AMAPE Machine Learning Prediction Model Report

Carson Wagner

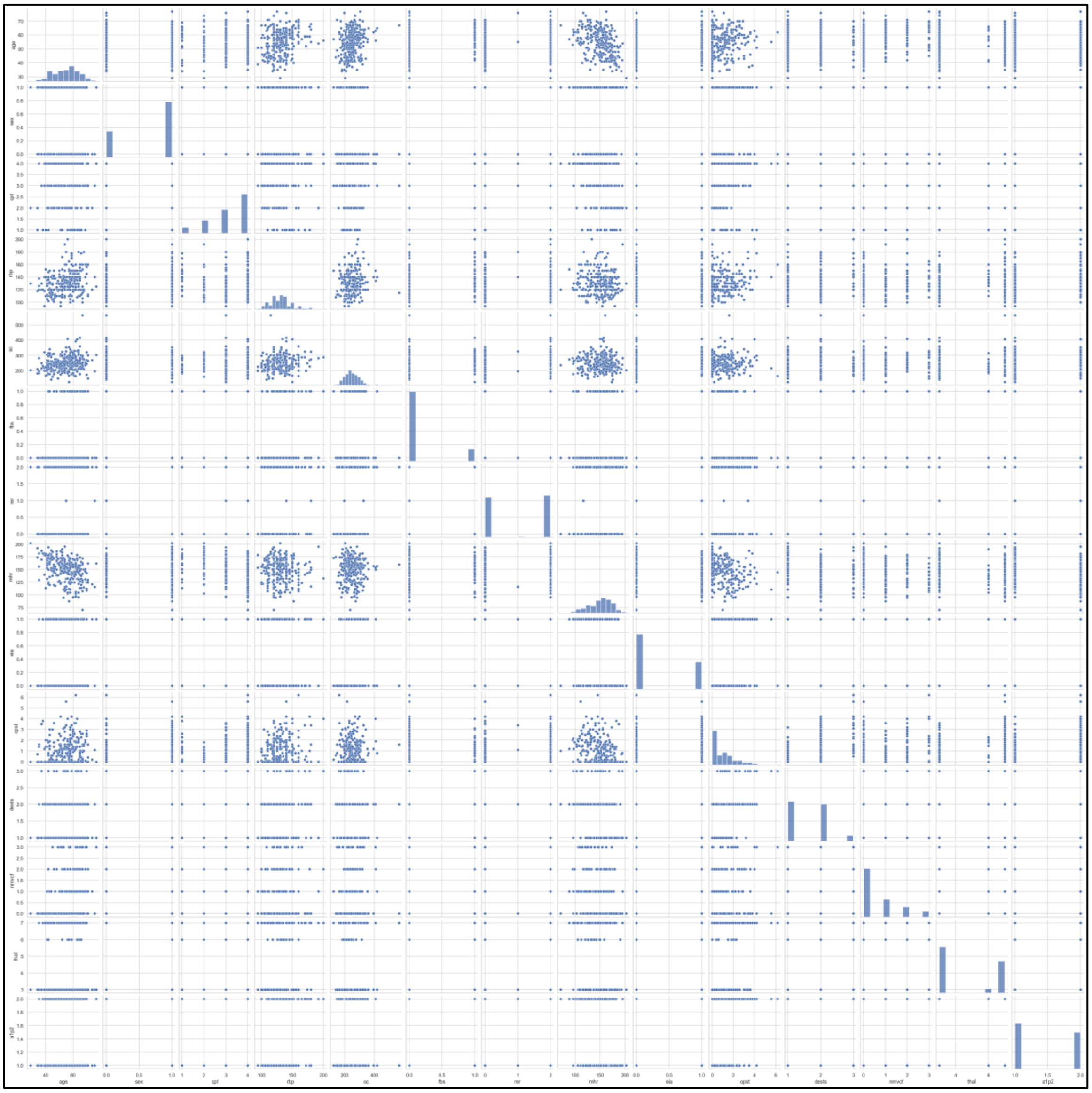
Part 1:

The goal of this script was to analyze heart1.csv and to provide tables to show correlation and covariance of the variables within the data set to find which variables are the best for training the machine learning model to predict heart disease in a patient.

Program Details

1. The first step of Problem1.py is to find the correlation matrix for the heart1.csv data set to find the correlation between the variables.
2. We also find the covariance matrix for heart1.csv data set to the variables with the covariance between each other.
3. The highest correlation is also found for the heart1.csv data set to find which variables have the most correlation between each other, which shows that **dests** and **opst** has the highest correlation between each other amongst all the variables with a correlation of 0.609712. It is also seen that the highest correlation for **a1p2** is **thal**
4. The highest covariance is found for heart1.csv data set to find which variables have the highest covariance between each other. This shows that the highest covariance between variables is **sc** and **rbp** with the covariance value of 159.731185. The highest covariance with **a1p2** is **mhr** with the value of 4.826518

Pair Plot



Analysis of Dataset:

After analyzing the heart1.csv dataset, the **thal** variable has the highest correlation with a1p2 variable with the correlation value being 0.525020 with a1p2. The highest correlation between two variables was **dests** and **thal** with a correlation value of 0.609712. Due to the high correlation between **dests** and **thal** this also means that **dests** is a valuable variable for predicting heart disease. As a final analysis, **thal, dests,** and **a1p2** are the best variables for predicting heart disease in a patient and training the machine learning model.

Problem 2:

Problem2.py uses machine learning algorithms to predict heart disease and display the test accuracy and the combined accuracy of the algorithms. There are 6 machine learning methods.

Machine Learning Method Result

Machine Learning Method	**Test Accuracy**	**Combined Accuracy**
Perceptron	0.84	0.86
Logistic Regression	0.85	0.87
Support Vector Machine	0.85	0.87
Decision Tree	0.51	0.90
Random Forest	0.83	0.95
K-Nearest Neighbor	0.84	0.85

Conclusion

Upon inspection of the Machine Learning Method Result the method with the best accuracy was Logistic Regression and Support Vector Machine with equal values of 0.85

test accuracy and 0.87 combined accuracy. This provides the best accuracy values between variables, unlike the Decision Tree with 0.51 test accuracy and 0.95 combined accuracy as the accuracies are not close in value. The Machine Learning Algorithms Logistic Regression and Support Vector Machine are the best methods for predicting the variable **a1p2** which is the variable of predicting heart disease in a patient