CW1 Lab Report

Name: Junhao Huang Student ID: 2256793 TA: Yu Kang

I. Introduction

In this report, machine learning techniques are used to create a table of final exam information for the ICS program based on the information stored in the student performance information file provided. The table contains personally identifiable information of 619 students who participated in the final exam and the corresponding final exam grade information. In this paper, we will record a series of algorithms to validate the valid data in the data set and carry out data cleaning and preprocessing, and then filter the suitable feature set for data analysis according to the correlation between the data, and then downsize the data in the feature data set by using the PCA dimensional reduction algorithm to convert the high-dimensional data into a low-dimensional space, in order to obtain the visualization image that can reflect the distribution of the information about the students' grades as much as possible. Distribution of student achievement information as much as possible to get the visualization image.

II. Data Analysis

A. Exploratory Data Analysis

First of all, after the retrieval of the table, it can be seen that there is no duplication or missing content in the data in the table. So the data as a whole is clean and does not contain useless columns or rows. The original test scores can be analyzed according to the table below. With respect to variability, the standard deviation indicates the degree to which scores differ, and the variability of the total score and MCQ is relatively high compared to the individual question scores. This scale change can skew the analysis, resulting in certain features having a disproportionate impact on the overall analysis. To mitigate this, subsequent consideration must be given to normalizing or standardizing the data to ensure that each feature contributes equally to subsequent analysis.

TABLE I: Summary Statistics of Exam Scores

Feature	Max Value	Min Value	Average Value
Total	100	11	62.87
MCQ	54	6	34.58
Q1	8	0	6.67
Q2	8	0	3.82
Q3	14	0	10.76
Q4	10	0	5.84
Q5	6	0	1.19

B. HeatMap

In order to gain a deeper understanding of the relationship between the elements. Combining the performance data in the table, the distribution of students' performance can be seen through the generated Fig. 1 to satisfy the Gaussian distribution, which is also known as the normal distribution.

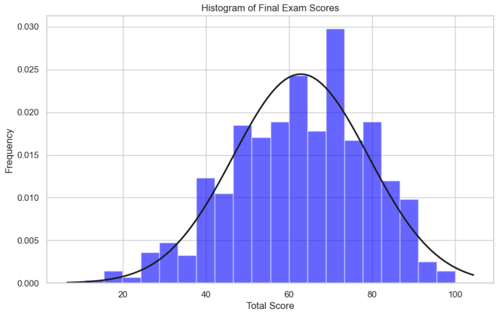
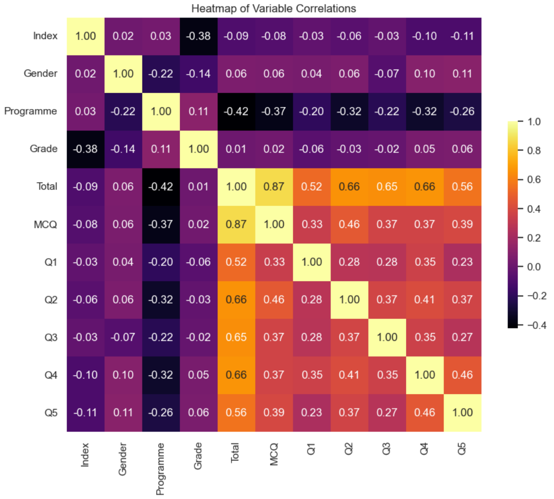


Fig. 1: Take the Total feature in the data table to check whether the scores distribution of students in the final exam conforms to the Gaussian distribution

When the Gaussian distribution is satisfied the Pearson algorithm can be used to find the correlation between the features, and the results can be generated by using the heatmap method in the seaborn library to generate a heat map for a clearer display of the correlation between the data.



	Index	Gender	Programme	Grade	Total	MCQ	Q1	Q2	Q3	Q4	Q5
Index	1.00	0.02	0.03	-0.38	-0.09	-0.08	-0.03	-0.06	-0.03	-0.10	-0.11
Gender	0.02	1.00	-0.22	-0.14	0.06	0.06	0.04	0.06	-0.07	0.10	0.11
Programme	0.03	-0.22	1.00	0.11	-0.42	-0.37	-0.20	-0.32	-0.22	-0.32	-0.26
Grade	-0.38	-0.14	0.11	1.00	0.01	0.02	-0.06	-0.03	-0.02	0.05	0.06
Total	-0.09	0.06	-0.42	0.01	1.00	0.87	0.52	0.66	0.65	0.66	0.56
MCQ	-0.08	0.06	-0.37	0.02	0.87	1.00	0.33	0.46	0.37	0.37	0.39
Q1	-0.03	0.04	-0.20	-0.06	0.52	0.33	1.00	0.28	0.28	0.35	0.23
Q2	-0.06	0.06	-0.32	-0.03	0.66	0.46	0.28	1.00	0.37	0.41	0.37
Q3	-0.03	-0.07	-0.22	-0.02	0.65	0.37	0.28	0.37	1.00	0.35	0.27
Q4	-0.10	0.10	-0.32	0.05	0.66	0.37	0.35	0.41	0.35	1.00	0.46
Q5	-0.11	0.11	-0.26	0.06	0.56	0.39	0.23	0.37	0.27	0.46	1.00

