JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3031-3041
https://publishoa.com
ISSN: 1309-3452

Using Traditional Machine Learning Algorithms and SMOTE Technique to Estimate Student's Academic Performance in Higher Education

E. Sandhya[1], Dr. SK Althaf Hussain Basha[2], E. S. Phalguna Krishna[3], Dr. V.Jyothsna[4], C.Silpa[5]

[1,5]Assistant Professor, Department of IT, Sree Vidyanikethan Engineering College, Tirupati
[2] Professor, Department of CSE, Krishna Chaitanya Institute of Technology and Sciences, Markapur
[3]Assistant Professor, Department of CSE, Sree Vidyanikethan Engineering College, Tirupati
[4]Associate Professor, Department of IT, Sree Vidyanikethan Engineering College, Tirupati

Received 2022 April 02**; Revised** 2022 May 20**; Accepted** 2022 June 18**.**

Abstract

Predictive analysis applications have become a revolutionary topic in higher education in the following days. Predictive analysis will use analytical models, which include machine learning applications, to help in producing the greater quality performance and useful student data at different stages of education. Most people know that a student's grade is the most important performance indicator that teachers can use to track their academic progress. Many learning algorithms have been proposed in the education field over the past decade. However, dealing with unequal databases in order to improve the efficiency of predicting student marks causes significant difficulties. Random Forest (RF), Naive Bayes (NB), Decision Tree (J48), K-Nearest Neighbour (kNN), Logistic Regression (LR), Support vector Machine (SVM), and a hybrid model that combines Random Forest and the XGBoost algorithm are all evaluated for accuracy. By using feature selection and Synthetic Minority Oversampling Technique (SMOTE) a multi-stage prediction model is proposed to reduce the effects of overlap and misalignment caused by multiple class inequalities. The SMOTE uses the random sampling method and in the feature selection the wrapper and filter methods are used. The proposed model produces promising and comparable results, which are used to develop a performance model for predicting the unequal distribution.

Keywords: Student Grade Prediction Dataset, Machine Learning, Random Sampling, SMOTE, Multi Class Classification

1. Introduction

Every Higher Education Institutions (HEI) will keep track of each student's academic records. The student record contains the academic result of final marks of exams and grades of various programs and courses. Every recorded student mark and grade is utilised to calculate the student's performance and evaluate the semester's course completion. The records are used to generate useful information about students' performance [8]. And not only the academic background but also other factors like family, socioeconomic, demographic etc will also contribute to the student performance. So here the prediction of student grade will depend on different factors and the prediction is made.

Predictive analytics has become one of the most potential approaches in every sector. When it comes to the educational sector predictive analytics will help in finding out the hidden patterns and prediction making even with the vast database. It also helps in solving different problems in the education field with program selection, dropout prediction, and semester early warning system. Furthermore, the use of predictive analytics-based solutions in the education sector has grown in popularity over time. [1].

One of the most essential aspects that will aid in enhancing student academic performance is the potential to estimate a student's grade. There are many previous studies which have found different machine learning methods[16] which are used for the prediction of student learning performance. However, the connection applied to the machine for this to

improve the imbalance of multi-phase drawback in predicting grade-level prediction is difficult for finding. As a result, in this model, a comparative study was performed to determine which predictor model is the best for predicting student's grade. [4]. Here, the model addresses the problems and questions like which predictive model will give higher accuracy among the taken machine learning algorithms in predicting the student's final grade. Also, how an unbalanced dataset will be balanced using machine learning approaches such as oversampling SMOTE and Feature Selection algorithms.

Introduce a descriptive analysis of the student dataset to depict student grade patterns, which can lead to strategic planning decisions that will help teachers assist students more effectively. [5]. Following that, the comparative analysis was used to compare seven machine learning algorithms, including J48, NB, LR, k-NN, XG-Boost, SVM and RF. With regard to addressing imbalanced multi-classification a method for each model speculative with data levels solutions using a FS [13] and SMOTE sample.

Today any machine learning specialist who specializes in binary separation problems has to deal with this common unequal data set. Class inequality is a rising situation in which there is an unequal distribution of class in a database i.e., a number. Data points in the negative category (majority category) are much larger compared to the positive category (sub-category) [9]. Generally, a minority / positive class is an interesting category and aims to achieve the best results in this class than that. If the inequality data is not treated earlier, this will reduce the effectiveness of the partition model. SMOTE is an over-sampling method where synthetic samples are made in the junior class. This technique works on overcoming the problem of overfitting which is caused by random sampling. [10]. It focuses on the feature of creating new situations with the help of adding among the best living conditions together.

