A Predictive Framework Based on Students' Academic Performance in Higher Education

"إطار تنبؤي قائم على الأداء الأكاديمي لطلاب التعليم العالي"

Mohamed Ali Elhayes[1], Osama Mohammed El-Deeb[2], Abdelaziz Fathy Abdelaziz[3]

[1]Faculty of Business, Economics & Information Systems Misr University for Science& Technology, Giza, 12572 Egypt.
[2] Giza Higher Institute for Managerial Sciences, Tomah, Egypt, 12913 Egypt
[3] Giza Higher Institute for Managerial Sciences, Tomah, Egypt, 12913 Egypt

Volume three - Issue seven - February 2025

المجلد الثالث - العدد السابع – فبراير ٢٠٢٥

ISSN-Online: 2812-6122 ISSN-Print: 2812-6114

موقع المجلة على بنك المعرفة المصري

https://aiis.journals.ekb.eg/contacts?lang=ar

Abstract

This research proposes a predictive framework for assessing the academic performance of students in higher education, leveraging data mining and machine learning techniques. The study addresses the challenge of imbalanced datasets, which often skew the performance of classification models, by employing the Synthetic Minority Oversampling Technique (SMOTE) to enhance prediction accuracy. The framework integrates various supervised learning algorithms, including J48, Random Forest (RF), and K-Nearest Neighbors (KNN), to predict Students' Academic Performance and recommend suitable academic paths based on historical data. The data set, collected from the Higher Institute for Management Sciences, spans multiple academic years, and includes student records from three departments: Information Systems, Management, and Accounting. The research demonstrates that handling class imbalance through SMOTE significantly improves model performance, with the Random Forest classifier achieving the highest accuracy of 93.70%. The study also highlights the importance of feature selection, data preprocessing, and normalization in optimizing predictive outcomes. The findings underscore the potential of educational data mining to support early academic interventions and personalized guidance, aiding students in making informed decisions about their academic trajectories. Future work will explore additional sampling techniques and expand the dataset to further enhance model accuracy.

Keywords: Students performance, SMOTE, Educational data mining, Data mining algorithms.

إطار تنبؤي قائم على أداء الطلاب في التعليم العالي

المستخلص:

يقترح البحث إطارًا تنبؤيًا لتقييم الأداء الأكاديمي للطلاب في التعليم العالي، والاستفادة من تقنيات استخراج البيانات والتعلم الآلي. وتتناول الدراسة تحدي مجموعات البيانات غير المتوازنة، والتي غالبًا ما تشوه أداء نماذج التصنيف، من خلال استخدام تقنية أخذ العينات الزائدة عن الحد من الأقليات الاصطناعية (SMOTE) لتعزيز دقة التنبؤ. يدمج الإطار خوارزميات التعلم الخاضع للإشراف المختلفة، بما في ذلك J48، Random Forest (RF)، K-Nearest Neighbors (KNN)، للتنبؤ بأداء الطلاب والتوصية بالمسارات الأكاديمية المناسبة بناءً على البيانات التاريخية. تمتد مجموعة البيانات، التي تم جمعها من المعهد العالي لعلوم الإدارة، على سنوات أكاديمية متعددة وتتضمن سجلات الطلاب من ثلاثة أقسام: نظم المعلومات والإدارة والمحاسبة. يوضح البحث أن التعامل مع اختلال التوازن الطبقي من خلال SMOTE يحسن بشكل كبير من أداء النموذج، حيث حقق Random Forest أعلى دقة بنسبة ٩٣,٧٠٪. وتسلط الدراسة الضوء أيضًا على أهمية اختيار الميزات ومعالجة البيانات مسبقًا والتطبيع في تحسين النتائج التنبؤية. تؤكد النتائج على إمكانات التنقيب عن البيانات التعليمية لدعم التدخلات الأكاديمية المبكرة والتوجيه الشخصي، ومساعدة الطلاب في اتخاذ قرارات مستنيرة حول مساراتهم الأكاديمية . سيستكشف العمل المستقبلي تقنيات أخذ العينات الإضافية وتوسيع مجموعة البيانات لزيادة تحسين دقة النموذج.

الكلمات المفتاحية: أداء الطلاب، SMOTE، التنقيب عن البيانات التعليمية، خوارزميات التنقيب عن البيانات.

1. Introduction

Higher education institutions face the difficulty of efficiently using enormous volumes of student data to improve educational outcomes in the age of abundant data. Predicting pupils' academic performance helps with both individualized learning experiences and early intervention techniques. Accurately forecasting future results and understanding the complexity of individual student paths are two areas where traditional evaluation approaches frequently fail. As a result, the demand for complex prediction frameworks that make use of developments in data mining and machine learning is rising.

To assess and anticipate students' academic performance in higher education settings, this study suggests a thorough predictive framework. Through the utilization of data mining tools, significant patterns can be extracted from past student data, coupled with machine learning algorithms capable of predictive modelling, this framework aims to provide actionable insights for educators and administrators. Such insights can enable proactive measures to support struggling students, optimize resource allocation, and ultimately foster a conducive learning environment.

The use of strong machine learning models like decision trees (J48), Random Forest, and K-Nearest Neighbors (KNN) for prediction, feature selection strategies to find important predictors of academic success, and data preprocessing to guarantee data quality and relevance are essential elements of this framework. This study aims to support a more data-driven approach to student support and institutional decision-making in higher education by combining these approaches.

This research is significant as it has the potential to change the way that educational institutions manage student assessment and support systems, making them more proactive rather than reactive. Ultimately, using such predictive frameworks could raise graduation and retention rates while also improving the general caliber of the educational process.

2. Problem Statement

Academic guidance is usually given based on the educational institute predefined rules and procedures. However, guidance would be more beneficial and effective if it relays on analyzing the student's academic performance along with his/her demographic data.

However, with the abundance of research in the field to determine the academic performance of students to help devices early and appropriate educational interventions, the gap between research findings and industry application continues to pervade. Most of the suggested papers solely address prediction methods in relation to the available datasets. However, very few are designed to be incorporated into the platforms that these institutions' educational decision-makers use. Most real-world data sets in this field are unbalanced, which makes quality time-series classification challenging because examples in each class are unevenly distributed. This is why this field of study is regarded as challenging. Due to the classifiers' tendency to Favor the majority classes over the minority classes, this problem impairs the performance of the classification algorithms.

3. Research Questions

- What are the appropriate data mining techniques used to improve the early prediction of Students' Academic Performance predictions?
- What is the appropriate method that can be used to overcome the problem of the imbalanced dataset?

4. Research Importance

- Guidance of the student to decide to choose the department that suits his capabilities.
- Instructing the student to choose the department's materials to improve his GPA.
- Instructing the student to choose the sub-section for the department to allocate the appropriate field for it.
- Provide time and effort for the student in the department counselling and specialization process.

5. Research Objectives

- After preprocessing the data to get the highest accuracy, propose an integrated model that uses regression techniques to provide an early prediction of Students' Academic Performance.
- Examine various approaches to dealing with unbalanced data and how they affect the accuracy of predictions.
- Using classification techniques, suggest a study path for a student based on the features that are available.

6. Research Methodology

This research will extract features of real dataset and classify it to a more accurate algorithms to improve prediction Students' Academic Performance accuracy.

7. Method of Data Collection

data is collected from the high institute for management sciences students' records. There are three different departments of the institute. Data was collected from between the semester (2015/2016) through to semester (2018/2019). with (2869 records of student attributes).

Secondary sources of data such as articles, periodicals, books, and websites are used to determine the theoretical framework of the study and identify the research problem.

8. Research Limitations

These search restrictions fall into three types:

- Apply the proposed framework in the higher education environment.

  During the work of this study, several experiments were conducted to implement this framework in one of the higher institutes.

- The education sector has changed students 'records (privacy).

  During the many negotiations that took place with the management of the Institute, the results demonstrated that all of them are keen on student data to maintain student privacy.

- Diversity of students and the size of the dataset.

The case study that was achieved in this research is to predict Students' Academic Performance, but this research aims to obtain datasets for a variety of students and to apply the model in other higher institute sectors.

9. Background

An outline of the basic theoretical underpinnings of data mining, classification, and methodology will be given in this part. Providing a summary of the main approaches currently used to address the imbalanced class distribution problem, which is used to address imbalanced data sets having two or more classes.

Supervised Learning

One kind of machine learning is supervised learning, in which a labelled dataset is used to train the model. The term "labelled" here refers to the fact that every training example has an output label (the right response). Using these sample input-output pairings, supervised learning aims to develop a mapping from inputs to outputs. One of

the most popular and well-understood forms of machine learning is supervised learning, which makes it a cornerstone of the discipline (Ramasubramanian & Singh, 2017).

