AMAPE - Heart Disease Report using Machine Learning

Name: Divyesh Rathod

ASU ID: 1225916954

Problem Statement 1: Read the heart1.csv database and analyze the data. Create matrices to show role of each variable and its impact on predicting heart disease.

Tables from Problem1.py

1. Correlation matrix to understand correlation coefficients between pairs of variables in a dataset.

Initial Correlation Matrix

	age	sex	cpt	rbp	sc	fbs	rer	mhr	eia	opst	dests
age	1.00	-0.09	0.10	0.27	0.22	0.12	0.13	-0.40	0.10	0.19	0.16
sex	-0.09	1.00	0.03	-0.06	-0.20	0.04	0.04	-0.08	0.18	0.10	0.05
cpt	0.10	0.03	1.00	-0.04	0.09	-0.10	0.07	-0.32	0.35	0.17	0.14
rbp	0.27	-0.06	-0.04	1.00	0.17	0.16	0.12	-0.04	0.08	0.22	0.14
sc	0.22	-0.20	0.09	0.17	1.00	0.03	0.17	-0.02	0.08	0.03	-0.01
fbs	0.12	0.04	-0.10	0.16	0.03	1.00	0.05	0.02	-0.00	-0.03	0.04
rer	0.13	0.04	0.07	0.12	0.17	0.05	1.00	-0.07	0.10	0.12	0.16
mhr	-0.40	-0.08	-0.32	-0.04	-0.02	0.02	-0.07	1.00	-0.38	-0.35	-0.39
eia	0.10	0.18	0.35	0.08	0.08	-0.00	0.10	-0.38	1.00	0.27	0.26
opst	0.19	0.10	0.17	0.22	0.03	-0.03	0.12	-0.35	0.27	1.00	0.61
dests	0.16	0.05	0.14	0.14	-0.01	0.04	0.16	-0.39	0.26	0.61	1.00
nmvcf	0.36	0.09	0.23	0.09	0.13	0.12	0.11	-0.27	0.15	0.26	0.11
thal	0.11	0.39	0.26	0.13	0.03	0.05	0.01	-0.25	0.32	0.32	0.28
a1p2	0.21	0.30	0.42	0.16	0.12	-0.02	0.18	-0.42	0.42	0.42	0.34

	nmvcf	thal	a1p2
age	0.36	0.11	0.21
sex	0.09	0.39	0.30
cpt	0.23	0.26	0.42
rbp	0.09	0.13	0.16
sc	0.13	0.03	0.12
fbs	0.12	0.05	-0.02
rer	0.11	0.01	0.18
mhr	-0.27	-0.25	-0.42
eia	0.15	0.32	0.42
opst	0.26	0.32	0.42
dests	0.11	0.28	0.34
nmvcf	1.00	0.26	0.46
thal	0.26	1.00	0.53
a1p2	0.46	0.53	1.00

