Zohaib Aamer 20F-0168

Junaid Farooq 20F-0427

Abdul Rehman 20F-0238

Muhammad Shayan 20F-0348

Data Mining Project Report

# Introduction

Understanding student performance and projecting their academic results are critical for educators and institutions in the modern educational environment. The emergence of data-driven methodologies has made it essential to use datasets that include student assessment results in this undertaking. The goal of the current initiative is to use this data to forecast students' grades prior to important exams, such as the Mid-II and Final exams, earlier in their academic careers.

The dataset provided for this project comprises students' assessment scores, including assignments, quizzes, and Mid-I and Mid-II exams, along with a predictor variable indicating the final grade. To facilitate analysis, the data is distributed across seven sheets, each varying in the number of assignments and quizzes included. Before calculating grades, only the best five assignments and quizzes for each student have been considered, as specified in the dataset.

Exploratory Data Analysis (EDA) plays a pivotal role in understanding and preprocessing the dataset, laying the foundation for subsequent phases of the project. Through EDA, we aim to unravel hidden patterns, detect anomalies, and gain insights into the relationships between different assessment types and the final grades. By exploring the data comprehensively, we can make informed decisions regarding preprocessing steps and model selection, ultimately enhancing the accuracy of grade predictions.

# Data Preprocessing

In preparation for model training and analysis, several preprocessing steps were applied to the dataset to ensure its suitability for predictive modeling. Below is a description of the preprocessing steps performed, along with the rationale behind each step:

## Combining Data from Multiple Sheets:

1. The dataset provided was distributed across seven sheets, each containing different subsets of assessment scores.
2. To streamline analysis and modeling, data from all sheets were extracted and combined into a single CSV file.
3. This consolidation ensures consistency in data handling and facilitates seamless processing during subsequent phases of the project.

## Creation of Two Separate Datasets:

1. Given the project's objectives to predict students' grades before and after the Mid-II exam, two distinct datasets were created.
2. The first dataset is intended for predicting grades before the Mid-II exam and includes the following attributes: 'As:1', 'As:2', 'As:3', 'As:4', 'Qz:1', 'Qz:2', 'Qz:3', 'Qz:4', and 'S-I'.
3. The second dataset is tailored for predicting grades after the Mid-II exam and contains attributes: 'As', 'Qz', 'S-I', and 'S-II'.
4. This segregation allows for focused analysis and model training based on the specific requirements of each prediction scenario.

## Handling Missing Values:

1. In any dataset, missing values can pose challenges during analysis and modeling.
2. For this project, missing values were identified and addressed by imputing zeros ('0').
3. Since the objective is to predict students' grades, alternative methods such as mean value imputation were deemed unsuitable, as they could potentially distort the predictive accuracy by artificially inflating or deflating scores.
4. Imputing zeros maintains the integrity of the dataset while ensuring that missing values do not adversely affect grade predictions.

# **Summary Statistics**

## Before Mid-II Exam:

### **Assignment Scores**:

### As:1: Mean = 66.12, Standard Deviation = 33.07

### As:2: Mean = 59.40, Standard Deviation = 25.83

### As:3: Mean = 84.62, Standard Deviation = 34.11

### As:4: Mean = 54.09, Standard Deviation = 31.70

* As:1: Min = 0, Max = 127, IQR = 48.375
* As:2: Min = 0, Max = 100, IQR = 32.750
* As:3: Min = 0, Max = 140, IQR = 39.875
* As:4: Min = 0, Max = 130, IQR = 36.000

### **Quiz Scores**:

### Qz:1: Mean = 5.55, Standard Deviation = 4.25

### Qz:2: Mean = 6.77, Standard Deviation = 6.39

### Qz:3: Mean = 3.39, Standard Deviation = 3.48

### Qz:4: Mean = 3.42, Standard Deviation = 3.82

* Qz:1: Min = 0, Max = 20, IQR = 4.500
* Qz:2: Min = 0, Max = 30.5, IQR = 7.875
* Qz:3: Min = 0, Max = 17.5, IQR = 4.400
* Qz:4: Min = 0, Max = 10, IQR = 7.000

### **Mid-I Score**:

* + - * Mean = 5.72, Standard Deviation = 2.348
      * Min = 0, Max = 13.87, IQR = 2.885

## Before Final Exam:

### **Assignment Scores**:

* + - * Mean = 11.08, Standard Deviation = 2.52
      * Min = 0, Max = 14.87, IQR = 2.88

### **Quiz Scores**:

* + - * Mean = 5.60, Standard Deviation = 1.97
      * Min = 0.2, Max = 10, IQR = 2.665

