INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

IJARIIT

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 9, Issue 2 - V9I2-1161)
Available online at: https://www.ijariit.com

Employing Machine Learning, A Multiclass Prediction Model For The Student Grading System

Jahnavi Sannidhi
jahnavisannithi@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

Dumpala Pavan Kumar Reddy
luckypavankumar666@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

Akkaladevi Lumbhini Madhuri
lumbhini9856@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

Donka Suresh
dsuresh6484@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

Nimmagallu Swetha
518543swetha@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

D. Sarika
sarikadaruru7790@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

ABSTRACT

In today's higher education institutions, predictive analytics applications have become a pressing need. In order to generate high-quality performance and valuable data for all educational levels, predictive analytics used sophisticated analytics that included the application of machine learning. the majority of people are aware that One of the main metrics that may be used by educators to track students' academic progress is their grades. In the last ten years, a wide range of machine learning algorithms has been proposed by researchers in the field of education. To improve the performance of predicting student grades, addressing imbalanced datasets presents serious difficulties. Therefore, this study gives a thorough review of machine learning algorithms to predict the final student grades in the first semester courses by enhancing the performance of prediction accuracy. In this study, we'll emphasize two modules. Using a dataset of 1282 genuine student course grades, we assess the accuracy performance of six well-known machine learning techniques: Decision Tree (J48), Support Vector Machine (SVM), Nave Bayes (NB), K-Nearest Neighbor (kNN), Logistic Regression (LR), and Random Forest (RF). In order to reduce overfitting and misclassification results brought on by imbalanced multi-classification based on oversampling Synthetic Minority Oversampling Technique (SMOTE) using two feature selection methods, we have suggested a multiclass prediction model. The outcomes demonstrate that the suggested model integrates with RF and gives a notable improvement with the greatest f-measure of 99.5%. This model's suggested findings are comparable and encouraging, and they have the potential to improve the model's performance predictions for imbalanced multi-classification for student grade prediction.

Keywords: *Predictive Model, Unbalanced Issue, Forecasting Student Grades, and Multi-Class Classification*

1. INTRODUCTION

Every institution in higher education institutions (HEI) has a system for managing the academic performance of its students. This system is used to keep track of all the information about students, including their grades on final exams and their performance in various courses and programmes. Every semester, a student academic performance report is produced using the total number of recorded student grades and marks to assess the course's success. Insightful data about student academic performance can be found using the repository's data. According to Solomon et al., a significant difficulty in HEI is assessing students' academic performance.The influencing factors that might significantly affect a student's academic performance have been well-defined as a result by numerous earlier researches. In contrast to final student scores on the final exam, the most frequent determinants depend on socioeconomic background, demographics, and learning activities. For this reason, we see that one strategy for enhancing students' academic performance is the tendency of anticipating their grades.

The HEI has successfully benefited from predictive analytics. Finding hidden patterns and making predictions about trends in a sizable database may be a promising strategy to help the competitive educational field. It has been applied to address issues in a number of educational fields, including course selection, academic early warning systems, dropout prediction, and student performance. Furthermore, over time, the use of predictive analytics to forecast student academic success has grown. One crucial aspect that might aid to enhance a student's academic success is their capacity to predict grades. Numerous studies conducted in the past have discovered that different machine learning techniques are effective at forecasting student academic achievement. However, it is challenging to locate the associated studies on mechanisms to enhance the imbalanced multi-classification problem in predicting

© 2023, www.IJARIIT.com All Rights Reserved

students' grades. In order to determine the best prediction model for predicting student grades, the following questions have been addressed in this study's comparison analysis:

RQ1: Which machine learning algorithm of the ones chosen has the highest accuracy in predicting students' final course grades?

RQ2: How can an unbalanced multi-classification dataset be handled utilizing particular machine learning techniques, such as the Synthetic Minority Oversampling Technique (SMOTE) and feature selection (FS) methods, oversampling, and feature selection?

