USING DATA MINING TECHNIQUES TO PREDICT STUDENT PERFORMANCE

A PROJECT REPORT

for

DATA MINING TECHNIQUES (ITE2006)

in

B.Tech (Information Technology)

by

VINAY RAVINDRA MASKE (20BIT0128)

CHARVI GARG (20BIT0160)

KARAN RAVI (20BIT0162)

5th SEMESTER, 2022

Under the Guidance of

Prof. B. VALARMATHI

Associate Professor (Senior), SITE

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

November 2022

DECLARATION BY THE CANDIDATE

We here by declare that the project report entitled **"USING DATA MINING TECHNIQUES TO PREDICT STUDENT PERFORMANCE"** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide project work carried out by us under the guidance of **Prof. B.Valarmathi.** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place : Vellore Signature

Date :

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology & Engineering [SITE]

CERTIFICATE

This is to certify that the project report entitled **"USING DATA MINING TECHNIQUES TO PREDICT STUDENT PERFORMANCE"** submitted by **Vinay Ravindra Maske (20BIT0128), Charvi Garg (20BIT0160), Karan Ravi (20BIT0162)** to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide work carried out by them under my guidance.

Prof. B.Valarmathi

GUIDE

Associate Professor (Senior), SITE

USING DATA MINING TECHNIQUES TO PREDICT STUDENT PERFORMANCE

Vinay Ravindra Maske[1], Charvi Garg[2], Karan Ravi[3]

[1,2,3] Department of Information Technology, VIT University, Vellore, Tamil Nadu, India

Abstract

The quality of education that a university offers its students can be used to gauge the institution's success. By examining the data relating to redirection regarding student performance, the best level of quality in the educational system is attained. The lack of a framework in place to evaluate student performance and growth is still an issue today. This occurs frequently for two reasons. First off, it is difficult to anticipate students' success using the current approach. Second, due to the neglect of several important factors that are influencing students' performance. Due to the abundance of data in academic databases, predicting students' performance is a more difficult undertaking. This suggested technique can aid in more precise performance prediction for students.
For this proper data mining approach will be employed. This method involves applying a preprocessing step to the raw dataset in order to appropriately apply the mining algorithm. The performance prediction for the student can aid in improving performance.
Previous study [1] that we are referring to for this purpose used various algorithms like Decision Tree, Random Forest, Nave Bayes, Multilayer Perceptron, and JRip with Wrapper Subset Feature Selection. The highest accuracy was found for Random Forest which was that of 96.05% for the Portuguese Dataset after applying WSM method. In our study we have used different preprocessing methods and algorithms and are obtaining better accuracies in doing so.

Keywords - Education, student, performance, data mining, pre-processing, database, prediction

I. INTRODUCTION

Higher education's academic community faces a challenge in raising students' academic performance. Engineering and science students' academic performance in their first year of college is a turning point in their educational path and typically has a significant impact on their General Point Average (GPA). The evaluation criteria for the students, such as midterm and final exams, assignments, and lab work, are examined. Before the final test is given, it is advised that the class teacher be informed of all this related material. The results of this study will assist teachers in raising student achievement and significantly lowering the dropout rate.

In this research, we provide a hybrid method based on Decision Tree of Data Mining and Data Clustering that enables academics to forecast student GPA (SGPA, CGPA), and based on that, instructors can take the necessary steps to enhance student academic performance.

A frequently used measure of academic performance is the grade point average (gpa). Many universities have a minimum GPA requirement that must be met. As a result, the academic planners continue to use grade point average as their primary indicator of academic achievement. Throughout their time in college, a student's ability to achieve and maintain a high GPA that accurately reflects their overall academic achievement may be hampered by a variety of issues.

The faculty members could focus on these elements while creating methods to enhance student learning and enhance their academic success through tracking the development of their performance. The crucial qualities for future prediction can be found using the data mining technique's such as clustering algorithm and decision tree. The technique of extracting previously undiscovered, reliable, strategically relevant, and concealed patterns from big data sets is known as data clustering.

II. BACKGROUND

It is important for us to know what are the factors that overall affects the performance of the students. Thus, our motive was to consider all the factors which may or may not affect the grades of a student.

III. LITERATURE SURVEY

The literature survey is as follows

Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance (2022) [1]

The results of this study show accuracy rates of three well-known classification algorithms, including Decision Tree, Random Forest, Nave Bayes, Multilayer Perceptron, and JRip, which were evaluated. Using the wrapper feature subset selection technique, classification performance was enhanced. Preprocessing procedures on the dataset, such as categorizing the final grade field into a fine and two groups, boosted the proportion of accurate categorization predictions. The wrapper attribute selection process considerably improved the precision of all methods. The binary class technique was found to be more accurate in both mathematical and Portuguese datasets.

Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis (2021) [2]

This paper provides a systematic review of the SPP (Student Performance Prediction) study from the perspective of machine learning and data mining. This review partitions SPP into five stages, i.e., data collection, problem formalization, model, prediction, and application. SPP makes great sense to aid all stakeholders in the educational process. For students, SPP could help them choose suitable courses or exercises and make their plans for academic periods. This study also divides data from both online, offline and blending courses while taking into consideration historical performance data and background information. The problem is later formalized into machine learning problems of clustering, classification and regression. While discussing future works and current ethical data requirements they conclude this paper by mentioning that current studies are limited in statistical methods or educational theory, while it does not attract attention to using the popular techniques, i.e., feature learning.

Students' Class Performance Prediction Using Machine Learning Classifiers (2021) [3]

This paper implements various machine learning classification techniques on students' academic records for results predication. For this purpose, data of MS(CS) students were collected from a public university of Pakistan through their assignments, quizzes, and concessional marks. The WEKA data mining tool has been used for performing all experiments namely, data per-processing, classification, and visualization. For performance measure, classifier models were trained with 3- and 10-fold cross validation methods to evaluate classifiers' accuracy. The results show that bagging classifier combined with support vector machines outperform other classifiers in terms of accuracy, precision, recall, and F-measure score. The obtained outcomes confirm that their research provides significant contribution in prediction of students' academic performance which can ultimately be used to assists faculty members to focus low grades students in improving their academic records.

Data Mining for Student Performance Prediction in Education (2020) [4]

In this paper they point out that recently online systems in education have increased, and student digital data has come to big data size. And this makes possible to draw rules and predictions about the students by processing educational data with data mining techniques.

In this study, the successes of the students at the end of the semester are estimated by using the student data obtained from secondary education of two Portuguese schools. There are basically three data mining methods used in this study: *classification, clustering,* and *association rule mining*. In this study, three well-known 2 Data Mining - Methods, Applications and Systems classification algorithms (decision tree, random forest, and naive Bayes) were employed on the educational datasets to predict the final grades of students.

Wrapper feature subset selection method was used to improve the classification performance. Pr-processing operations on the dataset, categorizing the final grade field into five and two groups, increased the percentage of accurate estimates in the classification. The wrapper attribute selection method in all algorithms has led to a noticeable increase in accuracy rate. Overall, better accuracy rates were achieved with the binary class method for both mathematics and Portuguese data-set.

Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems (2020) [5]

This study focuses on ways to support universities in admissions decision making using data mining techniques to predict applicants' academic performance at university. A data set of 2,039 students enrolled in a Computer Science and Information College of a Saudi public university from 2016 to 2019 was used to validate the proposed methodology. In this study they developed four prediction models by applying four well-known data mining classification techniques, namely: Artificial Neural Network (ANN), Decision Tree, Support Vector Machine (SVM), and Naive Bayes. The results also showed that Scholastic Achievement Admission Test score is the pre-admission criterion that most accurately predicts future student performance. Thus, helping them conclude that this score should be assigned more weight in admissions systems.

Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques (2020) [6]

In this study the believe that analyzing the footprints left behind from these online interactions is useful for understanding the effectiveness of this kind of learning. This study was carried out with 772

examples of students registered in e-commerce and e-commerce technologies modules at an HEI. The study aimed to predict student′s overall performance at the end of the semester using video learning analytics and data mining techniques. A supervised data classification technique was used to determine the best prediction model that fit the requirements for giving an optimal result. For analysis, the same set of classification algorithms, performance metrics and the 10-fold cross-validation method were used. Two modules were selected for the study based on the similarity of the course content.

Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation (2020) [7]

This paper develops a novel approach called hybrid educational data mining model (HEDM) for analyzing the student performance for effectively enhancing the educational quality for students.

