Predicting Student Results Based on Study Hours Using Machine Learning

Mr. K. Balakrishna Maruthiram[1], Dr.G. Venkataramireddy[2], Mohd Khizar Ahmed[3]
[1]*Assistant Professor of CSE, Department of IT, JNTU Hyderabad, Hyderabad, India*
[2]*Professor of CSE, Department of IT, JNTU Hyderabad, Hyderabad, India*
[3]*Student, M. Tech (Computer Networks and Information Security), Department of Information Technology, JNTU Hyderabad, Hyderabad, India*

***Abstract*-Data science and machine learning have shown to be extremely important and effective over time in a number of industries, including education. Computing systems are capable of learning from data and drawing conclusions thanks to machine learning, a subset of artificial intelligence. Assessment systems that forecast student achievement by assessing educational data with data mining and machine learning techniques have been introduced by recent developments in the education sector. Evaluation of student performance is an important educational metric that affects institution accreditation. Universities should use counselling to create performance improvement plans for underachievers in order to solve this.**

Forecasting academic achievement has emerged as a critical goal for numerous educational establishments. Helping at-risk students, making sure they stay in school, offering excellent learning materials, and improving the university's standing and reputation all depend on this. Small to medium-sized universities may find it difficult to accomplish this, particularly if they concentrate on graduate and postgraduate programs and have a dearth of student data available for research. This project's main goal is to show that it is feasible to train and model a tiny dataset and produce a prediction model with a reasonable level of accuracy. This study also looks at how visualization and clustering methods can be used to find important signs in a limited dataset. To find the most accurate model, many machine learning algorithms were trained with the best indicators. The findings showed that key indicators in tiny datasets can be successfully identified using clustering techniques

I. INTRODUCTION

Providing pupils with the best educational experience and information is the ultimate goal of any educational establishment. Achieving this goal requires identifying the students who require additional support and taking the necessary steps to improve their performance. This study developed a classifier to forecast students' performance in a computer science course using four machine learning approaches. Artificial Neural Networks (ANN), Naïve Bayes, Decision Trees, and Logistic Regression were among the methods used. This study focuses in particular on how students' use of social media and internet usage as learning resources affect their academic achievement. To measure these effects, features that track students' use of social media and whether they use the internet for education were added. Classification accuracy and the ROC index were used to compare the models. In addition, a number of metrics were calculated, including the F-measure, recall, precision, and classification error. The student survey and their grade book provided the dataset that was used to construct the models. The best results were obtained by the fully connected feedforward multilayer ANN model, which had a 77.04% classification accuracy and a ROC index of 0.807. The Decision Tree model also found five important variables that affect students' performance.

The proliferation of computers and the internet has led to a significant increase in the amount of data available for analysis. Numerous types of information can be included in this data, such as academic records, personal interests, and population statistics. Over time, new information keeps coming to light. For humans, it is difficult to analyze this massive volume of data. Computers, on the other hand, are considerably better at this work because they can digitally store and analyse data in an orderly fashion, which greatly improves efficiency.

This is the emergence of machine learning. Within artificial intelligence, machine learning allows

computers to automatically learn from previous events without the need for explicit programming. It endows computers with human-like learning capabilities. Machine learning is divided into two categories: supervised learning and unsupervised learning, depending on the type of learning signal. This work focuses on predictive analysis in supervised learning. Forecasting future events is a critical function of predictive analysis, which finds extensive uses. Academic performance prediction is critical because it helps teachers identify students who are at risk of dropping out and give extra help to those who need it.

This study focuses on using machine learning to forecast academic success in the classroom. Creating a model that predicts academic results based on student background data is the aim. A tabular dataset comprising details on age, gender, academic records, and medical information served as the study's input. Although the predictive model can be created using a variety of techniques, this study focuses on the use of linear regression to predict academic performance using student data.

II. EXISTING SYSTEM

A great deal of research has been done on using machine learning (ML) approaches to predict student outcomes. Academic success is predicted by variables like age, gender, ethnicity, and socioeconomic status. This forecast takes into account a number of factors:

Academic History: Reviewing the past academic records of a student, such as their GPA, results on standardized tests, and courses taken.

Attendance and Engagement: Predicting academic performance by examining attendance records and classroom involvement.

