

Student Performance Prediction Through Multiple Linear Regression

Submitted By

Student Name	Student ID
Md. Merajul Alam Meraj	221-15-5632
Abdullah Al Mamun Sheikh	221-15-5530
Likhon Bhuiyan	221-15-5966
Miasha Alam	221-15-5668

MINI LAB PROJECT REPORT

This Report Presented in Partial Fulfillment of the course **CSE326: Data Mining and Machine Learning Lab** in the **Computer Science and Engineering Department**



DAFFODIL INTERNATIONAL UNIVERSITY

Dhaka, Bangladesh

December 2, 2024

DECLARATION

We hereby declare that this lab project has been done by us under the supervision of **Md. Abdullah Al Kafi, Lecturer**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere as lab projects.

Submitted To:

Md. Abdullah Al Kafi

Course Teacher's Name

Lecturer

Department of Computer Science and Engineering
Daffodil International University

Submitted by

Md. Merajul Alam Meraj ID: 221-15-5632 Dept. of CSE, DIU	Abdullah Al Mamun Sheikh ID: 221-15-5532 Dept. of CSE, DIU
Likhon Bhuiyan ID: 221-15-5966 Dept. of CSE, DIU	Maisha Alam ID: 221-15-5668 Dept. of CSE, DIU

DECLARATION	2
1. Introduction	4
1.1 Introduction	4
1.2 Literature Review	4
1.3 Importance	4
1.4 Gap	4
1.5 Conclusion	4
2. Literature review	5
2.1 Introduction	5
2.2 Literature Review	5
2.3 Gap	5
3. Methodology	5
3.1 Introduction	5
3.2 Methodology	6
3.3 Conclusion	8
4. Experimental Result	8
4.1 Introduction	8
4.2 Result	8
4.3 Model Evaluation	9
4.4 Discussion	9
5. Conclusion	9
6. Reference	10

1. Introduction

1.1 Introduction

For an educational institution, the performance of a student has a significant amount of value as it also highlights the overall quality of the institute. So, it is very essential to calculate the outcomes of the student and take necessary measures to improve the quality of the teaching and learning methods. To do so, predicting the performance of a student is a necessity.

This paper aims to apply machine learning algorithms to predict students' performance based on different attributes and their importance to the performance. The data attributes used to predict students' performance can include many features such as sleep hours, study hours, extra curricular activities, previous marks etc.

One of the methods that was widely used by researchers to predict students' performance is regression. Regression is a supervised machine learning technique that shares the same concept as classification in using a training dataset to make a prediction. The difference between them is that the output variable in classification is categorical while in regression is numerical.

In this project, the Multiple Linear Regression (MLR) model is used to find the most significant features. This will help find the feature which contributes most to the performance and which feature does not have little to less contribution. The dataset used to train the model consists of 10,000 information of students.

1.2 Literature Review

Previous studies have used several techniques to accomplish the prediction of students' performance. One way to predict students' performance is by applying data mining techniques on data that comes from educational databases, this process called educational data mining. Many have used all the features to make the prediction without prioritizing the key features. In this paper, various methods are used to find the most important features and predict the performance accordingly.

1.3 Importance

Predicting a student's performance is very important for an educational institution. This will help the educational institution to take necessary measures to enhance the performance of a student and also learn their lacking. By predicting the causes, the tutors can provide the required help in order to overcome this.

1.4 Gap

Although there are several works based on predicting the performance of a student with different features, there is a lack of work done on finding the significant features or variables which are responsible mostly for the performance index. This study focuses on the feature selection and using MLR to predict the value.

1.5 Conclusion

The introduction section concludes with introducing the problem statement, importance of the problem and gap on this related work and objective which is building the MLR model to predict the students' performance based on different attributes.

2. Literature review

2.1 Introduction

This section of the paper covers the related literature work related to educational data mining and student prediction. Here different techniques and models used in other work are discussed thoroughly.

