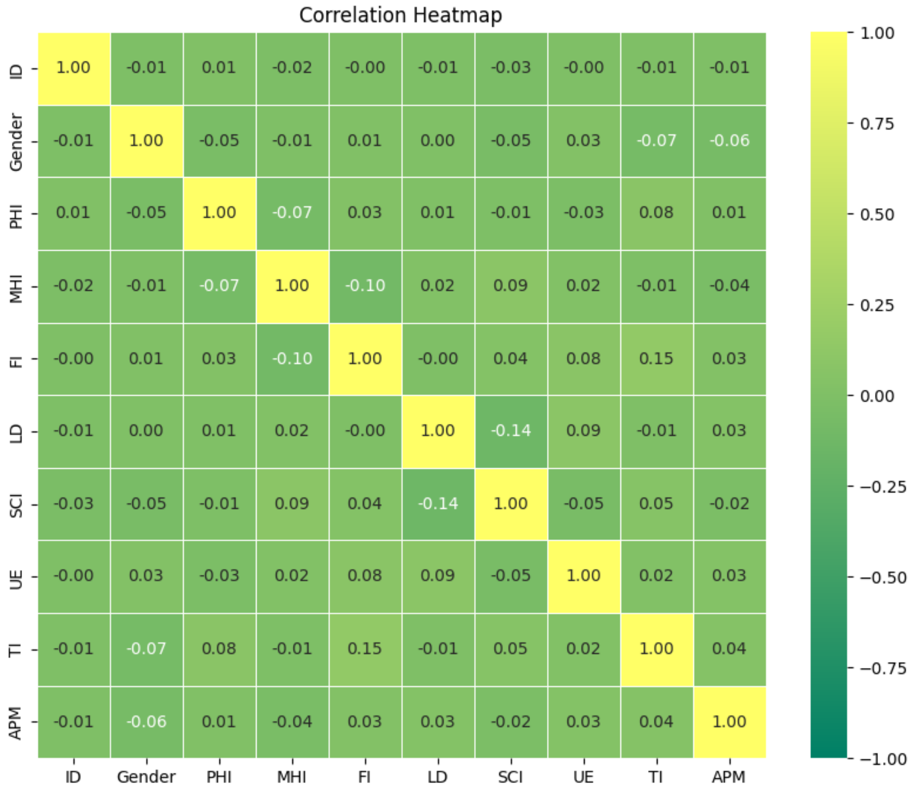


Figure 3: Correlation Matrix

This discovery underscored the complexity of factors contributing to student absenteeism, highlighting the need for a nuanced approach in interpreting and addressing educational challenges. By meticulously preparing and analyzing our dataset through these comprehensive steps, we established a robust foundation for subsequent stages of model development and evaluation. This approach not only enhanced our understanding of the dataset but also underscored the importance of methodical data preparation in achieving reliable and actionable insights in educational research and practice.

This discovery prompted us to adopt a different approach. Upon closer examination of our classes, we identified numerous classes based on the number of days a student was absent, revealing a significant imbalance in the data distribution across these classes. To address this imbalance without compromising data quality, we decided to employ SMOTE (Synthetic Minority Over-sampling Technique).

First, we identified the majority and minority classes. We found three majority classes and ten minority classes, with the latter having fewer than 50 samples each. To streamline our dataset, we combined the minority classes, reducing them to six major classes. These classes represented the number of days a student was absent, ranging from 1 day to more than 5 days, with the latter being denoted as "5+."

Implementing SMOTE allowed us to balance the dataset by oversampling the minority classes, ensuring that each class had a more equitable representation. This approach helped mitigate potential biases and improved the robustness of our predictive model [18]. Below is a pie chart illustrating the class distribution before and after applying SMOTE:

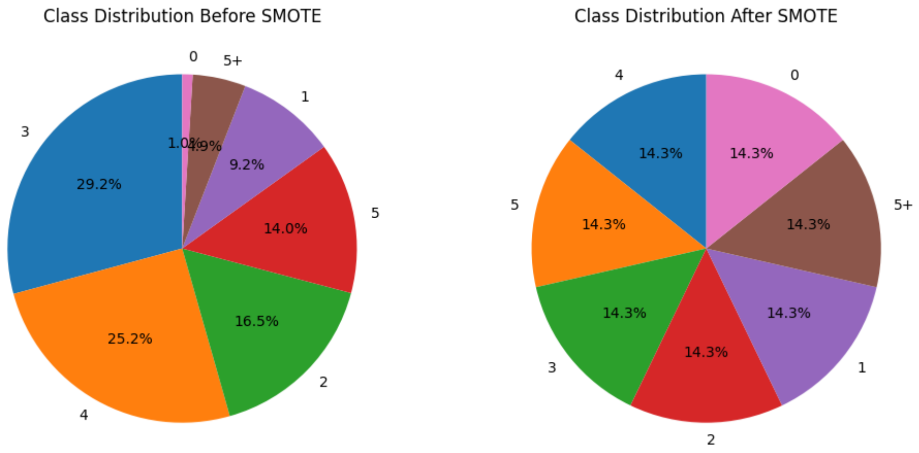


Figure 4: Class distribution before and after SMOTE

This balanced dataset provided a more reliable foundation for our machine learning models, enhancing their ability to accurately predict student absenteeism. Through this methodical approach, we ensured that our dataset was both representative and high-quality, setting the stage for more effective data analysis and model performance.

After applying SMOTE, we proceeded to normalize the data in preparation for running a stratified sampling process. Given that our dataset contained multiple classes, we determined that stratified sampling would be more effective than clustering. Unlike clustered sampling, stratified sampling ensures that samples are taken from all existing classes, thus preserving the class distribution in the training and testing sets.

Following the stratified sampling process, we normalized the data once more. This additional normalization step was crucial to ensure that the features were on a comparable scale, which is particularly important for the performance of various machine learning models. Normalizing the data helps to improve the accuracy and efficiency of the models by preventing any one feature from disproportionately influencing the results due to its scale.

Through these preprocessing steps, we ensured that our dataset was balanced, representative, and well-prepared for the subsequent machine learning modeling phase. As discussed above, we implemented six prominent machine learning algorithms: Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Random Forest, and Decision Tree. Among these models, the Random Forest algorithm achieved the highest accuracy, with a score of 77%.

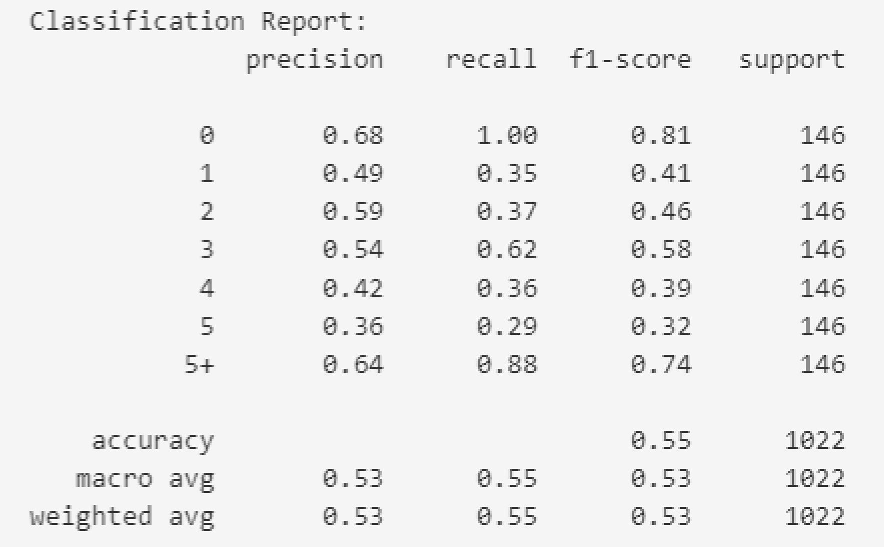
Among all of these models, there are some reasons why we think this model provided us the best result. Firstly, we need to talk about the ensemble learning process of Random forest. In this technique, random forest reduces the risk of overfitting by combining the predictions of multiple decision trees, which gives us more stable and reliable predictions. Also, this algorithm is well-suited for datasets with a large number of features, as it can handle feature interactions and dependencies effectively.

Chapter 4 Investigation/Experiment, Result, Analysis and Discussion

4.1 Results

After implementing and evaluating multiple machine learning models, we assessed their performance based on the F1 score, which balances precision and recall, providing a more comprehensive measure of model effectiveness in handling imbalanced datasets.

