Xi'an Jiaotong-Liverpool University
西交利物浦大學

Xi’an Jiaotong-Liverpool University

School of Advanced Technology

Report on Machine Learning

Author:
Zheyuan Cao

Student ID & Lab-Group:
2141805 & D1/4-C

A Lab Report Submitted For

INT104 Artificial Intelligence

April 30, 2023

1 Introduction

Relevant medical treatments should be applied to different patients according to individual conditions. Therefore, to achieve the target, this coursework consists of three tasks: Dimensionality Reduction, Classifier Construction and Unsupervised Patients Classification. For the first requirement, the transformed data should be generated through the PCA method. In the second phase, the theory of each classifier needs explanation. Results for K-ford cross validation and accuracy metrics of classifiers should also be provided to accomplish the selection process which ends up with the adoption of the SVM. Finally, unsupervised clustering should categorize patients into multiple cohorts on the basis of scores obtained from the questionnaire.

1.1 Data Dimension Reduction

In the original questionnaire, results derived from marks for each question can be divided into two primary labels, namely 0 and 1. However, there also exists an additional result label 2 which can be regarded as noise and require removal. After data cleaning, the remaining data contains scores attributed to questions from F1 to F15, encompassing 15 features in total, which may retard the speed of the training procedure. Accordingly, dimensionality reduction is imperative. In this coursework, principal component analysis(PCA) is employed to guarantee an appropriate data dimension that is exactly 12. Specifically, explained variance ratio, relevance and distribution analysis are implemented to assist the reduction process.

1.2 Classifier Selection

Upon the completion of data pre-processing and dimensionality reduction, the original dataset could be divided into two groups, which results in a binary classification issue [1]. Supervised learning algorithms such as support vector machine(SVM) [2], K-nearest-neighbour(KNN) [3] and decision tree(DT) [4] are reliable approaches that can effectively address binary classification problems [5]. Considering the nonlinear distribution of the transformed dataset, a SVM classifier using an RBF kernel [6] could be promoted. Moreover, Decision Tree(DT) and Naive Bayes(NB) [1,5] are two viable classification models that can classify dimension-reduced data. Finally, Deep Neural Networks(DNN) and Logistic Regression(LR) classifiers are also built in this lab. Following the model training and cross validation(CV) process, it is recommended that SVM is the optimal model for the given dataset. For the next task, K-means is an advisable unsupervised learning algorithm that can iteratively generate k distinct clusters based on similarity metrics [7]. In particular, to identify the optimal value of k, silhouette coefficient [8] is applied to evaluate the performance of algorithm with various values of k.

1.3 Classification Results

After the model fitting procedure, the accuracy of each supervised learning algorithm is presented as follow:

Table 1: Accuracy Score of Classifiers

Classifier	Accuracy of Train Set	Accuracy of Test Set
SVM	76.74%	70.93%
DT	89.06%	63.79%
NB	72.45%	70.23%
DNN	77.16%	69.45%
LR	73.30%	70.42%

To visualize the clustering result, dimensionality 2 is taken for instance. Figure 1 below describes the group distribution of the unsupervised clustering analysis with K-Means:

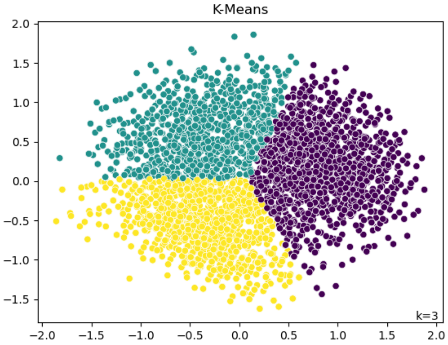


Figure 1: K-Means Clusters

2 Dimensionality Reduction

The initial dataset is comprised of 5344 samples and 17 distinct fields in total, including the patient index, individual scores for 15 questions and associated outcomes denoted by categorical labels 0, 1, and 2. During the data pre-processing, marks should be treated as data features and implemented as the input of the subsequent classifiers while respective results should be viewed as labels. Furthermore, data cleaning should be completed by means of deleting outliers and irrelevant columns. Correlation analysis is also necessary prior to the dimensionality reduction. After the data pre-processing, PCA is applied to reduce the data dimension to facilitate the training process. Specifically, bias detection, variance distribution and explained variance ratio are involved to assist the confirmation on the ultimate dimension.

