TIPR Assignment 1

Shilpa K K

S R No: 15598

Python Version: 3.6.5

Task I

Algorithm

1. Generate random matrix on Gaussian distribution with mean=0 and variance=1.
2. Multiply data matrix with random matrix and finally multiply with normalized value.
3. Save the output matrix in output file path.

Results

Implemented Random Projections algorithm to convert the high-dimensional data(K) into lower dimensions K=2, 4,…, (D/2) for all given datasets.

Task II

Designed Bayes classifier and nearest neighbor classifier from scratch. For Bayes classifier, estimated class conditional densities using training data. The prior probabilities are estimated using maximum likelihood approach. Used K=7 in the design of nearest neighbor classifier, though any value can be chosen as per convenience.

Task III

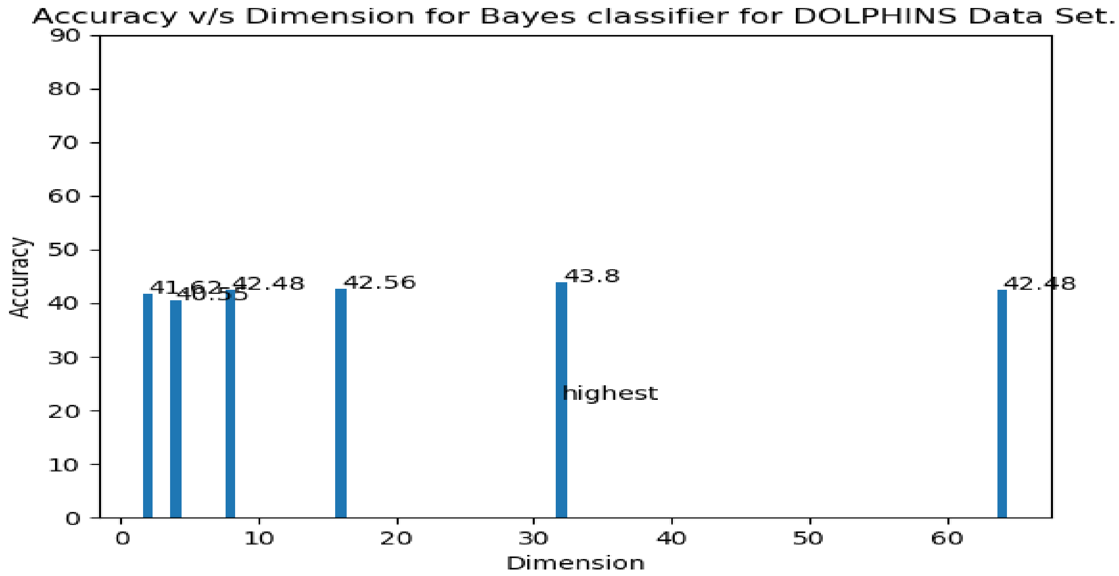
Divided the data (both the original/high and low-dimensional) into train and test set using cross-validation technique. Measured accuracy and F1-score(Macro and Micro) for all the data sets.

