Multiclass Prediction Model for Student Grade Prediction Using Machine Learning

[1]Shaik Mulla Almas, [2]G. Jayanth Satya, [3]D. Pavan Kumar, [4]Mohammed Asrar, [5]M. Yashwanth, [6]N. Sai Subhash

[1]Assistant Professor, [2,3,4,5,6]B.Tech. Students
Information technology
Vasireddy Venkatadri Institute of Technology
Guntur, India.

***Abstract*- Today, there is a growing demand for predictive analytics applications in higher education institutions. These applications utilize advanced analytics, including machine learning, to extract valuable insights and improve performance across all levels of education. Student grades are a crucial performance indicator that educators use to track academic progress. Over the past decade, various machine learning techniques have been proposed for educational purposes. However, challenges persist in dealing with imbalanced datasets to enhance the accuracy of predicting student grades. This study offers a detailed analysis of machine learning techniques to predict final student grades in first-semester courses, aiming to boost predictive accuracy. The paper focuses on two main modules. Firstly, it compares the performance of six popular machine learning techniques - Decision Tree (J48), Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (kNN), Logistic Regression (LR), and Random Forest (RF) - using a dataset of 1282 real student course grades. Secondly, a multiclass prediction model is proposed to address overfitting and misclassification issues in imbalanced multi-class scenarios, employing oversampling techniques like Synthetic Minority Oversampling Technique (SMOTE) along with feature selection methods. The results demonstrate that integrating the proposed model with RF leads to a significant improvement, achieving the highest f-measure of 99.5%. This model shows promising results in enhancing prediction performance for imbalanced multi-class student grade prediction.**

***Index Terms*-SVM, NB, KNN, LR, RF.**

I. INTRODUCTION (HEADING 1)

In higher education institutions (HEI), each institution has its own student academic management system to store all student data, including information about their final examination marks and grades in different courses and programs. These marks and grades are used to generate a student academic performance report, which evaluates their course achievements every semester. The data stored in the repository can provide valuable insights into student academic performance. Solomon et al. have emphasized that determining student academic performance is a significant challenge in HEI.

Consequently, previous researchers have identified various influential factors that can greatly impact student academic performance. However, the most common factors are typically related to socioeconomic background, demographics, and learning activities, as opposed to final examination grades. Therefore, it is evident that predicting student grades could be a viable solution to enhance student academic performance. Predictive analytics has proven to be beneficial in HEI, as it enables the identification of hidden patterns and the prediction of trends in a vast database, thereby benefiting the competitive educational domain. It has been successfully applied in various educational areas, such as student performance, dropout prediction, academic early warning systems, and course selection. Furthermore, the use of predictive analytics in predicting student academic performance has steadily increased over the years. The ability to predict student grades is an important area that can contribute to improving student academic performance. Previous research has explored different machine learning techniques for predicting student academic performance. However, there is a lack of research on addressing the challenges posed by imbalanced multi-classification problems in predicting students' grade prediction.

II. FRAMEWORK OF GRADE PREDECTION

This document aims to identify the most effective predictive model for student grade prediction, specifically in addressing imbalanced multi-classification. Our framework takes as input the final course grade of the student, which we extract from their academic spreadsheet document and academic repository. To tackle the issue of imbalanced multi-classification, we employ two data-level solutions: oversampling using SMOTE and two feature selection (FS) methods. These techniques help reduce overfitting and misclassification in the dataset. Next, we combine these techniques to

design our proposed model, which is then evaluated using performance metrics through a selected machine learning classifier. Finally, data visualization is utilized to depict the dataset trends and the final classification results.