2. Covariance Matrix

```
Covariance Matrix
          age    sex    cpt     rbp      sc    fbs    rer     mhr    eia   opst  \
age     82.98  -0.40   0.84   44.43  103.61   0.40   1.17  -84.87   0.42   2.03
sex     -0.40   0.22   0.02   -0.52   -4.88   0.01   0.02   -0.83   0.04   0.05
cpt      0.84   0.02   0.90   -0.73    4.44  -0.03   0.07   -6.99   0.16   0.18
rbp     44.43  -0.52  -0.73  319.04  159.73   0.99   2.07  -16.19   0.70   4.56
sc     103.61  -4.88   4.44  159.73 2671.47   0.46   8.65  -22.44   1.90   1.64
fbs      0.40   0.01  -0.03    0.99    0.46   0.13   0.02    0.19  -0.00  -0.01
rer      1.17   0.02   0.07    2.07    8.65   0.02   1.00   -1.73   0.04   0.14
mhr    -84.87  -0.83  -6.99  -16.19  -22.44   0.19  -1.73  536.65  -4.15  -9.26
eia      0.42   0.04   0.16    0.70    1.90  -0.00   0.04   -4.15   0.22   0.15
opst     2.03   0.05   0.18    4.56    1.64  -0.01   0.14   -9.26   0.15   1.31
dests    0.89   0.01   0.08    1.56   -0.18   0.01   0.10   -5.51   0.07   0.43
nmvcf    3.06   0.04   0.20    1.44    6.17   0.04   0.11   -5.80   0.07   0.28
thal     1.88   0.36   0.48    4.58    2.89   0.03   0.01  -11.39   0.29   0.72
a1p2     0.96   0.07   0.20    1.38    3.04  -0.00   0.09   -4.83   0.10   0.24

        dests  nmvcf    thal   a1p2
age      0.89   3.06    1.88   0.96
sex      0.01   0.04    0.36   0.07
cpt      0.08   0.20    0.48   0.20
rbp      1.56   1.44    4.58   1.38
sc      -0.18   6.17    2.89   3.04
fbs      0.01   0.04    0.03  -0.00
rer      0.10   0.11    0.01   0.09
mhr     -5.51  -5.80  -11.39  -4.83
eia      0.07   0.07    0.29   0.10
opst     0.43   0.28    0.72   0.24
dests    0.38   0.06    0.34   0.10
nmvcf    0.06   0.89    0.47   0.21
thal     0.34   0.47    3.77   0.51
a1p2     0.10   0.21    0.51   0.25
```

3. Variables highly correlated with each other

```
Most highly correlated with each other
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
dtype: float64
```

4. Variables highly correlated with a1p2 (variable to be predicted)

```
Correlation Matrix with a1p2
   Variable  Correlation with a1p2
0       age                   0.21
1       sex                   0.30
2       cpt                   0.42
3       rbp                   0.16
4        sc                   0.12
5       fbs                  -0.02
6       rer                   0.18
7       mhr                  -0.42
8       eia                   0.42
9      opst                   0.42
10    dests                   0.34
11    nmvcf                   0.46
12     thal                   0.53
```

