Journal of Engineering Education Transformations ,
Volume 38 , No. 2 , October 2024 , ISSN 2349-2473, eISSN 2394-1707

10.16920/jeet/2024/v38i2/24200

A Study on Regression Based Machine Learning Models to Predict the Student Performance

Kamlesh V. Patil[1], Kiran D. Yesugade[2], Kiran B. Naikwadi[3]
[1,2,3]Bharati Vidyapeeth's College of Engineering for Women, Pune
[1]kamlesh.patil@bharatividyapeeth.edu
[2]kiran.yesugade@bharatividyapeeth.edu
[3]kiran.naikwadi@bharatividyapeeth.edu

Abstract : This article discusses the use of three regression models (Linear Regression, Decision Tree Regression, and Random Forest Regression) to study the performance of high school students in India across three subjects: Physics, Chemistry, and Mathematics. The study identifies various factors that affect student performance, such as access to good internet connectivity, parental educational background, and lunch quality. The data was obtained from an educational firm and analyzed based on principles and methods that aid decision-making processes. The results showed that all three regression models produced accurate and plausible results, with an overall accuracy of approximately 95%. The study's primary objective was to provide a clear and concise comparative analysis of various Machine Learning techniques and their impact on the dataset and the predictive attributes analyzed. The findings from this study underscore the importance of considering various factors when analyzing student performance and highlight the effectiveness of Machine Learning techniques in this domain.

Kamlesh V. Patil
Bharati Vidyapeeth's College of Engineering for Women, Pune
kamlesh.patil@bharatividyapeeth.edu

Keywords: Online Courses, Learning Analytics Dataset, Machine Learning, Tutor Marked Assessment, Receiver Operating Characteristic (ROC).

1. Introduction

A. Relevance

In today's world, universities must navigate an environment that is both extremely complex and intensely competitive. The most important obstacle that contemporary educational institutions need to overcome is to conduct an in-depth analysis of their performance, to determine what makes them special, and to devise a strategy for future development and actions (Kabakchieva, D., 2013). Because the ability of higher education institutions to meet the needs of their students is a major factor in determining the quality of the teaching process, accurate predictions of student achievement are absolutely essential. In this way, significant data and information is collected on a consistent basis, and it is then reviewed at the proper authorities. Standards for continuing to maintain the quality of the product are then established. The provision of services that most likely satisfy the requirements of students, academic staff, and other individuals involved in the education system is an essential component of quality in higher education institutions (Osmanbegovic, E., et al 2012). In the process of developing computerised learning

environments, one of the most important questions that must be answered is whether or not success within a learning environment predicts success outside of the environment. A significant portion of the efforts that have been put into data mining have mostly been on modelling and predicting performance within the context of the learning environment (Käser, T., et al 2017). Using the vast amounts of information that modern computers are able to store in their databases, a significant amount of research has been conducted on the topic of determining the factors that contribute to the poor academic results of students (school dropout and relapse) at various levels of education (primary, secondary, and higher), including primary, secondary, and higher (Natarajan, N.et. al. 2024). These data are a "gold mine" of information that can be used to better understand the children. It is a challenging undertaking to locate and extract important information that is buried within enormous databases. The application of techniques for knowledge discovery in databases or data mining in education, often known as educational data mining (EDM) (Márquez-Vera et. Al. ,2013),, is one option that shows a lot of promise for accomplishing this objective. In addition, today's higher education institutions provide students with access to a greater variety of specialised degree programmes than ever before, including dual degrees, honours degrees, interdisciplinary degrees, and many others. Even if all of this information is collected and the instructors have access to it, the sheer volume of data makes it impossible for them to manage (Rovira, S., et al 2017). Poor academic performance on such standardised assessments can therefore lead to unfavourable educational outcomes such as grade retention. When this issue is not addressed properly, it can eventually trigger dropout or sub-optimal career pathways, both of which are linked to major personal and social costs (Tamhane, A. et. al. 2014, Kanchana, S., et al 2024). Analysis of educational data, including learning analytics, academic analytics, educational data mining, predictive analytics, and learners' analytics, has emerged as an innovative field of study in recent years. Related terms include: academic analytics, learning analytics, and educational data mining. One thing that all of these definitions have in common is the utilisation of educational data for a variety of applications. Recently, a new phrase known as "educational data science" has been coined. This word elucidates how various academic fields and academics with varying research interests and backgrounds can collaborate in this field (Daud, A., et al 2017)..

