Predicting Student Academic Performance with Ensemble Classification Method on Imbalanced Educational Data

E Deepak Chowdary, V Lakshmi Prasanna, V Vamsi Krishna T, Gokul Yenduri

Abstract*: Education benefits a person in sustaining his present and future by assuring the goal of life. At present universities and colleges are mainly focusing to improve the academic performance of the students. Recently, many studies have concentrated on employing several machine learning models in the field of higher education to assist both the teachers and the students to identify their problems and can take remedial measures to improve their performances. Some of the earlier studies have been discussed about class imbalance problem and achieved poor prediction outcomes due to low performance of the classifiers. In this paper, we aim to improve the classification/prediction outcomes with a rule-based ensemble model based on various sampling strategies by addressing the class imbalance problem. The dataset used for this study has been collected from Vignan's Lara and Vignan's Nirula institutions based on considering various factors such as Attendance Percentage, No of Backlogs, Adjustable Nature, Concentration, Result History, Discipline in class, Usage of Social Media, Degree of Intelligence, Understanding of Subjects, Event Participation, Time Management, Extra Classes, Alternative Learning Skills, Logical Thinking, Bad Habits, Parents Education, Health Condition, Planning for higher studies, Family Support, Time Management, and Aggregate. To evaluate the efficiency, we also considered and compared our original dataset with different benchmark datasets and the performance measures of the proposed method is also tested with various sampling methods based on a learning rate parameter ranging between 0.1 and 0.8. The original data set with the re-sampling method with the proposed method achieved maximum precision values at a learning rate 0.3 with an accuracy rate of 98.36%. Finally, the obtained results were also compared with several baseline classifiers like Naïve Bayes, SVM, MLP, KNN, and OneR on the collected original datasets.

Keywords: Education, Student Performance, Sampling, Class Imbalance, Classification, Prediction.

I. INTRODUCTION

Today, the academic performance of students has become more essential, particularly in a higher education organization [1]. The main objective of all educational organizations is to expand student performance in learning. The primary goal of this study is to examine the current research operations focused on the prediction of academic results of students. Student academic achievement provides useful data for instructional officials that offers different possibilities to decide and to support learners in their research [2]. The performance of students in organizations is assessed by their educational outcomes or by how well students fulfill the expectations set by the teacher and the organization itself. In addition, student performance can also be evaluated on how far students, teachers and institutions have accomplished their short-and long-term academic objective. There has been a growing interest in using the great amount of information gathered in the educational system. Some problems arise when large amounts of data are stored in an student data base known as Educational Data Mining (EDM) [3-5]. EDM in many organizations has been widely used. The implementation of education data mining can encourage educational institutions to predict the performance of their students. Today, institutions work in a highly competitive environment and have a sophisticated system. Modern universities need a thorough analysis of their performance to improve their strategy and future decisions. The university management should recognize the performance of their students as one of the main aspects of the marketing strategy and must address the institution's promising opportunities [6]. Predicting the performance of students can help to identify the poor student and help university management to adopt decisive policies for improving student performance. The management of institutions can take proper measures to prevent the failure of students or their dropout rate [7]. This type of analysis can assist to predict the poor student and motivate the institution management to take solutions and decisions to improve performance. The attributes considered for this analysis are Attendance Percentage, No of backlogs, Adjustable Nature, Concentration, Result History, Discipline in class, Degree of Intelligence, Social Networking sites, Understanding of Subjects, Event Participation and Aggregate. For this type of analysis, the university management can take early measures to prevent the student from failing or leaving. After understanding the prediction, the poor student is also anticipated to enhance their efficiency and attain better results. The main aim of this paper is to predict the performance of a student based on their educational data using various Data Mining techniques.

From the year 1990, the term "Data Mining" has become more popular and many people treat data mining as a synonym for Knowledge Discovery. It is mainly used to extract knowledge from large amounts of data for which interesting patterns are generated. These patterns are delivered to the user and may be stored in the Knowledge base.

Revised Manuscript Received on November 15, 2019
***** Correspondence Author
E Deepak Chowdary*, Department of CSE, VLITS, Vadlamudi, India.
V Lakshmi Prasanna, Department of CSE, VNITS, Vadlamudi, India.
V Vamsi Krishna T, Department of CSE, VLITS, Vadlamudi, India.
Gokul Yenduri, Department of IT, VFSTR, Vadlamudi, India.

Data mining has several functionalities like classification, Association Rule, Clustering etc., There are two main tasks in this data mining. They are classified as descriptive and predictive. Predictive is again classified into classification and prediction. In classification, mapping of a given data sample is predefined as a class label. The classification techniques are used to build models and predict the future data trends. In prediction the target class is continuous. The term "classifier" in classification implementation refers to a mathematical function that maps a category of input data. An exemplary measure used to compute the performance of classifiers is classification accuracy. There are many algorithms that were proposed over many years that bring out different knowledge representations. These algorithms may be effective for many classification problems. But every time they do not lead to good accuracy. As per theoretical studies, there is no better algorithm used for many data sets. Every algorithm has its own advantage to solve some problems. In this paper, we used a Rule Induction algorithm which defines formal rules that are obtained from a set of observations. These obtained rules may exhibit a complete data model with patterns. There are many rules which are generally considered and those rules are one of the most well-known types of knowledge used in practice. Many algorithms have been developed to involve those rules. We also utilized many Supervised Machine Learning algorithms to assess the prediction performance along with the proposed method.

II. RELATED WORK

Previously several attempts had made in advancement of model and techniques to predict the achievement of apprentice. Countless researches were presented for forecasting and scrutinizing conduct of learner. Y. Chen et al. [8] proposed a hybrid approach to enhance graduation rates from educational data. The method is explained with educational data from a public university of four years. They discovered the crucial part in developing graduation rates. The results showed that the average fall, housing, high school and average spring-term scores were the four most significant determinants and the ethnic background of the students was the least important factor for the student. The findings also showed that campus students could complete their studies more possibly within six years. T. Jiang et al. [9] discussed about finding the behavior of students by collecting card consumption data happened in supermarket, canteen and etc.... R. Asif et al. [10] tend to analyze the performance of various undergraduate students by using educational data. The achieved results prove that few important courses impacted the students.

