INT104 CW1 Lab Report

Mingxuan Hu 2252534
TA: Zhejun Yang

***Abstract*—This lab report documents an experiment that used machine learning methods to observe, process, analyze, and visualize a multidimensional dataset containing student score information, with the aid of Python. Firstly, after thorough data preprocessing, various techniques were used for data observation in the experiment, including bar charts, box plots, and scatter plots. These figures show the distribution of data before and after normalization. Then, the experiment achieved dimensionality reduction through principal component analysis (PCA), reducing the dataset from high-dimensional to four-dimensional. By selecting the most important principal components, students from different programmes were simply classified. The experiment also used t-SNE technology to observe similar scatter distribution patterns, validating the results of PCA. Finally, the experiment determined the contribution of each feature and extracted the main features by conducting correlation analysis and calculating the weight sum of each original feature in the principal component. The research results indicate that features such as Total, MCQ, Q2, and Q4 have significant contributions to principal components. And the above related features were visualized.**

***Index Terms*—Box Plot, Normalization, Normal Distribution, PCA, t-SNE, Correlation Heatmap, Python**

I. Data Preprocessing

In order to improve the quality of data, reduce errors and inconsistencies, and prepare for further analysis and processing, data preprocessing is necessary. In the data munging stage, a spreadsheet containing eleven columns of data is imported into Python. Next, Python checked the spreadsheet and found no missing values. In addition, all data in the table are within the relevant range (for example, all MCQ values should be within the range of 0-54). Moreover, considering the actual situation, different students have different levels of mastery of knowledge, so it is reasonable to have particularly high or low scores, and there is no need to remove outliers. Above, the work of data cleaning has been completed.

II. Data Observation

After completing data preprocessing, the next step is to observe the original features of the data. The given spreadsheet contains eleven dimensions of raw features, where the student's Index, Gender, Programme and Grade belong to qualitative data, while the remaining seven dimensions related to scores belong to quantitative data. A bar chart can be used to simply observe the properties of the raw data, such as mean and standard deviation, as shown in Figure 1.

However, bar charts are mainly used to display the count of data and are not intuitive for the distribution of data. A better alternative is the box plot. A box plot can display the distribution of data, including the median, upper and lower quartiles, outliers, etc., which helps to better understand

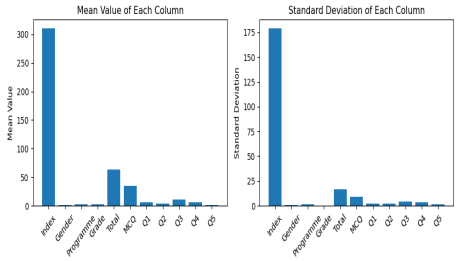

Fig. 1. Mean Value and Standard Deviation of Each Column

the overall situation of the data. Considering that the Index column only represents the student's serial number without any statistical significance, and the Programme column is the final classification standard, Figure 2 shows a box plot of the remaining nine features.

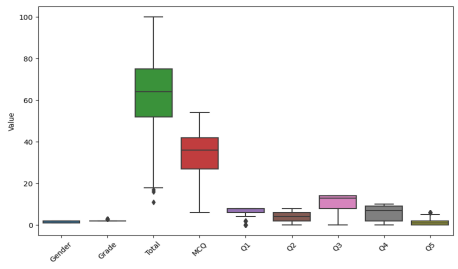

Fig. 2. Boxplot of Selected Columns

Nonetheless, due to the different scales or ranges of each feature, except for the Total and MCQ columns, all other features are compressed within a very small range, which is not conducive to data observation. Therefore, in order to better observe all data while preserving its features, data can be normalized. Here, the data will be subjected to Min-max normalization.

The following figures can clearly demonstrate the effect of normalization. Taking the Total column as an example, Figures 3 and 4 show the distribution of data before and after normalization using scatter plots and bar charts, respectively. Figure 3 shows that normalization reduces the scale of the data, making the distribution of the data denser. Figure 4 shows that after normalization, the features of the data (represented by a normal distribution in this example) are still preserved. Therefore, Figure 5 redrawn the box plots of those nine features using normalized data.

