A Method of Classification in Student Grade Prediction using SMOTE Oversampling Technique

M. S. Rekha[1], Meenakshi H[2], Dr. Anil Kumar C[3], Dr. Kumara Swamy J[4]
Assistant Professor, Department of Computer Science and Engineering[1]
Assistant Professor, Department of Allied Sciences[2]
Associate Professor, Department of Electronics and Communication Engineering[3]
Assistant Professor, Department of Mechanical Engineering[4]
R.L Jalappa Institute of Technology, Doddaballapur, India

Abstract*:* *The prediction of student academic performance is the most difficult task in educational institutions, which helps to improve the students performance in further semesters. Though we have many kinds of machine learning techniques we have disadvantages while handling the imbalanced data sets. It is also difficult to balance the large data sets which leads to obtain the predictions with less accuracy and false predictions. To overcome this disadvantages we will be using SMOTE technique which means Synthetic Minority Oversampling Method. It helps to obtain high accuracy by comparing the different machine learning algorithms used. The proposed methodology gives us the accurate results and best prediction of student grades*

Keywords: Machine Learning, Algorithms, SMOTE, Grade Prediction, multi-class classification.

I. INTRODUCTION

Prediction of Student grades based on their performance in previous semesters is a very difficult task in educational institutions. By predicting the student grade one can know where they are lacking and the part to be improved. Here we will be considering a dataset of student with various attributes where each and every attribute effects the student grades and their further studies. By letting them know this prediction they can improve which is benefit to a student to improve performance and also to the institution to obtain ranks. Here we will be using various machine learning algorithms and calculate the accuracy of each one and predict the grade. Once the prediction is done through each algorithm we will use SMOTE technique and then find the most accurate performance in grades. In this study, we aim to evaluate the performance of six popular machine learning techniques, namely Decision Tree, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors, Logistic Regression, and Random Forest, using a dataset of student's course grades. The primary objective is to compare the accuracy of these techniques and identify the most effective approach for student grade prediction.

Furthermore, we propose a multiclass prediction model to address challenges associated with imbalanced classification and potential issues of overfitting and misclassification. To mitigate these problems, we introduce the Synthetic Minority Oversampling Technique (SMOTE) in combination with feature selection methods. By applying SMOTE, we generate synthetic samples for minority classes to balance the dataset and provide more representative training data. This approach aims to improve the overall performance of the multiclass prediction model, reducing the biases caused by imbalanced class distributions. Additionally, we employ feature selection methods to identify the most relevant and informative features for the prediction model. This step helps in enhancing the model's efficiency and reducing the dimensionality of the dataset, ultimately improving prediction accuracy.

The results of this study have the potential to contribute to the field of student performance prediction by identifying the most accurate machine learning technique and proposing a robust multiclass prediction model. The combination of SMOTE and feature selection methods can help overcome the challenges associated with imbalanced datasets, ultimately leading to more reliable and precise student grade predictions. In the following sections, we will present the methodology employed for evaluating the six machine learning techniques and describe in detail the proposed multiclass prediction model incorporating SMOTE and feature selection methods.

DOI: 10.48175/IJARSCT-12719

II. LITERATURE SURVEY

[1]. "Predicting Student's Performance using Data Mining Techniques", By Nur'aini Abdul Rashida, Amirah Mohamed Shahiria, Wahidah Husaina.

The study explores the use of various data mining algorithms and methods to analyze student data and develop predictive models. The authors begin by emphasizing the importance of predicting student performance as it can aid educators in identifying students at risk of academic difficulties and implementing appropriate interventions. They highlight that data mining techniques offer valuable insights by extracting hidden patterns and relationships from large educational datasets. The literature survey and provides recommendations for future research. The authors suggest exploring ensemble methods that combine multiple data mining techniques to enhance prediction accuracy. They also advocate for the integration of real-time data streams and the use of more advanced techniques, such as deep learning and natural language processing, to improve prediction models further.

Overall, the literature survey presented in this paper provides an extensive overview of the research conducted on predicting student performance using data mining techniques. It offers valuable insights into the various algorithms, data sources, and challenges in this field, serving as a useful resource for researchers and educators interested in leveraging data mining for educational purposes. The authors emphasize the significance of predicting academic performance as it can assist educators in identifying students who may require additional support or interventions. They highlight the potential benefits of data-driven approaches, specifically the application of machine learning and statistical modeling techniques.