We compile the student final course marks from two core courses in the first semester of the final examination result in order to answer the questions indicated above. In order to visualize student grade trends and aid instructors in making more informed decisions, we propose a descriptive analysis of student statistics. Then, using actual student data from a Malaysia Polytechnic's Diploma in Information Technology (Digital Technology), we compare results using six well-known machine learning algorithms, including LR, NB, J48, SVM, kNN, and RF. Regarding the imbalanced multi-classification, we work to improve each predictive model's performance with data-level methods utilizing oversampling SMOTE and FS. To improve imbalanced multiclassification for predicting student grades, we combined modifications to two feature selection methods with oversampling SMOTE to automatically calculate the sampling ratio with the best features. Our proposed model shows different impact in improving the performance of student grade prediction model based on the versatility of two feature selection algorithm after implementing SMOTE.

2. RELATED WORKS

For forecasting student grades using various machine learning techniques, several studies have been undertaken at HEI. In order to forecast student grades for various outcomes, it includes the analysis process of many features and samples data from several sources. The effectiveness of predictive models for datasets with imbalances in the field of education, however, is still seldom ever explored. In order to increase the precision of students predicted final grades, a study from [12] used discretization and oversampling SMOTE approaches.

The final grade of pupils has been divided into five categories: A, B, C, D, and F using a variety of classification methods, including NB, DT, and Neural Network (NN). They demonstrated that the best equal width binning and SMOTE applied to NN and NB outperformed other approaches with comparable accuracy levels. They demonstrated that NN and NB applied with SMOTE and optimum With a similar best accuracy of 75%, equal width binning surpassed competing techniques. The best time to use the prediction models is faster with NB than NN, hence NB was determined to perform better. A system for forecasting future course grades has been created by research [13] from the University of Minnesota's Computer Science and Engineering (CSE) and Electrical and Computer Engineering (ECE) schools. The results showed that, compared to the current conventional methods, the proposed methods—Matrix Factorization (MF) and Linear Regression (LinReg)—performed more precise predictions. The author also discovered that using a subset of data that is specific to a given course can increase the predictability of future course grades. Another work in [14] used 225 genuine undergraduate student data sets to apply MF, Collaborative Filtering (CF), and Restricted Boltzmann Machines (RBM) algorithms in order to predict student grades in various courses. They note that utilizing CF does not necessarily indicate good performance, particularly when the dataset has more sparsity than MF.

However, their overall findings demonstrate that the suggested RBM, especially for modelling tabular data, offers effective learning and superior prediction accuracy compared to CF and MF with minimum Root Mean Squared Error (RMSE) 0.3. A predictive model that can predict students' final grades in introductory courses at an early stage of the semester was produced by a study in [15]. With the aid of WEKA, they compared eleven machine learning algorithms across five different categories, including Bayes, Function, Lazy (IBK), Rules-Based (RB), and Decision Tree (DT). They have used feature selection correlation-based and information-gain for data preparation to reduce high dimensionality and imbalanced data. The distribution instances of three separate classes were balanced by the author using SMOTE as well. When compared to other categories of algorithms, they found that the Decision Tree classifier (J48), one of the 11 algorithms, had the best accuracy (88%). Al-Barrak [16] developed classification criteria using the DT (J48) algorithm to estimate students' final Grade Point Average (GPA) based on their performance in earlier courses. 236 graduates of King Saud University's Computer Science College from 2012 were employed.

They discovered that the classification rule generated by J48 can identify early indicators and can extract valuable information for student GPA based on their grades in all required courses to enhance performance. Using three alternative DT algorithms—Random Tree (RT), RepTree, and J48—another study in [17] predicted the student's grade performance. The effectiveness of the prediction model is evaluated in this situation via cross-validation. According to the findings, RT outperformed the other algorithms in accuracy, achieving a value of 75.188%. By including more samples and attributes in the dataset, the prediction models' accuracy can be increased. At Malaysia's University Sultan Zainal Abidin (UniSZA), [18] has put up a system for forecasting students' academic achievement. The study used 399 student records from the academic department database for the eight years' intakes, which included data on the students' demographics, prior academic performance, and family history.

3. FRAMEWORK OF MULTICLASS PREDICTION MODEL FOR STUDENT GRADE PREDICTION

With imbalanced multi-classification for student grade prediction, this research tries to determine the most reliable predictive model. As seen in Figure 1, the framework is divided into four major phases. Our architecture takes as input the final course grade for each student, which we take from their academic spreadsheet and repository. To lessen the overfitting and misclassification of the imbalanced multi-classification dataset, we used two data-level solutions utilizing oversampling SMOTE and two FS techniques. Next, we create our suggested model by fusing the two approaches into a chosen machine learning classifier, with the goal of measuring its performance with performance measures. The trend of the dataset and the final categorization outcomes are lastly visualized using data visualization. Following subsection provides a description of each phase.

