Higher Education Students Performance Evaluation Dataset Using Machine Learning

Abstract : *Nearly all nations place a significant and growing emphasis on education to speed up their growth. The classification of students' performance before they enter tests or take courses has also acquired relevance because well-educated people benefit their countries more. To improve students' individual performance and meet this objective, education quality must be improved during the current semester. The students' personal information, educational preferences, and family assets are some of the key indicators for providing this. In order to categorize student final grade performances and to find the most effective machine learning algorithm for this task, artificial intelligence techniques are applied in this paper to the questionnaire results of three different courses from two faculties that consist of these main indicators.*

Keywords - Student performance, Machine Learning, Random Forest, K means, SVC, KNN, GaussianNB, Decision Tree

Abbrevatins -

ML - Machine Learning
SVC - Support Vector Classifier
RF - Random Forest
AI - Artificial Intelligence
KNN - K- Nearest Neighbors
GNB - Gaussian Naive Bayes

1. Introduction:

In the modern world, where there is an increasing need for high-quality education, higher education has taken on more significance. Higher education organizations are therefore under pressure to deliver a top-notch education and guarantee that their pupils are adequately prepared for their future employment. Evaluating students' academic performance in order to pinpoint areas that need improvement and create powerful tactics to enhance student success is a crucial part of reaching this goal.

The research presented in the article titled "*Prediction of Higher Education students performance dataset*" outlines a study that was carried out with the dataset includes several different types of information, such as demographic data, educational background, course enrolment, course performance, and other pertinent indicators. This study aims to pinpoint the elements that have a substantial impact on students' academic achievement using statistical analysis and data mining methods.

With the help of a sizable dataset, the Higher Education Students Performance Evaluation dataset project seeks to offer a thorough examination of the academic performance of higher education students. Given the rising demand for high-quality *instruction*, accountability, and data-driven decision-making in higher education, this project is especially crucial. The project's main goal is to investigate the connections between numerous variables and students' academic performance in order to support teachers in creating powerful methods for enhancing student outcomes. Policymakers can utilize the project's findings to inform data-driven decisions that could result in better educational policies. The project's results will be especially helpful to higher education institutions that want to improve their performance and keep up with stakeholders' and students' rising expectations. Institutions can discover areas for improvement and create efficient methods to enhance students' achievement by employing data-driven insights.

The study, in its entirety, demonstrates the potential of machine learning techniques in predicting the student performance in the dataset with various algorithms. This dataset having **145** rows, **33** columns which we used in the project.

Review:

There were a total of five phases in this particular project. A literature review was the focus of the first stage. The study reviewed prior studies in the form of journal articles, questionnaires, research papers, and electronic books. This procedure was carried out to familiarize the reader with the current state-of-the-art, to highlight a research gap, and to support the on going study. These sources were located by using Google Scholar. The experiment's planning and design made up the second part. The study then looked at several training and comparison of the

frameworks for ML algorithms. Also, various a "scoring" techniques were taken into consideration while assessing the trained ML models.

The second phase and the third phase, or the implementation, were interconnected. The majority of the difficult labour and plan execution were done in this area. It involved learning and training the machine learning models using the student dataset after cleaning it with Excel. The grading, graphing, and comparing of the generated data made up the fourth phase. This was the method used to arrange and display the outcomes of the obtained scores from the ML models. The process of creating values from fresh incoming data using a trained machine learning model is known as scoring, or prediction. The calculated values or scores can be used to forecast future values as well as to represent a predicted category or occurrence.

Analyzing the findings from the measured data made up the fifth and final stage. To examine the trained models, various analytical tools and methodologies were taken into consideration. Also, some tools were selected to visualize the model as diagrams and charts.

Random Forest: Many decision trees, which are themselves data constructions that choose the rules/patterns from the incoming data, are used to create the classification approach known as RF. It uses input randomization and bagging to create each individual tree, which together produce an uncorrelated forest of trees whose collective prediction is more accurate than any one tree.

SVC: The supervised machine learning technique known as SVC, or Support Vector Classifier, is frequently used for classification problems. SVC separates the data into two classes by mapping the data points to a high-dimensional space and then locating the best hyperplane.

Gussian NB: A classification method used in machine learning (ML) called Gaussian Naive Bayes (GNB) is based on the probabilistic method and Gaussian distribution. Each parameter (also known as a feature or a predictor) is presumed to have an independent capacity to predict the output variable via Gaussian Naive Bayes.

K means: Unlabeled data can be divided into a specified number of disjoint groups of equal variance, or clusters, using the data clustering method K-means for unsupervised machine learning. Due to its simplicity of use and speed on huge datasets, it is a well-liked algorithm.

KNN: The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a non-parametric, supervised learning classifier that relies on closeness to produce classifications or predictions about the grouping of a single data point.

Conclusion

Pre-assessing student performance will be made easier with the use of this study, which will also help students develop their performance and take timely action to maintain it. In this study, the effectiveness of machine learning algorithms has been compared for the purpose of data mining educational institutions, and it has been demonstrated that these algorithms are promising in the area of predicting student performance, with Deep Learning being the most suitable algorithm for this kind of dataset. The success of any educational process depends greatly on the ability to predict student achievement. utilizing data mining and machine learning techniques to forecast their success based on academic data. Student records can also be used to explain a student's conduct, the effect of each factor on the student's progress in the educational process, the relationship between age stage, parental follow-up, and absence days. The use of machine learning algorithms to forecast student performance and rank the significance of various performance-related factors is discussed in this research. evaluating how well the machine learning algorithms performs when investigating educational data.

References:

1. YÄ±lmaz N., Sekeroglu B. (2020) Student Performance Classification Using Artificial Intelligence Techniques. In: Aliev R., Kacprzyk J., Pedrycz W., Jamshidi M., Babanli M., Sadikoglu F. (eds) 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions - ICSCCW-2019. ICSCCW 2019. Advances in Intelligent Systems and Computing, vol 1095. Springer, Cham.
2. Sultana, J., Rani, M. U., & Farquad, M. A. H. (2019).Student’s performance prediction using deep Learning and data mining methods. International Journal of Recent Technology and Engineering (IJRTE), 8(1S4), 2277-3878