

Received May 26, 2021, accepted June 12, 2021, date of publication June 30, 2021, date of current version July 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3093563

Multiclass Prediction Model for Student Grade Prediction Using Machine Learning

SITI DIANAH ABDUL BUJANG¹, ALI SELAMAT^{1,2}, (Member, IEEE),
ROLIANA IBRAHIM², (Member, IEEE), ONDREJ KREJCAR³,
ENRIQUE HERRERA-VIEDMA⁴, (Fellow, IEEE),
HAMIDO FUJITA⁵, (Life Senior Member, IEEE),
AND NOR AZURA MD. GHANI⁶, (Member, IEEE)

¹Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Kuala Lumpur 54100, Malaysia

²Malaysia and Media and Games Center of Excellence (MagicX), Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia, Johor Baharu 81310, Malaysia

³Faculty of Informatics and Management, University of Hradec Kralove, 50003 Hradec Kralove, Czech Republic

⁴Department of Computer Science and AI, Andalusian Research Institute in Data Science and Computational Intelligence, University of Granada, 18071 Granada, Spain

⁵i-SOMET Incorporated Association, Morioka 020-0104, Japan

⁶Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor 40450, Malaysia

Corresponding author: Ali Selamat (aselamat@utm.my)


This work was supported in part by the Ministry of Higher Education through the Fundamental Research Scheme under Grant FRGS/1/2018/ICT04/UTM/01/1, in part by the Specific Research Project (SPEV) at the Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic, under Grant 2102-2021, in part by the Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-20H04, and in part by the Malaysia Research University Network (MRUN) under Grant Vot 4L876.

ABSTRACT Today, predictive analytics applications became an urgent desire in higher educational institutions. Predictive analytics used advanced analytics that encompasses machine learning implementation to derive high-quality performance and meaningful information for all education levels. Mostly know that student grade is one of the key performance indicators that can help educators monitor their academic performance. During the past decade, researchers have proposed many variants of machine learning techniques in education domains. However, there are severe challenges in handling imbalanced datasets for enhancing the performance of predicting student grades. Therefore, this paper presents a comprehensive analysis of machine learning techniques to predict the final student grades in the first semester courses by improving the performance of predictive accuracy. Two modules will be highlighted in this paper. First, we compare the accuracy performance of six well-known machine learning techniques namely Decision Tree (J48), Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (kNN), Logistic Regression (LR) and Random Forest (RF) using 1282 real student's course grade dataset. Second, we proposed a multiclass prediction model to reduce the overfitting and misclassification results caused by imbalanced multi-classification based on oversampling Synthetic Minority Oversampling Technique (SMOTE) with two features selection methods. The obtained results show that the proposed model integrates with RF give significant improvement with the highest f-measure of 99.5%. This proposed model indicates the comparable and promising results that can enhance the prediction performance model for imbalanced multi-classification for student grade prediction.

INDEX TERMS Machine learning, predictive model, imbalanced problem, student grade prediction, multi-class classification.

I. INTRODUCTION

In higher education institutions (HEI), every institution has its student academic management system to record all student

The associate editor coordinating the review of this manuscript and approving it for publication was Syed Islam .

data containing information about student academic results in final examination marks and grades in different courses and programs. All student marks and grades have been recorded and used to generate a student academic performance report to evaluate the course achievement every semester. The data keep in the repository can be used to discover

insightful information related to student academic performance. Solomon *et al.* [1] indicated that determining student academic performance is a crucial challenge in HEI. Due to this, many previous researchers have well-defined the influence factors that can highly affect student academic performance [2]. However, most common factors are relying on socioeconomic background, demographics [3] and learning activities [4] compared to final student grades in the final examination [5]. As for this reason, we observe that the trend of predicting student grades can be one of the solutions that are applicable to improve student academic performance [6].

Predictive analytics has shown the successful benefit in the HEI. It can be a potential approach to benefit the competitive educational domain to find hidden patterns and make predictions trends in a vast database [7]. It has been used to solve several educational areas that include student performance, dropout prediction, academic early warning systems, and course selection [8]. Moreover, the application of predictive analytics in predicting student academic performance has increased over the years [9].

The ability to predict student grade is one of the important area that can help to improve student academic performance. Many previous research has found variant machine learning techniques performed in predicting student academic performance. However, the related works on mechanism to improve imbalanced multi-classification problem in predicting students' grade prediction are difficult to found [10], [11]. Therefore, in this study, a comparative analysis has been done to find the best prediction model for student grade prediction by addressing the following questions:

RQ1: Which predictive model among the selected machine learning algorithms performs high accuracy performance to predict student's final course grades?

RQ2: How imbalanced multi-classification dataset can be addressed with selected machine learning algorithms using oversampling Synthetic Minority Oversampling Technique (SMOTE) and feature selection (FS) methods?

To address the above-mentioned questions, we collect the student final course grades from two core courses in the first semester of the final examination result. We present a descriptive analysis of student datasets to visualize student grade trends, which can lead to strategic planning in decision making for the lecturers to help students more effectively. Then, we conduct comparative analysis using six well-known machine learning algorithms, including LR, NB, J48, SVM, kNN and RF on the real student data of Diploma in Information Technology (Digital Technology) at one of Malaysia Polytechnic. As for addressing the imbalanced multi-classification, we endeavor to enhance the performance of each predictive model with data-level solutions using oversampling SMOTE and FS. The novel contribution of this paper are summarized as follows:

- We proposed combination of modification on oversampling SMOTE and two feature selection algorithms to automatically determine the sampling ratio with

best selected features to improve imbalanced multi-classification for student grade prediction.

- Our comparative analysis showed that the ratio between the minority class in imbalanced dataset does not necessarily to approach same ratio of majority class to obtain better performance in student grade prediction.
- 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.

This paper is organized as follows. Section II describes the related research work that has been conducted for student grade prediction. Section III illustrates the methodology of developing predictive models to predict final student grades by phases. Section IV and Section V present the descriptive analysis and prediction results of this study's findings, respectively. Section VI discusses the findings result. Lastly, the paper is highlighted with the main conclusions with some future directions in Section VII.

II. RELATED WORKS

Several studies have been conducted in HEI for predicting student grades using various machine learning techniques. It involves analytical process of many attributes and samples data from variety of sources for student grade prediction in different outcome. However, the performance of predictive model for imbalanced dataset in education domains are still rarely discussed. Related to this issues, a study from [12] used discretization and oversampling SMOTE methods to improve the accuracy of students' final grade prediction. Several classification algorithms have been applied such as NB, DT and Neural Network (NN) for classifying students' final grade into five categories; A, B, C, D and F. They showed that NN and NB applied with SMOTE and optimal equal width binning outperformed other methods with similar highest accuracy of 75%. However, NB found better compared to NN as the optimal time to utilize the prediction models are faster than NN. Research conducted by [13], has developed a method for predicting future course grades obtained from the Computer Science and Engineering (CSE) and Electrical and Computer Engineering (ECE) programs at the University of Minnesota. Based on the proposed methods, the results indicated that Matrix Factorization (MF) and Linear Regression (LinReg) performed more accurate predictions than the existing traditional methods. The author also found that the use of a course-specific subset of data can improve prediction accuracy for predicting future course grades. Another study in [14], applied MF, Collaborative Filtering (CF) and Restricted Boltzmann Machines (RBM) techniques on 225 real data of undergraduate students to predict student grade in different courses. They observe that using CF does not indicate good performance especially when there found many sparsity in the dataset compared to MF. However, their overall findings show that the proposed RBM provides efficient learning and better prediction accuracy compared to CF and MF with minimum Root Mean Squared