2.2 Literature Review

Modeling for predicting performance of a student is one of the main objectives in educational data mining. Mostly there are two main tasks to be performed; prediction and structural overview. In prediction there are also two sub categories which are predicting the undesirable students' behaviors, and predicting the students' characteristics such as learning styles and performance.

Basically two types of data mining techniques are used in this prediction; classification and regression. Classification models are applied where the output is categorical; on the other hand, regression models are applied to predict continuous variables. Many classifications are applied by the researchers like KNN, Decision Tree or Naive Bayes.

Decision Tree is widely used in performance prediction by the researchers. For example, Quadri, M.M. et al [1] applied a decision tree model to predict the drop out feature. Mishra, T. et al [2] in his research paper implemented a decision tree technique to build a performance prediction model based on students' social integration.

Naive Bayes is also popular among the researchers to predict students' performance. Mayilvaganan et al [3] used different classification models including naive bayes to predict the performance and also to compare the models with each other. Jishan, S.T. et al. [4] proposed a Naive Bayes model to improve the accuracy of the students' final grade prediction model for a particular course.

Apart from classification models, there are many researchers who used regression models to predict the students' performance. Arsal et al. [5] applied a comparison study between Artificial Neural Network (ANN) and Linear Regression (LR) in predicting academic performance. Yang, S.J.H. et al [6] combined a multiple linear regression and principal component analysis to establish a more accurate prediction model.

2.3 Gap

Although there are many research published to predict the students' performance using different techniques like classification and regression, these papers tend to use all the available attributes to predict the performance of a student. Before building a model, no attributes or variables are taken into consideration as an important one or not.

3. Methodology

3.1 Introduction

The methodology is the section where all the methods used to collect, preprocess the data, training and testing the model. The model used in this paper is a regression based mode.

In linear regression, there are mostly two models included, a simple linear regression model which predicts a dependent variable based on an independent variable and a multiple linear regression model which predicts the dependent variable based on multiple independent variables.

The equation for Multiple Linear Regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

Where,

Y is the output variable, also called response and dependent variable.

X_i ($i = 1, 2, 3 \dots p$) is the independent variable.

β_0 is the value of Y when independent variable are 0.

β_j ($j = 1, 2, 3 \dots p$) are the regression coefficients.

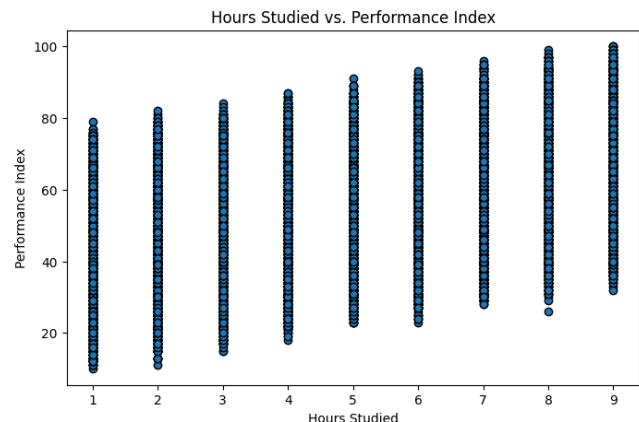
ϵ is a random error in the mode.

3.2 Methodology

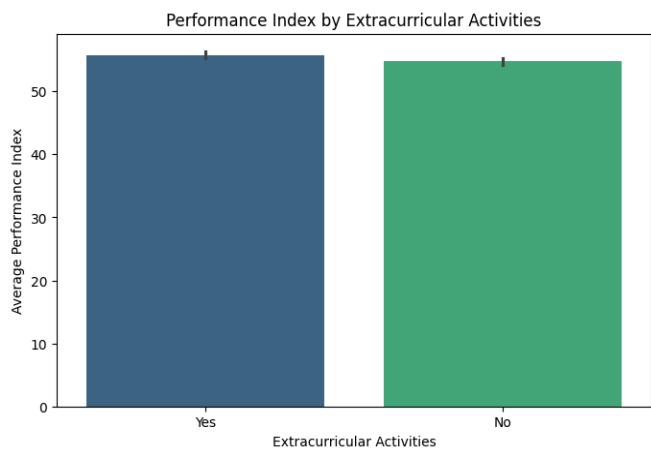
In order to build the model to predict the students' performance, the dataset is collected containing 10,000 student data. Then the dataset is analyzed using various techniques to check if there are any null values and dealt with them accordingly. Methods were also used to analyze the duplicate values and handle them with different techniques.