Although a large number of studies have used machine learning to make predictions about students' performance in (Xing, W. et. al 2019), very few works have been done to investigate the performance trajectories (Gallego Arrufat,,2015). As a direct consequence of this, teachers were unable to track their students' progress in real time. In the course of this investigation, two distinct groups of experiments are carried out. In the first group of trials, an application of regression analysis is made for the purpose of estimating the students' test results. In order to forecast student outcomes, we take into account not only the student's previous and current activities, but also their previous performance. This was done by using data from previous trials. There have been three categories of potential predictors looked at, beginning with behavioral features, then moving on to temporal features, and finally moving on to demographic features. The models that have been developed provide new insights into selecting which learning activities are the most important, and they aid educators in maintaining accurate records of current student achievement. According to our best knowledge, the performance of students in online courses has only been evaluated based on two possible outcomes: "success" and "fail." Our model makes predictions about the performance using three different class labels: "success," "fail," and "Quit." The structure of the course is as shown in Fig. 1.

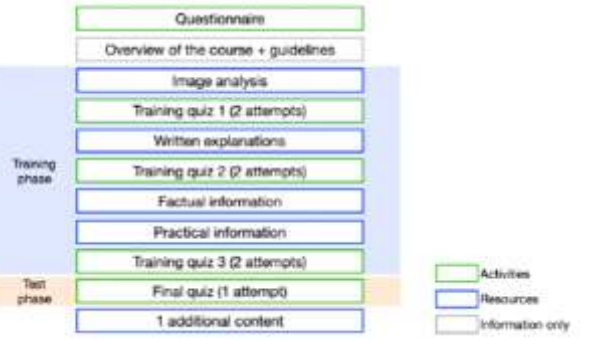


Fig. 1. Structure of the course. Uniformly with e-learning platforms common vocabulary, we call activity any content where participants need to interact with the course (questionnaires, quizzes, etc.) and +resource any content where participants only need to consult it (mostly learning material).

B. Research Hypothesis

The study suggests that various factors, such as the nutritional quality of school lunches, demographic variables (including gender and ethnicity, classified as rural, semi-urban, or urban), and the educational

attainment of parents (Diploma, Graduate, Postgraduate), collectively exert a substantial influence on students' academic performance. More precisely, the hypothesis suggests that there is a positive correlation between better meal quality, specific demographic backgrounds, higher levels of parental education, and improved academic achievements among students.