2.1 Data Pre-Processing

2.1.1 Data Cleaning

Patients are originally divided into three groups which are 0, 1 and 2. However, it is noted that the label 2 is regarded as the noisy data thus such rows should be identified and removed. In addition, the index column serves no purpose in model fitting and therefore warrants deletion. Finishing the data cleaning, the resultant dataset retains 5330 samples eventually.

2.1.2 Correlation Analysis

To estimate the relevance of two variables, Pearson and Spearman coefficients [9] are two common approaches. Nevertheless, the Pearson's r correlation requires the data to be continuously distributed [10] while scores for each question are discrete values such as 1 and 2. Therefore, the Pearson coefficient is inapplicable for these features while the Spearman correlation analysis can resolve this inconsistency [11]. Referencing the following function, the Spearman correlation coefficient could be calculated:

$$r_s = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

According to the results of the formula, the correlation heatmap between each feature is shown in figure 2.

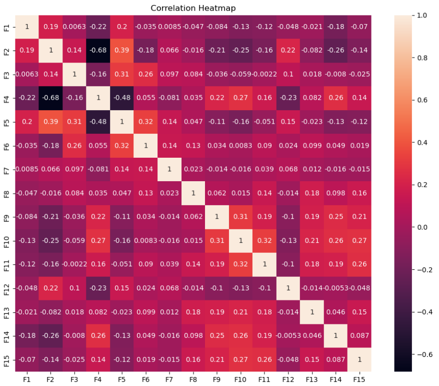


Figure 2: Correlation of each Feature

It could be inferred that 'F2' and 'F5' may have a relative positive effect mutually due to their positive correlation coefficient 0.39. Conversely, 'F2' and 'F4' have a negative value -0.68, which implies that a patient with a high mark for 'F2' may have poor performance in 'F4'. Other remaining coefficients are too slight to be emphasized. Therefore, it could be concluded that there are little correlation between every two columns except for 'F2', 'F4' and 'F5'.

2.2 Dimension Reduction with PCA

2.2.1 PCA

Altogether the dataset retains 15 features after the data preprocessing. Representative features should be selected to increase the precision of classifiers. Principal Component Analysis(PCA), one of the most prevalent dimension reduction methods, is adopted in this coursework. PCA is an algorithm that projects data onto certain dimensions which can preserve the maximum amount of variance, thereby preventing the information from loss as much as possible. The core processes are as follows:

1) Data normalization: before applying PCA to the dataset, it is crucial to normalize the mean value and variance using the zero-mean normalization:

$$x' = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the initial data. The objective of this step is to guarantee variables have an equivalent range of variation across all dimensions, which can avoid certain dimensions exerting dominant impacts on computations of principle components.

2) Covariance matrix calculation: the next procedure is to generate the covariance matrix according to normalized data. It is designed to evaluate the correlation of each feature to facilitate the principle components determination. The expression of the covariance and covariance matrix is shown below:

$$cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$\Sigma = \begin{bmatrix} cov(X_1,X_1) & cov(X_1,X_2) & \cdots & cov(X_1,X_n) \\ cov(X_2,X_1) & cov(X_2,X_2) & \cdots & cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n,X_1) & cov(X_n,X_2) & \cdots & cov(X_n,X_n) \end{bmatrix}$$

where μ_X and μ_Y are expected values of X and Y.

3) Eigenvalues and eigenvectors computation: eigenvalue is defined mathematically to characterize the dispersion degree of data in all directions [12], which is suitable

for variance measurement. Specifically, in PCA, eigenvalues represent the magnitude of principle components, which can be regarded as the importance of respective elements. Another concept is the eigenvector that corresponds to the projection coefficient [12], which refers to the direction of each principle components in the context of the PCA. In accordance with the covariance matrix, eigenvalues and eigenvectors can be calculated from the following function:

$$Cv = \lambda v$$

where C is covariance matrix, λ is eigenvalue and v is eigenvector. Additionally, singular value decomposition(SVD) is an alternative of obtaining eigenvalues and eigenvectors from the original dataset. In particular, in the formula below:

$$A = U\Sigma V^T$$

A is the data features matrix, the square of diagonal elements in Σ are eigenvalues and V contains all the principle components.

