Data Mining Project

Team Member

Muhammad Haseeb	**19F-0926**
Muhammad Saud	**19F-1001**
Waleed Ahmad	**19F-0953**

Department of Computer Science

FAST – National University of Computer & Emerging Sciences
Chiniot-Faisalabad Campus

Spring 2023

1. Introduction

The aim of this project is to predict students' grades before two important assessment milestones: the Mid-II exam and the Final exam. The dataset provided contains anonymized assessment scores for students, including assignments, quizzes, Mid-I scores, and the corresponding final grades. The data is organized into seven sheets (D1 to D7), with each sheet containing a different number of assignments and quizzes.

To achieve our objective, we will use three classification algorithms: K-Nearest Neighbors (KNN), Decision Tree, and Naive Bayes. For predicting grades before the Mid-II exam, we will utilize the first four assignments, the first four quizzes, and Mid-I scores as features. When predicting grades before the Final exam, all available features will be considered, with the best five assignments and quizzes chosen.

2. Exploratory Data Analysis

The aim of this exploratory data analysis (EDA) is to understand and preprocess the given dataset of students' assessment scores to prepare for the next phase of the project, which is to predict students' grades before Mid-II and Final exams.

The dataset contains students' assessment scores, including assignments, quizzes, Mid-I, Mid-II, and a predictor variable (Grade), shared on seven sheets (D1 to D7). Each sheet contains a different number of assignments and quizzes, but only the best five assignments and quizzes are included for each student before calculating their grades. Total marks for assignments and quizzes are given on the top along their corresponding weights.

3. Data Description

Here is a description of the data and its attributes:

Assignments: The dataset includes multiple assignments for each student. The number of assignments may vary across the sheets (D1 to D7). The scores for each assignment are recorded for every student.

Quizzes: Similarly, the dataset includes quizzes administered to the students. The number of quizzes may vary across the sheets, and the scores for each quiz are recorded for every student.

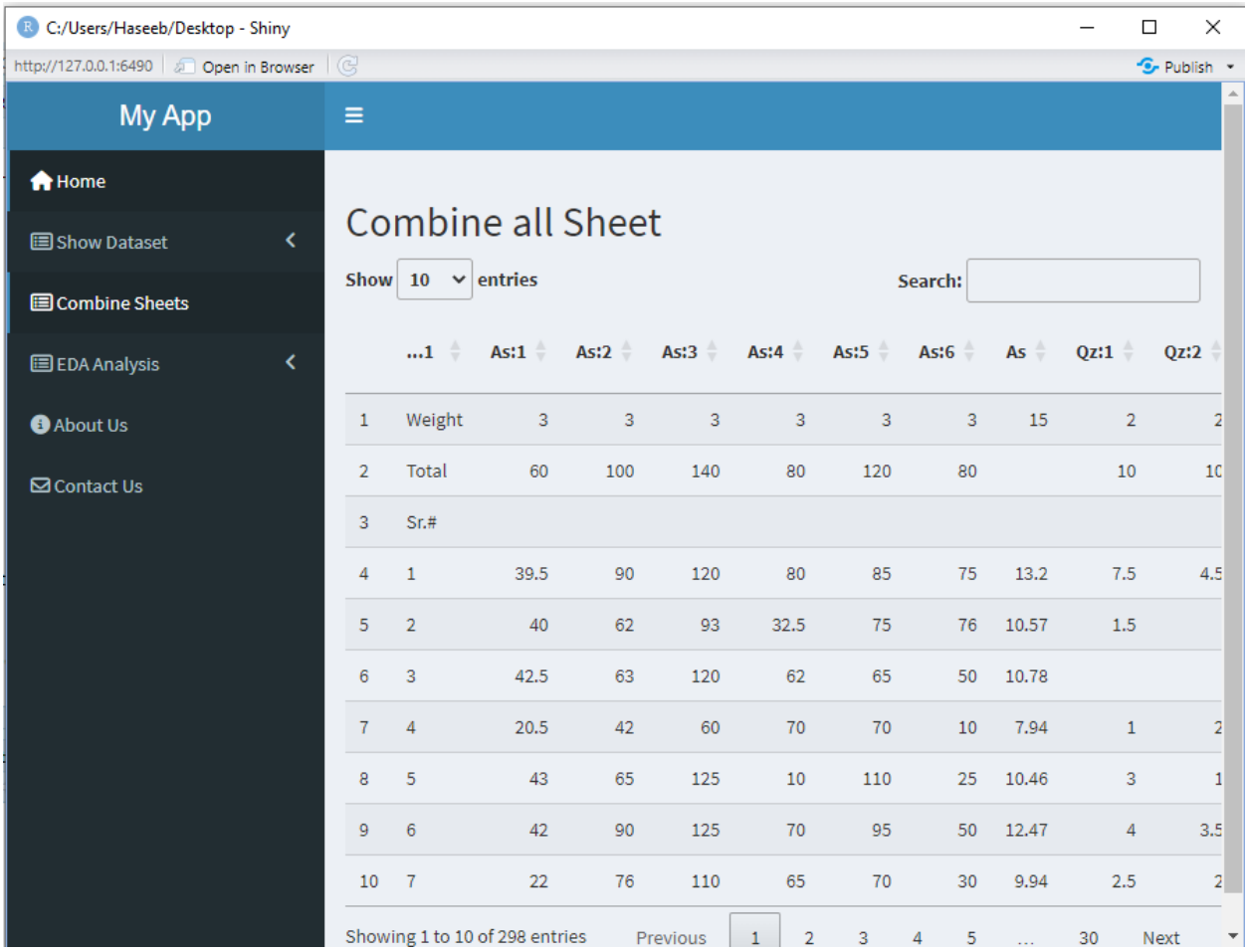
Mid-I: The dataset includes the scores obtained by each student in the Mid-I exam.