2. Literature Review

When it comes to determining how far forward a student is in their education, one of the most important indicators to look at is their success in virtual classrooms (Hew, K. F., et.al. 2014). Researchers have utilized a wide variety of approaches to monitor performance. The literature review that pertains to the relevant area is going to be discussed in this part. In the work (Osmanbegovic et. al. 2012,Käser, T. et. al 2017), Osmanbegovi and his colleagues conduct a comparison of various data mining strategies using a neural network, a Bayesian classifier, and with the assistance of decision trees. The challenge of classification has been overcome thanks to the neural network. It offers a variety of pattern accuracy methods and processes that are recognized, in addition to meaning algorithm. Because this platform is not very effective, the students will have the opportunity to select the area of the best subject in which they have the most interest through this implementation. As a result of these deficiencies, it does not offer a reliable method for determining the pupils' levels of performance. Another method that was more effective had been utilized. In the work that Christian and his colleagues (Renz, J., et al 2016) have done, the NBtree classification method is utilized to make predictions on the students' performance. The files include information regarding academics, education, and admissions, in addition to personal data collected by the students while they were enrolled in school. Weka toolbox for data mining is utilized. The paper is utilized in the process of constructing a model of categorization for the purpose of evaluating the performance of students. In addition to gender as an important attribute, GPA, credit, and test score are also utilized. When developing a model for predicting student performance, it is preferable to make advantage of the datasets made available by educational institutions of a higher level. When looking for more effective results, it is recommended to adopt strategies that use artificial intelligence. Grivokostopoulou et al. developed a method to assess the student's learning mechanism in addition to semantic rules; this method is helpful in estimating the levels of performance exhibited by pupils (Naren, J.,2014). You can find more information about this method here. Semantic rules and ontologies are also employed so that education can be improved, as well as the quality of what is delivered and the activities that are involved in learning. Adaptive learning makes use of a variety of different approaches to artificial intelligence. In order to answer in advance the question of whether or not a student will fail or pass the artificial intelligence class, an effective method known as a decision tree has been utilized, as have the scoring systems c4.5 and cart. Weka is utilized here for research purposes. The method that has been discussed up until this point does not offer a method that can be used on a bigger scale to evaluate and comprehend how a system functions. The study of the authors does not focus on examining the effectiveness of gender-related rules or the errors that occurred during examinations. In this approach, the management of the missing data is not done correctly. The cumulative GPA that was developed by Wang and his colleagues (Mishra, T., et al 2014) can be predicted using a straightforward model that is based on linear regression. In order to sketch the model, it took the analysis of the behaviors as an input. Students enrolled in Dartmouth College's undergraduate program have been used in research as data sets. For the purpose of making accurate predictions regarding the students' performances, a longitude measure of the students' lifestyle, style, and behaviors is utilized. There is data from Washington State University that can be used for the course that can be used to forecast the course. Based on the findings of an empirical investigation, Carter et al. (Arsad, P. M. et al 2013) suggested a normalized programming state model, from which a formula could be derived. This concept is inaccessible in programming environments suitable for beginners. The research is emphasized for programming ability and learning, but the findings of other studies are disregarded in terms of performance prediction. According to the proposed study that Mgala and his coworkers (Gray, 2014) have developed, it is based on the creation of a computer-based prediction tool. A model was developed on the basis of a huge dataset obtained from children in grades 3 through 8 who had participated in the Kenya certificate of primary education examination. The number of pupils who participated in the record was 2426. When dealing with missing data, the mean imputation method is utilized. Filters are employed in the feature selection procedure, and machine learning techniques, data mining algorithms, and other

preprocessing stages have been utilized to develop predictive models. Techniques for machine learning have also been applied. In some situations, such as an urban-based student analysis of college or university students, this method does not provide accurate results. Personal information about the student as well as information about the student's family that is important to expenditures are believed to be very helpful in determining the performance in advance, as was covered in in (Xing, Wet. Al.,2019, Verger, M., et al 2021). The primary objective of this term paper is to make predictions regarding the performance of learners and students by conducting a comparative examination of past work in the form of a survey. The purpose of this study is to provide a summary of the many methods that are utilized for performance prediction. Studies conducted for students' performance attributes from 2012 to the current day are dissected and debated, and this article also details the most accurate methods of prediction, which were not specifically mentioned in earlier polls.

The Item Response Theory concept (Cen, H., et al 2006) and the Factor Analysis Model (Koedinger, et al 2012) were proposed so that a student's progress in an Intelligent Tutoring System (ITS) may be predicted while also taking into account the level of difficulty of assessments (Cen, H., et al 2006, Koedinger, et.al.,2012). The level of difficulty of the tasks can be used to produce an evaluation of the correlation in between student's outcomes and the evaluation questions. This measurement can be used to determine the strength of the correlation. The FAM defines a set of response variable, such as the number of chances introduced to the student at every activity, the average length spent on each step, and the difficulty level of each question or latent variable, to determine the possibility of a student successfully completing a task can be computed. For example, the number of opportunities presented to the student at each task may be the same as the number of opportunities presented to the student at each step. The following are some examples of these variables: Based on the analyses, along with the latent variables in the predictions of student outcomes can make a significant contribution to the overall development of the model (Koedinger, et. al. 2012).The findings of the study lead the researchers to the conclusion that Learning Analytics (LAs), when combined with machine learning, are effective tools that have the capacity to monitor student knowledge. This was done so that the researchers could determine how much the activities of learners could affect the overall learning achievement of learners participating in MOOCs. The researchers demonstrated that machine learning could potentially help teachers by delivering information on the learning process to cohorts of students. Because of this, the researchers were able to visualise as well as assess the data that was gleaned from each level of the learner's progression. Attendance in such classes makes it possible to construct an accurate prediction model, hence it is recommended that you do so (Sinclair, J., et al 2016). It is possible to anticipate a student's eventual performance in an online course by using the student's marks from the initial assessment as well as their results on any quizzes given in conjunction with social factors (Maimon, O. Z., et al 2014). There were two different forecasting models presented. According to the findings, the number of peers who provided feedback was the factor that contributed the most to the overall quality of the distinction. It was determined that the most accurate predictor of who will obtain a certificate was the average score on each quiz. The accuracy of the distinction model was reported to have a percentage of 92.6%, whereas the accuracy of the normal model was reported to have a percentage of 79.6% (Sclater, N., et al 2016). At the University of Maryland, Baltimore County (UMBC), researchers looked into whether or not there was a correlation between the data collected from the Virtual Learning Environment (VLE) and student performance (Rastrollo-Guerrero,2020). Check My Activity (CMA) was able to make use of LA through its implementation of the tool. CMA is a learning analytics tool that may be defined as one that compares students' VLE activities with other activities and gives lecturers timely feedback on the emotional states of their students. According to the findings, students who participate in the class on a regular basis had a greater chance of earning a grade of C or above than their counterparts who did not participate on a regular basis (Breiman, L. ,2001).