Unsupervised Learning

Unsupervised learning is a type of machine learning where the model is trained on data that does not have labeled outputs. Unlike supervised learning, where the goal is to learn a mapping from inputs to outputs, unsupervised learning aims to find hidden patterns or intrinsic structures in the input data. Unsupervised learning is a powerful tool for exploring and understanding the underlying structure of data, making it invaluable in many real-world applications (Ramasubramanian & Singh, 2017).

Data Mining

Finding patterns, correlations, and anomalies in big data sets to glean valuable information and create prediction models is known as data mining. To evaluate data from many angles and condense it into actionable knowledge, it combines statistics, machine learning, database systems, and data visualization approaches (Ramasubramanian & Singh, 2017).

9.1Algorithms for Supervised Learning

– **Linear Regression**

A basic statistical and machine learning method for simulating the relationship between a dependent variable (goal) and one or more independent variables (features) is linear regression. The objective is to use the input information to determine which linear relationship best predicts the target variable. The mathematical function that links a dependent variable (y) and one or more independent variables (x) is known as linear regression. When we assume that there is an underlying reason in the data, we can utilize linear regression, which shows that when the value of the independent variable changes, the dependent variable also changes in a linear fashion. The correlation between variables is represented by the straight-line sloping in a positive way in the model that provides linear regression. A popular and adaptable method, linear regression forms the basis of many more intricate models in statistics and machine learning (Hussain, Gaftandzhieva, Maniruzzaman, Doneva, & Muhsin, 2021).

– **Logistic Regression**

A statistical and machine learning method for binary classification problems, logistic regression aims to forecast the likelihood that an instance will fall into a specific class. Logistic regression forecasts probabilities that are subsequently translated

to discrete classes (usually 0 or 1), in contrast to linear regression, which forecasts continuous values. For binary classification challenges, logistic regression is a strong and popular method that strikes a good compromise between interpretability and simplicity (logistic regression, 2024).

– **K-Nearest Neighbors (KNN)**

KNN is a straightforward, instance-based, non-parametric machine learning algorithm that may be applied to both regression and classification problems. The fundamental concept of KNN is to use the labels or values of the 'k' closest data points in the training dataset to predict the label or value of a new data point. For small to medium-sized datasets, KNN is a flexible and user-friendly method that can be highly successful, particularly when the data includes a distinct and significant distance metric (Zul, 2016).

– **Support Vector Machine (SVM)**

A potent supervised machine learning technique for classification, regression, and outlier detection applications is the Support Vector Machine (SVM). It operates by identifying the best hyperplane in the input data to divide the various classes. The primary goal is to increase the margin between classes to improve generalization to data that has not yet been observed. SVM uses many kernel functions to manage both linear and non-linear data. Because it works well in high-dimensional environments, it is frequently utilized in a variety of domains, including text categorization, bioinformatics, and image classification)Asaad و Abdulhakim(٢٠٢١ ،) Awad و Khanna(٢٠١٥ ،.

– **Naïve Bayes**

Based on Bayes' Theorem, the Naïve Bayes family of probabilistic algorithms is straightforward yet effective for classification applications. Given the class title, it presumes that the features are independent of one another (hence "naïve"). In fact, Naïve Bayes frequently outperforms this strong assumption, particularly in text classification, spam filtering, sentiment analysis, and medical diagnosis)Amra و Maghari(٢٠١٧ ،.

– **Decision Tree (J48)**

A supervised learning approach for classification and regression problems is called a decision tree. It uses a structure resembling a tree to model data, with each node standing for a feature-based decision and each branch for the decision's result. The anticipated class or value is represented by the tree's last leaves) Aljawarneh ،Yassein ، و Aljundi(٢٠١٩ ، (Patil & Hiremath, 2022).

– **Random Forest**

A potent ensemble learning technique for classification and regression applications is Random Forest. To increase accuracy and decrease overfitting, it constructs several decision trees and aggregates their results. One of the greatest general-purpose machine learning algorithms, particularly for structured data, is Random Forest. It is a great option for many real-world applications because of its excellent precision, durability, and adaptability) Abdulkareem و Abdulazeez(٢٠٢١ ،)Ghosh و Janan(٢٠٢١ ،.

9.2Techniques for Handling Class Imbalance Distribution

Biassed models that favor the majority class result from class imbalance, which happens when one class in a dataset outnumbers another. Here are a few practical methods to deal with this problem:

– **Under Sampling**

To balance the dataset and address class imbalance, under sampling is a technique that involves lowering the number of samples from the majority class. This lessens the likelihood that the model will favor the majority class. Some of the most popular and widely applied under-sampling techniques are listed below:

Random Under sampling

Samples from the majority classes are randomly removed using the random under sampling technique. Random under sampling is a non-heuristic method for under sampling. It is frequently used as a beginning point and is among the simplest tactics. In random under sampling, samples from the majority or negative class are selected at random and then eliminated until the number of samples from the minority or positive class equals that of the majority class. This technique is referred to as "random under sampling" since it selects samples from the majority class at random. With an equal number of positive and negative data samples, this approach will produce balanced data collection (Silveira, et al., 2022).

Tomek Link

Tomek Link is a distance-based heuristic under sampling algorithm. Tomek Connection establishes a link based on the distance between misclassified instances from two distinct classes, which is then remove majority class instances (Muntasir Nishat, Faisal, Jahan Ratul, Al-Monsur, & Ar-Rafi, 2022).

Edited Nearest Neighbors (ENN)

Another approach for picking cases for deletion is the Edited Nearest Neighbors rule (ENN). This rule includes finding misclassified cases in a dataset and eliminating them using k nearest neighbors =3. On the same dataset, The ENN technique can be repeated numerous times to improve the selection of samples in the majority class. This extension was originally known as "unlimited editing," but it is now more widely known as Repeatedly Edited Nearest Neighbors (Muntasir Nishat, Faisal, Jahan Ratul, Al-Monsur, & Ar-Rafi, 2022).

– **Over Sampling**

By increasing the quantity of samples in the minority class, oversampling is a technique used to address class imbalance. This keeps the model from being biassed towards the majority class and improves its ability to learn from under-represented data. The following are some of the more popular and applied oversampling techniques:

Random Over Sampling

Classes that use random oversampling locate samples at random and replicate them in the minority class. By randomly duplicating select data instances from minority classes to enhance their samples, we were able to address the dataset's class imbalance problem (Muntasir Nishat, Faisal, Jahan Ratul, Al-Monsur, & Ar-Rafi, 2022).

Borderline-SMOTE

The proposed algorithm was called Bordered-SMOTE, which is an Extended version of SMOTE. Only K nearest neighbor used by Borderline-SMOTE for make synthetic data along the decision boundary of two classes (Muntasir Nishat, Faisal, Jahan Ratul, Al-Monsur, & Ar-Rafi, 2022).

Adaptive Synthetic Sampling (ADASYN)

In ADASYN, the density distribution is used to specify how many synthetic samples (of minority class) will be generated. ADASYN generates minority data samples intelligently according to their distribution ratios from the overall datasets (Balaram & Vasundra, 2022).

Class imbalance is one of the main characteristics of educational data that is the effect of model performance. So, in this study, we solve this problem using SMOTE oversampling technique. SMOTE enhances the model's performance and gives the best results.

9.3Evaluations Metrics

Evaluation metrics measure a machine learning model's performance and effectiveness in making predictions. The most common metrics for classification problems include accuracy, precision, recall, and F-measure (F1-score) (Hersh, 2021).

– **Accuracy (AC)**: Measures the percentage of correctly predicted instances out of all instances (Patra & Patro, 2014) (Zeng, 2019).

$$\boldsymbol{Accuracy} = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Instances}$$

– **Recall**: Measures the proportion of correctly predicted positive instances out of all actual positive instances (Sokolova & Lapalme, 2009) (Gray, Bowes, Davey, Sun, & Christianson, 2011).

$$\boldsymbol{Recall} = \frac{\boldsymbol{TruePositives}}{\boldsymbol{TruePositives + falseNegatives}}$$

– **Precision:** Measures the proportion of correctly predicted positive instances out of all predicted positives (Sokolova & Lapalme, 2009) (Gray, Bowes, Davey, Sun, & Christianson, 2011).

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positives}}{\boldsymbol{True\ Positives + False\ Positives}}$$

– **F- Scores:** Harmonic mean of precision and recall, balancing both metrics (Ferri, Hernández-Orallo, & R, 2009).

$$\mathbf{F - Scores} = 2 * \frac{(\boldsymbol{precision})(\boldsymbol{recall})}{\boldsymbol{precision + recall}}$$

10.Literature Review

Various Data mining techniques and machine learning algorithms have been used by different researchers to predict academic performance in students.