Error (RMSE) 0.3 especially for modeling tabular data. A study in [15] has developed a predictive model that can predict student's final grades in introductory courses at an early stage of the semester. They have compared eleven machine learning algorithms in five different categories consist of Bayes, Function, Lazy (IBK), Rules-Based (RB) and Decision Tree (DT) using WEKA. To reduce high dimensionality and unbalanced data, they have performed feature selection correlation-based and information-gain for data-preprocessing. The author also applied SMOTE to balance the distribution instances of three different classes. Among the 11 algorithms, they indicated that Decision Tree classifier (J48) have the highest accuracy of 88% compared to other categories of algorithms. Al-Barrak [16] used DT (J48) algorithm to discover classification rules for predicting students' final Grade Point Average (GPA) based on student grades in previous courses. They have used 236 students who graduated from Computer Science College at King Saud University in 2012. They found that the classification rule produced from J48 can detect early predictors and can extract useful knowledge for final student GPA based on their grades in all mandatory courses to improve students' performance. Another study in [17] have predicted the student's grade performance using three different DT algorithms; Random Tree (RT), RepTree and J48. In this context, cross-validation is used to measure the performance of the predictive model. From the findings, the results indicated that RT obtained the highest accuracy of 75.188% better than the other algorithms. The accuracy of the predictive models can be improved by adding more number of samples and attributes in the dataset. [18] has proposed a framework for predicting student academic performance at University Sultan Zainal Abidin (UniSZA), Malaysia. The study applied 399 student records from the academic department database in the eight years' intakes that contained student demographics, previous academic records and family background information. The results indicated that the Rule-Based (PART) is the best model with 71.3% accuracy compared to DT and NB. However, using the small sample size has affected accuracy performance due to incomplete and missing value found in the dataset. Anderson and Anderson [19] performed an experimental study on 683 students at the Craig School of Business at California State University from 2006 to 2015 by applying three machine learning algorithms to predict student grades. The study found that SVM is the best classifier. It consistently outperforms a simple average approach that obtained the lowest error rate to optimize each data class. The result could be different for the large set of data due to significant changes in the historical grade dataset's structure and format. We have summarized related studies composed of sample size, data source, attributes, algorithm, best performance and limitation in Table 1.

III. FRAMEWORK OF MULTICLASS PREDICTION MODEL FOR STUDENT GRADE PREDICTION

This paper aims to identify the most effective predictive model especially in addressing imbalanced 95610

multi-classification for student grade prediction. The framework consists of four main phases is shown in Figure 1. The input of our framework contain student's final course grade that we extract from student's academic spreadsheet document and student academic repository. We applied two data-level solution using oversampling SMOTE and two FS methods to reduce the overfitting and misclassification of imbalanced multi-classification dataset. Then, we design our proposed model by combining both techniques into selected machine learning classifier to evaluate the performance using performance metrics. Finally, data visualization is used to visualize the trend of dataset and final classification results. The description of each phases is given in the following subsection.

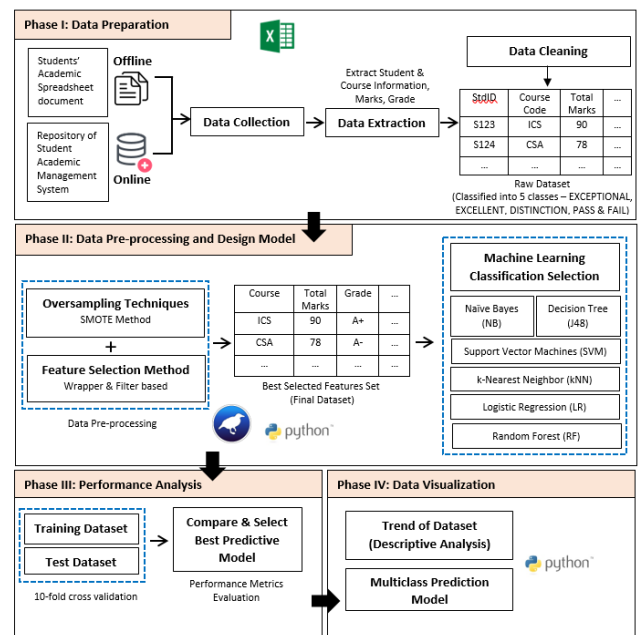


FIGURE 1. The framework of the proposed multiclass prediction model for predicting final student grades.

A. DATA PREPARATION

The dataset we used was collected by the Department of Information and Communication Technology (JTMK) at one of the Malaysia Polytechnics. The dataset contains 1282 instances which is the total course grades of the first semester students taken from the final examination during June 2016 to December 2019 session. Students need to take some compulsory, specialization and core courses modules to qualify them for the next academic semester. However, in this study we selected only two core courses that contained the percentage of final examination and course assessment marks. All features which are used for prediction are listed in the Table 2.

B. DATA PRE-PROCESSING AND DESIGN MODEL

In this phase, we applied data pre-processing for the collected dataset. For the convenience of data pre-processing, we have

TABLE 1. The taxonomy of related studies on student grade prediction.

Paper	Sample Size	Data Source	Attributes	Algorithm	Best Performance	Limitation
Jishan et al. [12]	180	Student Core Course offered at North South University, Bangladesh	CGPA, Quiz, Midterm, Lab, Attendance, Final grade	NB, DT, Neural Network Backpropagation with oversampling (SMOTE) and optimal binning	NB (optimal binning+SMOTE) Accuracy 75.28%	Small size of attributes that lead to high misclassification error
Polyzou et al. [13]	76,748	Student-course grade from 2002 to 2014	Historical student course grade information	LinReg, MF	LinReg	Not support a large number of latent factors
Iqbal et al. [14]	225	Undergraduate students of the Electrical Engineering Program from 2013 to 2015	Grades, GPA	CF, MF, RBM	RBM	Use limited attribute for analysis
Khan et al [15]	50	Student of Buraimi University College, Oman	Test1_marks, CGPA, Attendance, Major, Gender, Year	NB, MLP, SVM, Lazy (IBK), Rules-Based (Decision Table, JRIP, OneR, PART and ZeroR) DT ((J48), RF, RT SimpleCART)	DT (J48) (Feature Selection+SMOTE) Accuracy 88%	Small number of dataset
Barrak et al. [16]	236	A female student from Computer Sciences College at King Saud University 2012	Student name, student ID, final GPA, the semester of graduation, major, nationality, campus, courses are taken and course grade	DT (J48)	DT (J48)	Lack of experimental techniques for prediction
Abana [17]	133	Students of Computer Engineering program in 4 years	Research Method (RM) grade, Research Project (RP) grade, gender, backlog, programming proficiency	RT, RepTree DT (J48)	RT Accuracy 75.2%	Lack of experimental techniques for prediction
Ahmad et al. [18]	399	First-year bachelor students in Computer Science at UniSZA from 2006/2007 to 2013/2014	GPA, race, gender, family income, university entry mode, Malaysia Certificate of Education (SPM) grade in 3 subjects	DT, NB and Rule Based (PART)	RB Accuracy 71.3%	A small number of dataset due to incomplete and missing value
Anderson et al. [19]	683	Students of Craig School of Business at California State University, Fresno from 2006 to 2015	Historical grade data from 18 semester	NB, KNN, SVM	SVM	Some of the dataset are not available due to significant changes.

