Parichay: Maharaja Surajmal Institute Journal of Applied Research

Comprehensive Analysis of Marks Prediction Analysis Using Multiple Linear Regression Methodology

Mr. Harjender Singh*

ABSTRACT

One popular method for addressing the challenge of exam score prediction has been to base the prediction on the students' prior academic records. In this work, we introduce a model that builds this forecast on how well students do on multiple assignments given throughout the course of the semester. We used data from a semi-automated peer assessment system, which was used in two undergraduate computer science courses, to create our prediction model. In this system, students answer questions from their peers, ask questions about topics covered in class, and rate the answers given by their peers. After that, we create the attributes needed to create a number of multiple linear regression models.

To assess the performance of the prediction models, we use their Root Mean Squared Error (RMSE). Our final model was constructed utilizing 14 features that capture different student actions. It has reported an RMSE of 2.93 for one course and 3.44 for another on predicting grades on a scale from 18 to 30. Our research may have ramifications for MOOCs and other online course management platforms.

Keywords: Exam Marks Prediction, Multiple Linear Regression, Student Performance, Predictive Modeling, Educational Data Analytics

1. INTRODUCTION

Exam performance prediction is a vital component of education since it allows teachers to recognize pupils who may score poorly and adjust their activities accordingly. Precise forecasts have the potential to enhance educational results by providing valuable insights for resource allocation, curricular modifications, and individualized assistance plans. Multiple linear regression is a particularly useful predictive technique because of its ease of use, readability, and efficiency in addressing linear correlations between dependent and independent variables.

A statistical technique called multiple linear regression is used to model the connection between one or more independent variables and one or more dependent variables. test results are the dependent variable in the context of predicting test marks, whereas study hours, attendance, and past academic success are examples of independent variables (predictors). The regression model examines these factors in an effort to forecast exam scores by utilizing the known correlations.

This study examines the use of multiple linear regression to forecast exam scores using a dataset that contains important variables including study time, attendance, and prior grades. Data collection, preprocessing, model training, evaluation, and prediction are all included in the study's structure. The goals are to create a strong prediction model, assess how accurate it is, and determine how important each predictor is in affecting exam success.

First, the dataset is meticulously cleaned up by removing outliers and missing values and making sure that every variable is scaled and normalized correctly. In order to train the regression model, the most pertinent predictors are found using feature selection. To facilitate model validation and avoid overfitting, the dataset is divided into training and testing subsets.

Metrics like R-squared (R^2) and Root Mean Squared Error (RMSE) are used to evaluate the performance of the trained model and offer information about the model's correctness and explanatory capacity. To find out how each predictor affects exam scores, the regression coefficients are also looked at.

This study illustrates the potential advantages of multiple linear regression for educational stakeholders and shows that it is a viable method for predicting exam scores. Teachers are able to build a more productive learning environment and better support students' academic journeys by offering a data-driven method of predicting student performance. In order to improve predicted accuracy and reliability, future research topics can examine non-linear correlations, integrate more predictors, and compare multiple linear regression with other machine learning methods. By means of these endeavors, the predictive modeling of exam scores can be enhanced, providing more profound understandings and extensive backing for educational initiatives.

2. LITERATURE REVIEW

Exam marks have been predicted using multiple linear

*Assistant Professor, Department of Computer Applications, Maharaja Surajmal Institute

regression (MLR) models on a variety of criteria. Research has demonstrated the efficacy of machine learning (MLR) in predicting final test scores of students in various academic contexts, including corporate statistics courses [1], personalized learning systems [2], and even the concentration of chlorophyll-a in ocean water systems [3]. MLR has demonstrated its adaptability across various fields by being used to estimate stock trends for businesses such as NVDA, AMD, and INTC [4]. Teachers, researchers, and analysts can gain valuable insights into how to improve outcomes, customize learning experiences, and make well-informed decisions based on statistical analyses by utilizing multiple predictor variables such as test scores, homework assignments, and basic categories of academic topics in MLR models.