3. Proposed Methodology

Nowadays, machine learning is found to be the best effective method for data analysis and data prediction. The best thing about the machine learning method is the ability to apply it in any field like manufacturing, image processing, academics, chemistry, pharmacy, social, etc. this is possible because in machine learning there is no prerequisite to develop or use the system equation for the analysis purpose. The basic principle of the machine is to train the model where the output response is described by the relationship with the input factor after developing

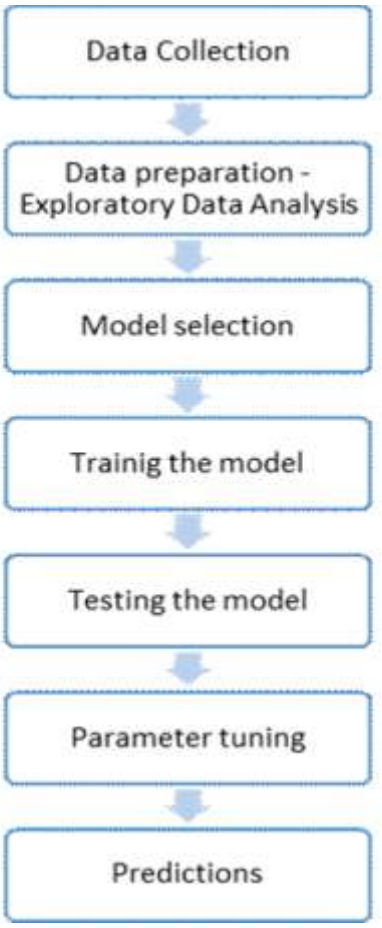


Fig. 2 : Basic steps in supervised machine learning

the training model developed relationship is applied to test with the testing purpose. The results of the testing phase decide the validity of the model for the selected problem study. Basically, there are two type of machine learning namely supervised learning, and unsupervised learning. Unsupervised learning requires the labeling of the factors or columns which are further grouped for cluster analysis. There is no training dataset thus there is no guarantee for the optimal solution validity (Maimon, O. Z., et al 2014). The student performance analysis comes under the supervised learning category where all the input and output factors are labeled. Simple steps in supervised machine learning are described in Fig. 2

A. Data Collection

The methodology utilizes an extensive data gathering framework, specifically targeting essential measures that are crucial for comprehending the success of students. The main data metrics comprised comprehensive evaluations of lunch quality derived from school records, surveys, and cafeteria observations; student demographics including gender and ethnicity, classified as rural, semi-urban, or urban based on school records; and parental education levels, divided into Diploma, Graduate, or Postgraduate categories through parent survey questions. In addition to these main measures, other variables such as enrollment in test preparation courses, scores in individual subjects (math, reading, writing), and results from IQ tests were first gathered. Nevertheless, in order to determine their significance to the primary goal of the study, additional assessment was necessary using feature selection and exclusion procedures.