4) K principle components determination: on account of maximizing the preservation of the residual information upon projecting down to k dimensions, the eigenvalues calculated should be arranged in a descending order. Axes with the first k largest explained variance will be selected to be the principle components. Figure 3 below illustrates the explained variance distribution of each principle components.

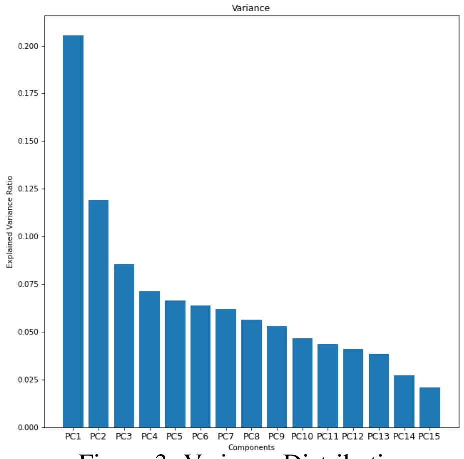


Figure 3: Variance Distribution

Similarly, it is noted that according to the cumulative sum of explained variance ratio shown in figure 4, 12 dimensions add up to a considerable portion that reaches up to nearly 92% of the total variance. Such high proportions occupied by the first 12 features guarantee the reliability and accuracy of classification, meanwhile, justify their selection as the principle components.

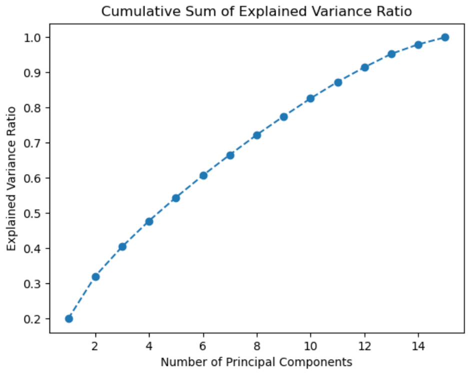


Figure 4: Explained Variance Ratio Sum

5) PCA visualization: the final dimensionality after the PCA analysis is 12, however, producing a 12-dimension image can be a laborious task. Hence, the dimensionality 2 is employed temporarily to simplify the data visualization in this section. The result is displayed in figure 5.

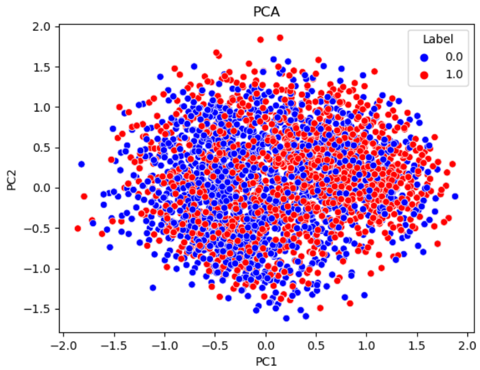


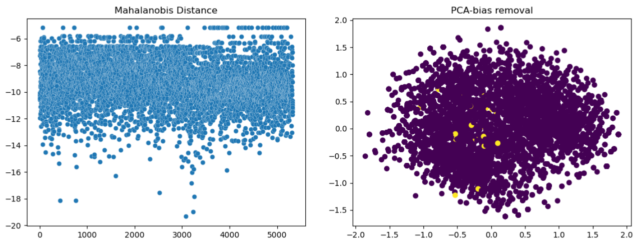
Figure 5: PCA (k=2)

2.2.2 Bias Removal

Although the PCA algorithm assists in completion of the data dimensional reduction, there still exits bias in the dataset, which could impede the precise classifier construction. Accordingly, bias removal is significant before model fitting. Specifically, the Mahalanobis Distance is applied to detect outliers. Following the formula below, the Mahalanobis Distance is a distance metric that indicates the distance between a point and certain aggregations [13].

$$d_M(\mathbf{x}, \mu) = \sqrt{(\mathbf{x} - \mu)^T \mathbf{S}^{-1} (\mathbf{x} - \mu)}$$

On the basis of calculated values, a proper threshold will be set and samples that exceed this threshold value will be viewed as outliers and warrant removal.