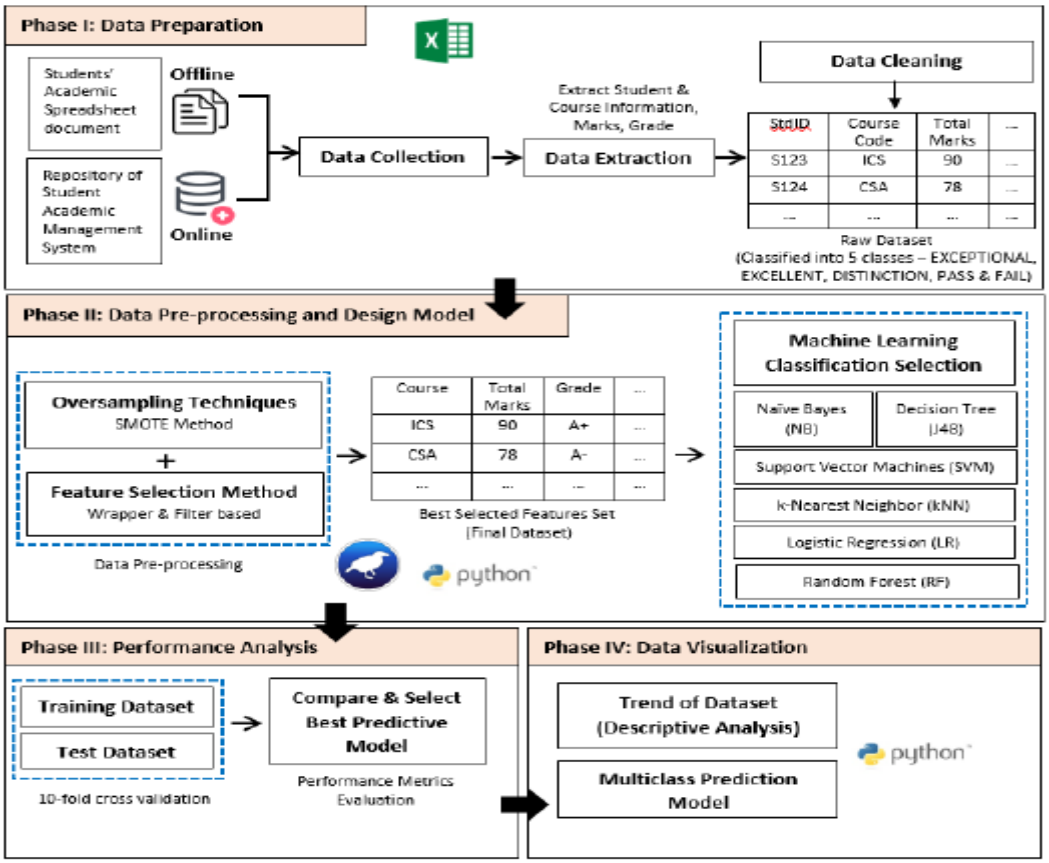

FIG: FRAME WORK OF GRADE PREDECTION

A. DATA PREPARATION

The dataset utilized in this study was obtained from the Department of Information and Communication Technology (JTMK) at one of the Malaysia Polytechnics. It comprises 1282 instances, representing the overall course grades of first-semester students who took the final examination between June 2016 and December 2019. To progress to the next academic semester, students are required to complete certain compulsory, specialization, and core course modules. However, for the purpose of this research, we focused solely on two core courses that included the percentage of final examination and course assessment marks.

B. DATA PRE-PROCESSING AND DESIGN MODEL

During this stage, data pre-processing was implemented on the dataset that was collected. To facilitate the pre-processing of data, we categorized the students into 5 different grade categories: Exceptional (AC), Excellent (A), Distinction (A-, B+, B), Pass (B-, C+, C, C-, D+, D), and Fail (E, E-, F). These categories were established as the output for the prediction class. However, upon analyzing the class distribution of the dataset, it was evident that there was an imbalance in the number of instances for each class. Specifically, there were 63 instances of Exceptional, 377 instances of Excellent, 635 instances of Distinction, 186 instances of Pass, and 21 instances of Fail, with a high ratio of 3:18:30:9:1. This imbalance could potentially lead to overfitting results. To address this issue, data-level solutions such as oversampling SMOTE and two feature selection methods, Wrapper and Filter based, were employed as benchmark methods in this study. The experiment utilized the open-source tool Waikato Environment for Knowledge Analysis (WEKA) version 3.8.3 due to its wide range of machine learning algorithms and user-friendly graphical interfaces for easy visualization.

III. PERFORMANCE ANALYSIS

The objective of this study is to forecast the final grades of students by analyzing their academic performance in the previous semester's final exams. The research utilized various machine learning algorithms to determine which algorithm yielded the most accurate predictions for the students' final grades. The study consisted of three experiments conducted in four distinct phases across five different classes. The accuracy of the predictions was assessed through ten-fold cross-validation, where 90% of the dataset was allocated for training and 10% for testing on the same dataset.

The theoretical models employed in constructing the multiclass prediction model included Logistic Regression (LR) and Naïve Bayes (NB). Logistic Regression utilizes a cost function that employs a logistic function to mathematically model

classification problems, making it ideal for analyzing categorical data and understanding the relationships between variables. On the other hand, Naïve Bayes is based on Bayesian theorem and is favored for its simplicity and ability to provide quick predictions. It is particularly suitable for small datasets, combining complexity with a flexible probabilistic model for accurate predictions.