Feature Selection is a way to minimize different inputs to your model by taking the necessary data and eliminating audio from the input data. It's a method of selecting relevant characteristics of your machine learning model automatically based on the type of problem you're trying to answer [1]. Here by adding or excluding key features without modifying them. It helps to reduce noise in the data and reduce the weight of our input data. Features are the variable input we supply to our machine learning models. A feature is created by each column in our database. To train the right model, it must make sure that only the most important features are used [11]. If we have too many features, the model will pick up on insignificant patterns and learn from them. Feature Selection is a strategy for selecting the most essential parameters in our data.

The proposed work is combination of SMOTE+FS (SFS) algorithms, calculates the ratio of sampling by selecting best features and improving multi-classification that is unbalanced in predicting the grade of student. So first the dataset is extracted and cleaned. And the dataset is also pre-processed using the S FS technique. The multiclass model is then developed, and the model is then trained and evaluated on the dataset before being finalised [3].

The following is the report's structure. The linked research endeavour, which is a literature survey for student grade prediction, is described in Section II. Section III demonstrates the suggested predictive model methodology for predicting final student grades by phases. The results are presented in Section IV. The conclusions and future directions are addressed in Section V.

2. Literature Survey

In distinctive fields, such as education, using system getting to know has resulted in widespread improvements in predicting responsibilities. Prediction models not only are simple to comprehend but also offer valuable insights in decision-making. Therefore, an approach is proposed to improve the predicting student performance by improvement in performance and explains how the student is able to achieve a certain score with this working. Machine learning methods were used to develop and test a prediction model. A technique called Synthetic minority oversampling technique (SMOTE) is used for balancing to oversample all the majority classes. It also uses ensemble methods [6] [15]. Coming to tune the parameters, it estimates the ideal parameters using a basic grid search method given by scikit.

The innovative model's dependability is demonstrated through hyperparameter optimization and a ten-fold cross-validation [2] approach. In addition, a novel visual and intuitive technique is used to help determine which factors have the biggest influence on the score. In educational institutions, student success is critical. Early discovery of pupils who

are performing poorly, as well as preventative measures, can aid in their improvement [13]. To produce predictions, machine learning techniques have been extensively used. Indeed, the effective and efficient use of data mining technologies, through which student characteristics to focus on ranging from how to define student achievement to which machine learning method is better appropriate for the specific difficulty. It assists educators in providing a step-by-step set of instructions as well as the use of data mining. So, first, the data must be collected and cleaned, and then the data must be transformed [7]. After that, feature selection is completed, and data mining models are implemented. The final outcome will be divided into two categories: successful and failed. It will make data mining techniques more accessible to teachers, allowing them to maximise their potential.

The data includes demographics, previous academic records, and family background information for students. To build the most accurate academic achievement prediction model possible, students' data is submitted to Decision Tree, Naive Bayes, and Rule Based classification techniques. The Rule Based strategy [8] receives the best accuracy value when compared to the other techniques, according to the results of the trial. Due to partial and missing values in the obtained data, the limitation is the tiny amount of the data. The gathered knowledge from the prediction model will aid in identifying and profiling pupils in order to estimate their success level.

Students' grades in such registered courses can be predicted using machine learning techniques. Such strategies would not only assist students in improving their performance based on expected grades, but they would also allow instructors to identify students who may require assistance in their enrolled courses. Restricted Boltzmann Machines, Matrix Factorization [3], and Collaborative Filtering will be in this case. The methodology is limited to small datasets, but it may be extended to huge datasets in the future. The RBM technique is proven to be superior to other techniques in terms of accuracy in forecasting students' success in a particular course.

In the subject of data mining, educational data mining is a relatively new phenomenon. The Optimal Equal Width Binning normalisation approach and SMOTE over-sampling strategy are used in this study to illustrate how data will be pre-processed to increase the accuracy of the students' final grade prediction model [7] for a specific course. Naive Bayes classification is used for individual classification of the attributes independent of each other. Decision tree classification is used to derive the rules, which are required for the prediction of student’s grade. The result obtained after discretization and applying over-sampling methods, gives clear evidence that the prediction model's accuracy is greatly improving.

The evaluation and prediction of college students' progress is at the centre of university student management. The traditional evaluation system focuses primarily on assessing students' past accomplishments, but it falls short of predicting students' future growth. As a result, classification models such as Multilayer Perceptron, Decision Tree, Naive Bayes [15], and Support Vector are employed to predict students' academic achievement. The multi-layer perceptron model has shown to be the most effective, with higher accuracy in both the training and test sets. The experimental results will be more accurate if the data sample is large enough.