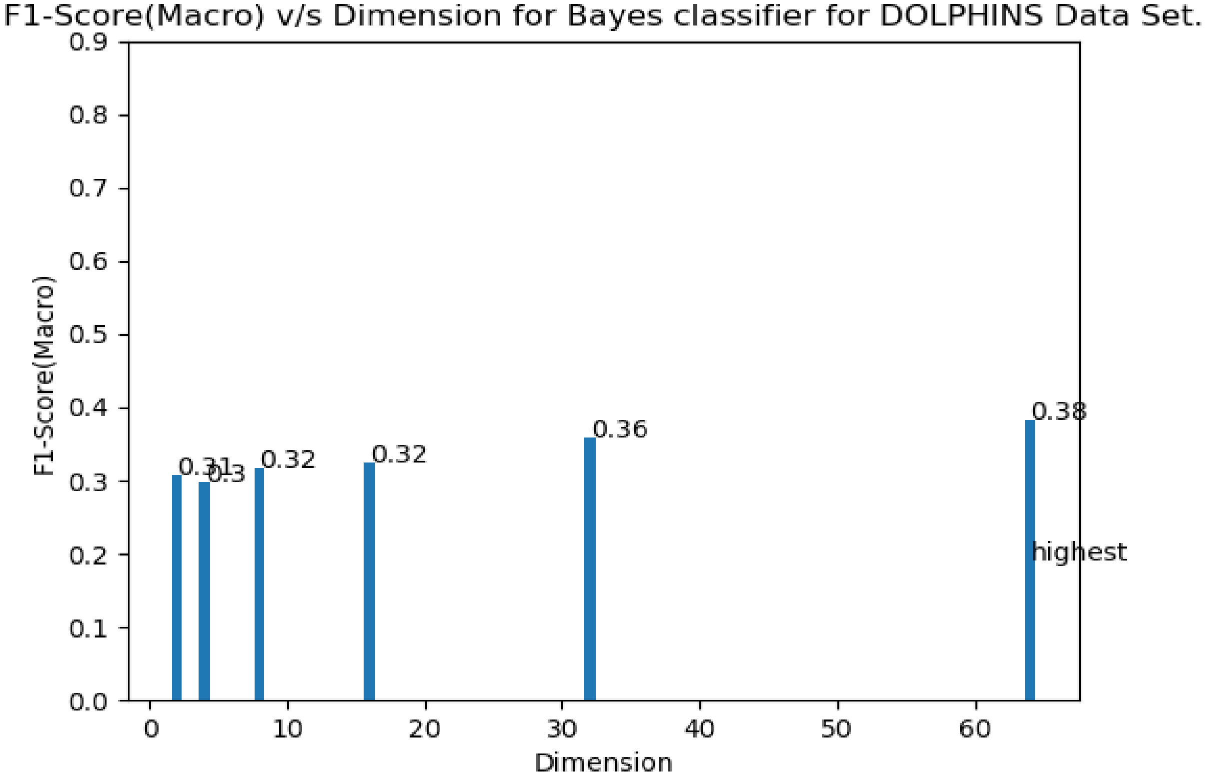
Results

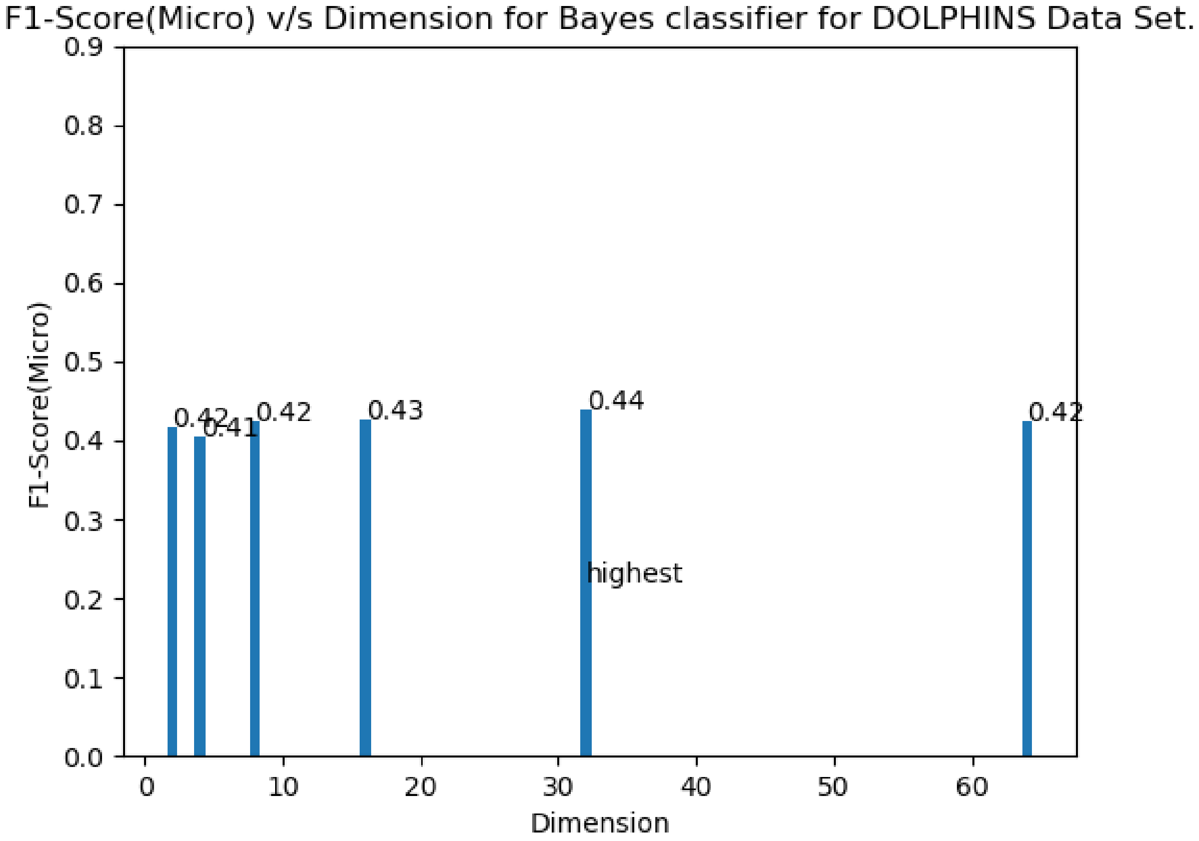
1. Dolphins data set

   a. Bayes Classifier

Dimension, D	2	4	8	16	32	64
Accuracy	41.62	40.55	42.48	42.56	43.80	42.48
F1-score(Macro)	0.307	0.298	0.317	0.324	0.357	0.382
F1-score(Micro)	0.416	0.405	0.425	0.426	0.438	0.425