The proposed model evaluates the student performances based on distinctive factors that provide appropriate results. Furthermore, the model combines the efficiencies of Naive Baye's classification technique and J48 Classifier for deriving the results and categorizing the student performance in precise manner. The model is evaluated with the benchmark education dataset that is available online in the WEKA environment. The results show that the proposed model outperforms the results of existing works in evaluating student performance in EDM.

Data Mining for Student Advising (2020) [8]

This paper illustrates how to use data mining techniques to help in advising students and predicting their academic performance. Data mining is used to get previously unknown, hidden and perhaps vital knowledge from a large amount of data. The key importance of this project is that it discusses different data mining techniques in the literature review to study student behaviour depending upon their performance. They have tried to identify the most suitable algorithms from the existing research methods to predict the success of students. In this paper, the J48 algorithm was applied to the data set, gathered from Umm Al-Qura University in Makkah. Decision Tree, Neural Networks, Naïve Bayesian Classification, Support Vector Machines, and K-Nearest Neighbour are common algorithms used to classify data in this study. The data was later divided and used to learn patterns on the bases pof many different parameters like gender, age, faculty etc.

Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining (2020) [9]

This study used 4413 student instances of two datasets: students' information system and e-learning from the Faculty of Engineering in a Malaysia university for First Semester 2017/2018. The research empirically analyzes five types of ensemble classifiers and seven sampling techniques. The experimental results show a hybrid technique ROS with AdaBoost produces the most excellent performance compared to the other benchmark techniques. SMOTEENN technique with ensembles classifiers consistently produces high results. This study is using five different techniques of ensemble learning. By default, all ensemble learning classifiers are using the Decision Tree method as a base algorithm. This study used four classification performance metrics to evaluate the performance of machine learning models. Those are accuracy, precision, recall and F1 measure. The experiment result shows a training model using the XGBoost classifier and SIS data produce the highest accuracy. However, accuracy is not suitable when the classes are imbalance since it does not distinguish correct classified data into multi-class. This study focuses on F-Measure that is mean of precision and recall.

Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student (2020) [10]

The data used in this research is collected from Federal Board of Intermediate and Secondary Education Islamabad Pakistan, there are 7 regions in FBISE i.e., Punjab, Sindh, Khyber Pakhtunkhwa, Balochistan, Azad Jammu and Kashmir and overseas. The aims of this work is to analyze the education quality which is closely tightened with the sustainable development goals. The implementation of the system has produced an excess of data which must be processed suitably to gain more valuable information that can be more useful for future development and planning. In our proposed methodology, the obtained data is preprocessed to improve the quality of data, the labeled student historic data (29 optimal attributes) is used to train decision tree classifier and regression model. The obtain results show the effectiveness and importance of machine learning technology in predicating the students performance. The presented work is a student marks and grade prediction system using supervised machine learning techniques, the system is developed on the historic performance of students.

Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses (2017) [11]

Evandro B. Costa et al. presents a comparative study on the effectiveness of educational data mining techniques to early predict students likely to fall in introductory programming courses. The main goal is to investigate the effectiveness of such techniques to identify students likely to fail at early enough stage for action to be taken to reduce the failure rate and analyze the impact of data pre-processing and algorithms fine-tuning tasks, on the effectiveness of the mentioned techniques.

Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout (2018) [12]

Concepción Burgos et al. proposed the use of knowledge discovery techniques to analyze historical student course grade data in order to predict whether or not a student will drop out of a course. Logistic regression models and predictive models were used to reduce the dropout rate in e-learning courses.

Early segmentation of students according to their academic performance: A predictive modelling approach (2018) [13]

V.L. Miguéis et al. proposed that the early classification of university students according to their potential academic performance can be a useful strategy to mitigate failure, to promote the achievement of better results and to better manage resources in higher education institutions. It used a two-stage model which are supported by data mining techniques, that used the information available at the end of the first year of students' academic career to predict their overall academic performance. Moreover, it proposed to segment students based on the dichotomy between the evidence of failure or high performance at the beginning of the degree program, and the students' performance levels predicted by the model.

Predicting academic performance by considering student heterogeneity (2018) [14]

Sumyea Helal et al. proposed a study that creates different classification models for predicting student performance, using data collected. The predictive models were constructed by considering the student

heterogeneity because based on the observation that students with different socio-demographic features or study modes may exhibit varying learning motivations.

Early detection of university students with potential difficulties (2017) [15]

Anne-Sophie Hoffait et al. proposed an early detection of potential failure using student data available at registration, i.e., school records and environmental factors, with a view to timely and efficient remediation and/or study reorientation. It adapts three data mining methods, namely random forest, logistic regression and artificial neural network algorithms.

Data mining models for student careers (2015) [16]

Renza Campagni et al. proposed a different approach based on clustering and sequential patterns techniques in order to identify strategies for improving the performance of students and the scheduling of exams. They introduced an ideal career as the career of an ideal student which has taken each examination just after the end of the corresponding course, without delays which is then compared to the career of a generic student with the ideal one by using the different techniques just introduced. Finally, they apply the methodology to a real case study and interpret the results which underline that the more students follow the order given by the ideal career the more they get good performance in terms of graduation time and final grade

Classification and prediction based data mining algorithms to predict slow learners in education sector (2015) [17]

Parneet Kaur et al. proposed a paper which r focus on identifying the slow learners among students and displaying it by a predictive data mining model using classification-based algorithms. The dataset of student academic records is tested and applied on various classification algorithms such as Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree using WEKA an Open-source tool. This paper showcases the importance of Prediction and Classification based data mining algorithms in the field of education and also presents some promising future lines.

Educational Data Mining Techniques and their Applications (2015) [18]

John Jacob et al. paper elaborates a study on various Educational Data Mining techniques and how they could be used for the benefit of all the stakeholders in the educational system. Correlation is used to see if a variation in one variable results in a variation in the other. Decision trees give possible outcomes and are used to predict students' performance in this study. Regression analysis is used in the construction of a model involving a dependent variable and multiple independent variables; if the model is satisfactory, then the value of dependent variable is determined using the values of the independent variables. Clustering finds groups of objects so that objects that are in a cluster are more like each other than to objects in another cluster, helping in arranging items under consideration; clustering would help in analyzing the job profiles that would be suited for each student.

Educational Data Mining & Students' Performance Prediction (2016) [19]

Amjad Abu Saa emphasizes on the importance of educational data especially students' performance. This paper is concerned with the students' performance and explores multiple factors theoretically

assumed to affect students' performance in higher education and finds a qualitative model which best classifies and predicts the students' performance based on related personal and social factors. The author uses Educational Data Mining (EDM).

Educational Data mining for Prediction of Student Performance Using Clustering Algorithms (2014) [20]

M. Durairaj et al. took the student details from there college for analysis and data mining methods have been employed to get vital information. There work aims to develop a trust model using data mining techniques which mines required information, so that the present education system may adopt this as a strategic management tool. They are able to predict the students' performance and their pass percentage with some degree of accuracy.

A Comparative Study of Predicting Student's Performance by use of Data Mining Techniques Education (2017) [21]

Data Mining (EDM) has boomed in the educational systems recently. EDM enables us to analyze and predict student performance so that measures can be taken in advance. This paper presents a comparative study of various EDM techniques, classification algorithms and their impact on datasets. EDM involves analysis and improvement in the prediction methods of student performance. Data mining techniques and classification algorithms are applied to get deeper insights and predictions.

Supervised data mining approach for predicting student performance (2019) [22]

Data mining applications can help the academic management systems to investigate and identify groups of excellent students and groups of dropped out students from the university. Currently, studies on existing prediction methods are still insufficient to identify the most suitable methods to predict student's achievement in particular courses. With the accurate data mining techniques prediction algorithms can help to identify the most important attributes in contributing to a student's performance. Higher institutions can gain deep and thorough knowledge to enhance their lesson plan, assessment, evaluation planning and decision-making based on the findings obtained.

Student's Performance Prediction using Deep Learning and Data Mining Methods (2019) [23]

Data Mining is the most prevalent technique to evaluate students' performance and is extensively used in the educational sector. EDM is a methodology or like a procedure used to mine valuable information and patterns or forms from a massive educational database. The prime motto of study is to discover the performance of students using some classification techniques and discovering the best one which yields optimal results.

Performance Analysis and Prediction Student Performance to Build Effective Student Using Data Mining Techniques (2019) [24]

In this paper, data mining techniques have been applied to construct a classification model to predict the performance of students. Data mining is a developing capable tool for examination and expectation. It is effectively applied in the field of fraud detection, marketing, promoting, forecast, and loan assessment. However, it is an incipient stage in the area of education. In this period of

computerization, schooling has additionally remodeled itself and is not restrained to old lecture technique. These days, masses of data are gathered in educational databases; however, it stays unutilized.

