**Journal of Business and Social Sciences (ISSN: 2350-8876)**
Research Department, Hetauda School of Management and Social Sciences (HSMSS), Hetauda

# Student Result Prediction System using Linear Regression

**Er. Sujan Devkota**

Lecturer, Hetauda School of Management & Social Sciences
HOD-IT
Email: sujandevkota@hsm.edu.np

**Cite this paper:**

Devkota, S. (2023). Student result prediction system using linear regression. *Journal of Business and Social Sciences, 5*(1), 49-54

https://doi.org/10.3126/jbss.v5i1.72445

**Abstract**:

The success of an academic institution depends heavily on the performance of its students. A student result prediction system can be beneficial in improving their performance. This study proposes a machine learning algorithm-based Linear regression model to predict the board exam CGPA. The research used a data set of 253 entries, each encoded using one-hot encoding. The Linear regression model was created using 80% of the data set, while the remaining 20% was used to test the model. The results show that the model can accurately predict the final exam's CGPA. This can be useful in identifying students who require additional support and enhancing teaching techniques.

**Keywords**: Regression model, Student Performance, one-hot encoding, predictive model

## Introduction

The academic performance of students is impacted by various factors, including personal, socio-economic, and environmental variables (Kumar & Pal, 2011). Knowledge of these factors and their effect on student achievement can help manage their impact. Educational mining research has recently received a lot of attention. Educational data mining refers to techniques, tools, and research designed to automatically extract meaning from large repositories of data generated by or related to the learning activities of people in an educational environment. Predicting student performance becomes more difficult due to the large volume of data in educational databases (Acharya & Sinha, 2014). The subject of explaining and predicting school performance is widely studied. The ability to predict student performance is very significant in educational institutions. Increasing student success is a long-term goal in all educational institutions. If educational institutions can predict the academic performance of students early before their final exam, extra efforts can be made to organize proper support for low-performing students to improve their studies and help them succeed (Raut & Nichat, 2017).

Data mining helps (Shrivas1 & Tiwari2, 2017) in extracting knowledge from available datasets and should be created as knowledge intelligence for the benefit of the institution. Higher education categorizes students based on their academic achievement. Many factors influence a student's academic performance. The model predicts final exam CGPA based on factors affecting student performance. In this study, a linear regression model is created to predict a student's final CGPA and evaluate their performance.

The objective of this study is to predict the board exam CGPA using various attributes from past records, such as first-term marks, second-term marks, age, gender, and +2 GPA. All these academic and non-academic records are collected from the Management Information System used by the Hetauda School of Management and Social Sciences. This study examines the accuracy of linear regression tasks to predict student results, making it useful in identifying weak students who can be individually supported by educators to improve their performance in the future.

## Literature Review

Research was conducted by Baradwaj and Pal (Kumar & Pal, 2011) on a group of 50 students who were enrolled in a specific course program for a period of 4 years (2007-2010). The study analyzed multiple performance indicators like "Previous Semester Notes," "Class Test Notes," "Seminar Performance," "Homework," "General Proficiency," "Attendance," "Lab Work," and "End of Semester Notes" using the ID3 decision tree algorithm. The objective of the study was to assist both instructors and students in understanding and predicting student performance at the end of the semester. Additionally, the study aimed to identify students who require special attention to reduce the failure rate and take appropriate measures for the next semester's examination. The ID3 decision tree was selected as the data mining technique due to its simplicity.

A study (Krishna et al., n.d.) in the International Journal of Innovative Technology and Exploring Engineering used the CART algorithm to predict student performance in a blended learning course. The algorithm categorized students based on their online activities and accurately predicted which students were at risk of failing the course.

In their paper titled "A CHAID-Based Performance Prediction Model in Educational Data Mining," Ramaswami and Bhaskaran (2010) explore the use of the Chi-squared Automatic Interaction Detection (CHAID) algorithm to predict the academic performance of higher secondary school students in India. The authors collected data from 1,000 students across five schools in three districts of Tamil Nadu, which included various factors like student demographics, academic performance, and socioeconomic status. Utilizing the CHAID algorithm, the authors were able to create a predictive model that identified several factors that significantly affected student performance. Furthermore, the authors suggested that this model could be used to develop early intervention programs for students at risk of underperforming.