Predicting a student’s grade is very important to know the performance of a student. By doing the prediction of a student's final grade in advance, it helps the student to take the needed steps to get best performance every time. So here the final grade of a student is anticipated based on their previous academic performance. The university's strategic method focuses offering high-quality education that is in line with the Sustainable Development Goals (SDGs) of the United Nations [13]. One of the most common applications of educational data mining is using previous grades to forecast a student's performance. In the first stage, a technique will be proposed for data collecting, data cleaning, and pre-processing, and in the second stage, students with comparable academic performance patterns will be grouped [7]. The supervised machine learning method will then be chosen based on the gathered and detected patterns, and the experimental phase will begin. Finally, the data has been collected and will be analysed. One of the most pressing future tasks is to analyse and create a big data infrastructure.

As demands to enhance student performance and the urge to help increase, it's harder to spot students who will perform poorly in a class. Traditional grading relied on intuition and complex averaging of exam data. With learning devices and data mining, we can use student performance data to make statistically sound and justified predictions beyond a teacher's intuition. It investigates the effects of a "basic average" method of forecasting student grades as a substitute for

a teacher's prediction, as well as the utility of three algorithms in predicting students' grades—Naive Bayes prediction, k nearest neighbours, and guiding vector system. It demonstrates that an assist vector device outperforms all other Predictors, including the basic average [10]. Even if machine learning is the primary strategy for obtaining more effective results, it requires the addition of additional strategies.

Data mining is the process of extracting previously unknown information from data in a non-trivial manner. Category is a supervised studying Approach in data mining. This is used for expecting a Specific unit label of a facts instance; therefore one can classify it into one of the predefined lessons. Here we are able to bear in mind magnificence Imbalance hassle [2] and solutions for dealing with class imbalance Hassle. Magnificence Imbalance problem happens where in comparison to the other elegances, this dataset has a fairly limited number of samples [5]. Here it uses a financial institution marketing Dataset for our evaluation that could be an elegance imbalance dataset.

Educational Data Mining (EDM) has emerged during the last two decades. The main goal of this investigation is to look at how Advanced Capacity is used in this setting machine learning techniques on educational settings from the attitude of hyperparameter Optimization. By the effectiveness of automated gadgets gaining knowledge of (auto ML) for the task of predicting students' studying outcomes based on their participation in online Mastering structures. At the same time, we restrict the hunt space to tree-based totally and rule-based models with a view to accomplishing transparent and interpretable consequences. The auto ML tools achieve consistent results. EDM [4] conducts experiments with suitable automated parameter configurations, for that reason accomplishing exceptionally accurate and understandable outcomes. Furthermore, the proposed approach may additionally function as a good size useful resource in the early estimation of college students Performance, and as a result permitting timely guidance and powerful intervention techniques [10]. Appropriate Software extensions inside learning management systems could be constructed, to allow non-expert users advantage from autoML

This article discusses the comparison of prediction classifiers in a combined Mastering environment. The proposed approach predicts students' final grades entirely based on sports in unusual academic situations. A comparison of classifier performance was carried out with the goal of determining which classifier is best suited for multiclass feature datasets [19]. The majority of people's voting method was used to form an ensemble based exclusively on Hidden Naive Bayes [6], Naive Bayes, Random Forest and J48 decision tree, which gave crucial results for specialised training. It calculates the accuracy and precision of the scholars' predication of grades in mixed learning of environment state of affairs based on experimental evaluation. The key contribution is a multi-class prediction system that works well.

2.1 Open issues

- The main goal was to improve the models they built by pre-processing them and then determining which model had the best accuracy. Oversampling was inevitable due to the minimal number of cases in the sample.
- The errors cannot be detected early. As a result, there will be less accuracy rate. When a dataset with a large number of records is used, the predicted result will have less accuracy.
- Models based on tensor factorization are not being studied to account for the temporal effect in the prediction of student performance.
- The model doesn't explain why a student is prone to fail and can't help to better align the learning activities and support with the student’s needs.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3031-3041
https://publishoa.com
ISSN: 1309-3452

3. Proposed Methodology

A model is built which helps to determine the best effective predictive model for dealing with unbalanced multi-classification in predication of student grade. Our model's input contains many data such as the student's name, course name, previous year grades, personal information, social information, and so on.

