“Machine Learning”

Shervin Halat

98131018

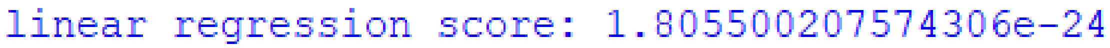
Project

(Final)

1.

a.

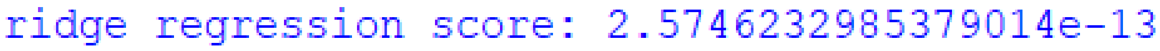
Linear-Regression:



```
linear regression score: 1.805500207574306e-24
```

b.

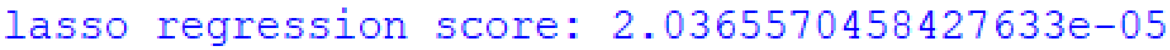
Ridge-Regression:



```
ridge regression score: 2.5746232985379014e-13
```

c.

Lasso-Regression:



```
lasso regression score: 2.0365570458427633e-05
```

d.

Bayesian Regression:

NO RESULTS!!

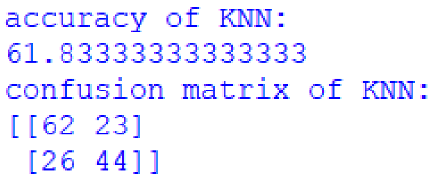
Each of the scores obtained above, are computed using 10-fold-cross-validation and feature selection among 5 to 14 best features of the data. Day data was used for all algorithms.

From the results above, it is obvious that **Lasso-Regression had the gained score.**

2.

a.

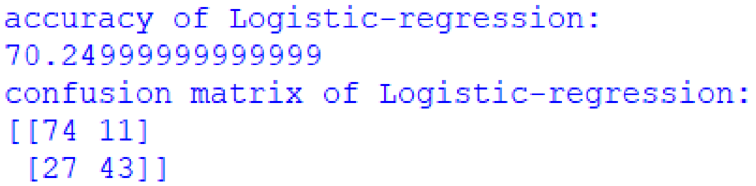
KNN:



```
accuracy of KNN:
61.83333333333333
confusion matrix of KNN:
[[62 23]
 [26 44]]
```

b.

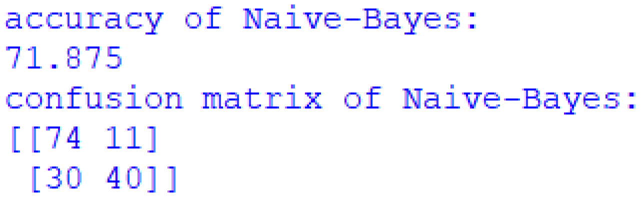
Logistic-Regression:



```
accuracy of Logistic-regression:
70.24999999999999
confusion matrix of Logistic-regression:
[[74 11]
 [27 43]]
```

c.

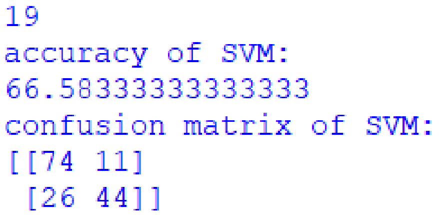
Naïve-Bayes:



```
accuracy of Naive-Bayes:
71.875
confusion matrix of Naive-Bayes:
[[74 11]
 [30 40]]
```

d.

SVM:



```
19
accuracy of SVM:
66.58333333333333
confusion matrix of SVM:
[[74 11]
 [26 44]]
```

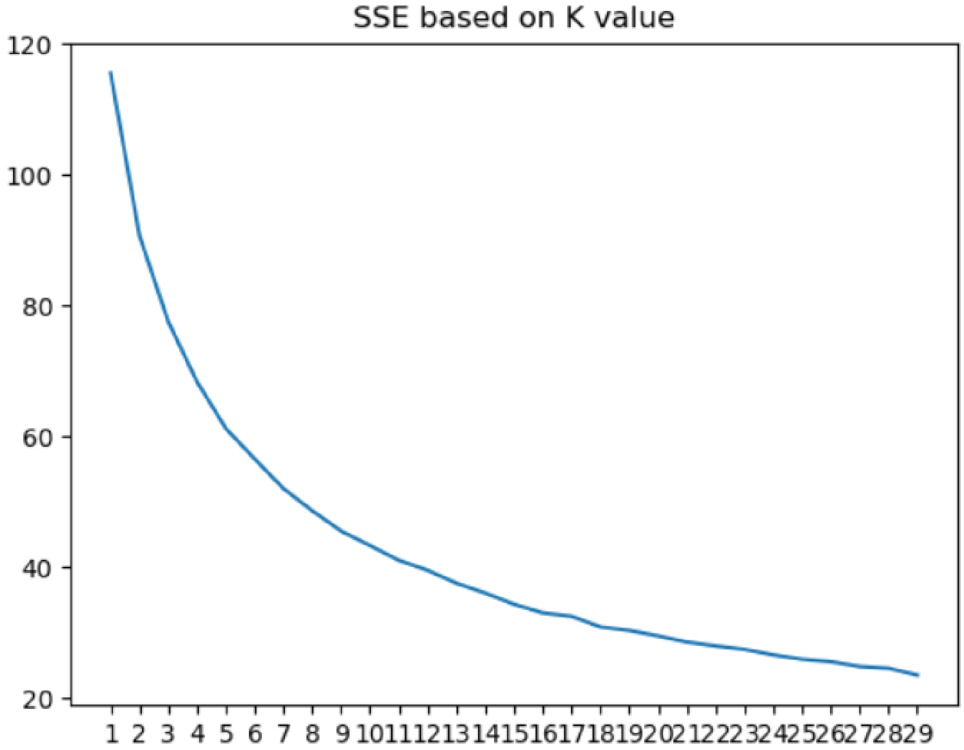
As it is clear from results above, **Naïve Bayes classifier gained the best accuracy.** For each classifier 10 to 20 first best features for selection were examined and accuracies are based on the best feature selections.

3.

a.

K-means:

For k-means, elbow method was applied to find the best k, value but the plot obtained was quite smooth. Therefore, k value of ‘7’ was considered.



The obtained clusters in the order of input data is written in file ‘3.kmeans.csv’.

b.

DBSCAN:

Min-points: As a rule of thumb, a minimum *minPts* can be derived from the number of dimensions *D* in the data set. Hence, minpoints value of ‘5’ was applied here.

ε: The value for ε can then be chosen by using a k-distacnce graph, plotting the distance to the *k* = *minPts*-1 nearest neighbor ordered from the largest to the smallest value.[5] Good values of ε are where this plot shows an "elbow".

The epsilon value was set to ‘0.2’ based on the mean of Euclidean distance between 100 random points with their closest neighbors.

The final labels are stored in file ‘3.dbscan.csv’

c.

Spectral clustering:

For number of clusters from 2 to 11 and defining error index for each clustering based on sum of each cluasters’ datapopints’ absolute distance form cluster’s mean, the number of clusters of ‘ ’ was determined as the best number of clusters. The final labels are stored in file ‘3.spectral.csv’