In a study conducted by Ahmed and Elaraby (Ahmed & Elaraby, 2014), they focused on creating classification rules and predicting student performance in a specific course curriculum based on previously registered student activities. Abeer and Elaraby analyzed data from students who had previously enrolled in the same course program over a span of 6 years (2005-2010), gathering multiple attributes from the university database. This study successfully predicted the final grades of students to some extent and provided insight to help students improve their performance. It also identified students who required special attention to reduce defective ratios and take appropriate measures at the right time.

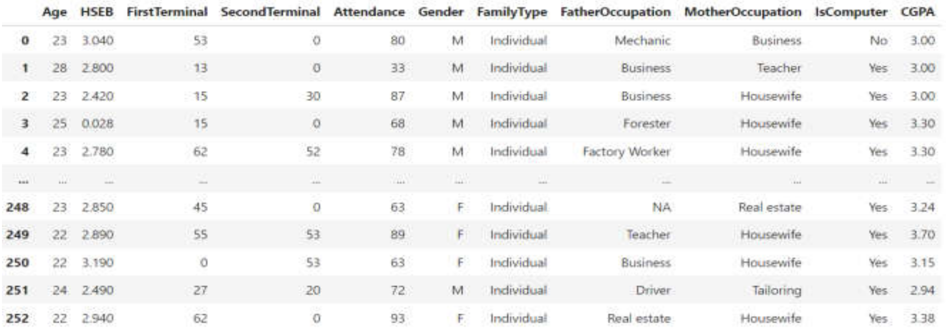
Research conducted by Pandey and Pal (2011) used Naïve Bayes classification for data mining to analyze, classify, and predict whether students were high achievers or underachievers. Naïve Bayes classification is a simple probability classification technique that assumes all attributes in a data set are independent of each other. Pandey and Pal conducted their research on a sample of students enrolled in a Graduate Diploma in Computer Applications (PGDCA) program at Dr. R. M. L. Awadh University, Faizabad, India. Their study was able to rank and predict students' grades to some extent based on their previous year's grades. The findings of this research can be used to assist students in their future education in various ways.

Researchers typically use academic and non-academic indicators to conduct academic data mining. These indicators mainly focus on predicting the factors that can affect academic success, rather than predicting the final CGPA. Most of these studies are based on classification algorithms. Therefore, this study aims to fill this gap and attempt to predict the numerical CGPA based on academic data.

## Methodology

### *Data Collection*

The dataset for this study was collected from the Hetauda School of Management and Social Sciences in Hetauda, Nepal. This college utilizes a management information system to store academic and demographic records. This system extracts information such as first-term marks, second-term marks, attendance, +2 CPA, age, father and mother's occupation, computer science studies in +2, and gender of BIM and BCA students. The Final Exam CGPA was obtained from the website of Tribhuvan University and used as the dataset label. The dataset includes 253 data with varying attributes, and a comprehensive description of the data is provided in the table.