1. Logistic Regression: Logistic Regression, a fundamental binary classification algorithm, achieved an F1 score of 55%. This model, while straightforward and interpretable, struggled to capture the complexity of the relationships in our dataset, resulting in moderate performance.



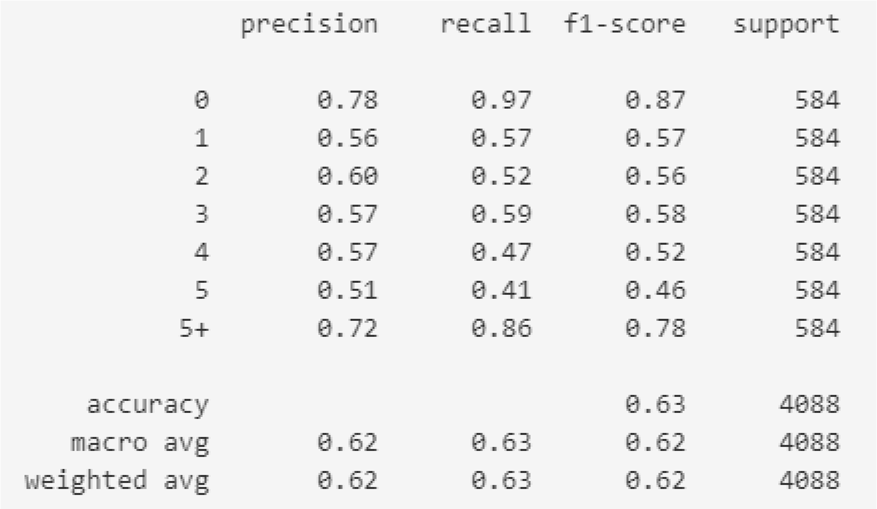
```
Classification Report:
              precision    recall  f1-score   support

           0       0.68      1.00      0.81       146
           1       0.49      0.35      0.41       146
           2       0.59      0.37      0.46       146
           3       0.54      0.62      0.58       146
           4       0.42      0.36      0.39       146
           5       0.36      0.29      0.32       146
          5+       0.64      0.88      0.74       146

    accuracy                           0.55      1022
   macro avg       0.53      0.55      0.53      1022
weighted avg       0.53      0.55      0.53      1022
```

Figure 5: Classification report for Logistic Regression

2. K-Nearest Neighbor (KNN): The K-Nearest Neighbor algorithm, known for its simplicity and instance-based learning approach, yielded an F1 score of 63%. KNN improved upon Logistic Regression by better capturing the local data structure, yet it was limited by its sensitivity to the high-dimensional feature space.



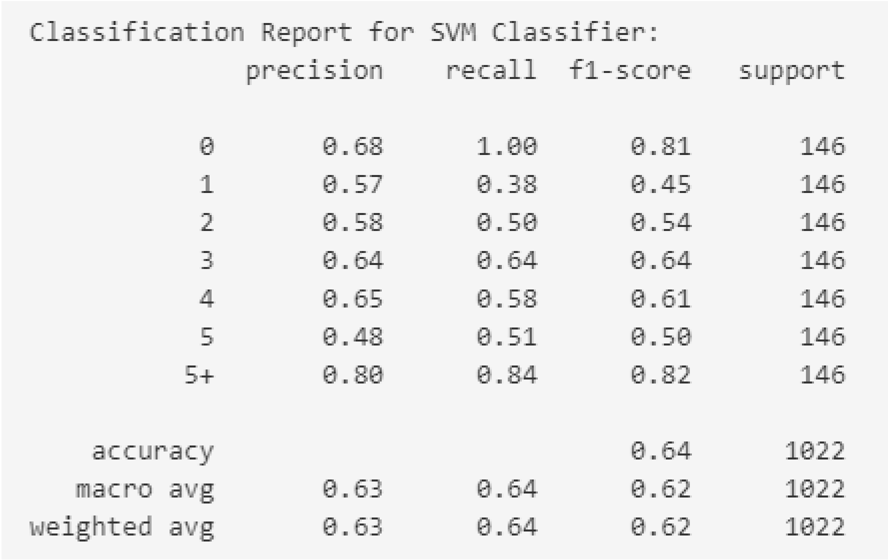
```
              precision    recall  f1-score   support

           0       0.78      0.97      0.87       584
           1       0.56      0.57      0.57       584
           2       0.60      0.52      0.56       584
           3       0.57      0.59      0.58       584
           4       0.57      0.47      0.52       584
           5       0.51      0.41      0.46       584
          5+       0.72      0.86      0.78       584

    accuracy                           0.63      4088
   macro avg       0.62      0.63      0.62      4088
weighted avg       0.62      0.63      0.62      4088
```