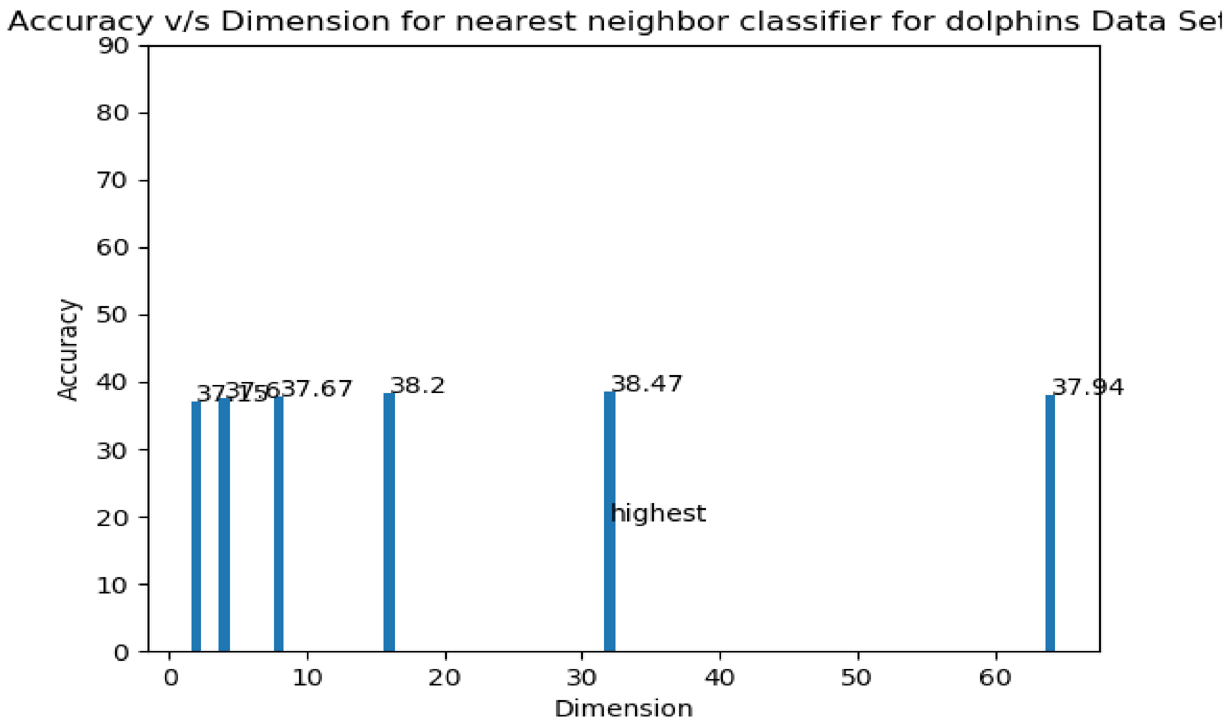

Accuracy v/s Dimension for Bayes classifier for DOLPHINS Data Set.
Accuracy
0
10
20
30
40
50
60
70
80
90
41.62
40.55
42.48
42.56
43.8
highest
42.48
0
10
20
30
40
50
60
Dimension