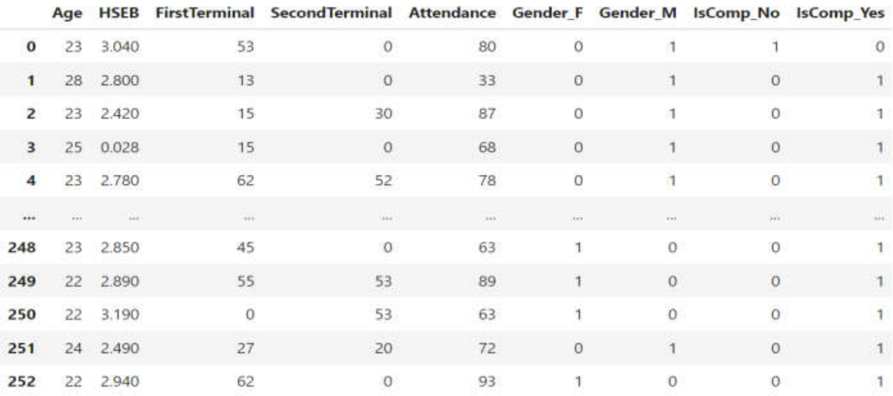


| | Age | HSEB | FirstTerminal | SecondTerminal | Attendance | Gender | FamilyType | FatherOccupation | MotherOccupation | IsComputer | CGPA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 23 | 3.040 | 53 | 0 | 80 | M | Individual | Mechanic | Business | No | 3.00 |
| 1 | 28 | 2.800 | 13 | 0 | 33 | M | Individual | Business | Teacher | Yes | 3.00 |
| 2 | 23 | 2.420 | 15 | 30 | 87 | M | Individual | Business | Housewife | Yes | 3.00 |
| 3 | 25 | 0.028 | 15 | 0 | 68 | M | Individual | Forester | Housewife | Yes | 3.30 |
| 4 | 23 | 2.780 | 62 | 52 | 78 | M | Individual | Factory Worker | Housewife | Yes | 3.30 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 248 | 23 | 2.850 | 45 | 0 | 63 | F | Individual | NA | Real estate | Yes | 3.24 |
| 249 | 22 | 2.890 | 55 | 53 | 89 | F | Individual | Teacher | Housewife | Yes | 3.70 |
| 250 | 22 | 3.190 | 0 | 53 | 63 | F | Individual | Business | Housewife | Yes | 3.15 |
| 251 | 24 | 2.490 | 27 | 20 | 72 | M | Individual | Driver | Tailoring | Yes | 2.94 |
| 252 | 22 | 2.940 | 62 | 0 | 93 | F | Individual | Real estate | Housewife | Yes | 3.38 |

**Figure 1: Original Data Set**

### *Data preprocessing:*

This dataset provides data on first-term and second-term grades and attendance, which differ based on subject weight and number of classes. The records are converted to a 100-point scale for analysis. To make the data usable for machine learning algorithms, one-hot encoding is applied to all attributes, which converts categorical variables to numerical ones to create a model. Unimportant features have been removed from the original dataset.



| | Age | HSEB | FirstTerminal | SecondTerminal | Attendance | Gender_F | Gender_M | IsComp_No | IsComp_Yes |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 23 | 3.040 | 53 | 0 | 80 | 0 | 1 | 1 | 0 |
| **1** | 28 | 2.800 | 13 | 0 | 33 | 0 | 1 | 0 | 1 |
| **2** | 23 | 2.420 | 15 | 30 | 87 | 0 | 1 | 0 | 1 |
| **3** | 25 | 0.028 | 15 | 0 | 68 | 0 | 1 | 0 | 1 |
| **4** | 23 | 2.780 | 62 | 52 | 78 | 0 | 1 | 0 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **248** | 23 | 2.850 | 45 | 0 | 63 | 1 | 0 | 0 | 1 |
| **249** | 22 | 2.890 | 55 | 53 | 89 | 1 | 0 | 0 | 1 |
| **250** | 22 | 3.190 | 0 | 53 | 63 | 1 | 0 | 0 | 1 |
| **251** | 24 | 2.490 | 27 | 20 | 72 | 0 | 1 | 0 | 1 |
| **252** | 22 | 2.940 | 62 | 0 | 93 | 1 | 0 | 0 | 1 |

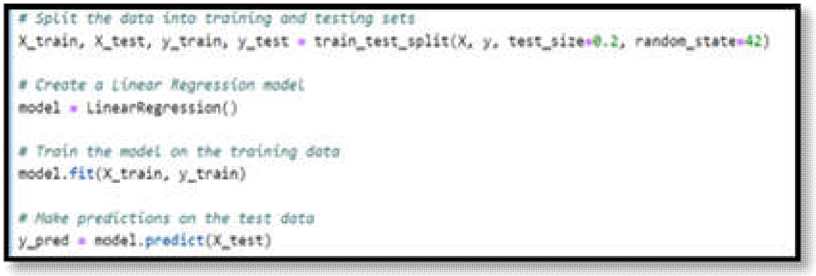
**Figure 2: preprocessed dataset**

### *Tools and Algorithm:*

The main objective of this study is to predict the results of the board exam based on past records. The total data was split into two parts. 80% of the data was used for model creation, and the remaining 20% was used for training. A linear regression model was used to predict the final CGPA, and this model was ultimately used to predict the final grades of new students. Python programming language was utilized to construct the model.

