## Project 2: Machine Learning

In this project, I am going to analyze data based on machine learning paradigms and try to find out the best method to modelized our problem and make prediction on future data.

In Problem 1 and 2, I will read data from “heart1.csv”, which the target result a1p2 only has 2 class: 1 and 2, which means absence of heart disease or not.

In Problem 3, I will read data from “heart2\_.csb”, which shares the same parameters as previous problems but the target variable a1p2 will have 5 different classes, which ranges 0 to 4.

### Problem 1:

First, I analysis the data by finding the correlation of each variables. Cross correlation matrix is showed as follows:

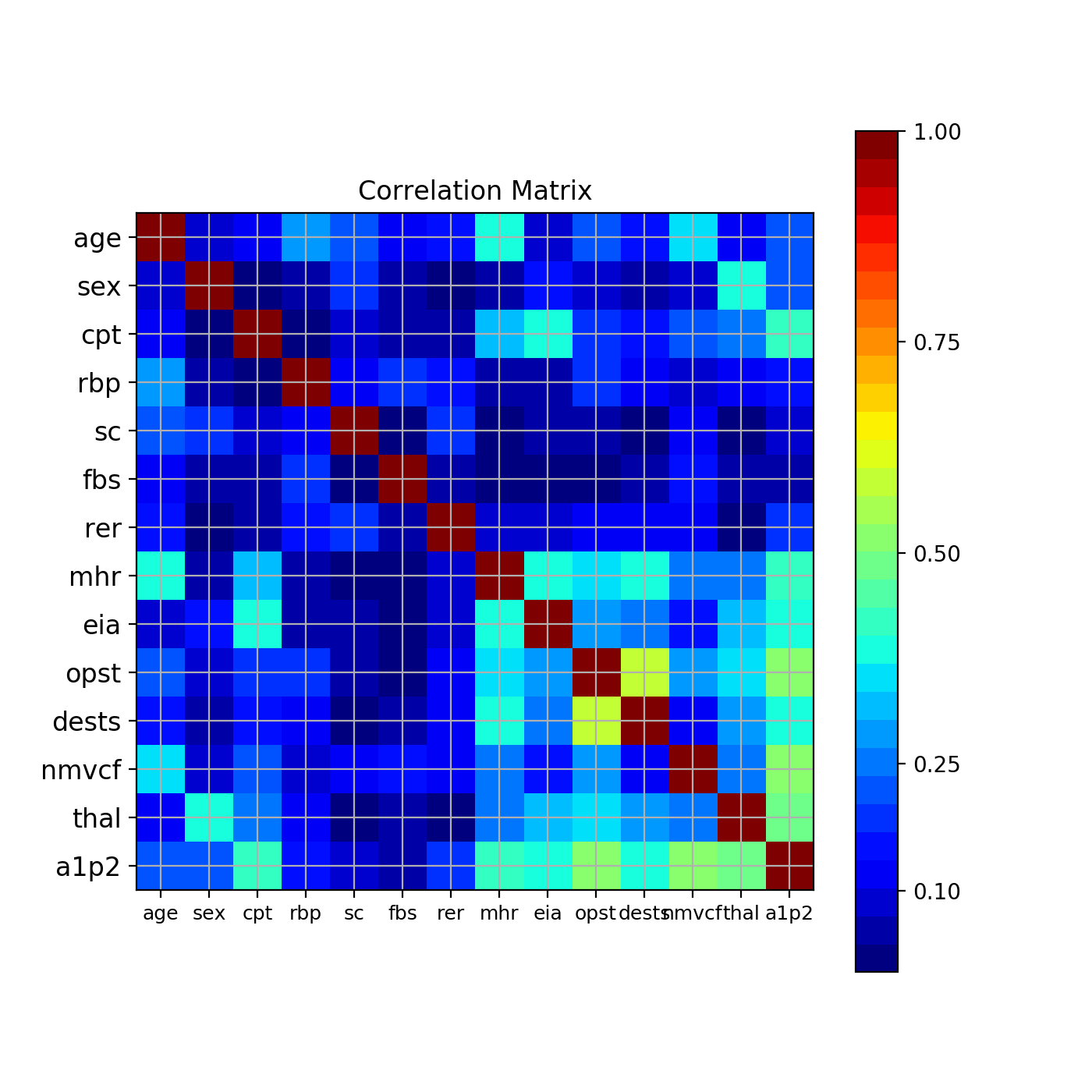


Fig 1-1: Correlation Matrix of heart1.csv

一張含有 黑色 的圖片



描述是以高可信度產生

Fig 1-2: Top 15 Cross Correlated Variables of heart1.csv

According to Fig 1-1 and 1-2, I choose dest, thal, nmvcf, eia, mhr, opst and cpt as our variables for further analysis. I also made a pair plot for those chosen variables as below:

一張含有 窗戶, 室內 的圖片



描述是以高可信度產生

Fig 1-3: Pair Plot for Chosen Variables of heart1.csv

From the pair plot, we can see that the combination of opst and mhr shows the less correlation to each other since their pair plot is distributed in a more random way. As a first step observation, I would take these two variables for further analysis.