Various supervised and unsupervised machine learning algorithms can extract hidden relationships in data to assist decision-making. A study introduced a model using machine learning techniques like support vector machine and logistic regression for predicting students' academic performance, with the sequential minimum optimization approach showing higher accuracy.[5]

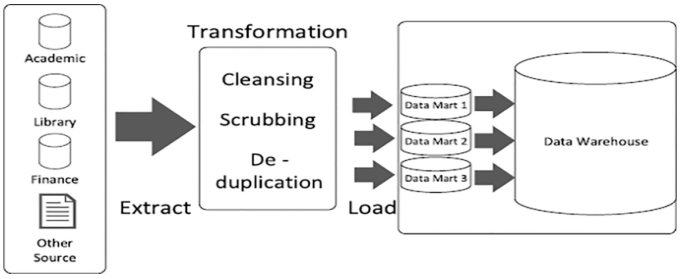
The research's objectives are to analyze the impact of the characteristics of online learning platforms and to create a prediction model to anticipate students' success (grade/engagement). The model developed for this study used machine learning techniques to forecast a learner's ultimate grade and level of engagement. The Random Forest classifier fared better than the others, according to the quantitative approach used by the students for their data analysis and processing. With characteristics connected to student profiles and interactions on a learning platform, grade and engagement prediction accuracy were found to be 85% and 83%, respectively.[6]

In this work, the independent sample t-test, Pearson correlation, means, frequencies, percentages, and standard deviations were all obtained by data analysis using IBM Statistical. According to the study's findings, Polytechnic Sultan Ibrahim's engineering students use CIDOS LMS at a high level as they learn. In light of the fact that the millennial generation continues to dominate the workforce, more research should be done on the necessity for instructors to determine how to best engage students in CIDOS LMS. [7]

Various strategies are offered in this work, including interactive movies, branching scenarios as tools for application-level learning, Speak the word for knowledge-level learning, interactive presentations, and image sequencing for analytical-level learning. The H5P plugin was selected as the tool for course delivery because of its many capabilities integrated within the Canvas LMS. 61 students took the elective "Routing and Switching Concepts," and the outcomes were surveyed both before and after utilizing the recommended framework. To assess the accuracy, the correlation between the students' performance and responses was computed. The section's average correlation was found to be satisfactory, which suggests that the framework of choice worked well.[8]

The current study set out to accomplish two goals: (1) to rank distance education platforms analytically using human-computer interaction criteria, and (2) to use multi-criteria decision-making techniques to determine which distance learning platform would be best for teaching and learning activities. Human-computer interaction-related selection factors, such as interactivity, ease of use, potential for mental strain, presentation style, and user-friendly interface design, were grouped together.[9]

In this work, data mining technology is applied to the college student information management system, data mining is used to mine student evaluation information, data mining is used to design student evaluation information modules, and various relationships between factors influencing student development are explored. The foundation is provided by personalized teaching decision-making and predictive knowledge assessment.[10]



3. RESEARCH METHODOLOGY

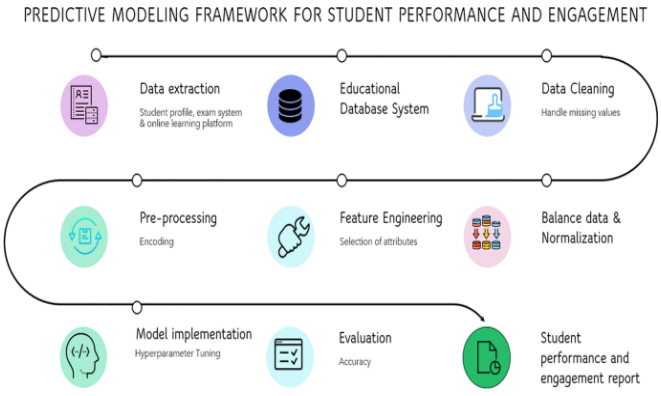
Research involves finding the dataset based on the study hours and exam marks that would be predicted.

It involves analyzing the previous results and marks obtained.

Multiple Linear regression method is used to examine the student final exam marks.

Test and train are also used to validate the dataset for determining the student final exam marks.

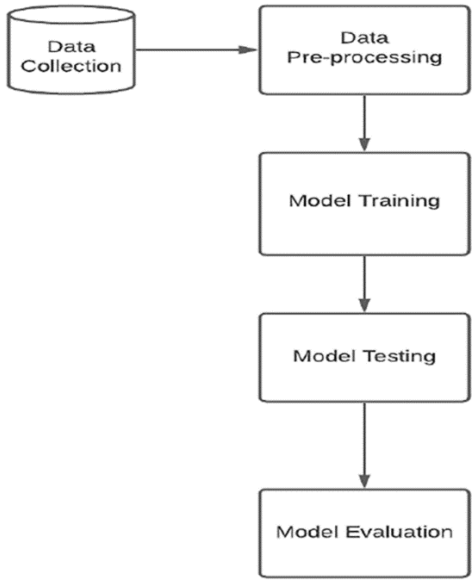
4. PROPOSED FRAMEWORK



A framework that reflects the procedures that underpin the idea of data processing, consolidation, and student performance evaluation was developed. In other educational contexts, the same idea might be used again.

