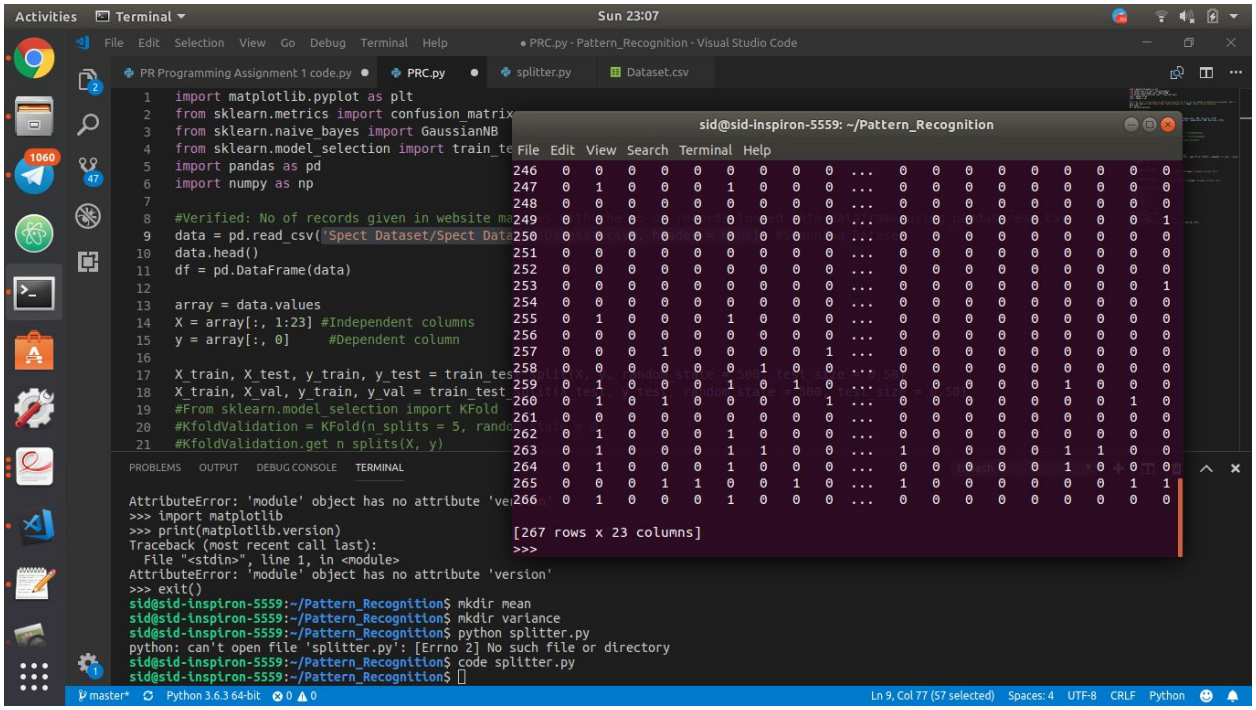
Pattern Recognition Assignment-1

Task 1

Dataset Background:

The dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images.
As a result, 44 continuous feature pattern was created for each patient. The pattern was further processed to obtain 22 binary feature patterns representing partial diagnosis and 1 binary feature representing an overall diagnosis. The first column depends on the remaining 22 columns and they are all categorically classified with values of 0 or 1.



Cross-Validation:

A resampling procedure used to evaluate machine learning models on a limited data sample. There are several types of Cross Validation, the most common one being the ‘K-Fold’ Cross-Validation.

Splitting dataset:

In order to split the data such that 50% is contained in the training set, 25% in validation set and 25% is in the test set, I am using a script that:

1. Couples the .test and .train data into one file
2. Takes 50% from new dataset randomly for training
3. Takes 25% from remaining rows in the test set and validation set.

Usage and Syntax in Python:

from sklearn.model_selection import train_test_split

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state =
500, test_size = 0.50)
To split the datset by 50%.
```

Precision: the proportion of the positive identifications which are actually positive.

Recall: Recall is the number of true positives divided by the sum of the total number of true positives and false negatives.

recall = (TP/TP+FN)

Sensitivity: Proportion of the actual positives identified correctly(used in medical terminology)

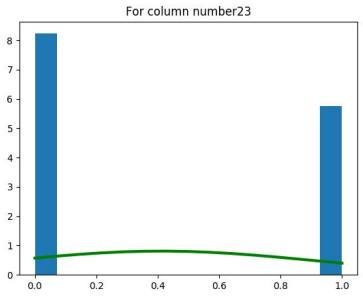
Specificity: Proportion of actual negatives that are actually identified (medical terminology)

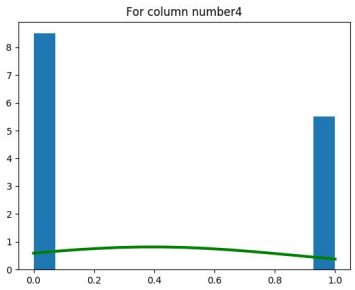
F Score: Measure of test's accuracy

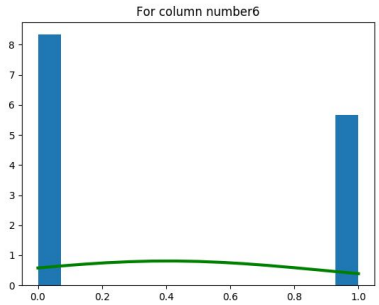
MCC - This is used to measure the quality of binary classifications. This coefficient returns a value between -1 and +1. If the return value is +1then the model is predicting perfectly and if the return value is -1 then the model is total disagreement between the prediction and the original observation. MCC = (TP * TN - FP * FN)/(sqrt((TP + FP)*(TP + FN)*(TN + FN)*(TN + FP)))

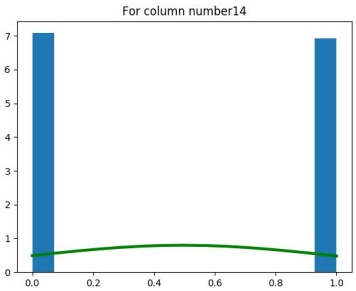
Task 2

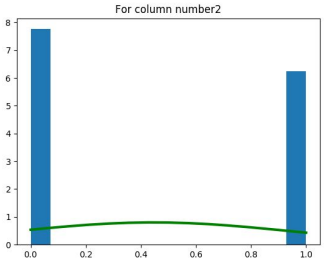
Features that roughly obey Gaussian Distribution are:
2, 4, 6, 9, 14, 23.
Below are the Plots for the above 5 features.



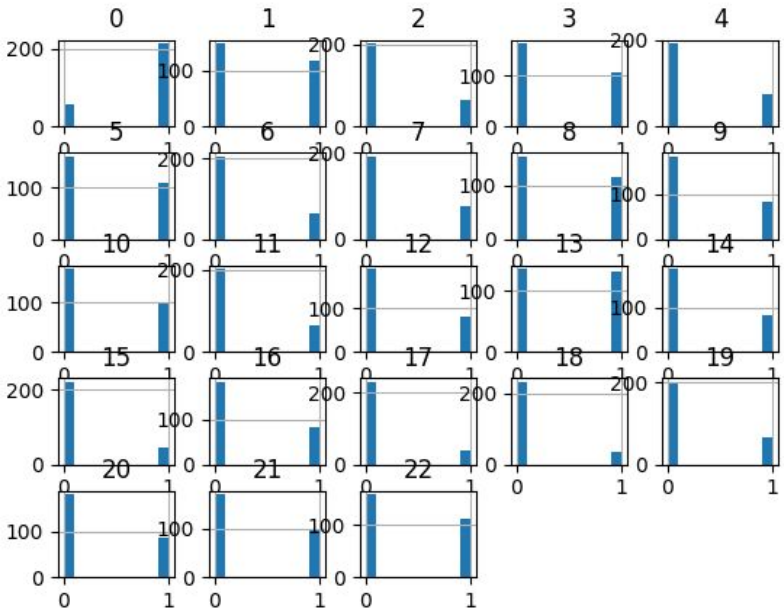




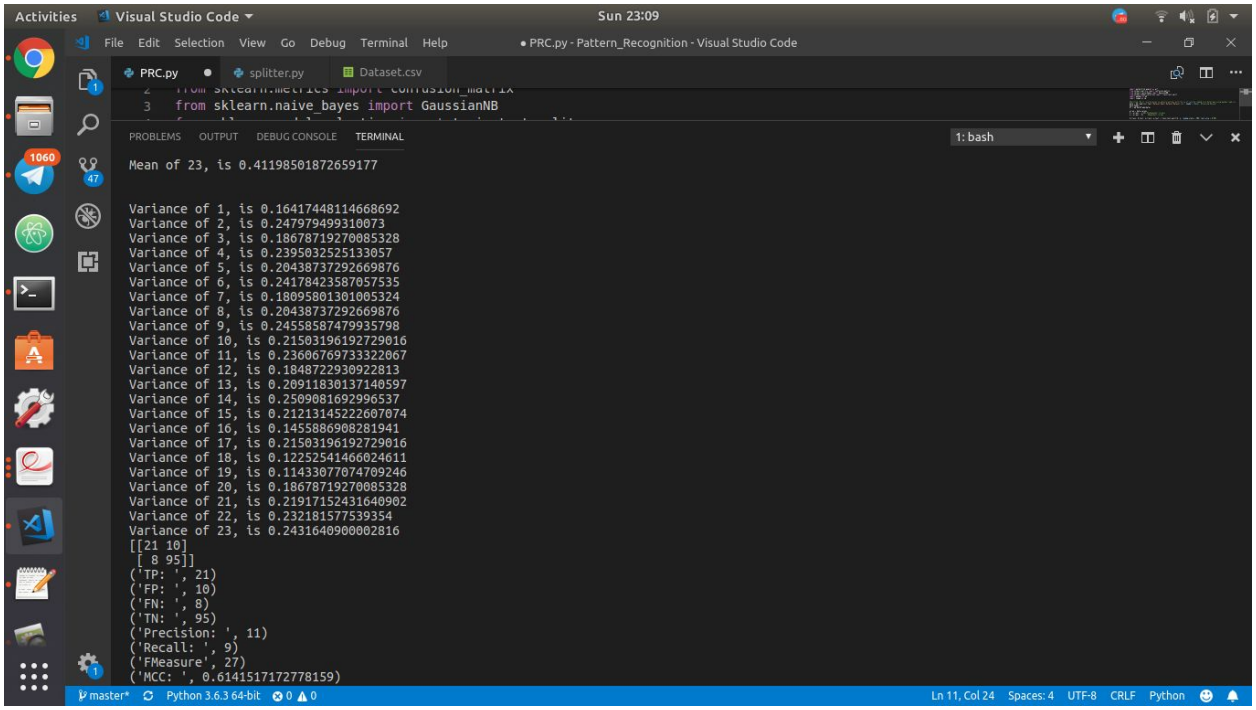


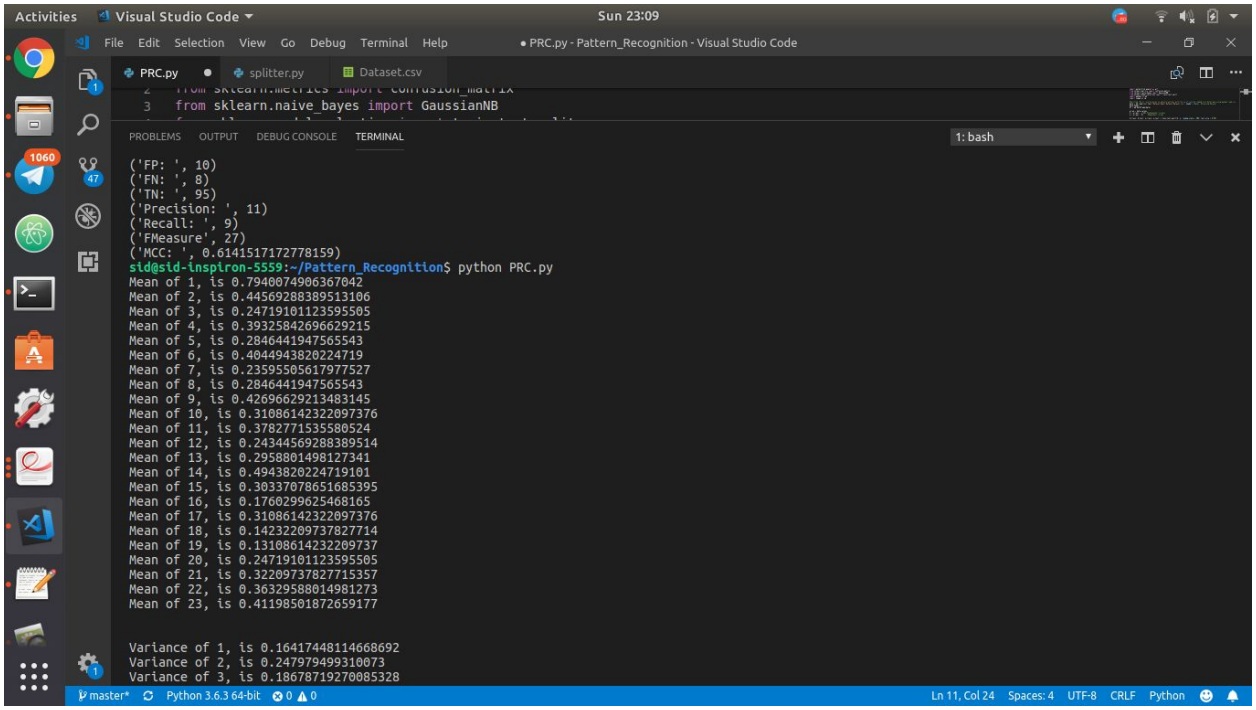