### Problem 2:

Based on our previous observation, I import all machine learning method I know: Perceptron, Logistic Regression, Random Forest Classifier, K Nearest Neighbors Classifier, Decision Tree Classifier, Support Vector Machine for both Linear and RBF kernel cases. The table of accuracy for each method is as below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Method** | **Accuracy** | **Combined Accuracy** |
| 0 | Perceptron | 0.679 | 0.711 |
| 1 | Logistic Regression | 0.716 | 0.741 |
| 2 | Random Forest Classifier | 0.617 | 0.837 |
| 3 | KNearestNeighbors Classifier | 0.654 | 0.731 |
| 4 | Decision Tree Classifier | 0.617 | 0.737 |
| 5 | Support Vector Machine (Linear) | 0.704 | 0.73 |
| 6 | Support Vector Machine (RBF kernl) | 0.679 | 0.744 |

Table 2-1: Method of Machine Learning for Variables opst and mhr of heart1.csv

一張含有 文字 的圖片



描述是以高可信度產生

Fig 2-1: Plot of Decision Region with Random Forest Tree of heart1.csv

一張含有 文字, 地圖 的圖片



描述是以非常高的可信度產生

Fig 2-2: Plot of Decision Region with Logistic Regression of heart1.csv

According to the table above, the best method to obtain the highest overall accuracy is Random Forest Classifier. But for prediction, I would choose Logistic Regression since I can obtain better prediction accuracy if the task is to predict for future diagnose. Note that if we use all the 7 most related variables, the result would be different:

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Method** | **Accuracy** | **Combined Accuracy** |
| 0 | Perceptron | 0.815 | 0.852 |
| 1 | Logistic Regression | 0.84 | 0.859 |
| 2 | Random Forest Classifier | 0.778 | 0.919 |
| 3 | KNearestNeighbors Classifier | 0.815 | 0.852 |
| 4 | Decision Tree Classifier | 0.79 | 0.859 |
| 5 | Support Vector Machine (Linear) | 0.84 | 0.856 |
| 6 | Support Vector Machine (RBF kernel) | 0.815 | 0.93 |

Table 2-2: Method of Machine Learning for 7 Most Correlated Variables of heart1.csv

The method that can obtain the highest overall accuracy is changed to Support Vector Machine (RBF kernel). However, the best method for prediction is still Logistic Regression.

### Problem 3:

In Problem3, I read in heart2\_.csv and try the same method to classify the new dataset. From Fig. 3-1 to 3-3, we can observe that the most correlated variables are the same as dataset in previous problem. However, the accuracy rate is not able to reach the same level as in previous dataset.