The authors in (Alla Vladova & Katsiaryna M. Borchyk, 2024) propose a predictive framework to study student performance at higher education by using a multi-method approach. It uses logistic and linear regressions, k-means clustering and data normalization of features to achieve more correct predictions. Essential elements comprise predicting students, at an academic risk, assessing feature importance, and

predicting term marks. They managed 90% predictive accuracy in determining whether students would pass an exam on the platform and 70% accuracy in determining performance by individual assessments, allowing for personalized consultations with students who progressed similarly.

The study (Ibrahim H. Ibrahim, Etemi J. Garba, Uwaisu A. Umar, & Adedeji A. Adejumo, 2024) offers a predictive model for assessing the academic achievement of students in higher education, specifically at Modibbo Adama University. Using machine learning algorithms like Random Forest, the model was able to attain 95% accuracy. Among the significant factors found are the professor explanation, infrastructure quality, sponsor support, and lecture attendance. These results emphasize the necessity of having a comprehensive understanding of the numerous factors that can impact on instruction and learning to create settings that will support students' success.

The paper (Jamal Eddine, Zakrani, Mohammed, Said, & Abdellah, 2025) uses machine learning techniques to develop a predictive model for academic achievement in online learning environments. To improve prediction accuracy, it incorporates demographic, social, emotional, and cognitive factors. Both explicit traces, such as demographic information, and implicit traces from student interactions are processed by the model. This method seeks to create a more flexible educational experience catered to the needs of individual students in higher education by not just forecasting performance but also providing insights into skill learning and personal growth.

In (Sixuan & Bin, 2024) Proposed XGB-SHAP which is the combination of extending a dumping, which can be carried out by XGBoost and SHAP (SHapley Additive exPlanations) to predict academic achievement of students. It works on a dataset from a public university and achieves an MAE of around 6 and R-square value of 0.82, higher than that of other machine learning models. The results highlight the importance of skills in self-directed learning and the impact of instructional modalities towards academic performance and recommend tailored feature selection in predictive models.

The study (Zheng, et al., 2024) proposes a predictive framework for analyzing Students' Academic Performance in higher education by incorporating multidimensional spatiotemporal features. They applied on a dataset by merging the fundamental data of students, their performance throughout the semester, and educational markers from their home countries. This dataset was used to train six machine learning models, all of which showed accuracy in predicting academic

achievement. By making appropriate use of these multifaceted characteristics in actual educational settings, the framework seeks to direct instructional practices.

In the study (Kenth C. Novo, 2024) Three algorithms—Naive Bayes, Multilayer Perceptron, and C4.5 Decision Tree—are used in data mining approaches to predict students' academic achievement. It provides some insightful information on how these methods might be applied to identify correlation in academic datasets and, in practice, categorize student achievement. The findings show that the Naive Bayes algorithm performs better than the other predictive techniques, suggesting that it is a viable predictive framework that may enhance learning outcomes and offer tailored interventions in institutions of higher learning.

In this research, (Vivi Nur Wijayaningrum, Ika Kusumaning Putri, & Annisa Puspa Kirana, 2024) proposed a predictive framework for Academics performance of Students in Higher Education which is picking Vocational College students. It uses SMOTE technique for handling imbalanced data and Random Forest for feature selection to create a MLP classification model. This MLP model has two hidden layers and yielded an accuracy of 0.8889 and an F1-Score of 0.9032, accounting for its powers of identifying whether a student is most likely to drop out and enabling teachers that they can take proactive actions.

This research (Isreal Ogundele, Olutosin Taiwo, Asegunloluwa Babalola, & Olumide Ayeni, 2024) provides a deterministic approach using linear regression model to analyze and predict education performance in higher education. There are several factors involved, such as length of lecture, attendance, gender, and mode of learning, all of which contribute to helping education outcomes. It was able to achieve 88.3% accuracy (1.6429 RMSE and 1.14043 MAE), demonstrating the ability of predicting a student's performance based on the historical data.

In (Abdullah, et al., 2024), a prediction framework for academic performance in higher education is proposed based on learning management system multimedia data. It uses features from a CNN combined with ML models, Specifically, a combination of random forest (RF) and support vector machine (SVM). With 97.88% prediction accuracy, high precision, recall, and an F1 score of 98%, it increases prediction accuracy and identifies pupils who are in danger of performing poorly.

In (Stapel, Zheng, & Pinkwart, 2016), a model is proposed to predict students' performance in an online learning environment. Using Math as teaching material, has

been evaluated the model. Prediction attempts are concerned about a formal way which the students are required to complete, such as an institute course or an online-only course with a complete structure. Being different from course-centered online-only or mixed forms of learning, this math learning platform does not have its user taking a formal course to achieve the learning objectives. The classifiers applied to the operation scopes are K-Nearest Neighbor, Logistic Regression, AdaBoost, Naïve Bayes implementation, Decision Tree, and Random Forest. Then finally, the previous classifiers are used to predict the performance of students in an ensemble that has achieved 73.5 % accuracy.

The student's academic fortune was investigated through (Dengen, Budiman, Wati, & Hairah, 2018), being measured with grade point average (GPA). The study used 279 student records with features of students' data consisting of gender, status, high school, place of birth, age, high school department, organizational behavior, age when starting the high school stage, and GPA averages and overall GPA. It was first normalized using integration, cleaned, and transformed. They used the Naïve Bayes algorithm, which yielded 76.79% accuracy. Using 80% of the dataset to train the model we achieved around 90% accuracy.

In (Razaque, et al., 2017) the authors Cognitive proposed to use Naïve Byes classifier as a technique for predicting student performance. Internal evaluation and the final semester examination based on the student's success in educational activities such as discussion, quizzes, lab work, attendance and assignments were found to decide the academic achievement of students. Data comes from the Department of Computer Science. Rows After Dealing Missing 500+ Students Used out of 660 rows, because 160 entries are with lost value. The volume of documents has been limited to 500 records. Accuracy of Naïve byes algorithm used in study was 93.79 %. However, limited knowledge exists regarding student features.

Research in (Widyahastuti & Tjhin, 2017) proposes a model for Students' Academic Performance predicting in the final exam using Multilayer Perceptron and Linear Regression. Data is collected from 50 undergraduate Information System Management departments. Using different attributes from the database such as students' information, students' demographics, behavior, socioeconomic, and psychological they applied linear regression and multilayer perception classifier and the accuracy was 0.82%, and 0.84%, respectively.

This study (Agung Triayudi, Rima Tamara Aldisa, & S. Sumiati, 2024) employs Educational Data Mining to anticipate higher education student learning outcomes. This uses algorithms like Decision Tree, Neural Networks, and Naïve Bayes for the analysis of academic records and socio-economic data from 300 students from Information Systems and Informatics programs. Students have many challenges to encounter, such as failure in study and drop out risks, providing significant impact for these students can be a solution, thus the framework is a solution to the academic guidance by accurate prediction. This approach emphasizes the importance of data-driven insights in educational settings.

The research (Ekemini A. Johnson, et al., 2024) suggests a sophisticated analytical framework for forecasting students' performance in higher education that uses the Random Forest (RF) and Multiple Linear Regression (MLR) methods. The study, which is based on the analysis of 664 datasets from Federal Polytechnic Ukana, demonstrated that the RF model performed better than MLR in terms of prediction accuracy and was able to capture non-linear interactions between predictor and response variables. Analyzing all this data can reveal information about Students' Academic Performance, assist in identifying those who are at risk, and implement suitable solutions that maximize learning results.