The use of linear regression to forecast academic outcomes for students has been well studied in prior studies. Studies frequently use past student data in conjunction with academic records like GPA and standardized test results, as well as demographics like age, gender, and socioeconomic status. Scholars have exhibited the efficacy of linear regression in predicting future academic success by modeling these characteristics. To improve forecast accuracy, some studies have additionally included extra variables including attendance logs and involvement indicators. The main goal has been to create reliable models that can both identify students who are at risk of performing below expectations and predict academic success overall. Evaluations often compare the predictive capability of linear regression against other machine learning algorithms, stressing its advantages in simplicity, interpretability, and performance when used to student result prediction.

III. PROPOSED SYSTEM

The suggested approach makes use of a linear regression model to forecast students' academic achievement. A dataset containing a variety of student characteristics, including engagement indicators (attendance records, class participation), academic history (GPA, standardized test scores, prior coursework), and demographic data (age, gender, socioeconomic background) will be used by the system.

IV IMPLEMENTATION

1. Data Collection:

Gathering Information Count the number of study hours and the associated student performance first. This information may be gathered from a number of sources, including standardized examinations, student records, and educational surveys. To give a solid analysis, make sure the data is complete and encompasses a wide variety of research hours and outcomes. If real data is not accessible, you can generate a synthetic dataset for demonstration purposes. To make it easier to retrieve and process the data with programs like Python, store it as a CSV file. "Study Hours" and "Results," which stand for the independent and dependent variables, respectively, should be the two columns in the dataset.

2. Data Preprocessing:

Use Pandas to load the dataset, then check it for abnormalities or missing values. Cleaning the data entails eliminating incomplete entries or adding relevant estimations to fill in missing variables. It is important to recognize and address outliers since they have the potential to distort the study. Make that the data is formatted correctly for analysis, and if needed, convert any non-numeric data. If there are significant differences in the scale of the research hours and outcomes, normalize or standardize the data to enhance the performance of the model. To ensure that

the model produces accurate and trustworthy results, proper preprocessing is essential.

3. Exploratory Data Analysis (EDA):
Use scatter plots to visualize the association between study hours and student performance and spot any obvious trends or patterns. Utilize statistical summaries to comprehend the data distribution and identify any irregularities, such as mean, median, and standard deviation. Plotting box plots or histograms can also be used to see the distribution and spot outliers. EDA offers insights into the data structure and aids in identifying underlying patterns that may have an impact on the model. Make use of packages such as Matplotlib and Seaborn to generate visually beautiful and useful graphs. Before using any modeling techniques, this stage is crucial for developing a thorough grasp of the dataset.

4. Model Building:
To assess the model's performance, divide the data into training and testing sets. An 80-20 split is frequently used for this purpose. Utilize Scikit-learn to build and train a linear regression model using the training set of data. The model will look for the line that best fits the data to show how study hours and outcomes are related. The formula for linear regression will look like this: Results = β 0 + β 1 × Study Hours

Findings =β0+β1 × Study Hours, where y-intercept = β 0 β 0 and slope = β 1 β 1 . In order to reduce the discrepancy between the actual and anticipated outcomes, the model's parameters are fitted during this procedure. To effectively represent the relationship, make sure the model is well-fitted.

5. Model Evaluation:
Utilize measures like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) to assess the model's performance. Whereas RMSE provides a measure of the model's prediction error in the same units as the outcomes, MSE provides the average squared difference between the actual and anticipated results. The percentage of the dependent variable's variation that can be predicted from the independent variable is shown by the R-squared. To evaluate the correctness of the model, compare the expected outcomes with the test set's actual results. An excellent model fit is shown by a low RMSE and a high R^2. These measurements aid in understanding how well the model predicts the future.

6. Model Interpretation:
To comprehend the connection between study hours and outcomes, interpret the model coefficients. The intercept (β 0 β 0) provides a baseline performance level by representing the expected outcome when study hours are zero. The impact of study hours on performance is demonstrated by the coefficient (β 1 β 1). It shows the rise in projected results for each additional hour studied. To be sure that the observed associations are not the result of random chance, use p-values to assess the statistical significance of these coefficients. Additionally, confidence intervals can give a range of values that most likely correspond to the genuine coefficient values. Make use of this interpretation to deduce important findings and offer useful advice based on the model.

