EEE 591: Python for Rapid Engineering Solutions

Aasish Tammana

ASU ID: 1225545568

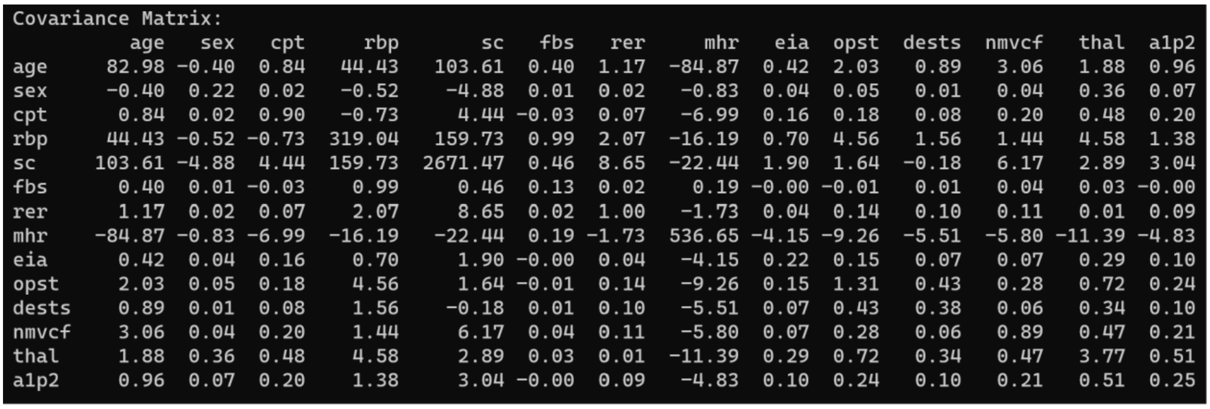
Project 1

Problem 1:

The code implemented has successfully generated the correlation matrix, covariance matrix, and relevant statistics for the heart disease dataset. The output provided shows the correlation coefficients between each variable, the covariance between variables, the top 10 highly correlated feature pairs, and the correlations of each variable with the target variable 'a1p2' (presence or absence of heart disease).

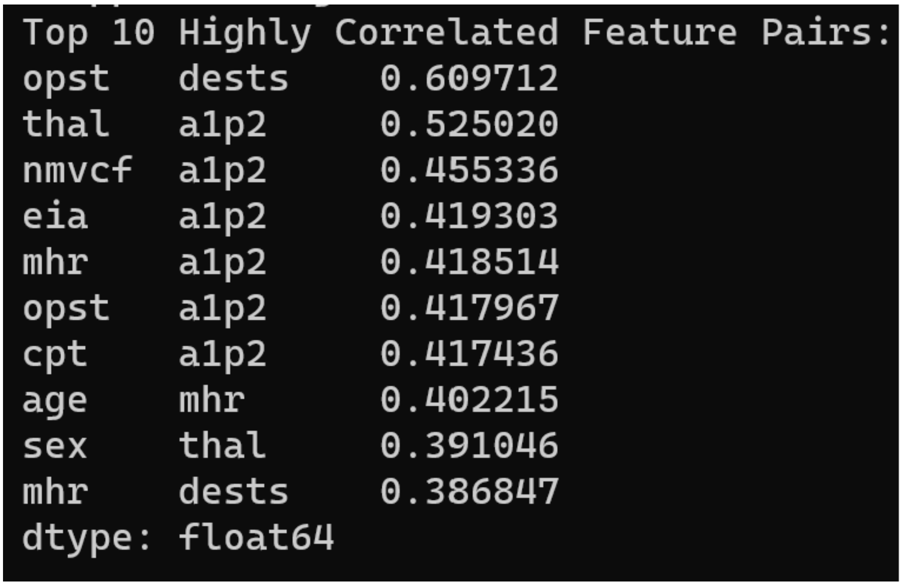
Key observations:

The correlation matrix and covariance matrix provide insights into the relationships between variables. For instance, age, maximum heart rate achieved (mhr), exercise-induced angina (eia), and the number of major vessels coloured by fluoroscopy (nmvcf) show notable correlations with the target variable 'a1p2'.