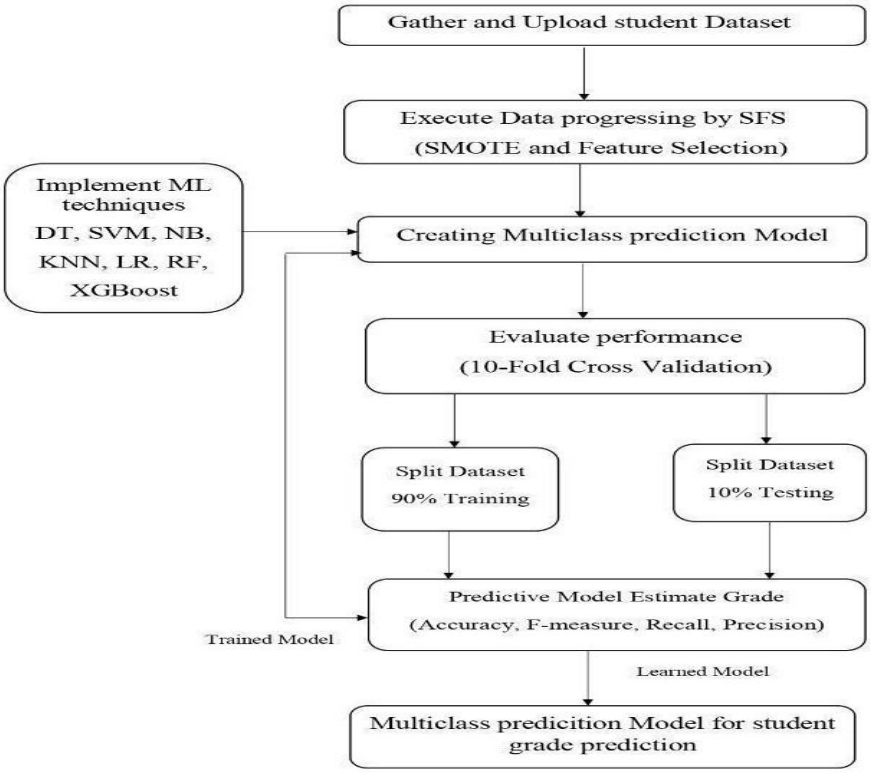

Fig 1: Proposed Framework for Multi Class-prediction Model

To reduce overfitting and misclassification, we employed oversampling SMOTE and Feature Selection (FS) [12]. To assess the student's success, combine both strategies into a chosen machine learning classifier to build the recommended model. The framework is depicted in Figure 1.

3.1 Data Preparation

From Kaggle website with Student Grade Prediction dataset is considered for predicting grade of high school students [17]. The dataset contains 33 attributes. The attributes include student grades, features related to school, social, demographic. The grades are related to two distinct subjects. The data was classified using a binary/five-level classification system. And the output attribute which is G3 has a strong correlation with attribute G1 and G2. Binary classification is used for the attributes like gender (F or M), Choosing the school between the given two schools, city (Urban or Rural) etc. And the five-level classification is used for choosing the one value among five different values. Some of the attributes of the dataset are school, sex, address, age, family size, travel time, study time, free time etc. The attributes take the numeric values, binary values, and nominal values based on the attribute type.

3.2 Data Pre-processing

The dataset which is collected in the previous stage undergoes the data pre-processing. Here the output class is divided into five ranks namely P, Q, R, S, and T. If the dataset is imbalanced, then the results are inefficient. If the results are inefficient then there is no use of the predicted results. So, the dataset needs to be pre-processed to get the accurate results. The data pre-process in this study mainly involves two techniques which are feature selection and oversampling technique.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3031-3041
https://publishoa.com
ISSN: 1309-3452

a. Oversampling Technique

The random sampling technique is used in SMOTE, which stands for Synthetic Minority Oversampling Approach. It is the most often used oversampling technique for addressing the overfitting problem. Today any machine learning specialist who specializes in binary separation problems has to deal with this common imbalanced dataset. Class imbalance is a condition that occurs when it has an unequal distribution of class in a database that is a number. Data points in the majority category are very large compared to the minority category. SMOTE is an over-sampling method where synthetic samples are made in the class which contribute less [8]. This approach helps to solve the problem of random overloading causing overfilling. It focuses on the space feature of creating new situations with the help of adding among the best living conditions together. So, this is how the oversampling technique is done.

