

AMAPE Executive Summary

Prepared by Ahmed Ismail Abu Qahf (1223023698)

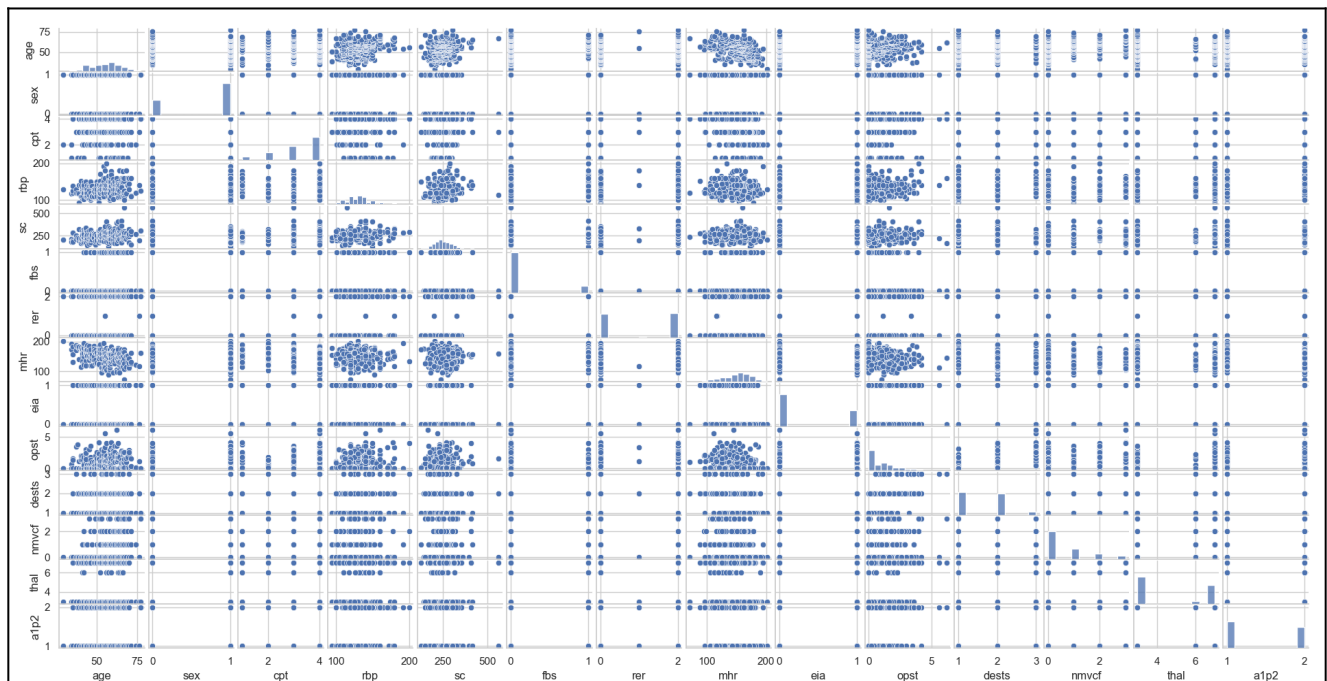
Part 1

The `prob1.py` script analysis on the file `heart1.csv` includes 6 tables and a pair plot for the features extracted from the data frame created from the file.

Tables

Table Name	Table Representation
Table 1	Correlation matrix of each variable with all other variables.
Table 2	Highly correlated variables.
Table 3	Highly correlated variables with ' <code>a1p2</code> '.
Table 4	Covariance matrix of each variable with all other variables.
Table 5	Variables with high covariance between them.
Table 6	Variables with high covariance with ' <code>a1p2</code> '.

Pair Plot



Observations

1. From table 3, we can see that the highly correlated variables are **(dests, opst)** with a correlation value of **0.6**, and also shows that the independent variable **'a1p2'** could be considered one of the best variables as it enjoys a good correlation with **(thal, nmvcf, eia, mhr)**.
2. From table 4, we can see that the variables with high correlation with the independent variable are **(thal, nmvcf, eia, mhr, opst)**.
3. From table 5, we can see that the variables with high covariance between them are **(sc, rbp)** with a value of **159**, and also shows that the variable age seems to vary with most of the other variables.
4. From table 6, we can see that the variables that vary from the independent variable are **(mhr, sc, rbp, age...etc.)**.

Conclusion

The variable **thal** seems to be in the best correlation with the independent variable **a1p2**. The variables **(dests, opst)** exhibit the highest correlation. So these variables **(thal, dests, opst)** are essentially the best variables to be considered for training the machine learning model.

Part 2

The `prob2.py` script analysis on the file `heart1.csv` includes 6 methods presented in the following table:

Table

Classifier Number	Classifier Name	Test Accuracy	Combined Accuracy
1	Perceptron	0.84	0.86
2	Logistic Regression	0.85	0.87
3	Support Vector Machine	0.85	0.87
4	Decision Tree	0.78	0.90
5	Random Forest	0.81	0.93
6	K-Nearest Neighbors	0.67	0.76

Observations

The above table gives the accuracy of the test and combined datasets in terms of the number of the classified items in each dataset. From the above table, we can infer the following inequality in terms of the test accuracy:

SVM > Logistic Regression > Perceptron > Random Forest > Decision Tree > KNN

Conclusion

From the above inferences, we can conclude that the **Support Vector Machine** algorithm is the best method of predicting the independent variable `'a1p2'`.