STUDENTS PERFORMANCE PREDICTION

A project report submitted to ICT Academy of Kerala in partial fulfillment of the requirements for the certification of

CERTIFIED SPECIALIST

IN

DATA SCIENCE & ANALYTICS

submitted by

Lakshmipriya K A

ICT ACADEMY OF KERALA

THIRUVANANTHAPURAM KERALA, INDIA

NOVEMBER 2024

List of Figures

TABLE OF CONTENTS

ABSTRACT

Predicting Student Eligibility Based on Assessment Scores and Attendance

This project aims to predict student eligibility for academic advancement using key performance indicators derived from both formative and summative assessments. The prediction model integrates various performance metrics, including Case Study Assessment Scores, Total Quiz Scores, Grand Total Coding Scores, Project Scores, and Attendance Records. The dataset, which contains these features, is analyzed to identify patterns and correlations between students' academic engagement and their eligibility outcomes. Multiple machine learning algorithms, such as decision trees, logistic regression, and random forests, are employed to develop a predictive model. The performance of the models is assessed through metrics like accuracy, precision, recall, and F1-score to ensure the reliability and robustness of the predictions. The results highlight the importance of consistent participation in both coursework and assessments, with attendance and project scores emerging as significant predictors of student eligibility. This project demonstrates the potential of data-driven approaches in predicting student outcomes, providing educators and institutions with valuable insights for identifying students at risk of not meeting academic criteria and offering early interventions to improve student success.

The findings from this project have significant implications for educational institutions seeking to implement more personalized and data-driven approaches to student assessment. By leveraging predictive models to identify students at risk of falling below eligibility thresholds, educators can intervene proactively, offering targeted support and resources to those who need it the most. This approach can help optimize student success, reduce dropout rates, and ensure that students are provided with the appropriate academic guidance at critical points in their education. Furthermore, the model can be adapted for different educational contexts, such as high schools, colleges, and vocational training centers, making it a versatile tool for improving educational outcomes. The project also highlights the importance of integrating diverse assessment metrics, such as quizzes, projects, and attendance, into a comprehensive framework for evaluating student progress and fostering a more holistic understanding of their academic performance.

1. INTRODUCTION

1.1 Problem Overview

This project involves analyzing the engagement and performance data of students who completed a Data Science course. The dataset includes student attendance, daily quiz marks, and a detailed score sheet containing scores from assignments, case studies, and the final assessment. The formative score is the sum of assignments, case studies, and projects, while the summative score is the combined score from the final MCQ and coding test. Participants will conduct data analysis to extract meaningful insights and develop predictive models to forecast student outcomes and success.

1.2 Problem Statement

- Analyze the relationship between student engagement (attendance, daily quizzes) and overall course performance.
- Identify patterns in formative and summative assessments that contribute to final student outcomes.
- Build predictive models to forecast student performance based on engagement and formative/summative assessments.
- Present findings and insights using appropriate visualization tools and predictive analytics techniques.

1.3 DATA DESCRIPTION

- **Dataset**: The dataset includes:
    - **Engagement Data**: Attendance records and daily quiz scores.
    - **Performance Data**: Scores from assignments, case studies, final assessments, and overall grades.
    - **Formative Score**: Sum of scores from assignments, case studies, and projects.
    - **Summative Score**: Combined score from final MCQ and coding test.

    - **TARGET** : Eligibility

1. **Machine Learning Model Training:** Train a classification-based machine learning model (such as LogisticRegression, svm,RandomForestRegressor, etc.) on the collected data to predict Eligibility.

2. **Web Application Development**:
   - Build a Python Flask web application that:
   - Accepts inputs (e.g : attendance percentage,grandtotal ,quiztotal) from users.
   - Processes these inputs and feeds them into the trained machine learning model. ● Displays the prediction to the user in a user-friendly format.

3. **Integration and Deployment:**
   Integrate the trained model with the Flask application, ensuring efficient prediction and response times. Deploy the application on a suitable platform to make it accessible via the web.

4. **User Interface (UI) Design:**
   Design an intuitive UI that allows users to easily input required parameters and view prediction. The UI should be responsive and provide feedback on data input errors or missing fields.

5. **Testing and Validation**:
   Validate the accuracy of the predictions against real-time data and known standards (e.g., EPA standards). Perform through testing to ensure the application functions correctly under different scenarios and edge cases.

By completing this project, users will have access to a reliable tool for predicting eligibility of students

DATA SCIENCE LIFECYCLE

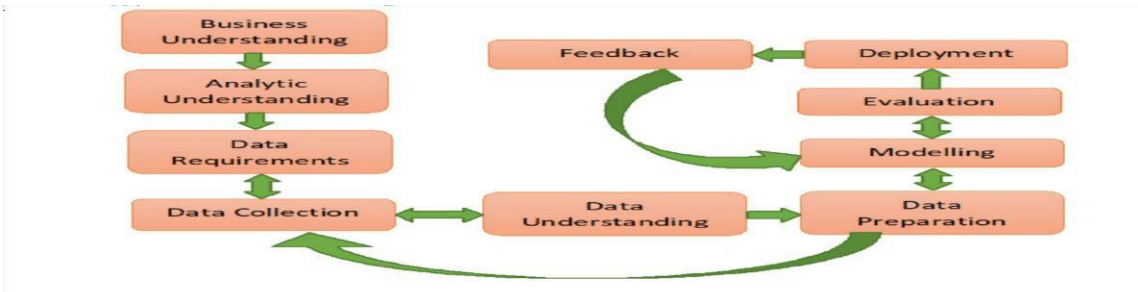


Fig. 1 data science lifecycle

2 Data Collection

Data Composition : The dataset includes the following information:

- **Dataset**: The dataset includes:
  - **Engagement Data**: Attendance records and daily quiz scores.
  - **Performance Data**: Scores from assignments, case studies, final assessments, and overall grades.
  - **Formative Score**: Sum of scores from assignments, case studies, and projects.
  - **Summative Score**: Combined score from final MCQ and coding test.

3 DATA PREPROCESSING

- The data underwent necessary preprocessing steps such as handling missing values, encoding categorical variables, and feature scaling. These steps are critical in ensuring the model's accuracy and reliability.

- It was observed that the dataset had outliers and some non-linear relationships that might have influenced the model's performance. The presence of such data points calls for further investigation or more sophisticated methods like outlier detection and non-linear transformation.

- Data preprocessing is essential to ensure the quality and reliability of the dataset used for training the machine learning model. The following steps are undertaken:

1. Handling Missing Values:
   - Missing values in the dataset are imputed using the mode and median strategy . This approach replaces missing values with the median/mode of the respective feature to maintain data integrity.