[2]. "Predicting Academic Performance", By Arto Hellas: University of Helsinki, Petri Ihantola: University of Helsinki, Andrew Petersen: University of Toronto Mississauga

The paper presents a comprehensive review of existing research and studies related to predicting academic performance. The authors discuss different factors and features that have been identified as influential in determining students' academic success. These factors encompass a wide range of variables, including demographic information, prior academic records, socio-economic background, and behavioral patterns. The paper concludes by summarizing the key findings from the literature review and identifying future directions for research. The authors suggest exploring the integration of additional data sources, such as social network data or sensor data, to enhance the predictive accuracy. They also highlight the importance of considering individual differences and personalized approaches in predicting academic performance.

[3]. "Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning", By Shah Hussain & Muhammad Qasim Khan

The authors highlight the importance of predicting academic performance as a means to identify students who may be at risk of underperformance or require additional support. They emphasize the potential of machine learning algorithms to analyze large datasets and extract meaningful patterns and insights. The study includes a comprehensive literature review that surveys previous research on predicting academic performance using machine learning. The authors discuss various factors that influence academic performance, including demographic information, prior academic achievements, socio-economic background, and learning behavior. Furthermore, the paper discusses the data sources commonly used for predicting academic performance, including student records, examination results, and socio-economic data. The authors emphasize the need for data privacy and ethical considerations when utilizing personal information. The study concludes by summarizing the key findings from the literature review and identifying potential areas for future research. The authors suggest exploring ensemble methods that combine multiple machine learning algorithms to enhance prediction accuracy. They also emphasize the importance of interpretability and transparency in the predictive models to facilitate their practical implementation in educational settings.

[4]. "Predicting academic success in higher education", By Eyman Alyahyan & Dilek Düştegör

The authors discuss different data sources that have been utilized for predicting academic success, such as student records, admission data, learning management systems, and surveys. They emphasize the importance of data quality and the need for appropriate data preprocessing techniques to ensure accurate and reliable predictions. Furthermore, Alyahyan and Düştegör explore the different prediction models and techniques employed in the field. They discuss the use of statistical methods, machine learning algorithms, and data mining techniques to develop predictive models. The authors provide insights into regression analysis, decision trees, support vector machines, neural networks, and ensemble methods, highlighting their applicability and limitations. The paper concludes by summarizing the findings from the literature survey and suggesting directions for future research. Alyahyan and Düştegör recommend exploring the integration of additional data sources, such as social network data and online learning analytics, to enhance prediction accuracy.

[5]. "Student Academic Performance Using Hybrid Deep Neural Network ", By Bashir Khan Yousafzai 1, Sher Afzal Khan 1, Taj Rahman 2, Inayat Khan 3, Inam Ullah, Ateeq Ur Rehman, Mohammed Baz, Habib Hamam and Omar Cheikhrouhou

The authors emphasize the significance of predicting student academic performance as it can assist in identifying students who may be at risk of underperformance and enable timely interventions to support their educational journey. They propose the use of a hybrid deep neural network, which combines the strengths of different deep learning models, to achieve accurate and reliable predictions. Throughout the literature survey, the authors critically analyze the strengths, limitations, and potential challenges of existing research in predicting student academic performance using deep neural networks. They highlight the need for interpretability and explainability of the models to provide actionable insights for educators and stakeholders. The paper concludes by summarizing the findings from the literature survey and providing recommendations for future research. The authors suggest exploring the integration of additional data sources, such as social network data and learning analytics, to further enhance the predictive accuracy. They also highlight the importance of considering ethical and privacy concerns when handling student data.

III. ALGORITHM

STEP 1: Start a Django project
STEP 2: Create an app
STEP 3: Write the HTML code
STEP 4: Hit the URL on app
STEP 5: Include the app URLs in Django project
STEP 6: write algorithms in python
STEP 7: Apply SMOTE
STEP 8: Save the model in .sav file
STEP 9: Dump the model file into views

The student grade prediction model with data collection,where relevant information such as previous grades,attendance recors,demographic details and other factors can be gathered from open source to train the modules and it has to be preprocessed in order to address overfitting and misclassification issues that can be arised from imbalanced multi classification. Next,the feature selection can be extracted from the use of SMOTE Technique to over sample the minority class.the finalised the result data has to be saved for model building and system takes input data from the users and produces the output. Hence,the user can view the generated files from the model.the system check the accuracy of the model based on the data set and classification model we built.

