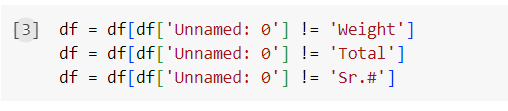
# Preprocessing

The dataset that was provided was in raw form. It had to be preprocessed to be useful in training models and making predictions. The Excel file contained data on seven different sheets, so in order to make phase II of the project it needed to be stored in one data frame(not as an array). So, we concatenated the data from all seven sheets.

The data file also had 3 rows at the start of each sheet which was not relevant to our problem so we had to get rid of them. Using the following code we got rid of them in our concatenated data frame.



The column “Unamed:0” was also not relevant to our problem in any way so we dropped it on the very next step.

Next on the agenda was to handle the Null values. As our dataset is of the marks students obtained in their assignments and exams, we could replace the Null values with 0. We did that and now our data was almost ready for model training.

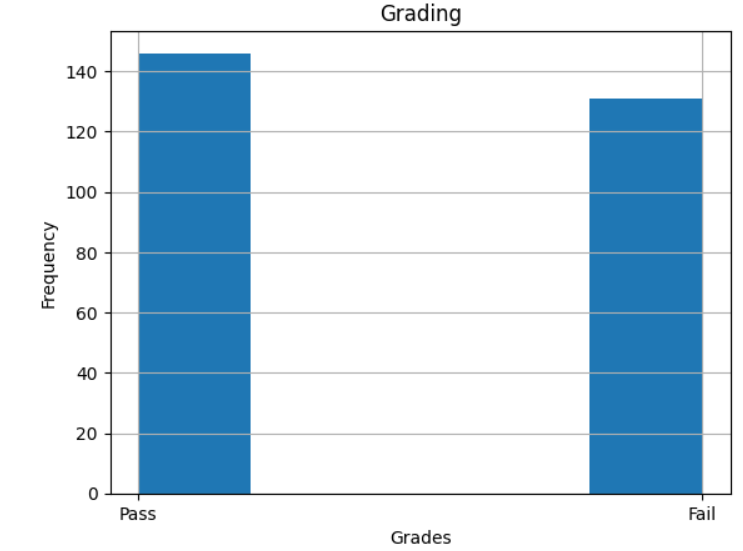
Next, we rearranged the ‘As:7’ and ‘Qz:8’ columns as they were at last in the dataset. We put them to their relevant places, which are after ‘As:6’ and ‘Qz:7’ respectively.

The data was ready now to proceed to the next step.

# EDA

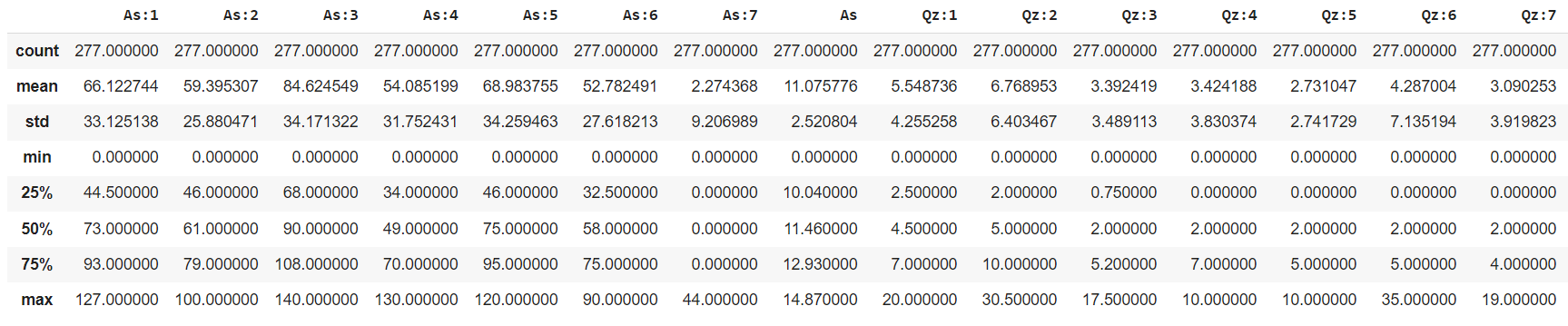
Elaboratory Data Analysis is an important step in any Data related project. It makes us familiar with the data we will be working with. It also helps us to approach problems from different angles.