The following are the plots of all the features:



The mean and variance of each feature is shown below along with the factors mentioned in Task 1


Activities
Visual Studio Code
Sun 23:09
File Edit Selection View Go Debug Terminal Help
• PRC.py - Pattern_Recognition - Visual Studio Code
PRC.py
splitter.py
Dataset.csv
from sklearn.naive_bayes import GaussianNB
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
1: bash
Mean of 23, is 0.41198501872659177
Variance of 1, is 0.16417448114668692
Variance of 2, is 0.247979499310073
Variance of 3, is 0.18678719270085328
Variance of 4, is 0.2395032525133057
Variance of 5, is 0.20438737292669876
Variance of 6, is 0.24178423587057535
Variance of 7, is 0.18095801301005324
Variance of 8, is 0.20438737292669876
Variance of 9, is 0.24558587479935798
Variance of 10, is 0.21503196192729016
Variance of 11, is 0.23606769733322067
Variance of 12, is 0.1848722930922813
Variance of 13, is 0.20911830137140597
Variance of 14, is 0.2509081692996537
Variance of 15, is 0.21213145222607074
Variance of 16, is 0.1455886908281941
Variance of 17, is 0.21503196192729016
Variance of 18, is 0.12252541466024611
Variance of 19, is 0.11433077074709246
Variance of 20, is 0.18678719270085328
Variance of 21, is 0.21917152431640902
Variance of 22, is 0.23218157539354
Variance of 23, is 0.2431640900002816
[[21 10]
[8 95]]
('TP: ', 21)
('FP: ', 10)
('FN: ', 8)
('TN: ', 95)
('Precision: ', 11)
('Recall: ', 9)
('FMeasure', 27)
('MCC: ', 0.6141517172778159)
master*