b. Feature Selection Method

Features refer to the variable input that machine learning models receive. A feature is created by each column in our database. It is necessary to ensure that just the most important features are included when training the proper model. If there are too many characteristics, the model will pick up on insignificant patterns and learn from them. Feature selection is the process of selecting the most essential parameters in our data. Using just relevant data and removing sounds from the data, feature selection is a strategy of minimising input variances in your model [11]. It's a way for automatically identifying relevant machine learning model attributes based on the sort of problem you're trying to solve. solve. This is accomplished by simply adding or removing key features without altering them. It assists us in reducing both the amount of noise in our data and the size of our input data. Wrapper and filter methods are two different types of feature selection approaches.

i. Wrapper Method

Depending on their relationship to the output, or how they relate to the output, features are eliminated in this manner. It checks whether features are positively or adversely connected to output labels using correlation and discards features accordingly. If the feature has no much relation with the output, then the feature is dropped. And also, if the attribute has a negative impact on the output and if it provides very less contribution then the feature is removed.

ii. Filter Method

The dataset is broken into smaller sets, which are then used to train the model. The smaller sets can add and remove features and train the model based on the model's output. Using a greedy method, create subsets and check the correctness of all conceivable feature combinations. So here the feature can either be added or removed. The features are made into subsets and then the subset is trained and the accuracy is calculated [13]. Finally based on the results it is decided whether the feature needs to be added or removed.

3.3 Performance Analysis

The Performance Analysis which is a very important factor to know the standards of the model build. So, when the performance is evaluated, it talks about the model ability. In the paper the student grade prediction is made and building the multi-class classification model. The model is developed using several machine learning approaches, and its performance is assessed using 10-fold cross validation [7], in which the dataset is divided into two parts: train and test datasets. The training dataset makes up 90% of the overall dataset, whereas the test dataset makes up 10%. The model is trained using the training dataset, which accounts for 90% of the total dataset, and the results are assessed. If the results are with more accuracy, then the model works on a test dataset.

a. Logistic Regression

The supervised learning categorization class approach of logistic regression is used to anticipate the likelihood of specified diversity. The target or dependent variable is dichotomous in nature, which means there are only two possible values.. There are two possible values of 0 (success or yes) and 1 (failure or no). The stage where the dependent variant may have 3 or more ordered types that may be 0, 1, 2..or bad, good, very good, very good, and then use the ordinal type of regression.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3031-3041
https://publishoa.com
ISSN: 1309-3452

b. Naive Bayes

The Naive Bayes algorithm is a supervised learning technique that is used to handle classification problems. It's commonly used to categorise text, including high-quality training databases. One of the most fundamental and successful classification methods is the Naive Bayes Classifier. It helps to construct faster machine learning models that can make better guess. It is a probabilistic classification system; therefore it generates predictions based on the likelihood that an object exists.

c. Decision Tree

The Decision Tree is a supervised learning method for classifying and predicting problems. Classification problems are its specialty. Core nodes represent database elements, branches represent decision rules, and leaf nodes reflect the conclusion. Decision Tree has two nodes: Decision and Leaf. Decision Nodes are used to make decisions and have many branches, while Leaf Nodes are the results. Information is utilised to make choices or take examinations. CART algorithm is used to build a tree. The decision tree asks a question and creates subtrees based on the answer (Yes/No).

d. Support Vector Machine

SVM is a prominent supervised learning technique for classification and regression. Machine Learning uses it to categorise. The SVM technique generates a superior line or decision line that divides n-dimensional space into classes with similar qualities, making it easier to place data points in the correct category. Hyperplanes are useful for decision-making. SVM chooses the hyperplane's extreme points/vectors.

e. K-Nearest Neighbour

K-Nearest Neighbours is a supervised machine learning methodology. The KNN algorithm compares new cases/data to old cases and assigns new examples to existing categories. KNN tracks existing data and sorts new data based on similarities. The KNN algorithm can quickly segment fresh data, independent of its source. KNN can be used for regression and classification, however classification is more prevalent. KNN is non-parametric; hence it makes no data assumptions.

f. Random Forest

Random Forest is a popular machine learning algorithm that is used to solve classification and regression problems. It's a part of a supervised learning approach. Random forest is based on ensemble learning, in which several divisions are combined to tackle complex problems and improve model performance [18]. According to the title, "The Random Forest is a subcategory that includes a number of decision trees for the different datasets offered and takes methods to improve the database's prediction accuracy." The random forest accepts a forecast from each tree and forecasts the eventual result based on a number of reliable votes rather than depending on a single decision tree. The forest's large number of trees allows for great accuracy while also reducing congestion.