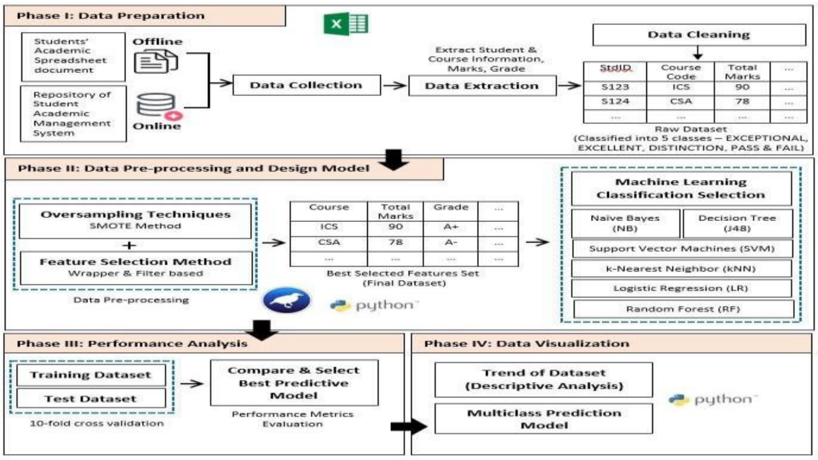

Figure1. The structure of the suggested multiclass prediction model for forecasting final student grades

Data Preparation

The Department of Information and Communication Technology (JTMK) at one of Malaysia's Polytechnics amassed the dataset that we utilized. The dataset, which includes 1282 occurrences, represents the cumulative course grades for first-semester students as determined by their final exam results during the June 2016 to December 2019 session. To qualify for the following academic semester, students must complete some required, specialized, and core course modules. The percentage of final exam and course assessment marks were only present in the two core courses that we chose for this study. The Table 2 lists every feature that is utilized for prediction.

Data Pre-Processing and Design Model

We pre-processed the acquired dataset during this phase. We have evaluated and divided the students into 5 grade categories for ease of data pre-processing: Exceptional (A+), Excellent (A), Distinction (A, B+, B), Pass (B, C+, C, C, D+, D), and Fail (E, E, F). To serve as the prediction class's output, the group was formed. The dataset's class distribution, however, showed an unbalanced class instance population that included a significant proportion of ratio 3:18:30:9:1, (186) pass, (21) fail, (377) excellent, (635) distinction, and (63) outstanding class instances, which can result in overfitting of the results.

As a result, the benchmark approaches employed in this work to address the issue of an unbalanced multi-classification dataset were a data-level solution employing oversampling SMOTE and two FS methods: Wrapper and Filter based. In the experiment3.8.3 of the opensource application Waikato Environment for Knowledge- edge Analysis (WEKA) was utilized because it offers several machine learning algorithms with user-friendly graphical interfaces for straightforward display

Performance Analysis

Based on the first semester's final exam results and the students' prior course performance records, this study attempts to forecast the final grades of the students. To determine which machine learning algorithm offered the best performance for the task at hand, the proposed model used a variety of algorithms prediction of final grades for students. According to four unique phases, three experiments were carried out the various classes, which are five. With our dataset divided into 90% for training set and 10% for testing set on the same dataset, ten-fold cross-validation is used to evaluate accuracy. The flowchart of the suggested multiclass prediction model used in this investigation is shown in Figure 2. The theoretical model that served as the foundation for the development of our multiclass prediction model is specifically as follows:

logistic regression (LR) in order to tackle classification issues, logistic function was utilized as a representation of mathematical modelling. Analysis comprehends the interrelationships of the factors.

Bayesian theorem, on which Nave Bayes (NB) is built Since it is straightforward and quick in making predictions, it is commonly employed. It works well with tiny datasets that combine using a flexible probabilistic model to reduce complexity.

