International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Academics Performance Prediction Using Machine Learning Algorithm

[1]Kalyani Darekar, [2]Shweta Khilari, [3]Hritik Shirsath, [4]Shweta Mate, [5]Prof. J. B. Jagadale

[1,2,3,4]PVGCOE & SSDIOM, Nashik, India
[5]Project Guide, PVGCOE & SSDIOM, Nashik, India

Abstract-

Applications for predictive analytics have become urgently desired in higher education. Predictive analytics employed sophisticated analytics that included the application of machine learning to produce high-quality performance and valuable information at all academic levels. Researchers have put forth numerous variations of machine learning approaches in the education area over the past ten years. By increasing the performance of predictive accuracy, the system will give a thorough study of machine learning approaches to forecasting the final grades of the students in the first semester courses. The suggested system will be emphasized by two components. First, using a dataset of actual student course grades, assess the accuracy performance of six well-known machine learning approaches, including Decision Tree (J48), Nave Bayes (NB), Logistic Regression (LR), and Random Forest (RF). Second, we provide a multiclass prediction model that may improve the prediction performance model for unbalanced multi-classification for student grade prediction to avoid overfitting and misclassification outcomes produced by imbalanced multi-classification.

Keywords- ***Machine learning, algorithms, educational data mining, predictive model, unbalanced issue, and student grade prediction.***

I. Introduction

One of the key components of our society and one that is essential to the expansion and advancement of any country is the educational sector. The use of data mining in education is called educational data mining. It is a young, multidisciplinary field of study that is concerned with the creation of approaches for investigating data coming from educational contexts. A new concept called educational data mining aims to automatically explore the many forms of data from sizable databases of information about education. This data is frequently comprehensive, exact, and fine-grained. The major goal of this research is to analyze student performance in the courses using machine learning models. Many activities are available from ML Technologies that may be used to examine student performance. Since there are several ways used for data regression, the linear regression, decision tree, etc. method is employed here to evaluate student success on the regression test. The accuracy of regression approaches for projecting student performance is examined. The faculty has a difficult time determining the skills and interests of the pupils so they can better them in it. As a result, it could have an impact on a person's career, placement, and university outcomes. The effect is that it aids in achieving the institute's goal and vision. Successful completion of the project will greatly aid faculty in improving the educational system.

II. LITERATURE SURVEY

In order to actively engage such students in a better learning experience, academics might use machine learning technologies to forecast the anticipated limitations in learning processes. On historical data of student grades in one of the undergraduate courses, we used logistic regression, linear discriminant analysis, K-nearest neighbors, classification and regression trees, gaussian Naive Bayes, and support vector machines to develop a model to forecast the grades of students taking the same course in the following term.[1] Based on the previous student's final examination score for the first-semester course, this study suggests a multiclass prediction model with six predictive models to forecast the grades of the final students. To assess the performance accuracy of student grade prediction, we specifically conducted a comparison analysis of integrating oversampling SMOTE with other FS approaches. We also demonstrated that oversampling was investigated. When employing all prediction models, SMOTE regularly outperforms FS alone.[2] This study presents a multiclass prediction model with six predictive models to anticipate the grades of the final students based on the prior student's final test scores for the first-semester course. We especially carried out a comparative analysis of integrating oversampling SMOTE with other FS techniques to evaluate the performance accuracy of student grade prediction. Additionally, we showed that oversampling was looked into. SMOTE consistently performs better when all prediction models are used than FS alone. [5] According to the literature, there is currently no accurate predictor of programming students' success because even extensive analyses of several characteristics have shown relatively modest predictive ability. Be aware that Deep Learning (DL) can deliver superior outcomes for both massive amounts of data and challenging issues. In this way, utilizing data gathered in the first two weeks from basic programming classes provided to a total of 2058 students over the course of six semesters, we used DL to predict students' success early on (longitudinal study). The cutting-edge Evolutionary Algorithm (EA), automatically builds and improves machine learning pipelines, as compared to our findings. The average accuracy of our DL model was 82.5%, which is significantly greater than the model developed and improved by the EA (p-value