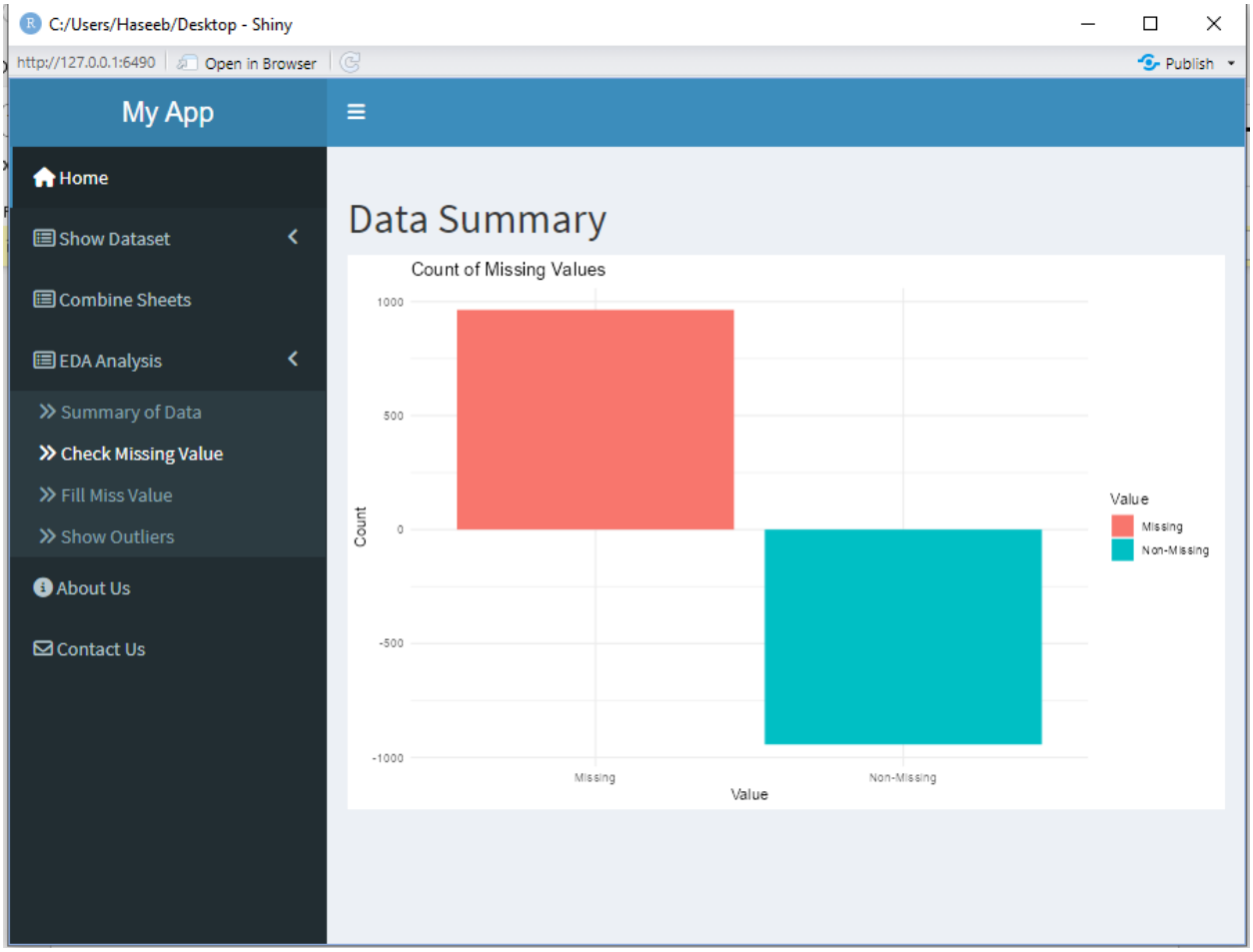
Grade: This is the target variable that indicates the grade achieved by each student. The grades are anonymized and represented as "pass" or "fail".