Many graphs are used to understand the relationship between the attributes. These graphs helped to understand the importance of an attribute and its significance to the performance index.

This graph is a scatter plot that illustrates the relationship between the number of hours studied and the performance index. The x-axis represents the hours studied, ranging from 1 to 9, while the y-axis indicates the performance index, ranging from 0 to 100. Each dot on the graph represents a data point, showing the performance index for a specific number of hours studied. This graph suggests that studying more hours generally provides more value to the performance index but the difference between the study hours is very little.

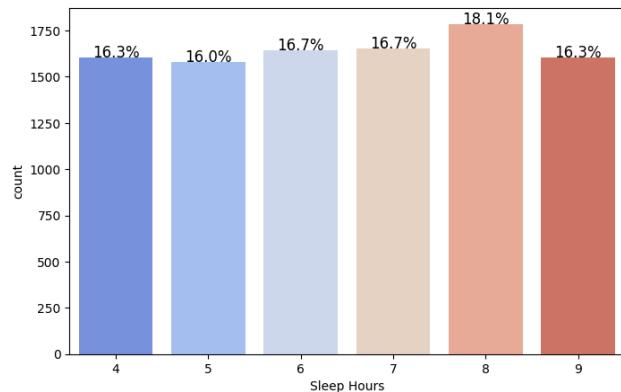


This graph is a bar chart that compares the average performance index of two groups: students who participate in extracurricular activities ("Yes") and those who do not ("No"). The x-axis represents the two groups, and the y-axis indicates the average performance index. The height of each bar corresponds to the average performance index for that group. The bars for "Yes" and "No" have nearly the same height, indicating that there is no significant difference in the average performance index

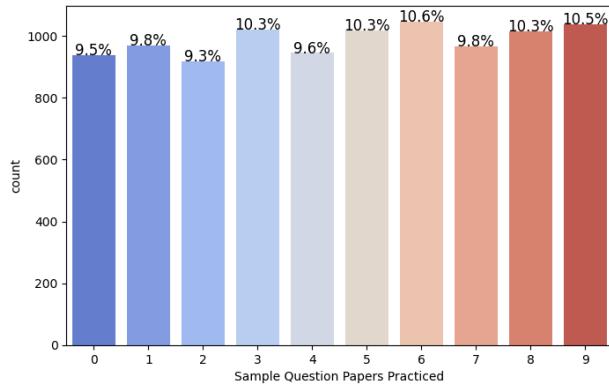


between students who participate in extracurricular activities and those who do not.

This graph is a bar chart that displays the distribution of sleep hours among a group of individuals. The x-axis represents the number of sleep hours, ranging from 4 to 9. The y-axis indicates the count or frequency of individuals who reported sleeping a certain number of hours. The height of each bar corresponds to the number of people who slept that specific number of hours. The distribution is somewhat skewed to the right and most of the values are on the 7 to 9 hours.