B. Feature Extraction and Selection Process

The process of feature extraction and selection began with an initial statistical study to evaluate the possible influence of each gathered parameter on student performance. This phase encompassed conducting correlation studies and performing exploratory data analysis. Variables such as test preparation course enrollment, subject-specific scores, and IQ test results, which showed weak connections or minor impacts on student performance, were identified for removal. This elimination was strengthened with the implementation of redundancy checks, which involved detecting and eliminating variables that contained overlapping information. The ultimate set of features was subsequently improved by the application of dimensionality reduction methods, such as Principal Component Analysis (PCA), with a specific emphasis on comprehensive characteristics like lunch quality, student demographics, and parental education. The chosen characteristics were considered highly important in their association with student achievement, and so constitute the key components of our analytical dataset. Throughout this procedure, ethical considerations and data integrity were of utmost importance, with rigorous steps implemented to guarantee the precision and confidentiality of the gathered data. Table 1 shows the major factors affecting students' performance in

Table 1 : Factors affecting students performance in subject marks

Sr. No.	Gender	Ethnicity	Parent's Education	Lunch Quality
1	Female	Rural	Diploma	Highly-Nutritious
2	Male	Semi-urban	Graduate	Less-Nutritious
3		Urban	Postgraduate	Nutritious

subject marks in HSC exam. Data is collected for the HSC students where every student is appearing for the three subject exams namely Mathematics, Physics, and Chemistry. Final marks were calculated simply by taking an average of the three subject marks. For each subject, every student undergoes unit test I, unit test II, and the assignments. Based on these assessment marks were calculated for each student which is out of

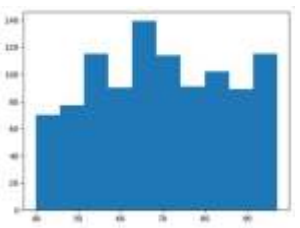
Fig. 3 : Steps in Exploratory Data Analysis

100 marks. To increase the accuracy of the machine learning model, it is necessary to have a large dataset therefore subject marks of 1002 students were selected for the analysis.

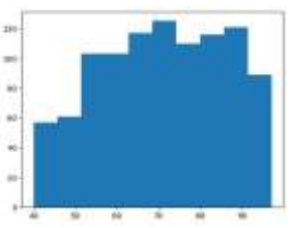
C. Data preparation – Exploratory Data Analysis (EDA)

This step plays the most important role in the machine learning process. In this step different operations are performed to make the data ready for the analysis where the machine learning model is applied. Various steps in the EDA are shown in Fig. 3.

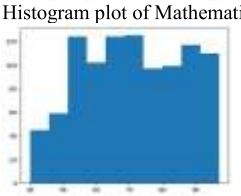
The first step in the data analysis includes the identification of the data where the nature of the data is studied for example; features having the data may be numerical, categorical, discrete, continuous, nominal, etc. Here, all the input features are having categorical data whereas all the output features are the subject marks that have the numerical data in a continuous manner. In data cleaning based on the interactions of the input features with the output features; unwanted



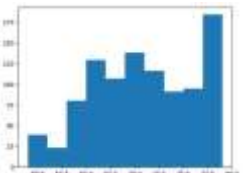
(a) Histogram plot of Mathematics



(b) Histogram plot of Physics



(c) Histogram plot of Chemistry



(d) Histogram plot of final marks

Fig. 4 : Histogram plots for different subjects

features are removed which do not affect the output features. Here all the input features are affecting the output features hence data cleaning was not required. To increase the accuracy; missing value analysis is performed where, missing values are identified and replaced by either the mean, mode, median, forward fill, or backward fill method. Outliers are simply identified by visualizing the box plot. For analyzing the student performance, all the students appeared for the exam, and the marks of the students are out of 100 hence missing value analysis and outlier analysis steps were avoided. In the univariate analysis, histogram plots, and bar plots of the feature are studied for the betterment of the model accuracy; data of the feature should be normally distributed and if the data are not normally distributed then by using scaling and normalizing data is made to be normally distributed.

The histogram plots of the students marks in mathematics, physics, chemistry, and final marks are shown in Fig. 4. From Fig 4, It is observed that all the data of the subject marks are almost normally distributed and thus it will increase the accuracy of the machine learning model. The bivariate analysis mainly includes three main analyses such as numeric-numeric analysis performed using scatter plots, numeric-categorical analysis performed using box plots and bar graphs, and lastly categorical-numeric analysis performed using bar graphs. The purpose of the bivariate analysis is to understand the relationship between the input and output features using which data cleaning is performed.