Python 3.6.3 64-bit
0 0
Ln 11, Col 24
Spaces: 4
UTF-8
CRLF
Python


Activities
Visual Studio Code
Sun 23:09
File Edit Selection View Go Debug Terminal Help
• PRC.py - Pattern_Recognition - Visual Studio Code
PRC.py
splitter.py
Dataset.csv
from sklearn.naive_bayes import GaussianNB
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
1: bash
('FP: ', 10)
('FN: ', 8)
('TN: ', 95)
('Precision: ', 11)
('Recall: ', 9)
('FMeasure', 27)
('MCC: ', 0.6141517172778159)
sid@sid-inspiron-5559:~/Pattern_Recognition$ python PRC.py
Mean of 1, is 0.7940074906367042
Mean of 2, is 0.44569288389513106
Mean of 3, is 0.24719101123595505
Mean of 4, is 0.39325842696629215
Mean of 5, is 0.2846441947565543
Mean of 6, is 0.4044943820224719
Mean of 7, is 0.23595505617977527
Mean of 8, is 0.2846441947565543
Mean of 9, is 0.42696629213483145
Mean of 10, is 0.31086142322097376
Mean of 11, is 0.3782771535580524
Mean of 12, is 0.24344569288389514
Mean of 13, is 0.2958801498127341
Mean of 14, is 0.4943820224719101
Mean of 15, is 0.30337078651685395
Mean of 16, is 0.1760299625468165
Mean of 17, is 0.31086142322097376
Mean of 18, is 0.14232209737827714
Mean of 19, is 0.13108614232209737
Mean of 20, is 0.24719101123595505
Mean of 21, is 0.32209737827715357
Mean of 22, is 0.36329588014981273
Mean of 23, is 0.41198501872659177
Variance of 1, is 0.16417448114668692
Variance of 2, is 0.247979499310073
Variance of 3, is 0.18678719270085328
master*
Python 3.6.3 64-bit
0 0
Ln 11, Col 24
Spaces: 4
UTF-8
CRLF
Python