IV. FLOW DIAGRAM

The figure 1 shows the block diagram of the process of predicting student grades in various stages. Initially the main requirement of the process is to gather the dataset. The dataset plays a vital role in which it contains student data in various attributes and helps in prediction. Then we do the preprocessing of data which means we need to filter the data and make the data ready for feature engineering. In feature Engineering the data will further undergone for Feature Selection.

The SMOTE technique is used in handling the imbalanced dataset in the stage of Feature Engineering. Then we will apply all the machine learning algorithms to the dataset. Here we will calculate the accuracy in each algorithm. Once we apply all the algorithms we do model building.

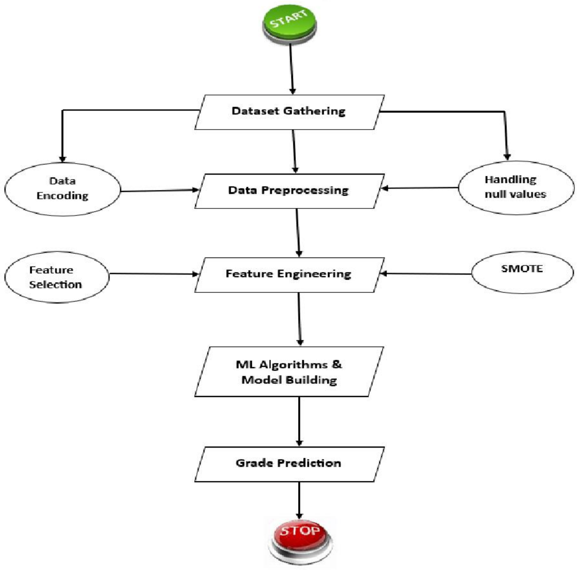


Fig 1:Flow Diagram of Student Grade Prediction

Model building is an important task where we implement our code here and check it with the system. The system checks the syntax and errors. If there is no errors it will execute the code and gives the output. Here we will be getting a grade as a prediction as output as a pop-up box. We will run our code on server where we the Django application will be running.

V. MODULES

USER MODULE

- Data Gathering: In this study, we will gather relevant information or data from open-source datasets, which will be used to train our machine learning models. These datasets may include historical student performance records, demographic information, socio-economic factors, and other related attributes. Open-source datasets provide a valuable resource for conducting research and developing prediction models.
- Pre-Processing: Data pre-processing is a crucial step to improve the accuracy and quality of the models. We will perform various pre-processing techniques such as handling missing values, data normalization, removing outliers, and dealing with categorical variables. Pre-processing ensures that the data is in a suitable format and removes any inconsistencies or noise that may affect the performance of the models.
- Feature Engineering: Feature engineering plays a significant role in selecting the most relevant features for our models. We will analyze the importance and correlation of each column in the dataset and prioritize them based on their significance in predicting student grades. This step helps in reducing the dimensionality of the dataset and optimizing the training process by focusing on the most informative features.
- Model Building: Once the data is prepared and features are engineered, we will proceed with building our prediction models. We will employ various classification and regression algorithms such as Decision Trees, Support Vector Machines, Naïve Bayes, K-Nearest Neighbors, Logistic Regression, and Random Forest. These models will be trained using the pre-processed data to accurately predict student grades based on the selected features.
- View Results: After the models are built and trained, the user will have the opportunity to view the generated results. This could include predicted student grades for future semesters or specific courses. The results will provide valuable insights into student performance and assist in making informed decisions for educational institutions, teachers, and students themselves.

By following this data gathering, pre-processing, feature engineering, model building, and result viewing process, we aim to develop an accurate and reliable prediction model for student grades using machine learning techniques.

DOI: 10.48175/IJARSCT-12719

SYSTEM MODULE

- Model Building: System Models model accuracy and it takes of the necessary for the model building.
- Generate Results: System takes the input data from the users and produce the output.

First,the system models that we built will model the accuracy and identify the results which leads to the selection of features or not which are necessary to build the model.

Later,it generate the resukts from the input data given from the user end.