With its ability to accommodate missing values and high dimensional data, the decision tree (J48) is a popular tool for multi-class categorization. It has been successfully implemented in order to provide the best accuracy with the fewest features.

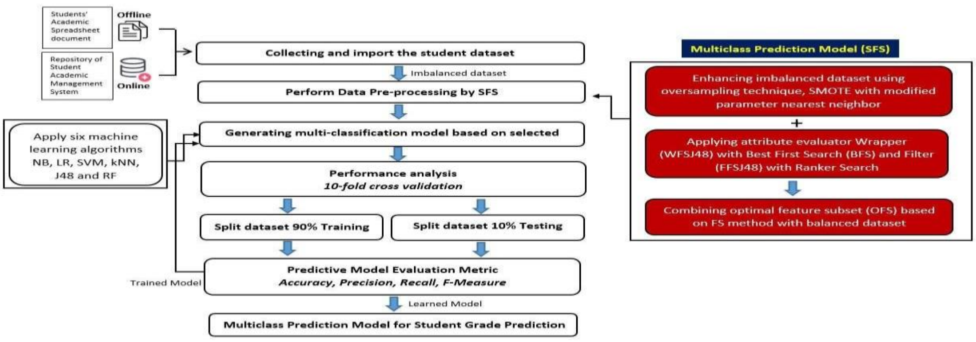

Figure 2. Flowchart of the proposed multiclass prediction model

Support Vector Machine (SVM) is built on the idea of decision planes that provide decision limits and successfully handles

© 2023, www.IJARIIT.com All Rights Reserved

categorization problems. The SVM is a non-probabilistic binary linear classifier since it uses a sorted dataset to predict which of two possible classes contains the data.]

The non-parametric approach K-Nearest Neighbor (kNN) categorizes instances in the dataset based on their closest vectors, where k denotes the distance in the n-dimensional space, and calculates the difference between examples based on that. The dataset's tiny features are appropriately performed using a distance function.

Random Forest (RF) is an ensemble learning-based classifier that uses a variety of decision trees on different subsets to discover the optimum features for high accuracy and avoids the overfitting issue. The RF performs well in classification and is relatively resilient to outliers and noise.

Table 1. The information of the input features

TABLE 2. The information of the input features.

Attribute	Type	Values	Description
StudID	Nominal	S1-S641	Student Identification
Year	Numeric	[2016,2019]	Year of student intake
Class	Nominal	DDT1A, DDT1B, DDT1C, DDT1D	Class of student
Session	Nominal	DEC, JUNE	Session of student intake per year
Credit Hour	Numeric	[3]	Credit hour of each course
Course Code	Nominal	[CSA, ICS]	Course ID of 2 courses
Total Marks	Numeric	[38,91]	Student Final Marks obtained from final exam and courses assessment
Grade Pointer Average	Numeric	[0.00,4.00]	Student course grade pointer
Grade	Nominal	[A+, A, A-, B+, B, B-, C+, C, C-, D+, D, E, F]	Student Final Grade of each course
Group	Nominal	EXCEPTIONAL, EXCELLENT, DISTINCTION, PASS, FAIL	Category of student academic performance

Algorithm 1 Algorithm for Multiclass Prediction Model (SFS)

Input: The training dataset

Output: The predicted Student's Grade label, *SG*

1 **Begin**
2. Import necessary library packages and select dataset
3. Perform data preprocessing
   3.1 Select filters for oversampling
   3.2 Set parameter of SMOTE (*nearest neighbor*, $k = 10$)
   3.3 Select features with attribute evaluator & search method
   3.4 Select attribute selection mode (*Use full training set*)
4. Use classification models to predict the results
   4.1. Splitting data into training and testing dataset using 10-fold cross validation
   4.2. Using well-known classification models (*J48, kNN, SVM, LR, NB, RF*) to predict the *SG (Exceptional, Excellent, Distinction, Pass, Fail)*
5. Evaluate the accuracy of well-known classification models
6. **end**