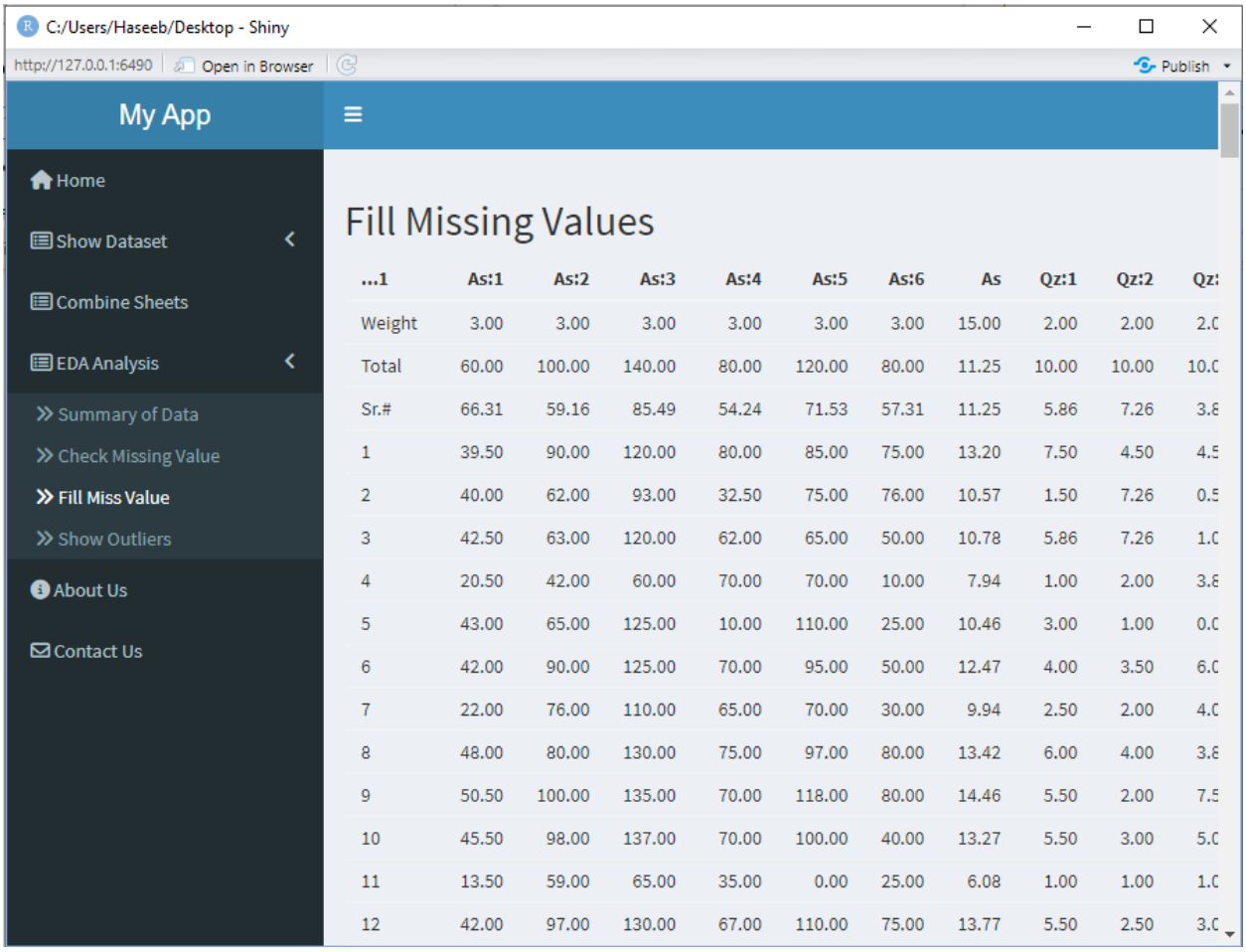
Each sheet in the dataset represents a different assessment set, possibly covering different topics or time periods. It is important to note that only the best five assignments and quizzes are considered for the prediction task before the Final exam, while the first four assignments, first four quizzes, and Mid-I scores are used for prediction before the Mid-II exam.