CF – Collaborative Filtering, MF-Matric Factorization, DT- Decision Tree, LinReg- Linear Regression, RBM - Restricted Boltzmann Machines, MLP- Multilayer Perceptron, NB – Naïve Bayes, RT – Random Tree, Lazy (IBK) – K-Nearest Neighbor, RF – Random Forest, SVM – Support Vector Machine

ranked and grouped the students into 5 categories of grades: Exceptional (A+), Excellent (A), Distinction (A–, B+, B), Pass (B–, C+, C, C–, D+, D) and Fail (E, E–, F). The group was created to be the output of the prediction class. However, the class distribution of the dataset indicated an imbalanced class instances containing number of (63) exceptional, (377) excellent, (635) distinction, (186) pass and (21) fail with high number of ratio 3:18:30:9:1 that can lead to overfit results.

Therefore, data-level solution using oversampling SMOTE and two FS methods; Wrapper and Filter based were used as the benchmark methods in this study to overcome the problem of imbalanced multi-classification dataset. The experiment used the open-source tool Waikato Environment for Knowledge Analysis (WEKA) version 3.8.3 because it provides many machine learning algorithms with easy graphical user interfaces for simple visualization [20], [21].

C. PERFORMANCE ANALYSIS

This paper aims to predict students’ final grades based on their previous course performance records in the first semester’s final examination. The proposed model applied

different machine learning algorithms to evaluate which of the algorithms performed the highest performance for predicting student’s final grades. There are three experiments were conducted in four distinct phases based on the five different classes. The accuracy is evaluated using ten-fold cross-validation which our dataset is partitioned into 90% for training set and 10% for testing set on the same dataset [22].

Figure 2 illustrates the flowchart of the proposed multiclass prediction model applied in this study.

In particular, the following are the theoretical model used as basis to construct our multiclass prediction model:

- Logistic Regression (LR) known as cost function that used logistic function as represent mathematical modeling to solve classification problems. The model performs great contextual analysis for categorical data to understand the relationship between variables [23].
- Naïve Bayes (NB) is based on Bayesian theorem that widely used as it is simple and able to make fast predictions. It is suitable for small datasets that combines complexity with a flexible probabilistic model [24].

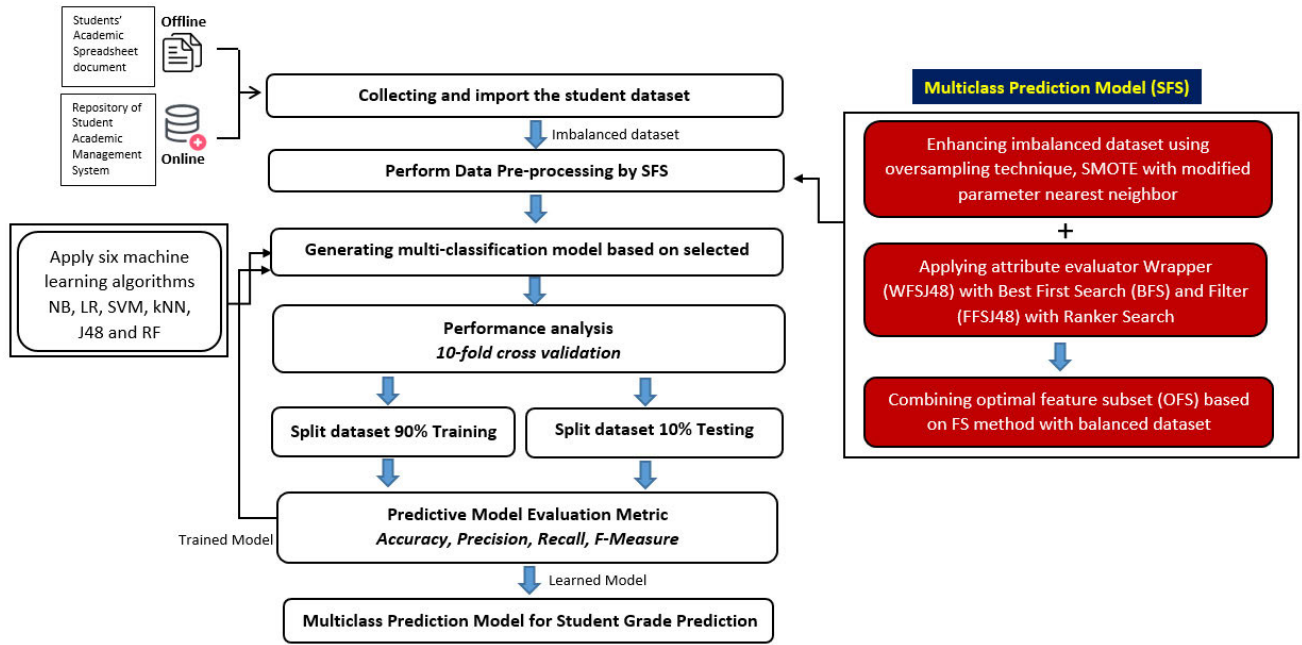


FIGURE 2. Flowchart of the proposed multiclass prediction model.

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

- Decision Tree (J48) a widely used in several multi class classification that can handle missing values with high dimensional data. It has been implemented effectively for giving an optimum results of accuracy with minimum number of features [25].
- Support Vector Machine (SVM) is based on the notion of decision planes that states decision boundaries which handle classification problem successfully [11]. It takes

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**

a sorted dataset and predicts, which of two conceivable classes includes the information, making the SVM a non-probabilistic binary linear classifier.

- K-Nearest Neighbor (kNN) is a non-parametric algorithm that classifies and calculate the difference between instances in the dataset based on their nearest vectors

where k refers to the distance in the n - dimensional space. It uses a distance function to suitability performs in small features of dataset [11].

- Random Forest (RF) is a classifier based on ensemble learning that used number of decision trees on various subset to find the best features for high accuracy and prevents the problem of overfitting. The RF is relatively robust to outliers and noise that operates effectively in classification [26].

A confusion matrix helps to visualize the classification performance of each predictive model. Table 3 presents the confusion matrix used for student grade prediction where A, B, C, D and E represent the classes for student grade (SG) level as being ‘exceptional’, ‘excellent’, ‘distinction’, ‘pass’ and ‘failure’. The class label represents in a form an expression:

$$SG \in \{A, B, C, D, E\} \tag{1}$$

TABLE 3. Confusion matrix for student grade prediction classification.

		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

The performance metrics of the confusion matrix is determined using accuracy, precision, recall and f-measure in the following equation:

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

where N is the number of samples

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

$$Recall (R) = \frac{1}{5} \left(\frac{AA}{AA + AB + AC + AD + AE} + \frac{BB}{BA + BB + BC + BD + BE} + \frac{CC}{CA + CB + CC + CD + CE} \right)$$

$$+ \frac{DD}{DA + BD + DC + DD + DE} + \frac{EE}{EA + EB + EC + ED + EE} \tag{4}$$

$$F - Measure = 2 \frac{PR}{P + R} \tag{5}$$

where the f-measure is weighted harmonic mean of precision and recall.