We needed to understand how many students have passed or failed the course, which is visualized through the following histogram.



The graph shows us that 140+ entries in our dataset are labeled as ‘Pass’ and around 130 are labeled as ‘Fail’.

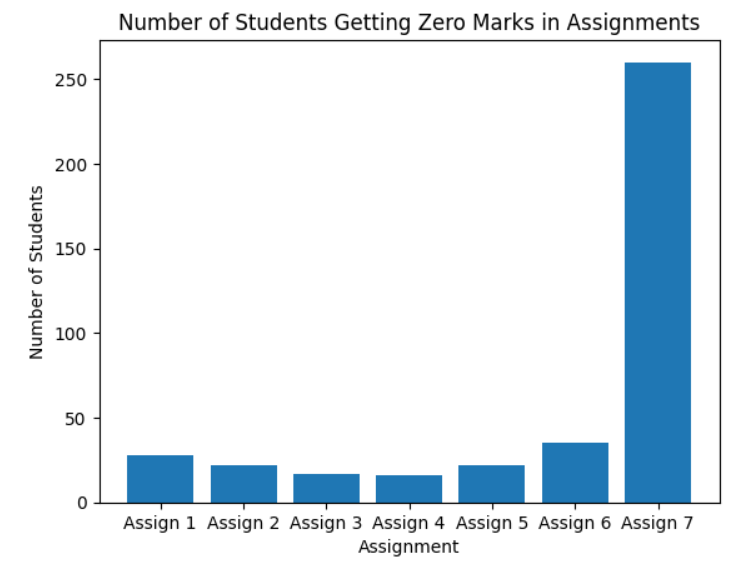
In order to find the basic information regarding each attribute in our dataset we used the “describe” function provided by pandas.



A table with numbers and letters

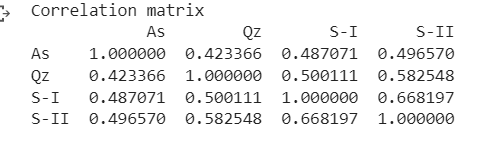
Description automatically generated with low confidence

Next on we checked which assignments and quizzes contained most students with zero marks.

A picture containing text, screenshot, font, diagram

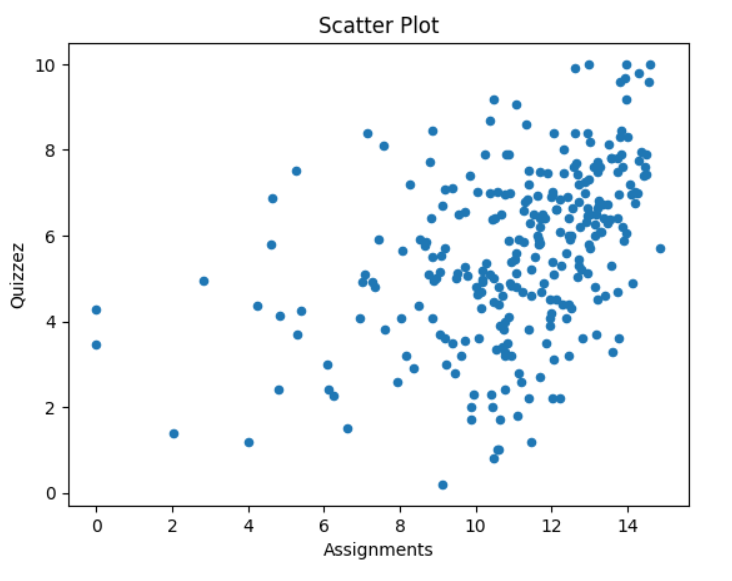
Description automatically generated

To get an idea of how the assignments, quizzes and exams are related to each other we used the correlation matrix.



The results show that all of them are positively correlated, which means if a student is doing good in assignments and quizzes it is more likely that they will be successful in exams.

We then through the scatter plot understood that it is not necessary that high-scoring students in assignments score high in quizzes.