The Decision Tree (J48) is a commonly used algorithm in various multi-class classification tasks, capable of handling missing values in high-dimensional data. It has been implemented successfully to achieve optimal accuracy results while using the minimum number of features.

The K-Nearest Neighbor (kNN) algorithm, on the other hand, is a non-parametric method that classifies instances in a dataset by calculating the differences between them and their nearest vectors. The value of k represents the distance in the n-dimensional space. kNN employs a distance function to perform well in datasets with small features.

Lastly, the Random Forest (RF) is a classifier that utilizes ensemble learning by combining multiple decision trees from various subsets of the data. This approach helps identify the best features for achieving high accuracy while mitigating the issue of overfitting. Additionally, RF demonstrates relative robustness to outliers and noise, making it an effective classification method.

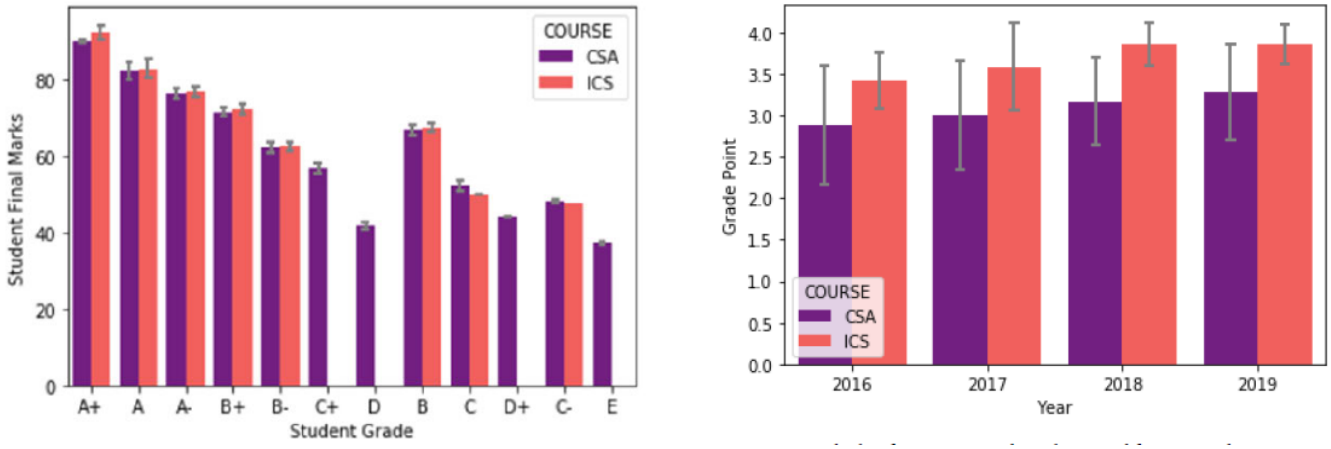

FIG: THE ABOVE FIGURES PERFORAMCE ANALYSIS

IV. EXPERIMENTAL RESULTS

The study's findings are categorized into two subsections based on the research questions. A thorough performance analysis was carried out through three experiments utilizing real data. The results of J48, kNN, NB, SVM, LR, and RF experiments were examined and contrasted. Additionally, the effects of employing oversampling SMOTE and FS techniques to address the imbalanced multi-classification issue with the identical dataset were also assessed.

The primary aim of this study is to evaluate the predictive model's accuracy performance by comparing six different machine learning algorithms. These algorithms were utilized to train a student dataset, and their prediction accuracy was assessed. The performance accuracy was analyzed using ten-fold cross-validation with stratification as the testing method to identify the most optimal predictive model. Various metrics such as classification accuracy, precision, recall (Sensitivity), and f-measure were employed to ensure the accuracy of the predictive model. The results of different classifiers on the student dataset are summarized.

Metric	J48	kNN	NB	SVM	LR	RF
Accuracy	**0.989**	0.985	0.978	0.984	0.984	**0.989**
Precision	0.989	0.985	0.978	0.981	0.983	0.989
Recall	0.990	0.985	0.977	0.984	0.984	0.989
F-Measure	0.989	0.985	0.978	0.979	0.983	0.989

FIG: Table Performance comparison of predictive models.