D. Su et al. [11] proposed a new technology to analyze the consumption behavior of students to analyze student attitudes by providing decision support in canteen and other places for school decision-makers. J. Xu et al. [12] proposed a novel approach to predict the performance of 1169 students in degree programs based on their current and past performance. They collected data was obtained from undergraduate student data over three years at University of California. They developed a clustering method based on a latent factor model to identify the relevant courses. After, an ensemble based progressive prediction framework has been developed to employ the student developing performance into prediction. M. Sagar et al. [13] through challenging programming sites intended to preface and formulate the capabilities of learner so that apprentice may habitude to work out on problems. This helps the students to estimate one and also benefit the professors to analyze their student's performance. The input was composed from challenging programming sites such as Hacker Earth and IGDTUW University. The following aspects like student register number, number of problems solved by students, regularity of student to classes, efficiency, precision, number of errors, type of problem solved are considered to enhance the capability of learner. The techniques adopted are Decision Tree, Perceptron learning, Random Forest, Multilayer perceptron, Naive Bayes, J48, SMO, Bayesian Nets, Instance Based Learning.

S. Venkatramaphanikumar et al. [14] presented a case study on students performance prediction . Now a days, Academies and Universities are struggling hard to improve the academic performance of students. Hence by using this model the development of learner may be anticipated prior to the course so that necessary restorative steps may be implemented to enhance the student performance. In this audit, data fragments are poised from 307 apprentices of Computer Science and Engineering students, in India. The virtues considered for this case study are Online learning skills, problem solving efficiency, sports participation, intension of doing higher studies, days scholar, self –learning, usage of library, versatile nature and some other aspects that effect the development of student. The approach adapted on the dossier is multivariate regression forecast model M5P which is a decision tree initiation. E. P. I. Garcia et al. [15] supervised a audit on students of UNAM engineering college. The main goal of Ernesto is to anticipate the apprentice collegiate performance .The stats was detached into three categories the primary one is apprentice who cleared nil and at most two subjects, the secondary is apprentice who cleared three and four subjects, the tertiary is apprentice who cleared all five subjects .The methods like k-NN, IBK, decision Trees, Naive Bayes, Logistic regression and Rapidminer groupware is composed to classify the features.

I. A. A. Amra et al. [16] have intended to improve the accomplishment of apprentices by prior indication as that may favor the ministry of education. In the research, educational data mining routines have been owned as its main goal is to explore the shrouded intelligence and patterns relating to apprentice performance. This paper illustrates a apprentice conduct precognition model by exercising a binary categorization algorithms –Naive Bayes and k-Nearest Neighbor (KNN) on informational data, composed from results of high school docents in Gaza for the year 2015.The incisiveness obtained by Naive Bayes is high so it is more preferable than KNN. Hence the preceptors can also endure the appropriate evaluation to improve novice knowledge.

O. A. Echegaray-Calderon et al. [17] proposed the most fascinating problem is to forecast the academic progress in higher education system. With help of genetic algorithms such as artificial neural networks, Bayesian networks and decision trees are used to classify the traits. The traits such as test score, Classroom Intervention ,lab work, attendance report , participation in extracurricular activities ,genre ,parents occupation ,parents Education profile, Residency ,work are taken into consideration to predict the performance of student in final exam inactions.

S. Huang et al. [18] contemplated that student performance prediction assist the docents to know how the apprentice will accomplish in examination that is how high and low so the docents can take proactive part to enhance apprentice learning. The data sets were collected from department of engineering and technology education, the course of engineering dynamics at, Utah state university. The models such as MLR, MLP, RBF and SVM are used to analyze the attributes. The attributes such as Learning Style, Motivation, interest, Time devoted to learning, family Background are considered. The incisiveness obtained by support vector machine is high so it is more preferable. Y. Meier et al. [19] have intended to propose an algorithm that forecast the end grades of students so that the preceptor can take necessary restorative measures prior to the course. The main objective is to improve the efficiency of apprentices in classical classroom courses by up-to-date accomplishment prognosis .The prognostication depends upon the previous history of learner's conduct , time spent on specific questions and the achievement of apprentice in class examinations. The audit is composed on the 700 UCLA first year apprentices who have undergone a preliminary course.

I. Hidayah et al. [20] developed a theory which is an association of fuzzy IF-THEN standards and neural network's capability to study, in order to obtain a student categorization model. This model is a solution to many difficulties like huge volume of apprentices and finite sum of conferences accomplishing hard to assess every apprentice and also helps to forecast apprentice's academic achievement. A Neuro-Fuzzy theory is also based on pattern recognition. This technique has capability to understand from the produced standards so that most excellent categorization model can be obtained. ANFIS Editor-Matlab Fuzzy logic has been applied on the data. Hence they have concluded that three attribute domains that affect the outcome of apprentice categorization model are interest, talent and motivation. D. Fensel et al. [21] had a study about a problem regarding rule induction. They proposed an algorithm called RELAX with a modification used for generalization and used for rule induction also known as "dropping condition rule". They also provided a minimal set of rules which allows the recognition of object with minimal costs. To restrict the first order rules and examples, they extended with the formal representation and integrated statistical information which guides the searching procedure.

M. Inuiguchi et al. [22] proposed a study about classes in cluster decision before the rule induction is applied. The similarity between these classes is defined by applying an agglomerative hierarchical clustering method. They also applied LEM2 rule induction algorithm to induce the decision rules for deducing the clusters. The results for this experiment is obtained from dendrograms which are randomly generated adopted with some operations. J. Stefanowski et al. [23] led a study of rule induction algorithm based on rough sets working with combiner, bagging and n^2 classifier. The results obtained from this experiment confirm the improvement of classification range of the combiner depending on errors generated by the component classifiers. They firstly observed that the classification accuracy was no improved by applying combiner strategy than other two classifiers. The accuracies with 15 datasets were not significant and improved with 4 datasets. A. Mahajan et al. [24] has led a study about "Performance evaluation of Rule Based Classification Algorithms" about five classifications algorithms OneR, PART, Decision Table, DTNB and Ridor. In this model Weka tool is used to measure accuracy and error rate. Cross Validation is used testing options. Based on the accuracy generated by all the five algorithms, OneR has generated with lower accuracy.

