ISSN No:-2456-2165

Academic Prediction

Angelo Jaison
Student, Computer science and Engineering Department
Sahrdaya College of Engineering Technology
Thrissur, India

Ashik V
Student, Computer science and Engineering Department
Sahrdaya College of Engineering Technology
Thrissur, India

Aswin E B
Student, Computer science and Engineering Department
Sahrdaya College of Engineering Technology
Thrissur, India

Daniyel Johnson
Student, Computer science and Engineering Department
Sahrdaya College of Engineering Technology
Thrissur, India

Priya K V
Asst. Professor, Computer science and Engineering Department
Sahrdaya College of Engineering Technology
Thrissur, India

Abstract:- Every educational institute maintains a proper database on their student performance and activities. This information is incredibly useful in the realm of education., particularly for evaluating the performance of students. It is true that evaluating student performance has grown difficult due to the lack of comparison between different resampling methods due to the imbalanced data sets in this discipline. Some of these resampling models such as Random Forest, Artificial Neural Network and Logistic Regression. Furthermore are compared in this paper and the model validation we used here is 5-fold cross validation. These resampling methods provide an accurate output on the current performance of students and state the variance in their performance. This provides a reliable source to view and check the performance of students.

I. INTRODUCTION

Educational quality is essential to a country's development. With the help of admissions systems, academic information systems, learning management systems, and e-learning, data in the education domain is growing by the day. The information gathered from pupils is typically utilised to answer simple questions and make decisions. However, due to the complexity and enormous size of the data sets, the majority of this data stays unusable.As a result, analysing this vast amount of educational data in order to predict student achievement is a hot topic. Data mining, also known as knowledge discovery in databases, is the process of extracting meaningful information from large collections of data (KDD). It's been used successfully in a variety of industries, including finance, medical, and business, and it's now being employed in education under the label Educational Data Mining.

Predicting student performance is a crucial topic that EDM is looking at. This task predicts the value of an unknown variable that characterises students in terms of outcome (Pass/Fail), grades, and marks, among other things. The literature review for this study focuses on predicting student attrition, failures, and success. The stakeholders in this area are all looking for an early warning system to predict learning at an early stage. Not only did this early warning system reduce the cost of learning, but it also reduced the amount of time and space needed.

Because of technological advancements, predicting student success has become an essential study issue. In this field, unbalanced datasets have made it difficult to forecast student performance, and there is no comparison between various resampling strategies. The majority of data on student performance and activities is not used. These data can be utilised to make accurate forecasts and evaluations of students if a good baseline is created. To deal with the imbalanced data, this paper compares a few resampling strategies such as Random Forest, Artificial Neural Network, and Logistic Regression. To fully verify the efficiency of the resampling strategies, balanced datasets are used to improve classifier performance. In addition, the 5-fold cross-validation approach is employed to validate the model.

Students may be able to have a better idea of how well or poorly they will perform in a course based on the prediction results, and then take efforts to improve their performance. Any educational institution around the world has a long-term goal of increasing student retention. Increased retention has numerous benefits, including improved college reputation, ranking, and job possibilities for alumni, among others.

ISSN No:-2456-2165

II. RELATED WORK

Here we look at some of the related works in the field of student performance prediction.

