Lab Report for CW1 DATA OBSERVATION

[redacted]

***Abstract*—This report aims to classify students from different programmes. Data Observation and Dimensionality Reduction are throughout the whole process and several techniques have been applied to output better visualization, including PCA (Principal Component Analysis), t-SNE (t-Distributed Stochastic Neighbor Embedding) and LDA (Linear Discriminant Analysis). The experiment results show that 'Programme3' has been extracted while other programmes are still mixed up. Further experiments should be taken in the latter coursework and attempt to design a more effective classifier.**

***Index Terms*—Data Observation, Dimensionality Reduction, LDA, PCA, t-SNE**

I. INTRODUCTION

THE data to process comes from INT104 final exam in 2023. It is in 'excel' format and contains 11 columns in total to show specific identification and grade of each student. Specifically speaking, students learning this course vary in 4 different programs. The aim of this report is to extract the potential relations between features, and design a classifier to determine which programme the student comes from in visualized graphics. Several techniques have been applied on the data including PCA, t-SNE and LDA to achieve the goal. For your information, this report should be considered as a primary observation and does not represent the definite result of the data processing. Further experiments will be done to demonstrate better data separation in latter coursework 2 and 3.

II. DATA OBSERVATION

To prepare for the latter dimensionality reduction, several necessary data visualizations are performed as follows.

A. Overall Observation

First of all, the feature 'Index' has been removed from the raw data, since it does not represent any meaningful information. The resulting data can be named as 'numerical data'. If generating Box Plot on the numerical data directly, the poor results as shown in Fig. 1. are predictable, since the data from different features is not in the similar range of values. The box of several features will become victims for better presentation of other features. Several solutions are performed to address this issue, including performing Log Scale, Z-score Normalization and Min-Max Score on the numerical data. The Min-Max Score graph is shown in Fig. 2. as an example.
For the outliers captured in the figure, not all of them represent mistakes in the data set. Taking 'Q1' for instance, when majority of the scores has reached 4 or even higher, 0 and 2 score will be considered as outliers. However, such data is actual collections from students from last year, and should not be removed entirely. The feature 'Grade' can be explained in a similar way, since it is composed of a large part of 2 score and a small part of 3 score.

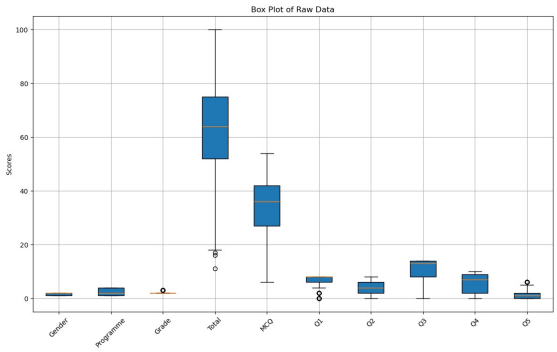

Fig. 1. Box Plot of Raw Data

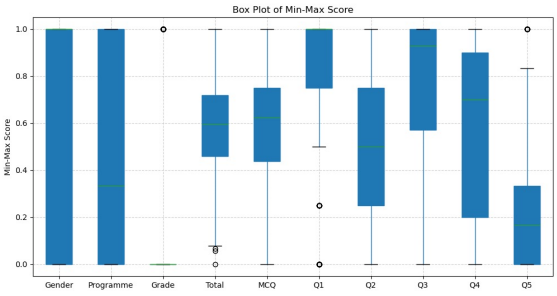

Fig. 2. Box-Plot of Min-Max Score

B. 'Total' and 'Programme'

With the final aim of extracting 'Programme', perform some visualizations as attempts to find potential relations. One of the attempts is on 'Total' and 'Programme' and the ideal result is to represent some linear relation, which will grealy benefit the extracting work. The Scatter Plot of different programmes and the Line Chart of the mean of total grades are shown in Fig. 3. and the distribution is not clear due to the overlap of scatter points. A new visualization called Violin Plot is performed as well, but no obvious linear relation has been captured.

III. EXTRACT 'PROGRAMME3'

This section includes different ways to successfully extract 'Prorgamme3' from the standardized data Moreover, some possible reasons are offered to explain why only 'Programme3' has been extracted.