C. Seiffert et al. [25] discussed about re-sampling or re-weighting with a boosting algorithm. In this study they had shown the performance improvement in classifiers in different situations. They implemented boosting by reweighting to base learners designed for handling example weights. By using imbalanced training data, the evaluation is performed between two boosting implementations. The performance by boosting is found by using ten boosting algorithms with 4 learners and 15 datasets. Finally, they concluded that boosting by resembling is adopted than boosting by weighting. S. Balakrishnan et al. [26] had a study on the performance of Rule-based classification by feature selection. They considered the Medical database which consists of huge amount of clinical data those provides valuable information like diagnosis, prognosis and treatment plan when these classification algorithms are used in a relevant manner. The features in this study are less influenced on the predicted output which was rejected. Instead of this, an optimal feature subset is recovered by enhancing the accuracy of the classifier and the performance of this classification model is proved with small amount of discriminatory features. I. Journal et al. [27] has explored in an article "Proficiency Comparison of ZeroR, RIDOR and PART classifiers for Intelligent Heart Disease Prediction" about comparative results of classifiers RIDOR, ZeroR, PART used in prediction of heart disease. The survey has proved that RIDOR is the best classifier to predict the heart disease. V. Veeralakshmi et al. [28] has intended in a paper "Ripple Down Rule learner (RIDOR) classifier for IRIS Dataset" about three Rule Based classifiers named JRIP, RIDOR AND Decision Table. This model uses an IRIS dataset for calculating the performance. The analysis of algorithms is done by using factors like execution time and classification accuracy. By observing the accuracy results RIDOR performs with best accuracy.

III. METHODOLOGY

In this study, we evaluated the performance of various students by collecting the real time data. The data sets are assembled from students of II Vignan's Lara Institute of Technology and Sciences. Some techniques like data scrubbing, data processing are used in this research paper. Data Scrubbing is used to refine the data by padding the left out values. The Data Preprocessing is the phenomenon in which data is organized, accumulated and derived.

In this study, we evaluated the performance of students with different machine learning classifiers by collecting the real time data. The dataset was prepared based on collecting the demographical information from various students according to the factors as described below. Data Discretization specifies transmissioning steady functions, elements and statistical algorithms. For example, Attendance is classified as the troop such as regular and Irregular.

A. Attendance Percentage: The attribute Attendance represents the regularity of the students. It is

tagged into two categories.

1 .Regular (>=75)
2 .Irregular (<75)

B. No of Backlogs: Due to backlog burden increases in students and could not focus properly in current academics. They are expressed in quantitative form.

C. Adjustable Nature: Students who can adapt the environment easily they can fixate on studies and tolerate all types of mindsets. They can be classified as the {Easy, Slow}.

D. Concentration: The students who Concentrates on classes need not to spend much time to study. This can be identified as {Yes, No}.

E. Result History: This plays a very important role for faculty to help the students who are poor in academic. This can be indicated in numerical form from {1 to 5}.

F. Discipline in class: The Discipline justifies the character of a student who behaves in class. This can be organized in the form of {Good, Bad}.

G. Usage of Social Media: If students are socially active than they are educationally inactive. This can be labeled as {High, Low}.

H. Degree of Intelligence: Degree of Intelligence can be measured in how easily a student can understand the subject. This can be analyzed as {High, Low}.

I. Understanding of Subjects: If students are attentive in class, it helps them to properly understandability of subjects they can gain some grip on it. They can be categorized as {Good, Bad}.

J. Event Participation: This helps students to get relaxation in free time, which helps to settle on studies effectively. This can be arranged as {Yes, No}.

K. Time Management: The students who manage time wisely be obedient, dependable, and attentive to classes. They are grouped as {Good, Bad}.

L. Extra Classes: Other than academics, students can learn some extra subjects based on their branch. This can be analyzed as {Yes, No}.

M. Alternative Learning Skills: Students who have alternative skills line Online Learning, Referring Library, etc… can learn effectively. These are tagged as {Yes, No}.

N. Logical Thinking: Logical Thinking in students helps them to solve and reason all kinds of problem in Logical way improve and problem solving capacity. This is arranged in the form of {Good, Bad}.

O. Bad Habits: Generally day bad habits mislead studies in students, which results degradation of their performance in academics. They are categorized as {Yes, No}.

P. Parents Education: If the parents are well educated, they can guide and suggest their children in the correct way. If parents are uneducated they cannot guide their children, so they have to be self learned. These can be assembled as {Post Metric, Pre Metric}.

1. Post Metric -PhD, UG, PG, BBA, M.tech, B.ED.
2. Pre Metric – Class X, Class XI

Q. Health Condition: Students with proper health condition can contemplate effectively on studies. Health can be grouped into {Good, Bad}.

R. Planning for higher studies: Students need good aggregate for doing higher studies, so they can make their graph in the increasing order. These can be classified as {Yes, No}.

S. Family Support: Students have good family support can properly manage their academics. This can be tagged as {Yes, No}.

T. Time Management: If the students regular study for minimum hours it would minimize the burden at the end and also help for good score in academics.

U. Aggregate: The Aggregate of apprentice is classified into five groups.

1. Exemplary (>=90)
2. Outstanding (>=75)
3. Satisfactory (>=65)
4. Average (<65)
5. Fail (<45)

These are again assembled into different categories as E, O, S, A, F. Additionally, this proposed work also made several assessments to predict the performance of students in the final examination and evaluate the effectiveness of various classifiers using different sampling techniques with various parameters based on error and accuracy as discussed below.