Predicting student performance of different regions of Punjab using classification techniques (2018) [25]

Data mining techniques can provide solutions and smooth functioning of data mining in the best possible manner. In this study, an algorithm is applied on students of different colleges of Punjab to predict their performance. This algorithm is predicted whether the student is going to pass or fail in the final examination. If the outcome of this test is predicted as failure then extra earlier effort can be provided to the student which will improve his result. Data Mining algorithms such as decision trees, neural networks and naïve bayes classification are very useful in the field of marketing, medicine, real estate, customer relationship management, financial management etc.

Student performance prediction based on data mining classification techniques (2018) [26]

The process of predicting student performance has become a crucial factor in the academic environment and plays a significant role in producing quality graduates. Several statistical and machine learning algorithms have been proposed for analyzing, predicting and classifying student performance. This paper presents a method to predict student performance using Iterative Dichotomiser 3 (ID3), C4.5 and Classification and Regression tree (CART).

Review on Predicting Students' Graduation Time Using Machine Learning Algorithms (2019) [27]

The ability of data mining to obtain meaningful information from meaningless data makes it very useful to predict students' achievement, university's performance, and many more. According to the Department of Statistics Malaysia, the numbers of students who do not manage to graduate on time rise dramatically every year. Findings of this research confirmed the usefulness of Neural Network and Support Vector Machine as the most competitive classifiers compared with Naïve Bayes and Decision Tree.

Predicting Student Performance Using Data Mining (2018) [28]

Supporting the goal of higher education to produce graduates who will be professional leaders is crucial. Most universities implement intelligent information systems to support their vision and mission. One of the features of the Intelligent Information System is student performance prediction. This feature could accurately predict the student' grade for their enrolled subjects and identify students at risk in failing a course. EDM is concerned with developing and applying data mining methods to detect patterns in large amounts of educational data. Data mining is a tool to improve the quality of education by identifying the students who are at risk in their study.

Using of data mining techniques to predict of student's performance in industrial institute of Al-Diwaniyah, Iraq (2019) [29]

The aim of this paper is to show the benefits of the educational data mining (EDM) techniques, in order to understand the factors which, lead to technical student's success and failure. We use the

individual data of 311 students and their grades that were collected in the Industrial Institute of Al-Diwaniyah city (Iraq) during 2015–2017 academic years. Using Microsoft SQL Server Business Intelligence Development Studio 2012 platform and based on Cross Industry Standard Process for Data Mining, we prepare of 13 nominal and numerical attributes for each student.

Educational data mining for students' performance based on fuzzy C-means clustering (2019) [30]

A study presents a fuzzy C-means clustering algorithm using 2D and 3d Clustering to evaluate students' performance based on their examination grades from College of Computer Science and Technology, Huaqiao University for students enrolled in 2014. Based on the experimental results, the researchers can better understand students' performances and build a pedagogical basis for decisions. Students can also receive some recommendations from the mining results about their performance.

Table 1: Literature review

S.No.	Title of the Paper and the year	Algorithms Used	Data set used	Performance Measures	Scope for future work
1.	Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance (2022)	Decision Tree, Random Forest, Nave Bayes, Multilayer Perceptron, and JRip	Two Portuguese secondary schools data, same as our project	Naïve Bayes generated the best results in the binary label dataset, as it did in the multiple-grade mathematics dataset. It went from 93.49 % to 94.10%.	The datasets can be subjected to a variety of categorization methods, different methods of feature selection may be employed
2.	Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis (2021)	Decision Trees, Linear Regressions, Support vector machines, Matrix factorization, Collaborative filtering, Artificial neural network, Deep Learning.	Composed of 1,325 students, and 832 courses, a typical higher education in China.	This work provides developments and challenges in the study task of SPP and facilitates the progress of personalized education.	Integrating these models into the expert system and using the conclusion guide the teaching procedure.

3.	Students' Class Performance Prediction Using Machine Learning Classifiers (2021)	J48, ID3, Naive Bayes, IB1, and OneR, implemented using WEKA explorer tool.	data of MS(CS) students from a Pakistan university through assignments, quizzes etc.	Naive Bayes algorithm and BayesNet algorithm achieved 68.33% and 65.28% on 3-cross and 10-Cross fold validation respectively.	The selection of more significant attributes from academic records which might affect students' class performance from academic records.
4.	Data Mining for Student Performance Prediction in Education (2020)	Decision Tree, Random Forest, and Naïve Bayes.	Related to mathematics lesson and Portuguese language lesson.	Accuracy rates have changed positively in all trials using wrapper subset attribute selection method.	Different feature selection method, different classification algorithms can also be utilized.
5.	Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems (2020)	Three EDM techniques (Naïve Bayes, ANN, and Decision Tree).	Data set of 2,039 students enrolled in a Computer Science and Information College of a Saudi public university.	Artificial Neural Network technique has an accuracy rate above 79% making it superior to the other ways.	Further studies are needed to consider more pre-admission factors that affect future student performance, such as student personality.
6.	Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques (2020)	Decision Tree, Random Forest, Naïve Bayes, Support Vector Machines, Linear Regression or Logistic Regression models, and K means.	772 examples of students registered in e-commerce and e-commerce technologies modules at an HEI.	Random Forest accurately predicted successful students at the end of the class with an accuracy of 88.3% with an equal width and information gain ratio.	A dashboard with data representations from these virtual learning environments would help in projecting the students' performance and interactions.
7.	Towards developing hybrid	Develops a hybrid educational data	Comprised of 2344 student instances, from various	The variant of Naive Bayes, and J48 algorithms provided 69.96% and 66%	Could consider the combination of EDM techniques and

	educational data mining model (HEDM) for efficient and accurate student performanc e evaluation (2020)	mining model (HEDM) combining J48, Naive Bayes.	courses in a higher educational university.	accuracy. Together the hybrid model achieved higher accuracy than other models.	educational theories and priors.
8.	Data Mining for Student Advising (2020)	Decision Tree, Neural Networks, Naive Bayes, SVM, J48.	Data set, gathered from Umm Al-Qura University in Makkah.	Accuracy of the prediction is high because it has identified that GPA and extra tests are not the only factors that affect the final results of the student.	More data sets by removing private information might be created and opened to use for this research field in the future.
9.	Improved students' performanc e prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining (2020)	Decision tree, Random Forest, Bootstrapping, Boosting, AdaBoost, Gradient Boosting, XGBoost.	4413 student instances of two datasets; students' information system and e-learning from the Faculty of Engineering in a Malaysia university.	The highest accuracy rate achieved s that of 71.8% for the AdaBoost. However, SMOTEENN frequently obtains the highest results among all imbalance techniques employed.	Need to use cross-domain data to have a broad overview of the benchmark, focus more on the hybrid of ensemble techniques with different base-classifiers, with more hypertuning parameters.
10.	Application of machine learning and data mining in predicting the performanc e of intermediat e and secondary education level student	(29 optimal attributes) is used to train decision tree classifier and regression model.	Collected from Federal Board of Intermediate and Secondary Education Islamabad Pakistan.	Accuracy of the classification system in predicting the grade and the regression model in predicting the marks is high.	More combinations of hybrid techniques to solve the multi-class classification problem.

	(2020)				
11.	Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses (2017)	Educational Data Mining techniques – SVM, Decision Tree, Neural Network and Naïve Bayes, Information Gain Algorithm, SMOTE Algorithm, J48 Algorithm, Multilayer Neural Network.	The first data source contains information about 262 undergraduate students that took the introductory programming course performed in a distance education. The second data source contains information about161students that took the introductory programming course performed on-campus.	The techniques present an effectiveness that varies from 0.55 to 0.82 in the distance education course, and from 0.50 to 0.79 in the on-campus course. The preprocessing on the distance education data was able to increase the effectiveness of most of the techniques, but the preprocessing on the on-campus data did not significantly impact the effectiveness of the techniques.	This study can be improved by considering other data sources from different universities as well as the use of other techniques of data preprocessing and algorithms fine-tuning.
12.	Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout (2018)	Regression Modelling – Training Algorithm and Classification Algorithm, Feed-Forward Neural Network, Support Vector Machine, (Probabilistic Ensemble Simplified Fuzzy ARTMAP –Adaptive Resonance Theory Mapping PESFAM) a Classifier, a System for Educational Data Mining.	Considered the historical data for students of five different BS in Computer Engineering programme courses, which were taught during the 2013/14 academic year at Madrid Open University (UDIMA). These courses, composed of 10 teaching unit.	Proved to output slightly better results than existing proposals in terms of accuracy, especially in the crucial weeks of the semester. The resulting models, combined with the special tutoring action plan, proved to be very useful for reducing the dropout rate during the 2014/15 academic year for the analyzed courses compared with previous academic years.	It would be better for the mechanism to be activated automatically.
13.	Early segmentation of students according to their	Binning Algorithm, Random Forests, Decision Trees, Support Vector	The study uses secondary data from an engineering and technology school, belonging to a	The results reveal that there is statistically significant difference between the models (significance level of 95%), with the exception of adaptive Boosting -	It is crucial to integrate the proposed method on the platforms already used by the institutions' educational decision makers, such