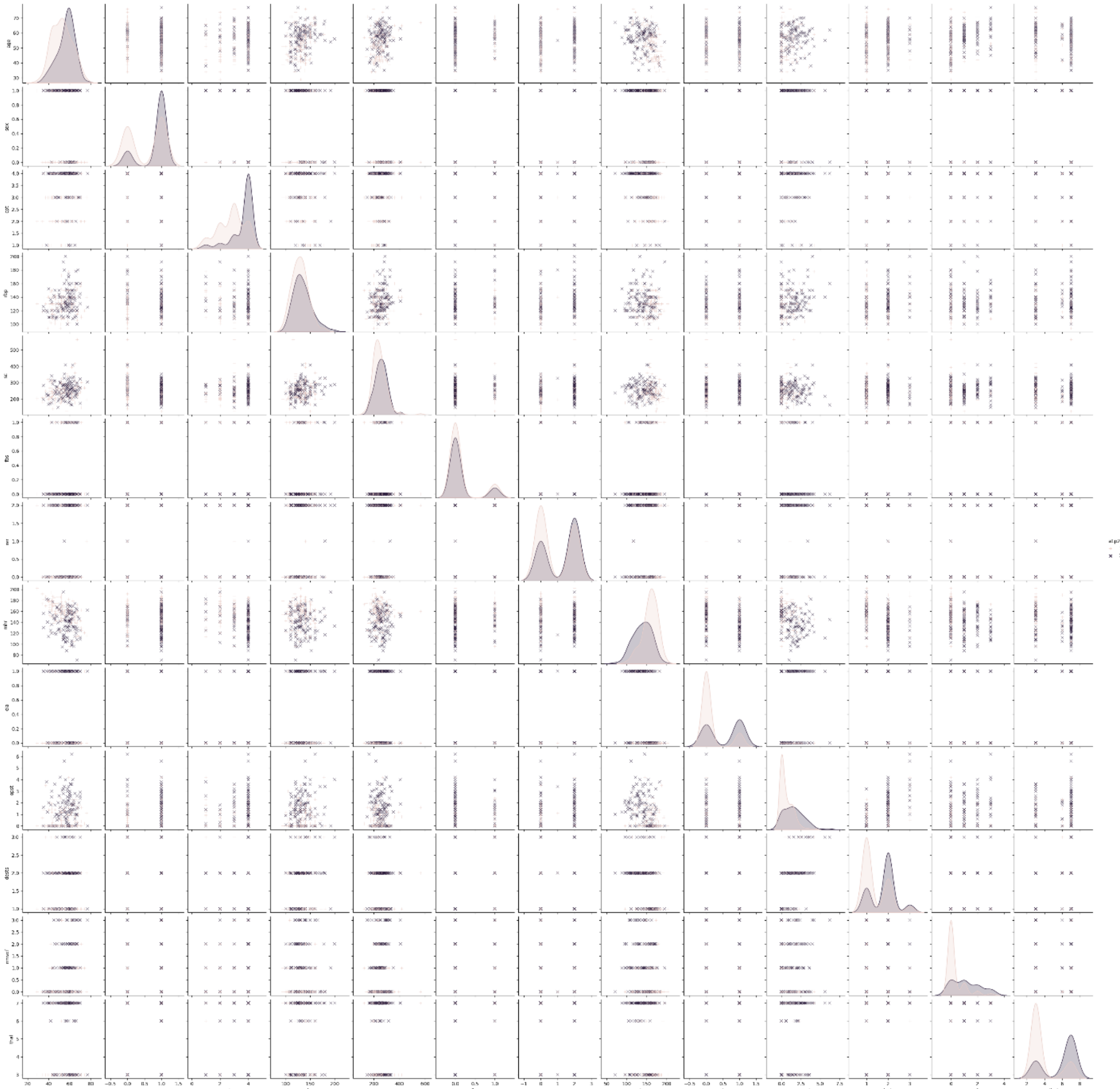
5. Highest correlation of every variable present in the dataset

```
Highest Correlation of each variable
   Variable 1 Variable 2  Correlation
0         age        mhr        -0.40
1         sex       thal         0.39
2         cpt       a1p2         0.42
3         rbp        age         0.27
4          sc        age         0.22
5         fbs        rbp         0.16
6         rer       a1p2         0.18
7         mhr       a1p2        -0.42
8         eia       a1p2         0.42
9        opst      dests         0.61
10      dests       opst         0.61
11      nmvcf       a1p2         0.46
12       thal       a1p2         0.53
13       a1p2       thal         0.53
```

Conclusion:

1. Based on the covariance matrix, it can be concluded that there is a high covariance between the variables **sc, rbp, age** indicating that these variables are not independent of each other.
2. Furthermore, these variables have the highest covariance with the target variable (heart disease), suggesting that they are the best predictors of heart disease.
3. **thal** has the highest correlation with a1p2
4. **opst** and **dests** variables are highly correlated to each other.

Pair Plot



Problem Statement2: Split your heart1.cvs data into training and test datasets. Use every method specified below. Create a report containing a table where you compare prediction percentages and based on this data. Choose the best method of predicting heart disease on this database.

The data was split into 70% training data and 30% test data. Six different machine algorithms

Output from Problem2.py

Table of Prediction Percentages

	Test Accuracy in %	Combined Accuracy in %
Perceptron	85.19	81.85
Logistic Regression	86.42	83.70
Support Vector	87.65	84.44
Decision Tree	74.07	86.67
Random Forest	77.78	92.59
K Nearest Neighbor	71.60	77.78

Conclusion:

The data of heart disease was split into 70% training data and 30% test data. The table presents the test and combined accuracy scores of six machine learning algorithms on this dataset.

The Support Vector algorithm achieved the highest test accuracy score of 87.65%. The Decision Tree algorithm had the lowest test accuracy score of 74.07%.

However, when considering the combined accuracy score, which considers the performance of the algorithm on both the training and testing data, the Random Forest algorithm achieved the highest score of 92.59%. The K Nearest Neighbor algorithm had the second-highest combined accuracy score at 89.26%, despite having the lowest test accuracy score of 64.20%.

These results indicate that the choice of machine learning algorithm can significantly impact the accuracy of the model, and that it is important to consider both test and combined accuracy scores when selecting the best algorithm to predict heart disease.