IV. SAMPLING TECHNIQUES

In many data mining implementations, the class imbalance problem is often a serious problem [29]. Traditional learning optimization algorithms do not classify imbalanced data. With the goal of improving overall performance, some training methods benefit majority-class training instance and thus reduces the prediction effectiveness of the minority class. Several researchers in the field of engineering then turned their attention on the issue of the class imbalance. The imbalanced data sets were handled by different techniques and procedures. At the data point, the goal is to re-sample the data space for the class distribution. One of the essential methods for handling of imbalanced data is sampling. The most effective way to address the class imbalance is to adjust the class distributions to a more evenly balanced distribution. Such methods include a variety of different re-sampling types, including random over-sampling, random under-sampling, enhanced sampling and variations of these techniques. In order to achieve a better distribution, a random over sampling is a non-heuristic method duplicating samples of the minority class.

The following methods are employed for sampling process in this work:

1). Resample is a method develops a unique dataset, which is replaced with a sample [30].

2). Synthetic Minority Over-sampling Technique (SMOTE) generates artificial information on the premise of the similarity of feature space between developed minority samples [31]. SMOTE is a novel sampling method which is used to over-sample the minority class by generating synthesized samples instead of over-sampling by eliminating them. By selecting each minority class, the sample is over-sampled by adding synthetic samples in the line fragments that unite all of the nearest minority classes. The neighbors from the nearest k neighbors are chosen randomly, based on the amount of sampling expected.

3). Spread Subsample: This method generates a unique sub-sample with a certain spread across class frequencies that are sampled with substitution frequencies.

Random sampling is also a non-heuristic technique focused on optimizing data by extracting samples from the main class. In the under-sampling process, representations of the main class are eliminated to ensure the data set is balanced. This approach attempts to balance class distribution by eliminating samples from large class randomly. The disadvantage of the sampling approach is that it is easy to exclude highly useful data, which may be essential for the classifications. Random experiments delete instances from the dominant group uniformly until the data set is balanced. The informative approach selects only the necessary major class samples focused on the selection criteria pre-specified to balance the data set. The sampling technique is a sampling method that incorporates the data set by replicating minority class samples. The benefit of this approach is that the analysis procedure does not result in data loss. The limitation of this approach is that if the data set is already very large or inconsistent and it can result in over fitting and may add additional computing cost. Oversampling is also classified into two different types, like under sampling:

1. Random Oversampling and
2. Informative-Oversampling.

Such randomly selected members will be replicated and transferred to the new training set. This strategy increases the total minority class instances synthetically; the informative process of oversampling produces samples for minority class based on the predefined rule. In short, sampling could lead to greater training time. Under-sampling is the alternative to over-sampling. In case of time and space complexity, this method is easier than sampling.

V. RULE BASED CLASSIFICATION

Stefanowski [32] had introduced a rule based algorithm called *MODLEM* which is an application of the rule induction algorithm. This is an exemplary algorithm for inducing minimal set of rules. The disjunctive set of conjunctive rules is known as a set of rules.

Rule: A Rule is a popular symbolic notation of knowledge from derived data. The standard form of these rules is represented as

$$IF\ P\ THEN\ Q \qquad (1)$$

where, *P* denotes a Condition part and *Q* denotes a Decision part. This algorithm is mainly based on the sequential covering an interesting minimal rules that are generated for every decision class. These types of rules cover most significant positive examples and do not cover negative examples. The Rule Induction strategy starts from building a first rule by selecting the best elementary conditions in a sequential order according to given criteria. If this criterion does not accept the rule, then the next best rule is added and evaluated in the elementary condition. If the rule is stored, then all the positive examples which match to this rule are eliminated from the deliberation. By repeating this process iteratively, some of the positive examples remain uncovered and the strategy for each set of examples is repeated sequentially. This algorithm is similar to LEM2 algorithm and also evaluated by either class entropy or Laplace class. The numerical attributes N_t in this algorithm are denoted as $(x<v)$. where v represents the attribute threshold. The representation of the given attribute is represented as

$$N_t = (x < v)\ or\ N_t = (x \geq v) \qquad (2)$$

By depending on $[N_t]$ block, more training samples are covered from *Y*. While building the single rule, if the same attribute is selected twice at the same time, then one of the rules may retrieve the condition as $(x = [v_1, v_2])$ and results in two conditions by intersecting $(x < v_2)\ and\ (x \geq v_1)$ Such that $v_1 < v_2$. The procedure of this algorithm is followed by sorting the numerical attribute values of "x" for all the examples in increasing order. The candidates are generated in sorted order for the mid points (cut-points) between consecutive values and these values are only considered only when the examples belong to the various decision classes. The evaluation of cut-points is done by using a class entropy technique to find the best cut-point. More positive examples are covered from the set *Y*, when the best cut-point is used to choose the condition $(x < v)\ or\ (x \geq v)$. Finally, to find the best condition, this procedure is repeated for all the other attributes unless the complete rule is induced. The schema for the rule based classification algorithm is represented as follows:

Input data: set Y, a set Z of attributes.
begin
$A := Y;$
$C := \emptyset;$
while $A \neq \emptyset\ do$
begin
$C := \emptyset;$
$W := U;$
while $(C = \emptyset) or$ **not** $([C] \subseteq Y)\ \ do$
begin
$N_t; = \emptyset;$
$N_{t_E} := \infty;$
for each attribute $x \in Z$ do
begin
F_B_C $\left(x, W, new_{N_t}, E_{new_{N_t}}\right);$
if $E_{new_{N_t}} < N_{t_E}$ the
begin
$N_t := new_t;$
$N_{t_E} := new_E;$
end;
end;
$C := C \cup \{N_t\};$
$W := W \cap [N_t];$
end;
for each elementary condition $N_t \in C$ do

if $[C - \{N_t\}] \subseteq Y$ then $C := C - \{N_t\}$;
$C := C \cup \{C\}$;
$A := Y - U_{C \in C}[C]$;
end;
for each $C \in C$ do
if $U_{W \in C - \{C\}}[W] = Y$ then $C := C - \{C\}$;
end
Output: single local covering *C* of set *Y*.