D. DATA VISUALIZATION

In this phase, after performed the data analysis, we extracted and visualized our findings to view the useful information and student grade performance trends in different courses using Python. Data visualization allows discovering all the features and insightful of the student dataset to help lecturers improve student academic performance for better decision making in the future. We also compare each the result of our proposed model in a better graphical approach to better understand the findings’ results.

IV. DESCRIPTIVE ANALYSIS OF STUDENT DATASET

Our dataset contains records of 641 students who taken two core courses namely Computer System Architecture (CSA) and Introduction to Computer System (ICS). Based on the analysis performed, we found 362 students obtained distinction grade (A–, B+, B) in CSA course, followed by the pass grade (B–, C+, C, C–, D+, D) with 176 students, the excellent grade (A) with 80 students, failed grade (E, E–, F) with 19 students and finally exceptional grade (A+) with 4 students. On the other hand, for the ICS course, the highest grades obtained by the students were in excellent grade (A) with 297 students, followed by distinction grade (A–, B+, B) with 273 students, exceptional 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. Correspondingly, we have investigated the mean and standard deviation of the final student grades for the CSA course were respectively 68.95 and 9.189, whereas for ICS course 79.62 and 7.379. Table 4 shows the number of students in both courses.

TABLE 4. Result of 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

Figure 3 shows the mean and standard deviation of students’ final marks and grade achievement according to the taken course. The students’ final marks were calculated based

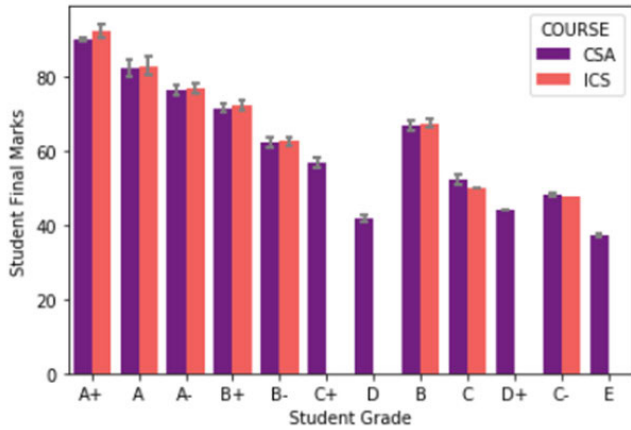


FIGURE 3. Mean and standard deviation of student's final marks against student's final grades achievement according to the taken courses.

on the total of percentage from continuous assessment marks evaluated during class and the final test marks in the final exam at the end of the semester. However, the students must earn more than 40 marks for both assessments in order to enable them to pass in both courses.

From the results, we recognize there is a difference in student achievement results between the CSA and ICS, where the students obtained higher marks better in ICS course compared to CSA. Figure 4 shows the normal trend of final marks distribution achieved by the students. Out of the total number of failure students, we found 3% of them are prominent in CSA compared to the ICS course. From these findings, we indicated that students who failed in both courses were not performed the minimum passing marks of the final examination, although their final marks classified as good and pass grades.

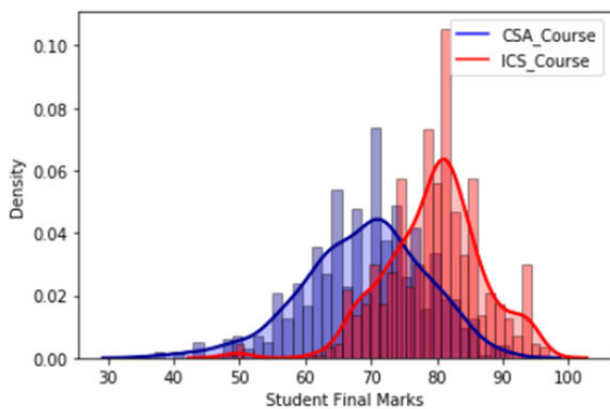


FIGURE 4. Graph plot of student's final marks distribution.

Furthermore, we also visualized the average grade point trend for ICS and CSA courses based on yearly achievement (2016 to 2019) as shown in Figure 5. From the observation, we found that the students' overall academic performance was improved yearly for both ICS and CSA courses. However, it is clearly shows that the grade point obtained from the

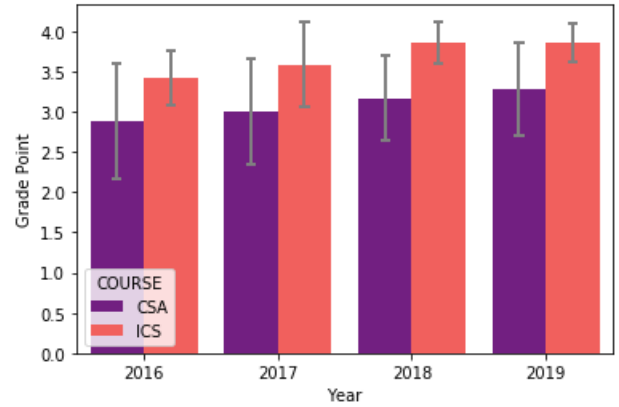


FIGURE 5. Analysis of average grade point trend for ICS and CSA courses by yearly basis.

ICS students is higher than the CSA. Therefore, from these findings, we indicated that CSA course is more challenging to those students who are weak in mathematics whereas the ICS course is more easy to understand for students who already have basic knowledge of computers before entering the polytechnic.

V. EXPERIMENTAL RESULTS

In this section, the results of this study are divided into two subsections according to research questions. We have conducted a comprehensive performance analysis with three experiments that run based on real dataset. The experiments' results of J48, kNN, NB, SVM, LR and RF were explored and compared. Then, we also compared and evaluate the impact of using oversampling SMOTE and FS methods in order to improve the imbalanced multi-classification problem with the same dataset.

A. RQ1: COMPARISON OF THE PREDICTIVE MODEL USING MACHINE LEARNING ALGORITHMS

Our main objective is to compare the predictive model based on the accuracy performance in this section. Here, six selected algorithms were used to train the student dataset and their prediction accuracy was evaluated. In order to analyze the differences, we compare the performance accuracy using the ten-fold cross-validation with stratification as a testing method to derive the best predictive model for optimal results. We measure the performance using various metrics including classification accuracy, precision, recall (Sensitivity) and f-measure to ensure the predictive model was fit to produce accurate results. Table 5 summarizes the prediction performance measures of different classifier on the student dataset.

It can be seen from Table 4 that the results indicated J48 and RF achieves the best prediction performance with precision score of 0.989 whereas followed by kNN with 0.985. Meanwhile, LR and SVM obtained precision 0.983 and 0.981 respectively. The lowest model is achieved by NB with 0.978. However, because of the classes in our dataset were highly imbalanced, the prediction results were often lead to misclassification decisions of the minority class that was

TABLE 5. Performance comparison of predictive models.

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

created while training the dataset. For generalizability purpose, another experiments in dealing with the issues were conducted to reduce the ratio of each classes which it is described in the next subsection.