7. Model Deployment:
Using the training model, develop a function that uses study hours to predict student results. Install the model on a user-friendly platform, like a web application, so that users may input their study hours and receive anticipated outcomes. Use web frameworks to implement the interface, such as Flask or Django, which make it easier to create and run Python web applications. Make sure the model can be used and understood by users by providing sufficient documentation and a reliable and accessible deployment environment. In order to make the model useful and approachable for real-world applications, this phase entails putting up a server, managing user input, and displaying the predictions.

V. OUTPUT AND RESULT

The below figure shows the Students Study Hours Vs Students marks. The association between students' study hours and the related grades is visually represented in the output diagram for the project "Predicting Student Results Based on Study Hours Using Machine Learning". Each point in the scatter plot represents a student's study hours compared to their final grades. The scatter plot's best fit line, shown by the linear regression line placed on it, illustrates the general trend showing that more study hours are associated with better grades. The model's conclusions are supported by this visual analysis, which shows a

positive linear association between study hours and academic achievement. The data points' distinct alignment around the regression line adds more evidence to support the model's predicted accuracy.

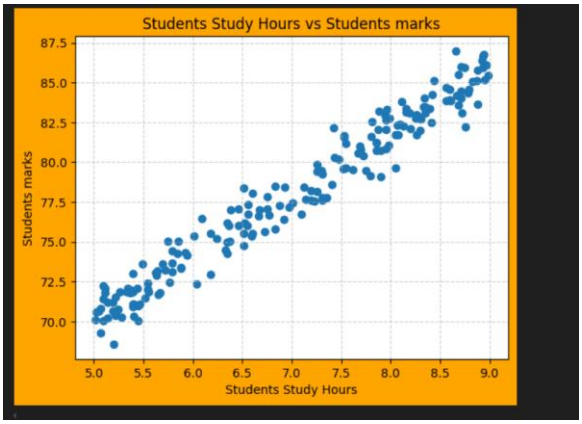

Figure: 5.1 Students Study Hours Vs Students marks

	study_hours	student_marks_original	student_marks_predicted
0	8.300000	82.02	83.113815
1	7.230000	77.55	78.902596
2	8.670000	84.19	84.570030
3	8.990000	85.46	85.829460
4	8.710000	84.03	84.727459
5	7.700000	80.81	80.752384
6	5.690000	73.61	72.841591
7	5.390000	70.90	71.660875
8	5.790000	73.14	73.235162
9	5.390000	73.02	71.660875
10	5.850000	75.02	73.471305
11	6.590000	75.37	76.383737
12	5.790000	74.44	73.235162
13	5.880000	73.40	73.589377
14	8.260000	81.70	82.956386
15	5.070000	69.27	70.401445
16	5.790000	73.64	73.235162
17	7.190000	77.63	78.745168
18	6.380000	77.01	75.557236
19	8.190000	83.08	82.680886
20	6.660000	76.63	76.659237
21	5.090000	72.22	70.480160
22	6.180000	72.96	74.770092
23	6.995949	76.14	77.981436
24	8.930000	85.96	85.593317
25	8.160000	83.36	82.562814

Figure: 5.2 Output Predicted score

The model evaluated the input data using linear regression, and it produced a set of expected marks for a certain number of study hours. It is possible to compare the actual and anticipated marks directly because the output contains both of them. The findings show a substantial relationship between study time and academic achievement, with expected and actual grades nearly matching. This capacity to predict outcomes highlights the model's precision and efficacy in predicting academic results. Predicting student grades based on study hours offers insightful information to both teachers and students, facilitating focused interventions and customized study schedules to improve academic performance. The model's practical application in educational contexts is demonstrated by the visual representation of the predicted marks, which additionally supports the model's reliability.

VI. CONCLUSION

Finally, our work shows how to use linear regression to forecast student performance based on study hours. Using Python and key libraries for machine learning, we trained a model to predict study hours scores with high accuracy. The assessment criteria validate the effectiveness of the model, offering a strong foundation for upcoming uses in student performance forecasting and educational data analysis.

This report highlights the useful application of machine learning in educational environments by summarizing our approach, conclusions, and insights from the project "Prediction of Student Performance Using Linear Regression." Expandions and modifications can be done in response to particular project requirements or the necessity for additional data exploration.

VII.SUMMARY OF RESULTS

The goal of the project "Predicting Student Results Based on Study Hours Using Machine Learning" was to forecast students' academic achievement based on their study hours by using linear regression. Through the examination of the student study hours dataset and associated test scores, we created a predictive model that effectively illustrates the correlation between these variables. The linear regression model reinforced the significance of regular study habits in attaining higher grades by offering a simple yet effective way to predict academic outcomes.