2. Outlier Handling:
   - Outliers in the dataset are managed using the Interquartile Range (IQR) method. This approach identifies and removes extreme values that lie beyond 1.5 times the IQR from the first (Q1) and third quartiles (Q3). By filtering out these outliers, the data becomes more representative of the typical distribution, which can enhance the robustness and performance of predictive models.

3. Encoding Categorical Features:
    - Categorical variables, such as location where the data was collected, are encoded using LabelEncoder. This step transforms categorical data into numerical values, allowing machine learning algorithms to operate on them effectively.
4. Feature Selection:
    - Based on Target Correlation: Choose features with high absolute correlation values with the target.
    - Select Features: Choose features that have a high correlation with the target variable, as these are more likely to be relevant.

4 Exploratory Data Analysis

Exploratory data analysis (EDA) is a most common approach, which helps in summarising the main characteristics of a dataset by analysing the characteristics, commonly with visual methods. Using a statistical model is optional. However, mainly EDA helps to seek for pieces of information from the data which are beyond the hypothesis testing task or formal modelling. fig 2 :Histogram to understand relationships

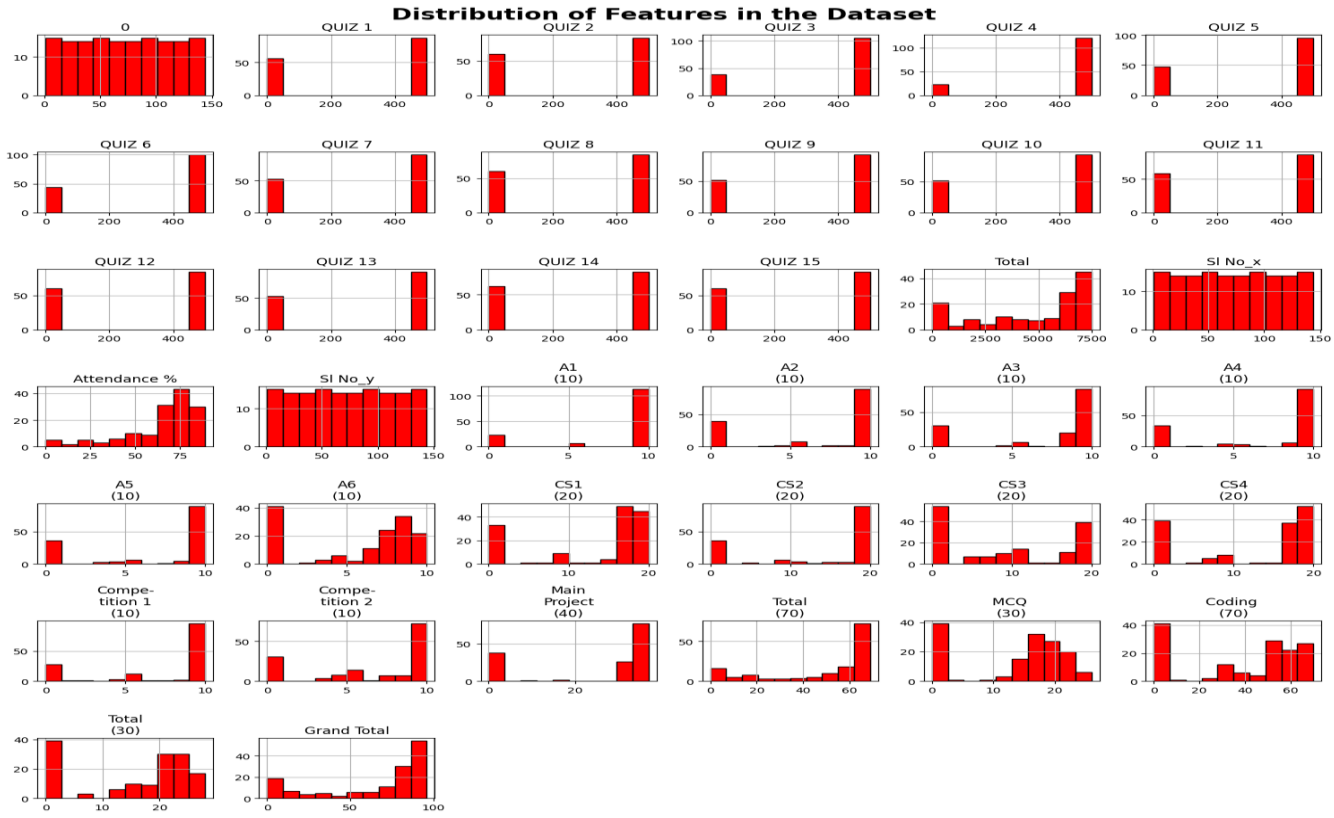


fig 2 :Histogram to understand relationships

The figure displays multiple histograms showing the distribution of various features in a dataset, including scores from quizzes (Quiz 1 to Quiz 15), attendance percentage, project scores, and other assessments (e.g., A1-A6, CS1-CS4, competitions, MCQ, coding). The x-axes represent different score ranges, while the y-axes indicate the frequency of occurrences.

Key observations:

- Many quiz distributions are bimodal, with peaks at lower and higher score ranges.
- Attendance percentages are right-skewed, peaking around 60%-80%.
- The total score shows a broad distribution, with some students scoring very high.
- Assignment and project scores are mostly concentrated at lower values, suggesting limited high scores across these components.