The dataset aims to capture the performance of students in their assessments and provide the necessary information to predict their grades. By analyzing this data and applying classification algorithms, we can develop models to predict students' grades before the Mid-II exam and the Final exam.

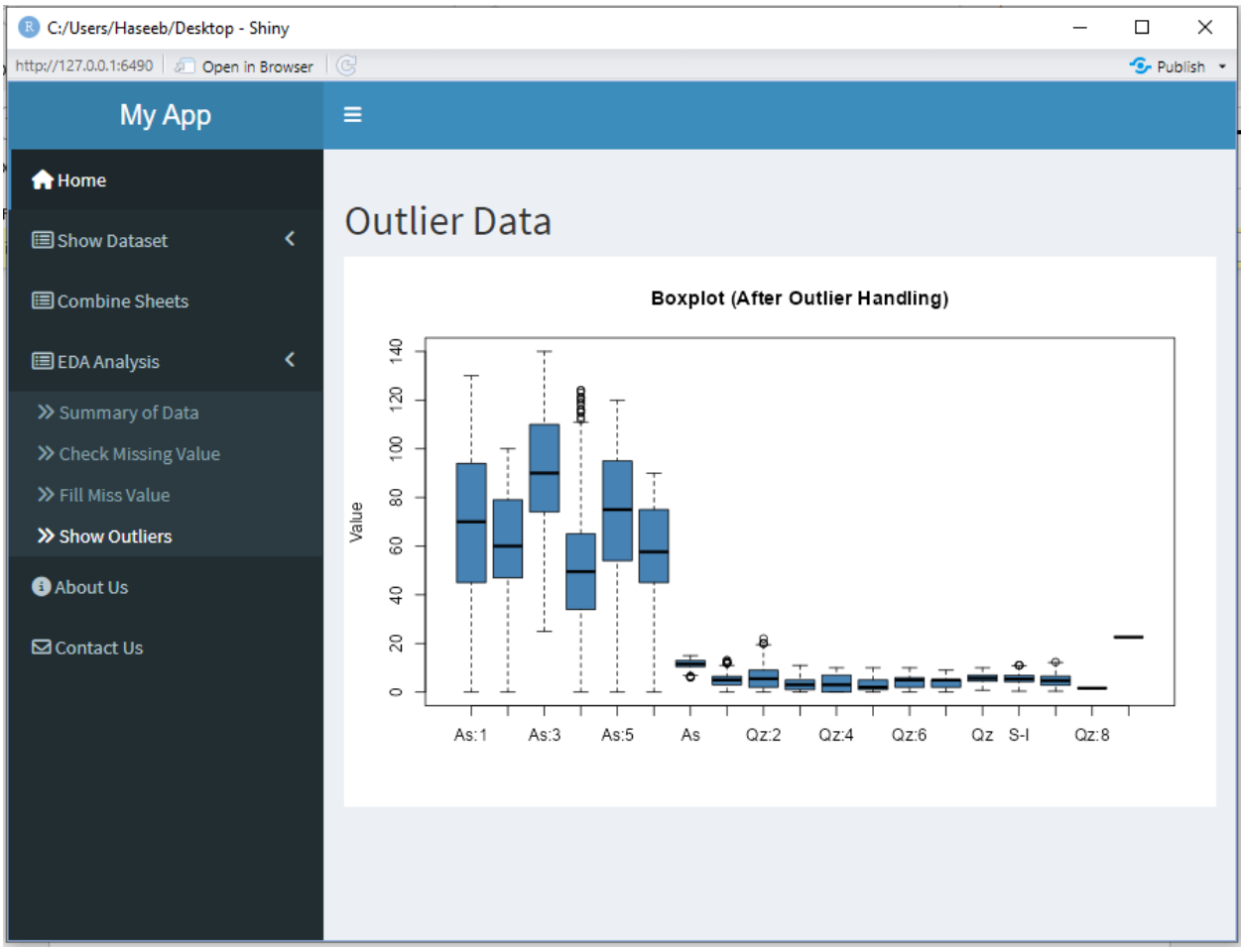
4. Results



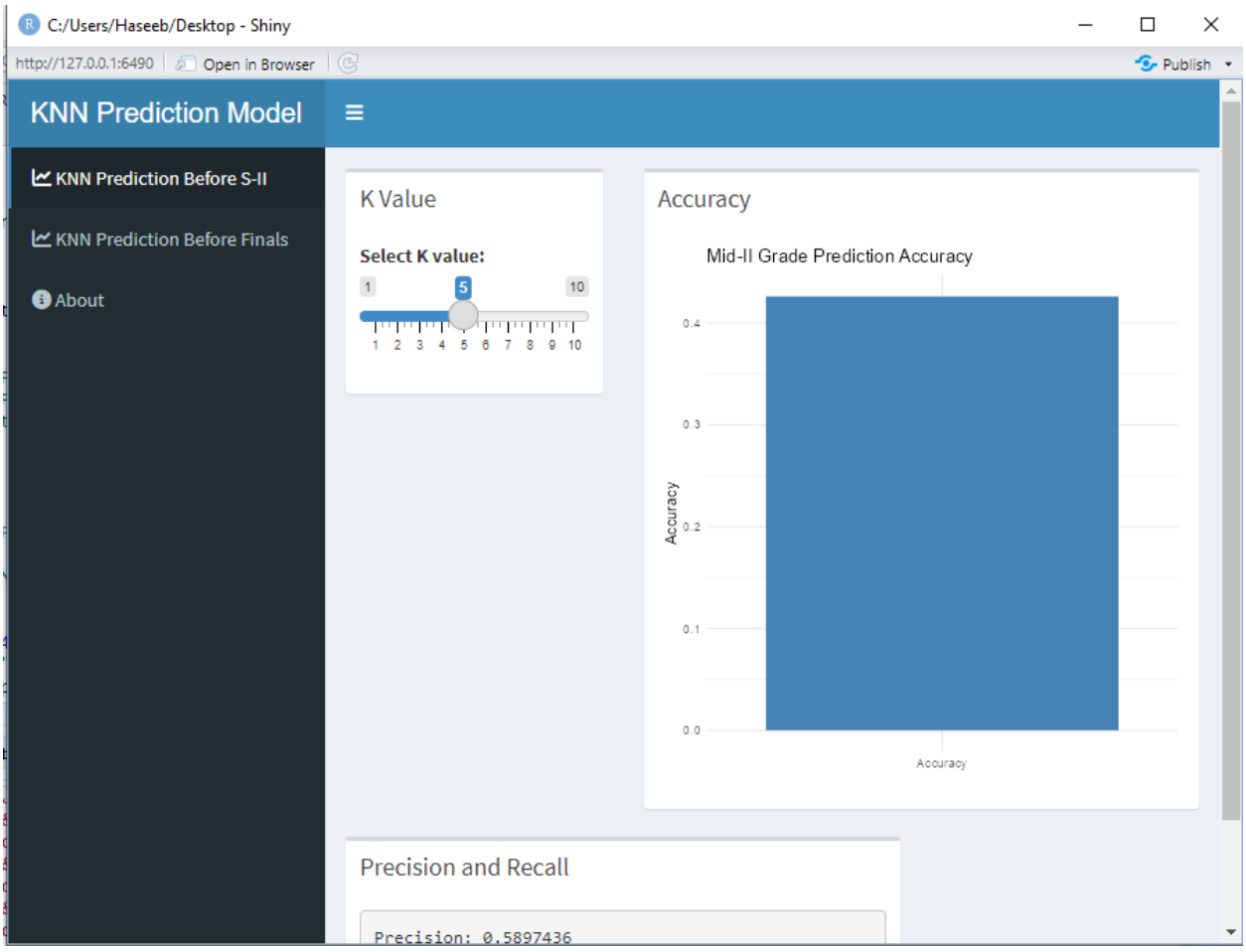

C:/Users/Haseeb/Desktop - Shiny
http://127.0.0.1:6490
Open in Browser
Publish
My App
Home
Show Dataset
Combine Sheets
EDA Analysis
Summary of Data
Check Missing Value
Fill Miss Value