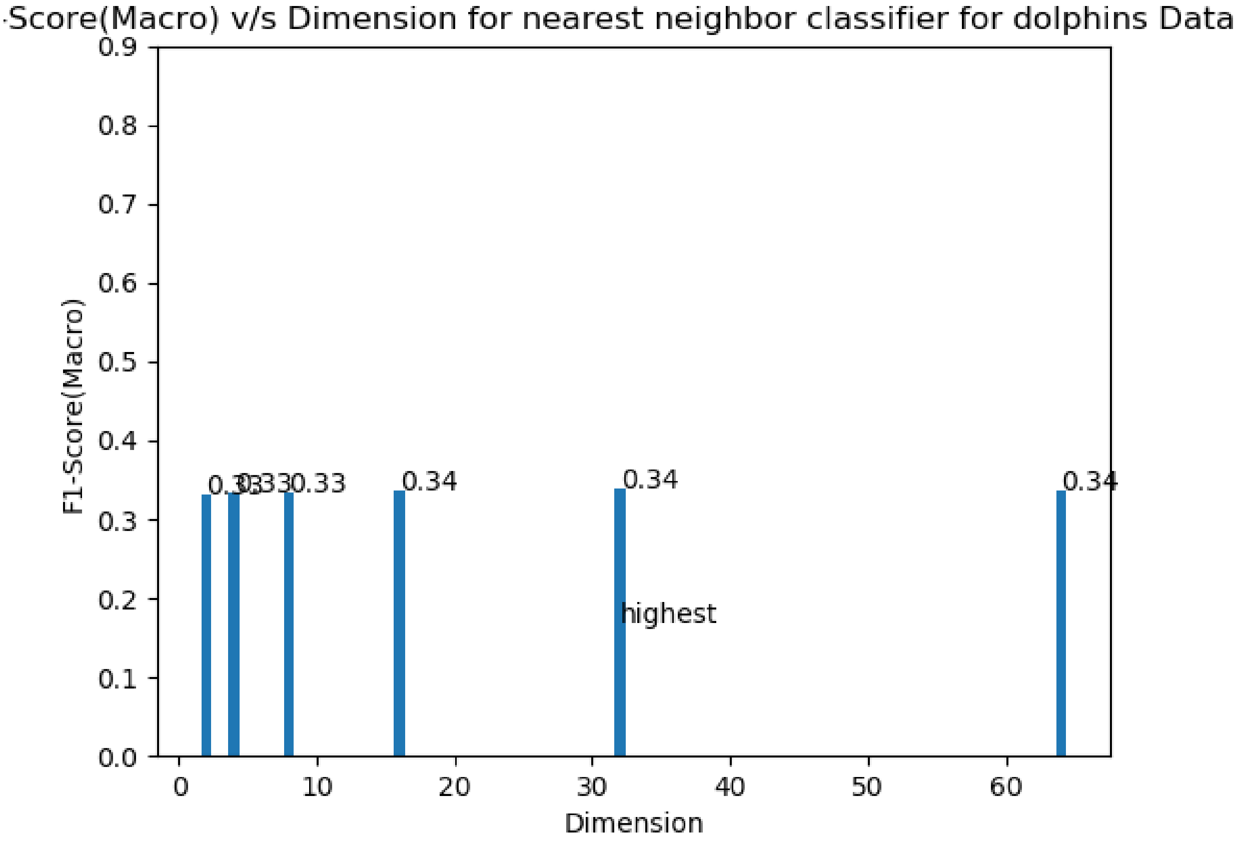

F1-Score(Macro) v/s Dimension for Bayes classifier for DOLPHINS Data Set.
F1-Score(Macro)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.31
0.3
0.32
0.32
0.36
0.38
highest
0
10
20
30
40
50
60
Dimension