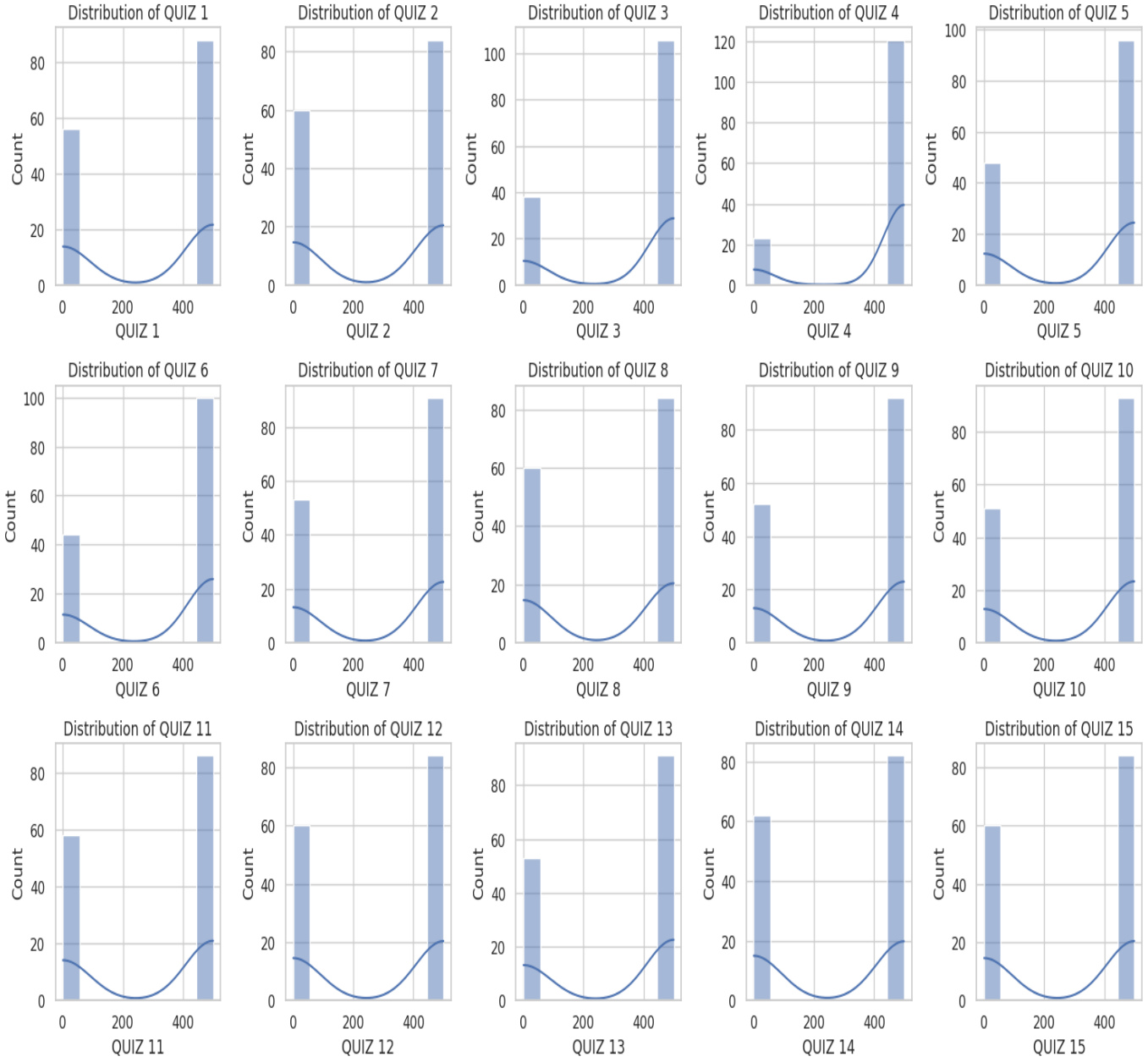


Fig 3:Histogram to visualize relationship

The figure shows a series of histograms for 15 quizzes, each representing the score distribution of the respective quiz. The histograms reveal a bimodal distribution with two distinct peaks: one at the lower end (close to 0) and one at the higher end (around 400-500). This suggests a polarized performance pattern, with many students scoring either very low or very high, and fewer scores in the middle range. A smooth density curve overlays each histogram, highlighting the distribution trend.