Figure 6: Classification report for K-Nearest Neighbor

3. Support Vector Machine (SVM): Support Vector Machine, effective for high-dimensional data, attained an F1 score of 67%. The SVM model demonstrated improved performance by optimizing the decision boundary, yet it required careful tuning of hyperparameters and was computationally intensive.



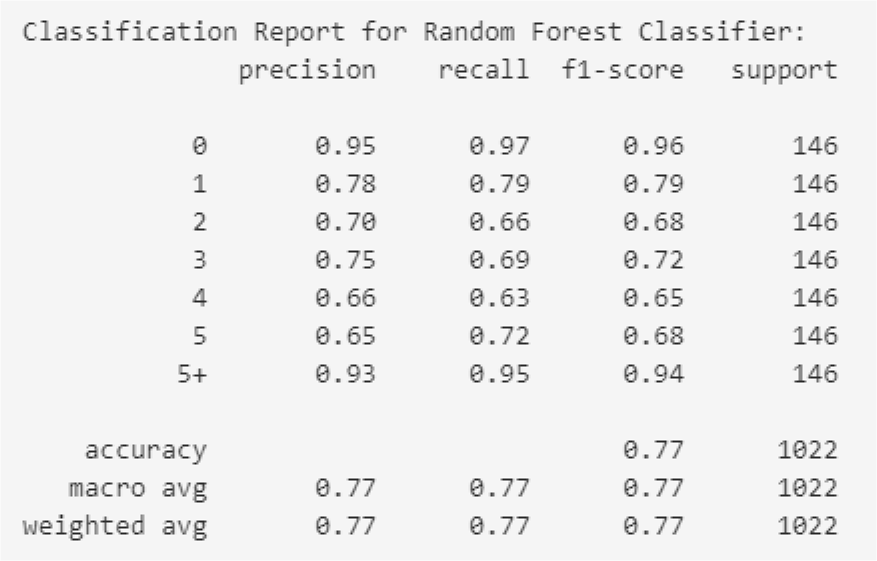
```
Classification Report for SVM Classifier:
              precision    recall  f1-score   support

           0       0.68      1.00      0.81       146
           1       0.57      0.38      0.45       146
           2       0.58      0.50      0.54       146
           3       0.64      0.64      0.64       146
           4       0.65      0.58      0.61       146
           5       0.48      0.51      0.50       146
          5+       0.80      0.84      0.82       146

    accuracy                           0.64      1022
   macro avg       0.63      0.64      0.62      1022
weighted avg       0.63      0.64      0.62      1022
```

Figure 7: Classification report for SVM Classifier

4. Random Forest: The Random Forest model, an ensemble learning method, achieved the highest F1 score of 77%. This superior performance can be attributed to the model's robustness and ability to aggregate multiple decision trees, thereby reducing variance and preventing overfitting. The ensemble approach enhanced predictive accuracy and provided a reliable classification outcome.



