

EEE 591

Project 1

Trisha Ashok

1226289911

AIM:

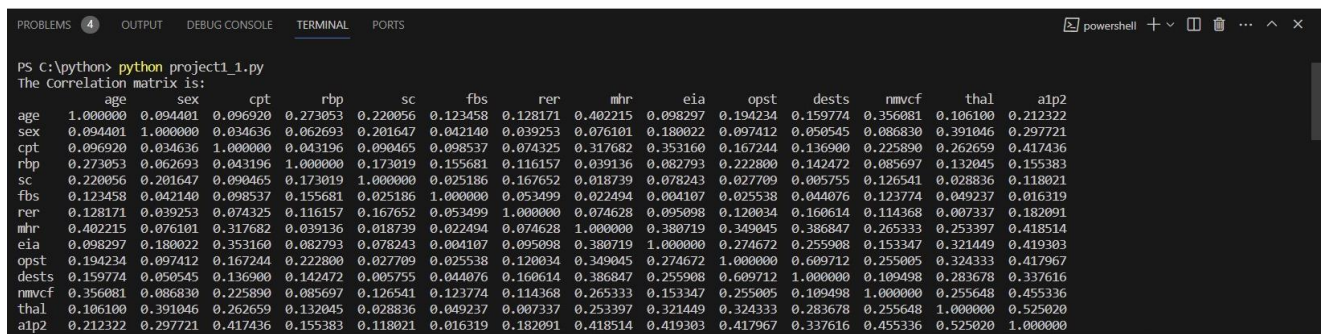
To obtain the correlation and covariance for each variable in the dataset pertaining to the heart disease prediction for the doctors from AMAPE. Also, to determine the highest achievable accuracy for predicting heart disease using machine learning algorithms.

The Machine Learning Algorithms to be used:

- Perceptron
- Logistic Regression
- Support Vector Machine
- Decision Tree
- Rain forest
- K Nearest Neighbor

Problem1:

Correlation matrix:



```
PS C:\python> python project1_1.py
The Correlation matrix is:
age      sex      cpt      rbp      sc      fbs      rer      mhr      eia      opst      dests      nmvcf      thal      a1p2
age      1.000000  0.094401  0.096920  0.273853  0.228056  0.123458  0.128171  0.402215  0.098297  0.194234  0.159774  0.356081  0.106100  0.212322
sex      0.094401  1.000000  0.034636  0.062693  0.201647  0.042140  0.039253  0.076101  0.180022  0.097412  0.050545  0.086830  0.391046  0.297721
cpt      0.096920  0.034636  1.000000  0.043196  0.090465  0.098537  0.074325  0.317682  0.353160  0.167244  0.136900  0.225890  0.262659  0.417436
rbp      0.273853  0.062693  0.043196  1.000000  0.173019  0.155681  0.116157  0.039136  0.082793  0.222800  0.142472  0.085697  0.132045  0.155383
sc      0.228056  0.201647  0.090465  0.173019  1.000000  0.025186  0.167652  0.018739  0.078243  0.027709  0.005755  0.126541  0.028836  0.118021
fbs      0.123458  0.042140  0.098537  0.155681  0.025186  1.000000  0.053499  0.022494  0.004187  0.025538  0.044076  0.123774  0.049237  0.016319
rer      0.128171  0.039253  0.074325  0.116157  0.167652  0.053499  1.000000  0.074628  0.095098  0.120034  0.160614  0.114368  0.007337  0.182091
mhr      0.402215  0.076101  0.317682  0.039136  0.018739  0.022494  0.074628  1.000000  0.380719  0.349045  0.386847  0.265333  0.253397  0.418514
eia      0.098297  0.180022  0.353160  0.082793  0.078243  0.004187  0.095098  0.380719  1.000000  0.274672  0.255908  0.153347  0.321449  0.419303
opst     0.194234  0.097412  0.167244  0.222800  0.027709  0.025538  0.120034  0.349045  0.274672  1.000000  0.609712  0.255005  0.324333  0.417967
dests    0.159774  0.050545  0.136900  0.142472  0.005755  0.044076  0.160614  0.386847  0.255908  0.609712  1.000000  0.109498  0.283678  0.337616
nmvcf    0.356081  0.086830  0.225890  0.085697  0.126541  0.123774  0.114368  0.265333  0.153347  0.255005  0.109498  1.000000  0.255648  0.455336
thal     0.106100  0.391046  0.262659  0.132045  0.028836  0.049237  0.007337  0.253397  0.321449  0.324333  0.283678  0.255648  1.000000  0.525020
a1p2     0.212322  0.297721  0.417436  0.155383  0.118021  0.016319  0.182091  0.418514  0.419303  0.417967  0.337616  0.455336  0.525020  1.000000
```

Result:

It can be observed from the dataset that the **a1p2**(absence and presence of a heart disease) is highly correlated with **thal**. Also, there is a slight dependency of **mhr,eia, opst,cpt** with **a1p2** and all of them almost has same correlation. In contrast, any changes in **fbs** will not affect **a1p2** since they are independent to each other.