To see how well each predictive model does in categorizing data, use a confusion matrix. The confusion matrix for predicting student grades is shown in Table 3, where the letters A, B, C, D, and E stand in for the classes that are "extraordinary," "outstanding," "distinction," "pass," and "failure" for the purposes of student grade (SG) level. A form of the expression is represented by the class label:

$$SG\{A,B,C,D,E\} \qquad (1)$$

Table 2. Confusion matrix for the classification of student grade predictions

		Predicted				
		A	B	C	D	E
Actual Label	A	AA	AB	AC	AD	AE
	B	BA	BB	BC	BD	BE
	C	CA	CB	CC	CD	CE
	D	DA	DB	DC	DD	DE
	E	EA	EB	EC	ED	EE

Accuracy, precision, recall, and f-measure are used in the following equation to calculate the confusion matrix's performance metrics:

$$\textbf{\textit{Accuracy}}\ (A) = \frac{(AA + BB + CC + DD + EE)}{\sum N} \qquad (2)$$

where N is the number of samples

$$\textbf{\textit{Precision}}\ (P) = \frac{1}{5}\left(\frac{AA}{AA + BA + CA + DA + EA} + \frac{BB}{AB + BB + CB + DB + EB} + \frac{CC}{AC + BC + CC + DC + EC} + \frac{DD}{AD + BD + CD + DD + ED} + \frac{EE}{AE + BE + CE + DE + EE}\right) \qquad (3)$$

$$\textbf{\textit{Recall}}\ (R) = \frac{1}{5}\left(\frac{AA}{AA + AB + AC + AD + AE} + \frac{BB}{BA + BB + BC + BD + BE} + \frac{CC}{CA + CB + CC + CD + CE} + \frac{DD}{DA + BD + DC + DD + DE} + \frac{EE}{EA + EB + EC + ED + EE}\right) \qquad (4)$$

$$F - Measure = 2\frac{PR}{P + R} \qquad (5)$$

Data Visualization

Using Python, we extracted and visualized our findings in this phase after doing the data analysis to evaluate the pertinent data and trends in student grade performance across various courses. To assist lecturers in enhancing students' academic performance for better decision-making in the future, data visualization makes it possible to uncover all the features and insightful facts in the student dataset. Additionally, in order to better comprehend the outcomes of the data, we compare each result of our suggested model using a better graphical manner.

4. DESCRIPTIVE ANALYSIS OF STUDENT DATASET

641 students who took the core courses Computer System Architecture (CSA) and Introduction to Computer System are represented in our dataset (ICS). Following the pass grade (B, C+, C, C, D+, D) with 176 students, the excellent grade (A) with 80 students, the failed grade (E, E, F) with 19 students, and the outstanding grade (A+) with 4 students, we discovered that 362 students received distinction grades (A, B+, B) in CSA courses. However, for the ICS course, excellent grade (A) with 297 students came out on top, followed by distinction grade (A, B+, B) with 273 students, outstanding grade (A+) with 59 students, pass grade (B, C+, C, C, D+, D), and failed grade (E, E, F) with 10 and 2 students, respectively. In accordance with our investigation, the final student grade mean and standard deviation for the CSA course were 68.95 and 9.189, respectively, and 79.62 and 7.379 for the ICS course. The total enrollment for both courses is shown in Table 4.

Table 3. Results of the student performance by course

Student Final Grade	No. of Student	
	CSA	ICS
Exceptional	4	59
Excellent	80	297
Distinction	362	273
Pass	176	10
Fail	19	2

5. EXPERIMENTAL RESULTS

Smote Oversampling Technique

The method most frequently employed to solve the overt problem based on random sampling algorithm is called SMOTE, or Synthetic Minority Oversampling Technique [29]. In order to make the distribution more balanced, it can change an unbalanced dataset and generate new instances of minority classes from already existing ones. By raising the nearest neighbours' default value (k) in the minority class sample SG sample, N samples were randomly chosen and recorded as SGi. By using the following expression, the new sample SGnew is rejected. where the ratio of creating new samples is roughly 100% and rand is a seed used for random sampling in the range (0,1) and class index value 0. Weka. filters.s upervised. instance was implemented. SMOTE to add artificial instances between neighbouring dataset samples from the minority class. To automatically discover the non-empty minority class, we set the index class option to value 0. Then, the SMOTE filter was applied ten times during iteration, with the number of nearest neighbours' k value set up to equal 10 (k D 10) and the percentage of instances set to 100%.

$$SG_{new} = SG_{origin} + rand \times \left(SG_i - SG_{origin}\right), \quad i = 1, 2, 3, \ldots n \quad (6)$$

The number of instances where the SG class distribution using SMOTE becomes (504) exceptional, (377) excellent, (635) distinction, (744) pass, and (672) fail has increased due to the oversampled dataset, going from 1282 up to 2932 instances in total. This is due to a reduction in the ratio to 1:1:2:2:2. The results of all the predictive models' detailed comparisons with all the performance measures are shown in Table 4.