11. The Proposed Model of Performance Student

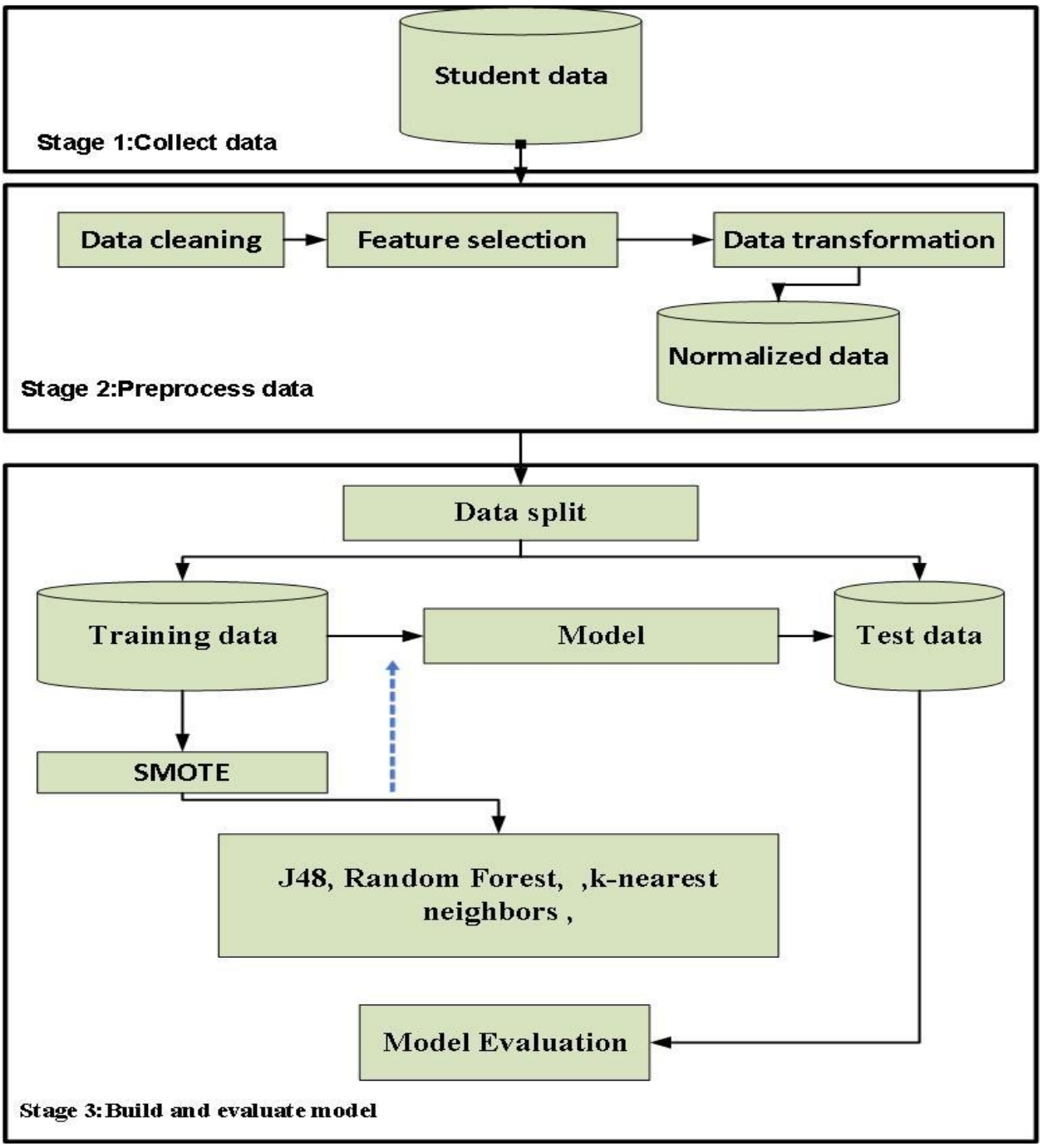

Figure 1: The Proposed Model of Performance Student

In this way the proposed model predicts the academic performance of students in the multi-disciplinary institute. It also suggests a learning track for the student depending on his performance alongside multiple features. The models are based on the standard steps to build a classification model. As shown in figure 1, the model comprises 3 main phases: data collection, data preprocessing. My dataset is divided

into the training data section, and the test data section. Then, it performs data preparation related to train and addresses unbalanced classes with SMOTE method on it. The second part includes applying different classification algorithms to predict students' performance and to recommend the study path. Third, Data evaluation has a maximum performance accuracy.

11.1 Data Collection

An overview of the dataset's main characteristics is presented in this section.

Data was collected from records of students at the Higher Institute for Management Sciences. The institute has three unique disciplines, namely Information Systems, Management, as well as Accounting which are each part of specialized departments within the institute. Excel is the tool for data collection from October 2023. The data was collected after the semesters of 2015/2016 and 2018/2019. Record Number is 2869 records.

11.2 Preprocessing

In the pre-processing phase, the dataset was subjected to 4 stages, including data cleaning, feature selection, and data transformation. As the dataset obtained from the Student Management System Student data from Higher Institute is highly informative, we preprocessed them as follows:

Step 1: Removing features which contain redundancy (Student code, Student name, place of birth, lecturer name, course name, expenses when you apply for a course, etc.)

Step 2: Eliminate redundant or noise records including courses that the student is enrolled in but did not take the examination (i.e., null marks) or exempted courses, etc.

Step 3: We need to check for availability for the courses which do not have enough registration (considering some universities courses are removed if there are less than 25 students registered).

Step 4: Convert categorical values into numbers or diverse types.

11.3 Data Cleaning

Data cleaning is a crucial operation which prepares the data for the analysis process by modifying or removing corrupt, incomplete, duplicate, or incorrectly formatted data in the dataset. It enhances precise results post preparing data for analysis. Additional pre-processing such as removing redundant samples/duplicated features etc., removing outliers, managing missing values etc. Outlier removal processes include a column that contains the same value in the data. Data redundancy elimination is the process of eliminating repeated rows in your data. Missing process removal is to delete empty values. We use the learnt model to fill in the values of the missing dataset as shown in the Figure2.

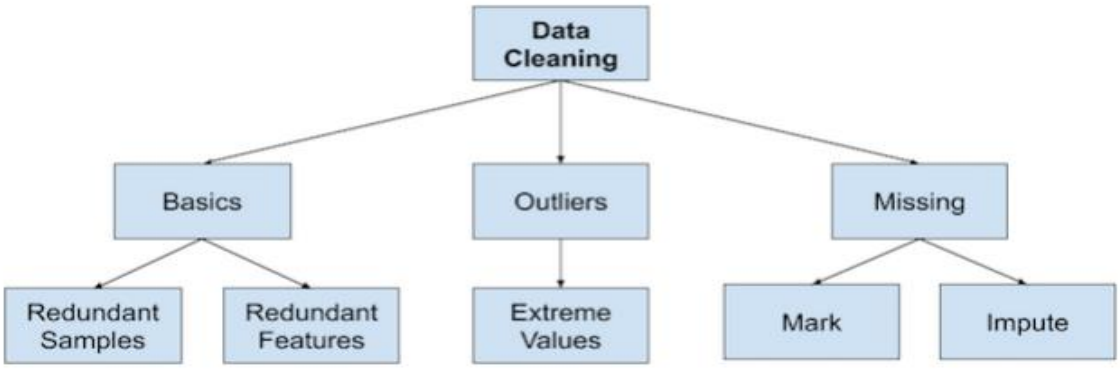

Figure 2: Data Cleaning Techniques. (Brownlee, 2020)

11.4 Feature Selection Using Data Mining

Wrapper method, filter method, and embedded method are just a few types of feature selections. Feature selection involves the process of identifying the subset of features from a larger feature set that can be effectively used to represent the same input data, by changing the dimension of feature space and eliminating unnecessary data. The approaches are depicted in Figure 3.

But there may be many features available in Students' Academic Performance dataset when the data set was collected the idea was to focus on student's features from the academic guidance perspective, which there exist interpretation in the register off subject courses which make chain connected with each other for the improvement of student's performance.

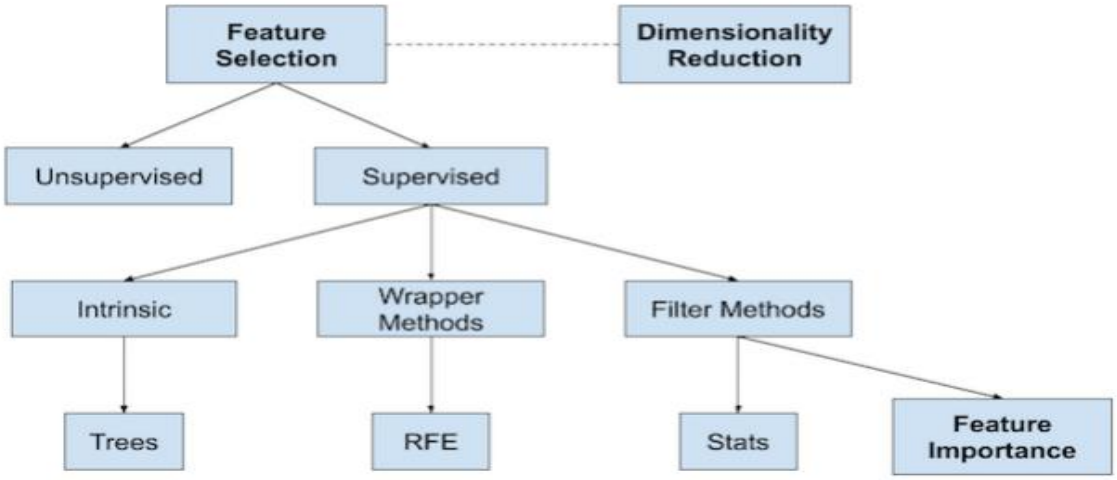

Figure 3: An Overview of Feather Selection Techniques (Brownlee, 2020)

11.5 Data Transformation

Data transformation refers to the process of converting data from one format or structure into another format or structure that is more suitable for analysis, storage, or presentation. It is a crucial step in data processing and is often necessary to ensure that data is clean, organized, and ready for further analysis or modeling.