	academic performance: A predictive modelling approach (2018)	Machines, Naïve Bayes, Bagged Trees, Adaptive Boosted Trees.	European public research University. The data refers to the student information for those enrolled between 2003 and 2007, i.e., 2459 students. It encompasses all academic information obtained until either the conclusion of their degree, or until the academic year 2014/2015.	decision trees and both decision trees and naive Bayes; random forests and decision trees; and SVM and naive Bayes.	as programme directors and committees, to support the development of early and appropriate educational measures, targeting the specific segmented groups of students.
14.	Predicting academic performance by considering student heterogeneity (2018)	Naïve Bayes, SMO, J48, JRip.	The datasets used in this paper were collected from 2011 to 2013 from a division (akin to a faculty comprising multiple disciplinary schools) in an Australian university regarding their first-year domestic undergraduate students.	This study demonstrates that the results derived from the submodels produce a higher degree of accuracy than the base model.	Generate student profiles using unlabeled data to discover interesting student clusters and their characteristics. It would be very useful to consider the combined features for a particular module and the categorization features (e.g. social and information) in terms of student participation in LMS activities. Identify the risk indicators of international students, as they may possess some different features from domestic students, such as diverse ethnic origins, funding opportunities, native languages and other factors.
15.	Early detection of university students	Logit Regression, Artificial Neural Network,	Real data pertaining to the University of Liège. Final sample	On a real data set, they are now able to identify with a high rate of confidence (90%) a subset of 12.2% of	The rate of correct prediction can be increased. Accuracy can be increased with a

	with potential difficulties (2017)	Decision Trees and Random Forests.	counts 6845 students.	students facing a very high risk of failure.	larger class size.
16.	Data mining models for student careers (2015)	K-Means Clustering Algorithm, Sequential Pattern Mining.	The dataset is based on 141 graduated students, enrolled for the first time at the degree course in Computer Science at the University of Florence (Italy) from 2001–2002 up to 2007–2008 academic years.	The usage of sorting, clustering and pattern analysis have given the insightful information of the database, on how a particular student is and how their grades affect the performance.	A possible development of this research could be the identification of students at risk of dropping out by using a methodology similar to that proposed in this paper. It could compute the ideal career sorting the courses relative to the same semester according to the preference of students.
17.	Classification and prediction based data mining algorithms to predict slow learners in education sector (2015)	Multilayer Perception, Naïve Bayes, SMO, J48, REPTree, Regression and Density Estimation	A record 152 students of high school is used as dataset.	Multi Layer Perception technique performs best with accuracy 75%. Multi Layer Perception performs best among all classifiers with F-Measure 82%.	Integration of data mining techniques with DBMS and Elearning techniques is merged together on different datasets to find accuracy and predictions of desired results.
18.	Educational Data Mining Techniques and their Applications (2015)	K – Means Clustering, Multiple Regression, J48.		The accuracy in predicting GPA's of students, predictions were found to be correct 80% of the time when compared with actual results.	Lowering the costs to store logged data and cost associated with hiring staff dedicated to managing data systems. Integrating multiple data systems with the support of statistical and visualization tools, creating one simplified version of the data.
19.	Educational Data Mining & Students' Performanc	Decision Tree (C4.5, CART, CHAID, ID3), Naïve Bayes Classification.	This study was conducted on a group of students enrolled in different colleges in Ajman University of	CART had the best accuracy of 40%, which was significantly more than the expected (default model) accuracy, CHAID and C4.5 was next with	It would be possible to do more data mining tasks on it, as well as, apply more algorithms. It would be interesting to apply association

	e Prediction (2016)		Science and Technology. The initial size of the dataset is 270 records.	34.07% and 35.19% respectively, and the least accurate was ID3 with 33.33%. The Naïve Bayes classifier was able to predict the class of 95 objects out of 270, which gives it an Accuracy value of 36.40%.	rules mining to find out interesting rules in the students' data. It could be better if the data was collected as part of the admission process of the university, that way, it would be easier to collect the data, as well as, the dataset would have been much bigger, and the university could run these data mining tasks regularly on their students to find out interesting patterns and maybe improve their performance.
20.	Educational Data mining for Prediction of Student Performanc e Using Clustering Algorithms (2014)	K – Means Clustering Algorithm, Naïve Bayes Clustering, Naïve Bayes Probabilistic Algorithm.	The dataset contains students' details of different subject marks in semester wise have been recorded and subjected to the data mining process.	The Naive bayes algorithm gives more accurate results than decision tree. The results are predicted within 0 seconds.	
21.	A Comparative Study of Predicting Student's Performance by use of Data Mining Techniques (2017)	Naïve Bayes, Decision tree, neural networks, outlier's detections and advanced statistical techniques.	Data sets of 1547 record used to predict the performance, longitudinal data derived from Gwinnett County Public Schools data,	Decision tree algorithm is 88%. naive Bayes provides 75%. K-nearest neighbor provides 83%	Will implement in future by use of real datasets of Fast University and take the student's attributes. We will find more efficient techniques based on other execution measure like recall and other in future.
22.	Supervised data mining approach for predicting student performance (2019)	Decision Trees, Naïve Bayes, K-Nearest Neighbour, Logistic Regression.	Collected 631 transcripts from 2013 to 2016.	The Information Gain algorithm has about 82.15% for the 10-fold cross validation testing, Gini Decision Tree algorithm is 80.15%. The precision of Information Gain is (78.61%) and non-Excellent class (83.78%) compared to GINI with 75.25% and 83.78% for Excellent and	Some high influence attribute to predict student performance can be considered by the university to plan further action for improvement. The study can be further extended to predict student's performance of other courses using other

				Non-Excellent class respectively.	attributes.
23.	Student's Performance Prediction using Deep Learning and Data Mining Methods (2019)	KNN, Naïve Bayes and Kappa-statistic.	Dataset was collected from an educational institute of Saudi University. Academic data set consisting of 473 instances, and found that 70% accuracy was yielded by Bayesian classifier.	MLP, Decision trees and Random Forest with maximum accuracies of 99.45%, 99.81% and 100%.	MLP technique is more efficient compared to other technique in prediction of students' performance. Rules can be mined, and accuracy needs to be improved in SVM, K-NN as part of the future work.
24.	Performance Analysis and Prediction Student Performance to build effective students Using Data Mining Techniques (2019)	Naïve Bayes classifiers and decision trees.	The survey was filled by 130 students, from the first, the second, the third institutions, and the rest from a few different institutions using the net questionnaire.	Efficiency percentage ranges about 36%–45%, which are low percentages.	it is recommended to gather more appropriate data from several universities and to have the right performance rate for students. the software could be created to be used by the universities and institutions, including the rules generated for foreseeing the performance of students.
25.	Predicting Student performance of different regions of Punjab using classification techniques (2018)	Decision trees, neural networks and naïve bayes classification.	Taken from malva region of Punjab.	The Naive bayes algorithm gives more accurate results than decision tree.	Many factors may affect the students' performance and if that has been observed properly in advance, ways can be suggested to improve it.
26.	Student performance prediction based on data mining classification techniques (2018)	ID3, C4.5 and CART algorithms.	The data set used in this research was obtained from private University in Northern part of Nigeria. Initially the size of the data was 234.	The Simple CART tree algorithm used gave a prediction accuracy of 98.3%, incorrectly classification instance of 1.70%, recall of 98.3%, specificity of 98.3%, precision of 98.4%, F-measure of 98.3%, and the time taken 0.58sec sequentially.	Future work would be to investigate other machine learning algorithms to predict the student performance and extending the coverage of the dataset used in this paper.
27.	Review on Predicting	Decision Tree, Naïve Bayes,	Five major departments at	SVM scored the second highest of the prediction	The impact of this work laid in intention to help