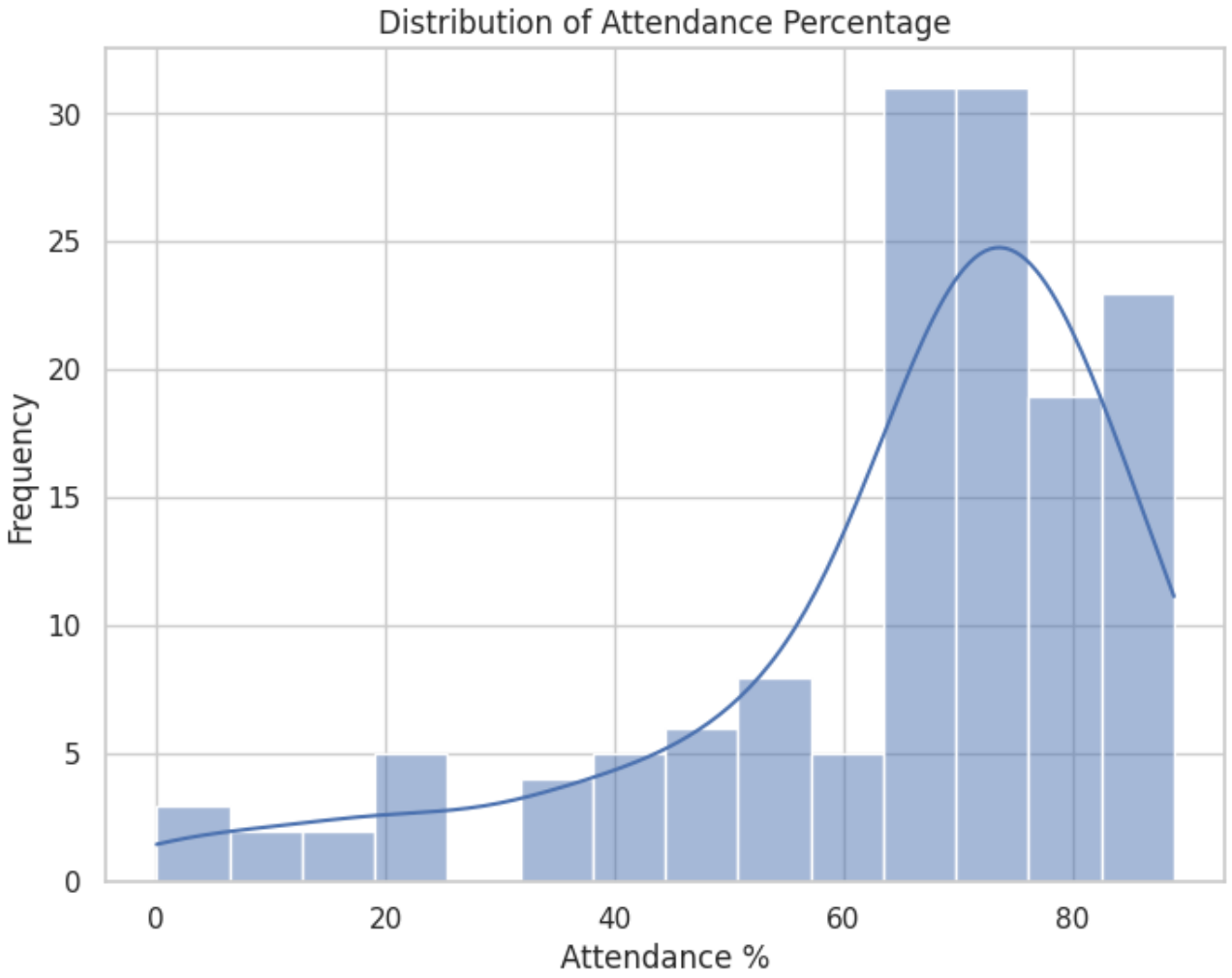


Fig 4: Distribution of attendance percentage

The histogram illustrates the distribution of Attendance Percentage among students. The data is right-skewed, with most students having attendance between 60% and 80%. The attendance percentage peaks around 60%-70%, indicating that this is the most common attendance range. Fewer students have very low (0%-20%) or perfect (near 100%) attendance. A smooth density curve highlights the trend, showing a peak followed by a decline as attendance reaches higher percentages.

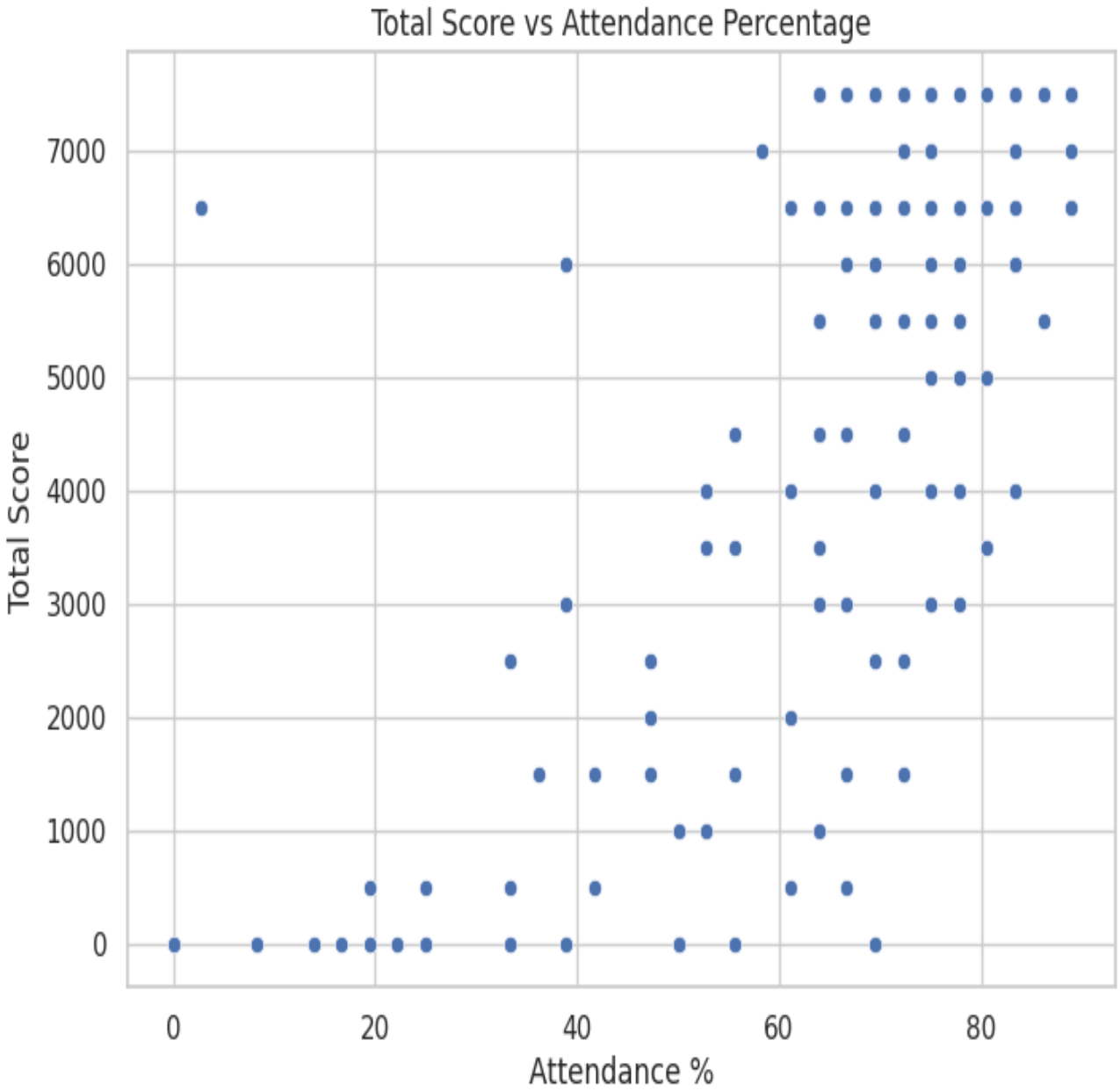


Fig 5:Scatter plot to visualize relationship

The scatter plot shows the relationship between Total Score and Attendance Percentage for a group of students.The data indicates a positive correlation: as attendance percentage increases, the total score also tends to increase. However, there is considerable variability, especially at higher attendance levels, suggesting that while better attendance generally leads to higher scores, some students perform well or poorly irrespective of their attendance.