### *Model development:*

A linear regression model was used to create a model for result prediction. Linear regression (Linear Regression in Machine Learning - Geeks for Geeks, n.d.) is an example of a supervised machine-learning technique that maps data points to the best possible linear functions by gaining knowledge from labeled datasets. It may be used to make predictions using new data.



```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create a Linear Regression model
model = LinearRegression()

# Train the model on the training data
model.fit(X_train, y_train)

# Make predictions on the test data
y_pred = model.predict(X_test)
```

**Figure 3: Model Creation**

### *Model evaluation:*

In this study, the precision of regression models is measured using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R Squared matrix.

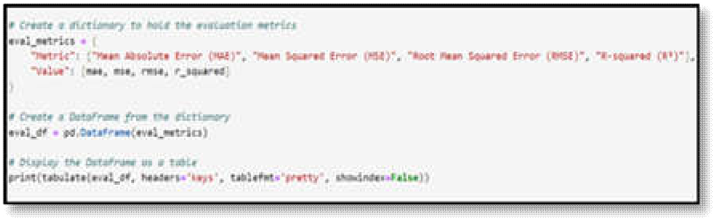
MAE (*Evaluation Metrics for Your Regression Model - Analytics Vidhya*, n.d.) calculates the absolute difference between actual and predicted values, The MAE metric is easy to understand and a lower value suggests higher model performance.

$$MAE(f) = \frac{1}{N}\sum_{i=1}^{N} |f(x_i) \ - \ y_i|$$

The concept of mean squared error (*Evaluation Metrics for Your Regression Model - Analytics Vidhya*, n.d.) involves calculating the squared difference between the actual and predicted values.

$$MSE(f) = \frac{1}{N}\sum(f(x_i) \; - \; y_i)^2$$

R-squared (RSME - Root Mean Square Error in Python - Java point, n.d.) is a statistical technique used to determine the quality of a fit. A high $R^2$ value indicates that the regression model fits the data well. It indicates how well the model can predict the variable's variation.

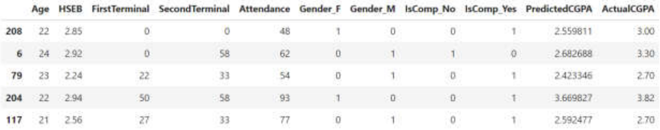


```
# Create a dictionary to hold the evaluation metrics
eval_metrics = {
    "Metric": ["Mean Absolute Error (MAE)", "Mean Squared Error (MSE)", "Root Mean Squared Error (RMSE)", "R-squared (R²)"],
    "Value": [mae, mse, rmse, r_squared]
}

# Create a DataFrame from the dictionary
eval_df = pd.DataFrame(eval_metrics)

# Display the DataFrame as a table
print(tabulate(eval_df, headers='keys', tablefmt='pretty', showindex=False))
```

**Figure 4: Model Evaluation**

***Model Deployment***

The model predicts board exam CGPA using past academic and non-academic records.



| | Age | HSEB | FirstTerminal | SecondTerminal | Attendance | Gender_F | Gender_M | IsComp_No | IsComp_Yes | PredictedCGPA | ActualCGPA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **208** | 22 | 2.85 | 0 | 0 | 48 | 1 | 0 | 0 | 1 | 2.559811 | 3.00 |
| **6** | 24 | 2.92 | 0 | 58 | 62 | 0 | 1 | 1 | 0 | 2.682688 | 3.30 |
| **79** | 23 | 2.24 | 22 | 33 | 54 | 0 | 1 | 0 | 1 | 2.423346 | 2.70 |
| **204** | 22 | 2.94 | 50 | 58 | 93 | 1 | 0 | 0 | 1 | 3.669827 | 3.82 |
| **117** | 21 | 2.56 | 27 | 33 | 77 | 0 | 1 | 0 | 1 | 2.592477 | 2.70 |