5. PROPOSED SOLUTION

The study of manipulating one or more variables (dependent variables) in order to determine the impact on one or more variables (independent variables) is the essence of experimental research. The conclusion of the many relationships that a product, theory, or idea can produce is based on the cause-and-effect relationship on a selected subject matter (Jongbo, 2014). The precise and methodical manipulation of the variables establishes their nature.



6. ALGORITHM

Algorithm to Calculating the coeffeciants of the simple linear regression equation: y = C0 + C1.x (C1: Is the Slope, C0:Is the Intercept)

Algorithm

Step1: Look up the prior outcomes.

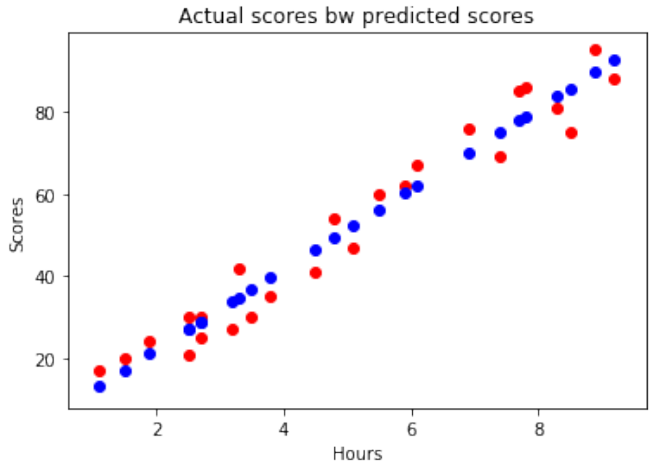
Step 2: Examine the number of hours required to earn a given grade.

Step 3: Calculate the correlation between the number of study hours per day and the final grade.

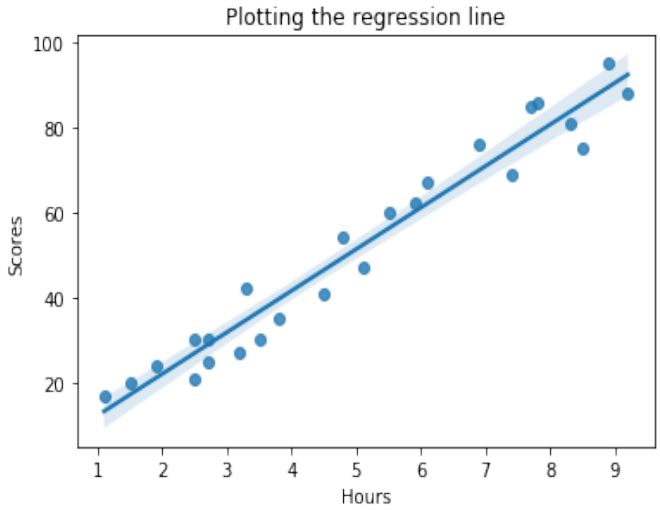
Step 4: Use multiple variable linear regression to calculate exam mark prediction.

```
mean_x = np.mean(df['Hours'])
mean_y = np.mean(df['Scores'])
num = 0
den = 0
```