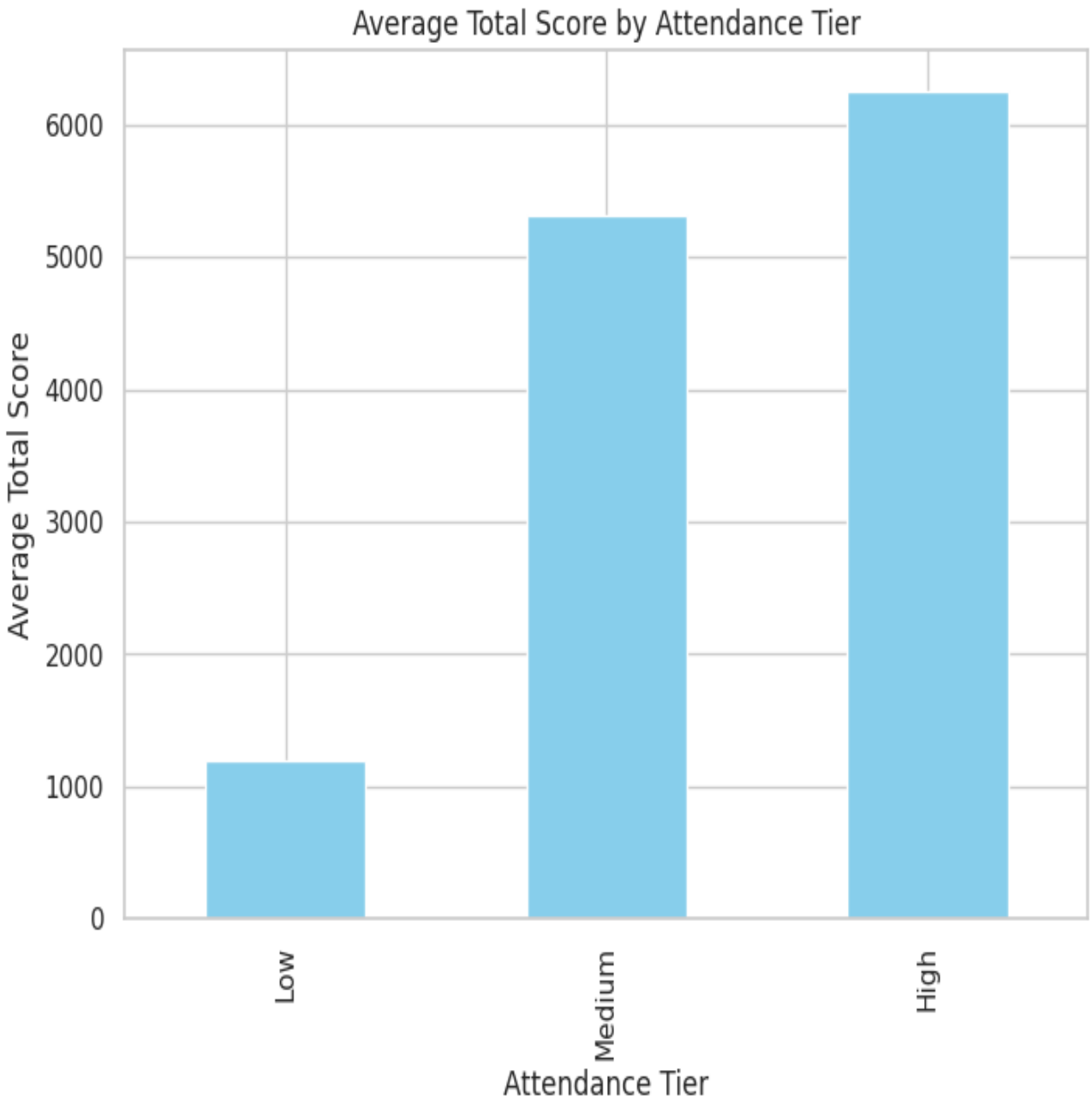


Fig 6:Attendence Tier

The bar chart displays the average total score categorized by attendance tier. There are three tiers: Low, Medium, and High. The average total score increases with each tier, with the Low attendance tier scoring the lowest, Medium attendance scoring higher, and High attendance tier achieving the highest average score, surpassing 6000. This suggests a positive correlation between attendance level and total score.

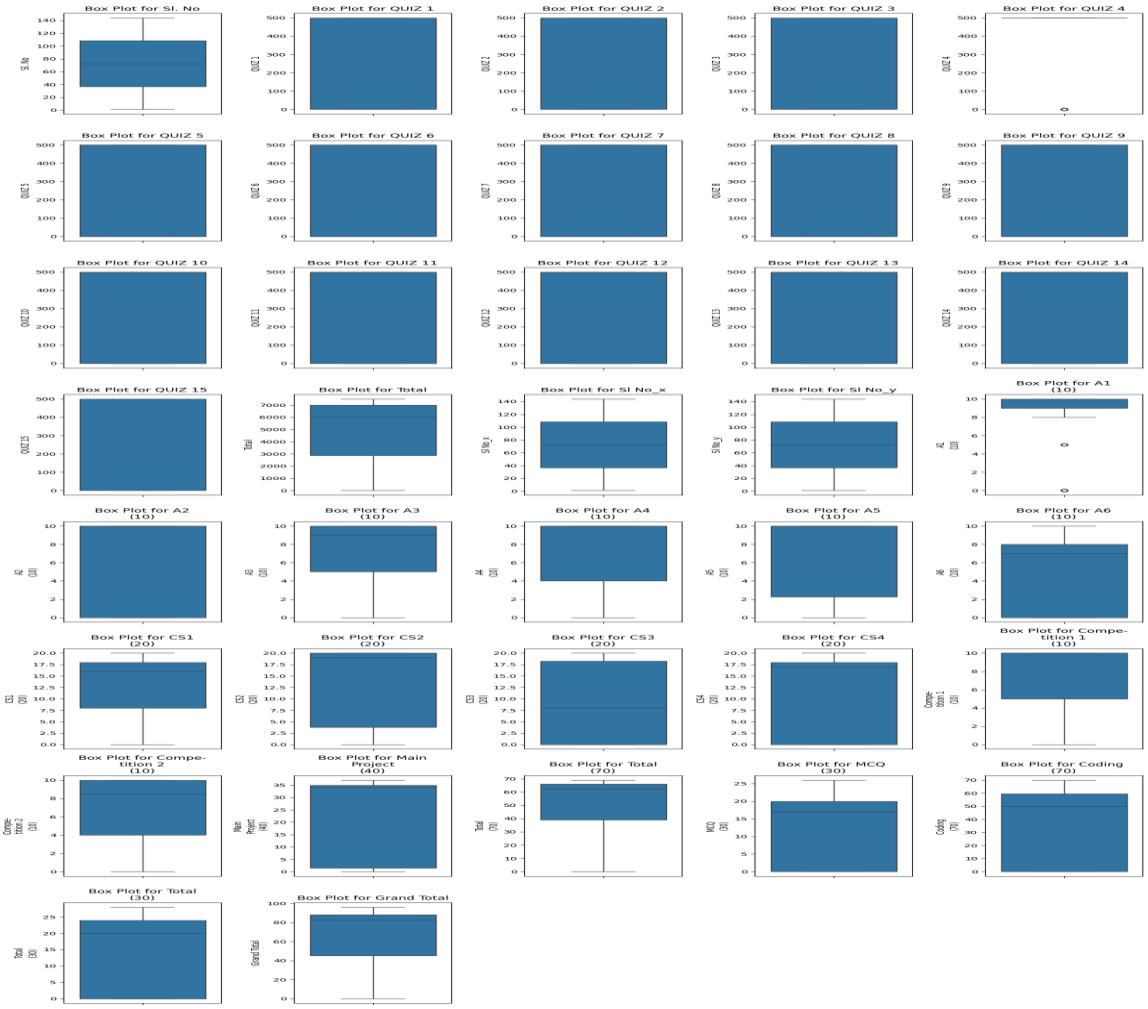


Fig 7 : boxplot

The collection of box plots provides an overview of the distribution of scores across different quizzes, assessments, and components in the dataset. Most of the distributions are centered around the median, with a few outliers observed in some assessments The box plots indicate the presence of outliers in the following:

1. **A1** - Contains a few outliers.
2. **Total (10)** - Shows some outliers.
3. **Coding** - Outliers are visible in this component.
4. **SI No. y** - Contains outliers as well.

