Data Mining Project

M.Wahaj Tahir
F191014@cfd.nu.edu.pk
wahajt@acm.org

April,2023

National University of Computer
and Emerging Sciences

Department of Computer Science
National University of Computer and Emerging Science
Chiniot Faisalabad Campus, Pakistan 2023

Contents

1 Exploratory Data Analysis

The dataset which is provided to us is in excel format and it consist of 7 sheets in it.I'm assuming these different sheets are representing different courses and different student are enrolled in these courses. I have loaded these sheets in different data frames and these are the common data dictionary(columns) I have found.

1.1 Data Dictionary

A data dictionary is a type of dictionary that tells us what type of data is describing itself.

Table 1: Data Dictionary

Dictionary	Data Type	**Description**
As:1,..,6	Float64	The marks of assignments of each student in a particular assignment.
As	Float64	The total number obtained from all assignments.
Qz:1,..,6	Float64	The marks of quizzes of each student in a particular Quiz.
Qz	Float64	The total number obtained from all quizzes.
S-I	Float64	The exam marks of individual from total.
S-II	Float64	The Mid-exam marks of individual from total
Grade	Object	Either the student is pass or fail in that particular subject

The given table is telling us what each column is telling about the data.

1.2 Data Pre-processing

I loaded these files **jupyter note book** and start working on it and check ambiguities in the Excel file and did some pre-processing on it and found some errors. Which can be found in the given code.

After doing this, I tried to understand the data and found **Mean**,**Median**,**Mode**,**Min**,**Max**,**total counts etc**, after that I found out the total number of null values. By looking at the data there are not many null values that can affect the model training.I thought of changing the **NAN** value with zero. After that, I did some drawn histograms of each sheet and displayed the data which is giving the information about each course.

1.3 Data Visualization:

The data is Visualized.

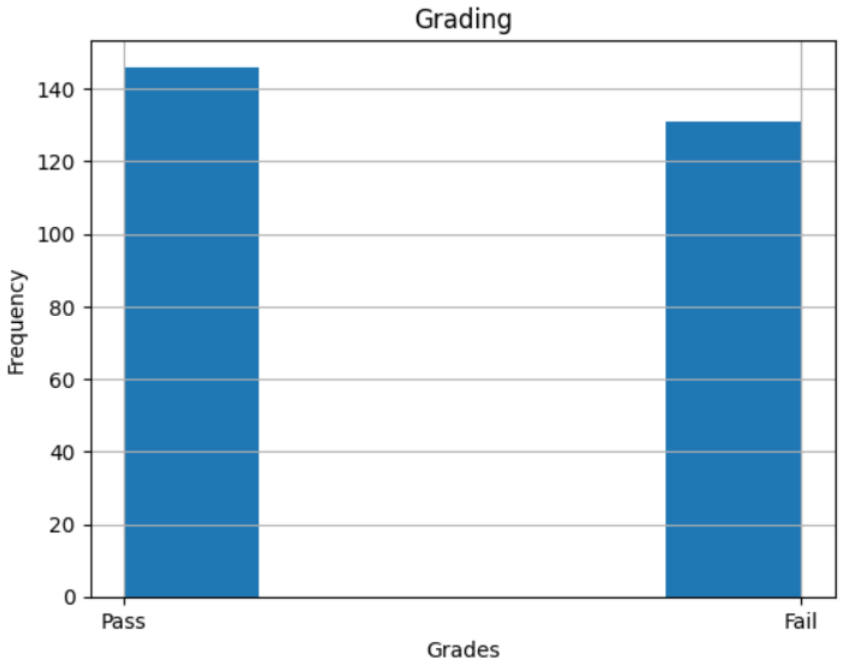

Figure 1: Grading

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

	As:1	As:2	As:3	As:4	As:5	As:6	As:7	As	Qz:1	Qz:2	Qz:3	Qz:4	Qz:5	Qz:6	Qz:7
count	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000
mean	66.122744	59.395307	84.624549	54.085199	68.983755	52.782491	2.274368	11.075776	5.548736	6.768953	3.392419	3.424188	2.731047	4.287004	3.090253
std	33.125138	25.880471	34.171322	31.752431	34.259463	27.618213	9.206989	2.520804	4.255258	6.403467	3.489113	3.830374	2.741729	7.135194	3.919823
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	44.500000	46.000000	68.000000	34.000000	46.000000	32.500000	0.000000	10.040000	2.500000	2.000000	0.750000	0.000000	0.000000	0.000000	0.000000
50%	73.000000	61.000000	90.000000	49.000000	75.000000	58.000000	0.000000	11.460000	4.500000	5.000000	2.000000	2.000000	2.000000	2.000000	2.000000
75%	93.000000	79.000000	108.000000	70.000000	95.000000	75.000000	0.000000	12.930000	7.000000	10.000000	5.200000	7.000000	5.000000	5.000000	4.000000
max	127.000000	100.000000	140.000000	130.000000	120.000000	90.000000	44.000000	14.870000	20.000000	30.500000	17.500000	10.000000	10.000000	35.000000	19.000000