/* Finding Best Condition*/
Input data: attribute x, set *W* of objects;
begin
$B_t := \emptyset$:
$E_{B_t} := \infty$;
sort H;
for $i := 1\ to\ (H) - 1$ do
begin
$v := (H(i) + H(i+1))/2$;
$W_1 := \{a \in W | x(a) < v\}$;
$W_2 := \{a \in W | x(a) \geq v\}$;
$E_{N_t} := (|W_1|/|W_1 \cup W_2|) * E_{Nt}(W_1) + (|W_2|/|W_1 \cup W_2|) * E_{Nt}(W_2)$;
if $E_{N_t} < E_B$ then
begin
if $|A \cap W_2| \geq |A \cap W_2|$ then
$B_t := (c \geq v)$ else $B_t := (c < v)$;
$E_{B_t} := E_{N_t}$
end
end
end
Output: $B_t\ and\ E_{B_t}$

In the above algorithm, the class entropy measure is used to compute the conditions.
$C := \emptyset$ represents the candidates for condition part of the rule.
$W := U$ represents the set of objects covered by *C*.
$N_t := \emptyset$ represents the Candidate for elementary condition.
$N_{t_E} := \infty$ represents an evaluation measure for *t*.
F_B_C represents to find the best condition.
E_{N_t} represents the evaluation of the new condition.
E_{B_t} represents the evaluation measure for Best condition.

The Laplacian measure is the measure that is optionally used and preferred to be lower. Laplacian measure is represented as

$$(N_c + 1)/N_{TOT} + n) \quad (3)$$

where *n* represents the number of classes in the given dataset. N_c represents the number of examples in class *Y*. N_{TOT} represents the total number of examples.

VI. ENSEMBLE CLASSIFICATION MODEL

Freund and Schapire introduced AdaBoost algorithm in the year 1995. The main goal of this algorithm is to solve several practical issues caused in boosting algorithms. Adaboost is also a traditional boosting technique which is similar to bagging method and integrates the number of machine learning methods [33-34]. This method predicts the efficiency of the base classifiers by making them sensitive at misclassified instances. The input of this algorithm is taken as $(a_i,b_i), \ldots (a_n,b_n)$, where a_i belongs to sample space *A* and b_i belongs to the label set *B*. $t = 1, \ldots, T$ indicates the number of rounds. The base classifier $f_t: A \rightarrow R$ is initiated by a base learner in the distribution D_t. The Ensemble system in Machine learning selects a classifier from various algorithms for training based on the following eq (4).

$$c = (M, T) \quad (4)$$

where T represents the data sample at training level, M represents a model, and c is a classifier. The general representation of Adaboost is defined as follows.

Input: $(a_1, b_1), \ldots \ldots, (a_n, b_n), a_i \in A, b_i \in B = \{-1, +1\}$
Initialize $D_1(i) = \frac{1}{n}$.
for $t = 1, \ldots, T$
Using distribution D_t train the base learner
$f_t: A \rightarrow R$ is a base classifier
Choose $\alpha_t \in R$ then
Update:
$D_{t+1}(i) = (D_t(i)\ \exp(-\alpha_t b_i f_t(a_i)))/N_t$
Output: $F(a) = sign(\sum_{t=1}^{T} \alpha_t f_t(a))$

From the above algorithm, D_t indicates the Distribution and N_t denotes the Normalization part. There are a large number of methods for building an ensemble of classifiers. By re-sampling the training dataset all the base classifiers are generated in this boosting algorithm. The most crucial consideration in building the ensemble method is to meet with AdaBoost. Additionally, we also used several machine learning classifiers like Naive Bayes [35], Support Vector Machine (SVM) [36], K-Nearest neighbor (KNN) [37], Multi-layer Perceptron (MLP) [38], and OneR [39].

VII. EXPERIMENTAL RESULTS

This proposed model is evaluated on real-time student data collected from Vignan Groups. Several Classifiers were also used in this study to assess the performance of the proposed approach when compared to each other. The Imbalanced data set is a common problem with most of the data sets. Sampling methods play a key role in the efficiency of classification when most classification frameworks cannot accommodate an imbalanced class distribution.
A comparative analysis with sampling algorithms for predicting student performance was performed in this study. For the proposed method, the impact of the class imbalance in the training data was evaluated. The evaluation of the data set was conducted using three sampling methods, which were tested by the proposed method using learning rate. The Resample approach shows better performance for data sets among all the sampling strategies. For sampling, the methods Resample, SMOTE, and Spread Sub Sample are simultaneously used. The same test was conducted by adjusting the learning rate between 0.1 and 0.8 in multiple incremental levels. For this analysis, we used open-source software (WEKA) that contains a list of machine learning algorithms for data extraction tasks. The platform provides several mechanisms for preprocessing the data, feature extraction, association rules, clustering, classification, and visualization. We also compared our approach with various classification models with respect to Resample, SMOTE and Subsample. Once the class labels are identified, different classification models like

Naïve Bayes (NB), MultiLayer Perceptron (MLP), Support Vector Machines (SVM), K-Nearest Neighbor (KNN) and OneR were applied on this dataset. In order to compute the features and analyze their effect on class labels of student performance, we use Root Mean Square Error (RMSE) technique and obtained positive rates among all independent class attributes as shown in Table- II. We also tend to compare our proposed method on various datasets, as shown in Table. I. So, we collected different datasets like and Tic-Tac-Toe, Breast Cancer, Annealing, Abalone, German-Credit, Ionosphere, Vehicle, and Balance-scale from UCI repository web sites.