IMPACT OF OVERSAMPLING AND FEATURE SELECTION FOR IMBALANCED MULTI-CLASS DATASET

In this study, we concentrate on the impact of oversampling and feature selection techniques on imbalanced multi-class datasets. Specifically, we utilize oversampling SMOTE and two feature selection algorithms to address the imbalanced classification issue. To evaluate the performance of various predictive models, we conduct three experiments using six machine learning algorithms. Initially, we apply SMOTE to the dataset with each of the six machine learning algorithms independently. Subsequently, we employ two feature selection algorithms separately with three different attribute evaluators. Lastly, we implement and test the proposed multiclass prediction model (SFS) using the same dataset with the six selected machine learning algorithms. In addition to accuracy, we also consider other performance metrics such as precision, recall, and f-measure to ensure the effectiveness of our predictive model in predicting the dimensionality accurately .

SMOTE OVERSAMPLING TECHNIQUE

The Synthetic Minority Oversampling Technique (SMOTE) is widely utilized to address the issue of overfitting in machine learning. It employs a random sampling algorithm to modify imbalanced datasets and generate new instances of the minority class using synthetic sampling techniques. This helps to create a more balanced distribution. In this study, the default parameter for the number of nearest neighbors (k) in the SG sample of the minority class was increased. N samples were then randomly selected and recorded as SGi. The new sample, SG_{new}, is defined by the following expression.

$$SG_{new} = SG_{origin} + rand \times \left(SG_i - SG_{origin}\right), \quad i = 1, 2, 3, \ldots n$$

The random seed, 'random', is used for random sampling within the range of (0,1). We implemented the SMOTE filter in Weka, specifically the weak filters. supervised. Instance SMOTE' filter, to insert synthetic instances between the minority class samples and our dataset. The parameter for the index class value 0 was set to auto-detect the non-empty minority class. We also set the 'k' value for the number of nearest neighbors to 10 (k = 10), with a percentage of instances set to 100%. The SMOTE filter was applied in ten iterations. The oversampled dataset increased the number of instances from 1282 to 2932. The class distribution using SMOTE became (504) exceptional, (377) excellent, (635) distinction, (744) pass, and (672) fail, reducing the ratio to 1:1:2:2:2. In Table 6, we present the detailed comparison results of all predictive models with their performance measures. When the classifiers were used with oversampling SMOTE, we consistently observed an improvement in the effectiveness of all predictive models. Among these models, RF achieved the most promising f-measure of 99.5%, followed by kNN with 99.3%, J48 with 99.1%, SVM with 98.9%, LR with 98.8%, and NB with 98.3%. This result was statistically significant with a confidence level of 95% using the Paired T-Tester (corrected), as shown in Figure 6. We also observed that when the SMOTE method was applied, the number of minority class instances increased to balance with the other classes through the iteration and 'k' value. The accuracy performance was analyzed in detail based on the confusion matrix.

DISCUSSION

This research aimed to tackle the issue of imbalanced multi-classification problems in student grade prediction by focusing on data-level solutions. To address this problem, we utilized a real dataset of final course grades from JTMK at one of Malaysia Polytechnics and analyzed the results of our proposed model. A similar study conducted in the past also highlighted the importance of course grades in decision making within the educational domain. In order to answer our research question, we conducted a comprehensive experiment on the real student dataset, comparing the accuracy performance of our prediction model with a selected machine learning algorithm. Additionally, we applied oversampling SMOTE and two FS methods to assess the effectiveness of the predictive model, using evaluation metrics such as accuracy, precision, recall, and f-measure to demonstrate the performance of the predictive models.

The results of the study revealed that predictive models generated from J48, NB, kNN, SVM, LR, and RF exhibited enhanced performance when SMOTE was applied individually to address the imbalanced dataset. However, when the FS method was applied to the imbalanced dataset using a wrapper-based approach, only kNN and NB demonstrated significant improvement, while SVM remained unchanged. Additionally, it was observed that SVM struggled to independently address imbalanced multi-classification due to limitations in computing the optimal hyperplane for high-dimensional imbalanced datasets. On the other hand, NB's utilization of FS for predicting student grades was supported by previous research, which highlighted NB's superior accuracy performance when employing wrapper-based subset feature selection. Nevertheless, it was noted that FS alone did not enhance the accuracy performance of RF, possibly due to the imbalanced nature of the dataset. Therefore, while FS facilitated quicker interpretation of the predictive model, its impact on performance was not solely dependent on a few features.