D. Model Selection

The detailed EDA clearly decides which type of machine learning model is selected. If the feature which is to be analyzed has continuous numerical data then regression models are applied and if the data is discrete or categorical then various classification models are applied. Model selection is mostly decided by univariate and bivariate data analysis. Student marks are having numerical data hence regression models such as linear regression, decision trees, and random forests are applied to the student data.

E. Training the model

Once the model is selected the dataset for input and output features is divided into a training dataset and a testing dataset. Generally, the ratio of the training dataset to the testing dataset is considered 70:30 or

80:20. If the training dataset has a large data then it may cause the overfitting of the model and if the training dataset has inadequate data then it may cause the underfitting of the model. Here the ratio of 70:30 was selected for the analysis where 70 % of the data is considered for the training purpose and 30 % of the data is considered for the testing purpose.

F. Testing the model

In machine learning, the actual model is developed in the training of the dataset step where the model establishes the relationship between the input features with the output feature. This developed model needs to be tested to validate its performance which is carried out in the testing of the model.

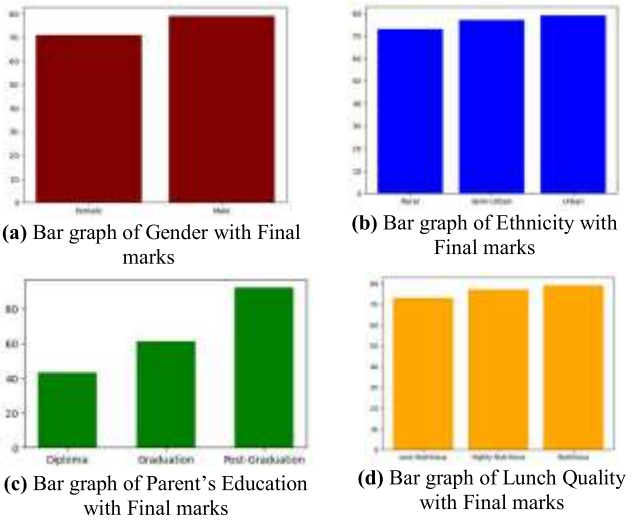
G. Parameter tuning

Parameters are the variables present in the model in which the programmer decides the accuracy of the model at a particular value of the parameter; model accuracy is maximum and such values are considered in the parameter tuning.

H. Predictions

This is the last step in machine learning where the prediction of the feature is performed based on the testing dataset. Errors in the predicted values and the actual values decide the accuracy of the model.

4. Results and Discussion



(a) Bar graph of Gender with Final marks

(b) Bar graph of Ethnicity with Final marks

(c) Bar graph of Parent's Education with Final marks

(d) Bar graph of Lunch Quality with Final marks

Fig. 5 : Bar graphs for final marks obtained in different groups

The objective of this research work is to identify the importance of the parent's education, Ethnicity, and quality of lunch provided to students on their marks. This section of the paper evaluates the determinants of student performance, with a focus on gender, socioeconomic background, parental education, and nutritional quality of lunches. Each factor is analyzed in relation to the students' final marks.

The analysis demonstrates a disparity in performance based on gender, with males attaining higher final scores than females (Figure 5a). The discovery implies that gender-related factors may play a role in the observed disparity in performance. It prompts significant inquiries on the fairness of educational resources and potential societal biases that may advantage one gender over the other in academic environments. Figure 5b of the study reveals a significant association between the socioeconomic backgrounds of the students and their academic performance. Students residing in urban settings demonstrate superior academic performance compared to their peers from semi-urban and rural locations. The difference can be ascribed to the enhanced availability of educational tools, such as the internet and specialized study areas, in metropolitan settings. These findings support the existing literature's claim that there is a direct relationship between the availability of resources and academic accomplishment. Moreover, there is a direct correlation between the educational attainment of parents and the academic achievement of students, as illustrated in Figure 5c. Students with parents who have greater educational levels, particularly post-graduate degrees, generally obtain higher grades. This result may indicate the amplified educational assistance and enhanced learning atmosphere that parents may offer, emphasizing the impact of cultural knowledge and resources on the achievement of students. The study suggests that students who have access to nutritious lunches tend to achieve higher scores in their final marks, as seen in Figure 5d. This finding corroborates the concept that sufficient nourishment is crucial for cognitive growth and academic achievement, underscoring the necessity for efficient nutritional initiatives in educational establishments.