Covariance matrix:

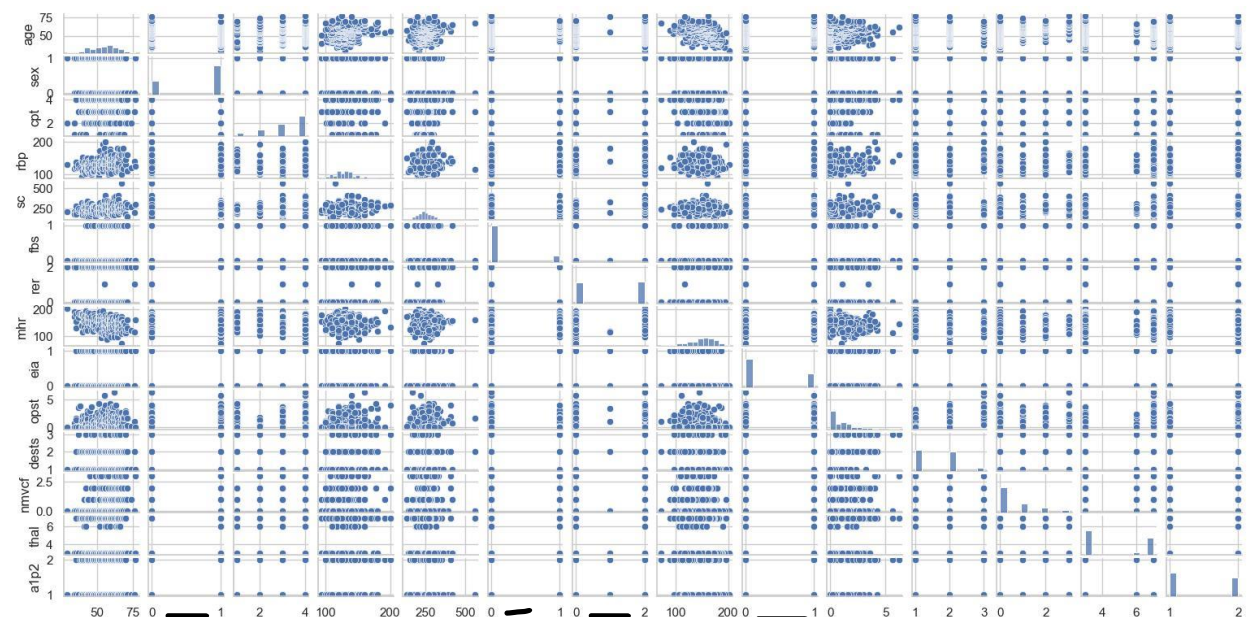
Covariance matrix is:

	age	sex	cpt	rbp	sc	fbs	rer	mhr	eia	opst	dests	nmvcf	thal	aip2
age	0.0	0.402602	0.838786	44.426394	103.605452	0.400248	1.165056	84.874721	0.421685	2.026208	0.894176	3.061586	1.875589	0.962825
sex	0.0	0.000000	0.015407	0.524287	4.879719	0.007022	0.018340	0.825403	0.039694	0.052230	0.014539	0.038373	0.355308	0.069393
cpt	0.0	0.000000	0.000000	0.733044	4.442434	0.033320	0.070467	6.992028	0.158020	0.181970	0.079912	0.202575	0.484290	0.107439
rbp	0.0	0.000000	0.000000	0.000000	159.731185	0.989674	2.070384	16.193432	0.696448	4.557435	1.563486	1.444816	4.577117	1.381660
sc	0.0	0.000000	0.000000	0.000000	0.000000	0.463307	8.647005	22.437340	1.904557	1.640149	0.182762	6.173510	2.892414	3.036762
fbs	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.019000	0.185461	0.000688	0.010409	0.009638	0.041581	0.034008	0.002891
rer	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.725155	0.044692	0.137175	0.098472	0.107724	0.014209	0.090458
mhr	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	4.153614	9.260037	5.505907	5.801776	11.391904	4.826518
eia	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.148141	0.074047	0.068167	0.293790	0.098306
opst	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.428996	0.275651	0.720818	0.238290
dests	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.063500	0.338235	0.103263
nmvcf	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.468291	0.213961
thal	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.507228
aip2	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Result:

It can be observed from the above matrix that the prediction of heart disease is has highest covariance with **sc**. Also, **thal** has slightly measurable covariance for prediction. In contrast, **sex** doesn't matter for this prediction, so it has zero covariance.

Pair plot:



Result:

All the correlations can be observed here with the pair plot. This shows all the best correlated predictions for heart disease. More information can be obtained from **aip1** with those that are underlined because the data is not overlapped with each other. At some point, it is very difficult to obtain the information as the data seems to be overlapped too much.

Problem2:

Algorithm	Accuracy	Combined Accuracy
Perceptron	75%	80%
Logistic Regression	77%	86%
SVM	75%	86%
Decision Tree	77%	92%
Random Forest	75%	93%
K Nearest Neighbor	73%	92%

Result:

Based on the observations from the table above, it appears that RANDOM FOREST is the most accurate Machine Learning Algorithm for heart disease. So, it will predict the heart disease for a person accurately using the app. Other models are comparatively producing less accurate results for the doctors but still they can be used.

Remarks:

Hence the correlation and covariance between each and every variable from the data set has been predicted along with the accuracy of the disease using the algorithms.