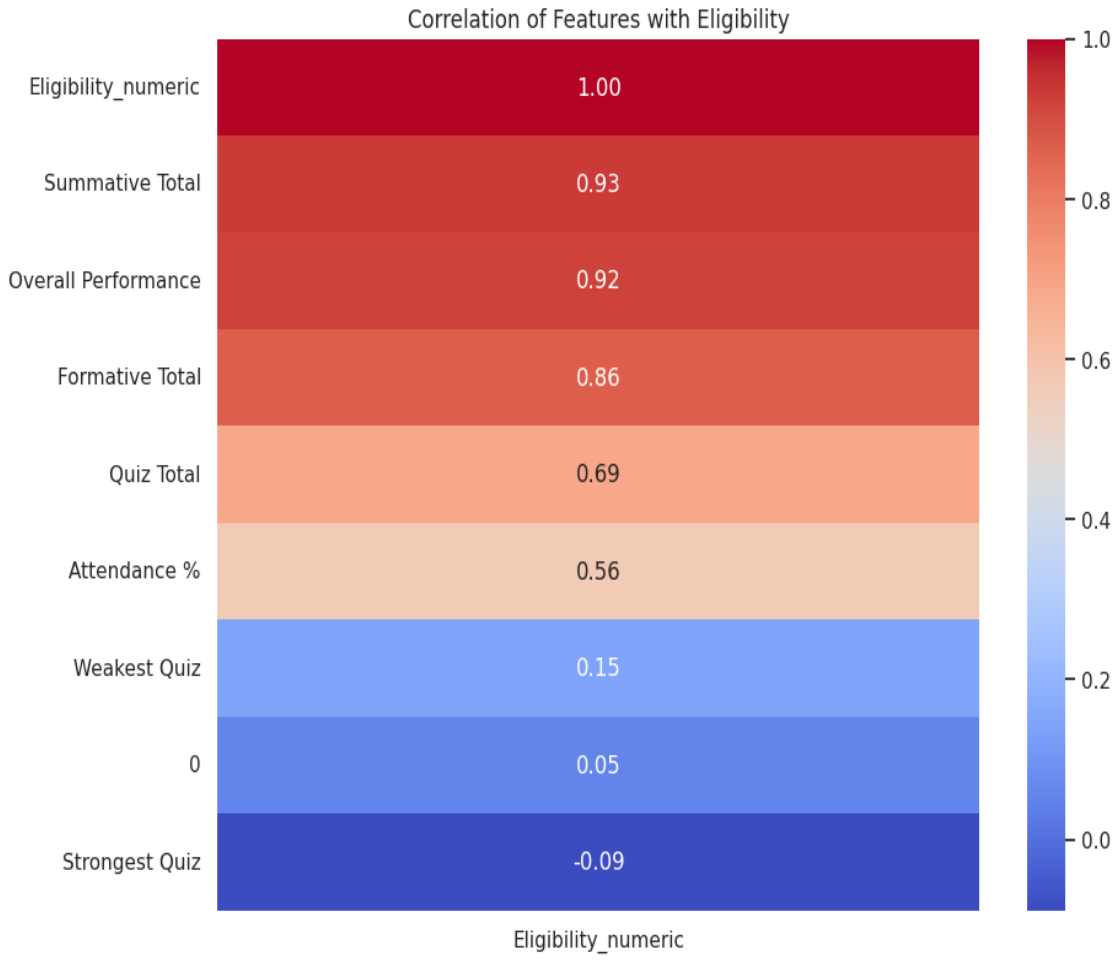


Fig 8:HeatMap

Based on the figure, several key factors influence prediction. The strongest positive correlation was found between eligibility and Overall performance(Grand Total) .Other factors like quiz total attendance showed positive relationships. Further analysis is needed to quantify these impacts and consider potential non-linearrelationships or interaction effects.

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate)

1. Positive correlation: A positive correlation is a relationship between two variables that move in tandem—that is, in the same direction. A positive correlation exists when one variable decreases as the other variable decreases, or one variable increases while the other increases.

2. Negative correlation: Negative correlation is a relationship between two variables in which one variable increases as the other decreases, and vice versa.In statistics, a perfect negative correlation is represented by the value -1.0, while a 0 indicates no correlation, and +1.0 indicates a perfect positive correlation. A perfect negative correlation means the relationship that exists between two variables is exactly opposite all of the time. These are two types of correlation are represented numerically and as well as by shade of color in the heat map.

5 Literature Survey

Machine Learning Techniques for House Price Prediction

[1] J. Xu, K. H. Moon, and M. Van Der Schaar, "A Machine Learning Approach for Tracking Process., vol. 11, no. 5, pp. 742–753, 2017.
[2] K. P. Shaleena and S. Paul, "Data mining techniques for predicting student performance," in ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology, 2015, no. March, pp. 0–2.

Data Preprocessing and Feature Engineering

- **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R.** *Springer.*

  This book covers essential data preprocessing techniques, including handling missing values and encoding categorical variables, which are crucial for preparing datasets in house price prediction.

- **Li, Y., & Yao, Y. (2016). Feature selection in property value prediction.** *Journal of Real Estate Research, 38(2), 213-238.*

This paper provides insights into feature selection methods specific to real estate data, such as identifying the most influential features for predicting property prices.

Comparison of Different Machine Learning Models

Key Papers:

- **Bishop, C. M. (2006).** *Pattern Recognition and Machine Learning.* Springer.
  - A detailed discussion of various machine learning models, including regression, boosting, and ensemble methods, helping to understand the theoretical background behind the models used in your project.

6 DATA SPLITTING

Split into Train and Test Set](#Split-into-Train-and-Test-Sets) we will split the dataset into training, and testing sets. This step is essential to evaluate the model's performance, tune hyperparameters, and ensure its generalizability to unseen data.80% of the dataset is allocated for training, allowing the model to learn patterns and relationships from the data. The remaining 20 % is reserved for testing, used to assess how well the model generalizes to unseen data.