All predictive models consistently performed better when the classifiers were combined with oversampling SMOTE, as was observed. The most promising of these prediction models, RF, produced an f-measure of 99.5%. It was followed by kNN, J48, SVM, LR, and NB, which each produced an f-measure of 99.3%. Figure 6 illustrates the statistical significance of this finding with a 95% confidence level using paired T-tester (adjusted). We also noticed that, after applying the SMOTE approach, the minority class instance increased by the number of iterations and the amount of k values to our dataset in order to balance with the other classes. Based on the confusion matrix shown in Table 7, a detailed study of the accuracy performance was presented. It is clear that the confusion matrix for all predictive models developed from J48, NB, kNN, SVM, LR, and RF improves the accuracy of accurately classifying grades of "Pass" and "Fail." Though there is a slight decline in performance from SVM, the predictive model nevertheless accurately identified 97.2% of students who received "Pass" marks, down from 99.5% when SMOTE was not used. Figures 7 and Figure 8 show real test scores and projections based on four grade-level categories before and after the SMOTE, respectively, for comparison examination. With the exception of the minority class, each predictive model's performance significantly improves for the groups that make up the majority.

Feature Selection

Feature selection (FS), which is effective in lowering dimensionality, eliminating pointless data, and improving learning accuracy, is a further experiment we conducted. The performance of six prediction models was maximised in this experiment by using two

© 2023, www.IJARIIT.com All Rights Reserved

FS approaches, wrapper and filter based, as the benchmark methods. The J48 classifier and two attribute evaluators, WrapperSubsetEval (FS-1) and Classifier SubsetEval (FS-2), are utilised in the FS wrapper technique, which was used in this study to select the optimal feature set. The InfoGainAttributeEval (FS-3) feature set was chosen as the best one for the FS filter algorithm since it had a ranker search method greater than 0.5. Table 8 lists how many characteristics there are in each of the two FS algorithms.

```
Dataset                      (1) trees.Ra | (2) trees (3) lazy. (4) funct (5) funct (6) bayes
--------------------------------------------------------------------------------------------
new5-weka.filters.supervi(100)   99.53 |    99.16 *   99.35      98.86 *   98.82 *   98.35 *
--------------------------------------------------------------------------------------------
                                 (v/ /*) |  (0/0/1)   (0/1/0)    (0/0/1)   (0/0/1)   (0/0/1)
```

Figure 6. Result of predictive model performance with smote.

To find the best predictive model that satisfied the criteria for producing an ideal result, we conducted the study using the same dataset.

Results from several predictive models using all FS algorithm measurements are summarised in Table 9. In contrast to other prediction models, the results showed that kNN had the highest performance f-measure score, reaching 98.8% and 98.9%, respectively, with the best feature sets chosen.

NB exhibits the lowest accuracy performance, as seen in Table 9, yet the f-measure for NB shows a little improvement when the FS-2 method was used, increasing from 97.8% to 98.2%. On the other hand, as compared to no FS was used, the performance of J48, LR, SVM, and RF revealed only modestly promising results. The learning performance to better forecast student grade is hampered by limiting the amount of features for the multi-class dataset's high imbalance ratio. In Figure 9, a comparison of the best accuracy and f-measure score with various FS is shown.

5. CONCLUSION AND FUTURE DIRECTIONS:

One of the important performance indicators that might assist instructors in keeping track of students' academic progress is predicting their grades performance. As a result, it's crucial to have a predictive model that can lower the degree of uncertainty in the result for a dataset that is unbalanced. Based on the prior student final examination results from the first-semester course, this study suggests a multiclass prediction model with six predictive models to forecast the final student grades.