### **Mid-II Score**:

* + - * Mean = 4.89, Standard Deviation = 2.71
      * Min = 0, Max = 12.37, IQR = 4.108

# **Exploratory Data Analysis**

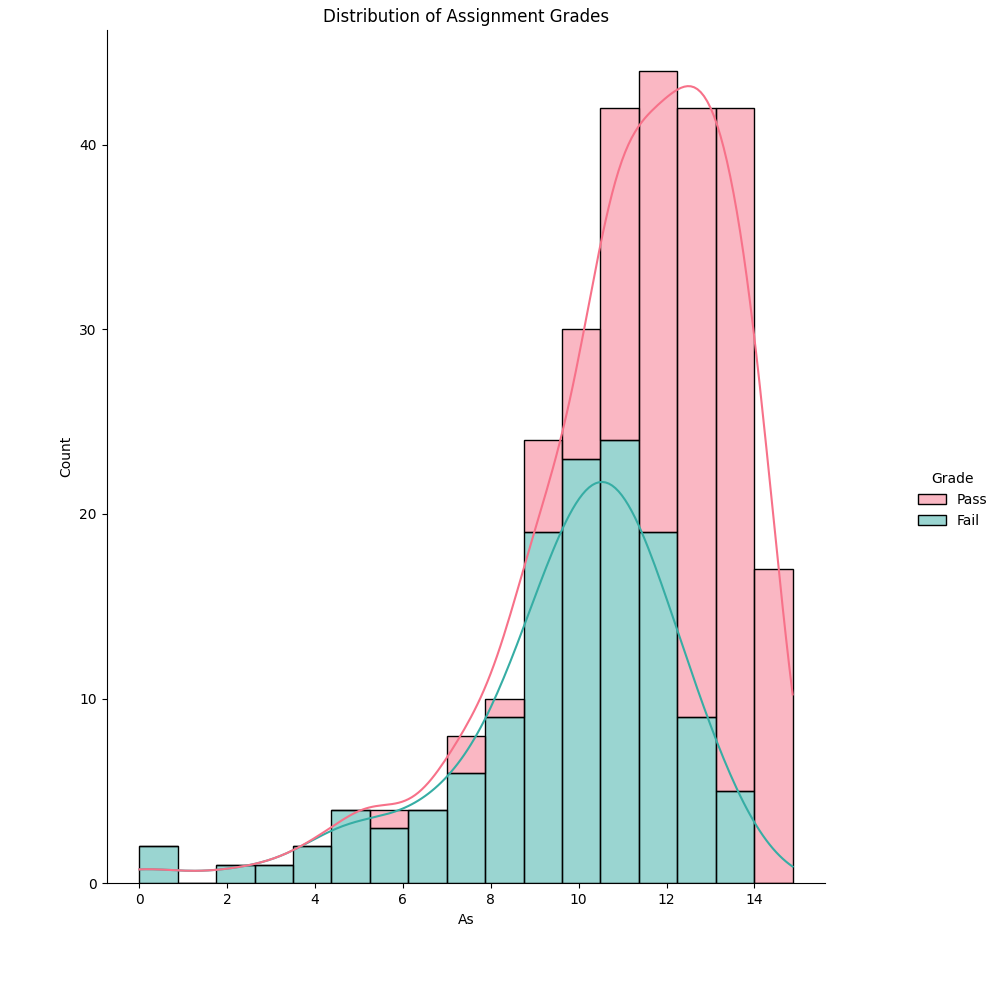
# 

## Percentage of Students Who Passed the Course

* The pie chart illustrates the proportion of students who passed and failed the course.
* Approximately 52.71% of students passed the course, while 47.29% failed.