B. RQ2: IMPACT OF OVERSAMPLING AND FEATURE SELECTION FOR IMBALANCED MULTI-CLASS DATASET

Here, we only focus on data-level solution using oversampling SMOTE and two FS algorithms for addressing imbalanced multi-classification dataset [27], [28]. To see the performance of each predictive model, we have performed three experiments on six selected machine learning algorithms to reduce the imbalanced problem. First, we performed SMOTE on our dataset with six selected machine learning algorithms independently. Secondly, the dataset was executed on two FS algorithms independently using three different attribute evaluators, and thirdly the proposed multiclass prediction model (SFS) was performed and tested using the same dataset in six selected machine learning algorithms. For a better view of the dimensionality prediction accuracy, other performance metrics on precision, recall and f-measure were used to ensure that our predictive model was fit to produce accurate results.

1) SMOTE OVERSAMPLING TECHNIQUE

SMOTE known as Synthetic Minority Oversampling Technique is the most commonly used to improve the overfitting problem based on random sampling algorithm [29]. It can modify an imbalanced dataset and generates new existing minority class instances by using synthetic sampling technique to create the distribution more balanced. This study was taking into consideration by increasing the default parameter of nearest neighbors (k) in sample SG in the minority class, select N samples randomly and record them as SG_i . The new sample SG_{new} is defined by the follows expression:

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

where $rand$ is a seed used of random sampling within range (0,1) and index of class value 0 with the ratio of generating new samples approximates 100%. In Weka, we implemented `weka.filters.supervised.instance.SMOTE` to insert synthetic instances between minority class samples of neighbors to our dataset. We set parameter of index class value 0 to

auto-detect the non-empty minority class. Then, the number of nearest neighbor' k value was set up to equal 10 ($k = 10$) with percentage of instances 100% and SMOTE filter was applied in ten times of iteration. The impact of oversampled dataset has increased the number of instances from 1282 up to 2932 where the SG class distribution using SMOTE becomes (504) exceptional, (377) excellent, (635) distinction, (744) pass and (672) fail by reducing the ratio to 1:1:2:2:2. In Table 6 we present the details comparison results of all predictive models with all performance measures. When the classifiers were used with oversampling SMOTE, we found that the effectiveness of all predictive models were consistently improved.

Among these predictive models, RF generated the most promising f-measure of 99.5%, whereas 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 confidence level of 95% using Paired T-Tester (corrected) as showed in Figure 6. We also observed when SMOTE method was applied, the minority class instance has increased to balance with other classes by number of iteration and number of k value to our dataset. The detailed analysis of the accuracy performance was presented based on confusion matrix as reported in Table 7.

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 *
	(w/ /*)	(0/0/1)	(0/1/0)	(0/0/1)	(0/0/1)	(0/0/1)

FIGURE 6. Result of predictive model performance with SMOTE.

It is obviously seen that confusion matrix of all predictive models derived from J48, NB, kNN, SVM, LR and RF shows improvement results of correctly classified for 'Pass' and 'Fail' grades.

However, there is small decrease performance from SVM where the predictive model correctly classified 97.2% of student who obtained 'Pass' grades compared to 99.5% when applied without SMOTE. For comparative analysis, Figures 7 and Figure 8 illustrate actual scores and predictions based on five categories of grade before and after applying the SMOTE respectively. Each predictive model performance shows the significant improvement for the majority classes except for minority class.

2) FEATURE SELECTION

Another experiment that we applied is feature selection (FS) which is effective in reducing dimensionality, removing irrelevant data and learning accuracy [30], [31]. In this experiment, two FS methods consist of wrapper and filter based were used as the benchmark methods to maximize the performance of six predictive models. The FS wrapper algorithm used to identify the best features set in this study consist of two attribute evaluator using J48 classifier; WrapperSubsetEval (FS-1) and ClassifierSubsetEval (FS-2) with

TABLE 6. Result of oversampling SMOTE with different predictive models.

Predictive Model	Oversampling	Accuracy	Precision	Recall	F-Measure
J48	None	0.989	0.989	0.990	0.989
	SMOTE	0.992	0.991	0.991	0.991
kNN	None	0.985	0.985	0.985	0.985
	SMOTE	0.993	0.993	0.993	0.993
NB	None	0.978	0.978	0.977	0.978
	SMOTE	0.983	0.983	0.983	0.983
SVM	None	0.984	0.981	0.984	0.979
	SMOTE	0.989	0.989	0.989	0.989
LR	None	0.984	0.983	0.984	0.983
	SMOTE	0.988	0.988	0.988	0.988
RF	None	0.989	0.989	0.989	0.989
	SMOTE	0.995	0.995	0.995	0.995

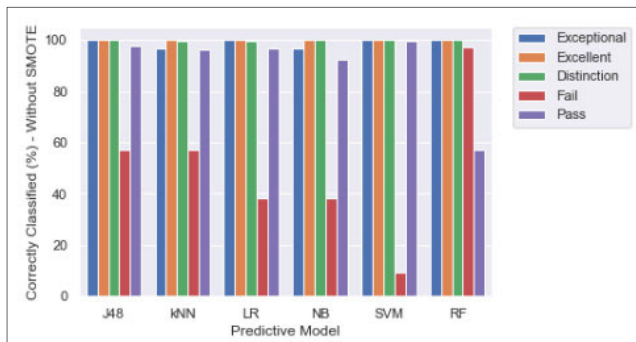


FIGURE 7. Comparison of correctly classified by class without applied SMOTE.

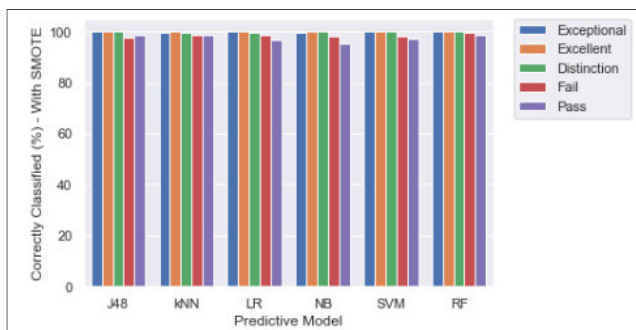


FIGURE 8. Comparison of correctly classified by class with applied SMOTE.

BestFirst search method. While for the FS filter algorithm, InfoGainAttributeEval (FS-3) with ranker search method more than 0.5 were selected as best feature set. The number of features in both FS algorithms are presented in Table 8. For the analysis, we have used the same dataset to find the best predictive model that fit with the requirements for giving an optimal result.

Table 9 shows overall results of different predictive model with all measurement of FS algorithms. The result showed

TABLE 7. Analysis of correctly classified based on confusion matrix.

Predicted Class	J48	NB	kNN	SVM	LR	RF	SMOTE
Exceptional	100	96.8	96.8	100	100	100	Before
Excellent	100	100	100	100	100	100	
Distinction	100	100	99.8	100	99.8	100	
Pass	97.8	92.5	96.2	99.5	96.8	97.3	
Fail	57.1	38.1	57.1	9.5	38.1	57.1	
Exceptional	100	99.6	99.8	100	100	100	After
Excellent	100	100	100	100	100	100	
Distinction	100	100	99.8	100	99.7	100	
Pass	98.7	95.3	98.8	97.2	96.9	99.5	
Fail	97.8	98.1	98.7	98.2	98.5	98.5	

All values are measure in percentage (%)

TABLE 8. Detailed selected features over different FS algorithms.