Split Data into Features (X) and Target (y)]

In this section, we will divide our dataset into two main components: Features (X) and the target variable (y). The features (X) consist of all the independent variables that will be used as input to the model, while the target variable (y) represents the outcome we aim to predict—in this case, customer churn. This separation is crucial for training and evaluating the model effectively.

7 MODEL EVALUATION

To compare the performance of models, various evaluation criteria are employed. The key performance evaluation indices used in this analysis include Cross-Validation Score and Best Parameters for each model. The cross-validation score reflects how well the model generalizes to an independent dataset, providing a more reliable estimate of model performance..

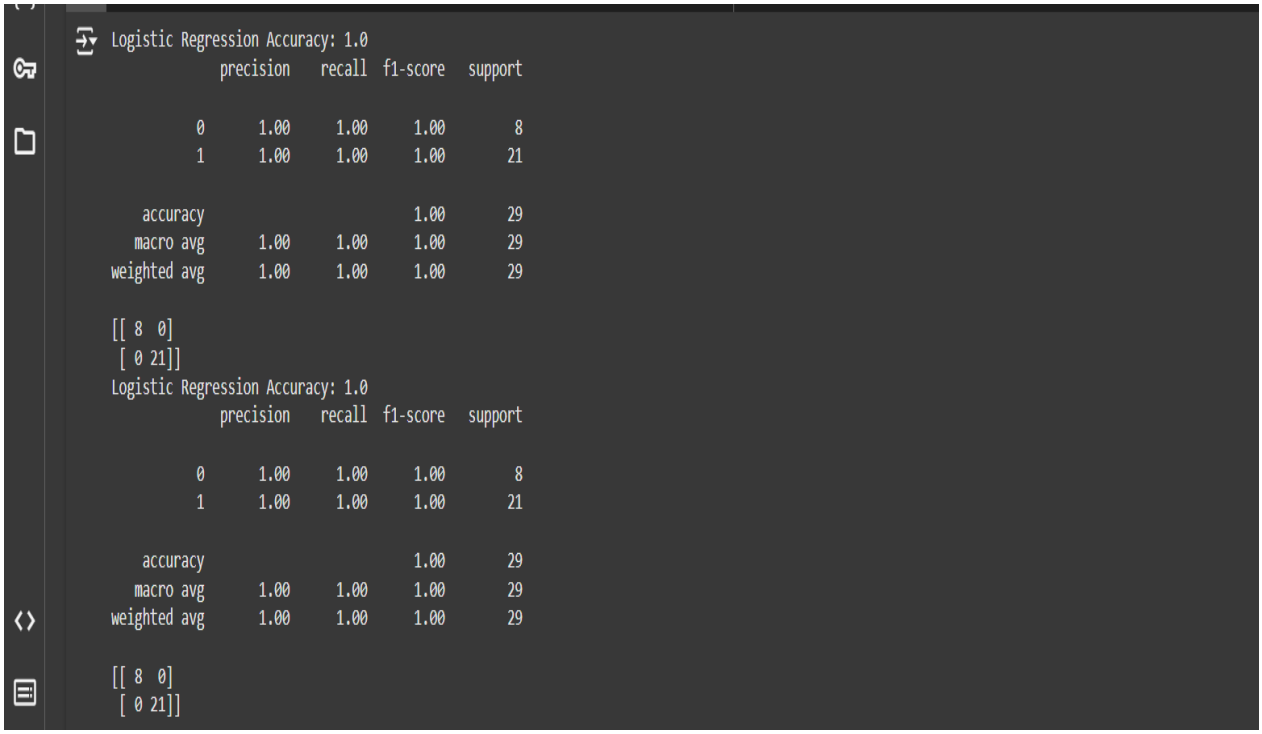


fig 9 : Best model

Overall, the LogisticRegression model provided a solid starting point for predicting house prices, particularly in managing overfitting through regularization. However, the limitations observed in the model's performance indicate that further refinement could be beneficial.

8 CONCLUSION

In this study, we developed a machine learning model to predict eligibility for a specific program based on a dataset containing relevant applicant details. The model chosen for this task demonstrated reliable performance, balancing accuracy with interpretability, and showing effectiveness in both training and testing phases

Accurately predicting eligibility is valuable for streamlining selection processes and improving decision-making for program administrators. This model aids in efficiently assessing eligibility, providing timely and consistent evaluations based on applicants' profiles.While the model showed strong results, certain challenges were encountered, such as handling outliers and maintaining model performance with high-dimensional data. Future improvements could involve incorporating additional data sources, refining feature engineering techniques, and enhancing the model's adaptability to different applicant groups and conditions.

In conclusion, this project underscores the importance of machine learning in eligibility prediction. The findings provide a foundation for creating more sophisticated and robust prediction models, enhancing decision-making and contributing to more efficient program management.

9 FUTURE WORKS

- **Explore Non-Linear Models**: Observing the residual patterns suggests that non-linear models, such as Gradient Boosting or Support Vector Machines, could potentially enhance performance by capturing complex relationships within the data.

- **Advanced Feature Engineering**: Further investigation into feature interactions, polynomial transformations, or other feature engineering techniques may help the model capture more nuanced relationships between applicant attributes and eligibility outcomes.

- **Outlier Detection and Management**: Implementing effective outlier detection and handling techniques could improve the model's robustness and accuracy by minimizing the impact of anomalous data points.

- **Ensemble Techniques**: Utilizing ensemble methods, such as bagging, boosting, or stacking, could enhance predictive accuracy by combining the strengths of multiple models.