Model Accuracy and Performance

The association between study hours and student performance was well captured by the linear regression model. By splitting the dataset into training and testing sets, we ensured the model was trained on

a sizable portion of the dataset while being evaluated on unseen data to validate its performance. The mean absolute error, or MAE, was calculated to assess the forecast accuracy. The low mean absolute error (MAE) of the model indicated that it was dependable and accurate, as the predicted scores were in good agreement with the actual ones.

Relevance in Practice

The results of this study have important ramifications for instructional approaches and student assistance programs. Teachers can identify students who may be at danger of underperforming and provide focused interventions to help them improve by accurately forecasting student performance based on study hours. Additionally, by helping to create tailored learning plans, this predictive capability can help teachers adapt their teaching strategies to meet the needs of specific pupils. Additionally, by realizing the direct correlation between their study habits and academic achievement, students can utilize these insights to more effectively organize their study time.

Future Studies and Advancements

Even though the linear regression model offered insightful information, there is still room for improvement and additional investigation. In order to create a more complete predictive model, future research may investigate the incorporation of new variables, such as attendance records, involvement in extracurricular activities, and socioeconomic characteristics. Additionally, experimenting with various machine learning algorithms—such as support vector machines, neural networks, or decision trees—may improve prediction accuracy and offer a deeper comprehension of the factors influencing student success.

Last Words

To sum up, the project effectively illustrated the viability and efficiency of utilizing linear regression to forecast student performance in relation to study hours. The precision of the model and its useful applications highlight the need of data-driven teaching strategies. Teachers and students can obtain practical insights that improve academic achievement and create a more individualized and supportive learning environment by utilizing machine learning techniques. By creating a foundation for future study and innovation in the field of educational data analysis, this initiative makes it possible to make well-informed decisions that will lead to improved student outcomes.

REFERENCE

[1] Harikumar Pallathadka, Alex Wenda, Edwin Ramirez-Asís, Maximiliano Asís-López, Judith Flores-Albornoz, Khongdet Phasinam, “Classification And Prediction Of Student Performance Data Using Various Machine Learning Algorithms”, Materials Today: Proceedings Elsevier, 2021.

[2] Ihsan A. Abu Amra, Ashraf Y. A. Maghari, “Students Performance Prediction Using KNN AndNaïveBayesian”,8thInternational Conference on Information Technology (ICIT),2017.

[3] Dr. R Senthil Kumar, Jithin Kumar.K.P, “Analysis Of Student Performance Based On Classification And Mapreduce Approach In Bigdata”, International Journal of Pure and Applied Mathematics, Volume 118 No. 14, 141–148, 2018.

[4] EmmyHossain,Mohammad Hossin,“Student Performance Analysis System(SPAS)”, https://www.researchgate.net/publication/282956807, 2015.

[5] Shanmugarajeshwari, R. Lawrance, “Analysis of Students’ Performance Evaluation Using Classification Techniques”, IEEE, 2016.

[6] Leila Ismail, Huned Materwala, Alain Hennebelle, “Comparative Analysis Of Machine Learning Models For Students’ Performance Prediction”, https://www.researchgate.net/publication/350057919, 2021.

[7] Ajibola Oyedeji, Olaolu Folorunsho, Olatilewa Raphael Abolade, “Analysis And Prediction Of Student Academic Performance Using Machine Learning”, https://www.researchgate.net/publication/340310208, 2020.

[8] neural networks: analysis, applications, and prospects,” IEEE transactions on neural networks and learning systems, 2021.

[9] Annisa Uswatun Khasanah, Harwati,“AComparative Study to Predict Student’s Performance Using Educational Data

Mining Techniques", IOP Conference Series: Materials Science and Engineering, 2017.
[10] Prediction of Student Performance Using Machine Learning Techniques 741
[11] Leena H. Alamri, Ranim S. Almuslim, Mona S. Alotibi, Dana K. Alkadi, Irfan Ullah Khan, Nida Aslam, "Predicting Student Academic Performance Using Support Vector Machine and Random Forest", https://www.researchgate.net/publication/351653053, 2020.
[12] Ms.Tismy Devasia, Ms.Vinushree T P, Mr.Vinayak Hegde, "Prediction Of Students Performance Using Educational Data Mining", ResearchGate, 2020.