The student's marks in their subjects are having numerical data, therefore, three regression models such as linear regression model, Decision trees regression, and random forests regression. In linear

regression; the model performs the regression task of the output feature with the input feature. In linear regression, there is a line that describes the relationship behavior of the output feature. In a simple way, linear regression is described by the equation of the line represented in equation 1.

$$Y=(m \times X)+c \quad \ldots\ldots.(1)$$

Where Y is the output feature and X is the input feature. 'm' is the slope of the line and c is the intercept. Linear regression describes the best fit of the line as compared with the other regression models. If the scatter plot of the output feature shows the nonlinear behavior then other regression models such as polynomial regression, and exponential regression models are used. Here in the data analysis of the student's marks, there are no interactions of the input features among each other therefore a simple relationship is required to describe the behavior of the output response hence simple linear regression model was used for the analysis.

The Decision Tree Regression is a supervised machine learning algorithm where the explanatory variable is divided into a number of decision branches. These branches resulted in the most relevant frequencies which are further divided into sub-branches based on the results of the parent node and this continues up to the final desired output where output frequencies are no longer desired (Koedinger, K. R., et al 2012). The linear regression model fails if the output feature has categorical data but in the case of the Decision Tree it is found to be equally beneficial for the continuous numerical data as well as categorical data. The Random Forest Regression is an improved or further modified version of the Decision Tree which includes the number of trees (Sclater, N., et al 2016). Random forest requires three main hyperparameters for training the model namely node size, number of trees, and the number of features sampled.

Table 2 : Regression Model Results for Final Marks

	Linear Regression			Decision Tree Regression			Random Forest Regression		
	M	F	Overall	M	F	Overall	M	F	O
Absolute Percentage Error (MAPE) (%)	0.9781	1.4726	1.1918	0.9769	1.5062	1.1959	0.9737	1.5045	1.1968
Model Accuracy	86.71	83.22	95.81	86.43	82.95	95.83	86.47	82.95	95.83

Table 2. Shows the results of the models obtained for the students with their marks. The student's marks dataset was analyzed in the python programming language of version 3.8.12.

From Table 2. It is observed that Random Forest regression and Decision Tree Regression have the same results where the overall accuracy of the final marks is 95.83 %. The overall accuracy of the linear regression is 95.81 % which is nearly the same as the overall accuracy of the Decision Tree Regression and Random Forest Regression. The mean absolute percentage error (MAPE) is a loss function that defines the accuracy of the machine learning model. The more MAPE less the accuracy of the model and vice versa. The MAPE is calculated by using equation (2).

$$MAPE = \frac{1}{n} \times \sum_{i=1}^{n} \left| \frac{Actual\ value - Predicted\ value}{Actual\ value} \right| \qquad (2)$$

The performance of the male student was found to be better than the female student. For this research work, the accuracy of all three models is approximately 95.00 %. However, Decision tree model allows more flexibility and scalability as compared to the rest of two models making it more suitable to be used in this study. A comparative study of all three models thus validates the selection of the model for predicting the final marks of the students.

Conclusion

The objective of this research work is to apply the machine learning model to predict student marks. Based on the proposed methodology and discussions, the following conclusions are determined.

- The EDA plays important role in any machine learning analysis which helps to decide the appropriate model to analyze the given data.
- Obtained overall accuracy of the Linear Regression, Decision Tree Regression, and Random Forest Regression for the student's marks is nearly equal to 95.00 %
- The Female students and the male student's data analysis shows that the male students have been found better performance as compared with the

female students.

- The number of female students and the number of male students was equal in size that is 501 each and applying the machine learning model reduces the accuracy of all three models which concludes that by increasing the data size the accuracy of the machine learning model increases proportionally.

- The parental level of education shows most significant impact on the students performance in subject marks as compared to other metrices.

- Comparative study of the machine learning model helps to validate the results and gives a choice to select the best model among all applied machine learning models.