In order to assess the performance accuracy of student grade prediction, we specifically conducted a comparison analysis of integrating oversampling SMOTE with other FS approaches. We also demonstrated that utilizing SMOTE with oversampling is consistently better than using FS alone with all prediction models. However, employing oversampling SMOTE and FS alone with specific parameter settings that can affect the performance accuracy of all predictive models, our suggested multiclass prediction model outperformed them. Here, based on the data-level solution for student grade prediction, our findings help to provide a useful strategy for dealing with the imbalanced multi-classification.

Predictive analytics is a key component of governance in the HEI, improving valuable data and fostering reliable decision-making that advances data science. One of the difficult problems that persists in choosing the pertinent and useful predictive models is determining the quality of the acquired dataset to reduce the imbalance and missing values challenges. As a result, more research into the application of suitable new predictive techniques in such sophisticated machine learning algorithms as well as more ensemble algorithms are advised for future works to maximize the outcome for predicting student grades. Additionally, it is crucial to choose a number of multi-class imbalanced datasets for analysis using appropriate sampling strategies and various evaluation metrics, suitable for the imbalanced multi-class domain, such as Kappa, Weighted Accuracy, and other measures. Therefore, utilizing machine learning for student grade prediction in higher education institutions will ultimately improve the decision support system to improve their students' academic performance in the future.

6. REFERENCES

[1] D. Solomon, S. Patil, 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. Düştegör, ‘‘Predicting academic success in higher education: Literature review and best practices,’’ Int. J. Educ. Technol. Higher Educ., vol. 17, no. 1, Dec. 2020.

[2] V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, ‘‘Early segmentation of students according to their academic performance: A predictive modelling approach,’’ Decis. Support Syst., vol. 115, pp. 36– 51, Nov. 2018. [4] P. M. Moreno-Marcos, T.-C. Pong, P. J. Munoz-Merino, and C. D. Kloos, ‘‘Analysis of the factors influencing Learners’ performance prediction with learning analytics,’’ IEEE Access, vol. 8, pp. 5264–5282, 2020.

[3] A. E. Tatar and D. Düştegör, ‘‘Prediction of academic performance at undergraduate graduation: Course grades or grade point average?’’ Appl. Sci., vol. 10, no. 14, pp. 1–15, 2020.

[4] Y. Zhang, Y. Yun, H. Dai, J. Cui, and X. Shang, ‘‘Graphs regularized robust matrix factorization and its application on student grade prediction,’’ Appl. Sci., vol. 10, p. 1755, Jan. 2020.

[5] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, ‘‘Educational data mining and learning analytics for 21st century higher education: A review and synthesis,’’ Telematics Informat., vol. 37, pp. 13–49, Apr. 2019.

[6] K. L.-M. Ang, F. L. Ge, and K. P. Seng, ‘‘Big educational data & analytics: Survey, architecture and challenges,’’ IEEE Access, vol. 8, pp. 116392–116414, 2020.

[7] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, ''Predicting academic performance: A systematic literature review,'' in Proc. 23rd Annu. Conf. Innov. Technol. Comput. Sci. Educ., Jul. 2018, pp. 175–199.

[8] L. M. Abu Zohair, ''Prediction of student's performance by modelling small dataset size,'' Int. J. Educ. Technol. Higher Educ., vol. 16, no. 1, pp. 1–8, Dec. 2019, doi: 10.1186/s41239-019-0160-3.

[9] X. Zhang, R. Xue, B. Liu, W. Lu, and Y. Zhang, ''Grade prediction of student academic performance with multiple classification models,'' in Proc. 14th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD), Jul. 2018, pp. 1086–1090.

[10] 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

[11] A. Polyzou and G. Karypis, ''Grade prediction with models specific to students and courses,'' Int. J. Data Sci. Anal., vol. 2, nos. 3–4, pp. 159–171, Dec. 2016.

[12] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, ''Machine learning based student grade prediction: A case study,'' 2017, arXiv:1708.08744. [Online]. Available: https://arxiv.org/abs/1708.08744

[13] 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.

[14] M. A. Al-Barrak and M. Al-Razgan, ''Predicting students final GPA using decision trees: A case study,'' Int. J. Inf. Educ. Technol., vol. 6, no. 7, pp. 528–533, 2016.

[15] 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.

