Data Mining Project

PART 1: EDA ON GIVEN DATASET

Group Member Roll Number	Group Member Name
19F-0916	Muhammad Abdullah
19F-0917	Zainab Ijaz
19F-0965	Mahnoor
19F-0908	Maria Fatima

Section= SE(8A)

MA'AM NASREEN AKHTAR

Introduction:

This document contains the pre-processing phase (part 1) of the Data Mining Project, which contains Python code and data analysis. We've applied different data mining techniques to normalize the data and make it ready for phase-2. We have performed EDA on the given dataset and tried to extract important information from it.

Problem:

To predict students' grades as "pass" or "fail" before: (a) Mid-II, and, (b) Final exams. For Mid-II grade prediction, use the following features: first four assignments, first four quizzes, and Mid-I score; and, for grade prediction before the final exam, use all the features (take the best 5 assignments and quizzes).

Summary of Dataset:

The dataset contains students' assessment scores including <Assignments, Quizzes, Mid- I, Mid-II>, and a predictor variable <Grade>. The data has been anonymized to hide the identities of the students and course(s). The data is shared on seven sheets **(D1 to D7),** where each sheet contains a different number of assignments and quizzes. However, only the best 5 assignments and quizzes are included for each student before calculating their grades. Also, note that total marks for assignments and quizzes are given on the top along with their corresponding weights.

Tools Used:

Tools	Purpose
Python	Data normalization using Code.
Orange	Data normalization using Software
MS Word	For generating report

Working:

We've tried to perform the following tasks to achieve phase 1:

Importing File while considering useful attributes only:

The dataset file starts from 1st row, which shows the name of the columns. 2nd row which shows the weightage of tasks. 3rd column which shows the total marks of the tasks. 4th column is null/free. And main data starts from the 5th column. So, due to this, we've started taking data while neglecting the 2nd, 3rd, and 4th row which contains nothing which can contribute to our analysis.

A5 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	C
1		As:1	As:2	As:3	As:4	As:5	As:6	As	Qz:1	Qz:2	Qz:3	Qz:4	Qz:5	Qz:6	Qz:7	Qz	S-I
2	Weight	3	3	3	3	3	3	15	2	2	2	2	2	2	2	10	
3	Total	60	100	140	80	120	80		10	10	10	10	10	10	2		
4	Sr.#																
5	1	39.5	90	120	80	85	75	13.2	7.5	4.5	4.5	0	1	5		4.5	9
6	2	40	62	93	32.5	75	76	10.57	1.5		0.5	0	1	2		1	3
7	3	42.5	63	120	62	65	50	10.78			1	0	1	0	2	2.4	6
8	4	20.5	42	60	70	70	10	7.94	1	2		0	0		2	2.6	5

Data Duplication:

We have also performed a data duplication analysis to check the redundant data. Because the dataset is on the student's record hence, we don't need to extra check any duplicity. Yet, we need to check for duplicate serial numbers. There were no duplications in the serial numbers and our data was perfectly passed through this phase.

Null Values:

The given dataset has some missing values, some missing assignments marks, and the same for the quizzes too. There were also some missing values available under the Grade column that also needs to be handled.

For missing numerical values, we've opted them to be replaced with the median of that column which also handles the outliers' issues.

For missing categorical values, we've opted them to be replaced with the mode of that column.

Displaying Null Values using Heatmap:

We've used Heatmap to illustrate the missing values inside the dataset and those missing values are available in word file that contains full working.

Handling Null Values:

As mentioned earlier, we've handled these null values in the following way:

For missing numerical values, we've opted them to be replaced with the median of that column which also handles the outliers' issues.

For missing categorical values, we've opted them to be replaced with the mode of that column.

Data Reduction:

In this phase, we performed data reduction on 1st column of the data set as this was the only column that was not playing any role in our dataset. This was just portraying serial numbers of students.

Apart from 1st column, all of the remaining columns were playing important role in the dataset and they were required for further processing of the data.

Correlation among attributes:

After data reduction, our data was ready to step on next level which represents the correlation phase. In this phase, we've tried to check out the correlation of attributes (assignments, quizzes, and sessional) to take note of how they are related to each other.

We've used the Seaborn library to illustrate the correlation of attributes in all sheets.

Relation of attributes with main class:

After performing correlation among attributes, now it was time to relate attributes with the main class. As our dataset is on the student's record, our main class shows binary values, Fail and Pass. We've tried to map our Grade column with assignments, quizzes, and sessional exams to have an abstract view of their relation.

Performing EDA:

After performing all the previous steps, now it was time to move on to EDA. We've performed Univariate EDA on our dataset to extract important information from it, which tells us about the features and their count.

We've only performed it for sheet D1 because the process for the other is the same. It only differs in the results. The following images show the EDA of different attributes:

This analysis shows us the skewness of the attributes depending upon their count and allows us to take steps to normalize the values using data transformation and make it eligible to perform Bivariate EDA on it.

But as we know our data only consists of numerical values hence there is no need to perform Bivariate EDA for further analysis. At this time, our data is clean and ready to proceed in the next phase.

PS: Full Report also contains working with Orange. Please check it out for more details.