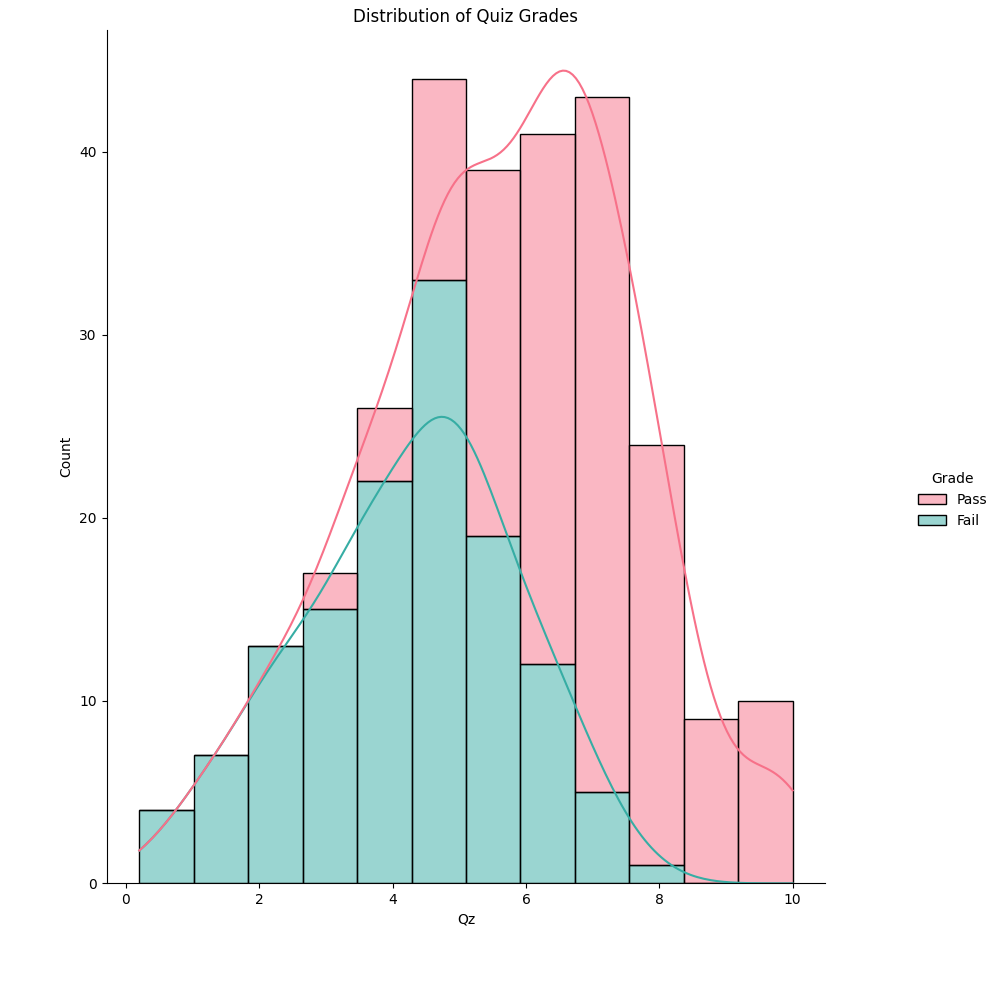
## Distribution of Assignment and Quiz Grades

The distributions of assignment and quiz grades are visualized using histograms (Image 1 and Image 2, respectively). These visualizations provide insights into the performance of students across different assessments.



### Assignment Grades

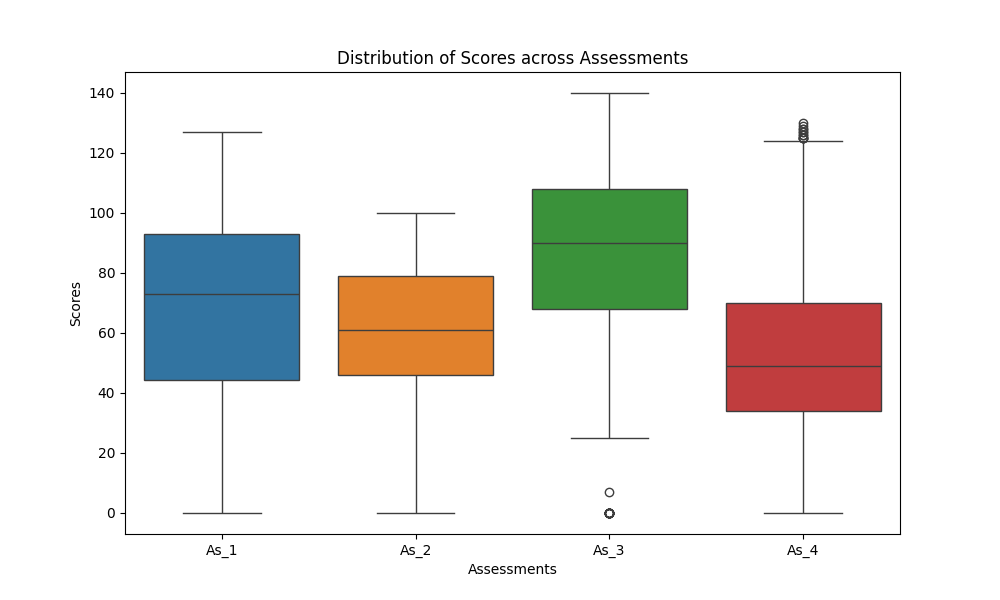
* The distribution of assignment grades shows a bimodal pattern, with peaks at lower and higher scores.
* There is a clear separation between students who passed and those who failed, with the majority of failing students scoring lower on assignments.