0.05).[6] The survey's findings show that the data-level technique employing SMOTE oversampling is frequently used to identify unbalanced issues for predicting student grades. To improve the effectiveness of student grade prediction, hybrid and feature selection approaches are often not applied to enhance the generalization of the predictive model. In addition, a few of the proposed methodologies' advantages and disadvantages are reviewed and summarised to guide future study.[7] In this essay, we investigate the institutional data of a Nigerian university and analyze how its characteristics affect student achievement as indicated by cumulative grade point average (CGPA). Age and marital status show no statistically significant link with final CGPA, whereas gender and pre-entry score have a modest relationship but are not effective predictors, according to the statistical analysis of the data set. The use of four sample machine learning techniques—linear regression, support vector regression, decision trees, and random forests—shows that the third-year cumulative grade point average is a good indicator of the final year cumulative grade point average. Using the sum of the CGPAs from the preceding three years results in greater accuracy. The decision tree is the model that performs the worst in terms of forecasting the final CGPA, whereas support vector regression performs the best.[9] To forecast the grades of university students, we compare several cutting-edge machine-learning approaches in this work. In the end, we discover that Restricted Boltzmann Machines (RBM) are better at predicting student grades. These methodologies' anticipated grades represent the degree of uncertainty in student learning and may be applied to individualized advising, degree planning, confidence building, and the identification of potential students who might want extra help in particular courses.[10] Student retention is a major issue in higher education. While numerous approaches have been put out to address this issue, early and ongoing input can be very powerful. In this article, we provide a technique for estimating students' final course marks based only on their performance information from the most recent semester. It helps instructors recognize certain student types early on in the course to better support them and helps students analyze how much work they need to put into the course to get the grades they want. Our approach divides students initially into several groups depending on the experience points (XP) they accumulate over the course of a semester. Then, by creating and adding virtual students to the smaller clusters, we estimate cluster size and balance clusters. Finally, we use a feature selection strategy to remove irrelevant student traits. Then, using three alternative procedures, we project their ultimate grades.[11] Utilizing predetermined computer-marked tests, the learning progress of the pupils is assessed. For instance, when the student completes the online examinations, the computer provides rapid feedback. According to the study, in addition to the degree of involvement, the student success rate in an online course can be correlated with their performance in the preceding session. The literature hasn't received enough attention to determine if student interest and performance on earlier exams might have an impact on their performance on upcoming exams. While GBM produces the best accuracy in final student performance, an average value of 0.086 was attained, and the lowest RSME gain by RF obtained a value of 8.131 for the students' assessment grades model.[12] This study aims to identify vulnerable students and provide them with critical support by determining which characteristics of a student best predict whether they will graduate. The outputs of several machine learning algorithms are compared after being applied to the data. A Bayesian network was used to artificially produce the data utilizing factors such as a student's chosen major, their school quintile, their high school grades, and their NBT scores. The best results were with bagging, which accurately classified 75.97% of the data. [13]

III. Methodology

1. Data Cleansing
2. Evaluation Matrix
3. Algorithm Selection
4. Regression Models
5. Model Training
6. RMSE
7.

IV. ALGORITHMS

Regression Analysis: The link between a dataset's dependent (goal) and independent variables may be discovered using predictive modeling approaches like regression analysis. It is frequently employed when the goal variable has a range of continuous values and the dependent and independent variables are connected either linearly or non-linearly. To model time series, predict, and demonstrate causal links between variables, regression analysis techniques are used. Regression analysis is the ideal tool for a business to use when assessing the relationship between sales and advertising costs.