1) KWOK TAI CHUIL, RYAN WEN LIU,MINGBO ZHAO and PATRICIA ORDONEZ DE PABLOS– "Using a Deep Support Vector Machine based on a Generative Adversarial Network to Predict Student Performance with School and Family Tutoring," Automating student evaluation entails the development and application of machine learning technologies to aid in student analysis. Conditional GAN-related deep support vector machine algorithm has been improved. ICGAN solves the problem of low data volume by duplicating a new training dataset, whereas DSVM extends SVM from shallow to deep learning. The ICGAN approach is given as a way to generate more data on training student performance. The DSVM is in charge of the performance prediction model for pupils. The collecting of data on pupils and their performance can be done more consistently and accurately.
2) HANAN ABDULLAH MENGASH– "Using Data Mining Techniques to Predict Student Performance to Assist University Admissions Decision Making," This research aimed to assist universities in making admission decisions by using data mining techniques to forecast students' academic achievement before they were admitted. Using data mining classification approaches, we design and test four prediction models to predict early academic success among candidates based on their preadmission profiles. Artificial Neural Network (ANN), Decision Tree, Support Vector Machine (SVM), and Naive Bayes are four well-known data mining approaches. The university where this study was done chose to adjust the weighting of its admission standards based on the findings and recommendations of this study.
3) ABDULLAH ALSHANQITI AND ABDALLAH NAMOUN DLT– "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," The educational system recognises the potential of data mining to significantly improve its performance. This paper compares different resampling techniques like Borderline To solve the unbalanced data problem for forecasting students' performance using two different datasets, SMOTE, Random Over Sampler, SVM-SMOTE, SMOTE-ENN, SMOTE, and SMOTE-Tomek were developed. Although data mining aids in knowledge discovery, it is not without its drawbacks, machine learning algorithms gives the required tools for this purpose. Model validation is done using 5-fold cross-validation, which divides the dataset into ve subsets and utilises one of the ve subsets as the testset and the other four subsets as the training set.
4) RAMIN GHORBANI AND ROUZBEH GHOUSI– "Using Hybrid Regression and Multi-Label Classification to Predict Student Performance and Its Influential Factors," Students have a wide range of characteristics and past behaviours, and using a single model may result in erroneous predictions. This paper contributes to a hybrid regression model that improves the accuracy of predicting student academic performance, as measured by future grades in various courses, as well as an optimised multi-label classifier that predicts qualitative values for the effect of multiple factors related to the student results received.

III. PROPOSED SYSTEM

As we have seen there are quite a lot of different approaches to predicting student performance.. What we hope to achieve is a performance predicting system that has all the latest features of any modern software and combining it with machine learning to further enhance its capabilities.

Student performance is the most essential aspect in determining the quality of a university in higher education. Because of its importance in decision making, EDM is currently the most extensively used method for evaluating and predicting student performance by researchers. There are two key things to consider when predicting student performance: characteristics and prediction methods. Student CGPA has been demonstrated to be the most commonly used indicator in predicting university performance. It has been utilised in numerous studies. Assessments, quiz grades, lab work, and final exam marks are some of the other characteristics that researchers use to predict student performance at university. Other factors such as extracurricular activities, student demographics, and social contact networks have been used by a few study.

Several data mining classification algorithms have been used to predict student performance in some of the articles mentioned above. In one study, 505 eighth-semester students' academic progress was predicted using ANN. Using Decision Trees, the study created a way to predict student achievement in specific courses using small student sample sizes. (32 and 42 students, respectively). In a study of 1,600 students, Naive Bayes was used to predict achievement in a specific subject. SVM was used to predict students at risk's performance in their first year of study on a data set of 1,074 students in a research.

According to a review of the literature, the majority of research do not look into employing numerous prediction models and using the most accurate results. This research aims to analyse the various resampling approaches for dealing with the unbalanced data problem in order to determine the optimal technique and classifier for forecasting student performance. This study also aims to look into the differences between multiclass and binary classification, as well as the significance of feature structure. Models with fewer classes and nominal features perform better, according to the findings obtained using multiple assessment measures. Furthermore, classifiers do not perform well when data is imbalanced, thus this problem must be addressed. A balanced dataset improves the performance of classifiers.

ISSN No:-2456-2165

IV. METHODOLOGY

This research aims to analyse various machine learning approaches for dealing with balanced data in order to determine the optimal methodology and classifier for predicting student performance. Figure 1 depicts the steps of the applied approach used to attain the paper's objectives.