Main Attribute	Wrapper-based		Filter-based
	FS- 1	FS- 2	FS-3
StudID	/		/
Year		/	
Class	/	/	
Session			
Credit Hour			
Course Code			
Total Marks	/		/
Grade Pointer Average	/	/	/
Grade	/	/	/
Group (Class)	/	/	/
Total Features Selected	6	5	5

that kNN exhibited the highest performance f-measure score up to 98.8% and 98.9% with the optimal selected features set obtained from FS-2 and FS-3 algorithm respectively compared to others predictive models.

As we also can see from Table 9, NB shows the lowest performance of accuracy but the f-measure for NB shows slightly

TABLE 9. Classification performance of FS in different predictive models.

Predictive Model	Feature Selection	Accuracy	Precision	Recall	F-Measure
J48	None	0.989	0.989	0.99	0.989
	FS-1	0.984	0.983	0.984	0.983
	FS-2	0.984	0.984	0.984	0.984
	FS-3	0.987	0.985	0.987	0.986
kNN	None	0.985	0.985	0.985	0.985
	FS-1	0.988	0.988	0.988	0.988
	FS-2	0.989	0.988	0.989	0.988
	FS-3	0.989	0.989	0.989	0.989
NB	None	0.978	0.978	0.977	0.978
	FS-1	0.976	0.978	0.976	0.977
	FS-2	0.984	0.981	0.984	0.982
	FS-3	0.977	0.978	0.977	0.977
SVM	None	0.984	0.981	0.984	0.979
	FS-1	0.984	0.981	0.984	0.979
	FS-2	0.984	0.981	0.984	0.979
	FS-3	0.984	0.981	0.984	0.979
LR	None	0.984	0.983	0.984	0.983
	FS-1	0.984	0.982	0.984	0.983
	FS-2	0.986	0.984	0.986	0.983
	FS-3	0.983	0.978	0.983	0.980
RF	None	0.989	0.988	0.989	0.989
	FS-1	0.989	0.989	0.989	0.989
	FS-2	0.988	0.986	0.988	0.987
	FS-3	0.988	0.988	0.988	0.988

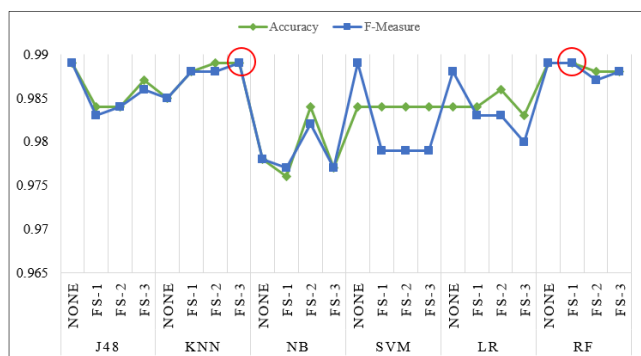


FIGURE 9. Classification performance of accuracy and f-measure with different FS.

improvement varied from 97.8% to 98.2% after FS-2 algorithm was undertaken. On the other hand, the performance of J48, LR, SVM and RF showed low promising performance when compared to without applied any FS. By reducing the number of features for the high imbalance ratio in multi-class dataset hinders the learning performance to predict student grade better. The comparison of the highest accuracy and f-measure score with different FS are highlighted in Figure 9.

3) PROPOSED MULTICLASS PREDICTION MODEL (SFS)

Then, we performed and tested the third experiments of the proposed SFS model by combining SMOTE oversampling and FS on the same dataset for pre-processing. The visualization of the comparison performance accuracy and f-measure rate of the proposed SFS model with all predictive models are presented in Figure 10. The highest score of accuracy and f-measure of each predictive models are presented in highlighted red. The results show that the proposed model with RF and J48 outperformed the highest f-measure performance up to 99.5% and 99.3% with SFS-1 algorithms, whereas kNN and SVM obtained the highest f-measure of 99.4% and 98.9% with SFS-2 algorithms respectively. LR and NB shares the result of f-measure up to 98.7% with SFS-2. The integration of the oversampling SMOTE and FS improves the performance of imbalanced multi-classification in our dataset where the oversampling SMOTE can balanced the selected features by increasing the number of features from minority class to equal with majority class. The details performance results of the proposed model SFS are presented in Table 10.

VI. DISCUSSION

This study was conducted to address the imbalanced multi-classification problems focus on data-level solution for

TABLE 10. Classification performance of the proposed SFS in different predictive models.

Predictive Model	SMOTE+ Feature Selection (SFS)	Accuracy	Precision	Recall	F-Measure
J48	None	0.989	0.989	0.990	0.989
	SFS-1	0.993	0.993	0.993	0.993
	SFS-2	0.993	0.993	0.993	0.993
	SFS-3	0.991	0.991	0.991	0.991
kNN	None	0.985	0.985	0.985	0.985
	SFS-1	0.993	0.993	0.993	0.993
	SFS-2	0.994	0.994	0.994	0.994
	SFS-3	0.993	0.993	0.993	0.993
NB	None	0.978	0.978	0.977	0.978
	SFS-1	0.975	0.976	0.975	0.975
	SFS-2	0.987	0.987	0.987	0.987
	SFS-3	0.958	0.959	0.958	0.958
SVM	None	0.984	0.981	0.984	0.979
	SFS-1	0.989	0.989	0.989	0.989
	SFS-2	0.989	0.989	0.989	0.989
	SFS-3	0.948	0.956	0.948	0.948
LR	None	0.984	0.983	0.984	0.983
	SFS-1	0.986	0.986	0.986	0.986
	SFS-2	0.987	0.987	0.987	0.987
	SFS-3	0.964	0.966	0.964	0.964
RF	None	0.989	0.988	0.989	0.989
	SFS-1	0.995	0.995	0.995	0.995
	SFS-2	0.994	0.994	0.994	0.994
	SFS-3	0.993	0.993	0.993	0.993

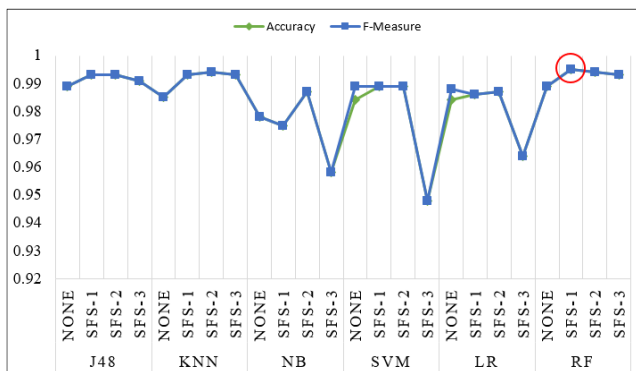


FIGURE 10. Comparison of accuracy and f-measure of proposed SFS model.

student grade prediction. For handling the imbalanced problem, we used the real student’s final course grades dataset from JTMK at one of Malaysia Polytechnics to analyze and compare the results of the proposed model. A similar study conducted in [7], [32], also mentioned the significant course grades can help in decision making in the educational domains. To answer our research question, we conducted a comprehensive experiment on real student dataset by

comparing the accuracy performance of the prediction model in a selected machine learning algorithm. Then, we also applied oversampling SMOTE and two FS methods to compare the effectiveness of the predictive model by using evaluation metrics of accuracy, precision, recall and f-measure to show the predictive models’ performance.