VI. FIGURES AND TABLES

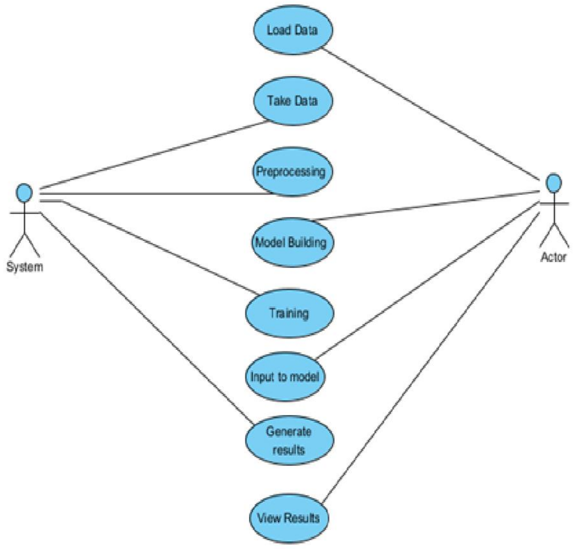


Figure 2: Use Case Diagram

The figure 2 depitcs about the visualizing and communicating the intended behavior of a system and can be used to support requirements gathering,design,testing activities. Actors are the external entites that interact with the system and trigger one or more use cases actors can be human users other system or external devicies Use case represent the specific functionality or behavior of a system that is triggered by an actors request or interaction.Each use case should have a clear purpose or goal that is meaningful to its stake holders.

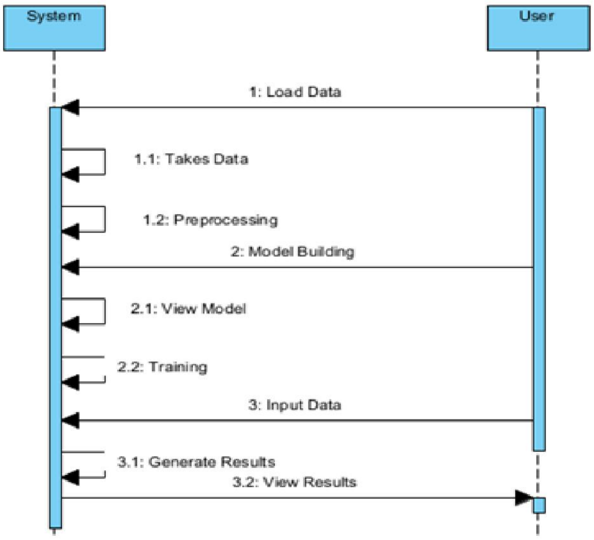
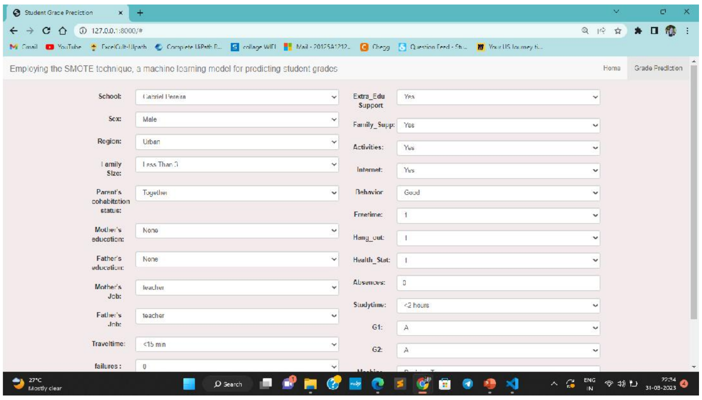


Figure 3: Sequence Diagram

DOI: 10.48175/IJARSCT-12719

The sequence diagram illustrates how a multi-class prediction model can be used to predict a student's grade based on their historical data and how that information can be displayed to the student. It will shows the flow of data and interactions between the student and the machine learning model and provide a clear visualization of the process. The sequence diagram only includes the specific interactions between the student and the machine learng model and excludes any oher functionality or external systems.