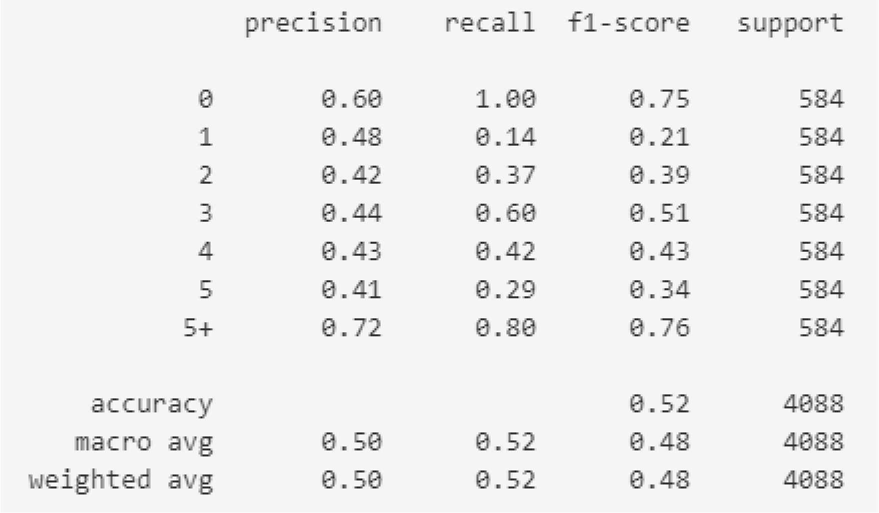
```
Classification Report for Random Forest Classifier:
             precision    recall  f1-score   support

           0       0.95      0.97      0.96       146
           1       0.78      0.79      0.79       146
           2       0.70      0.66      0.68       146
           3       0.75      0.69      0.72       146
           4       0.66      0.63      0.65       146
           5       0.65      0.72      0.68       146
          5+       0.93      0.95      0.94       146

    accuracy                           0.77      1022
   macro avg       0.77      0.77      0.77      1022
weighted avg       0.77      0.77      0.77      1022
```

Figure 8: Classification report for Random Forest

5. Decision Tree: Lastly, the Decision Tree model recorded an F1 score of 52%. Despite its interpretability and simplicity, the Decision Tree's tendency to overfit the training data led to the lowest performance among the models tested, highlighting its limitations in handling the dataset's complexity.



```
             precision    recall  f1-score   support

           0       0.60      1.00      0.75       584
           1       0.48      0.14      0.21       584
           2       0.42      0.37      0.39       584
           3       0.44      0.60      0.51       584
           4       0.43      0.42      0.43       584
           5       0.41      0.29      0.34       584
          5+       0.72      0.80      0.76       584

    accuracy                           0.52      4088
   macro avg       0.50      0.52      0.48      4088
weighted avg       0.50      0.52      0.48      4088
```

Figure 9: Classification report for Decision Tree

Overall, the Random Forest model outperformed the others, demonstrating its effectiveness in capturing intricate patterns and relationships within the data. Its superior F1 score underscores the benefits of ensemble learning techniques in improving classification performance, especially in datasets with diverse and imbalanced classes.

4.2 Analysis and Discussion

Student absenteeism is a critical issue in educational institutions, affecting not only individual academic performance but also overall educational outcomes. High rates of absenteeism can lead to lower grades, increased dropout rates, and diminished prospects for future success [19]. By understanding the factors that contribute to absenteeism, educational institutions can develop targeted interventions to improve attendance and support student achievement.

Our study aimed to predict student absenteeism using machine learning models, focusing on seven key features: Physical Health Issues, Mental Health Issues, Financial Issues, Learning Difficulties, Social Communication Issues, University Environment, and Transport Issues. These features were identified as potential contributors to student absenteeism based on their impact on students' ability to attend classes regularly. After implementing various machine learning models, we evaluated their performance using the F1 score, a metric that balances precision and recall [20]. Out of these models, the Random Forest model demonstrated the highest F1 score, indicating its superior ability to handle the complexities and nuances within the dataset. This can be attributed to the model's ensemble nature, which combines multiple decision trees to reduce variance and improve predictive accuracy. The Random Forest model's robustness and capacity to capture complex interactions among features made it the most effective tool for predicting student absenteeism in our study.

This study demonstrates the applicability of machine learning techniques in predicting student absenteeism, offering a robust framework for educational institutions to identify and address factors contributing to absenteeism proactively. The comprehensive analysis of different models and their performance highlights the importance of choosing appropriate algorithms and preprocessing techniques to achieve optimal results.

Furthermore, the study emphasizes the critical role of data preprocessing, such as balancing the dataset using SMOTE and normalizing the data, in enhancing model performance. The insights gained from feature importance analysis can guide future research and interventions aimed at improving student retention and academic success. In summary, the Random Forest model emerged as the most effective tool for predicting student absenteeism, providing a balance of accuracy, robustness, and interpretability. Future research could explore hybrid models or advanced techniques such as deep learning to further enhance predictive performance and uncover deeper insights into the factors influencing student attendance. By addressing the key features identified in this study, educational institutions can take proactive steps to reduce absenteeism and support student success.