	Students' Graduation Time Using Machine Learning Algorithms (2019)	support Vector Machine, Neural Network.	California State University Northridge.	accuracy which is 93.95%. NN has the highest prediction accuracy which is 95%.	and assist other researchers in developing a real model that can predict students' graduation time easily and accurately.
28.	Predicting Student Performance Using Data Mining (2018)	Multi-regression model.	Dataset extracted from the Petra Christian University' Moodle. The dataset spans four semesters, and it contains 486 courses, 7,563 students, and 109,231 activities.	The RMSE of multi-regression model with Lentera features with one linear model is 0.17. On the other hand, the RMSE of single regression model is 0.3. By accompanying student-bias term and course-bias term, multi regression model could better capture student performances in their course.	Lentera interaction features could improve the accuracy of prediction of student performance.
29.	Using of data mining techniques to predict of student's performance in industrial institute of Al-Diwaniyah, Iraq (2019)	Association rules detection, clustering, classification, Anomaly detection.	The dataset divided into two samples, the first sample of the dataset consist of 70% for 218 students to represent the modules for training algorithms, as for the second sample consist of 30% for 93 students to testing algorithms.	The usage of sorting, clustering and pattern analysis have given the insightful information of the database, on how a particular student is and how their grades affect the performance.	Hope that further research in the field of EDM will help us to resolve the principal problems of computer systems of individual instruction
30.	Educational data mining for students' performance based on fuzzy C-means clustering (2019)	Fuzzy C-means clustering algorithm, FCM clustering algorithm.	The total number of students is 246. Derived from the student achievement management system of Huaqia Uni versity.	Performance levels in the experiments are based on the students' overall performance in the class or the college. Thus, it is not the same as the grade.	More sources and structures of educational data can be utilized for clustering. In addition, regression and classification algorithms like support vector machines and artificial neural networks can be applied to predict students' grades in the next examination.

Table 1 shows literature review for the 30 papers reviewed on this topic while discussing the dataset used, performance measures and scope of future work for each paper in a tabular form to aid with easier understanding

IV. DATASET DESCRIPTION & SAMPLE DATA

1. **DATASET INFORMATION**

The data was received from UCI Machine Learning Repository. The information about the dataset is below. This data is of student's achievement in secondary education of Portuguese school. The data attributes include student grades, demographic, social and school related features) and it was collected by using questionnaires and school reports. Dataset are provided regarding the performance in subject: Mathematics. The target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade, while G1 and G2 correspond to the 1st and 2nd period grades.

2. ATTRIBUTE INFORMATION

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

33 letter_grade - classification of G3 accordingly : 15-20 : A 10-15 : B 5-10 : C 0-5 : D

3. SAMPLE DATASET

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltim	studytim	failures	schoolsu	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	letter_grade
2	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	6	5	6	6	C
3	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	4	5	5	6	C
4	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no	yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10	B
5	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	yes	yes	yes	yes	yes	yes	3	2	2	1	1	5	2	15	14	15	A
6	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	yes	no	yes	yes	no	no	4	3	2	1	2	5	4	6	10	10	B
7	GP	M	16	U	LE3	T	4	3	services	other	reputatic	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	4	2	1	2	5	10	15	15	15	A
8	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11	B
9	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	6	6	5	6	C
10	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	yes	no	yes	yes	yes	no	4	2	2	1	1	1	0	16	18	19	A

Fig 1: Math Course Datasets

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	
1	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltim	studytim	failures	schoolsu	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	letter_grade
2	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	4	0	11	11	B
3	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	2	9	11	11	B
4	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	0	yes	no	no	no	yes	yes	yes	no	4	3	2	2	3	3	6	12	13	12	B
5	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	no	yes	yes	yes	yes	yes	3	2	2	1	1	5	0	14	14	14	B
6	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	no	no	yes	yes	no	no	4	3	2	1	2	5	0	11	13	13	B
7	GP	M	16	U	LE3	T	4	3	services	other	reputatic	mother	1	2	0	no	yes	no	yes	yes	yes	yes	no	5	4	2	1	2	5	6	12	12	13	B
8	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	13	12	13	B
9	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	2	10	13	13	B
10	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	no	no	yes	yes	yes	no	4	2	2	1	1	1	0	15	16	17	A

Fig 2: Portuguese Course Datasets

V. PROPOSED ALGORITHM WITH FLOWCHART

We have pre-processed the data by first checking for missing values and then performed variable transformation using two different techniques: One Hot Transformation and 0-1 Transformation (as shown in Fig 3 and Fig 4).
Next, we have checked for the outlier data and if present we have used suppression method with the help of interquartile range and to bring back the outlier data in the acceptable range.

FIRST LAST

reason	reason_course	reason_home	reason_other	reason_reputation
course	1	0	0	0
course	1	0	0	0
other	0	0	1	0
home	0	1	0	0
home	0	1	0	0

Fig 3: One Hot Transformation

FIRST LAST

Pstatus	Pstatus
A	0
T	1
T	1
T	1
T	1

Fig 4: 0-1 Transformation

We have also created a heat map to determine the correlation between various attributes.

Next, we had determined the distribution of the class label in both the Mathematics and Portuguese dataset using WEKA tool as shown here

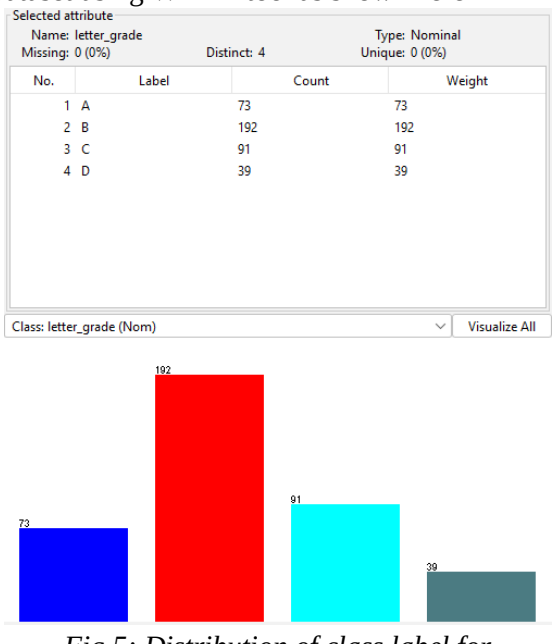

Fig 5: Distribution of class label for Mathematics Dataset

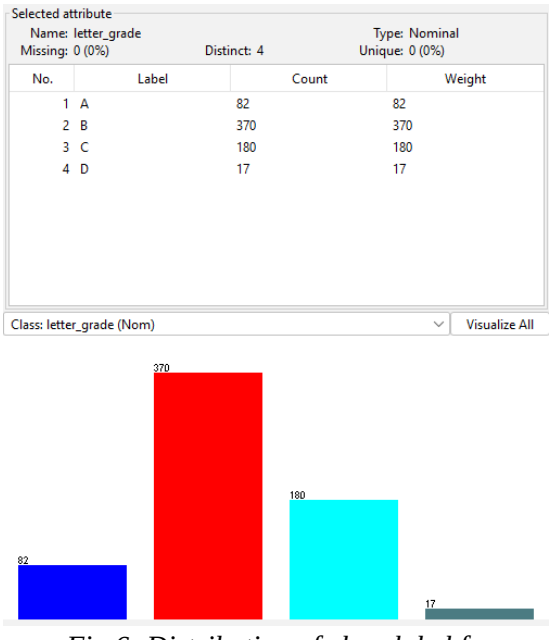

Fig 6: Distribution of class label for Portuguese Dataset

Next, we decided to apply PCA (Principle Component Analysis) on our dataset which resulted in the reduction of number of columns of 47 from that of the pre-processed model to 35 for the Mathematics dataset and from 47 to 36 columns for the Portuguese Dataset.

Now as *Fig 5 and Fig 6* indicate our data is imbalanced as the count for each class label in our dataset vary considerably from each other. So, in order to deal with this imbalance we have applied SMOTE to both the Mathematics and Portuguese dataset.