(a) Mahalanobis Distances (b) PCA Bias Removal (k=2)

Figure 6: Bias Detection and Removal

The Mahalanobis Distance distribution of the original data is shown in figure 6(a). It is noted that the majority of samples gather within the boundary '-16' and therefore this score could be regarded as the threshold for bias detection. Samples that beyond this value will be eliminated as noisy data. Specifically, through the PCA inverse transformation, 12-dimension data will be mapped to the initial space and conducted to the bias detection. Upon the bias removal, the corrective dataset will be again projected to 12 dimensional space with PCA. Likewise, to make the data visualization intelligible, the dimensionality 2 is taken as an instance in figure 6(b) to display the removal results.

3 Classifier Construction

There are five categorical classifiers which are built in total in this lab. Specifically, including SVM, DT, NB, DNN and LR. Ahead of delivering the dimension-reduced data to respective classifiers, it is fundamental to split the dataset into the training set and test set. Additionally, K-fold cross validation is applied to evaluate the performance of classification models.

3.1 Classifier Principles

1) Support Vector Machine(SVM) is a supervised learning algorithm with the aim of seeking for a maximum margin hyperplane to separate data points of different types [14], where the margin refers to the distance between the hyperplane and the closest data of various classes. Hyperplane follows the below function:

$$w^T x + b = 0$$

The next step is to optimize the margin which can be defined as:

$$\gamma = \frac{y_i(w^T x_i + b)}{||w||}$$

The closest data points relative to the decision boundary are called support vectors [14]. To prevent the model from the overfitting, soft margin classification is necessary to increase the flexibility and robustness. Additionally, RBF kernel is applicable for the linear inseparable dataset in this report:

$$k(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2), \gamma > 0$$

2) Decision Tree(DT) is a top-down classification model that bases on the tree data structure to recursively classify the dataset through a series of decision nodes. To partition the input feature vector into the concrete category, each node associates with a optimal attribute partitioning method which can be divided into three categories: ID3, C4.5 and CART [4]. More specifically, for ID3, partitioned features are selected based on the criterion of maximum information gain [4] which can be obtained from the information entropy and conditional information entropy using the following equations:

$$Ent(D) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

$$Ent(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} Ent(D_i)$$

$$Gain(D,A) = Ent(D) - Ent(D|A)$$

However, the ID3 is inclined to select features that account for a high proportion in the training set [4] which leads to the latent risk of overfitting and may reduce the generalization ability of the classifier. Therefore, C4.5 is invented to eliminate these drawbacks through the intrinsic value, grain ratio and pruning [15]:

$$IV(A) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|}$$

$$Gain\_ratio(D,A) = \frac{Gain(D,A)}{IV(A)}$$

Another optimal classification standard is CART algorithm that adopts the gini index [16] to measure the impurity of nodes, which is applicable for both classification and regression:

$$Gini(D) = 1 - \sum_{k=1}^{|n|} p_k^2$$

3) Naive Bayes(NB) is a Bayes' Theorem-based classifier which can be applied to handle with the dataset in this report. According to the following theorem:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

where $P(c)$ is named as prior probability, $P(x|c)$ is class-conditional probability(CCP), $P(x)$ is considered as evidence factor and the result $P(c|x)$ is called the posterior probability, the feature passed to the model will be categorized to the type with the largest $P(c|x)$. Nevertheless, under special circumstances, it is possible that certain features are not split into the training set, which may result in 0 for corresponding conditional probabilities and impair the classifier. Laplace Smoothing [17] with the function below is therefore used to resolve this disparity:

$$P(x_i|c) = \frac{N_{ic} + k}{N_c + kv}$$

where k is the Laplace coefficient.

4) The perception indicates a model that incorporates several inputs and a single output while Deep Neural Networks(DNN) which is also named as Multi-Layer Perception(MLP) [18], is an extension of the perception. The architecture of the DNN contains multiple layers of interconnected neurons, which can be specifically divided into three types: input layer, hidden layer and output layer. The DNN structure with 12-dimension inputs and 2 outputs is shown in figure 7:

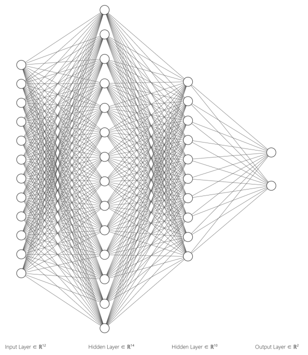

Input Layer ∈ $\mathbb{R}^{12}$
Hidden Layer ∈ $\mathbb{R}^{14}$
Hidden Layer ∈ $\mathbb{R}^{10}$
Output Layer ∈ $\mathbb{R}^{2}$