Chapter 5 Conclusions

5.1 Summary

From response of questions and findings, it can be concluded that absenteeism effect on student's performance. Student who became absent have poor class participation, poor coordination with teachers and peers and poor CGPA. Student's class participation becomes affected due to absenteeism [21]. The effects of absenteeism in class participation, they miss the chance to become a part in class participate, can't raise questions about any confusion regarding topics. As per data, 77% participants are strongly agree that the absenteeism has effects on students' class participation. Absenteeism leads the students to drop out graded activities. Poor performances in class quiz which lead to poor CGPA. They are unable to As a result of absenteeism, the relationship between the student and the teacher as well as between the school and the parents may be damaged. Due to absent students' learning deficiencies, teachers face difficult situations in classroom management, and schools deviate from their goals. The workload of school administrators increases, their time is spent on procedures related to absenteeism instead of instructional issues, and therefore, some of the resources become wasted. According to Taymaz (2000), one of the tasks that school administrations deal with most regarding student services is monitoring student attendance [22]. The compromises that school administrators will make on this issue may lead to the continuation of the absenteeism problem.

Prepare and the assignments on due dates which badly effect the GPA of students. Therefore, it is concluded independent variable Absenteeism is strongly impacted on dependent variable on Student Performance.

5.2 Limitations

- Conceptual Framework: The theoretical framework used to analyze absenteeism might have limitations in its applicability or relevance to the specific context of the project. For example, if the framework is based on studies conducted in different universities or even in different workplaces, it's applicability to the project's setting may be limited.
- Generalizability: Theoretical models of absenteeism may be based on studies with specific populations, such as employees in a particular university student in a certain

demographic. The extent to which findings from these studies can be generalized to the population or context of the project may be limited.

- Diverse Student Population: Universities typically have a diverse student body with varying demographics, academic backgrounds, and personal circumstances. This diversity can make it challenging to identify common causes of absenteeism or implement uniform interventions that cater to all students' needs.
- Student Mental Health: Increasing awareness of mental health issues among students may contribute to absenteeism. However, addressing these issues requires sensitive handling and consideration of privacy and confidentiality concerns.
- Academic Calendar: The academic calendar, including breaks, exams, and holidays, can influence patterns of absenteeism. Seasonal variations may affect attendance rates differently across different academic programs and student groups.
- Impact of External Factors: External factors such as commuting distance, health issues, employment obligations, and family responsibilities can significantly influence student attendance and absenteeism rates.
- Student Engagement: Levels of student engagement and motivation can affect attendance rates. Addressing absenteeism may require strategies to enhance student motivation and involvement in academic and extracurricular activities.

Legal and Ethical Considerations: Universities must adhere to legal requirements and ethical standards when collecting and using data related to student attendance and reasons for absenteeism. Privacy laws and regulations, such as GDPR or FERPA, must be respected.

5.3 Future Improvement

Improving a project focused on absenteeism in a university (or any organization) involves continuous refinement and adaptation based on ongoing evaluation and feedback. Here are several future improvements that could enhance the effectiveness of an absenteeism project:

- Enhanced Data Collection and Analysis: Conduct more detailed analysis to identify patterns and trends in absenteeism across different student demographics, courses, and time periods.

- Utilization of Technology: Utilize data analytics and predictive modeling to forecast absenteeism trends and allocate resources more effectively.
- Continuous Monitoring and Evaluation: Establish regular monitoring and evaluation processes to assess the impact of interventions on absenteeism rates and student outcomes.
- Adaptation to Changing Circumstances: Continuously update strategies and interventions based on new research findings, best practices, and emerging trends in higher education and student support.

Longitudinal Studies and Outcome Measurement: Measure outcomes beyond attendance rates, such as graduation rates and career readiness, to gauge the comprehensive impact of the project.

References

[1]. "Definition of absenteeism." https://www.merriam-webster.com/dictionary/absenteeism.

[2]. "Addressing Absenteeism: Causes and Teacher Solutions," Varthana, Sep. 10, 2023. https://varthana.com/school/breaking-down-the-causes-of-absenteeism-and-solutions-teachers-can-use-to-prevent-it/.

[3]. H. Patel, "(PDF) Truancy Prevention Efforts".

