ML ASSIGNMENT 2

(Accuracy is calculated on a scale of 100)

Problem 1

1. Primal SVMs (**Refer to PrimalSlack.py**)

1b) The accuracy of the learned classifier on the training set

for c = 1 is 88.46153846153845
for c = 10 is 88.46153846153845
for c = 100 is 88.46153846153845
for c = 1000 is 89.74358974358975
for c = 10000 is 88.46153846153845
for c = 100000 is 89.74358974358975
for c = 1000000 is 94.87179487179486
for c = 10000000 is 97.43589743589743
for c = 100000000 is 98.71794871794873

1c) The accuracy of the learned classifier on the training set

for c = 1 is 86.20689655172413
for c = 10 is 86.20689655172413
for c = 100 is 86.20689655172413
for c = 1000 is 81.03448275862068
for c = 10000 is 81.03448275862068
for c = 100000 is 84.48275862068965
for c = 1000000 is 81.03448275862068
for c = 10000000 is 82.75862068965517
for c = 100000000 is 81.03448275862068

The best values of c that are tuned from Validation set are [1, 10, 100]

1d) Accuracy on testing set is 84.7457627118644 with the selected classifier from Validation set tuning.

Dual SVMs with Gaussian Kernels (**Refer to DualSVMGK.py**)
1b)
Accuracy on Training set for c= 1 and sigma= 0.1 is 30.76923076923077
Accuracy on Training set for c= 1 and sigma= 1 is 30.76923076923077
Accuracy on Training set for c= 1 and sigma= 10 is 28.205128205128204
Accuracy on Training set for c= 1 and sigma= 100 is 30.76923076923077
Accuracy on Training set for c= 1 and sigma= 1000 is 78.2051282051282
Accuracy on Training set for c= 10 and sigma= 0.1 is 30.76923076923077
Accuracy on Training set for c= 10 and sigma= 1 is 30.76923076923077
Accuracy on Training set for c= 10 and sigma= 10 is 28.205128205128204
Accuracy on Training set for c= 10 and sigma= 100 is 30.76923076923077
Accuracy on Training set for c= 10 and sigma= 1000 is 78.2051282051282
Accuracy on Training set for c= 100 and sigma= 0.1 is 30.76923076923077
Accuracy on Training set for c= 100 and sigma= 1 is 30.76923076923077

Accuracy on Training set for c= 100 and sigma= 10 is 28.205128205128204
Accuracy on Training set for c= 100 and sigma= 100 is 30.76923076923077
Accuracy on Training set for c= 100 and sigma= 1000 is 78.2051282051282
Accuracy on Training set for c= 1000 and sigma= 0.1 is 30.76923076923077
Accuracy on Training set for c= 1000 and sigma= 1 is 30.76923076923077
Accuracy on Training set for c= 1000 and sigma= 10 is 28.205128205128204
Accuracy on Training set for c= 1000 and sigma= 100 is 30.76923076923077
Accuracy on Training set for c= 1000 and sigma= 1000 is 78.2051282051282
Accuracy on Training set for c= 10000 and sigma= 0.1 is 30.76923076923077
Accuracy on Training set for c= 10000 and sigma= 1 is 30.76923076923077
Accuracy on Training set for c= 10000 and sigma= 10 is 28.205128205128204
Accuracy on Training set for c= 10000 and sigma= 100 is 30.76923076923077
Accuracy on Training set for c= 10000 and sigma= 1000 is 78.2051282051282
Accuracy on Training set for c= 100000 and sigma= 0.1 is 30.76923076923077
Accuracy on Training set for c= 100000 and sigma= 1 is 30.76923076923077