Fig. 2: Heatmap containing all the features, showing the correlation between the features by means of the Pearson algorithm. The warmer the color, the higher the positive correlation, and the colder the color, the higher the negative correlation.

From the Fig. 2, it can be seen that the correlation between Index, Gender and Grade is relatively small and weak for other features, where Index can be seen to have the least correlation with other features, so it can be removed first. Programme is negatively correlated, but negative correlation doesn't mean that it is not correlated, and at the same time, because the coursework uses the Programme features to act as a labeled feature set to divide the data, so Programme features can also be extracted.

C. Random Forest

The final determination of whether Gender and Grade still need to be taken out can be done by using the Random Forest algorithm.

$$z = \frac{(x - \mu)}{\sigma} \tag{1}$$

x is the original data point, μ is the mean of the data set, σ is the standard deviation of the data set, and z is the standardized data point.

This can be done by using the RandomForestClassifier method in the sklearn library to show that the importance of each feature to the result can be assessed by analyzing the predictions produced by the decision tree constructed by the random forest algorithm.

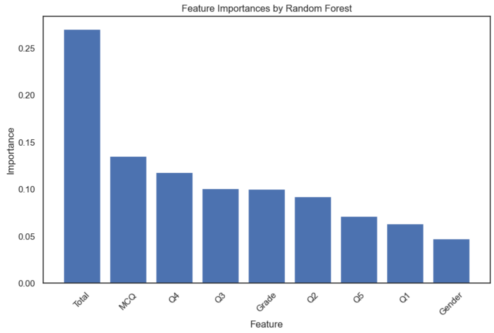


Fig. 3: The remaining 9 sets of features generate a histogram of the magnitude of the impact on the results

The Fig. 3 shows that the influence of Gender is minimized, so the removal of the Gender feature can also be considered.

D. Standardization

Principal component analysis is very sensitive to the scale of the feature data in the calculation process, so before the operation of principal component analysis, it is also necessary to standardize the selected data, by using the StandardScaler method in the sklearn library to standardize the current data, convert the data to a distribution with a mean of 0 and a standard deviation of 1 as the following Fig. 4 shows, from reducing the impact of the data differences on the accuracy of the the accuracy of the next operation.

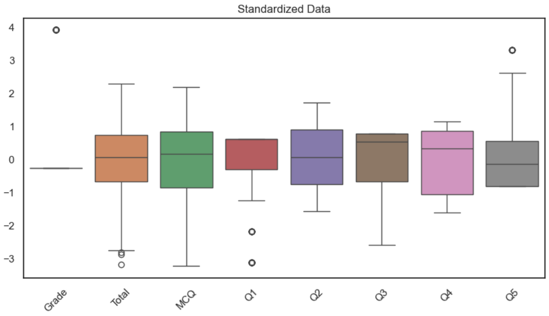


Fig. 4: A boxplot of the results obtained after standardizing the existing 8 sets of feature data

III. Dimensionality Reduction

A. Explained Variance Ratio

Performing PCA dimensional reduction on the remaining feature dataset still retains relatively enough information to show the relationships between the original data.The basic idea of PCA is to reconstruct the n-dimensional feature set into k dimensions to generate new orthogonal features.The PCA operation selects the first principal component axes in the direction where the data has the largest variance, and the second principal component axes can be detected in the planes that are orthogonal to the first principal component axes. These new axes may show that most of the variance is contained in the first k axes, with less variance in the other axes. This is why only the first k axes containing most of the variance are retained as representative features of the data set.

Component	EVR (%)	CVR (%)
1	46.59	46.59
2	12.96	59.54
3	9.59	69.13
4	9.43	78.56
5	8.19	86.76
6	7.06	93.82
7	6.17	99.99
8	0.006	100.00

TABLE II: Explained Variance Ratio(EVR) and Cumulative Variance Ratio (CVR) by PCA Components

As shown in Table 2 for this dataset, based on the variance rates for all 8 axes, k was chosen as 2 because the majority of the variance is in the first and second axes.