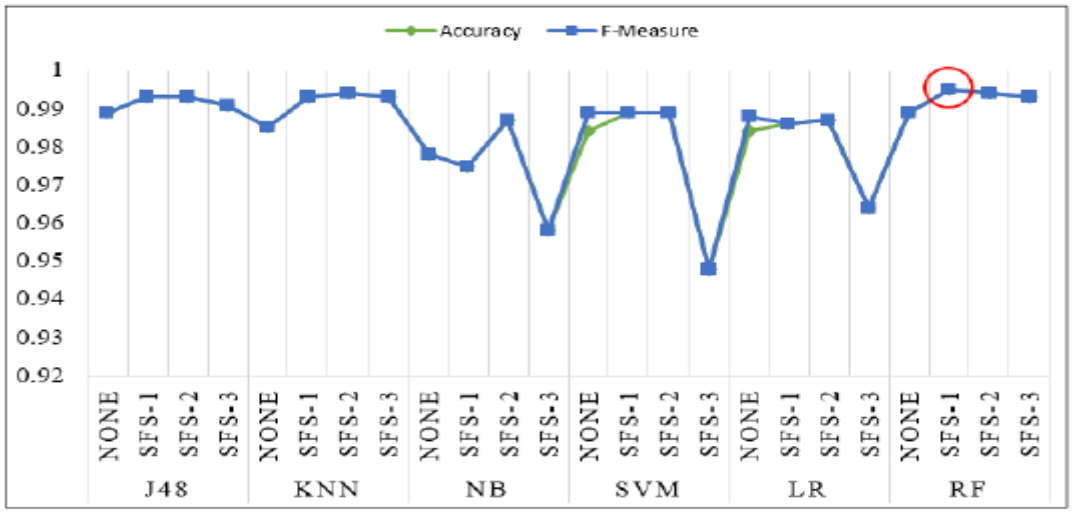

FIG: Accuracy and f-measure of proposed SFS model.

V. CONCLUSION AND FUTURE DIRECTIONS

Forecasting student grades serves as a crucial metric for educators to monitor academic progress effectively. To address the uncertainty in outcomes for imbalanced datasets, the implementation of a predictive model is essential. This study introduces a multiclass prediction model comprising six predictive models to anticipate final student grades based on the previous semester's final examination results. Through a comparative analysis, the combination of oversampling SMOTE with various FS methods was evaluated to enhance the accuracy of student grade prediction.

The findings reveal that the utilization of oversampling SMOTE consistently outperforms using FS alone across all predictive models. Moreover, the proposed multiclass prediction model demonstrates superior performance compared to employing oversampling SMOTE and FS independently, under specific parameter settings that influence the accuracy of predictive models. This research contributes a practical approach to addressing imbalanced multi-classification challenges through data-level solutions for student grade prediction in higher education institutions. Furthermore, future studies are recommended to explore emerging predictive techniques in advanced machine learning algorithms and ensemble algorithms to optimize student grade prediction outcomes. Additionally, the selection of diverse multi-class imbalanced datasets, along with appropriate sampling techniques and evaluation metrics suitable for the imbalanced multi-class domain, such as Kappa and Weighted Accuracy, is crucial for further analysis and improvement in predictive modeling.

VI. ACKNOWLEDGMENT

We are grateful to pervious authors a organization staff for helping us.

REFERENCES:

1. D.Solomon, S.pital, and P.Agrawal," Predicting performance and potential difficulties of university student using classification: Survey paper *Int. J. Pure Appl. Math*, vol. 118, no. 18, pp. 2703_2707, 2018.
2. E. Alyahyan and D. Dü³tegör, ``Predicting academic success in higher education: Literature review and best practices," *Int. J. Educ. Technol.Higher Educ.*, vol. 17, no. 1, Dec. 2020.
3. S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, ``Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decis.Anal.*, vol. 2, no. 1, pp. 1_25, Dec. 2015.
4. H. Sun, M. R. Rabbani, M. S. Sial, S. Yu, J. A. Filipe, and J. Cherian, ``Identifying big Data's opportunities, challenges, and implications in _finance, '*Mathematics*, vol. 8, no. 10, p. 1738, Oct. 2020.
5. I. Khan, A. Al Sadiri, A. R. Ahmad, and N. Jabeur, ``Tracking student performance in introductory programming by Means of machine learning," in *Proc. 4th MEC Int. Conf. Big Data Smart City (ICBDSC)*, Jan. 2019,pp. 1_6.
6. M. A. Al-Barrak and M. Al-Razgan, ``Predicting students _nal GPA using decision trees: A case study," *Int. J. Inf. Educ. Technol.*, vol. 6, no. 7,pp. 528_533, 2016.
7. E. C. Abana, ``A decision tree approach for predicting student grades in research project using WEKA," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 285_289, 2019.