Accuracy on Training set for c= 100000 and sigma= 10 is 28.205128205128204
Accuracy on Training set for c= 100000 and sigma= 100 is 30.76923076923077
Accuracy on Training set for c= 100000 and sigma= 1000 is 78.2051282051282
Accuracy on Training set for c= 1000000 and sigma= 0.1 is 30.76923076923077
Accuracy on Training set for c= 1000000 and sigma= 1 is 30.76923076923077
Accuracy on Training set for c= 1000000 and sigma= 10 is 28.205128205128204
Accuracy on Training set for c= 1000000 and sigma= 100 is 30.76923076923077
Accuracy on Training set for c= 1000000 and sigma= 1000 is 78.2051282051282
Accuracy on Training set for c= 10000000 and sigma= 0.1 is 30.76923076923077
Accuracy on Training set for c= 10000000 and sigma= 1 is 30.76923076923077
Accuracy on Training set for c= 10000000 and sigma= 10 is 28.205128205128204
Accuracy on Training set for c= 10000000 and sigma= 100 is 30.76923076923077
Accuracy on Training set for c= 10000000 and sigma= 1000 is 78.2051282051282
Accuracy on Training set for c= 100000000 and sigma= 0.1 is 30.76923076923077
Accuracy on Training set for c= 100000000 and sigma= 1 is 30.76923076923077

Accuracy on Training set for c= 100000000 and sigma= 10 is 28.205128205128204
Accuracy on Training set for c= 100000000 and sigma= 100 is 30.76923076923077
Accuracy on Training set for c= 100000000 and sigma= 1000 is 78.2051282051282

1c)

Accuracy on Validation set for c= 1 and sigma= 0.1 is 25.862068965517242
Accuracy on Validation set for c= 1 and sigma= 1 is 25.862068965517242
Accuracy on Validation set for c= 1 and sigma= 10 is 22.413793103448278
Accuracy on Validation set for c= 1 and sigma= 100 is 25.862068965517242
Accuracy on Validation set for c= 1 and sigma= 1000 is 74.13793103448276
Accuracy on Validation set for c= 10 and sigma= 0.1 is 25.862068965517242
Accuracy on Validation set for c= 10 and sigma= 1 is 25.862068965517242
Accuracy on Validation set for c= 10 and sigma= 10 is 22.413793103448278
Accuracy on Validation set for c= 10 and sigma= 100 is 25.862068965517242
Accuracy on Validation set for c= 10 and sigma= 1000 is 74.13793103448276

Accuracy on Validation set for c= 100 and sigma= 0.1 is 25.862068965517242
Accuracy on Validation set for c= 100 and sigma= 1 is 25.862068965517242
Accuracy on Validation set for c= 100 and sigma= 10 is 22.413793103448278
Accuracy on Validation set for c= 100 and sigma= 100 is 25.862068965517242
Accuracy on Validation set for c= 100 and sigma= 1000 is 74.13793103448276
Accuracy on Validation set for c= 1000 and sigma= 0.1 is 25.862068965517242
Accuracy on Validation set for c= 1000 and sigma= 1 is 25.862068965517242
Accuracy on Validation set for c= 1000 and sigma= 10 is 22.413793103448278
Accuracy on Validation set for c= 1000 and sigma= 100 is 25.862068965517242
Accuracy on Validation set for c= 1000 and sigma= 1000 is 74.13793103448276
Accuracy on Validation set for c= 10000 and sigma= 0.1 is 25.862068965517242
Accuracy on Validation set for c= 10000 and sigma= 1 is 25.862068965517242
Accuracy on Validation set for c= 10000 and sigma= 10 is 22.413793103448278
Accuracy on Validation set for c= 10000 and sigma= 100 is 25.862068965517242
Accuracy on Validation set for c= 10000 and sigma= 1000 is 74.13793103448276