Table- I: Statistics of Datasets

Dataset	Instances	Attributes
Tic-Tac-Toe	**958**	**9**
Breast Cancer	**699**	**10**
Annealing	**798**	**38**
Abalone	**4177**	**8**
German-Credit	**1000**	**20**
Ionosphere	**351**	**34**
Vehicle	**946**	**18**
Proposed	**2310**	**21**

The description of the above datasets used for this study is described as follows: Tic-Tac-Toe dataset is considered with 958 instances and 9 attributes based on collecting data and the class label is categorized into 3 values, Breast Cancer dataset is considered with 699 instances and 10 attributes with the class labels categorized into 2 values, Annealing dataset is considered with 798 instances and 38 attributes for this evaluation, based on collecting data, the class label is categorized into 6 values, Abalone dataset is considered with 4177 instances and 8 attributes for this evaluation, based on collecting data, the class label is categorized into 29 values, German-Credit dataset is considered with 1000 instances and 20 attributes for this evaluation, based on collecting data, the class label is categorized into 2 values, Ionosphere dataset is considered with 351 instances and 34 attributes for this evaluation, based on collecting data, the class label is categorized into 2 values, and Vehicle dataset is considered with 946 instances and 18 attributes for this evaluation, based on collecting data, the class label is categorized into 4 values. We considered various evaluation metrics to analyze the efficiency of the model built at each iteration. On the other hand, the classifier efficiency is estimated by averaging the estimated values of all iterations into a evaluation metric value which is to be considered as a final output. In this regard, a set of metrics like Precision (P), Recall (R), F-measure, Matthews correlation coefficient (MCC), ROC area and PRC area were used to evaluate the proposed method. Precision is used to compute the percentage of significant samples of the extracted ones. Recall is used to compute the percentage of significant and extracted samples over the total number of significant ones. F-measure (F) is used to compute the harmonic mean of both precision and recall. MCC computes a correlation coefficient using all four values in a confusion matrix. ROC area is used to identify a threshold for separating positive and negative samples by analyzing a constructed classifier. All the above evaluation metrics are simplified as follows:

$$P = \frac{TP+TN}{(TP+TN+FP+FN)}$$

$$R = \frac{TP}{(TP+FN)} \qquad (5)$$

$$F = 2 * \frac{P*R}{P+R}$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

True Positive (TP) defines the number of correctly labeled samples which belongs to positive class. True Negative defines the number of correctly classified negative samples which belongs to negative class. False Positive (FP) defines the number of incorrectly classified positive samples that belongs to a negative class. False Negative (FN) defines the number of incorrectly classified negative samples that belongs to a positive class. The following Table-II presents performance measures of the proposed method with various sampling methods and a learning rate ranging between 0.1 and 0.8. The data set with the re-sampling approach with the proposed method of learning rate 0.3 have obtained the maximum precision values as presented in the following Fig.1.

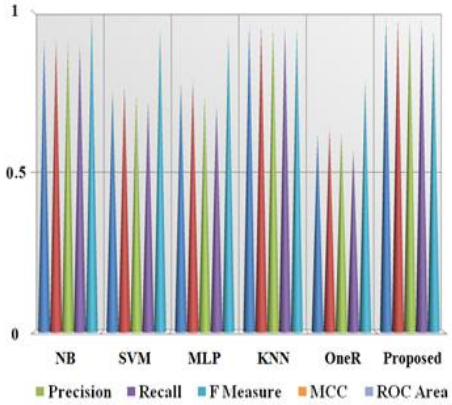

Fig. 1 Performance comparison of the proposed model with various models based on different Evaluation Metrics.

Table II. Error Rates obtained with the proposed method with various Sampling Strategies based on Learning Rate Parameter.

Sampling Methods	Leaning Rate	RMSE
No Sampling	**0.1**	**0.0614**
	0.3	0.0733
	0.5	0.0766
	0.8	0.0803
Re-sampling	0.1	0.0547
	0.3	**0.0514**
	0.5	0.0521
	0.8	0.0556
SMOTE	0.1	0.0687
	0.3	0.0677
	0.5	**0.0662**
	0.8	0.0698
Sub-Sampling	**0.1**	**0.0611**
	0.3	0.0627
	0.5	0.0638
	0.8	0.0654

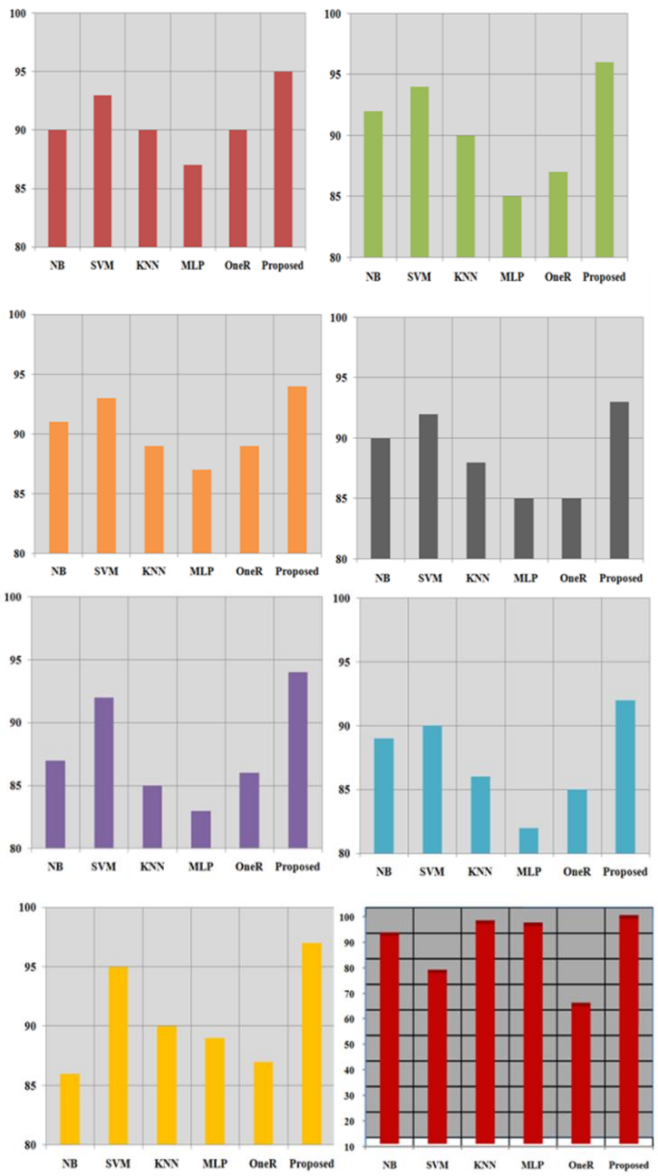

Fig. 1. Performance comparison of the proposed model with different models based on various Evaluation Metrics.