Figure 2: Data Description

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

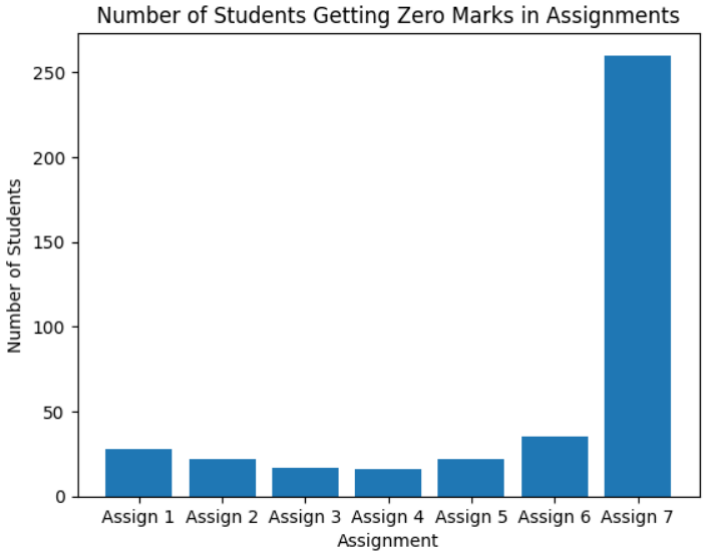

Figure 3: Assignment

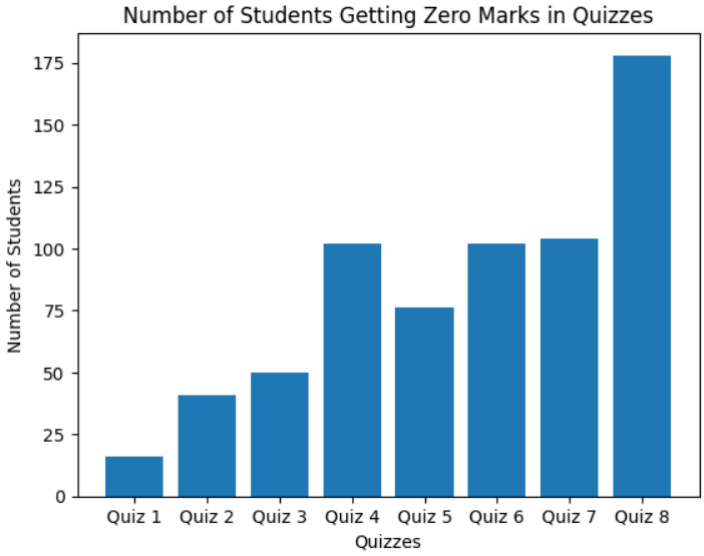

Figure 4: Quizes

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

```
Correlation matrix
             As        Qz       S-I      S-II
As     1.000000  0.423366  0.487071  0.496570
Qz     0.423366  1.000000  0.500111  0.582548
S-I    0.487071  0.500111  1.000000  0.668197
S-II   0.496570  0.582548  0.668197  1.000000
```

Figure 5: Correlation matrix

is doing good in assignments and quizzes it is more likely that they will be successful in exams.

1.4 Modeling 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.

1.5 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.

```
Accuracy: 0.79
---------------------------------
Confusion Matrix:
[[20  6]
 [ 6 24]]
---------------------------------
Classification Report:
              precision    recall  f1-score   support

        Fail       0.77      0.77      0.77        26
        Pass       0.80      0.80      0.80        30

    accuracy                           0.79        56
   macro avg       0.78      0.78      0.78        56
weighted avg       0.79      0.79      0.79        56

---------------------------------
Sensitivity: 0.6129032258064516
```

Figure 6: KNN

1.6 Decision Tree

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

```
Accuracy :  0.7678571428571429
---------------------------------
Confusion Matrix:
[[21  5]
 [ 8 22]]
---------------------------------
Classification Report:
              precision    recall  f1-score   support

        Fail       0.72      0.81      0.76        26
        Pass       0.81      0.73      0.77        30

    accuracy                           0.77        56
   macro avg       0.77      0.77      0.77        56
weighted avg       0.77      0.77      0.77        56

---------------------------------
Sensitivity: 0.5483870967741935
```

Figure 7: Decision tree

1.7 Naive Bayes

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

```
Accuracy: 0.79
---------------------------------
Confusion Matrix:
[[21  5]
 [ 7 23]]
---------------------------------
Classification Report:
              precision    recall  f1-score   support

        Fail       0.75      0.81      0.78        26
        Pass       0.82      0.77      0.79        30

    accuracy                           0.79        56
   macro avg       0.79      0.79      0.79        56
weighted avg       0.79      0.79      0.79        56

---------------------------------
Sensitivity: 0.5806451612903226
```

Figure 8: Naive Bayes

2 After Results

2.1 KNN

```
Accuracy: 0.80
--------------------------------
Confusion Matrix:
[[20  6]
 [ 5 25]]
--------------------------------
Classification Report:
              precision    recall  f1-score   support

        Fail       0.80      0.77      0.78        26
        Pass       0.81      0.83      0.82        30

    accuracy                           0.80        56
   macro avg       0.80      0.80      0.80        56
weighted avg       0.80      0.80      0.80        56

--------------------------------
Sensitivity: 0.5806451612903226
```

Figure 9: KNN

2.2 Decision Tree

```
Accuracy :  0.7857142857142857
--------------------------------
Confusion Matrix:
[[17  9]
 [ 3 27]]
--------------------------------
Classification Report:
              precision    recall  f1-score   support

        Fail       0.85      0.65      0.74        26
        Pass       0.75      0.90      0.82        30

    accuracy                           0.79        56
   macro avg       0.80      0.78      0.78        56
weighted avg       0.80      0.79      0.78        56

--------------------------------
Sensitivity: 0.6451612903225806
```

Figure 10: Decision tree

2.3 Naive Bayes

```
Accuracy: 0.82
--------------------------------
Confusion Matrix:
[[21  5]
 [ 5 25]]
--------------------------------
Classification Report:
              precision    recall  f1-score   support

        Fail       0.81      0.81      0.81        26
        Pass       0.83      0.83      0.83        30

    accuracy                           0.82        56
   macro avg       0.82      0.82      0.82        56
weighted avg       0.82      0.82      0.82        56

--------------------------------
Sensitivity: 0.5806451612903226
```

Figure 11: Naive Bayes