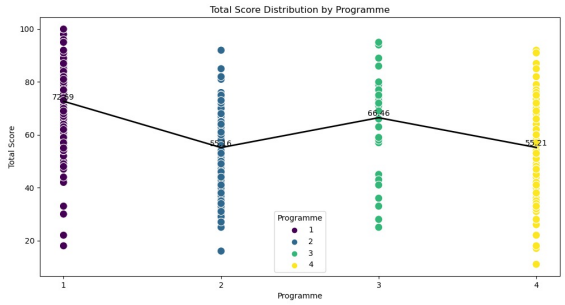

Fig. 3. 'Total' Score Distribution by 'Programme'

A. PCA

The first step is to remove 'Index' and 'Programme' from raw data. The reason to remove 'Index' has been explained earlier, and since the aim is to classify different programmes, there is no reason to carry the standard answer and keep 'Programme' in the data set. After the removal, perform Z-score Normalization and name it as 'standardized data'. Hereby the data pre-processing has been done.

Now perform PCA on the standardized data. Since 2 features has been removed and there are 9 features left, the result should contain 9 principal components, named from 'PC0' to 'PC8'. Before visualization, which two components to choose should be determined. By calculating the correlations between 'Programme' and each principal component, 'PC0' and 'PC1' are mostly correlated to 'Programme'. Consider 'PC0' and 'PC1' as horizontal and vertical coordinates respectively, generate the Scatter Plot to show the distribution of 'Programme', as shown in Fig. 4.. It is clearly demonstrated that 'Programme3' is mostly distributed above other 3 programmes, all of which are mixed up below.

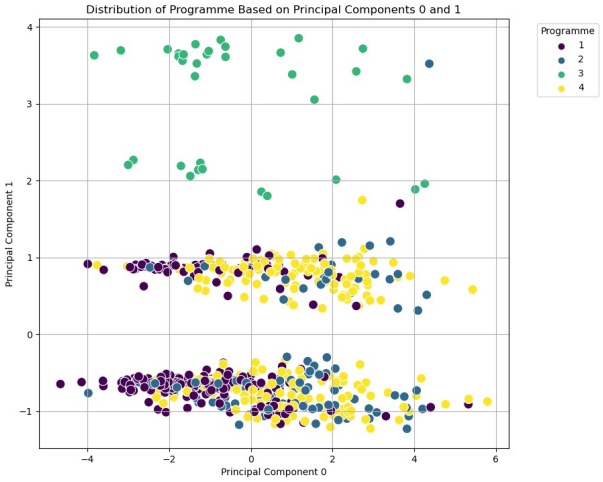

Fig. 4. Distribution of 'Programme' Based on 'PC0' and 'PC1'

B. t-SNE

Another way to visualize high-dimensional data by dimensionality reduction is t-SNE, and it can also be applied to extract 'Programme3'. Performing t-SNE requires the developer to choose several features as input data to reduce to 2 dimensions. To determine such set of features, the correlation between other features and 'Programme' has been calculated. The top features with highest correlations are 'Total', 'MCQ', 'Q2', 'Q4' and so on. The output of the top 4 features fails to classify any programme separately. However, if adding 'Grade' to the input feature, the output shown in Fig. 5. has successfully extract 'Programme3'.

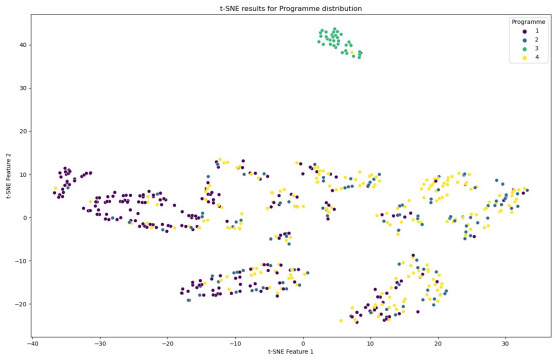

Fig. 5. t-SNE result for 'Grade', 'Total', 'MCQ', 'Q2', 'Q4'

C. Potential Explanation

Go back to the raw data, some characteristic of 'Grade' can be captured through simple observation. Comparing 'Grade' with 'Programme', it is implied that students who belong to 'Programme3' have all got 'Grade3', and thereby 'Grade' can be a good candidate to extract 'Programme3' effectively. However, not all students who got 'Grade3' are assigned to 'Programme3', which is why a small amount of scatter points of other programme also appear in the group of 'Programme3'.

IV. Further Extraction

After extracting 'Programme3', the next step is to extract the other 3 programmes from each other. Firstly, some adjustment on the data set should be done, including removing columns 'Index', 'Programme', and students in 'Programme3' from the raw data and performing Z-score Normalization one more time. Name the resulting data 'adjusted data'.

A. PCA

Perform similar steps stated above on adjusted data. The mostly correlated principal components are 'PC3' and 'PC6'. However, the resulting Scatter Plot fails to separate different programmes. Other methods should be done to achieve this goal.