Due to the fact that the Gender and Grade columns are qualitative labeled data, there is no need to analyze the

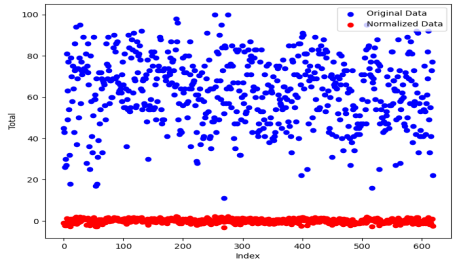

Fig. 3. Scatter Plot Before and After Normalization

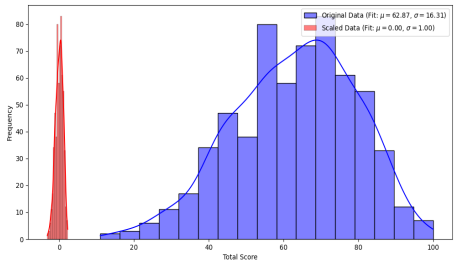

Fig. 4. Normal Distribution of Total Scores Before and After Normalization

corresponding box plot. The remaining seven box plots all display the key mathematical statistics corresponding to the features, such as quartiles or median. In addition, the last observation is to determine whether the data conforms to a normal distribution. As shown in Figure 6, the Total, MCQ, and Q2 columns follow a normal distribution, while the rest of the data is non normally distributed, which brings more difficulties for further data processing in the future.

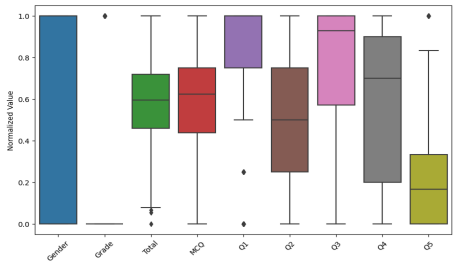

Fig. 5. Boxplot of Normalized Data

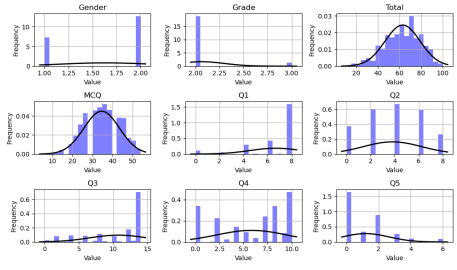

Fig. 6. Analysis of Data Normality Using Bar Charts

III. Dimensionality Reduction

Generally speaking, a dataset may contain a large number of features (including the dataset provided in this experiment). But some features may be redundant or unrelated, which will reduce the model's generalization ability. Therefore, in order to extract the most important and relevant features from complex data, dimensionality reduction is necessary to improve the accuracy of the model. Firstly, the dataset should undergo Z-score normalization.

PCA maps the original high-dimensional array to a low dimensional space with linear variation for better visualization. Specifically, the covariance matrix is first used to calculate standardized data, which describes the correlation between various dimensions. Then, the covariance matrix is decomposed using eigenvalues to obtain its eigenvectors and eigenvalues. Then, sort the feature vectors in descending order based on their corresponding eigenvalues, and select the first k feature vectors as the main components. Finally, by linearly transforming the selected principal components, the original data is mapped to a new low dimensional space.

In dimensionality reduction, PCA uses sample variance as an indicator to measure the amount of information, also known as explained variance. The greater the change, the more information the feature possesses.

When selecting the first k principal components, a threshold of 70% to 90% can usually be selected, so that the proportion of the total explained variance of the first k principal components to the total variance reaches the preset threshold. Usually, the appropriate dimensionality can be selected by plotting a curve of cumulative explanatory variance. In this experiment, the relevant data is shown in Figure 7:

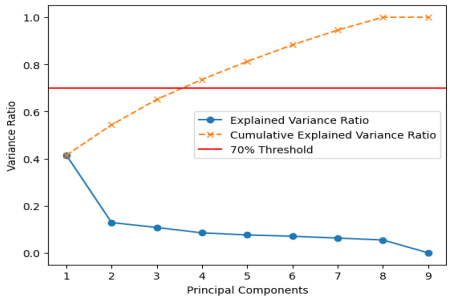

Fig. 7. Explained Variance Ratio and Cumulative Explained Variance Ratio

Figure 7 shows the explained variance ratio of each principal component and the cumulative explained variance ratio. In addition, 70% is set as a threshold to determine how many principal components should be selected. As shown in the figure, the four principal components with the highest contribution can be selected.