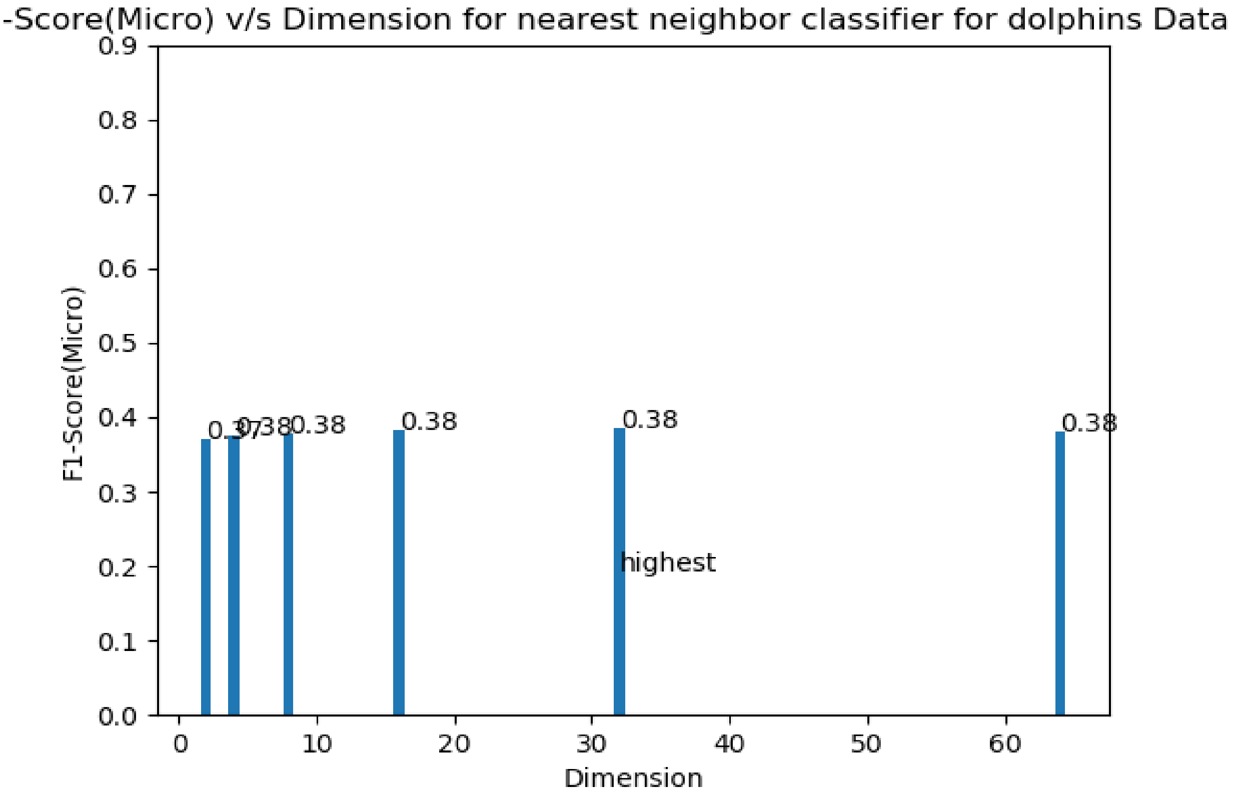


b. Nearest neighbor Classifier

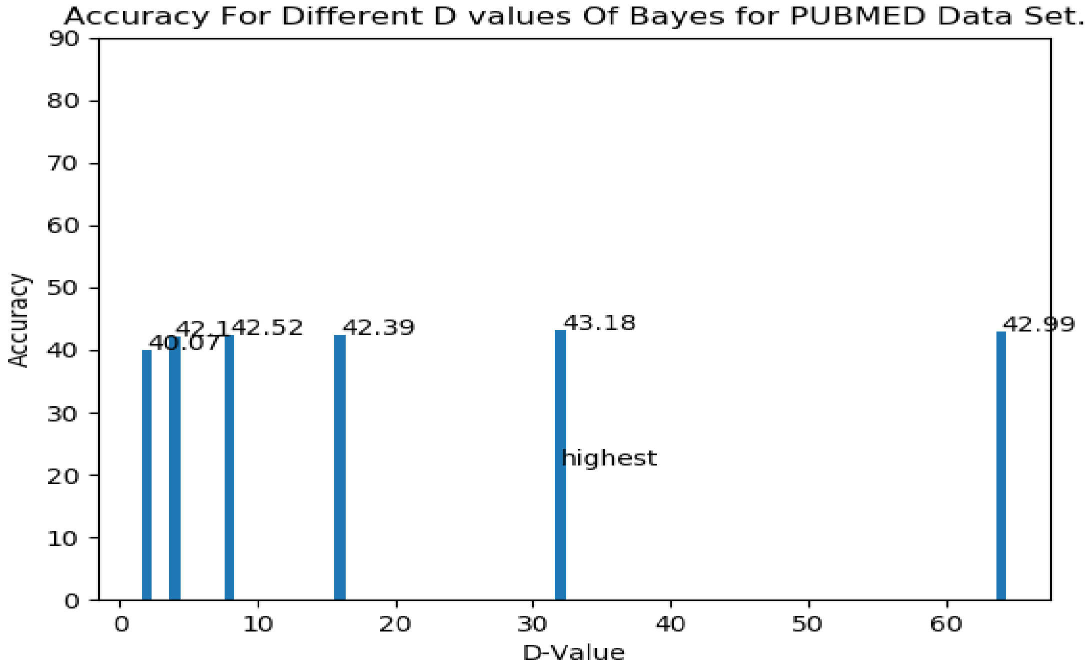
Dimension, D	2	4	8	16	32	64
Accuracy	37.15	37.60	37.67	38.19	38.47	37.94
F1-score(Macro)	0.331	0.334	0.333	0.337	0.338	0.335
F1-score(Micro)	0.372	0.375	0.377	0.382	0.385	0.379