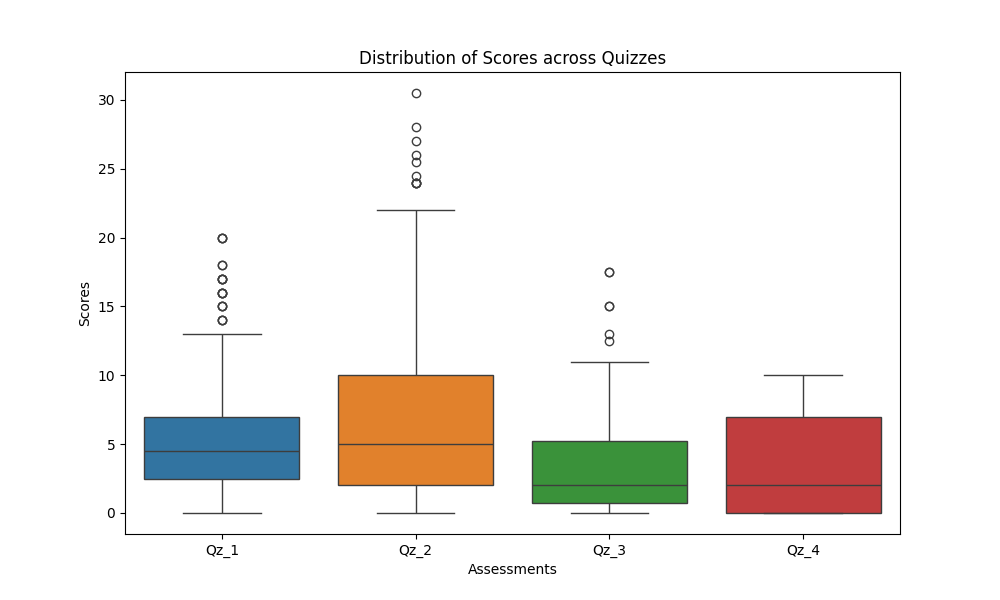


### Quiz Grades

* The distribution of quiz grades also exhibits a bimodal pattern, with peaks at lower and higher scores.
* Like assignment grades, there is a distinction between passing and failing students, with failing students generally scoring lower on quizzes.

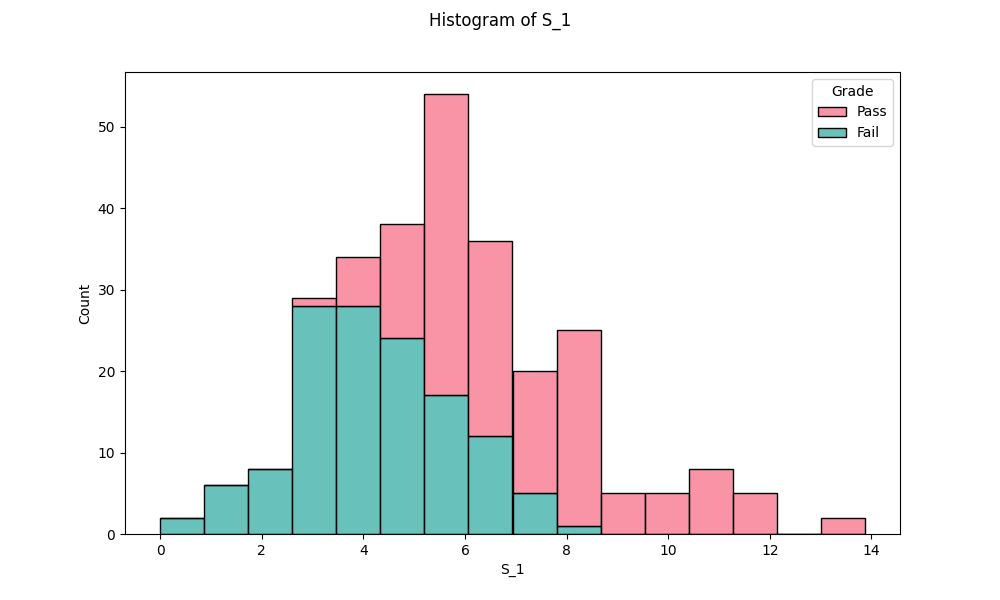
## Distribution of Scores across Assessments





* These box plots provide an overview of the score distributions for individual assignments and quizzes.
* For assignments, the distributions vary, with Assignment 3 (As\_3) having a higher median score compared to others.
* For quizzes, the distributions are more consistent, with similar median scores across all quizzes.
* Outliers are present in both assignments and quizzes, indicating some students performed exceptionally well or poorly on specific assessments.