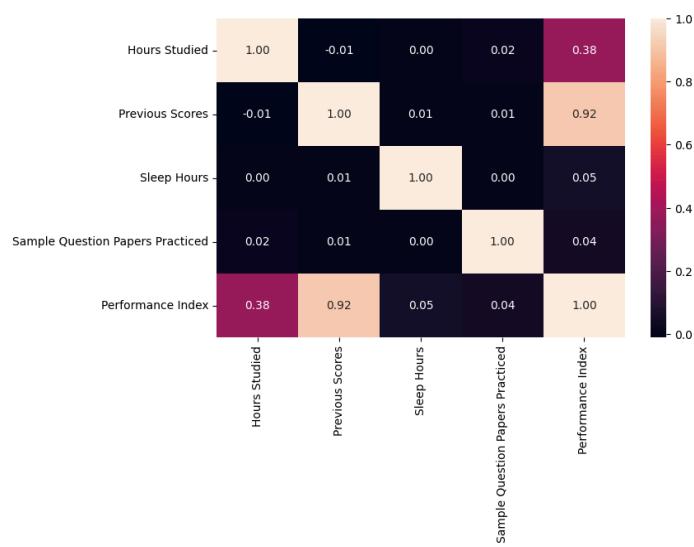


This bar chart displays the distribution of the number of sample question papers practiced by a group of individuals. The x-axis represents the number of sample question papers practiced, ranging from 0 to 9. The y-axis indicates the count or frequency of individuals who practiced that specific number of papers. The height of each bar corresponds to the number of people who practiced that number of papers. This graph suggests that most of the students practiced 6 sample questions to prepare for the exam.



This correlation matrix is a visual representation of the correlations between different variables. Each cell in the matrix shows the correlation coefficient between two variables. The correlation coefficient ranges from -1 to 1, where:

- **1:** Perfect positive correlation (as one variable increases, the other also increases)
- **-1:** Perfect negative correlation (as one variable increases, the other decreases)
- **0:** No correlation (no linear relationship between the variables)



From the matrix, we can observe that the most significant correlation is 0.92 which is between the Previous Score and Performance Index which indicates that if the Previous Score is high then the Performance will have the most significant rise. Hours Studied has moderate correlation with the Performance Index which is 0.38. Lastly, Sleep Hours and Sample Questions Papers Practiced have very weak correlation 0.05 and 0.04 respectively. This indicates that these variables have very less significance to the Performance Index variable.

After all the findings, the dataset is divided into two parts to make the training and testing dataset and trained into the regression model and checked the model score, prediction, error to understand the models efficiency.

3.3 Conclusion

This methodology section involves understanding the dataset and its attributes. The attributes are analyzed and refined to learn the significance of each attribute and its contribution to the Performance Index. Then the attributes are preprocessed and trained and tested to the model.

4. Experimental Result

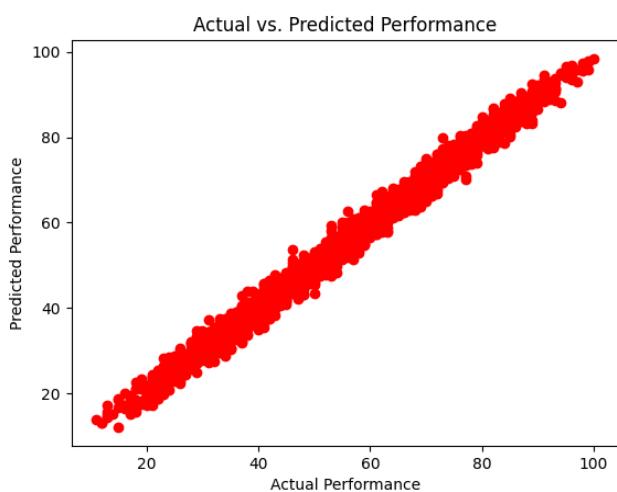
4.1 Introduction

This section covers the results of the experiment for the dataset collected from the source containing 10,000 rows of student information. Here, all the actual and predicted values are compared to understand the model's efficiency. The accuracy of the model is tested with different techniques which will be discussed.

4.2 Result

After testing the model with the testing dataset, various techniques were applied to learn the accuracy of predicted value including, Coefficient of Determination, Mean Absolute Error, R-squared. To visualise the predicted and actual values, scatter plots were also implemented to the experiment.