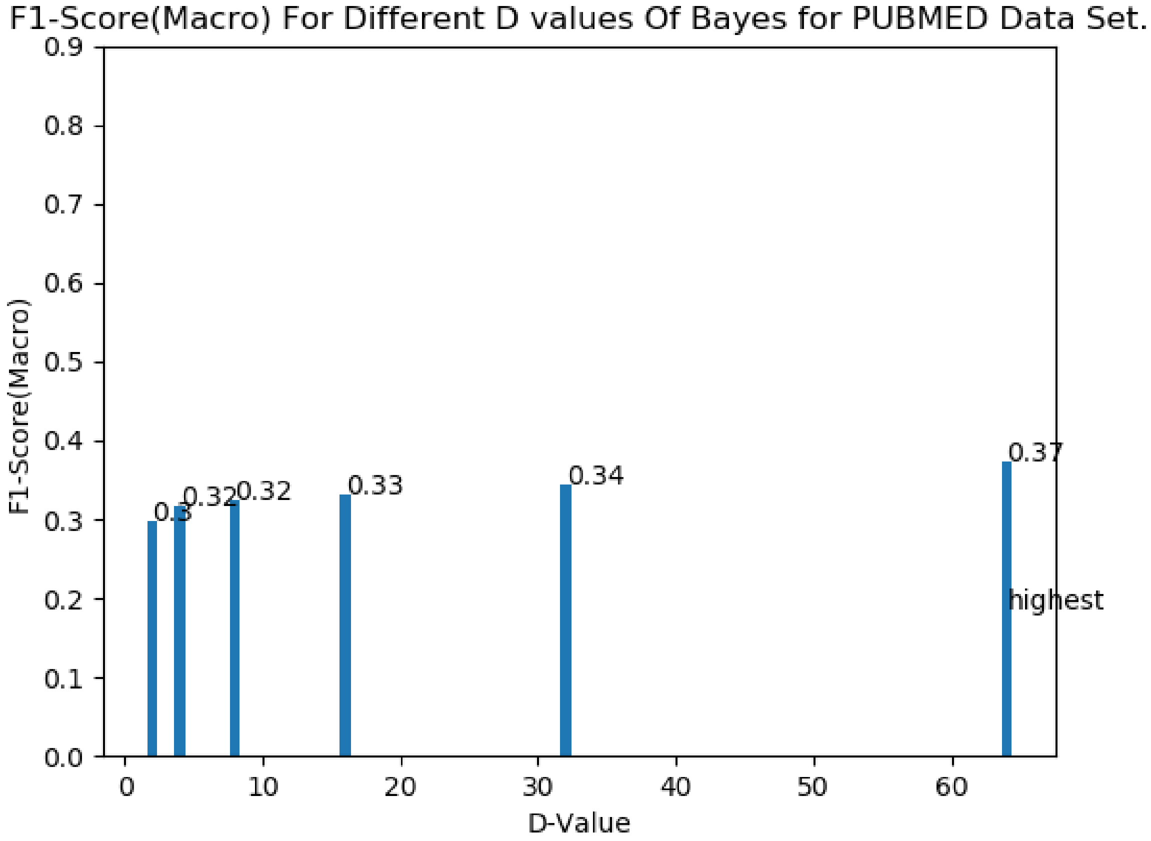

Accuracy v/s Dimension for nearest neighbor classifier for dolphins Data Set
Accuracy
90
80
70
60
50
40
30
20
10
0
37.15
37.56
37.67
38.2
38.47
37.94
highest
0
10
20
30
40
50
60
Dimension