11.6 Handling Imbalances Classes

Data balancing techniques are used to manage imbalanced datasets where one class significantly outnumbers another. This is crucial in machine learning tasks, where an imbalanced dataset can lead to biased predictions. The data collected is disproportionate as different classes are not well represented, the students enrolled in the Management Department are not evenly distributed among the two specializations. From figure 4, it is clear that the Marketing specialization forms the majority class and overshadows Finance, which is the minority class. The class imbalance problem was solved using under and oversampling for resampling the classes.

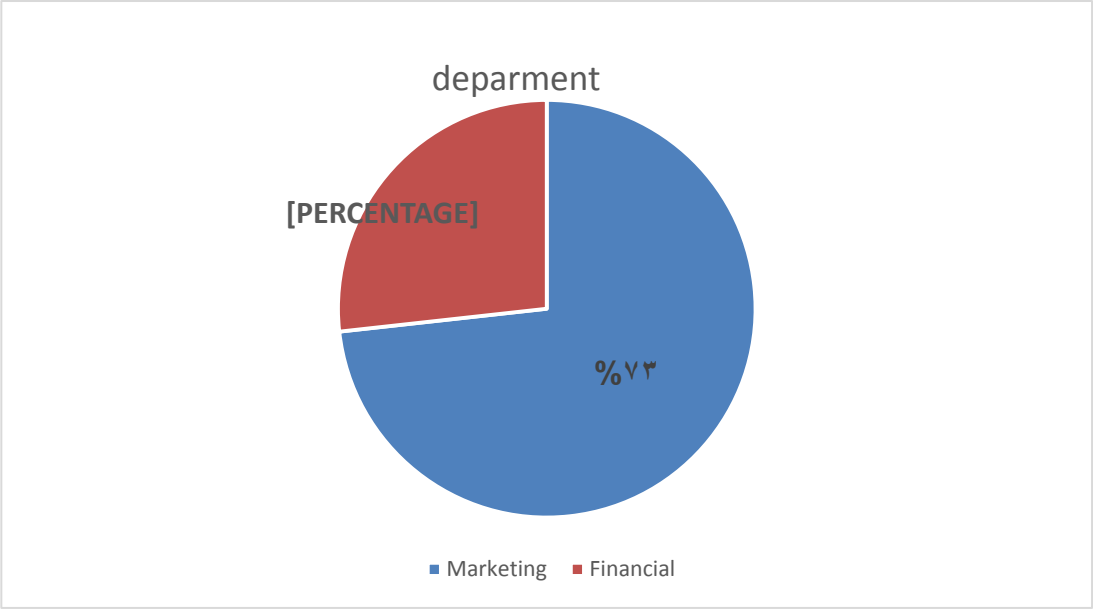

Figure 4: The Class Label's Distribution

Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is an oversampling technique that generates synthetic data points for the minority class instead of simply duplicating existing ones. It helps balance imbalanced datasets by making the minority class more representative (Chawla, 2009).

Steps in SMOTE

- Select a Minority Class Sample: A random sample from the minority class is chosen.
- Find k-Nearest Neighbors (k-NN): Identify the k nearest neighbors of the selected sample based on Euclidean distance.
- Generate Synthetic Samples: Create new synthetic samples by interpolating between the selected sample and one of its nearest neighbors.
- Repeat Until the Desired Balance is Achieved: The process is repeated for multiple samples until the dataset becomes balanced.

12.Experimental Results

12.1 J48 classifier

Table shows the performance accuracy of the J48 classifier. The J48 classifier relates to the class of trees while classifying the instances of the dataset before and after handling class imbalance. We used SMOTE technique to manage the class imbalance problem to enhance the classifier accuracy. Moreover, the number of instances that were classified correctly increased by **90.3361%** before using SMOTE and **91.6022%** after using SMOTE. So, the number of instances that were classified incorrectly decreased by **9.6639%** before using SMOTE and **8.3978%** after using SMOTE. The Relative absolute error decreased after using SMOTE with **31.124%** before and **19.5681%** after using SMOTE. as seen in figure 5.

Table 1: Accuracy of the J48 Classifier

Name Classifier	Classified Correctly	Classified Incorrectly	Relative absolute error
J48 before SMOTE	90.3361 %	9.6639 %	31.124 %
J48 After SMOTE	91.6022 %	8.3978 %	19.5681 %

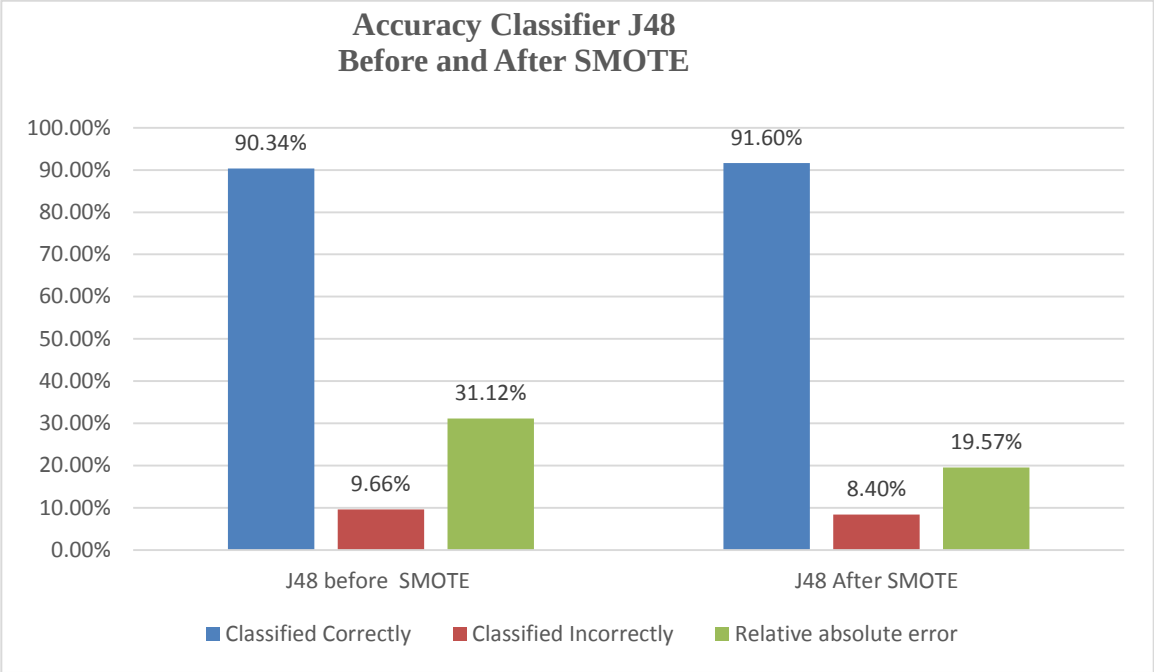

Figure 5: Accuracy of Classifier J48 Before and After SMOTE

Table 2 shows the performance metrics of comparative classification of the J48 classifier before and after handling class imbalance. The J48 classifier relates to the class of trees while classifying the instances of the data set into the marketing or finance department. The numbers of marketing instances exceed the number of finance instances which led to the existence of imbalance in the dataset. So, we used SMOTE technique to solve this imbalance to enhance the performance of the J48 classifier. The performance of the J48 classifier increased of marketing than finance after using SMOTE in terms of recall, precision, and F-measure with **0.939, 0.918**, and **0.928** to marketing and 0.885**, 0.914**, and **0.899** to finance respectively as seen in figure 6.