B. t-SNE

1) Direct t-SNE: Perform t-SNE on adjusted data directly. Take similar steps and choose top correlated features as input. Though after countless attempts to adjust different perplexity and learning rate, which are two important coefficients, the resulting graph is still not ideal. Since t-SNE might have bad performance if the input data is of high dimensions and simply choosing less features will lead to loss of important information, consider performing PCA first to reduce dimensionality, then apply t-SNE on the chosen principle components.

2) PCA and t-SNE: First of all, perform PCA on adjusted data and calculate the explained variance ratio of each principal component. As result shows, 'PC0' and 'PC1' have explained 88.59% and 6.19% of the variance respectively, which in total has reached up to nearly 95.00% and can cover most information of the adjusted data. The resulting Scatter Plot of dimensionality reduction is shown in Fig. 6..
The next step is to perform t-SNE on the 2-dimensional data. The result shown in Fig. 7. is better than direct t-SNE, but still not a good enough representation in respect to classification.

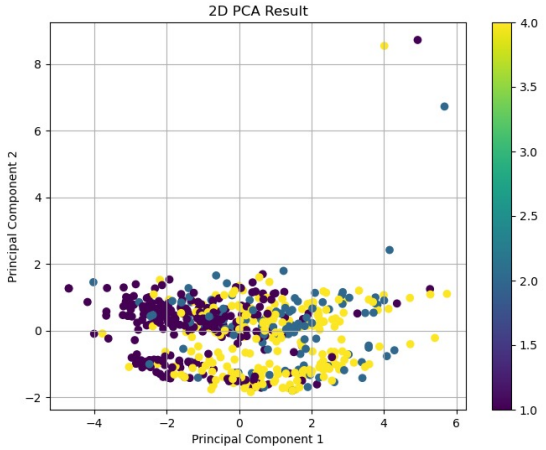

Fig. 6. 2D PCA Result

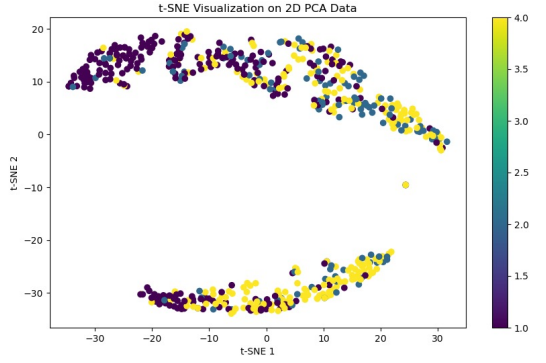

Fig. 7. t-SNE Performed on 2D PCA Result

C. LDA

1) Direct LDA: LDA aims to find a linear combination of features to separate data from different classes. Perform LDA on adjusted data directly, and the result does not show clear classification.

2) PCA and LDA: Similarly, perform PCA first to reduce dimensionality, and perform LDA on the 2D data. The output graph shown in Fig. 8. still fails to fully classify different programmes.

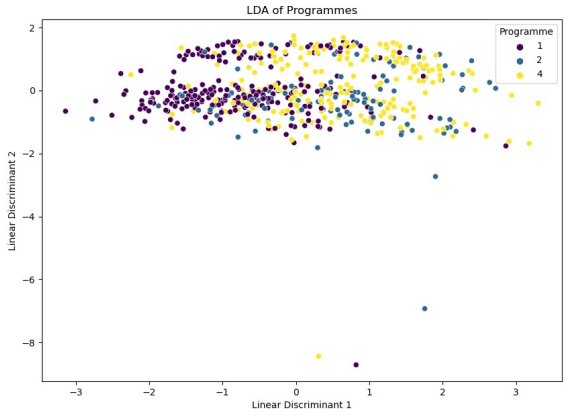

Fig. 8. LDA Performed on 2D PCA Result

D. Experiment on 'Grade'

Since 'Programme3' has already been extracted, it is normal to wonder if the feature 'Grade' still has its contribution to the classification of the other 3 programmes. Some experiments have been performed by removing 'Grade' from adjusted data, and perform PCA, t-SNE and LDA similarly. However, all of the three output graph have much less classified the different programs, in other words, the programmes are even more mixed up. It can be confirmed that the column 'Grade' should not be removed entirely.

V. Result

The data observation gives a general overview and attempts have been made to find linear relations. Due to the strong correlation between 'Grade3' and 'Programme3', 'Programme3' has been successfully classified. Several methods to process linear data have been performed including PCA and LDA. There are also attempts to perform methods for non-linear data, such as t-SNE. However, these 3 methods mentioned above fail to fully classify the left 3 programmes.

VI. Recommendation

This data of final exam grade from last year does not seem to show a strong linearity itself, and more non-linear techniques should be done to represent better classification, including Support Vector Machine, Random Forest and Neural Network.