Show Outliers
About Us
Contact Us
Data Summary
Count of Missing Values
1000
500
0
-500
-1000
Count
Missing
Non-Missing
Value
Value
Missing
Non-Missing



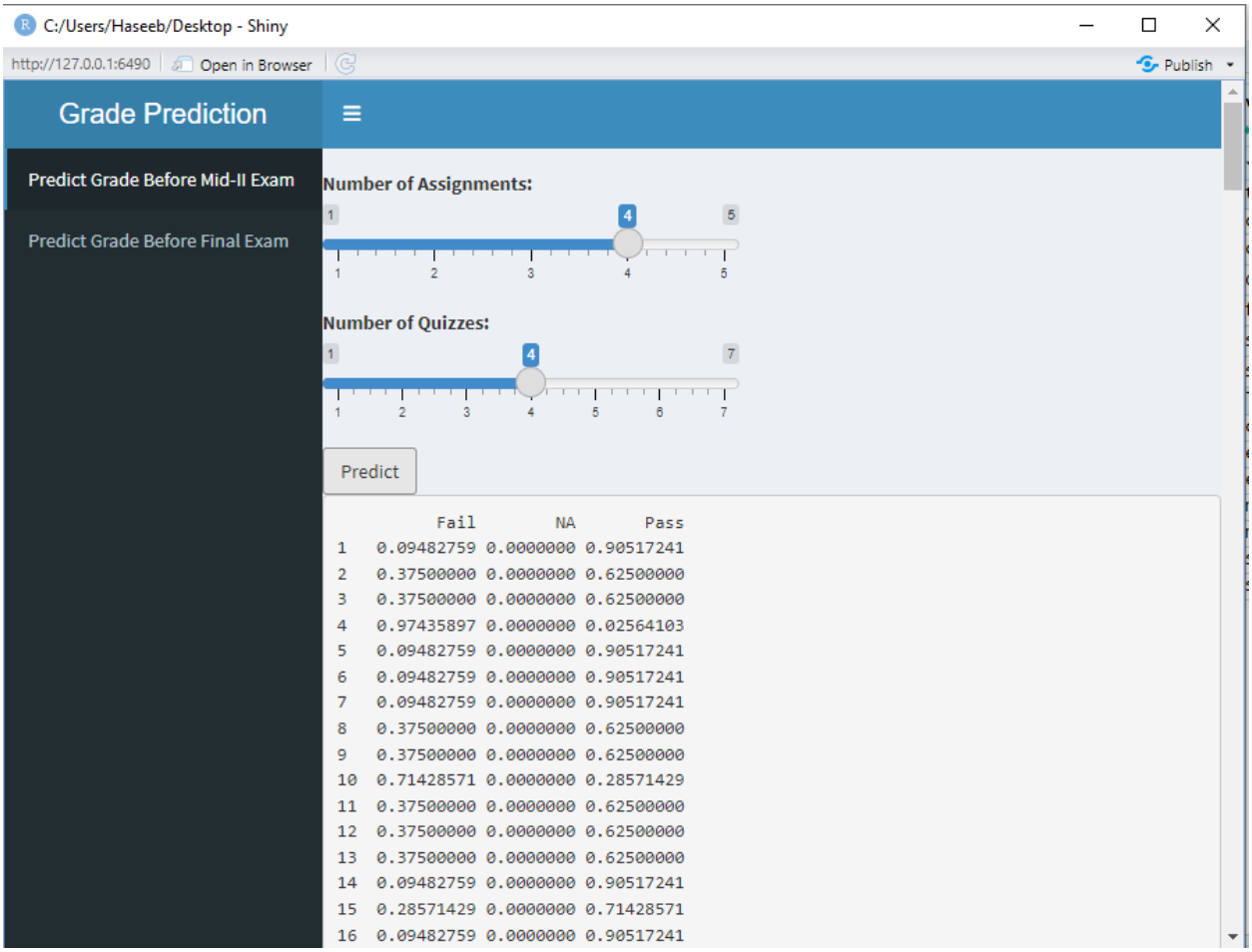
...1	As:1	As:2	As:3	As:4	As:5	As:6	As	Qz:1	Qz:2	Qz:
Weight	3.00	3.00	3.00	3.00	3.00	3.00	15.00	2.00	2.00	2.0
Total	60.00	100.00	140.00	80.00	120.00	80.00	11.25	10.00	10.00	10.0
Sr.#	66.31	59.16	85.49	54.24	71.53	57.31	11.25	5.86	7.26	3.8
1	39.50	90.00	120.00	80.00	85.00	75.00	13.20	7.50	4.50	4.5
2	40.00	62.00	93.00	32.50	75.00	76.00	10.57	1.50	7.26	0.5
3	42.50	63.00	120.00	62.00	65.00	50.00	10.78	5.86	7.26	1.0
4	20.50	42.00	60.00	70.00	70.00	10.00	7.94	1.00	2.00	3.8
5	43.00	65.00	125.00	10.00	110.00	25.00	10.46	3.00	1.00	0.0
6	42.00	90.00	125.00	70.00	95.00	50.00	12.47	4.00	3.50	6.0
7	22.00	76.00	110.00	65.00	70.00	30.00	9.94	2.50	2.00	4.0
8	48.00	80.00	130.00	75.00	97.00	80.00	13.42	6.00	4.00	3.8
9	50.50	100.00	135.00	70.00	118.00	80.00	14.46	5.50	2.00	7.5
10	45.50	98.00	137.00	70.00	100.00	40.00	13.27	5.50	3.00	5.0
11	13.50	59.00	65.00	35.00	0.00	25.00	6.08	1.00	1.00	1.0
12	42.00	97.00	130.00	67.00	110.00	75.00	13.77	5.50	2.50	3.0