[4]. M. J. Fornander and C. A. Kearney, "Family Environment Variables as Predictors of School Absenteeism Severity at Multiple Levels: Ensemble and Classification and Regression Tree Analysis," Frontiers in psychology, vol. 10, p. 2381, Oct. 2019, doi: 10.3389/fpsyg.2019.02381.

[5]. HamoyeHQ, "Machine Learning for Attendance Rate Prediction (Educational infrastructure)," Medium, Feb. 06, 2024. [Online]. Available: https://hamoyehq.medium.com/machine-learning-for-attendance-rate-prediction-educational-infrastructure-86ec4a3c33b0

[6]."(PDF) Predicting and Analyzing Student Absenteeism Using Machine Learning Algorithm," ResearchGate. Available: https://www.researchgate.net/publication/361647479_Predicting_and_Analyzing_Student_Absenteeism_Using_Machine_Learning_Algorithm

[7]."(PDF) Predicting Absenteeism at Work Using Machine Learning Algorithms," ResearchGate. https://www.researchgate.net/publication/350955612_Predicting_Absenteeism_at_Work_Using_Machine_Learning_Algorithms

[8]."(PDF) Incorporating a Machine Learning Model into a Web-Based Administrative Decision Support Tool for Predicting Workplace Absenteeism," ResearchGate. https://www.researchgate.net/publication/361669573_Incorporating_a_Machine_Learning_Model_into_a_Web-Based_Administrative_Decision_Support_Tool_for_Predicting_Workplace_Absenteeism

[9]. "Feature importances with a forest of trees," scikit-learn. https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html.

[10]. GeeksforGeeks, "SMOTE for Imbalanced Classification with Python," GeeksforGeeks, May 03, 2024. [Online]. Available:

https://www.geeksforgeeks.org/smote-for-imbalanced-classification-with-python/.

[11]. “What is Normalization in Machine Learning? Techniques &#38; Uses,” Deepchecks, Jul. 16, 2021. https://deepchecks.com/glossary/normalization-in-machine-learning/.

[12]. “Stratified Sampling - an overview,” ScienceDirect Topics. https://www.sciencedirect.com/topics/mathematics/stratified-sampling.

[13]. GeeksforGeeks, “Logistic Regression in Machine Learning,” GeeksforGeeks, May 09, 2017. [Online]. Available: https://www.geeksforgeeks.org/understanding-logistic-regression/.

[14]. GeeksforGeeks, “K-Nearest Neighbor (KNN) Algorithm,” GeeksforGeeks, Apr. 14, 2017. [Online]. Available: https://www.geeksforgeeks.org/k-nearest-neighbours/.

[15]. GeeksforGeeks, “Support Vector Machine (SVM) Algorithm,” GeeksforGeeks, Jan. 20, 2021. [Online]. Available: https://www.geeksforgeeks.org/support-vector-machine-algorithm/.

[16]. Simplilearn, “Random Forest Algorithm,” Simplilearn, Apr. 22, 2020. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm.

[17]. “Exploratory Data Analysis,” US EPA, Apr. 02, 2015. https://www.epa.gov/caddis/exploratory-data-analysis.

[18]. S. SATPATHY, “SMOTE for Imbalanced Classification with Python,” Analytics Vidhya, Oct. 06, 2020. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/.

[19]. J. Dräger, M. Klein, and E. Sosu, “The long-term consequences of early school absences for educational attainment and labour market outcomes,” British Educational Research Journal, Feb. 2024, doi: 10.1002/berj.3992.

[20]. GeeksforGeeks, “F1 Score in Machine Learning,” GeeksforGeeks, Dec. 27, 2023. Accessed: Jun. 19, 2024. [Online]. Available: https://www.geeksforgeeks.org/f1-score-in-machine-learning/

[21]. M. AKKUŞ and Ş. ÇINKIR, “The Problem of Student Absenteeism, Its Impact on Educational Environments, and The Evaluation of Current Policies,” International Journal

of Psychology and Educational Studies, vol. 9, pp. 978–997, Oct. 2022, doi: 10.52380/ijpes.2022.9.4.957

[22]. "A Principal-Led Team Overseeing Attendance," Attendance Works, Oct. 03, 2017. https://www.attendanceworks.org/resources/toolkits/mentoring-elementary-success-mentors/what-support-is-needed-from-schools/a-principal-led-team-overseeing-attendance/.