g. XG-Boost Algorithm

The Extreme Gradient Boosting (XG-Boost) machine learning library is a modular, widely disseminated gradient-boosted gradient (GBDT) machine learning framework. It is a powerful collection of machine learning, reconditioning, and level problems that allows for simultaneous tree creation. To begin grasping machine learning ideas and techniques that XG-Boost draws on, you must first understand XG-Boost: supervised machine knowledge, decision trees, cooperative learning, and gradient building. Supervised machine learning use algorithms to train the model to detect patterns in a database of labels and features, and then employs the learned method to forecast labelling in new system characteristics.

h. Hybrid Method

In the hybrid method different algorithms are combined and then worked together to produce an output. In this model by combining different algorithms the accuracy rate may be increased and works effectively. The methods combined

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3031-3041
https://publishoa.com
ISSN: 1309-3452

are random forest and XGBoost algorithm. It could have blended approaches from other domains or employed methods that were a mixture of current ones. Before giving our data to an ML technique, it may use data transformation methods like statistical methods or simple linear correlation coefficients. Hybrid method techniques are based on a machine learning architecture that differs slightly from the standard workflow.

3.4 Confusion Matrix

The confusion matrix is used to calculate the pre-processed dataset's values such as Recall, Accuracy, F-Measure, and Precision. The confusion matrix is used to assess the prediction model's performance. The student grade prediction parameters for confusion matrix are P, Q, R, S, T. It means that level of grades (LG) are 'exceptional', 'excellent', 'distinction', 'pass' and 'failure'. The following expression represents class label: LG ∈ {P, Q, R, S, T}.

Confusion Matrix for classification of student grade

Labels	**Predicted**				
	P	**Q**	**R**	**S**	**T**
P	PP	PQ	PR	PS	PT
Q	QP	QQ	QR	QS	QT
R	RP	RQ	RR	RS	RT
S	SP	SQ	SR	SS	ST
T	TP	TQ	TR	TS	TT

The performance is calculated by using the confusion matrix and evaluating the precision, f-measure, recall, accuracy.

$$\text{Accuracy (A)} = \frac{(PP+QQ+RR+SS+TT)}{\sum X}$$

Where X is the sample count

$$Precision(B) = \frac{1}{5}\left(\frac{PP}{PP+QP+RP+SP+TP} + \frac{QQ}{PQ+QQ+RQ+SQ+TQ} + \frac{RR}{PR+QR+RR+SR+TR} + \frac{SS}{PS+QS+RS+SS+TS} + \frac{TT}{PT+QT+RT+ST+TT}\right)$$

$$Recall(C) = \frac{1}{5}\left(\frac{PP}{PP+PQ+PR+PS+PT} + \frac{QQ}{QP+QQ+QR+QS+QT} + \frac{RR}{RP+RQ+RR+RS+RT} + \frac{SS}{SP+SQ+SR+SS+ST} + \frac{TT}{TP+TQ+TR+TS+TT}\right)$$

$$F-Measure = 2\frac{BC}{B+C}$$

So, finally the values are calculated by using the above formula and the model is tested.

4. Results

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3031-3041
https://publishoa.com
ISSN: 1309-3452

Student grade is one of the main criteria to know the performance of a student. By predicting the student grade, the person will be able to know the capacity and will help in the improvement of the student performance. Here, in this methodology we implanted different types of algorithms. Here by comparing the results of each and every algorithm and algorithm which gives best results is identified and shown in Table1.

Table 1. Comparison of various algorithms with and without SFS

Algorithm	None/SFS	Accuracy	Precision	Recall	F-Measure
J48	None	92.0	92.0	91.0	92.0
	SFS	94.5	94.5	94.5	94.5
KNN	None	91.4	91.4	91.4	91.4
	SFS	93.4	93.4	93.4	93.4
NB	None	91.8	91.8	91.5	91.8
	SFS	92.8	92.8	92.8	92.8
SVM	None	91.3	91.0	91.3	90.9
	SFS	93.2	93.2	93.2	93.2
LR	None	91.5	91.3	91.5	91.3
	SFS	93.5	93.5	93.5	93.5
RF	None	94.3	94.1	95.3	95.3
	SFS	95.6	95.1	96.2	94.6
XG BOOST	None	93.1	92.9	93.1	93.0
	SFS	94.6	94.0	94.1	94.1
Hybrid Method	None	95.3	92.6	92.6	91.6
	SFS	96.6	95.6	94.6	94.6