Figure 7: DNN Structure

During the model training in this lab, the activation function ReLU is employed to conduct non-linear transformation among each layers:

$$ReLU(x) = \max(0, x)$$

Moreover, Back-Propagation algorithm [18] led by the Chain Rule is widely used to achieve parametric optimization:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

5) Logistic Regression(LR) is one of the most common binary classifiers with the essence of constructing a parametric model that primarily inputs a linear function. Subsequently a logistic function, for example, the sigmoid function, will be applied to map the results to a likelihood score that provides a probabilistic interpretation of a sample belonging to a particular class separated by a decision boundary [19]. The sigmoid function is shown as follow and according to the calculated values, the sample can be classified to 1 or 0.

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$\hat{y} = \begin{cases} 1, & h_\theta(x) \geq 0.5 \\ 0, & h_\theta(x) < 0.5 \end{cases}$$

Minimizing the following log loss function using the gradient descent can assist in optimizing the parameters of model:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))]$$

$$\theta_j = \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

Moreover, in case of the overfitting, regularization methods such as Lasso Regression(L1) and Ridge Regression(L2) are essential and significant, which can be applied through the combination of cost function and L1-norm or L2-norm:

$$L_1 = \|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad L_2 = \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

3.2 Data Features Selection

During the data pre-processing phase, there remains a total of 5330 samples which consists of 2990 participants labeled as 0 and 2340 patients for label 1 after the data cleaning. The resultant classification distribution is reflected in figure 8. In order to alleviate the potential risks with high dimensions data, for instance, high computational complexity and overfitting issues, features extraction is the core procedure prior to the model fitting stage. Specifically, in this report, dimensionality reduction and bias removal are conducted with the application of the PCA algorithm and Mahalanobis Distance, which generates a 12-dimension dataset as the ultimate training set for candidate classifiers.

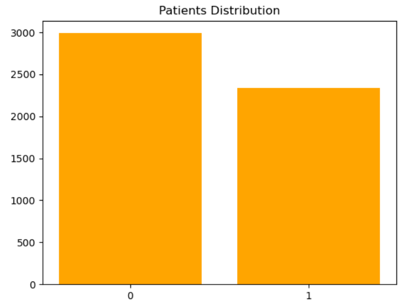
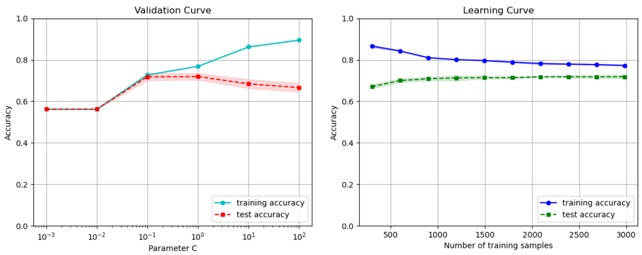


Figure 8: Participants

3.3 Classifier Selection

3.3.1 Classifier Training

1) Hyperparameter implies to parameters that needs manual specification instead of setting through automatic machine learning such as learning rate, batch size and maximum depth. For SVM specifically, ahead of inputting the 12-dimension training set, the ’rbf’ is used as the kernel function while the parameter ’gamma’ is set to ’auto’ which represents the reciprocal of the number of features. In addition, another hyperparameter ’c’ which called penalty parameter, is used to control the degree of punishment of violation. It can be viewed as a ’margin size’ and ’noise tolerance’ trade-off. The validation curve is applied to determine the optimal value of ’c’ between 10^{-3} and 10^{2}, where the validation curve [20] is a visualization method that can reflect the performance of classifiers with a variety of hyperparameter values. Figure 9(a) reveals that performance optimization in an accuracy of nearly 71% on the test set is achieved by setting ’c’ to 1 while the learning curve shown in figure 9(b) has a similar eventual score for the test dataset, which complies with the concrete value 70.93% presented in table 1 in the introduction section above.



(a) Validation Curve

(b) Learning Curve

Figure 9: Curve for SVM

2) During building the DT classifier, ’max_depth’ and ’criterion’ are two significant parameters that set to ’10’ and ’gini’ respectively. The first parameter imposes restrictions on the maximum depth of the tree, which can reduce the potential risks of overfitting. To simplify the visualization of decision-making procedure, maximum depth 2 is taken as an instance to generate a graph in figure 10. The second hyperparameter applies the Gini Index to calculate the impurity of nodes.