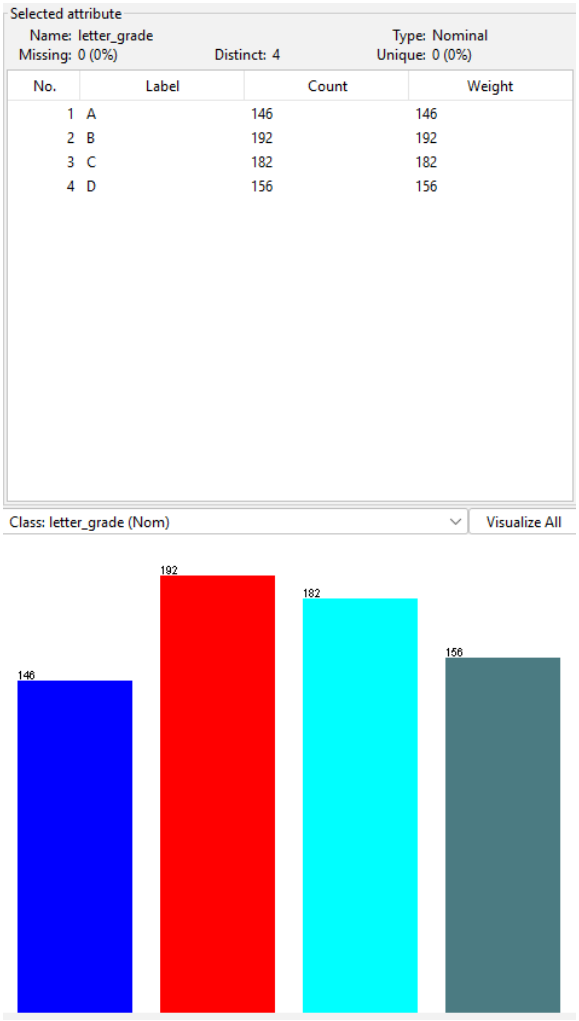

Fig 7: Mathematics Dataset after PCA and SMOTE

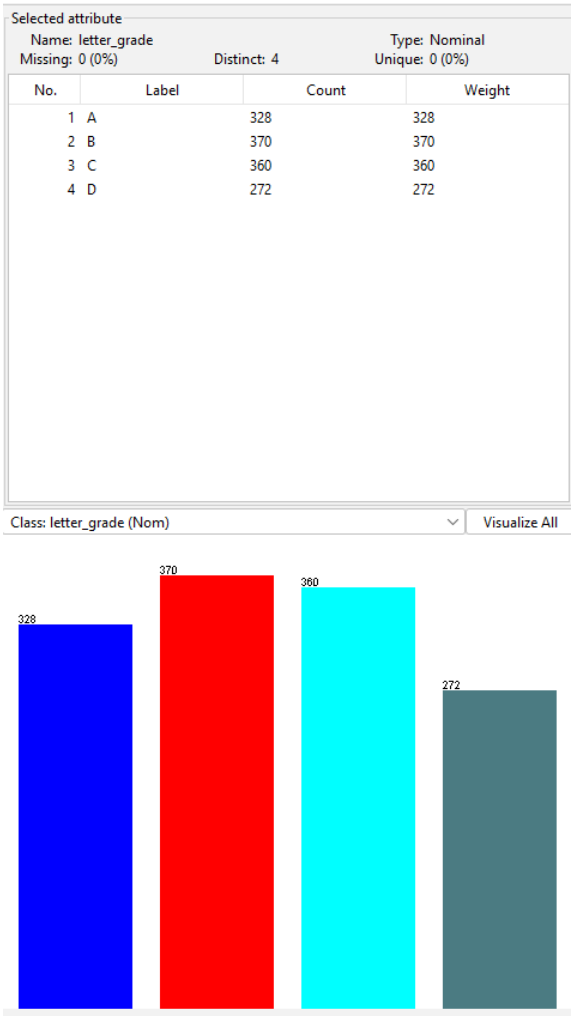

Fig 8: Portuguese Dataset after PCA and SMOTE

Figure 7 and Figure 8 show how both the datasets have changed after applying SMOTE to them using the WEKA tool. We observe that the imbalance previously observed is reduced significantly and the total number of entries in Mathematics and Portuguese Dataset have now become 676 and 1330, respectively.

The various algorithms used are as follows:

1. Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other.

2. K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. Here, the algorithm uses minkowski distance for n = 5 by default.

3. Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

4. Random forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. It selects the majority vote for classification and average in case of regression.

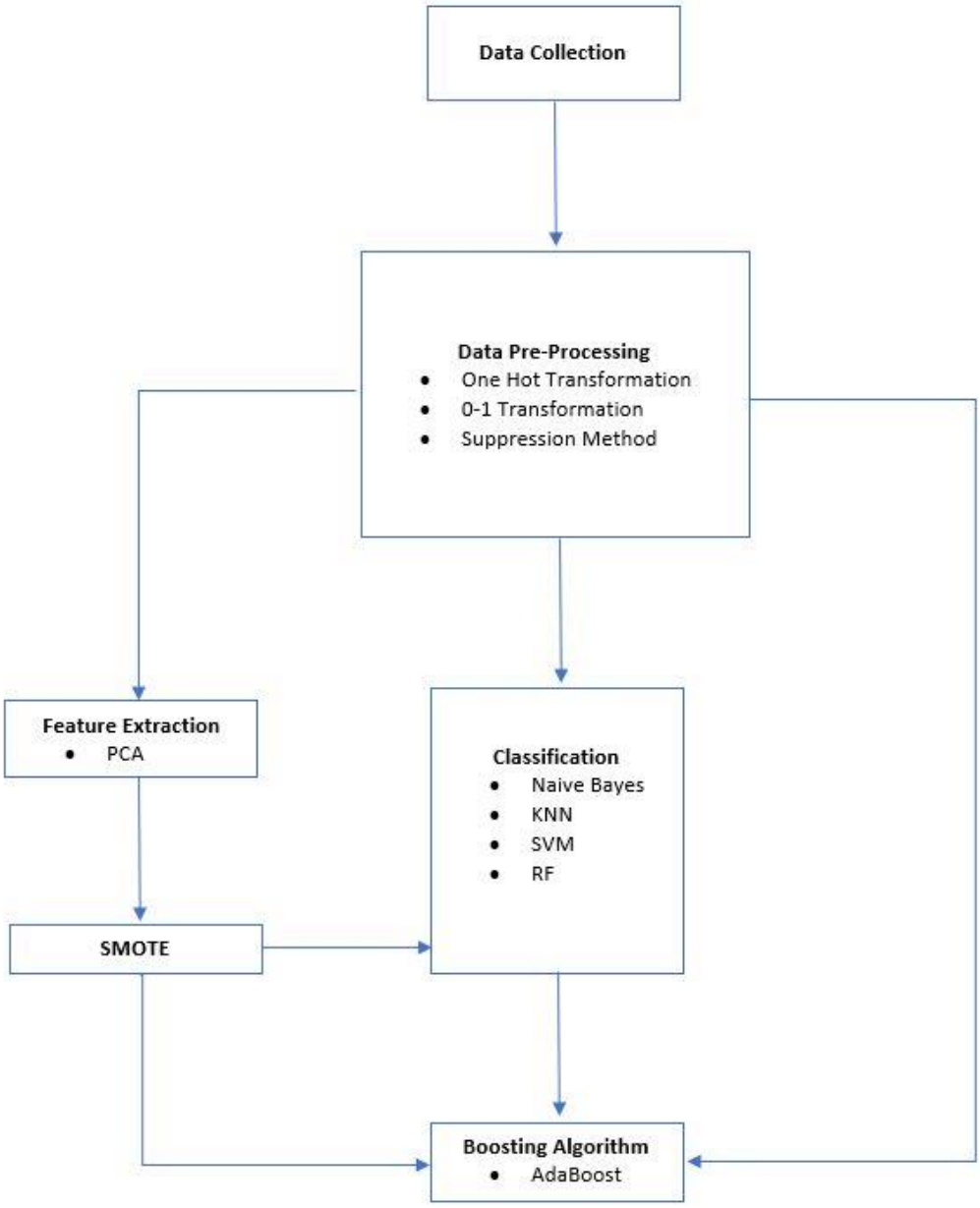

Fig 9: Flowchart of the Proposed Algorithm

VI. EXPERIMENTS RESULTS

In this section, the result is organized thus far into tables to better visualize the entire situation.

Table 2: Accuracy chart for the Mathematics course dataset

Classification Algorithm	Accuracy before PCA and SMOTE	Accuracy after PCA and SMOTE	Accuracy after PCA, SMOTE and AdaBoost	Accuracy before PCA, SMOTE with AdaBoost
Naïve Bayes	86.55%	72.90%	60.09%	51.26%
KNN	74.78%	29.55%	-	-
SVM	96.63%	87.68%	-	-
Random Forest	95.79%	87.19%	88.17%	96.64%

Table 3: Accuracy chart for the Portuguese course dataset

Classification Algorithm	Accuracy before PCA and SMOTE	Accuracy after PCA and SMOTE	Accuracy after PCA, SMOTE and AdaBoost	Accuracy before PCA, SMOTE with AdaBoost
Naïve Bayes	71.28%	79.69%	69.92%	62.56%
KNN	68.20%	48.12%	-	-
SVM	100.00%	91.22%	-	-
Random Forest	98.46%	90.72%	90.72%	98.97%

Table 4: Accuracy found when applying only AdaBoost

Dataset (Language)	Accuracy with only AdaBoost (only pre-processed)	Accuracy with PCA, SMOTE and AdaBoost only
Mathematics	100.00%	51.23%
Portuguese	100.00%	42.35%

Table 2, 3 and 4 shows us the accuracies obtained after running our model. From the above results, we can come to conclusion that Support Vector Machine (SVM) classifier works best in all the scenarios and Random Forest (RF) is able to achieve better accuracy after using AdaBoost algorithm.