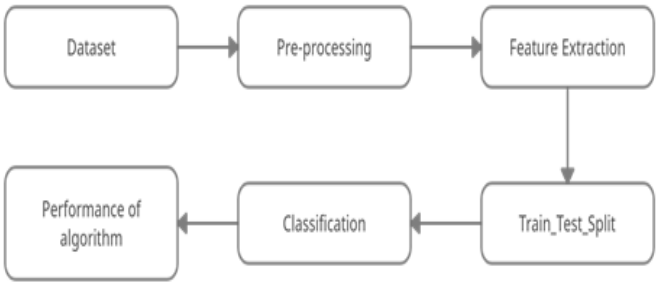

Fig. 1: Architecture diagram

A. Data Collection

A student performance data set used in this study has been collected from kaggle. It contains 1044 instances, each with 33 properties, such as student grades, demographic, social and school-related characteristics. The following are the detailed description of attributes present in our dataset they are Sex, Age, School, Address, Parent status, Mother's education, Mother' job, Father's education, Father's job, Student guarduia, Family size, Quality of family relationship, Reason to choose school, Travel time, Study time, Failures, School support, Family support, Activities, Tuition details, Internet, Nursery, Highers education, Relationship status, Free time, Going out with friends, Alcohol consumption weekend, Workday alcohol consumption, Health status, Absences, First period grade, Second period grade, Final grade. The dataset is downloaded from an Open Source Kaggle kernel database. The dataset attributes are independent of each other except the student's final grade which depends on all other attributes.

B. Data Preprocessing

Data preparation is one of the most important phases in machine learning. This process converts the raw data into a format that can be understood. In the real world, datasets contain several flaws; as a result, this phase can eliminate the errors, making datasets easier to manage. Because the datasets utilised in this study include no missing data, treating missing data as a stage of data preparation is not necessary.

C. Feature Extraction

Feature extraction, often known as data normalisation, is a technique for normalising a dataset's range of independent variables or features. The Euclidean distance between two data points is used in most machine learning models, hence they may not operate well without feature extraction. Standardization, Mean Normalization, Min-Max Scaling, and Unit Vector are four prominent techniques to implement Feature extraction. The student performance dataset values used in this study span a wide range. To rescale the features, this research used the Standardization approach. As a result, all of the features have the conventional normal distribution properties of = 0 and 1, where is the average and is the standard deviation from the average. Equ 1 defines the formula for scaling the values.

$$\mathscr{Z} = \frac{x - \mu}{\sigma} \qquad \text{(Equ 1)}$$

D. Random Forest

Random Forest is a common algorithm for machine learning that is part of the supervised learning process.It can be used for problems in machine learning in both classification and regression. It is based on the abstract of ensemble learning, which is a process of combining different classifiers to resolve a complex problem and to improve the efficiency of the model. Random Forest is a controlled machine learning approach that combines numerous decision trees on distinct subsets of the dataset and averages them to improve the data set's forecasting accuracy. Instead of relying on a single decision tree, random forest gathers forecasts from all trees and forecasts the ultimate result based on the most votes.

E. Logistic Regression

In its most basic form, logistic regression is a statistical model that represents a binary dependent variable with a logistic function, while there are many more complex extensions available. In regression analysis, logistic regression estimates the parameters of a logistic model. The logistic regression, like all regression studies, is a predictive analysis. Logistic regression, ordinal, interval, or ratio-level independent variables are used to analyse data and explain the relationship between one dependent binary variable and one or more nominal variables. Logistic regressions are notoriously difficult to comprehend; but, using the Intellectus Statistics technique, you may quickly conduct the study and analyse the results in plain English.

F. Artificial Neural Network

Artificial neural networks are comparable to biological neural networks in form, function, and data processing, and they are a relatively good methodology for solving classification and prediction problems. ANN is a set of mathematical models that can imitate a number of biological neural system properties and are similar to adaptive human learning.

They are made up of a large number of linked neurons connected by connections that carry permeability (weight) coefficients that are similar to synapses in function. Input layer, one or more hidden layers, and output layer are the three levels in which the neurons are organised.

ANNs handle data in the same way that biological neural networks do, with the added capability of remembering, learning, and correcting errors at a high rate, allowing neural networks to be utilised to solve complicated tasks like classification and prediction. ANNs have been effectively employed to model complicated and real-world situations in a variety of areas.

G. Model Validation

Cross-validation is a model validation approach for determining how statistical analysis results are generalised within a single dataset. This research uses precision data from shuffle 5-fold cross-validation for each resampling procedure. This divides the dataset into five subgroups, with one of the subsets serving as the assessment set and the other four serving as the training set. Set each time and then repeat the process five times more.

The results of shuffle 5-fold cross-validation are more reliable and appropriate because of the way this methodology works. After addressing the unbalanced data query, the results show that several of the model accuracies have improved slightly. Regarding the output of classifiers utilising other classifiers and resampling techniques, it should be mentioned that Random Forest has performed admirably in almost all balanced datasets.