Table 2: Metrics of the J48 Classifier

Classifier	Recall	Precision	F-Measure	TP Rate	FP Rate	Class
J48 before SMOTE	0.956	0.916	0.935	0.956	0.241	Marketing
	0.759	0.863	0.901	0.759	0.044	Financial
J48 After SMOTE	0.939	0.918	0.928	0.939	0.115	Marketing
	0.885	0.914	0.899	0.885	0.061	Financial

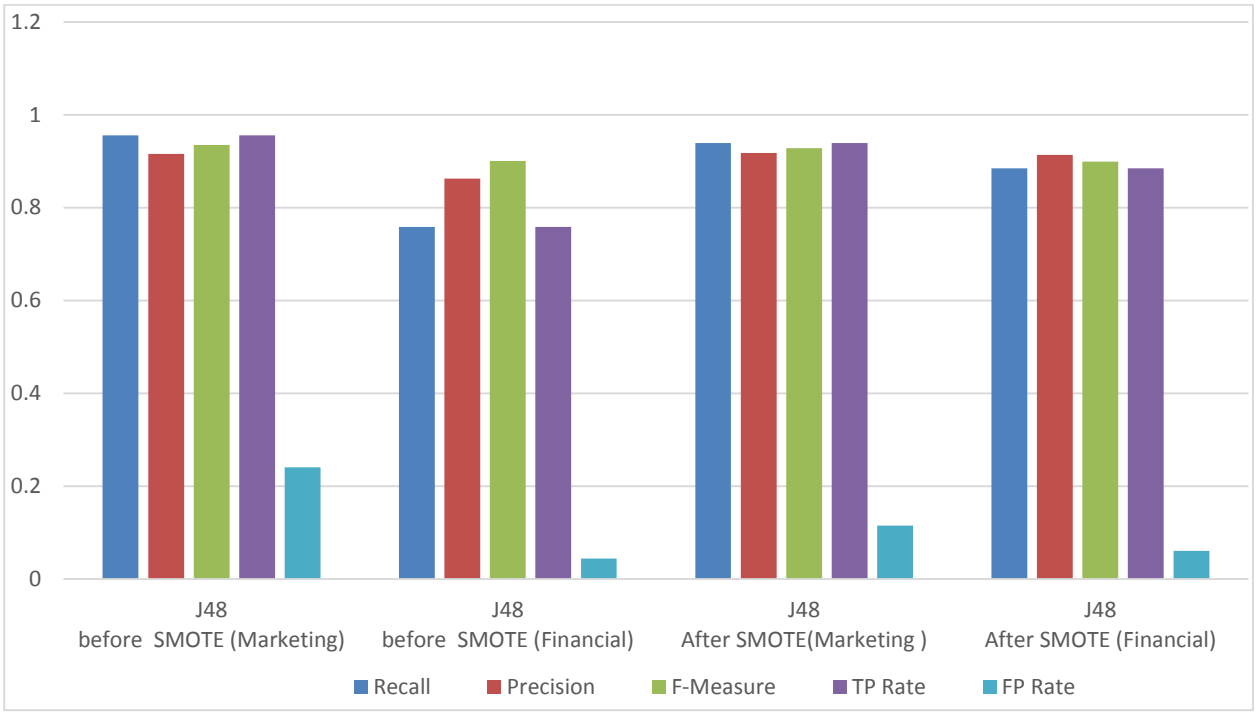

Figure 6: Comparison of The Metrics Performance on the Balanced Data with the J48 Classifier.

12.2 Random Forest Classifier

Table 3 shows the performance accuracy of the Random Forest classifier. The Random Forest classifier relates to the class of trees while classifying the instances of the dataset before and after handling class imbalance. We used SMOTE technique to manage the class imbalance problem to enhance the classifier accuracy. Moreover, the number of instances that were classified correctly increased by **92.1569 %** before using SMOTE and **93.7017 %** after using SMOTE. So, the number of instances that were classified incorrectly decreased by **7.8431 %** before using SMOTE and **6.2983 %** after using SMOTE. The Relative absolute error decreased after using SMOTE with **48.5513 %** before and **34.9409 %** after using SMOTE. as seen in figure 7.

Table 3: Accuracy of the Random Forest Classifier

Name Classifier	Classified Correctly	Classified Incorrectly	Relative absolute error
Random Forest before SOMTE	92.1569 %	7.8431 %	48.5513 %
Random Forest After	93.7017 %	6.2983 %	34.9409 %

SOMTE			

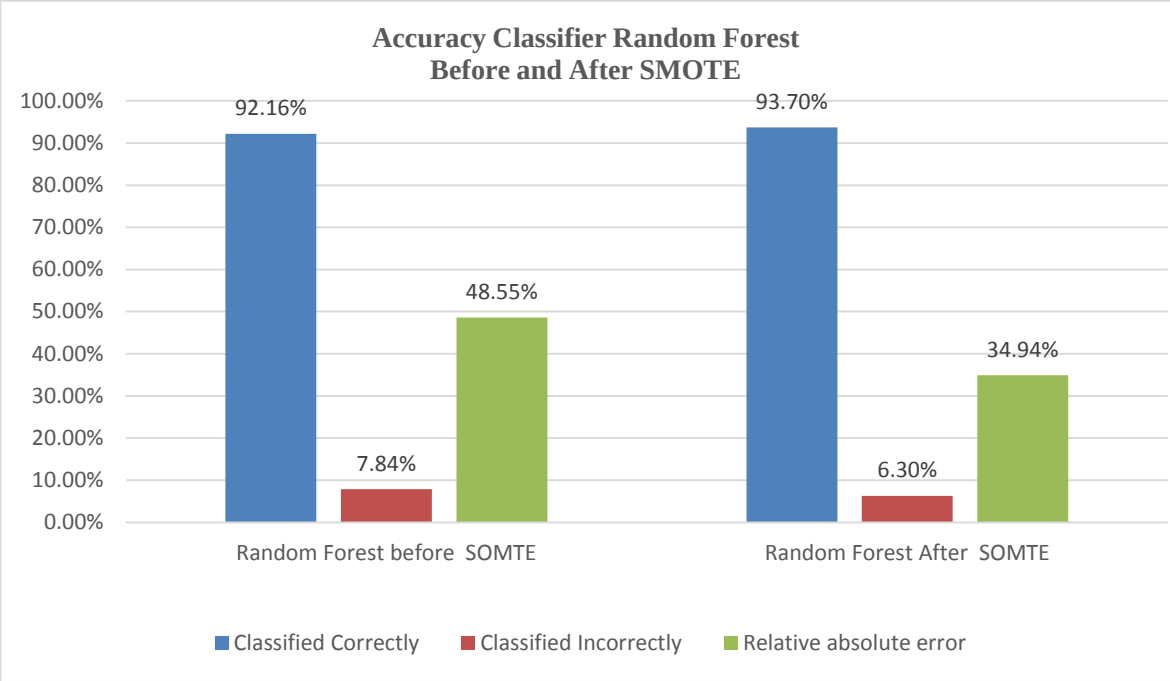

Figure 7: Accuracy of Classifier Random Forest Before and After SMOTE

Table 4 shows the performance metrics of comparative classification of the Random Forest classifier before and after handling class imbalance. The Random Forest classifier relates to the class of trees while classifying the instances of the dataset into the marketing or finance department. The numbers of marketing instances exceed the number of finance instances which led to the existence of an imbalance in the dataset. So, we used SMOTE technique to solve this imbalance to enhance the performance of the Random Forest classifier increased of marketing than finance after using SMOTE in terms of recall, precision, and F-measure with **0.971, 0.924**, and **0.947** to marketing and **0.890, 0.958**, and **0.923** to finance respectively as seen in figure 8 . It is observed from the results in Table 4 that Random Forest achieved the best metrics of recall, precision, and F-measure from other classifiers.

Table 4: Metrics of the Random Forest Classifier

Classifier	Recall	Precision	F-Measure	TP Rate	FP Rate	Class
Random Forest	0.979	0.919	0.948	0.979	0.236	Marketing

before SMOTE	0.764	0.930	0.839	0.764	0.021	Financial
Random Forest	0.971	0.924	0.947	0.971	0.110	Marketing
After SMOTE	0.890	0.958	0.923	0.890	0.029	Financial

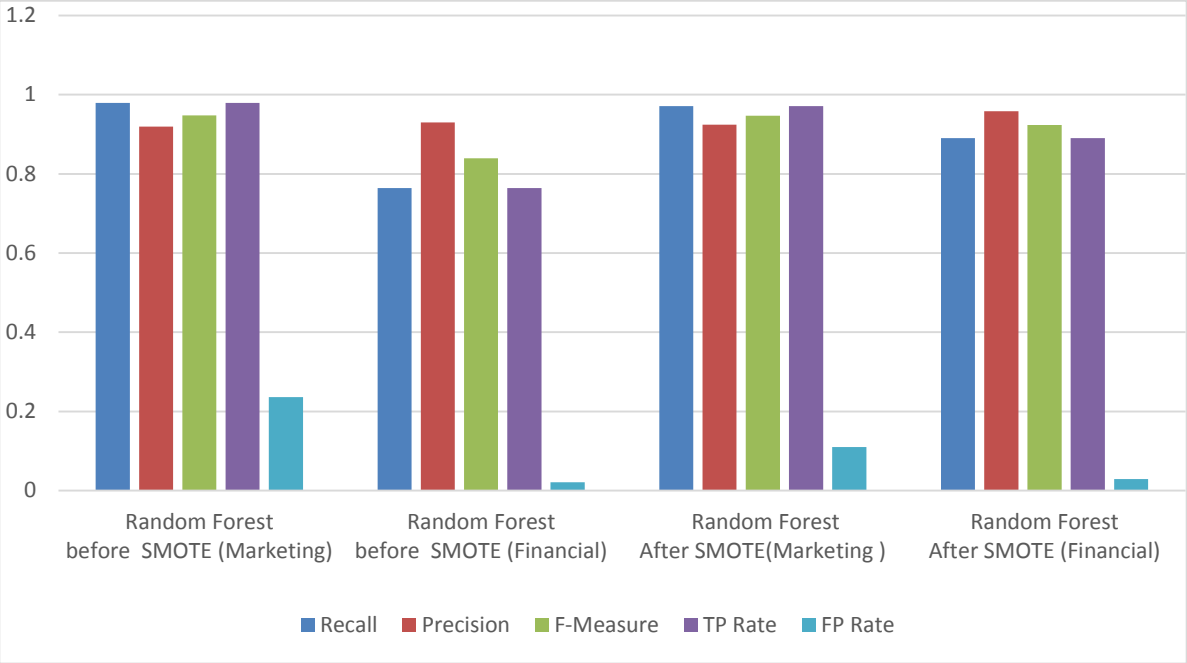

Figure 8: Comparison of The Metrics Performance on The Balanced Data with Random Forest Classifier