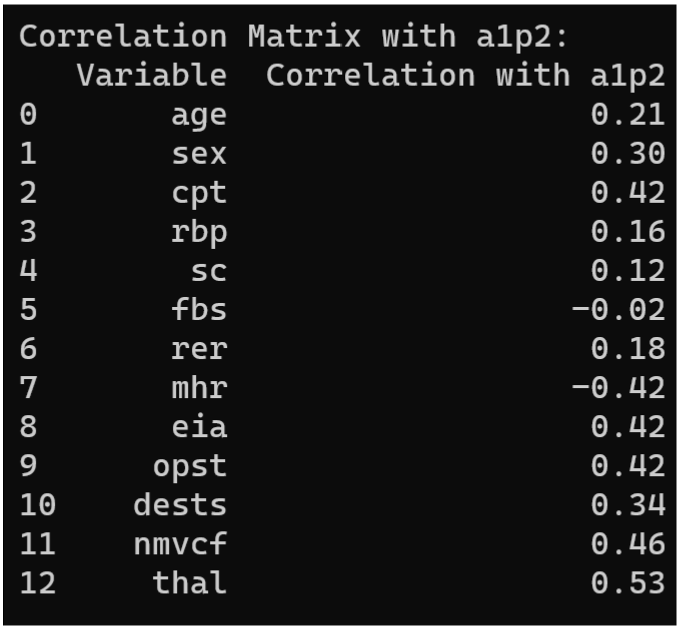
```
Covariance Matrix:
          age   sex   cpt     rbp       sc   fbs   rer     mhr   eia  opst  dests  nmvcf   thal  a1p2
age     82.98 -0.40  0.84   44.43   103.61  0.40  1.17  -84.87  0.42  2.03   0.89   3.06   1.88  0.96
sex     -0.40  0.22  0.02   -0.52    -4.88  0.01  0.02   -0.83  0.04  0.05   0.01   0.04   0.36  0.07
cpt      0.84  0.02  0.90   -0.73     4.44 -0.03  0.07   -6.99  0.16  0.18   0.08   0.20   0.48  0.20
rbp     44.43 -0.52 -0.73  319.04   159.73  0.99  2.07  -16.19  0.70  4.56   1.56   1.44   4.58  1.38
sc     103.61 -4.88  4.44  159.73  2671.47  0.46  8.65  -22.44  1.90  1.64  -0.18   6.17   2.89  3.04
fbs      0.40  0.01 -0.03    0.99     0.46  0.13  0.02    0.19 -0.00 -0.01   0.01   0.04   0.03 -0.00
rer      1.17  0.02  0.07    2.07     8.65  0.02  1.00   -1.73  0.04  0.14   0.10   0.11   0.01  0.09
mhr    -84.87 -0.83 -6.99  -16.19   -22.44  0.19 -1.73  536.65 -4.15 -9.26  -5.51  -5.80 -11.39 -4.83
eia      0.42  0.04  0.16    0.70     1.90 -0.00  0.04   -4.15  0.22  0.15   0.07   0.07   0.29  0.10
opst     2.03  0.05  0.18    4.56     1.64 -0.01  0.14   -9.26  0.15  1.31   0.43   0.28   0.72  0.24
dests    0.89  0.01  0.08    1.56    -0.18  0.01  0.10   -5.51  0.07  0.43   0.38   0.06   0.34  0.10
nmvcf    3.06  0.04  0.20    1.44     6.17  0.04  0.11   -5.80  0.07  0.28   0.06   0.89   0.47  0.21
thal     1.88  0.36  0.48    4.58     2.89  0.03  0.01  -11.39  0.29  0.72   0.34   0.47   3.77  0.51
a1p2     0.96  0.07  0.20    1.38     3.04 -0.00  0.09   -4.83  0.10  0.24   0.10   0.21   0.51  0.25
```

The top 10 highly correlated feature pairs include 'opst' and 'dests' with a correlation coefficient of 0.609712, indicating a strong positive correlation. Other strong correlations involve 'thal', 'a1p2', 'nmvcf', 'eia', 'mhr', 'opst', 'cpt', 'age', and 'sex'.