This graph is a scatter plot that compares the actual performance values to the predicted performance values generated by a model. Each point on the graph represents a data point, where the x-coordinate corresponds to the actual performance and the y-coordinate corresponds to the predicted performance. The points on the graph are in a diagonal line with a positive slope. This indicates a strong positive correlation between the actual and



predicted performance values. As the actual performance increases, the predicted performance also tends to increase. As the points from the Actual Performance and Predicted Performance are very close to each other, it can be said that roughly the model generates accurate predictions.

4.3 Model Evaluation

1. **Mean Absolute Error (MAE):** 1.6469. The value of mean absolute error indicates the average absolute difference between the actual and predicted values. Here, the MAE is 1.6469 which indicates that the model's prediction's average difference is 1.6469 from the actual value.
2. **R-squared (R²):** 0.9884. This value indicates the variance in the dependent variable (the target variable) that is explained by the independent variables (the features) in the model. The R² value 0.9884 means that 98.84% of the dependent variable can be explained by the independent variables.
3. **Model Coefficients:** [2.851, 1.0184, 0.5738, 0.47207, 0.1887]. These coefficients represent the impact of each independent variable on the dependent variable. The independent variables Hours Studied, Previous Scores, Extracurricular Activities ,Sleep Hours, Sample Question Papers Practiced are the independent variables respectively with the coefficients 2.851, 1.0184, 0.5738, 0.47207, 0.1887.
4. **Intercept:** -33.9813. The value of the intercept represents that if all the independent variables' values tend to 0 then -33.9813 will be the value of the dependent variable.

4.4 Discussion

The result of the model for predicting students' performance indicates that the multiple linear regression model is a good fit for the dataset. The high R-squared value indicates that the predicted value of the model covers a significant amount of the actual dataset. The low MAE indicates that the error percentage of the model is reasonable for the model. The magnitude of the coefficient also represents the significance of all the features of the dataset. The scatter plot visualizes that the predicted and actual values are on the same positive slope and very close to each other which means that the model represents the dataset.

5. Conclusion

Determining the factors that affect the students' performance in academic institutions is a very interesting task since it will help the teachers to enhance their learning and teaching process. Our research examines the student's behavior and attributes to predict the outcome or the result of the examination. Our methodology consists of taking different attributes and understanding their significance to the dependent variable and using the attributes to predict.

In future work, we would like to use a different dataset that records different attributes for the student performance, in order to identify the factors that influence their performance. We also would like to compare the performance of multiple regression techniques with others regression and classification models.

6. Reference

1. Quadri1, M.M., Kalyankar, N.V.: Drop out feature of student data for academic performance using decision tree techniques. *Glob. J. Comput. Sci. Technol.* 10(2), 2–5 (2010)
2. Mishra, T., Kumar, D., Gupta, S.: Mining students' data for prediction performance. In: 2014 Fourth International Conference on Advanced Computing & Communication Technologies, pp. 255–262 (2014)
3. Mayilvaganan, M., Kalpanadevi, D.: Comparison of classification techniques for predicting the performance of students' academic environment. In: 2014 International Conference on Communication and Network Technologies, pp. 113–118 (2014)
4. Jishan, S.T., Rashu, R.I., Haque, N., Rahman, R.M.: Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority oversampling technique. *Decis. Anal.* 2(1), 1 (2015)
5. Arsal, P.M., Buniyamin, N., Manan, J.A.: Prediction of engineering students' academic performance using artificial neural network and linear regression: a comparison. In: 2013 IEEE 5th Conference on Engineering Education (ICEED), pp. 43–48 (2013)
6. Yang, S.J.H., Lu, O.H.T., Huang, A.Y.Q., Huang, J.C.H., Ogata, H., Lin, A.J.Q.: Predicting students' academic performance using multiple linear regression and principal component analysis. *J. Inf. Process.* 26, 170–176 (2018)