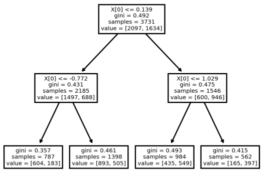


Figure 10: Decision with Gini

3) For NB, the argument ’alpha’ that named prior smoothing factor is equal to 1, which represents the Laplace Smoothing to guarantee the generalization ability and precision of the prediction model.

4) While constructing DNN classifiers, ’hidden layer sizes’ is set to 2 layers with 12 and 10 neurons respectively. ’Adam’ is specified for gradient descent optimization and the activation function is ’ReLU’ for non-linear mapping. In addition, to prevent the overfitting, ’shuffle’ is recommended to be ’True’ to randomly disorder the training set during each epoch.

5) In LR model, to ensure the classifier to generalize to new dataset, ’penalty’ is set to ’l1’ that implies the Lasso Regression and optimizer algorithm ’saga’ is selected due to the large-scale dataset in this report. The hyperparameter ’c’ follows the same meaning as in SVM and set to 1 upon the similar determination process.

3.3.2 K-Fold Cross Validation

Cross Validation(CV) is a statistic technique [21] that used to evaluate the performance and ensure the generalization ability of classifiers. The training set will be split into the training fold for model fitting and validation fold for estimation of classification prediction error. K-Fold Cross Validation is by far one of the most prevalent methods, which is also selected in this coursework. Specifically, the training set is firstly partitioned randomly into 10 folds which are mutually exclusive. Subsequently, one of them will be treated as the validation set successively while other 9 parts will be used to fit the classifier. The Mean Squared Error(MSE) of each validation set will be recorded repeatedly. The final stage is to calculate the

CV score that refers to the average of MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

The K-Fold CV scores of each model are displayed as followings:

Table 2: CV Scores

Classifier	Score on Train Set	Score on Test Set
SVM	71.94%	70.37%
DT	64.97%	61.72%
NB	72.02%	70.65%
DNN	71.05%	68.67%
LR	73.09%	69.98%

Model fitting conducts on the train set, which makes it possible that classifiers could learn certain rules or latent structures of training samples. On the contrary, properties of the test set may differ from the training set, which is normally unknown for classifiers. Hence, to avoid overestimation and obtain an objective assessment on the performance of classifiers, CV scores on the test set should be regarded as the evaluation criteria. It can therefore be concluded that the Naive Bayes model gains the highest score (70.65%) according to the table 2 above. Corresponding confusion matrix displayed in figure 11 provides a detailed illustration of the classification result distribution.

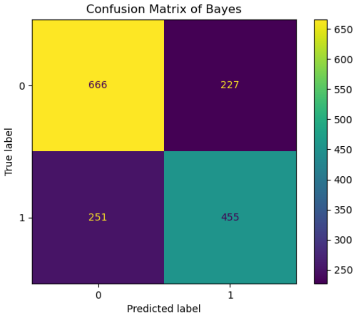


Figure 11: Confusion Matrix

3.3.3 Supervised Classifier Selection

Receiver Operating Characteristic Curve(ROC) is an evaluative graph that represents the discriminative power of classification models, which describes the variation tendency of recall and fall-out at different thresholds. Area Under the ROC Curve(AUC), a measure that ranges from 0 to 1, can be adopted to compare the discrimination ability of various classifiers. More specifically, a higher AUC value normally indicates a better predictive performance. On the basis of the predictions generated from the training process of each model, respective ROC curve with a concrete AUC value is shown in figure 12. It could be concluded that SVM, NB and LR achieve the same highest score of 0.70, while DT performs relatively poorly with an AUC value 0.63.