Linear Regression: The modeling method that is most frequently employed assumes a linear relationship between an independent variable (V) and a dependent variable (Y) (X). It employs a regression line, often known as a best-fit line. The equation for the linear relationship is Y = c+m*X + e, where c stands for the intercept, m for the slope, and e for the error term. The basic and complicated versions of the linear regression model, respectively, include different numbers of dependent and independent variables (with one dependent variable and more than one independent variable).

Decision Tree Regression: It is a tool with uses in several different fields. Both classification and regression issues may be solved using decision trees. The term itself implies that it displays the predictions that come from a sequence of feature-based splits using a flowchart that resembles a tree structure. The choice is made by the leaves at the end, which follow the root node.

LASSO Regression (Least Absolute Shrinkage and Selection Operator): Shrinkage is used in the linear regression method known as lasso regression. Shrinkage is the term used when data values move closer to a middle value, such as the mean. The lasso technique encourages simple, sparse models. (i.e. models with fewer parameters). This specific type of regression is well suited when models show high levels of multicollinearity or when you want to automate key aspects in the model selection procedure, such as variable selection and parameter removal.

Ridge Regression: Any data that exhibits multicollinearity can be analyzed using the model tuning technique known as ridge regression. This technique carries out L2 regularisation. Predicted values differ much from real values when the problem of multicollinearity arises, least-squares are unbiased, and variances are significant.

Random Forest Regression: As its name suggests, a random forest is made up of several independent decision trees that work together as an ensemble. The class with the highest votes becomes the prediction made by our model. The random forest's trees each spit forth a class prediction.

Gradient Boosting Regression: One well-liked boosting approach is gradient boosting. Each predictor in gradient boosting corrects the mistake of its predecessor. Unlike Adaboost, each predictor is trained using the residual errors of the predecessor as labels rather than adjusting the weights of the training examples. The Gradient Boosted Trees approach uses CART as its basic learner (Classification and Regression Trees).

Working of the proposed system: The proposed system uses labeled data for training a supervised machine learning model, which then generates results. a dataset including student data that is utilized throughout the training process. The dataset comprises tuples, connected to the student's academic as well as intimate data. The machine learning module is designed to forecast the student's success by monitoring their recreational and academic activities. To anticipate the grades of students by taking into account their most recent marks, the module will employ several machine learning algorithms, such as random forest, decision tree, linear regression, ridge regression, and lasso regression.

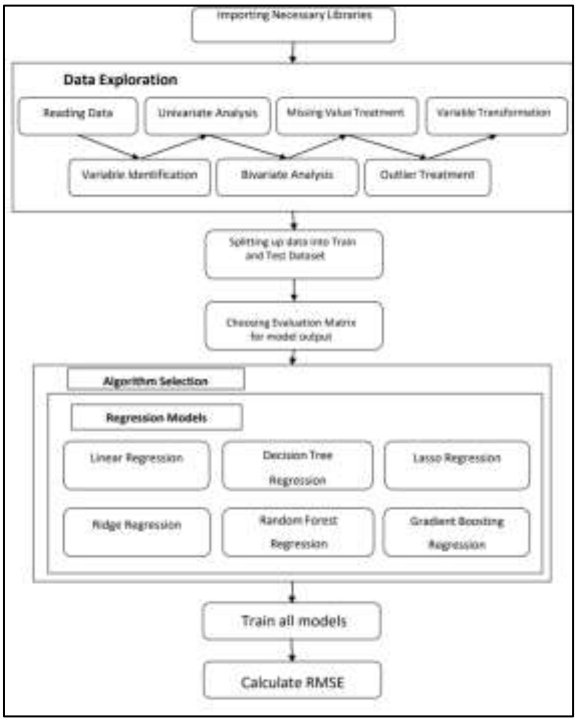

Fig. 1. Block Diagram of Academics Performance Prediction Using Machine Learning Algorithm

V. Conclusion

In conclusion, we can say that, surprisingly, the variables Dalc (Workday Alcohol Consumption) and Wale (Weekend Alcohol Consumption) were not significant, so I can say that alcohol consumption has little effect on students' grades. Instead, variables like failures, health, and free time had a much greater impact on students' grades and were significantly more significant than Dalc and Wale. We also discovered that all of the classification algorithms I used in my research yielded results with nearly identical accuracy levels, with Logistic Regression performing best with a score of 96.19, so I eliminated it from the study. Additionally, we discovered that PCA produced results that were identical to those of Logistic Regression and had no effect on them.