**Figure 5: Model testing**

## Result

For this study, a regression model was developed to predict the Final CGPA. The model was created using 80% of the available data and 20% was used for model evaluation. Python programming was used to build the model. Mean absolute error, mean squared error, root mean squared error, and R squared were assessed as the evaluation metrics.

**Table 1: Performance matrix**

| Matric | Value |
|---|---|
| Mean Absolute Error (MAE) | 0.7122 |
| Mean Squared Error (MSE) | 1.2176 |
| R-squared(R2) | 0.0992 |

The table above shows the MEA, MSE, and R-squared values as 0.7722, 1.217 and 0.009, respectively. In this study, the model's predictions have an average deviation of approximately 0.7123 units from the actual values. The average deviation of the model's predictions was 1.2176 squared units, as shown by the mean squared error (MSE) value of 1.2176. The R-squared ($R^2$) statistic measures the proportion of the dependent variable's variance described by the predictive model (Evaluation Metrics for Your Regression Model - Analytics Vidhya, n.d.). This study shows that $R^2$ value of around 0.0991, indicating that the model only explains a small portion of the total variation in the target variable.

## Conclusion and Discussion

Through evaluation metrics analysis, we can gain a better understanding of our predictive model's performance. Although the model displays some predictive ability, as shown by the MAE and MSE values, the R² value indicates a significant amount of unexplained variance in the target variable. This implies that this model could benefit from more improvement or the addition of more features to increase its predictive accuracy.

## References

Acharya, A., & Sinha, D. (2014). Early Prediction of Students Performance using Machine Learning Techniques. *International Journal of Computer Applications*, *107(1), 37-4.*

Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology, 2*(2), 43–47. https://doi.org/10.13189/wjcat.2014.020203

Evaluation Metrics for Your Regression Model - Analytics Vidhya. (n.d.). Retrieved October 5, 2023, from https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

Krishna, M., Rani, B. S. B. P., Chakravarthi, K., Madhavrao, B., & Chowdary, S. M. B. (n.d.). Predicting Student Performance using Classification and Regression Trees Algorithm. *International Journal of Innovative Technology and Exploring Engineering, 9(3), 96-100* https://doi.org/10.35940/ijitee.C8964.019320

Kumar, B., & Pal, S. (2011). Mining Educational Data to Analyze Student's Performance. International Journal of *Advanced Computer Science and Applications, 2(*6, 63-69), https://doi.org/10.14569/ijacsa.2011.020609

Linear Regression in Machine learning - Geeks for Geeks. (n.d.). Retrieved October 5, 2023, from https://www.geeksforgeeks.org/ml-linear-regression/

Pandey, U. K., & Pal, S. (2011). A Data Mining view on Class Room Teaching Language. *IJCSI International Journal of Computer Science Issues, 8*(2), 273-276, www.IJCSI.org

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID Based Performance Prediction Model in Educational Data Mining. *IJCSI International Journal of Computer Science Issues, 7*(1), 10-18. www.IJCSI.org

Raut, A. B., & Nichat, A. A. (2017). Students Performance Prediction Using Decision Tree Technique. In International *Journal of Computational Intelligence Research, 13*(7), 1760-1768 http://www.ripublication.com

RSME - Root Mean Square Error in Python - Java point. (n.d.). Retrieved October 5, 2023, from https://www.javatpoint.com/rsme-root-mean-square-error-in-python

Shrivas1, A. K., & Tiwari2, P. (2017). Comparative Analysis of Models for Student Performance with Data Mining Tools. *International Journal of Computer Trends and Technology, 46*(1, 44-49), http://www.ijcttjournal.org