H. Flowchart

Figure 2 represents the flowchart of our project. The flowchart shows you faces of our project training phase and prediction phase. In the training phase the data is collected and it is preprocessed and feature selected to obtain a clean dataset for the training phase. The normalised data is then trained by three different algorithms. The next is the prediction phase. During this phase an input of a student with relevant details is collected for the purpose of prediction; the collected details are also feature extracted to make it compatible for ML algorithms. The output will be in the form of a text file which has the prediction result of the student in percentage and a description of whether he/she has passed or not.

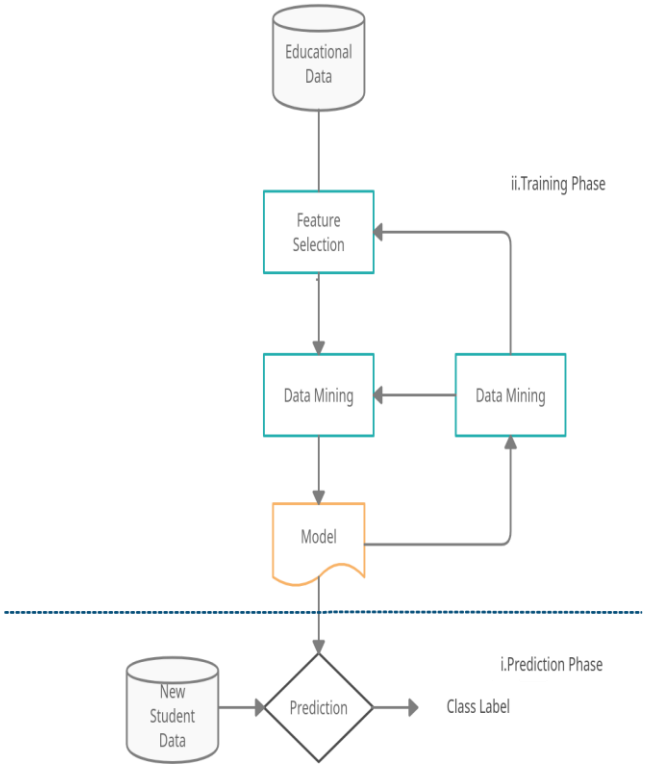

Fig. 2: Flowchart

I. Use Case Diagram

Figure 3 shows the use case diagram of our project. It shows the different users possible interaction with our system. There are two users : user and administrator both have equal interaction with the system but only the administrator has administrative power over the system.. The user group has a student and teacher and the student can input their details and make predictions. The teacher and inspect the individual results of students.