Accuracy on Validation set for c= 100000 and sigma= 0.1 is 25.862068965517242
Accuracy on Validation set for c= 100000 and sigma= 1 is 25.862068965517242
Accuracy on Validation set for c= 100000 and sigma= 10 is 22.413793103448278
Accuracy on Validation set for c= 100000 and sigma= 100 is 25.862068965517242
Accuracy on Validation set for c= 100000 and sigma= 1000 is 74.13793103448276
Accuracy on Validation set for c= 1000000 and sigma= 0.1 is 25.862068965517242
Accuracy on Validation set for c= 1000000 and sigma= 1 is 25.862068965517242
Accuracy on Validation set for c= 1000000 and sigma= 10 is 22.413793103448278
Accuracy on Validation set for c= 1000000 and sigma= 100 is 25.862068965517242
Accuracy on Validation set for c= 1000000 and sigma= 1000 is 74.13793103448276
Accuracy on Validation set for c= 10000000 and sigma= 0.1 is 25.862068965517242
Accuracy on Validation set for c= 10000000 and sigma= 1 is 25.862068965517242
Accuracy on Validation set for c= 10000000 and sigma= 10 is 22.413793103448278
Accuracy on Validation set for c= 10000000 and sigma= 100 is 25.862068965517242
Accuracy on Validation set for c= 10000000 and sigma= 1000 is 74.13793103448276

Accuracy on Validation set for c= 100000000 and sigma= 0.1 is 25.862068965517242
Accuracy on Validation set for c= 100000000 and sigma= 1 is 25.862068965517242
Accuracy on Validation set for c= 100000000 and sigma= 10 is 22.413793103448278
Accuracy on Validation set for c= 100000000 and sigma= 100 is 25.862068965517242
Accuracy on Validation set for c= 100000000 and sigma= 1000 is 74.13793103448276

The best value of c and sigma that are tuned from Validation set are 1, 1000
C = 1 and sigma = 1000

1d) Accuracy on testing set is 72.88135593220339 with the selected classifier that is tuned in Validation set

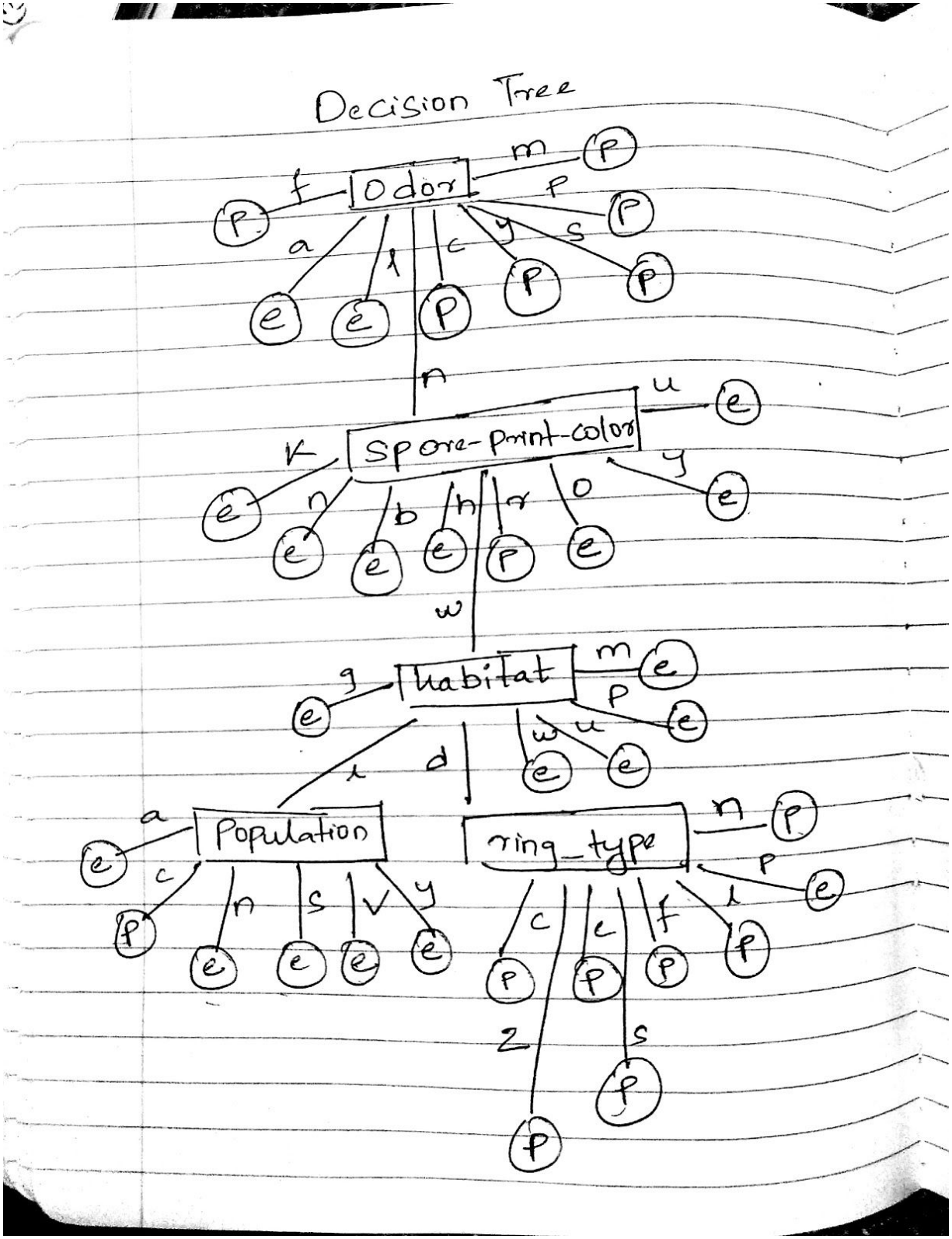
3. k-nearest neighbor classifier (**Refer to KNN.py**)

