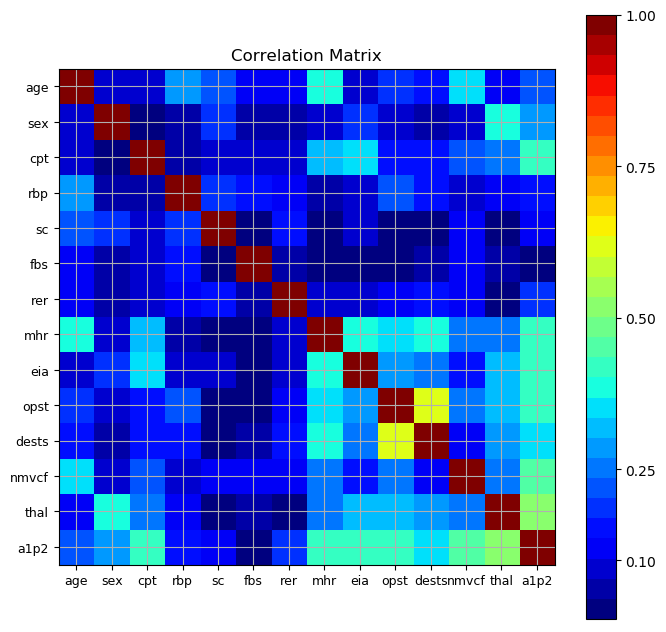
EEE591 Python for Rapid Engineering Solution | Project 1

Po-Kan Shih (1214361589)

**Problem 1:**

Covariance matrix of all variables:

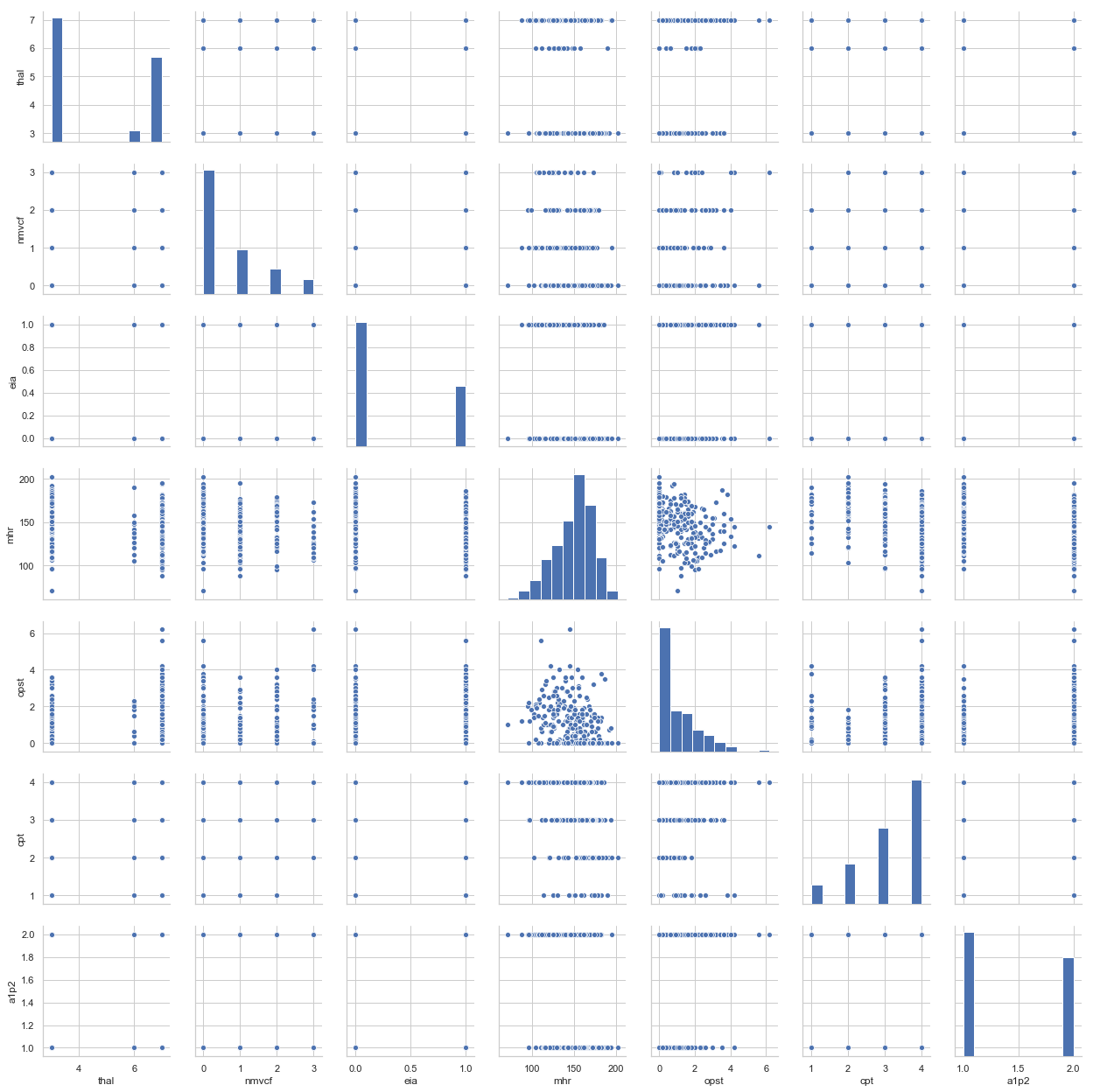


Calculate the pairwise correlation, and show the first 10 variable pairs with most highly correlation:

|  |  |  |
| --- | --- | --- |
| **First variable name** | **Second variable name** | **correlation** |
| opst | dests | 0.609712 |
| thal | a1p2 | 0.525020 |
| nmvcf | a1p2 | 0.455336 |
| eia | a1p2 | 0.419303 |
| mhr | a1p2 | 0.418514 |
| opst | a1p2 | 0.417967 |
| cpt | a1p2 | 0.417436 |
| age | mhr | 0.402215 |
| sex | thal | 0.391046 |
| mhr | dests | 0.386847 |

According to the correlation table listed above, the most highly correlated pair of variables is opst and dests, but dests does not have high correlation with desired result (a1p2), so they will not be selected for classifiers. We choose (“thal”, “nmvcf”, “eia”, “mhr”, “opst”, “cpt”) as best predictors, because their correlation values with the desired result (a1p2) are higher than 0.4, and plot the pairwise distributions of thes variables and desired result (a1p2):

Pair plot



**Problem 2:**

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Testing data set | Misclassified | Accuracy |
| Perceptron | 270 | 54 | 80% |
| Logistic Regression | 270 | 38 | 86% |
| Support Vector Machine (linear kernel function) | 270 | 41 | 85% |
| Support Vector Machine (nonlinear kernel function) | 270 | 14 | 95% |
| Decision Tree | 270 | 32 | 88% |
| Random Forest | 270 | 6 | 98% |
| K Nearest Neighbor | 270 | 4 | 99% |

It can be seen that the random forest and K nearest neighbor both rank the highest level, and then support vector machine with nonlinear kernel function. First, from the correlation table, it is evident that there is no variable that is very highly dependent with the desired result, so it is needed to combine multiple variables to increase the accuracy. Second, the pair plot shows that there is no linear relationship between variables and desired result, so nonlinear methods work better than linear ones.