VII. RESULTS AND SNAPSHOTS

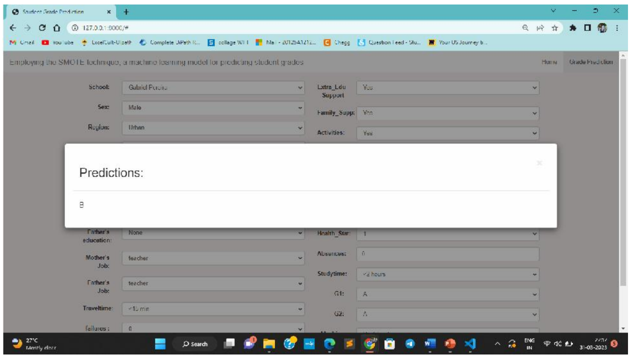


Employing the SMOTE technique, a machine learning model for predicting student grades

School: Gabriel Pereira
Sex: Male
Region: Urban
Family Size: Less Than 3
Parent's cohabitation status: Together
Mother's education: None
Father's education: None
Mother's Job: teacher
Father's Job: teacher
Traveltime: <15 min
failures : 0
Extra_Edu Support: Yes
Family_Supp: Yes
Activities: Yes
Internet: Yes
Behavior: Good
Freetime: 1
Hang_out: 1
Health_Stat: 1
Absences: 0
Studytime: <2 hours
G1: A
G2: A

Figure 4: Inputs for grade prediction

This page shows the grade prediction screen with student input attributes.



Predictions:

B

School: Gabriel Pereira
Sex: Male
Region: Urban
Father's education: None
Mother's Job: teacher
Father's Job: teacher
Traveltime: <15 min
failures : 0
Extra_Edu Support: Yes
Family_Supp: Yes
Activities: Yes
Health_Stat: 1
Absences: 0
Studytime: <2 hours
G1: A
G2: A

Figure 5: Predicted grade

This page shows the predicted grade of the student with given input attributes The results shows that the predicted grade for the given input attributes and selected algorithm is B grade.

VIII. CONCLUSION

Predicting Student grades is mostly used technique used by educational institutions to keep track of their students. Therefore, having the best predictive model helps to improve the accuracy of the outcome by depleting the uncertainty in the outcome of the model is important to avoid getting an imbalanced dataset as an outcome. Here considered a multiclass forecasting system which includes three analyzing systems to analyze the student's or user's final results by considering many factors and previous grades of students. Specifically, this is a comparison between factors that influences the output combining oversampling method SMOTE with various FS strategies to analyze the working rate of student data regarding grade prediction. Here, the findings address the data which was not in a balanced manner multi-classification considering the data-level arrangement for student grade forecasting.

REFERENCES

[1]. D. Solomon, S. Patil, and P. Agrawal, ''Predicting performance and potential difficulties of university student using classification: Survey paper,'' Int. J. Pure Appl. Math, vol. 118, no. 18, pp. 2703–2707, 2018

[2]. E. Alyahyan and D. Dustegor , ''Predicting academic success in higher education: Literature review and best practices,'' Int. J. Educ. Technol. Higher Educ., vol. 17, no. 1, Dec. 2020.

DOI: 10.48175/IJARSCT-12719

[3]. H. Sun, M. R. Rabbani, M. S. Sial, S. Yu, J. A. Filipe, and J. Cherian, ‘‘Identifying big Data's opportunities, challenges, and implications in finance,’’ Mathematics, vol. 8, no. 10, p. 1738, Oct. 2020.

[4]. A. E. Tatar and D. Düştegör, ‘‘Prediction of academic performance at undergraduate graduation: Course grades or grade point average?’’ Appl. Sci., vol. 10, no. 14, pp. 1–15, 2020.

[5]. T. Anderson and R. Anderson, ‘‘Applications of machine learning to student grade prediction in quantitative business courses,’’ Glob. J. Bus. Pedagog., vol. 1, no. 3, pp. 13– 22, 2017.

[6]. A. Polyzou and G. Karypis, ‘‘Grade prediction with models specific to students and courses,’’ Int. J. Data Sci. Anal., vol. 2, nos. 3–4, pp. 159–171, Dec. 2016.

[7] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, ‘‘Machine learning based student grade prediction: A case study,’’ 2017, arXiv:1708.08744. [Online]. Available: https://arxiv.org/abs/1708.08744.

[8] N. V. Chawla, “Data Mining for Imbalanced Datasets: An Overview,” Data Mining and Knowledge Discovery Handbook, pp. 853-867, 2005

DOI: 10.48175/IJARSCT-12719