The prediction of student grade is implemented by using eight algorithms and SMOTE and feature selection techniques. The above table shows precision, accuracy, f-measure and recall calculated results for each and every algorithm. The Hybrid method gives an accuracy of 96.6%, Random Forest with an accuracy of 95.6% and XGBoost with an accuracy rate of 94.6%. J48, LR, KNN, SVM and Naive Bayes achieved an accuracy rate of 94.5%, 93.55, 93.4%, 93.2% and 92.8% respectively. Predicting academic achievement is one of the performance indicators that educators can use to keep track of their students' academic progress. The graphical representations of comparison of accuracy values for various algorithms are shown in fig2.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 3031-3041
https://publishoa.com
ISSN: 1309-3452

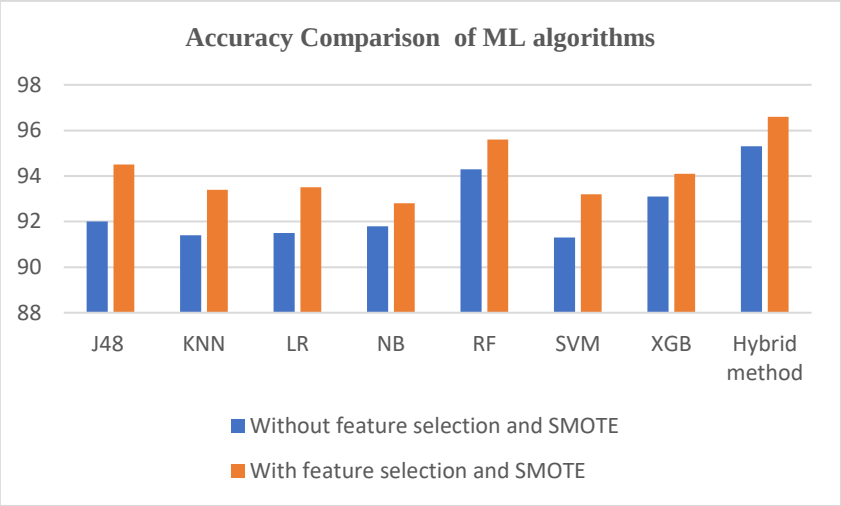

Fig 2: Comparison of Accuracy among various algorithms with and without SFS

5. Conclusion

Predictive analytic applications are a popular topic in higher education. Predictive analytics involves use of analytical models, including machine learning applications that can provide greater quality performance and data for students at every stage of learning. Students at all levels of study may expect high-quality performance and data. Most people are aware that a student's grade is one of the critical success factors that instructors could have used to monitor a student's academic achievement. Several machine learning methods have been reported in the education industry over the last decade. However, dealing with skewed data in order to improve the accuracy of projecting student grades is difficult. Machine learning algorithms are being used for predicting grade of a student with a higher degree of accuracy. The model is trained and tested using Student Grade Prediction dataset. The accuracy of the most popular machine learning methods, including Random Forest ,Naive Bayes, Decision Tree, K-Nearest Neighbour, Logistic Regression, Support vector Machine and a hybrid model which is the combination of Random Forest and XGBoost algorithm are tested. A multiclass prediction model is developed that combines Synthetic Minority Oversampling Technique and a Feature Selection technique to reduce the problems caused by misclassification and overfitting results given by asymmetrical multi-classification. The Random Sampling algorithm supports SMOTE. The Random Sampling algorithm is a method used for randomly selecting subsets from a sample by the user. Wrapper and Filter methods are also used in feature selection. The features are either eliminated or included by comparing the testing results with the output. The results of this proposed model are comparable and acceptable, and they are being used to improve the asymmetrical multi-classification student grade prediction, the prediction performance model was developed. As for future work, Investigation can be done using the emerging predictive techniques and ensemble techniques by using advanced machine learning techniques. Ensemble technique is essential for selecting several multi-classes imbalanced datasets. Additionally, it assists in identifying acceptable sampling procedures as well as various assessment measures that can be utilised for asymmetrical multi-class sectors, such as weighted accuracy, Kappa, and so on. Therefore, using machine learning in the education sector for prediction of a student's grade will help in improving the student's academic performance in future.

References

[1] Hayat Sahlaoui, El Arbi Abdellaoui Alaoui, Anand Nayyar, Said Agoujil, Mustafa Musa Jaber, ‘‘Predicting and Interpreting Student Performance Using Ensemble Models and Shapley Additive Explanations”, IEEE Access, Vol no 9, 2021. doi: https://doi.org/10.1109/ACCESS.2021.3124270

[2] E. Alyahyan, Dilek Düştegör ‘‘Predicting academic success in higher education: Literature review and best practices’’, International Journal of Educational Technology in Higher Education, Vol. No. 17, Dec. 2020.