The above Fig. 1 shows the performance of proposed method based on the accuracy evaluation metric with the comparison of different classifiers performed on the proposed educational dataset. The lowest RMSE is obtained with re-sampling by proposed method is to be 0.0514 and the highest precision is to be 0.9842. Therefore, the findings show that sampling can considerably improve the effectiveness of classification studies and the learning rate parameter has to be chosen precisely in order to achieve higher precision for the proposed classification problem. Furthermore, these findings showed that they resample is inferior to all the others. Likewise, the proposed method with no sampling for learning rate 0.1 and 0.3 is equally accurate, recall and f-measure. As these tables demonstrate, low learning rates will generally lead to high RMSE values. There is no assumption, though low learning rates often cause low RMSE values.

VIII. CONCLUSION

Education plays a vital role in every individual's life by developing personality, knowledge and analytical skills. Many universities and colleges are mainly focused to improve the academic performance of the students. Hence, by predicting the grade of the student we can prevent student dropout rate. In this study, we employed several machine learning methods like NB, SVM, MLP, KNN and OneR along with proposed ensemble model to predict the student academic performance. The data was collected from the student of Vignan groups of colleges based on several factors like Attendance Percentage, Aggregate, No of backlogs, Adjustable Nature, Concentration, Previous Academic History, Results, Discipline in class, Degree of Intelligence, Social Networking sites, Understanding of Subjects, Participation in Events. The original data set experiences a class imbalance problem in the training phase and has been overcome by using various sampling strategies based on learning parameter and RMSE. Therefore, the findings show that sampling can considerably improve the effectiveness of the proposed model by precisely selecting the learning parameter. The lowest RMSE is obtained with re-sampling technique by proposed method is to be 0.0514 and the highest precision is to be 0.9842. Moreover, these findings also show that the re-sampling technique with proposed method is inferior to all the other models. At the end, we also tested the proposed approach on various benchmark datasets and compared the efficiency in all the aspects by considering various evaluation metrics. The achieved results proved that the proposed method outperforms with different sampling and classification methods and achieved 98.36% of accuracy. The achieved results proves that, this research can also be applied to corporate data mining studies and demonstrates how sampling methods can enhance the efficiency of the proposed method with an optimal learning rate parameter. In the future, we tend to develop a novel systematic approach to predict the performance of students by collecting a large number of educational data.

REFERENCES

1. J. Luo, S. E. Sorour, K. Goda, and T. Mine, "Predicting Student Grade based on Free-style Comments using Word2Vec and ANN by Considering Prediction Results Obtained in Consecutive Lessons," Proc. 8th Int. Conf. Educ. Data Min., pp. 396–399, 2015.
2. A. Sharabiani, F. Karim, A. Sharabiani, M. Atanasov, and H. Darabi, "An enhanced bayesian network model for prediction of students' academic performance in engineering programs," IEEE Glob. Eng. Educ. Conf. EDUCON, no. April, pp. 832–837, 2014.
3. A. Abu Saa, "Educational Data Mining & Students' Performance Prediction," IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 5, pp. 212–220, 2016.
4. A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," IEEE Access, vol. 5, pp. 15991–16005, 2017.
5. A. Pradeep, S. Das, and J. J. Kizhekkethottam, "Students dropout factor prediction using EDM techniques," Proc. IEEE Int. Conf. Soft-Computing Netw. Secur. ICSNS 2015, pp. 1–7, 2015.
6. D. Kabakchieva, "Predicting student performance by using data mining methods for classification," Cybern. Inf. Technol., vol. 13, no. 1, pp. 61–72, 2013.
7. C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," Comput. Electr. Eng., vol. 66, pp. 541–556, 2018.
8. Y. Chen, Y. Chen, and A. Oztekin, "A hybrid data envelopment analysis approach to analyse college graduation rate at higher education institutions," INFOR, vol. 55, no. 3, pp. 188–210, 2017.
9. T. Jiang, J. Cao, D. Su, and X. Yang, "Analysis and Data Mining of Students' Consumption Behavior Based on a Campus Card System," Proc. - 2nd Int. Conf. Smart City Syst. Eng. ICSCSE 2017, pp. 58–60, 2017.