12.3 K-Nearest Neighbor (KNN) Classifier

Table 5 shows the performance accuracy of the K-Nearest Neighbor classifier. The K-Nearest Neighbor classifier is the best technique for classifying data and can achieve high accuracy. The KNN predicts the values of new data points based on "feature similarity" which implies that the new data point will be assigned a value depending on how closely it resembles the points in the training set. We used SMOTE technique to manage the class imbalance problem to enhance the classifier accuracy. Moreover, the number of instances that were classified correctly increased by **88.09%** before using SMOTE and **93.701%** after using SMOTE. So, the number of instances that were classified incorrectly decreased by **11.904%** before using SMOTE and **6.29 %** after using SMOTE. The Relative absolute error decreased after using SMOTE with **30.6526 %** before and **13.128%** after using SMOTE. as seen in figure 9.

Table 5: Accuracy of the K-Nearest Neighbor Classifier

Name Classifier	Classified Correctly	Classified Incorrectly	Relative absolute error
KNN before SMOTE	88.0952 %	11.9048 %	30.6526 %
KNN After SMOTE	93.7017 %	6.2983 %	13.1286 %

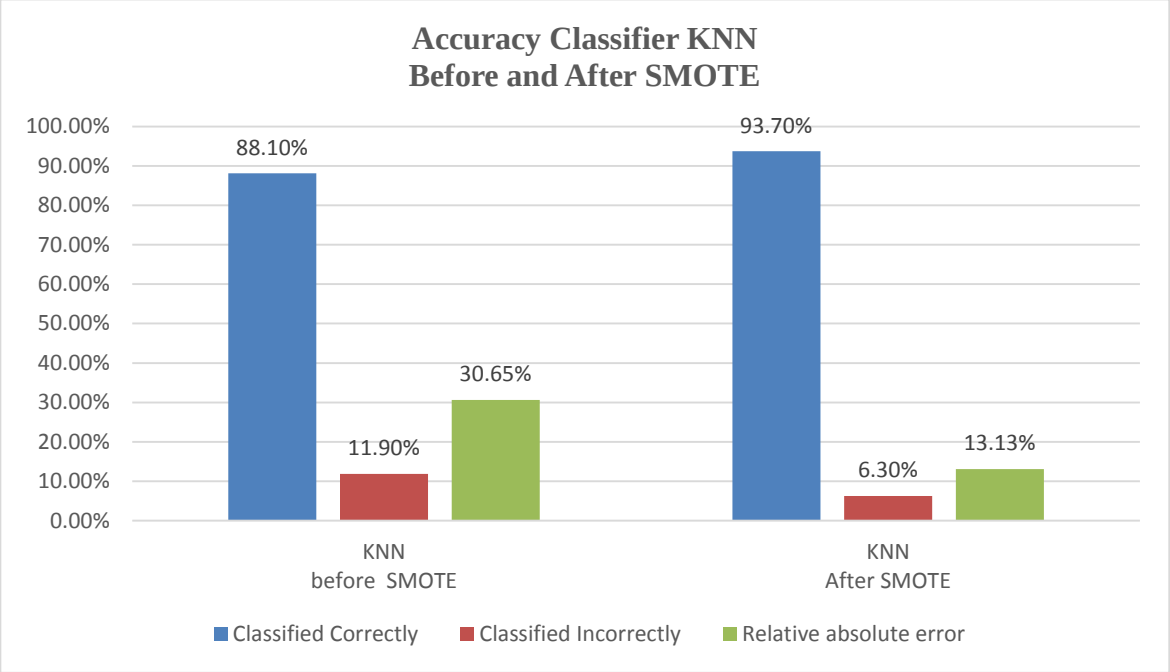

Figure 9: Accuracy of Classifier KNN Before and After SMOTE

Table 6 shows the performance metrics of comparative classification of the K-Nearest Neighbor classifier before and after handling class imbalance. The numbers of marketing instances exceed the number of finance instances which led to the existence of an imbalance in the dataset. So, we used SMOTE technique to solve this imbalance to enhance the performance of the K-Nearest Neighbor classifier increased of marketing than finance after using SMOTE in terms of recall, precision, and F-measure with **0.899**, **0.992**, and **0.943** to marketing and **0.990**, **0.877**, and **0.930** to finance respectively as seen Figure 10 .

Table 6: Metrics of the K-Nearest Neighbor Classifier

Classifier	Recall	Precision	F-Measure	TP Rate	FP Rate	Class

KNN before SMOTE	0.935	0.906	0.920	0.935	0.267	Marketing
	0.733	0.805	0.767	0.733	0.065	Financial
KNN After SMOTE	0.899	0.992	0.943	0.899	0.010	Marketing
	0.990	0.877	0.930	0.990	0.101	Financial

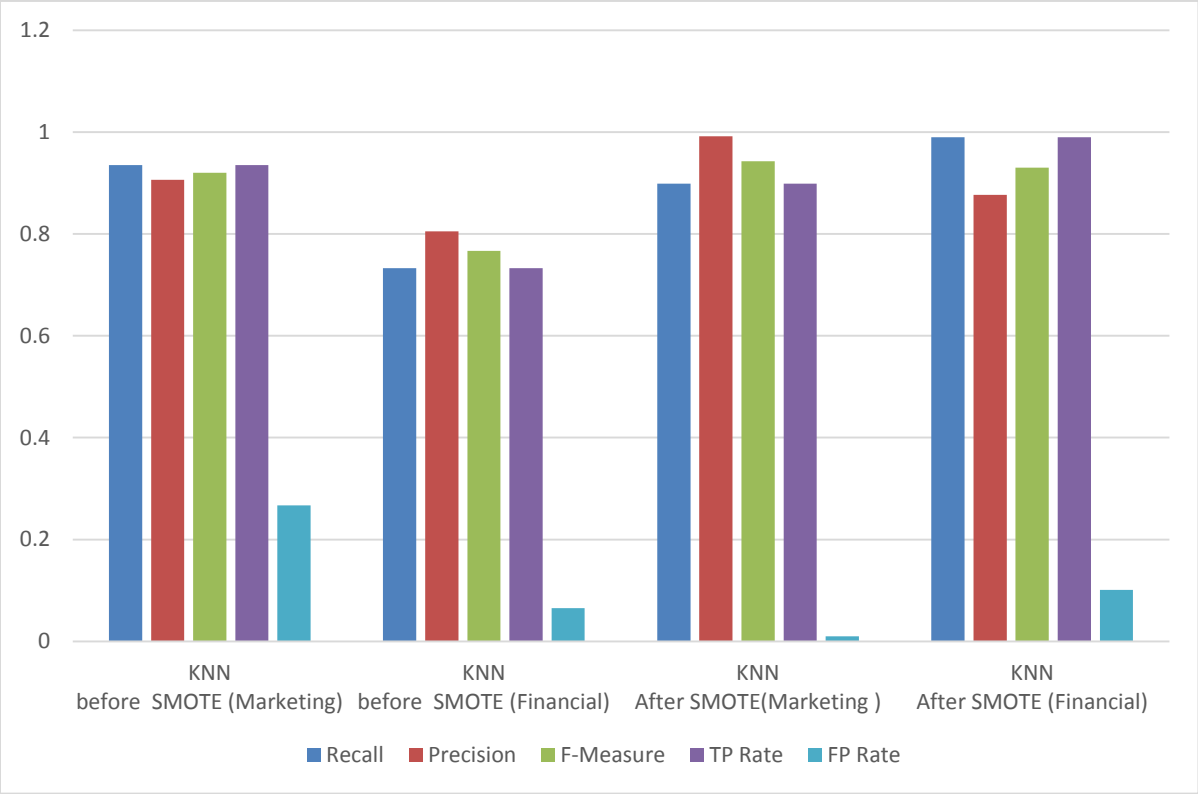

Figure 10: Comparison of The Metrics Performance on The Balanced Data with KNN Classifier

13.Conclusion

Educational data mining is a valuable analytical technique for assessing the large volumes of educational data collected in educational settings for decision-making purposes, the prediction of Students' Academic Performance at an early stage of a semester and identifying the hidden patterns and significant information from educational data. There are various issues with the prediction of Students' Academic Performance, such as an imbalanced dataset, which have more challenges that lead to deficient performance.