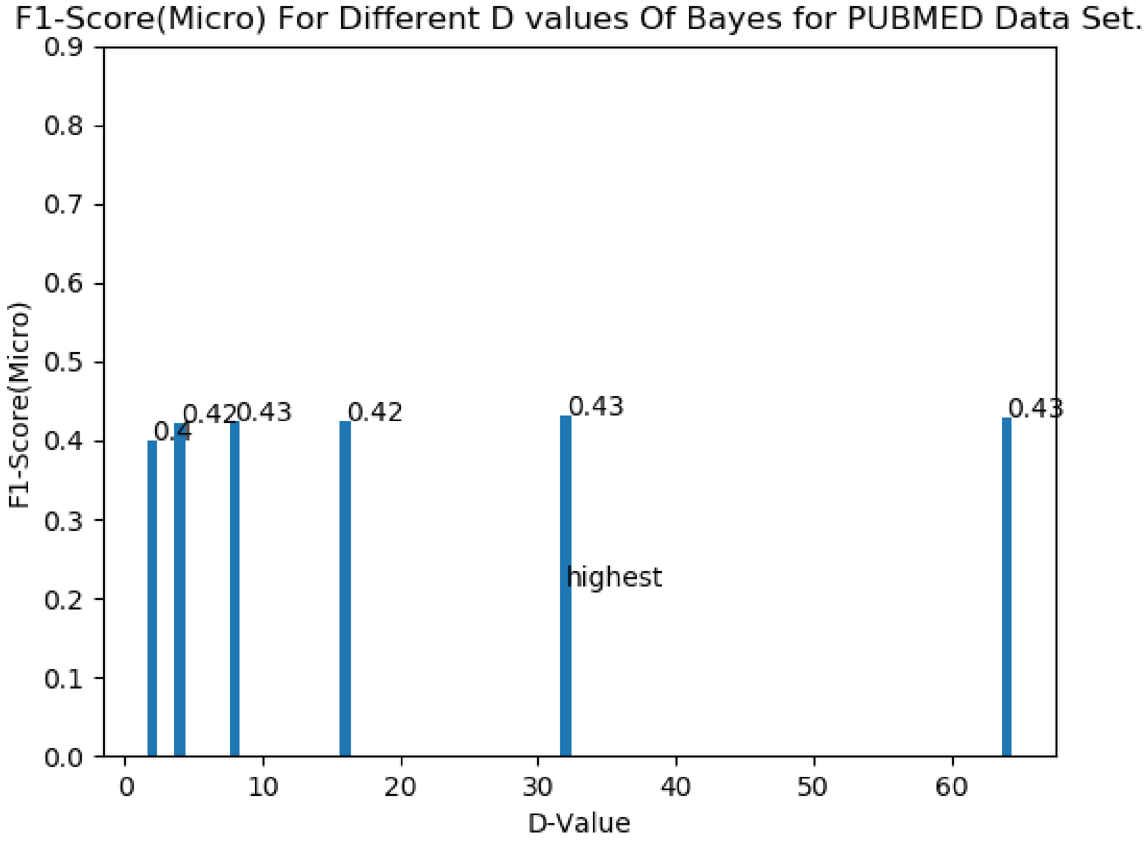

Score(Macro) v/s Dimension for nearest neighbor classifier for dolphins Data
F1-Score(Macro)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.33
0.33
0.33
0.34
0.34
0.34
highest
0
10
20
30
40
50
60
Dimension



2. Pubmed data set
    a. Bayes Classifier


Accuracy For Different D values Of Bayes for PUBMED Data Set.
Accuracy
0
10
20
30
40
50
60
70
80
90
40.07
42.1
42.52
42.39
43.18
42.99
highest
0
10
20
30
40
50
60
D-Value


F1-Score(Macro) For Different D values Of Bayes for PUBMED Data Set.
F1-Score(Macro)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.3
0.32
0.32
0.33
0.34
0.37
highest
0
10
20
30
40
50
60
D-Value



b. Nearest neighbor classifier

Dimension, D	2	4	8	16	16	32
Accuracy	39.12	37.79	38.14	38.06	38.89	39.10
F1-score(Macro)	0.346	0.335	0.331	0.337	0.343	0.340
F1-score(Micro)	0.391	0.378	0.381	0.381	0.389	0.391

Task IV

Divided the data (both the original/high and low-dimensional) into train and test set using cross-validation technique. Measured accuracy and F1-score(Macro and Micro) for all the data sets using scikit-learn library.

Results

1.Dolphins data set

Nearest neighbor classifier

Dimension, D	2	4	8	16	32	64
Accuracy	36.88	36.98	36.73	36.88	38.08	37.73
F1-score(Macro)	0.336	0.336	0.332	0.333	0.343	0.341
F1-score(Micro)	0.369	0.370	0.367	0.369	0.381	0.377

2.Pubmed data set

Nearest neighbor classifier

Dimension, D	2	4	8	16	32	64
Accuracy	38.45	36.90	37.42	37.36	37.71	38.43
F1-score(Macro)	0.348	0.332	0.333	0.338	0.341	0.343
F1-score(Micro)	0.384	0.369	0.374	0.374	0.377	0.384

Task V

Similar results are obtained using library function. Naïve Bayes classifier is implemented using library.

Task VI

Implemented Locality Sensitive Hashing(LSH) to reduce the dimensionality of the data without using any library.
For twitter dataset, without using library, 53.17% accuracy is obtained and using library, 50.08% accuracy is obtained.

Task VII

LSH involves randomized approach. Hence, for limited amount of data, inconsistent performance is obtained. PCA involves statistical approximation. Consistent performance is achieved. Improved accuracy is obtained.