[3] Fadhilah Ahmad, Nur Hafieza Ismail and Azwa Abdul Aziz, ‘‘The prediction of students’ academic performance using classification data mining techniques’’, Applied Mathematical Sciences, Vol. 9, Pg no. 6415 - 6426 , 2015.

[4] Zafar Iqbal, Junaid Qadir, Adnan Noor Mian, Faisal Kamiran, ‘‘Machine learning based student grade prediction”, Computers and Society. https://doi.org/10.48550/arXiv.1708.08744

[5] Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque and Rashedur M Rahman, ‘‘Improving accuracy of students’ final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique’’, Decision Analytics , 2015. DOI 10.1186/s40165-014-0010-2

[6] X. Zhang, Ruojuan Xue, Bin Liu, Wenpeng Lu, Yiqun Zhang, ‘‘Grade prediction of student academic performance with multiple classification models’’, 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2018.

[7] Diego Buenaño-Fernández, David Gil, Sergio Luján-Mor, ‘‘Application of machine learning in predicting performance for computer engineering students: A case study’’, Sustainability, Vol no.11, 2019; doi:10.3390/su11102833.

[8] Timothy Anderson, Randy Anderson, ‘‘Applications of machine learning to student grade prediction in quantitative business courses’’, Global Journal of Business Pedagogy, Volume 1, Number 3, 2017.

[9] Archit Verma, ‘‘Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA’’, International Research Journal of Engineering and Technology, Vol no.06 Issue: 03,2019.

[10] Maria Tsiakmaki, Georgios Kostopoulos, Sotiris Kotsiantis * and Omiros Ragos, ‘‘Implementing autoML in educational data mining for prediction tasks’’, Appl. Sci. 2020; doi:10.3390/app10010090.

[11] Bratislav Predić,Gabrijela Dimić,Dejan Rančić,Perica Štrbac,Nemanja Maček,Petar Spalević, ‘‘Improving final grade prediction accuracy in blended learning environment using voting ensembles’’, Comput Appl Eng Educ, pg no 1–13, 2018. doi: 10.1002/cae.22042.

[12] Raza Hasan, Sellappan Palaniappan, Salman Mahmood, Ali Abbas, Kamal Uddin Sarker, Mian Usman Sattar, ‘‘Predicting student performance in higher educational institutions using video learning analytics and data mining techniques’’, Appl. Sci., Vol no 10; 2020. doi:10.3390/app10113894

[13] Thiago M. Barros, Plácido A. Souza Neto, Ivanovitch Silva, Luiz Affonso Guedes, ‘‘Predictive models for imbalanced data: A school dropout perspective’’, Educ. Sci. Vol no 9, 2019. doi:10.3390/educsci9040275

[14] Ankit Kumar Srivastava, Devender Singh, Ajay Shekhar Pandey, Tarun Maini, ‘‘A novel feature selection and short-term price forecasting based on a decision tree (J48) model’’, Energies, Vol no 12, 2019.doi:10.3390/en12193665.

[15] Jyothsna, V., Prasad, K.M., Rajiv, K. et al. Flow based anomaly intrusion detection system using ensemble classifier with Feature Impact Scale. Cluster Computing 24, 2461–2478,2021. https://doi.org/10.1007/s10586-021-03277-5

[16] Silpa C, RamPrakash Reddy Arava , Dr K.K. Baseer "Agri Farm: Crop And Fertilizer Recommendation System For High Yield Farming Using Machine Learning Algorithms" International Journal of Early Childhood Special Education (INT-JECSE) Vol no 14, Issue 02, 2022. DOI: 10.9756/INT-JECSE/V14I2.740 ISSN: 1308-5581.

[17] Kaggle, Kaggle.com[online], Available: https://www.kaggle.com/datasets.

[18] E. Krishna, T. Arunkumar, “Hybrid Particle Swarm and Gray Wolf Optimization Algorithm for IoT Intrusion Detection System”, International Journal of Intelligent Engineering and Systems, Vol.14, No.4, 2021 DOI: 10.22266/ijies2021.0831.07

[19] Sri Lalitha, Y., Gayathri, Y., Aditya Nag, M.V., Althaf Hussain Basha, S, “Student Performance Prediction—A Data Science Approach”, Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough. Studies in Computational Intelligence, vol 956. Springer, 2021 Cham. https://doi.org/10.1007/978-3-030-68291-0_10