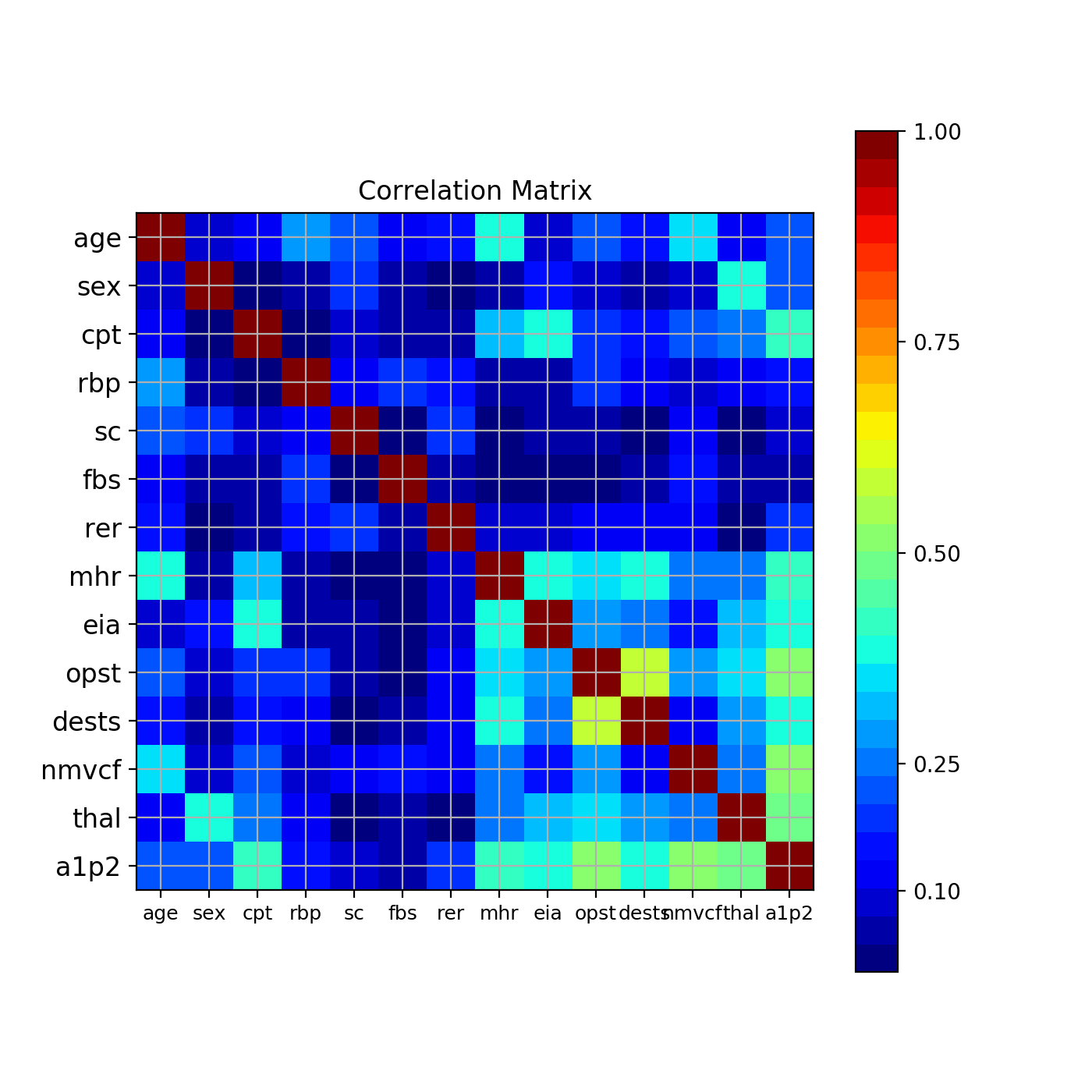


Fig 3-1: Correlation Matrix of heart2.csv

一張含有 黑色 的圖片



描述是以高可信度產生

Fig 3-2: Top 15 Cross Correlated Variables of heart2.csv

一張含有 窗戶, 建築物 的圖片



描述是以高可信度產生

Fig 3-3: Pair Plot for Chosen Variables of heart2.csv

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Method** | **Accuracy** | **Combined Accuracy** |
| 0 | Perceptron | 0.352 | 0.429 |
| 1 | Logistic Regression | 0.495 | 0.584 |
| 2 | Random Forest Classifier | 0.407 | 0.759 |
| 3 | KNearestNeighbors Classifier | 0.451 | 0.607 |
| 4 | Decision Tree Classifier | 0.462 | 0.594 |
| 5 | Support Vector Machine (Linear) | 0.451 | 0.587 |
| 6 | Support Vector Machine (RBF kernel) | 0.451 | 0.614 |

Table 3-1: Method of Machine Learning for Variables opst and mhr of heart2.csv

一張含有 文字 的圖片



描述是以高可信度產生

Fig 3-4: Plot of Decision Region with Random Forest Tree of heart2.csv

一張含有 文字, 地圖 的圖片



描述是以非常高的可信度產

Fig 3-5: Plot of Decision Region with Logistic Regression of heart2.csv

According to Table 3-1, even though we can’t make the same level of the prediction, the best method to obtain the highest overall accuracy is still Random Forest Classifier. And Logistic Regression will still make the best prediction.

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Method** | **Accuracy** | **Combined Accuracy** |
| 0 | Perceptron | 0.516 | 0.581 |
| 1 | Logistic Regression | 0.495 | 0.64 |
| 2 | Random Forest Classifier | 0.505 | 0.845 |
| 3 | KNearestNeighbors Classifier | 0.549 | 0.647 |
| 4 | Decision Tree Classifier | 0.473 | 0.607 |
| 5 | Support Vector Machine (Linear) | 0.516 | 0.653 |
| 6 | Support Vector Machine (RBF kernel) | 0.429 | 0.818 |

Table 3-2: Method of Machine Learning for 7 Most Correlated Variables of heart2.csv

Even we try to apply all of the highly correlated variables, the predicting accuracy is still low. The possible reason is that the a1p2 variable is imbalance, which would result in unsatisfactory classifier. To avoid this condition, more data of a1p2 with values other than 0 is required.

一張含有 路面, 黑色 的圖片



描述是以高可信度產生

Fig 3-6: Statistics of a1p2 Variable