We applied classifiers of different machine learning algorithms, including J48, Random Forest, and KNN, to predict the best suited specialization department, for a student to join based on the current grades in their first- and second-year courses in other to have an idea of proper recommendation of specialization department to them based on the available dataset generated from an institute of management sciences. and predicting models by the various machine learning methods such as Random Forest (RF), K-Nearest Neighbour (KNN) the prediction performance of students in terms of GPA based on grades of previous courses first, second, third and fourth, which is the GPA in semester first to semester four academic years. We also used a resampling method, like the SMOTE, to address the imbalanced dataset problem, to improve the models' performance.

We check the predicting models and get not a bad result. The class imbalance problem had been solved, which resulted in improved performance of the models. The class imbalance problem was solved with SMOTE as an over sampling technique. Random Forest classifier gave the best result among other classifiers with an accuracy of 93.70 %, F-measure of 0.947, precision of 0.924 and recall of 0.971 corresponding to marketing class, and F-measure of 0.923, precision of 0.958 and recall 0.890 corresponding to financial class.

14. Future Work

- Expand the Dataset: The current dataset is limited to one institute. To improve the generalizability of the model, future research should include data from multiple institutions and a wider range of academic disciplines.

- Incorporate Additional Features: The study primarily focuses on academic performance metrics (e.g., GPA, course grades). Future research should consider incorporating additional features such as Behavioral Data, Psychological Factors, and Socioeconomic Factors.

- Integration with Learning Management Systems (LMS): The studies should explore integrating the predictive framework with existing Learning Management Systems (LMS) used by educational institutions. This would allow for seamless data collection and real-time feedback to both students and educators.

References

Alla Vladova, & Katsiaryna M. Borchyk. (2024). Predictive analytics of student performance: Multi-method and code. *JRAMathEdu (Journal of Research and Advances in Mathematics Education), 9*(4).

Ferri, C., Hernández-Orallo, J., & R, M. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 12.

Abdulkareem, N., & Abdulazeez, A. (2021). Machine learning classification based on Radom Forest Algorithm: A review. *International Journal of Science and Business, 5*(2), 128-142.

Abdullah, Waleed, Aniello, Michele, Chiara, & Muhammad. (2024). Student Academic Success Prediction Using Learning Management Multimedia Data With Convoluted Features and Ensemble Model. *ACM Journal of Data and Information Quality*, 16.

Agung Triayudi, Rima Tamara Aldisa, & S. Sumiati. (2024). New Framework of Educational Data Mining to Predict Student Learning Performance. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 15*(1), 115-132.

Aljawarneh, S., Yassein, M., & Aljundi, M. (2019). An enhanced J48 classification algorithm for the anomaly intrusion detection systems. *Cluster Computing, 22*(5), 10549-10565.

Amra, I., & Maghari, A. (2017). Students performance prediction using KNN and Naïve Bayesian. *2017 8th International Conference on Information Technology (ICIT)* (pp. 909-913). IEEE.

Asaad, R., & Abdulhakim, R. (2021). The Concept of Data Mining and Knowledge Extraction Techniques. *Qubahan Academic Journal, 1*(2), 17-20.

Awad, M., & Khanna, R. (2015). Support vector machines for classification. In *Efficient Learning Machines* (pp. 39-66). Springer.

Balaram, A., & Vasundra, S. (2022). Prediction of software fault-prone classes using ensemble random forest with adaptive synthetic sampling algorithm. *Automated Software Engineering, 29*(1), 1-21.

Brownlee, J. (2020). *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python.* Machine Learning Mastery.

Chawla, N. (2009). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, *18*(6), 875-886.

Dengen, N., Budiman, E., Wati, M., & Hairah, U. (2018). Student Academic Evaluation using Naïve Bayes Classifier Algorithm. *2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT)* (pp. 104-107). IEEE.

Ekemini A. Johnson, Jude A. Inyangetoh, Habeeb A. Rahmon, Tope G. Jimoh, Eduediuyai E. Dan, & Mfon O. Esang. (2024). An Intelligent Analytic Framework for Predicting Students Academic Performance Using Multiple Linear Regression and Random Forest. *European Journal of Computer Science and Information Technology*, *12*(3), 56-70.

Ghosh, S., & Janan, F. (2021). Prediction of Student's Performance Using Random Forest Classifier. *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management, Singapore*, (pp. 7-11).

Gray, D., Bowes, D., Davey, N., Sun, Y., & Christianson, B. (2011). Further Thoughts on Precision. *15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011)*.

Hersh, W. (2021). Information retrieval. In *Biomedical Informatics* (Vol. 18, pp. 755-794). Springer.

Hussain, S., Gaftandzhieva, S., Maniruzzaman, M., Doneva, R., & Muhsin, Z. (2021). Regression analysis of student academic performance using deep learning. *Education and Information Technologies*, *26*(1), 783-798.

Ibrahim H. Ibrahim, Etemi J. Garba, Uwaisu A. Umar, & Adedeji A. Adejumo. (2024). Predictive Model for Identification and Analysis of Factors Impacting Students Academic Performance Using Machine Learning Algorithms. *Kasu Journal of Computer Science*, *1*(3), 567-592.

Isreal Ogundele, Olutosin Taiwo, Asegunloluwa Babalola, & Olumide Ayeni. (2024). Prediction of Student Academic Performance Based on Machine Learning Model. *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, (pp. 1-11).

Jamal Eddine, Zakrani, Mohammed, Said, & Abdellah. (2025). Predicting academic performance: toward a model based on machine learning and learner's intelligences.. *international Journal of Electrical & Computer Engineering (2088-8708)*, *15*(1), 645:653.

Kenth C. Novo. (2024). Predicting Students' Academic Performance Using Data Mining Method. *International Journal of Latest Technology in Engineering Management & Applied Science, 13*(10), 127-131.

logistic regression. (2024, 4-). (techtarget) Retrieved 8- 2024, from https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression

Muntasir Nishat, M., Faisal, F., Jahan Ratul, I., Al-Monsur, A., & Ar-Rafi, A. (2022). A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset. *Scientific Programming, 2022*(5), 1-13.

Patil, P., & Hiremath, R. (2022). Big Data Mining—Analysis and Prediction of Data, Based on Student Performance. In *Pervasive Computing and Social Networking* (Vol. 317, pp. 201-215). India: Springer.

Patra, M. R., & Patro, V. M. (2014). Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy. *Transactions on Machine Learning and Artificial Inteligence, 2*(4), 15.

Ramasubramanian, K., & Singh, A. (2017). *Machine learning using R.* Springer.

Razaque, F., Soomro, N., Shaikh, S., Soomro, S., Samo, J., Kumar, N., & Dharejo, H. (2017). Using naïve bayes algorithm to students' bachelor academic performances analysis. *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)* (pp. 1-5). IEEE.

Silveira, A., Sobrinho, Á., Silva, L., Costa, E., Pinheiro, M., & Perkusich, A. (2022). Exploring Early Prediction of Chronic Kidney Disease Using Machine Learning Algorithms for Small and Imbalanced Datasets. *Applied Sciences, 12*(7), 1-14.

Sixuan, & Bin. (2024). Academic achievement prediction in higher education through interpretable modeling. *Plos one, 19*(9), 1-18.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*, 11.

Stapel, M., Zheng, Z., & Pinkwart, N. (2016). An ensemble method to predict student performance in an online math learning environment. *International Educational Data Mining Society, 17*(6), 231-238.

Vivi Nur Wijayaningrum, Ika Kusumaning Putri, & Annisa Puspa Kirana. (2024). Student academic performance prediction framework with feature selection and imbalanced data handling. *Jurnal Ilmiah Kursor, 12*(3), 123-134.

Widyahastuti, F., & Tjhin, V. (2017). predicting students' performance in final examination using linear regression and multilayer perceptron. *2017 10th International Conference on Human System Interactions (HSI)* (pp. 188-192). Institute of Electrical and Electronics Engineers Inc. doi:1509046887

Zeng, G. (2019). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics - Theory and Methods*, 15.

Zheng, Jiahao, Deyao, Jingyu, Yunhong, & Zhanbo. (2024). A Method for Prediction and Analysis of Student Performance That Combines Multi-Dimensional Features of Time and Space. *Mathematics, 12*(22), 1-26.

Zul, M. (2016). Prediction of Student Final Grade by using k-Nearest Neighbor Algorithm. *The First International Conference on Technology, Innovation, and Society (ICTIS) 2016.*