10. R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, “Analyzing undergraduate students’ performance using educational data mining,” Comput. Educ., vol. 113, pp. 177–194, 2017.
11. D. Su, X. Liu, T. Jiang, and Z. Li, “Research on the Application of Data Mining Technology in Campus Card System,” Proc. - 2nd Int. Conf. Smart City Syst. Eng. ICSCSE 2017, pp. 199–201, 2017.
12. J. Xu, K. H. Moon, and M. Van Der Schaar, “A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs,” IEEE J. Sel. Top. Signal Process., vol. 11, no. 5, pp. 742–753, 2017.
13. M. Sagar, A. Gupta, and R. Kaushal, “Performance prediction and behavioral analysis of student programming ability,” 2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2016, pp. 1039–1045, 2016.
14. S. Venkatramaphanikumar, S. C. Kumar, E. D. Chowdary, and K. V. K. Kishore, “M5P model tree in predicting student performance : A case study,” 2016 IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. RTEICT 2016 - Proc., pp. 1103–1107, 2017.
15. E. P. I. Garcia and P. M. Mora, “Model prediction of academic performance for first year students,” Proc. - 2011 10th Mex. Int. Conf. Artif. Intell. Adv. Artif. Intell. Appl. MICAI 2011 - Proc. Spec. Sess., pp. 169–174, 2011.
16. I. A. A. Amra and A. Y. A. Maghari, “Students performance prediction using KNN and Naïve Bayesian,” ICIT 2017 - 8th Int. Conf. Inf. Technol. Proc., pp. 909–913, 2017.
17. O. A. Echegaray-Calderon and D. Barrios-Aranibar, “Optimal selection of factors using Genetic Algorithms and Neural Networks for the prediction of students’ academic performance,” 2015 Latin-America Congr. Comput. Intell. LA-CCI 2015, pp. 1–6, 2016.
18. S. Huang and N. Fang, “Work in progress: Early prediction of students’ academic performance in an introductory engineering course through different mathematical modeling techniques,” Proc. - Front. Educ. Conf. FIE, vol. 1, pp. 1–2, 2012.
19. Y. Meier, J. Xu, O. Atan, and M. Van Der Schaar, “Predicting grades,” IEEE Trans. Signal Process., vol. 64, no. 4, pp. 959–972, 2016.
20. I. Hidayah, A. E. Permanasari, and N. Ratwastuti, “Student classification for academic performance prediction using neuro fuzzy in a conventional classroom,” Proc. - 2013 Int. Conf. Inf. Technol. Electr. Eng. "Intelligent Green Technol. Sustain. Dev. ICITEE 2013, pp. 221–225, 2013.
21. D. Fensel and J. Klein, “A new approach to rule induction and pruning.” pp. 538–539, 1992.
22. M. Inuiguchi and D. Fukuda, “LEM2-Based Rule Induction via Clustering,” pp. 1–6.
23. J. Stefanowski and S. Nowaczyk, “On using rule induction in multiple classifiers with a combiner aggregation strategy,” Proc. - 5th Int. Conf. Intell. Syst. Des. Appl. 2005, ISDA ’05, vol. 2005, pp. 432–437, 2005.
24. A. Mahajan and A. Ganpati, “Performance evaluation of rule based classification algorithms,” Int. J. Adv. Res. Comput. Eng. Technol., vol. 3, no. 10, pp. 3546–3550, 2014.
25. C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “Resampling or reweighting: A comparison of boosting implementations,” Proc. - Int. Conf. Tools with Artif. Intell. ICTAI, vol. 1, pp. 445–451, 2008.
26. S. Balakrishnan, M. R. Babu, and P. V. Krishna, “An empirical study on the performance of rule-based classification by feature selection,” Proc. - 2014 World Congr. Comput. Commun. Technol. WCCCT 2014, pp. 147–149, 2014.
27. I. Journal, C. Science, and L. Devasena, “Proficiency Comparison of ZeroR , RIDOR and PART Classifiers for Intelligent Heart Disease Prediction,” vol. 3, no. 11, pp. 12–18, 2014.
28. V. Veeralakshmi, “Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset,” vol. 4, no. 03, pp. 79–85, 2015.
29. H. L. Yin and T. Y. Leong, “A model driven approach to imbalanced data sampling in medical decision making,” Stud. Health Technol. Inform., vol. 160, no. PART 1, pp. 856–860, 2010.
30. H. M. Nguyen, E. W. Cooper, and K. Kamei, “A comparative study on sampling techniques for handling class imbalance in streaming data,” 6th Int. Conf. Soft Comput. Intell. Syst. 13th Int. Symp. Adv. Intell. Syst. SCIS/ISIS 2012, pp. 1762–1767, 2012.
31. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 2009 Jun 26;21(9):1263-84.
32. J. W. Grzymala-Busse and J. Stefanowski, “Three discretization methods for rule induction,” Int. J. Intell. Syst., vol. 16, no. 1, pp. 29–38, 2001.
33. R. E. Schapire, “Explaining adaboost,” Empir. Inference Festschrift Honor Vladimir N. Vapnik, pp. 37–52, 2013.
34. K. W. Hsu and J. Srivastava, “Improving bagging performance through multi-algorithm ensembles,” Front. Comput. Sci. China, vol. 6, no. 5, pp. 498–512, 2012.
35. F. Razaque et al., “Using naïve bayes algorithm to students’ bachelor academic performances analysis,” 4th IEEE Int. Conf. Eng. Technol. Appl. Sci. ICETAS 2017, vol. 2018-January, pp. 1–5, 2018.
36. E. Deepak, G. Sai Pooja, R. N. S. Jyothi, S. V. Phani Kumar, and K. V. Kishore, “SVM kernel based predictive analytics on faculty performance evaluation,” Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016, vol. 2016, pp. 1–4, 2016.
37. M. A. jabbar, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” Procedia Technol., vol. 10, pp. 85–94, 2013.
38. J. S. Sonawane and D. R. Patil, “Prediction of heart disease using multilayer perceptron neural network,” 2014 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2014, no. 978, pp. 1–6, 2015

AUTHORS PROFILE

E Deepak Chowdary received his M. Tech degree in Computer Science and Engineering from Vignan's Foundation for Science, Technology and Research in 2015. He is currently working as Assistant Professor in the Department of Computer Science and Engineering in Vignan's Lara Institute of Technology and Science. His research interests include Data Mining and Natural Language Processing.

V Lakshmi Prasanna received her M. Tech degree in Computer Science and Engineering from Vignan's Foundation for Science, Technology and Research in 2015. He is currently working as Assistant Professor in the Department of Computer Science and Engineering in Vignan's Nirula Institute of Technology and Science. His research interests include Machine Learning.

V Vamsi Krishna T received his M. Tech degree in Computer Science and Engineering from Acharya Nagarjuna University in 2010. He is currently Head of the Department in the Computer Science and Engineering in Vignan’s Lara Institute of Technology and Science. His research interests include Data Mining and Machine Learning.

Gokul Yenduri received his Master’s degree (MTech, IT) from VIT University. Currently, he is a Research Scholar at Vignan’s Foundation for Science, Technology, and Research (Deemed to be University), Vadlamudi, India. His areas of interest are in software engineering, computer networks, network security, machine learning, and predictive analysis.