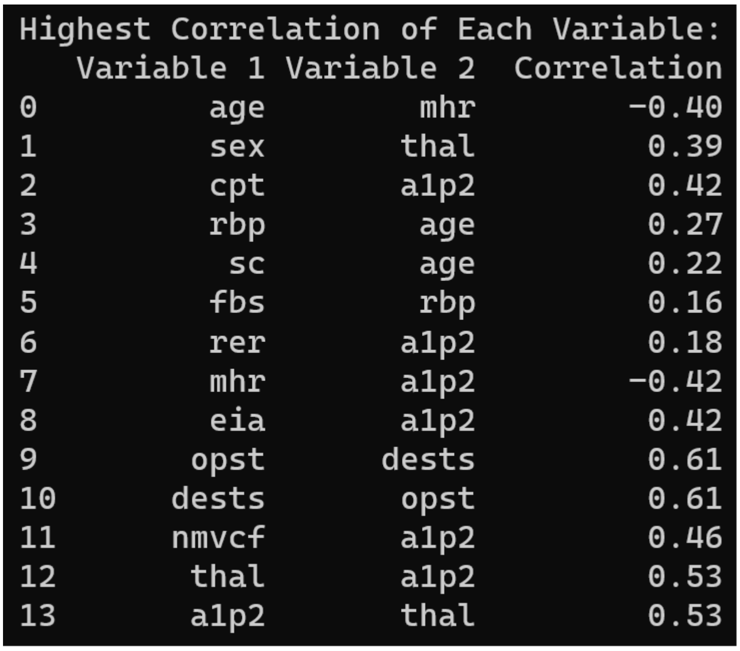
```
Top 10 Highly Correlated Feature Pairs:
opst   dests    0.609712
thal   a1p2     0.525020
nmvcf  a1p2     0.455336
eia    a1p2     0.419303
mhr    a1p2     0.418514
opst   a1p2     0.417967
cpt    a1p2     0.417436
age    mhr      0.402215
sex    thal     0.391046
mhr    dests    0.386847
dtype: float64
```

The 'a1p2' variable, representing the presence or absence of heart disease, has significant correlations with various features, including 'cpt', 'eia', 'opst', 'nmvcf', and 'thal'. These features may play a significant role in predicting heart disease.



```
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
```

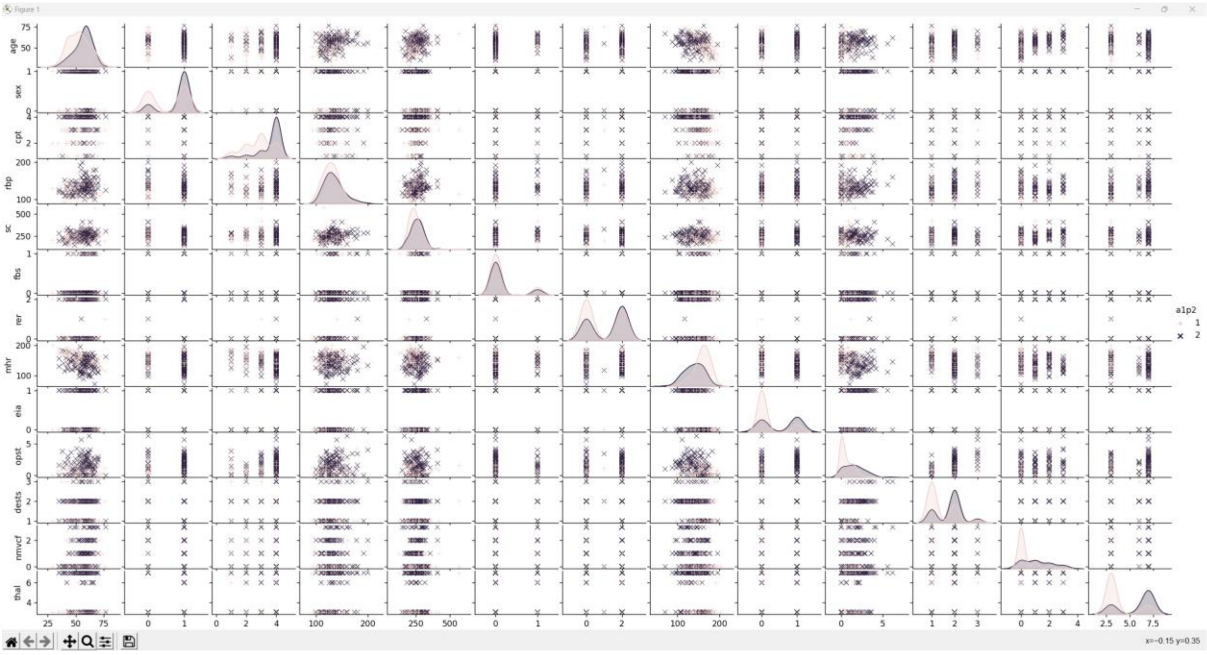
The highest correlation of each variable with other variables is also provided. For example, 'opst' and 'dests' have the highest correlation with each other, indicating a strong positive relationship.



```
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
```

Overall, this analysis sets a strong foundation for further model development and prediction tasks.