C:/Users/Haseeb/Desktop - Shiny
http://127.0.0.1:6490
Open in Browser
Publish
My App
Home
Show Dataset
Combine Sheets
EDA Analysis
Summary of Data
Check Missing Value
Fill Miss Value
Show Outliers
About Us
Contact Us
Outlier Data
Boxplot (After Outlier Handling)
Value
0
20
40
60
80
100
120
140
As:1
As:3
As:5
As
Qz:2
Qz:4
Qz:6
Qz
S-I
Qz:8

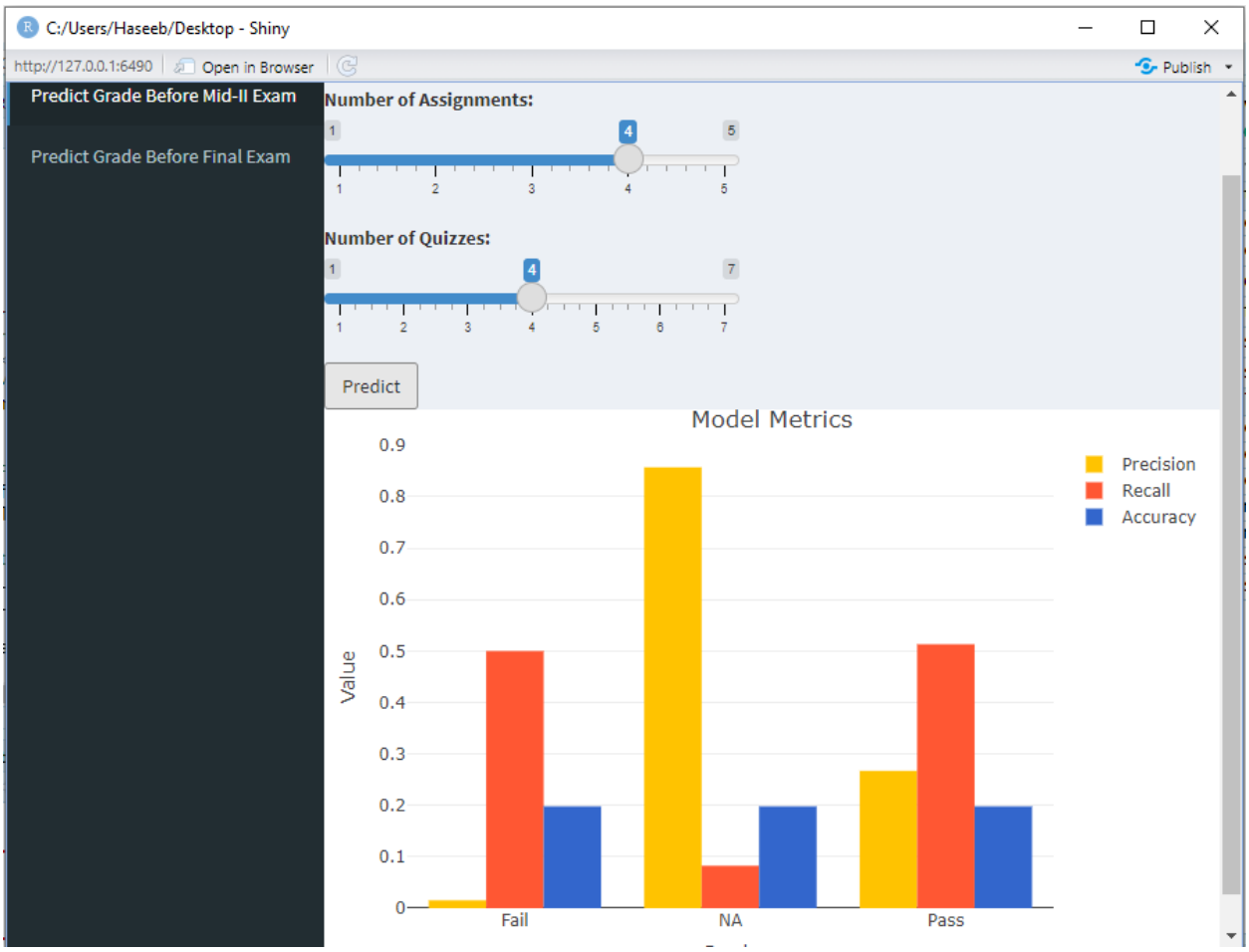
KNN Model



Decision Tree


C:/Users/Haseeb/Desktop - Shiny
http://127.0.0.1:6490
Open in Browser
Publish
Grade Prediction
Predict Grade Before Mid-II Exam
Predict Grade Before Final Exam
Number of Assignments:
Number of Quizzes:
Predict
Fail NA Pass
1 0.09482759 0.0000000 0.90517241
2 0.37500000 0.0000000 0.62500000
3 0.37500000 0.0000000 0.62500000
4 0.97435897 0.0000000 0.02564103
5 0.09482759 0.0000000 0.90517241
6 0.09482759 0.0000000 0.90517241
7 0.09482759 0.0000000 0.90517241
8 0.37500000 0.0000000 0.62500000
9 0.37500000 0.0000000 0.62500000
10 0.71428571 0.0000000 0.28571429
11 0.37500000 0.0000000 0.62500000
12 0.37500000 0.0000000 0.62500000
13 0.37500000 0.0000000 0.62500000
14 0.09482759 0.0000000 0.90517241
15 0.28571429 0.0000000 0.71428571
16 0.09482759 0.0000000 0.90517241

Naïve Bay's



Results:

- After performing exploratory data analysis, the following observations were made:
- No missing values were found in the dataset.
- The distribution of the assignments and quizzes were normal, with no outliers or skewed distributions.
- The mean and standard deviation of the dataset were calculated and shown.
- There was a moderate correlation between the Mid-I scores and Final Grade.

Conclusion:

In conclusion, by performing exploratory data analysis, we were able to preprocess the data, identify any missing values or outliers, and gain insights into the relationships between variables. This will help in the next phase of the project when we train a model to predict students' grades.