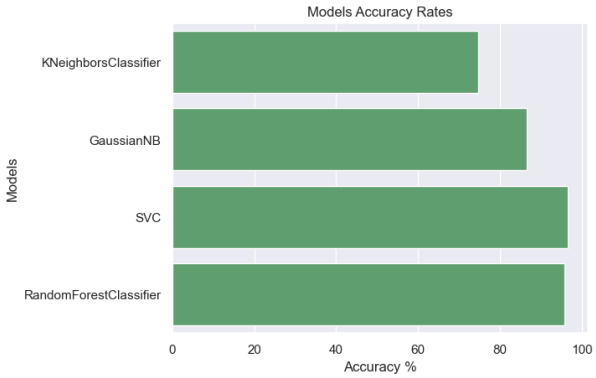

Fig 10: Mathematics (only pre-processed)

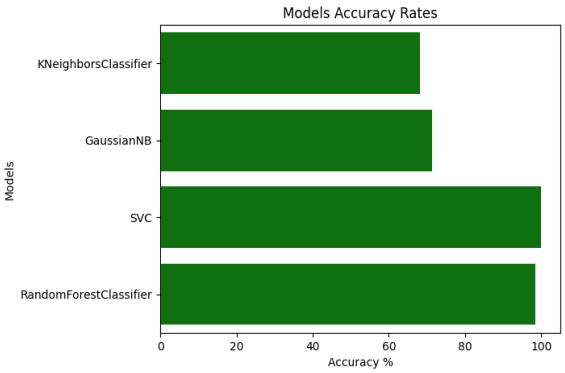

Fig 11: Portuguese (only pre-processed)

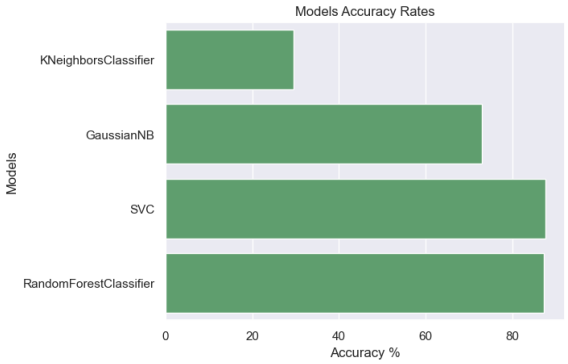

Fig 12: Mathematics (after PCA and SMOTE)

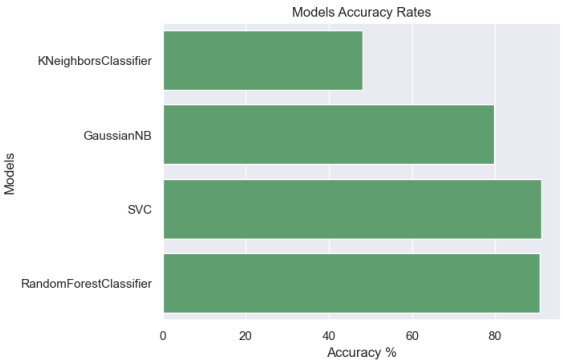

Fig 13: Portuguese (after PCA and SMOTE)

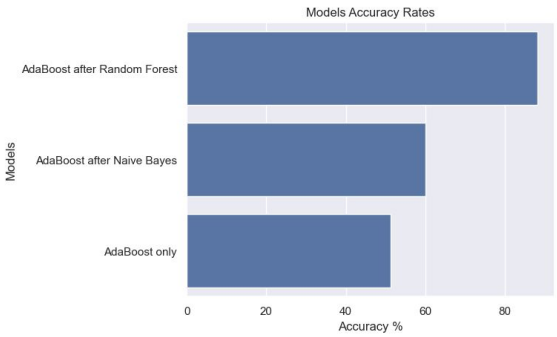

Fig 14: Mathematics (after PCA, SMOTE and AdaBoost)

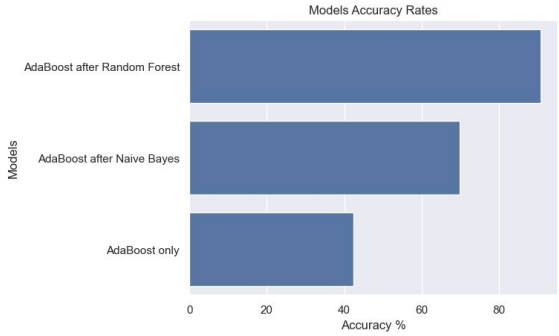

Fig 15: Portuguese (after PCA, SMOTE and AdaBoost)

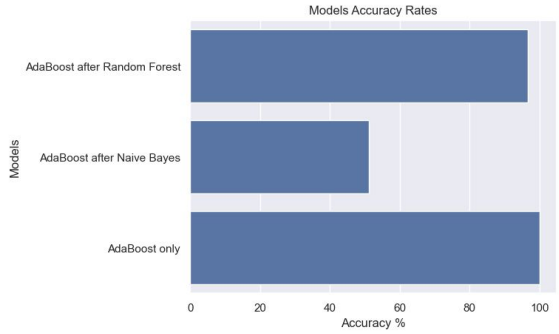

Fig 16: Mathematics (pre-processed and AdaBoost)

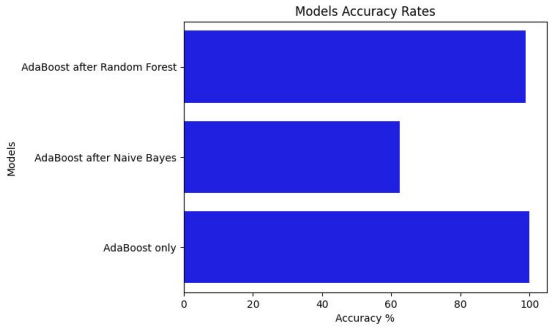

Fig 17: Portuguese (pre-processed and AdaBoost)

Fig 10 to Fig 17 is graphical representation of the above tables.

VII. COMPARATIVE STUDY / RESULTS AND DISCUSSION

The results as that in the base paper is as shown here

Classification Algorithms	Accuracy before by WSM	Accuracy after by WSM
Decision Tree Algorithm	89.11%	90.45%
Random Forest Algorithm	**93.49%**	94.10%
Naïve Bayes Algorithm	86.33%	89.50%
Multi-layer Perceptron Algorithm	85.67%	87.70%
JRip Algorithm	87.56%	89.76%

Fig 18: Maths course results from base paper [1]

Classification Algorithms	Accuracy before by WSM	Accuracy after by WSM
Decision Tree Algorithm	92.07%	92.76%
Random Forest Algorithm	**95.08%**	96.05%
Naïve Bayes Algorithm	88.74%	93.70%
Multi-layer Perceptron Algorithm	85.75%	85.50%
JRip Algorithm	88.90%	87.07%

Fig 19: Portuguese course results from base paper [1]

When comparing the results with the results from the base paper (*Fig 18 and Fig 19*), we can see that Random Forest and SVM are able to find a much better accuracy more than 96% on the pre-processed data.

VIII. CONCLUSION AND FUTURE WORK

An essential component of our society is education. The field of education can benefit from business intelligence (BI)/data mining (DM) approaches, which enable the high-level extraction of knowledge from unstructured data.
In example, several studies have employed BI/DM techniques to boost school resource management and educational quality. Using past school grades (first and second periods), demographic, socioeconomic, and other school-related data, we have addressed the prediction of secondary student grades in two key classes (Mathematics and Portuguese). Four alternative DM techniques, including KNN, Support Vector Machine (SVM), Naive Bayes (NB), and Random Forests (RF), were investigated. However, the use of a student prediction engine as a component of a school administration support system has the potential to create an automatic online learning environment. This will enable the gathering of extra information (such as grades from prior academic years) and the acquisition of insightful input from the school personnel. To improve the student databases, we also plan to expand the experiments to new schools and academic years. Because just a subset of the input variables taken into consideration appear to be pertinent, automatic feature selection techniques (such as filtering or wrapping) will also be investigated. This should particularly help nonlinear function approaches (like NN and SVM), which are more susceptible to irrelevant inputs. To comprehend why and how particular factors (such as motivation for choosing school, parent's employment, or alcohol intake) effect student performance, more research is also required (e.g., sociological studies).