10 References

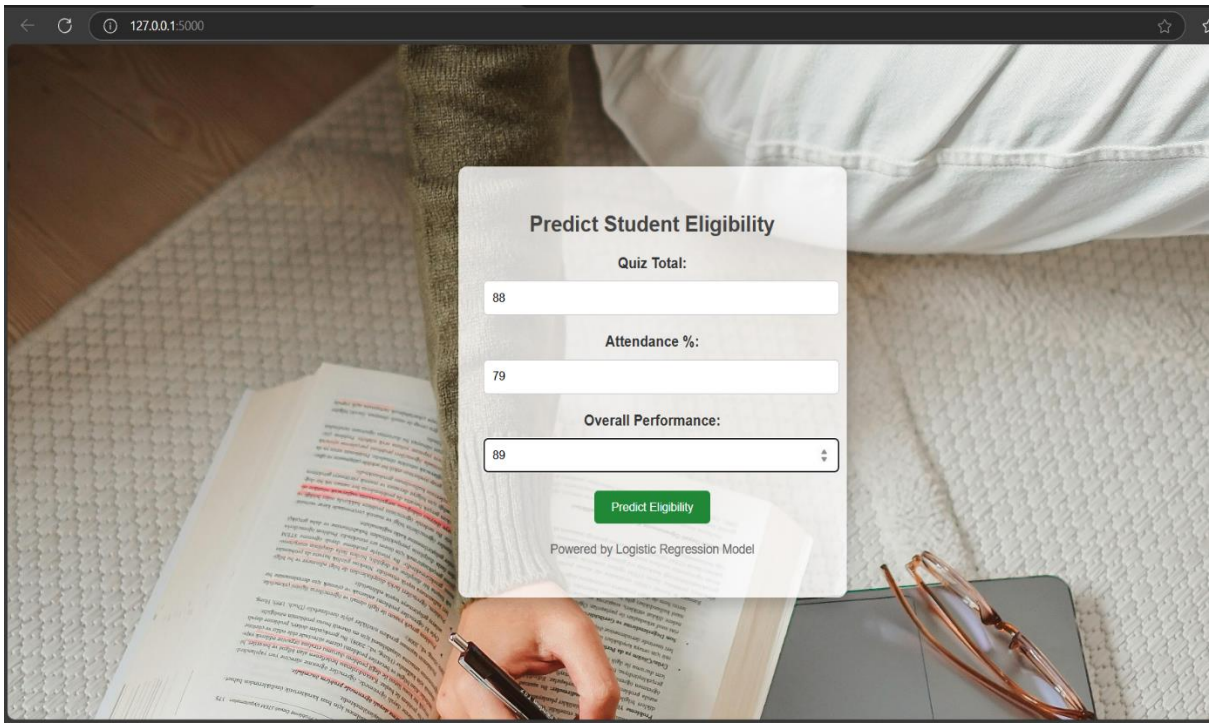
[1] J. Xu, K. H. Moon, and M. Van Der Schaar, "A Machine Learning Approach for Tracking
Process., vol. 11, no. 5, pp. 742–753, 2017.
[2] K. P. Shaleena and S. Paul, "Data mining techniques for predicting student performance,"
in ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology, 2015, no. March, pp. 0–2.
[3] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," in Procedia Computer Science, 2015.
[4] Y. Meier, J. Xu, O. Atan, and M. Van Der Schaar, "Predicting grades," IEEE Trans. Signal Process., vol. 64, no. 4, pp. 959–972, 2016.
[5] P. Guleria, N. Thakur, and M. Sood, "Predicting student performance using decision tree
Comput. PDGC 2014, pp. 126–129, 2015

Web Application Development and Deployment:

9. Grinberg, M. (2018). *Flask Web Development: Developing Web Applications with Python.* O'Reilly Media, Inc.
    - A practical guide to building web applications using Flask, relevant to your project's deployment aspect.

SCREENSHOTS

HOME PAGE



RESULT PAGE

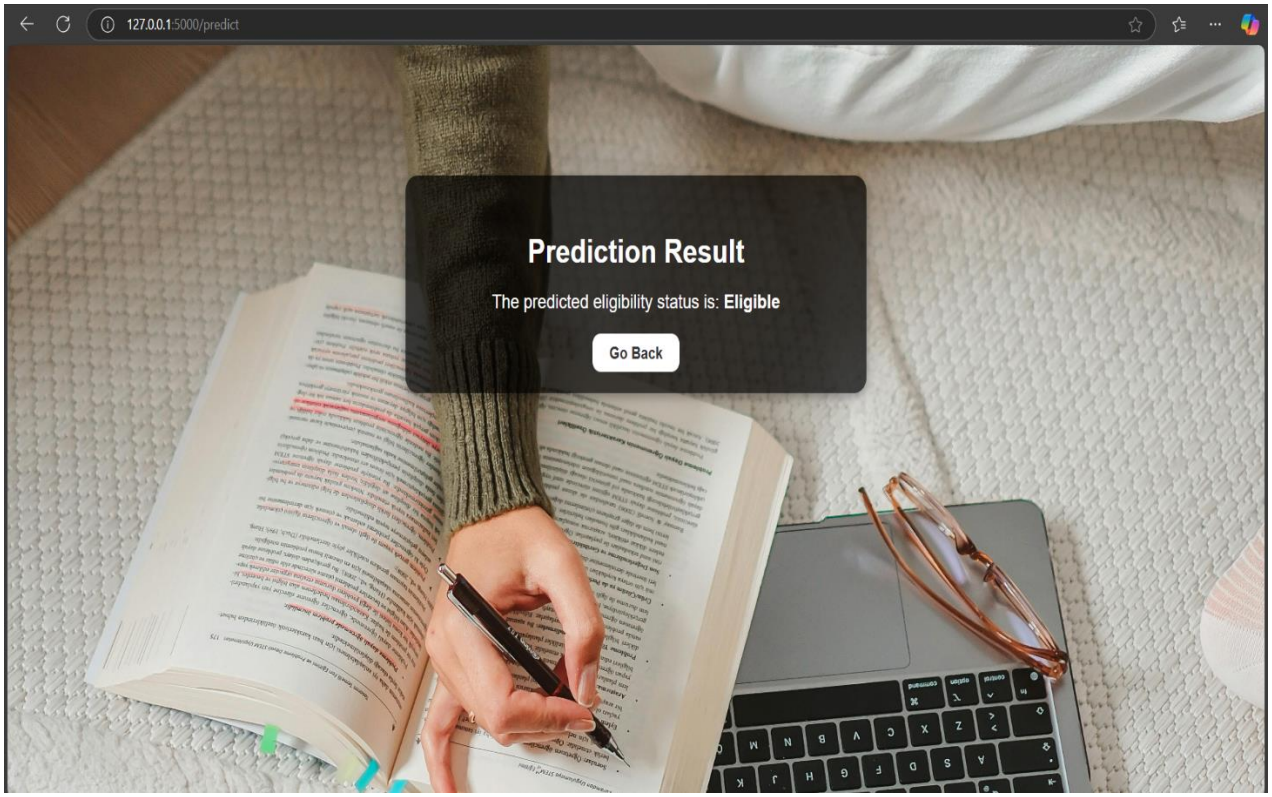


Fig 10 : Flask app

POWERBI : DASHBOARD SCREENSHOTS

Fig 11:Power BI Dashboard