[16] F. Ahmad, N. H. Ismail, and A. A. Aziz, ''The prediction of students' academic performance using classification data mining techniques,'' Appl. Math. Sci., vol. 9, pp. 6415–6426, Apr. 2015.

[17] T. Anderson and R. Anderson, ''Applications of machine learning to student grade prediction in quantitative business courses,'' Glob. J. Bus. Pedagog., vol. 1, no. 3, pp. 13–22, 2017.

[18] [20] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, ''Educational data mining and analysis of students' academic performance using WEKA,'' Indonesian J. Electr. Eng. Comput. Sci., vol. 9, no. 2, pp. 447–459, 2018.

[19] [21] A. Verma, ''Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA,'' Int. Res. J. Eng. Technol., vol. 6, no. 3, D. Berrar, ''Cross-validation,'' Comput. Biol., vols. 1–3, pp. 542–545, Jan. 2018, doi: 10.1016/B9780-12-809633-8.20349-X. pp. 54–60, 2019

[20] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, ''Using machine learning to predict student difficulties from learning session data,'' Artif. Intell. Rev., vol. 52, no. 1, pp. 381–407, Jun. 2019

[21] B. Predić, G. Dimić, D. Rančić, P. Štrbac, N. Maček, and P. Spalević, ''Improving final grade prediction accuracy in blended learning environment using voting ensembles,'' Comput. Appl. Eng. Educ., vol. 26, no. 6, pp. 2294–2306, Nov. 2018, doi: 10.1002/cae.22042.

[22] K. Srivastava, D. Singh, A. S. Pandey, and T. Maini, ''A novel feature selection and short-term price forecasting based on a decision tree (J48) model,'' Energies, vol. 12, p. 3665, Jan. 2019.

[23] L. E. O. Breiman, ''Random Forests,'' Mach. Learn., vol. 45, pp. 5–32, Oct. 2001.

[24] T. M. Barros, P. A. SouzaNeto, I. Silva, and L. A. Guedes, ''Predictive models for imbalanced data: A school dropout perspective,'' Educ. Sci., vol. 9, no. 4, p. 275, Nov. 2019.

[25] T. Alam, C. F. Ahmed, S. A. Zahin, M. A. H. Khan, and M. T. Islam, ''An effective recursive technique for multi-class classification and regression for imbalanced data,'' IEEE Access, vol. 7, pp. 127615– 127630, 2019.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''SMOTE: Synthetic minority oversampling technique,'' J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[27] C. Jalota and R. Agrawal, Feature Selection Algorithms and Student Academic Performance: A Study, vol. 1165. Singapore: Springer, 2021.

[28] G. A. Sharifai and Z. Zainol, ''Feature selection for high-dimensional and correlation based redundancy and binary,'' Genesm, vol. 11, pp. 1–26, Jun. 2020.

[29] Buenaño-Fernández, D. Gil, and S. Luján-Mora, ''Application of machine learning in predicting performance for computer engineering students: A case study,'' Sustain., vol. 11, no. 10, pp. 1–18, 2019.

[30] S. Chinna Gopi, B. Suvarna, and T. Maruthi Padmaja, ''High dimensional unbalanced data classification Vs SVM feature selection,'' Indian J. Sci. Technol., vol. 9, no. 30, Aug. 2016.

[31] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K. U. Sarker, and M. U. Sattar, ''Predicting student performance in higher educational institutions using video learning analytics and data mining techniques,'' Appl. Sci., vol. 10, no. 11, p. 3894, Jun. 2020.

[32] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, ''Learning k for kNN Classification,'' ACM Trans. Intell. Syst. Technol., vol. 8, no. 3, pp. 1–9, 2017.

[33] [36] P. Nair and I. Kashyap, ''Optimization of kNN classifier using hybrid preprocessing model for handling imbalanced data,'' Int. J. Eng. Res. Technol., vol. 12, no. 5, pp. 697–704, 2019.

[34] P. Brous and M. Janssen, ''Trusted decision-making: Data governance for creating trust in data science decision outcomes,'' Administ. Sci., vol. 10, no. 4, p. 81, Oct. 2020. [39] 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.

[35] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, ''Implementing autoML in educational data mining for prediction tasks,'' Appl. Sci., vol. 10, no. 1, pp. 1–27, 2020.

© 2023, www.IJARIIT.com All Rights Reserved