Accuracy on Validation data set is 81.03448275862068 for value of K = 1
Accuracy on Validation data set is 77.58620689655173 for value of K = 5
Accuracy on Validation data set is 81.03448275862068 for value of K = 11
Accuracy on Validation data set is 74.13793103448276 for value of K = 15

Accuracy on Validation data set is 74.13793103448276 for value of K = 21
The best value K can take is [1, 11]
Accuracy on test data set is 81.35593220338984

4. The best approach that can be used for the classification purpose is **Primal SVM** because it has the best accuracy of 84.745 on training set and it can be seen that for each value of c it maintains accuracy above 80% thus it is not overfitting data.

Problem 2 (Refer to DecisionMushTree.py)



2.2) The number of nodes in the decision tree are 39
2.3) Height of the learned decision tree is 4
2.4) The accuracy of your learned decision tree on the training set is 100%
2.5) The accuracy of your learned decision tree on the testing set is 100%
2.6)
The decision tree works very well for this problem as the accuracy achieved is 100% on training and testing.

Apart from that even if the whole data is split into 15% of training data and 85% of testing data(Can be seen in next page), the accuracy on testing and training is 100%. So, with a simple set of rules that is only 5 attributes are used from 22 to build the decision tree.

2.7)
with 10 % split of training and 90 % split of testing data
Testing accuracy is 99.89056087551299
Training accuracy is 100.0

with 20 % split of training and 80 % split of testing data
Testing accuracy is 100.0
Training accuracy is 100.0

with 30 % split of training and 70 % split of testing data
Testing accuracy is 100.0
Training accuracy is 100.0

with 40 % split of training and 60 % split of testing data
Testing accuracy is 100.0
Training accuracy is 100.0

with 50 % split of training and 50 % split of testing data
Testing accuracy is 100.0
Training accuracy is 100.0

with 60 % split of training and 40 % split of testing data
Testing accuracy is 100.0
Training accuracy is 100.0

with 70 % split of training and 30 % split of testing data
Testing accuracy is 100.0
Training accuracy is 100.0

with 80 % split of training and 20 % split of testing data
Testing accuracy is 100.0
Training accuracy is 100.0

with 90 % split of training and 10 % split of testing data
Testing accuracy is 100.0
Training accuracy is 100.0

From above it can infer that irrespective of split on training and testing data the accuracy for training and testing is 100% except for lower split on training data. So, it is not dependent upon the split.

2.8)(**Refer to DecisionMushTree.py)**

No because the decision tree is dependent upon atleast 4 attributes which can be inferred from the decision tree program.

_____***END**_____