Overall results indicated that all predictive model derived from J48, NB, kNN SVM, LR and RF deliver a better performance when we applied SMOTE independently to the imbalanced dataset. However, after we applied FS method on imbalanced dataset using wrapper-based, only kNN and NB shows significant improvement whereas SVM remain same with none changes. This happened due to the tendency of overfitting and bias result caused by imbalanced data created when selecting the subgroup features. Other than that, we noticed the SVM not able to work independently in solving imbalanced multi-classification due to limitation for computing the best hyperplane for high dimensional imbalanced dataset [33]. As for NB, the used of FS for predicting student’s grade also supported in [30] where the author found NB shows the highest accuracy performance when wrapper-based subset feature selection was undertaken. However, we identified that FS independently not able to

improve the accuracy performance of RF that might be due to imbalanced dataset. Thus, we indicated FS enabled the predictive model to be interpreted more quickly, but the improvement was not depending on few features [34].

Then, we attempted to reduce the overfitting and misclassification of the minority class by combining SMOTE with a selection of appropriate features for all predictive models by introducing the SFS model. Here, the overall performance indicated the proposed SFS model outperformed with RF higher than previous study conducted by [12], [15]. The best accuracy obtained by the RF with 99.5% slightly higher than kNN and J48 shows that the RF algorithm was the ideal solution algorithm to predict student final grade. Meanwhile, kNN was the ideal solution that can work with the best value of k and optimal features [35]. The experiment results revealed that the proposed SFS model had more significant effect on kNN depending on the selected of FS algorithms. Certainly, these result also similar to the best performance of kNN in handling imbalanced data with different case studies as depicted in [36]. In this context, we also observe that most of the predictive models considered benefit when performing oversampling SMOTE but integrating the accurate features with different FS algorithms can influence the prediction effectiveness as well.

Despite these findings, we have identified several limitations to this fact; (1) the analysis is based on a defined dataset, but other dataset should be tested for data generalization that could affect the analysis results; (2) the analysis is only carried out with the certain well-known algorithms but can be analyzed with ensemble or advanced machine learning algorithms to compare the effectiveness for imbalanced multi-classification prediction model. (3) we used only one method of oversampling SMOTE, more method could be used to analyze whether they can improve the multi-class imbalanced problem.

Therefore, this study still needs to be improved in predicting students' final grades by improving the sampling techniques for imbalanced multi-class dataset that might affect the accurate prediction results. In addition, we also be considered to use SVM ensemble to be as part of the analysis since it has produced greater accuracy when predicting students' final grades as mentioned in [37].

VII. CONCLUSION AND FUTURE DIRECTIONS

Predicting student grades is one of the key performance indicators that can help educators monitor their academic performance. Therefore, it is important to have a predictive model that can reduce the level of uncertainty in the outcome for an imbalanced dataset. This paper proposes a multiclass prediction model with six predictive models to predict final student's grades based on the previous student final examination result of the first-semester course. Specifically, we have done a comparative analysis of combining oversampling SMOTE with different FS methods to evaluate the performance accuracy of student grade prediction. We also have shown that the explored oversampling

SMOTE is overall improved consistently than using FS alone with all predictive models. However, our proposed multiclass prediction model performed more effectively than using oversampling SMOTE and FS alone with some parameter settings that can influence the performance accuracy of all predictive models. Here, our findings contribute to be a practical approach for addressing the imbalanced multi-classification based on the data-level solution for student grade prediction.

In HEI, predictive analytics plays a significant role in governance for improving valuable information and developing trusted decision-making that contributes to data science [38]. Determining the quality of the collected dataset to reduce the imbalance and missing values difficulties is part of the challenging issues that adhere to select the relevant and valuable predictive models [39]. Therefore, as for future works, further investigation on the use of appropriate emerging predictive techniques in such advanced machine learning algorithms [40] and more ensemble algorithms are recommended to optimize the result for predicting student grades. It is also essential to select several multi-class imbalanced datasets to be analyzed with appropriate sampling techniques and different evaluation metrics which suitable for the imbalanced multi-class domain such as Kappa, Weighted Accuracy and other measures. Thus, using machine learning in higher learning institutions for student grade prediction will ultimately enhance the decision support system to improve their student academic performance in the future.

ACKNOWLEDGMENT

The authors are grateful for the support of Student Sebastien Mambou in consultations regarding application aspects.

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. Düşteğör, "Predicting academic success in higher education: Literature review and best practices," *Int. J. Educ. Technol. Higher Educ.*, vol. 17, no. 1, Dec. 2020.
- [3] 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.
- [5] A. E. Tatar and D. Düşteğör, "Prediction of academic performance at undergraduate graduation: Course grades or grade point average?" *Appl. Sci.*, vol. 10, no. 14, pp. 1–15, 2020.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] A. Hellas, P. Ihanntola, 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.
- [10] 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](https://doi.org/10.1186/s41239-019-0160-3).

- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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>
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [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.
- [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, pp. 54–60, 2019.
- [22] D. Berrar, "Cross-validation," *Comput. Biol.*, vols. 1–3, pp. 542–545, Jan. 2018, doi: [10.1016/B978-0-12-809633-8.20349-X](https://doi.org/10.1016/B978-0-12-809633-8.20349-X).
- [23] 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.
- [24] B. Predić, G. Dimić, D. Rančić, P. Šurbac, 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](https://doi.org/10.1002/cae.22042).
- [25] 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.
- [26] L. E. O. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
- [27] 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.
- [28] 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.
- [29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
- [30] C. Jalota and R. Agrawal, *Feature Selection Algorithms and Student Academic Performance: A Study*, vol. 1165. Singapore: Springer, 2021.
- [31] G. A. Sharifai and Z. Zainol, "Feature selection for high-dimensional and correlation based redundancy and binary," *Genes*, vol. 11, pp. 1–26, Jun. 2020.
- [32] 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.
- [33] S. Chinna Gopi, B. Suvama, and T. Maruthi Padmaja, "High dimensional unbalanced data classification Vs SVM feature selection," *Indian J. Sci. Technol.*, vol. 9, no. 30, Aug. 2016.
- [34] 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, X. Zhu, 2020.
- [35] S. Zhang, X. Li, M. Zong, X. Zhou, and D. Cheng, "Learning k for kNN Classification," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, pp. 1–9, 2017.
- [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.
- [37] Brodic, A. Amelio, and R. Jankovic, "Comparison of different classification techniques in predicting a university course final grade," in *Proc. 41st Int. Conv. Inf. Commun. Technol. Electron. Microelectron.*, 2018, pp. 1382–1387.
- [38] 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.
- [40] 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.



SITI DIANAH ABDUL BUJANG received the B.S. degree in science (computer science) and the M.S. degree in science from Universiti Teknologi Malaysia (UTM), in 2006 and 2010, respectively. She is currently pursuing the Ph.D. degree in software engineering with the Malaysia-Japan International Institute of Technology, UTM, Kuala Lumpur. Her thesis focuses on the application of predictive analytics on student grade prediction in a higher education institution. From 2010 to 2019, she was a Senior Lecturer of Information and Communication Technology Department, Polytechnic Sultan Idris Shah, Sabak, Selangor, Malaysia. She has experience in developing the polytechnic curriculum for Diploma in information technology (technology digital), 2.5 years' program. She is one of the book authors that contribute for the Department of Polytechnic and Community College Education. Her research interests include data analytics, predictive analytics, learning analytics, educational data mining, and machine learning.