Therefore, PCA can reduce the original high-dimensional dataset of this experiment to four dimensions, and select the first and second principal components to visualize the dimensionality reduction results in a two-dimensional plane. The visualized results are shown in Figure 8:

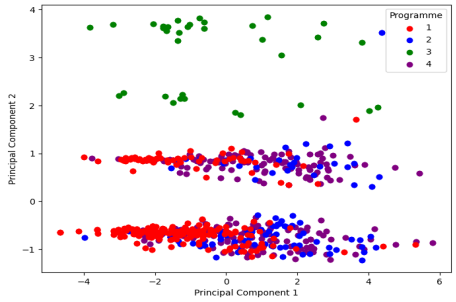

Fig. 8. PCA Plot

In Figure 8, the green dots (i.e. Programme 3) are well separated. Most of the red dots (i.e. Programme 1) also have a clear distinction. However, a small number of red dots are intertwined with blue dots (i.e. Programme 2) and purple dots (i.e. Programme 4). The insufficiency of the experiment's dataset size or the existence of overlapping or non-linearly separable data might be the reasons why PCA failed to effectively separate all points. In subsequent experiments, more classifiers will be attempted to find a better solution for classification.

In addition, another visualization method called t-Distributed Stochastic Neighbor Embedding (t-SNE) can be used for cross validation with previous PCA results. t-SNE maps samples into a two-dimensional graph to reflect the distance between sample pairs. By performing t-SNE dimensionality reduction on the same data, the results shown in Figure 9 can be obtained:

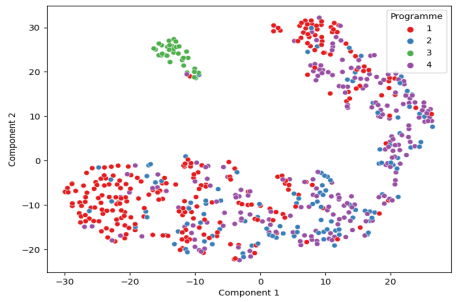

Fig. 9. t-SNE Result

It can be observed that the distribution of different Programme points in Figure 9 is almost consistent with the previous PCA results. The above completes the dimensionality reduction work.

IV. Feature Extraction and Analysis

After performing PCA dimensionality reduction, 9 principal components were generated from linear combinations of original features. Below, an analysis will be conducted to determine which original features play a greater and more important role in principal components.

Firstly, correlation analysis will be conducted. Pearson correlation serves as a measure of linear correlation between two variables. In this case, for this dataset, Pearson correlation can be utilized to assess the correlation of the data.

Figure 10 is the corresponding correlation heatmap. A correlation heatmap is a chart used to visualize the correlation between variables. Generally, red represents positive correlation, blue represents negative correlation, and the depth of the color represents the strength of the correlation. Figure 10 illustrates that the correlation between each pair of features is not particularly strong. Relatively speaking, the columns that are highly correlated with the Programme are Total, MCQ, Q4, and Q2.

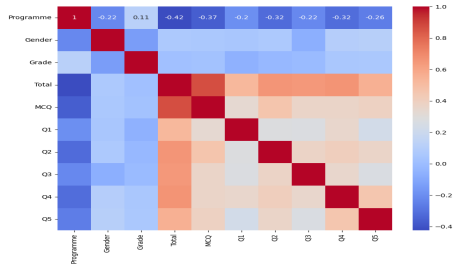

Fig. 10. Heatmap of Feature Correlation

In addition, by calculating the sum of weights of each original feature among the top N principal components, we can determine which features have the greatest contribution among the top N principal components.

N can be taken as 1, because according to Figure 7 shown earlier, the explained variance ratio of the first principal component is higher than that of the other principal components, reaching over 40%. By visualizing the corresponding comprehensive contribution values, we obtained Figure 11. Figure 11 shows that Total, MCQ, Q2, and Q4 have the highest contribution values to the first principal component, which is consistent with the conclusion drawn from the correlation heatmap.

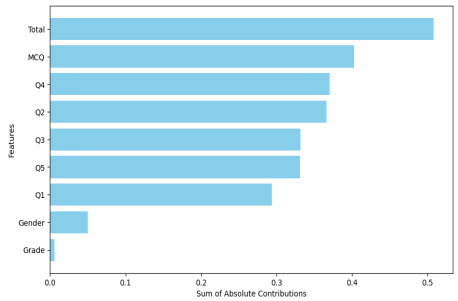

Fig. 11. Top 1 Principal Components Feature Contributions in PCA

Of course, N can also be taken as 4, as PCA ultimately reduces the dimensionality of the data to 4 dimensions. N can also be taken as 2, as the results of PCA are visualized in a two-dimensional plane. Different values of N will lead to different conclusions, which will not be elaborated here. The above completes the part of feature extraction and analysis, with relevant features being visualized and compared.