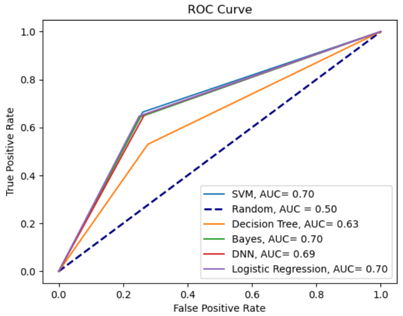


Figure 12: ROC Curve

In addition to AUC values, the accuracy should also be considered as a crucial performance criterion to select an appropriate classifier. The experimental results from table 1 above show that SVM realizes the most precise classification in this lab with an accuracy of 70.93%. Furthermore, according to the CV scores, NB scores the highest with 70.65% while SVM also yield well finishing with 70.37%. However, ultimate determination on classifiers cannot rely on a single evaluation standard. Although NB scores slightly higher than SVM, the difference reaches simply 0.28% which is relatively negligible. Therefore, considering the eminent performance of SVM in the first two metrics and marginal weakness in CV score, SVM is comprehensively selected to be the eventual recommendation on the classification model.

4 Unsupervised Learning

4.1 K-Means Clustering Principles

Clustering is an unsupervised leaning method that classifies objects into certain groups based on the similarity of sample attributes. In this report, K-Means, a typical distance-based [22] unsupervised clustering algorithm, is introduced to categorize the patients into k number of clusters. K-Means mainly follows the procedures below:

1) Center initialization: completing the data preprocessing, the number of initial clustering centers k should be selected randomly from the dataset, which can be recorded

as:

$$C = \{c_1, c_2, \ldots, c_k\}$$

2) Cluster allocation: for the remaining samples, on the basis of calculating distance to each center, assigning each point to the cluster with the nearest distance:

$$S_i = \{x_p | x_p \in X, \arg\min_j d(x_p, c_j) = i\}$$

3) Center recomputation: when distributing a new sample to a specific cluster, the average of all data points in the group will be computed as the updated clustering center:

$$c_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j$$

4) Iteration: repeating step 2 and 3 until reaching the limited number of iterations or the variation of within-cluster sum of square error(SSE) is less than a specified threshold, which means that the stabilization of clusters distribution:

$$SSE = \sum_{i=1}^{k} \sum_{x \in S_i} ||x - c_i||^2$$

4.2 Determination on K

The 'n_clusters', an indispensable parameter for K-Means algorithm to determine the number of classifications, is set an initial range from 3 to 7(excluded). However, it is arduous to achieve the 12 dimensional dataset visualization and consequently the dimension 2 is taken for example in figure 13 for clustering results.

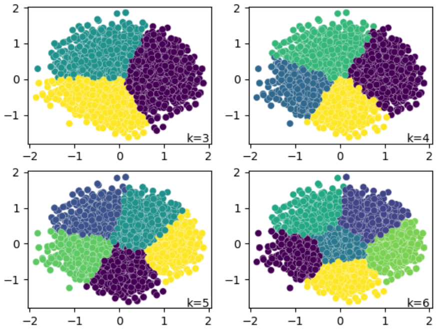


Figure 13: K = 3 to 6

To optimalize the clustering result, the Silhouette Coefficient is an applicable approach for determining the value of k. Cohesion and Separation are two criteria that respectively refer to the aggregation degree of within-class and extra-class elements. Silhouette Coefficient between -1 and 1 generated from the following formula synthesizes these two factors to evaluate the capacity of clustering algorithm:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} = \begin{cases} 1 - \frac{a_i}{b_i}, & a_i \leq b_i \\ \frac{b_i}{a_i} - 1, & a_i > b_i \end{cases}$$

where a_i represents the inner-class distance while b_i illustrates the between-class distance. More specifically, a high positive coefficient generally means samples within the cluster gather compactly and boundaries of distinct categories are evident, which corresponds to a well-performed group distribution. For this 12 dimensions dataset, figure 14 displays relevant Silhouette Coefficients with various k and the highest score 0.138 is related to the value 3 for k:

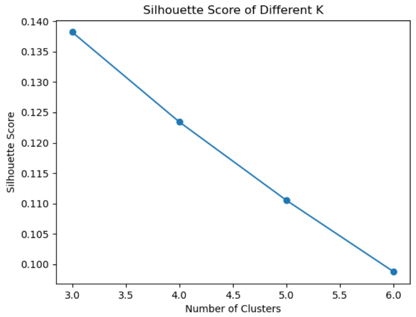


Figure 14: Silhouette Coefficient

Therefore, it is recommended to separate the patients into 3 groups in this coursework. Moreover, due to the above-mentioned reason which is the difficulty in visualizing the 12-dimension data, figure 1 takes the dimensionality 2 as an example to achieve the clustering result visualization.