The study focused primarily on assessing multiple factors that impact student performance in High School Certificate (HSC) tests. After considering several characteristics such as gender, ethnicity, the quality of school lunches, and the education level of parents, it was determined that parental education level had the most substantial influence on student performance. This discovery emphasizes the significance of the educational background of parents as a crucial aspect that can be enhanced to improve students' academic achievements. The study findings emphasize the significant impact of familial educational environments on students' academic accomplishments, indicating that interventions focused on improving parental involvement and support could play a crucial role in strengthening student performance in HSC exams..

Declarations

Funding – Not applicable

Conflicts of interests - The authors declare that they have no conflict of interest.

Ethics approval – Not applicable

Consent for publication – Not applicable

Availability of data and material – Not applicable

Code availability – Not applicable

References

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. Cybernetics and information technologies, 13(1), 61-72.

Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. Economic Review: Journal of Economics and Business, 10(1), 3-12.

Käser, T., Hallinen, N. R., & Schwartz, D. L. (2017, March). Modeling exploration strategies to predict student performance within a learning environment and beyond. In Proceedings of the seventh international learning analytics & knowledge conference (pp. 31-40).

Natarajan, N., Vasudevan, M., & Venkatesh, A. (2024). A Simple Computer Code for Result Analysis of University Graduates. Journal of Engineering Education, 38(1).

Márquez-Vera, C., Morales, C. R., & Soto, S. V. (2013). Predicting school failure and dropout by using data mining techniques. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 8(1), 7-14.

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout. PLoS one, 12(2), e0171207.

Tamhane, A., Ikbal, S., Sengupta, B., Duggirala, M., & Appleton, J. (2014, August). Predicting student risks through longitudinal analysis. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1544-1552).

Kanchana, S., & Cherukuri, J. (2024). Augmenting Teaching-Learning Process Through Discussion Forum and Padlet. Journal of Engineering Education Transformations, 38(1).

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017, April). Predicting student performance using advanced learning analytics. In Proceedings of the 26th international conference on world wide web companion (pp. 415-421).

Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized

intervention. Journal of Educational Computing Research, 57(3), 547-570.

Gallego Arrufat, M. J., Gamiz Sanchez, V., & Gutierrez Santiuste, E. (2015). Trends in assessment in massive open online courses. Educacion Xx1, 18(2), 77-96.

Hew, K. F., & Cheung, W. S. (2014). Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. Educational research review, 12, 45-58.

Ramesh, V. A. M. A. N. A. N., Parkavi, P., & Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. International journal of computer applications, 63(8).

Naren, J. (2014). Application of data mining in educational database for predicting behavioural patterns of the students.

Renz, J., Schwerer, F., & Meinel, C. (2016). openSAP: Evaluating xMOOC Usage and Challenges for Scalable and Open Enterprise Education. Int. J. Adv. Corp. Learn., 9(2), 34-39.

Mishra, T., Kumar, D., & Gupta, S. (2014, February). Mining students' data for prediction performance. In 2014 Fourth International Conference on Advanced Computing & Communication Technologies (pp. 255-262). IEEE.

Arsad, P. M., & Buniyamin, N. (2013, November). A neural network students' performance prediction model (NNSPPM). In 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA) (pp. 1-5). IEEE.

Gray, G., McGuinness, C., & Owende, P. (2014, February). An application of classification models to predict learner progression in tertiary education. In 2014 IEEE International Advance Computing Conference (IACC) (pp. 549-554). IEEE.

Verger, M., & Escalante, H. J. (2021). Predicting students' performance in online courses using multiple data sources. arXiv preprint arXiv:2109.07903.

Cen, H., Koedinger, K., & Junker, B. (2006, June). Learning factors analysis–a general method for cognitive model evaluation and improvement. In International conference on intelligent tutoring systems (pp. 164-175). Springer, Berlin, Heidelberg.

Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (2012). Automated Student Model Improvement. International Educational Data Mining Society.

Sinclair, J., & Kalvala, S. (2016). Student engagement in massive open online courses. International Journal of Learning Technology, 11(3), 218-237.

Maimon, O. Z., & Rokach, L. (2014). Data mining with decision trees: theory and applications (Vol. 81). World scientific.

Sclater, N., Peasgood, A., & Mullan, J. (2016). Learning analytics in higher education. London: Jisc. Accessed February, 8(2017), 176.

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. Applied sciences, 10(3), 1042.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.