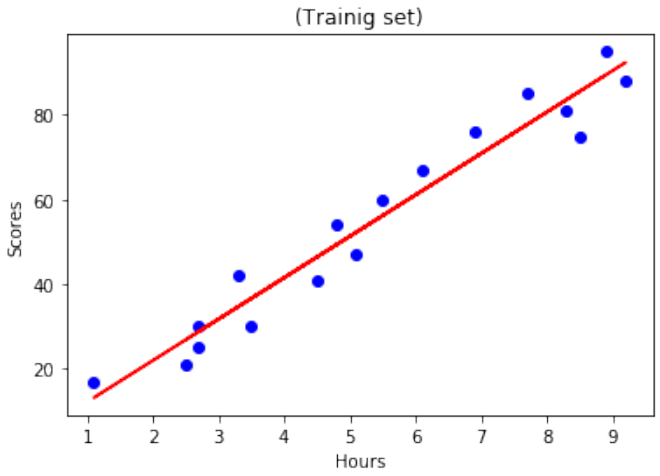
```
x = list(df['Hours'])
y = list(df['Scores'])
for i in range(len(df)):
    num += (x[i]-mean_x)*(y[i]-mean_y)
    den += (x[i]-mean_x)**2
B1 = num/den
B0 = mean_y - B1*mean_x
df['predicted_Scores'] = B0 + B1*df['Hours']
df.head()
plt.scatter(df['Hours'], df['Scores'], c='red', label='Aactual Marks')
plt.scatter(df['Hours'], df['predicted_Scores'], c='blue', label='Predected Marks')
plt.title('Actual scores bw predicted scores')
plt.xlabel('Hours')
plt.ylabel('Scores')
plt.plot()
```



Comparison of Actual score (Red dot) with Predicted Marks (Blue Dot)



Visualizing How Scores and Hours are correlated to each other through linear regression line. Straight Regression lines represent Predicted Score and Dot represent Actual score.



Training Split Ratio = 70:30, which means that 70% of data is used for training and Remaining data (30%) is used for testing. In the above Training set graph dot (blue color) represent actual data While red straight line represents the predicted value.

7. CONCLUSION

The major goal of our work was to take advantage of such information in order to predict student performance. In this paper, we presented a linear regression model for predicting final exam scores of students.

The preliminary results of our prediction model are encouraging.

Linear Regression method helped us to determine the study hours needed to obtain higher exam marks.

REFERENCES

[1] Egodawatte, G. (2021). Forecasting Students' Final Exam: Results Using Multiple Regression Analysis in an Undergraduate Business Statistics Course. *Asian Journal of Economics, Business and Accounting*, *21*(14), 30-40.

[2] Abirami, T., & Vadivel, R. (2023). Student semester marks prediction using linear regression algorithms in machine learning. *World Journal of Advanced Research and Reviews*, *18*(1), 469-475.

[3] Jyothsna, T., & Chitreddy, S. R. (2022). Identification of Implicit Subject Categories Responsible for Academic Test Scores Using Multiple Linear Regression. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021* (pp. 97-102). Springer Singapore.

[4] Lola, M. S., Ramlee, M. N. A., Gunalan, G. S., Zainuddin, N. H., Zakariya, R., Idris, M., & Khalil, I. (2016). Improved the prediction of multiple linear regression model performance using the hybrid approach: a case study of chlorophyll-a at the offshore Kuala Terengganu, Terengganu. *Open Journal of Statistics*, *6*(5), 789-804.

[5] Bhutto, E. S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020, February). Predicting students' academic performance through supervised machine learning. In *2020 International Conference on Information Science and Communication Technology (ICISCT)* (pp. 1-6). IEEE.

[6] Badal, Y. T., & Sungkur, R. K. (2023). Predictive modelling and analytics of students' grades using machine learning algorithms. *Education and Information Technologies*, *28*(3), 3027-3057.

[7] Shida, N., Osman, S., Halim, A., & Sultan, P. I. (2018). Students' perceptions of the use of asynchronous discussion forums, quizzes, and uploaded resources. *Int. J. Eng. Technol*, *7*, 201-204.

[8] Chilukuri, K. C. (2020). A novel framework for active learning in engineering education mapped to course outcomes. *Procedia Computer Science*, *172*, 28-33.

[9] Adem, A., Çakıt, E., & Dağdeviren, M. (2022). Selection of suitable distance education platforms based on human–computer interaction criteria under fuzzy environment. *Neural Computing and Applications*, *34*(10), 7919-7931.

[10] Yin, X. (2021). [Retracted] Construction of Student Information Management System Based on Data Mining and Clustering Algorithm. *Complexity*, *2021*(1), 4447045.

[11] Student Marks Predictor using Machine Learning - Goeduhub Technologies

[12] Prediction Using Supervised ML (Prediction Of Marks) (c-sharpcorner.com)

[13] Predicting students performance in final examination using linear regression and multilayer perceptron IEEE Conference Publication | IEEE Xplore

[14] Social Network for Programmers and Developers (morioh.com)

[15] GitHub - Govind155/Students-Mark-Predictor: End to end implementation and deployment of Machine Learning based Student Mark Prediction.

[16] https://technicalhub.io/blog/student-grade-prediction/