ALI SELAMAT (Member, IEEE) has been the Dean of the Malaysia-Japan International Institute of Technology (MJIT), an academic institution established under the cooperation of the Japanese International Cooperation Agency (JICA) and the Ministry of Education Malaysia (MOE) to provide the Japanese style of education in Malaysia, Universiti Teknologi Malaysia (UTM), Malaysia, since 2018. He is currently a Full Professor with UTM, where he is also a Professor with the Software Engineering Department, Faculty of Computing. He has published more than 60 IF research articles. His H-index is 20, and his number of citations in WoS is more than 800. His research interests include software engineering, software process improvement, software agents, web engineering, information retrievals, pattern recognition, genetic algorithms, neural networks, soft computing, computational collective intelligence, strategic management, key performance indicator, and knowledge management. He is on the Editorial Board of the journal *Knowledge-Based Systems* (Elsevier). He has been serving as the Chair of the IEEE Computer Society Malaysia, since 2018.



ROLIANA IBRAHIM (Member, IEEE) received the B.Sc. degree (Hons.) in computer studies from Liverpool John Moores University, the M.Sc. degree in computer science from Universiti Teknologi Malaysia (UTM), and the Ph.D. degree in systems engineering from Loughborough University. She is currently the Director of applied computing at the School of Computing, formerly known as the Faculty of Computing. Previously, she was the Head of the Information Systems Department, for three years, at the Faculty of Computing, UTM. She has been an Academic Staff at the Information Systems Department, since 1999. She was previously a Coordinator of the B.Sc. Computer Science (Bioinformatics) Program and the Master of Information Technology (IT Management).



ONDREJ KREJCAR received the Ph.D. degree in technical cybernetics from the Technical University of Ostrava, Czech Republic, in 2008, and the Ph.D. degree in applied informatics from the University of Hradec Kralove. He is focusing on lecturing on smart approaches to the development of information systems and applications in ubiquitous computing environments with the University of Hradec Kralove. He is a Full Professor of systems engineering and informatics at the Center

for Basic and Applied Research, Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic, and a Research Fellow at the Malaysia-Japan International Institute of Technology, University Technology Malaysia, Kuala Lumpur, Malaysia. From 2016 to 2020, he was the Vice-Dean of Science and Research at the Faculty of Informatics and Management, UHK. He has been the Vice-Rector of science and creative activities at the University of Hradec Kralove, since June 2020. He is also currently the Director of the Center for Basic and Applied Research, University of Hradec Kralove. His H-index is 20, with more than 1300 citations received in the Web of Science, where more than 100 IF journal articles is indexed in JCR index. In 2018, he was the 14th Top Peer Reviewer in multidisciplinary in the world according to Publons and a Top Reviewer in the Global Peer Review Awards 2019 by Publons. His research interests include control systems, smart sensors, ubiquitous computing, manufacturing, wireless technology, portable devices, biomedicine, image segmentation and recognition, biometrics, technical cybernetics, and ubiquitous computing. His second area of research interests include biomedicine (image analysis), biotelemetric system architecture (portable device architecture and wireless biosensors), and development of applications for mobile devices with use of remote or embedded biomedical sensors. He is currently on the Editorial Board of the *Sensors* (MDPI) IF journal (Q1/Q2 at JCR) and several other ESCI indexed journals. He has been the Vice-Leader and Management Committee Member at WG4 at Project COST CA17136, since 2018. He has also been a Management Committee Member Substitute at Project COST CA16226, since 2017. Since 2019, he has been the Chairman of the Program Committee of the KAPPA Program, Technological Agency of the Czech Republic, and was a Regulator of the EEA/Norwegian Financial Mechanism in the Czech Republic, from 2019 to 2024. Since 2020, he has been the Chairman of the Panel 1 (computer, physical and chemical sciences) of the ZETA Program, Technological Agency of the Czech Republic. From 2014 to 2019, he was the Deputy Chairman of the Panel 7 (processing industry, robotics, and electrical engineering) of the Epsilon Program, Technological Agency of the Czech Republic.



ENRIQUE HERRERA-VIEDMA (Fellow, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Granada, Spain, in 1993 and 1996, respectively. He is currently a Professor of computer science AI and the Vice-President of research and knowledge transfer, University of Granada. He has authored or coauthored more than 300 articles in JCR journals. In 2013, he has published in the prestigious journal science about the new role of digital libraries in

the era of the information society. He was the Vice-President for Publications with the IEEE System Man and Cybernetics Society from 2019 to 2020. He is currently the VP of cybernetics, the Founder of the IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, and a Highly Cited Researcher by Clarivate Analytics in computer science and engineering from 2014 to 2020. His H-index is 101 in Google Scholar (more than 33000 citations) and 85 in WoS (more than 23000 citations). He is also an Associate Editor of several

AI journals like IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, *Knosys*, *Applied Soft Computing*, *Fuzzy optimization and Decision Making*, *Information Sciences*, and *Soft Computing*. He has also been a Guest Lecturer in plenary lectures and tutorials in multiple national and international conferences related to artificial intelligence.



HAMIDO FUJITA (Life Senior Member, IEEE) received the title of Honorary Professor and the Doctor Honoris Causa degree from Óbuda University, Budapest, Hungary, in 2011 and 2013, respectively, and the Doctor Honoris Causa degree from Timisoara Technical University, Timișoara, Romania, in 2018. He is an Emeritus Professor with Iwate Prefectural University, Takizawa, Japan. He is currently the Executive Chairman of i-SOMET Incorporated Association,

www.i-SOMET.org, Morioka, Japan. He is a Distinguished Research Professor at the University of Granada and an Adjunct Professor with Stockholm University, Stockholm, Sweden; the University of Technology Sydney, Ultimo, NSW, Australia; National Taiwan Ocean University, Keelung, Taiwan, and others. He has supervised Ph.D. students jointly with the University of Laval, Quebec City, QC, Canada; the University of Technology Sydney; Oregon State University, Corvallis, OR, USA; the University of Paris 1 Pantheon-Sorbonne, Paris, France; and the University of Genoa, Italy. He has four international patents in software systems and several research projects with Japanese industry and partners. He was a recipient of the Honorary Scholar Award from the University of Technology Sydney, in 2012. He was a Highly Cited Researcher in cross-field for the year 2019 and in computer science field for the year 2020 by Clarivate Analytics. He headed a number of projects, including intelligent HCI, a project related to mental cloning for healthcare systems as an intelligent user interface between human-users and computers, and a SCOPE Project on virtual doctor systems for medical applications. He collaborated with several research projects in Europe. He is recently collaborating in OLIMPIA Project supported by Tuscany Region on therapeutic monitoring of Parkinson's disease. He has published more than 400 highly cited articles. He is the Emeritus Editor-in-Chief of *Knowledge-Based Systems* and currently the Editor-in-Chief of *Applied Intelligence* (Springer).



NOR AZURA MD. GHANI (Member, IEEE) is a Professor with the Center for Statistical Studies and Decision Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia, and the Chair of IEEE Computer Society Malaysia Chapter. She currently serves as the Deputy Director (Research Impact) at the Research Management Center, Universiti Teknologi MARA, Malaysia. Her expertise is big data, image processing, artificial neural networks,

statistical pattern recognition, and forensic statistics. She is the author or coauthor of many journals and conference proceedings at national and international levels.

• • •