References

1. H. Gull, M. Saqib, S. Z. Iqbal, and S. Saeed, "Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning," 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-4, DOI: 10.1109/INOCON50539.2020.9298266.

2. S. D. A. Bujang et al., "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," in IEEE Access, vol. 9, pp. 95608-95621, 2021, DOI: 10.1109/ACCESS.2021.3093563.

3. H. Li, W. Li, Z. Zhang, H. Yuan and Y. Wan, "Machine learning analysis and inference of student performance and visualization of data results based on a small dataset of student information," 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2021, pp. 117-122, DOI: 10.1109/MLBDBI54094.2021.00031.

4. R. Hegde, A. G. V, S. Madival, S. H. S and S. U, "A Review on Data Mining and Machine Learning Methods for Student Scholarship Prediction," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 923-927, DOI: 10.1109/ICCMC51019.2021.9418376.

5. F. D. Pereira, E. H. T. Oliveira, D. Fernandes, and A. Cristea, "Early Performance Prediction for CS1 Course Students using a Combination of Machine Learning and an Evolutionary Algorithm," 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), 2019, pp. 183-184, DOI: 10.1109/ICALT.2019.00066.

6. R. Ayienda, R. Rimiru, and W. Cheruiyot, "Predicting Students Academic Performance using a Hybrid of Machine Learning Algorithms," 2021 IEEE AFRICON, 2021, pp. 1-6, DOI: 10.1109/AFRICON51333.2021.9571012.

7. Y. Luo, N. Chen and X. Han, "Students' Online Behavior Patterns Impact on Final Grades Prediction in Blended Courses," 2020 Ninth International Conference of Educational Innovation through Technology (EITT), 2020, pp. 154-158, DOI: 10.1109/EITT50754.2020.00034.

8. T. Hongthong and P. Temdee, "The classification-based machine learning algorithm to predict students' knowledge levels," 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer, and Telecommunications Engineering (ECTI DAMT & NCON), 2022, pp. 501-504, DOI: 10.1109/ECTIDAMTNCON53731.2022.9720334.

9. R. Suleiman and R. Anane, "Institutional Data Analysis and Machine Learning Prediction of Student Performance," 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2022, pp. 1480-1485, doi:10.1109/CSCWD54268.2022.9776102.

10. Z. Iqbal, A. Qayyum, S. Latif, and J. Qadir, "Early Student Grade Prediction: An Empirical Study," 2019 2nd International Conference on Advancements in Computational Sciences (ICACS), 2019, pp. 1-7, DOI: 10.23919/ICACS.2019.8689136.

11. H. Nabizadeh, D. Gonçalves, S. Gama and J. Jorge, "Early Prediction of Students' Final Grades in a Gamified Course," in IEEE Transactions on Learning Technologies, vol. 15, no. 3, pp. 311-325, 1 June 2022, DOI: 10.1109/TLT.2022.3170494.

12. R. Alshabandar, A. Hussain, R. Keight and W. Khan, "Students Performance Prediction in Online Courses Using Machine Learning Algorithms," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-7, DOI: 10.1109/IJCNN48605.2020.9207196.

13. N. Philippou, R. Ajoodha, and A. Jadhav, "Using Machine Learning Techniques and Matric Grades to Predict the Success of First Year University Students," 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (LIMITED), 2020, pp. 1-5, DOI: 10.1109/IMITEC50163.2020.9334087.