## **Distribution of Mid 1 (S-1) Grades**



* The histogram displays the distribution of scores for "S\_1".
* The distribution is bimodal, with peaks around scores of 3-4 and 7-8.
* There is a clear separation between the passing (green bars) and failing (pink bars) scores.

## Distribution of Scores across Mid 1(S-1) and Mid 2 (S-2)

## 

* This scatterplot shows the relationship between scores on two Mids, "S\_1" (x-axis) and "S\_2" (y-axis).
* The data points are color-coded to indicate whether the student passed (blue) or failed (orange) based on their combined scores on these two assessments.
* There appears to be a positive correlation between the scores on the two assessments, with students who scored higher on one assessment generally scoring higher on the other as well.
* However, there is also a significant amount of variability and overlap between the passing and failing groups.

# Model Evaluation

## Predicting Grades before Mid-II Exam

We use the first four assignments, first four quizzes, and the Mid-I score as features for this prediction task.

### Nearest Neighbor Classifier:

* Accuracy: 69.64%
* Precision:
  + Class 0: 67%
  + Class 1: 72%
* Recall:
  + Class 0: 69%
  + Class 1: 70%
* F1-score:
  + Class 0: 68%
  + Class 1: 71%

### Decision Tree Classifier:

* Accuracy: 76.79%
* Precision:
  + Class 0: 72%
  + Class 1: 81%
* Recall:
  + Class 0: 81%
  + Class 1: 73%
* F1-score:
  + Class 0: 76%
  + Class 1: 77%

The Decision Tree classifier outperforms the Nearest Neighbor classifier in predicting students' grades before the Mid-II exam, with an accuracy of 76.79% compared to 69.64%.

While both classifiers exhibit similar precision and recall for class 1 (passing students), the Decision Tree classifier demonstrates higher precision and recall for class 0 (failing students). This indicates that the Decision Tree classifier is better at correctly identifying failing students, contributing to its overall higher accuracy.

## Predicting Grades after Mid-II Exam

We use the best five assignments, best five quizzes, and the Mid-I score as features for this prediction task.

### Nearest Neighbor Classifier:

* Accuracy: 87.5%
* Precision:
  + Class 0: 88%
  + Class 1: 87%
* Recall:
  + Class 0: 85%
  + Class 1: 90%
* F1-score:
  + Class 0: 86%
  + Class 1: 89%

### Decision Tree Classifier:

* Accuracy: 87.5%
* Precision:
  + Class 0: 100%
  + Class 1: 81%
* Recall:
  + Class 0: 73%
  + Class 1: 100%
* F1-score:
  + Class 0: 84%
  + Class 1: 90%

Both classifiers achieved an accuracy of 87.5% in predicting students' grades after the Mid-II exam. However, there are differences in precision, recall, and F1-score for each class between the two classifiers.

The Nearest Neighbor classifier demonstrates balanced performance with high precision and recall for both classes, indicating its effectiveness in correctly classifying both passing and failing students. On the other hand, the Decision Tree classifier exhibits slightly imbalanced performance, with perfect precision for class 0 (fail) but lower recall compared to class 1 (pass), suggesting potential issues in correctly identifying failing students.

# Conclusion

Based on the evaluation of two different classifiers for predicting students' grades before and after the Mid-II exam, several key insights emerge.

Firstly, when predicting grades before the Mid-II exam, the Decision Tree classifier outperforms the Nearest Neighbor classifier with an accuracy of 76.79% compared to 69.64%. This indicates that the Decision Tree model is better at identifying failing students, as reflected in its higher precision and recall for class 0.

Conversely, when predicting grades after the Mid-II exam, both classifiers achieve the same accuracy of 87.5%. However, they exhibit different strengths and weaknesses. The Nearest Neighbor classifier demonstrates balanced performance with high precision and recall for both passing and failing students, indicating its effectiveness in classifying both classes accurately. In contrast, the Decision Tree classifier shows perfect precision for failing students but lower recall, suggesting potential issues in identifying failing students correctly.

In conclusion, while both classifiers achieve high accuracy in predicting grades after the Mid-II exam, the Nearest Neighbor classifier shows more balanced performance across both classes. However, the Decision Tree classifier excels in identifying failing students before the Mid-II exam. Depending on the specific requirements and priorities of the prediction task, either classifier could be chosen for implementation. Additionally, further investigation into the Decision Tree classifier's recall for failing students after the Mid-II exam may be warranted to address potential issues in correctly identifying students at risk.

You can find the code for the student grade prediction task at the following GitHub repository: [**StudentGradePrediction**](https://github.com/mshayan3/StudentGradePrediction)