# Model Training and Testing

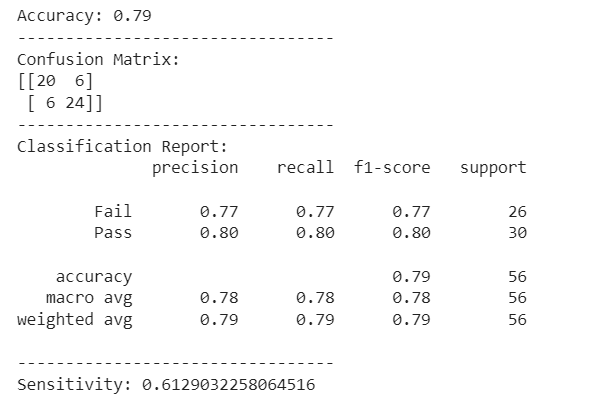
For all the models we used 80% of the data for training and the remaining 20% for testing. We also extracted the basic measuring values for each model.

## Before Mid II

1. **KNN:**

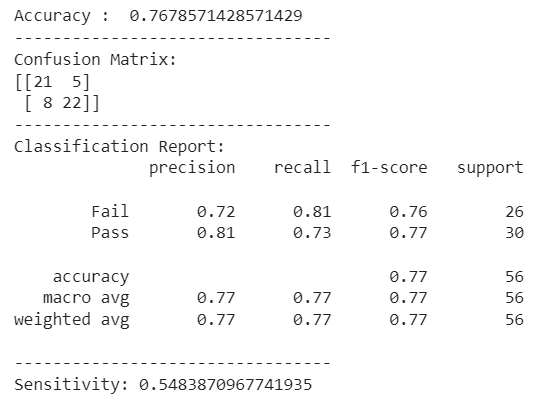
KNN is one of the most common models used in industrial problems. We used the first four assignments and quizzes, and the Mid I exams as features. The ‘Grade’ column served as the target.

Our model was giving the following figures when we applied the measuring metrics to our model.



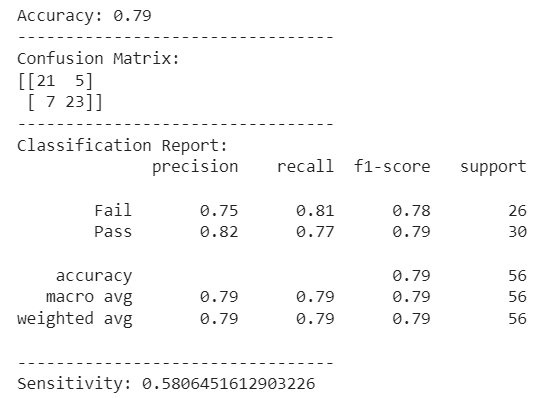
1. **Decision Tree:**

Our Decision Tree Classifier was giving the following figures when we applied the measuring metrics to it.



1. **Naïve Bayes**

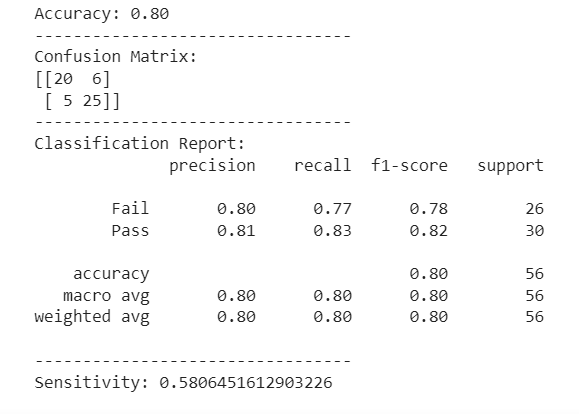
Our Naïve Bayes model was giving the following figures when we applied the measuring metrics to it.



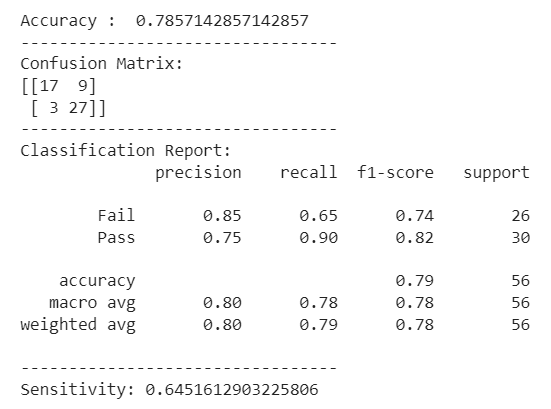
## After Mid II

Before training model we had to preprocess the data further. As per our problem we need to select the best 5 assignments and quizzes. So we did that developed a new data frame for this part of project.

1. **KNN**



1. **Decision Tree**



1. **Naïve Bayes**