B. Cumulative Variance Ratio

The cumulative explained variance ratio is used to show the ratio of the eigenvalues of the final selected principal component to the sum of all eigenvalues. Often if the ratio is higher, it indicates that the corresponding principal component captures more information By calculating the cumulative explained variance ratios of the 8 principal components retained and finally presented on Table 2, the ratio of explained variance ratios is close to 1, implying that after reduce the dimensions of original data from

11 to 8, these principal components have captured most of the information about the data, and thus it can be finally confirmed that the table generated by the PCA performed through the 8 principal components can be chosen as the result of the final analysis.

C. Principal Component Analysis

Calculate the covariance matrix of the data. If the data has n features, the covariance matrix will be an n-by-n matrix, where each element represents the covariance between pairs of response features. Covariance measures how two variables change together. For features X and Y, the covariance is calculated by:

$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \tag{2}$$

where X and Y are the two variables, $\bar{X}$ and $\bar{Y}$ are their means, respectively, and n are the number of observations.

$$Cv = \lambda v \tag{3}$$

where C is the covariance matrix, v is the eigenvector, and λ is the corresponding eigenvalue, representing the data variance in the v direction.

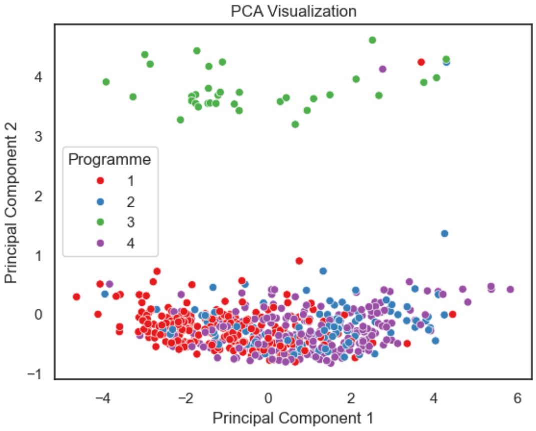


Fig. 5: Principal Component Analysis $k = 2$

After the data preprocessing, the PCA method in sklearn library will be used to perform the computation, and the final result is shown in Fig. 5. The calculation of PCA requires the covariance matrix and the determination of the direction of the data. The covariance matrix will show the correlation between different features, and the largest two sets of eigenvalues can be found by eigen-decomposition of the covariance matrix. The eigenvalues indicate the importance of the corresponding eigenvectors in the representation of principal components, so the corresponding principal components can be found according to the descending order of the eigenvalues, and the data will be mapped to the new coordinate system constructed by the principal components.

D. tSNE

In order to compare with the results obtained from the final PCA verification. Use the same data for tSNE method dimensionality reduction. tSNE is also a dimensionality reduction algorithm, but it is slightly different from PCA. PCA is a linear dimensionality reduction method that relies on the main direction of change of the data to reduce the dimensionality of the data, with a focus on preserving the global structure of the data. PCA, on the other hand, maps the data to a new coordinate system after orthogonal exchange.

t-SNE Target Function:

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{4}$$

The cost function to be minimized, the sum of Kullback-Leibler divergences between p_{ij} which isHigher dimensional space and q_{ij} which is Lower dimensional space

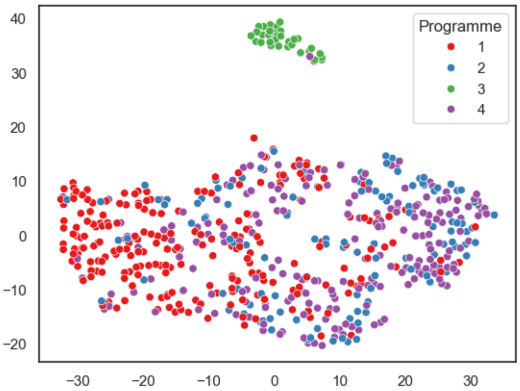


Fig. 6: tSNE is a nonlinear dimensionality reduction technique that maps high-dimensional data to low-dimensional space by preserving the similarity between data points, and pays more attention to preserving the local structure of the data, making it easier to represent clusters or groups in high-dimensional data in low-dimensional space.

As can be seen from Fig. 6, tSNE provides a clearer separation and clustering of the student data for Program 3. However, the distribution of the remaining three majors is more aggregated on the second principal component and also has a certain positional distribution based on grades compared to the resulting image of PCA.

IV. Conclusion

There are two hypothesis based on result images. First, both PCA and tSNE basically separate Programme 3 students completely, and this may be because the vast majority of Programme 3 students are also Grade 3 students, which may be more relevant in themselves. Another piece of information is that the two students with full scores in this exam both appeared in Programme 1, while the lowest scores appeared in Programme 4. The distribution after dimensionality reduction can also roughly show that Programme 1 is distributed on the left and Programme 4 is distributed on the right.