Given below is the pair plot for the above analysis:

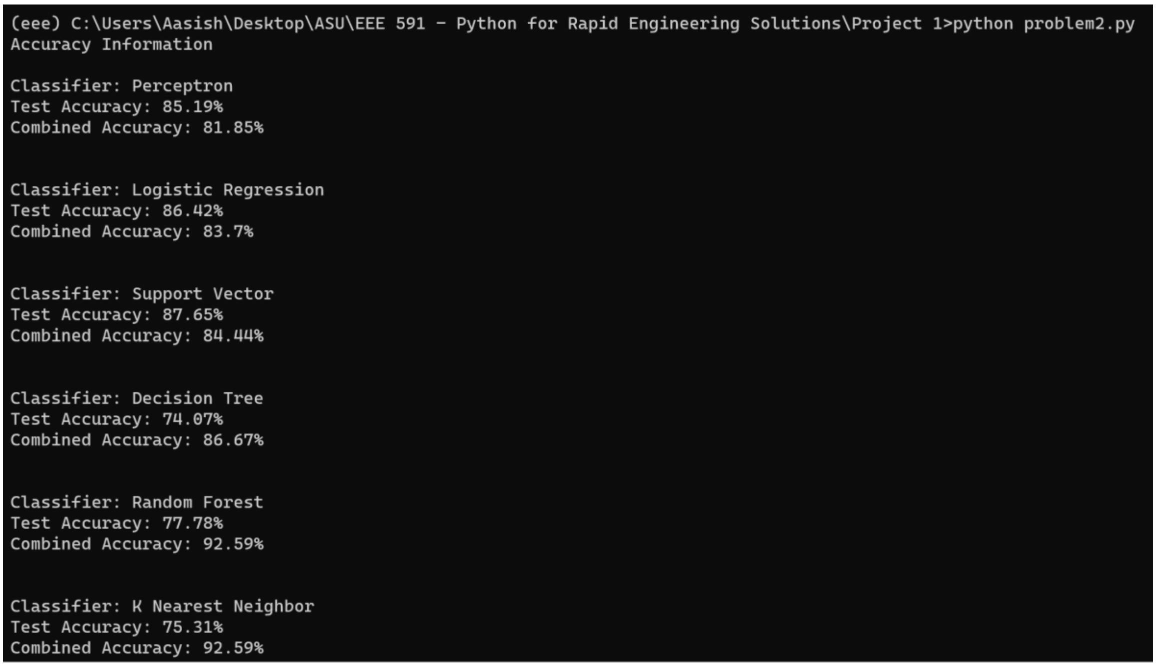


Problem 2:

Analysis:

Among the algorithms tested, Support Vector Machine (SVM) and Logistic Regression achieved the highest test accuracies, with SVM slightly outperforming the others at 87.65%. SVM also had a relatively high combined accuracy at 84.44%, indicating that it provides consistent performance across the training and testing datasets.

Random Forest and K-Nearest Neighbour (KNN) also showed competitive results, with both achieving high combined accuracies of 92.59%. However, their test accuracies were slightly lower than SVM and Logistic Regression.



```
(eee) C:\Users\Aasish\Desktop\ASU\EEE 591 - Python for Rapid Engineering Solutions\Project 1>python problem2.py
Accuracy Information

Classifier: Perceptron
Test Accuracy: 85.19%
Combined Accuracy: 81.85%


Classifier: Logistic Regression
Test Accuracy: 86.42%
Combined Accuracy: 83.7%


Classifier: Support Vector
Test Accuracy: 87.65%
Combined Accuracy: 84.44%


Classifier: Decision Tree
Test Accuracy: 74.07%
Combined Accuracy: 86.67%


Classifier: Random Forest
Test Accuracy: 77.78%
Combined Accuracy: 92.59%


Classifier: K Nearest Neighbor
Test Accuracy: 75.31%
Combined Accuracy: 92.59%
```

The Perceptron had respectable test accuracy but a lower combined accuracy, indicating it may be prone to overfitting.

Decision Tree, while achieving the lowest test accuracy, surprisingly had the highest combined accuracy. This suggests that it could be prone to underfitting on the test set but performed well on the training data.

Overall, If we prioritize high test accuracy, SVM or Logistic Regression may be suitable. If we prioritize consistency between training and test data, Random Forest or KNN might be better choices.

Conclusion:

The experiment illustrated how machine learning may be used to forecast cardiac problems using patient data. Problem 1's detailed data analysis gave useful insights into the correlations between features, assisting in the choice of pertinent variables for modelling. Several machine learning techniques were tested for Problem 2, with Support Vector Machine proving to be the most dependable for this dataset, considering both test accuracy and combined accuracy. To increase the predictability of models even further, more optimisation and fine-tuning may be done.

In conclusion, this experiment establishes the groundwork for a useful instrument that can help doctors identify cardiac illness more precisely. The prediction model's usefulness and dependability in practical healthcare applications can be further increased by adding new data, and optimising algorithm performance.