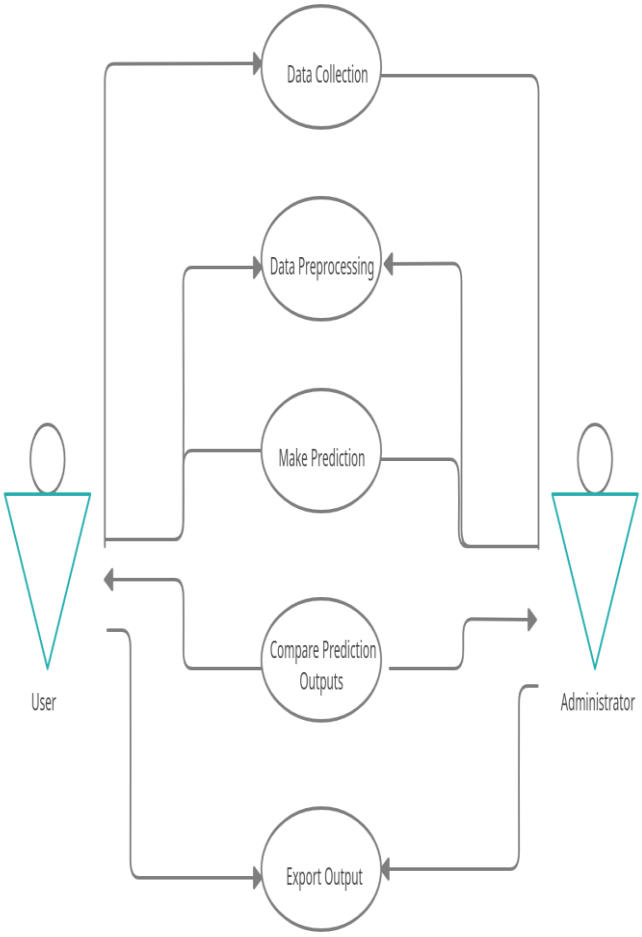

Fig. 3: Use case diagram

V. RESULT

Data Mining is extremely helpful, especially for analyzing students' performance. Our project helps us to get a futuristic information of students whether he/she may pass or fail in the upcoming series of assessment. Our system provides the result in the form of text which is generated after the prediction phase; it includes the predicted percentage of his/her next series assessment. It also has a textual output describing whether he/she will be passed or failed in the form of (Pass/Fail) this helps the student to foresee their future and be prepared for their upcoming series assessment.

It also has a feature that recommends to the student what all things they need to do for improving their results in future. This recommendation module is created by comparing the students feature extracted value with the

clothes students values by comparing this the system gets an overview of what needs to be improved for increasing their results. This is a useful feature for students who get “fail” as prediction output.

VI. CONCLUSION

Predicting student performance is one of the most important study subjects that should be investigated right now. Data Mining is tremendously useful in the realm of education, particularly for analysing student performance. Because of the imbalance in data sets in this field, forecasting students' performance has become a huge challenge, and there is no comparison between different machine learning algorithms. The project will analyze the given datasets and perform various ML algorithms, compare their outputs and give the most accurate results. It's worth noting that two separate datasets linked to student performance are employed, as well as the differences between multiclass and binary classification and feature structure. To improve the conclusion of resampling approaches, several classifiers can be utilised. On the imbalanced dataset, all of the classifiers are first run using the random hold-out approach. The results reveal that when dealing with unbalanced data, classifiers are unable to make correct predictions and are unable to predict some of the classes at all. Furthermore, the findings collected using various assessment criteria show that having fewer classes leads to greater performance with machine learning models.

This research can be expanded in a variety of ways, and future work could go in the following lines. For a better comparison and improved performance, new ensemble and hybrid classifiers could be developed. Additionally, feature selection approaches can be used to improve model results and gain a better understanding of the important features.

REFERENCES

[1]. KWOK TAI CHUIL, RYAN WEN LIU,MINGBO ZHAO and PATRICIA ORDONEZ DE PABLOS, PredictingStudents' Performance with School and Family Tutoring using Generative Adversarial Network based Deep SupportVector Machine Digital Object Identifier 10.1109/ACCESS.2019.Doi Number Shanghai 200051, China

[2]. ABDULLAH ALSHANQITI AND ABDALLAH NAMOUN, Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification Digital Object Identifier November19,2020.Digital Object Identifier 10.1109/ACCESS.2020.3036572 Madinah 42351, Saudi Arabia

[3]. RAMIN GHORBANI AND ROUZBEH GHOUS,[1] Comparing Different Resampling Methods in Predicting Students’ Performance Using Machine Learning Techniques Digital Object Identifier 10.1109/ACCESS.2020 April 22, 2020. Digital Object Identifier 10.1109/ACCESS.2020.2986809 Tehran 16846-13114, Iran

[4]. HANAN ABDULLAH MENGASH, Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems March 30, 2020. Digital Object Identifier 10.1109/ACCESS.2020.2981905 Riyadh 11351, Saudi Arabia

[5]. AGORITSA POLYZOU AND GEORGE KARYPIS, Feature Extraction for Next-Term Prediction of Poor Student Performance IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 12, NO. 2, APRIL-JUNE 2019

[6]. SANA BHUTTO,DR. QASIM ALI ARAIN,MALEEHA ANWAR, Predicting Students' Academic Performance Through Supervised Machine Learning 2020 International Conference on Information Science and Communication Technology

[7]. Y.-H. HU, C.-L. LO, AND S.-P. SHIH, ‘‘Developing early warning systems to predict students’ online learning performance,’’ Comput. Hum. Behav., vol. 36, pp. 469–478, Jul. 2014.

[8]. M.ZIĘBA, S. K. TOMCZAK, AND J. M. TOMCZAK, ‘‘Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction,’’ Expert Syst. Appl., vol. 58, pp. 93–101, Oct. 2016.