5 Conclusion

According to the results of data pre-processing, scores of 15 questions are correlated with the patients distribution in this report. After dimensionality reduction using PCA and bias removal with Mahalanobis Distance, the 12-dimension dataset is applied to conduct the model training, where SVM is the recommended classifier with an accuracy of 70.93%. Moreover, during the unsupervised clustering with K-Means, it is suggested to partition the patients into 3 groups on the basis of the Silhouette Coefficient. However, optimization is also necessary. Specifically, for the unsupervised learning phase, the number of initial centers of K-Means is inestimable while different values of k will lead to a variety of clustering performances. To balance this deficiency, an enhanced algorithm K-Means++ could be used to select initial centers as far as possible to increase the precision and reliability of distribution results.

References

[1] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, 2020. 1.2

[2] R. R. Sharma and A. Sungheetha, "An efficient dimension reduction based fusion of cnn and svm model for detection of abnormal incident in video surveillance," *Journal of Social and Clinical Psychology*, vol. 3, pp. 55–69, 2021. 1.2

[3] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction," *Scientific Reports*, vol. 12, 2022. 1.2

[4] B. Charbuty and A. M. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, 2021. 1.2, 3.1

[5] F. Shahi and A. T. Rezakhani, "Fidelity-based supervised and unsupervised learning for binary classification of quantum states," *The European Physical Journal Plus*, vol. 136, 2021. 1.2

[6] G. Sharma, A. Panwar, I. Nasiruddin, and R. C. Bansal, "Non-linear ls-svm with rbf-kernel-based approach for agc of multi-area energy systems," *IET Generation, Transmission & Distribution*, 2018. 1.2

[7] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng, "A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots," *Applied Sciences*, 2021. 1.2

[8] S. M. Aziz and N. A. K. Rifai, "Pengelompokkan ekspor kopi menurut negara tujuan menggunakan metode k-means clustering dengan silhouette coefficient," *Bandung Conference Series: Statistics*, 2022. 1.2

[9] A. Rovetta, "Raiders of the lost correlation: A guide on using pearson and spearman coefficients to detect hidden correlations in medical sciences," *Cureus*, vol. 12, 2020. 2.1.2

[10] G. Nahler, "Pearson correlation coefficient," *Definitions*, 2020. 2.1.2

[11] K. A. A. Al-Hameed, "Spearman's correlation coefficient in statistical analysis," 2022. 2.1.2

[12] Y. He, M. Wu, and Y.-H. Xia, "The eigenvector-eigenvalue identity for the quaternion matrix with its algorithm and computer program," 2022. 2.2.1

[13] J. J. Ren, S. Fort, J. Z. Liu, A. G. Roy, S. Padhy, and B. Lakshminarayanan, "A simple fix to mahalanobis distance for improving near-ood detection," *ArXiv*, vol. abs/2106.09022, 2021. 2.2.2

[14] M. Fan and A. Sharma, "Design and implementation of construction cost prediction model based on svm and lssvm in industries 4.0," *Int. J. Intell. Comput. Cybern.*, vol. 14, pp. 145–157, 2021. 3.1

[15] M. M. Mijwil and R. A. Abttan, "Utilizing the genetic algorithm to pruning the c4.5 decision tree algorithm," *Asian Journal of Applied Sciences*, 2021. 3.1

[16] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: Cart model," *Computer methods and programs in biomedicine*, vol. 192, p. 105400, 2020. 3.1

[17] E. R. Setyaningsih and I. Listiowarni, "Categorization of exam questions based on bloom taxonomy using naïve bayes and laplace smoothing," *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, pp. 330–333, 2021. 3.1

[18] C. C. Aggarwal, "Neural networks and deep learning," in *Cambridge International Law Journal*, 2018. 3.1, 3.1

[19] R. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," *International Journal of Environmental Research and Public Health*, vol. 18, 2021. 3.1

[20] F. Mohr, "Towards model selection using learning curve cross-validation," 2021. 3.3.1

[21] X. Zhang and C.-A. Liu, "Model averaging prediction by k-fold cross-validation," *SSRN Electronic Journal*, 2022. 3.3.2

[22] A. M. Ikotun, E. E. Absalom, L. M. Abualigah, B. Abuhaija, and H. Jia, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, pp. 178–210, 2022. 4.1