IX. REFERENCES

[1] Kumar, Mukesh & Sharma, Chetan & Sharma, Shamneesh & , Nidhi & Islam, Nazrul. (2022). Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance. 10.1109/DASA54658.2022.9765236. https://www.researchgate.net/profile/Shamneesh-Sharma/publication/359520060_Analysis_of_Feature_Selection_and_Data_Mining_Techniques_to_Predict_Student_Academic_Performance/links/6242b4fd57084c718b72cabc/Analysis-of-Feature-Selection-and-Data-Mining-Techniques-to-Predict-Student-Academic-Performance.pdf

[2] Zhang, Yupei, et al. "Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis." *Frontiers in psychology* 12 (2021). https://www.frontiersin.org/articles/10.3389/fpsyg.2021.698490/full

[3] Ahmed, Adeel, et al. "Students' Class Performance Prediction Using Machine Learning Classifiers." *Quaid-E-Awam University Research Journal of Engineering, Science & Technology, Nawabshah.* 19.1 (2021): 112-121. https://pdfs.semanticscholar.org/0e6b/f2516ecb3eebdd1e1f35e16a26ce4d830769.pdf?_ga=2.77183125.1291724342.1660019231-263005371.1660019231

[4] Ünal, Ferda. "Data mining for student performance prediction in education." *Data Mining-Methods, Applications and Systems* (2020). https://pdfs.semanticscholar.org/8b76/9fd122c361ee695d71cc1e9aeac36f0cae67.pdf?_ga=2.14277687.1291724342.1660019231-263005371.1660019231

[5] Mengash, Hanan Abdullah. "Using data mining techniques to predict student performance to support decision making in university admission systems." *IEEE Access* 8 (2020): 55462-55470. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9042216

[6] Hasan, Raza, et al. "Predicting student performance in higher educational institutions using video learning analytics and data mining techniques." *Applied Sciences* 10.11 (2020): 3894. https://www.mdpi.com/2076-3417/10/11/3894

[7] Karthikeyan, V. Ganesh, P. Thangaraj, and S. Karthik. "Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation." *Soft Computing* 24.24 (2020): 18477-18487. https://link.springer.com/article/10.1007/s00500-020-05075-4

[8] Alhakami, Hosam, Tahani Alsubait, and Abdullah Aljarallah. "Data mining for student advising." *International Journal of Advanced Computer Science and Applications* 11.3 (2020). https://pdfs.semanticscholar.org/6a37/d4fca2abe00eba300d89668965dee95660ab.pdf?_ga=2.72594451.1291724342.1660019231-263005371.1660019231

[9] Hassan, Hasniza, Nor Bahiah Ahmad, and Syahid Anuar. "Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining." *Journal of Physics: Conference Series*. Vol. 1529. No. 5. IOP Publishing, 2020. https://iopscience.iop.org/article/10.1088/1742-6596/1529/5/052041

[10] Yousafzai, Bashir Khan, Maqsood Hayat, and Sher Afzal. "Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student." *Education and Information Technologies* 25.6 (2020): 4677-4697. https://link.springer.com/article/10.1007/s10639-020-10189-1

[11] Costa, Evandro B., et al. "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses." *Computers in human behavior* 73 (2017): 247-256. https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0747563217300596

[12] Burgos, Concepción, et al. "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout." *Computers & Electrical Engineering* 66 (2018): 541-556. https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0045790617305220

[13] Miguéis, Vera L., et al. "Early segmentation of students according to their academic performance: A predictive modelling approach." *Decision Support Systems* 115 (2018): 36-51. https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0167923618301428

[14] Helal, Sumyea, et al. "Predicting academic performance by considering student heterogeneity." *Knowledge-Based Systems* 161 (2018): 134-146. https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0950705118303939

[15] Hoffait, Anne-Sophie, and Michael Schyns. "Early detection of university students with potential difficulties." *Decision Support Systems* 101 (2017): 1-11. https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0167923617300817

[16] Campagni, Renza, et al. "Data mining models for student careers." *Expert Systems with Applications* 42.13 (2015): 5508-5521. https://www.sciencedirect.com/science/article/abs/pii/S0957417415001591?fr=RR-2&ref=pdf_download&rr=74775891d8826eb9

[17] Kaur, Parneet, Manpreet Singh, and Gurpreet Singh Josan. "Classification and prediction based data mining algorithms to predict slow learners in education sector." *Procedia Computer Science* 57 (2015): 500-508. https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S1877050915019018

[18] Jacob, John, et al. "Educational data mining techniques and their applications." *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*. IEEE, 2015. https://ieeexplore.ieee.org/abstract/document/7380675/

[19] Saa, Amjad Abu. "Educational data mining & students' performance prediction." *International Journal of Advanced Computer Science and Applications* 7.5 (2016). https://pdfs.semanticscholar.org/b280/216a1d63015afc6a3d1aac9595aeb2b7dd5a.pdf

[20] Durairaj, M., and C. Vijitha. "Educational data mining for prediction of student performance using clustering algorithms." *International Journal of Computer Science and Information Technologies* 5.4 (2014): 5987-5991. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.567.8824&rep=rep1&type=pdf

[21] Ashraf, Aysha, Sajid Anwer, and Muhammad Gufran Khan. "A Comparative study of predicting student's performance by use of data mining techniques." *American Academic Scientific Research Journal for Engineering, Technology, and Sciences* 44.1 (2018): 122-136. https://www.asrjetsjournal.org/index.php/American_Scientific_Journal/article/view/4170/1486

[22] Yaacob, Wan Fairos Wan, et al. "Supervised data mining approach for predicting student performance." *Indones. J. Electr. Eng. Comput. Sci* 16.3 (2019): 1584-1592. https://www.researchgate.net/profile/Wan-Fairos-Wan-Yaacob/publication/335541406_Supervised_data_mining_approach_for_predicting_student_performance/links/5d6c5c2a458515088606595f/Supervised-data-mining-approach-for-predicting-student-performance.pdf

[23] Sultana, Jabeen, M. Usha Rani, and M. A. H. Farquad. "Student's performance prediction using deep learning and data mining methods." *Int. J. Recent Technol. Eng* 8.1S4 (2019): 1018-1021. https://www.researchgate.net/profile/J-Sultana/publication/335234927_Student's_Performance_Prediction_using_Deep_Learning_and_Data_Mining_methods/links/5d5a6b3e92851c3763694c8f/Students-Performance-Prediction-using-Deep-Learning-and-Data-Mining-methods.pdf

[24] Aziz, Sirwan M., and Ardalan Husin Awlla. "Performance analysis and prediction student performance to build effective student using data mining techniques." *UHD Journal of Science and Technology* 3.2 (2019): 10-15. https://journals.uhd.edu.iq/index.php/uhdjst/article/view/332/196

[25] Garg, Rajni. "Predicting student performance of different regions of Punjab using classification techniques." *Int. J. Adv. Res. Comput. Sci* 9 (2018): 236-241. http://ijarcs.info/index.php/Ijarcs/article/view/5234/4486

[26] Saheed, Y. K., et al. "Student performance prediction based on data mining classification techniques." *Nigerian Journal of Technology* 37.4 (2018): 1087-1091. https://www.ajol.info/index.php/njt/article/view/179736

[27] Suhaimi, Nurafifah Mohammad, et al. "Review on Predicting Students' Graduation Time Using Machine Learning Algorithms." *International Journal of Modern Education & Computer Science* 11.7 (2019). https://www.mecs-press.org/ijmecs/ijmecs-v11-n7/IJMECS-V11-N7-1.pdf

[28] Santoso, Leo Willyanto. *Predicting student performance using data mining*. Diss. Petra Christian University, 2018.http://repository.petra.ac.id/18007/1/Publikasi1_03023_4169.pdf

[29] Salal, Y. K., and S. M. Abdullaev. "Using of Data Mining techniques to predictof student's performance in industrial institute of Al-Diwaniyah, Iraq." *Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника* 19.1 (2019): 121-130. https://cyberleninka.ru/article/n/using-of-data-mining-techniques-to-predictof-student-s-performance-in-industrial-institute-of-al-diwaniyah-iraq

[30] Li, Yu, Jin Gou, and Zongwen Fan. "Educational data mining for students' performance based on fuzzy C-means clustering." *The Journal of Engineering* 2019.11 (2019): 8245-8250. https://ietresearch.onlinelibrary.wiley.com/doi/epdf/10.1049/joe.2019.0938