Centre of Excellence in Artificial Intelligence

AI42001:Machine Learning Foundations and Applications

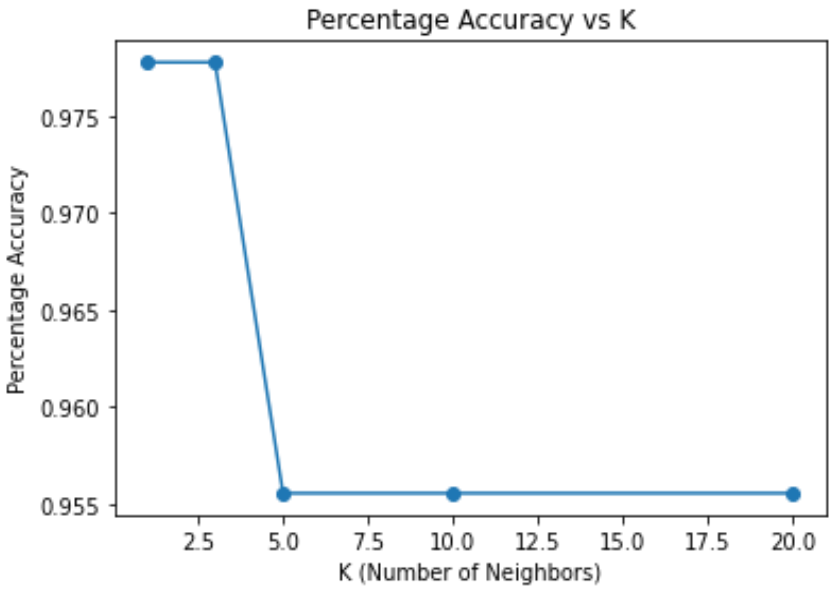
Name-Vaibhav Gupta **Roll Number**-20IE10041

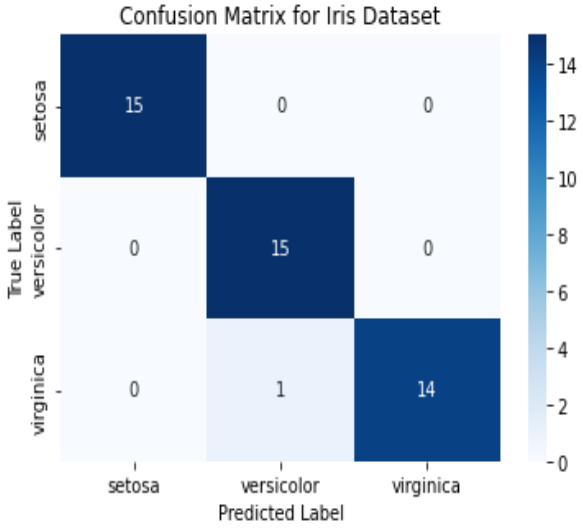
Date:24-01-2024 **Assignment -**2

1. ***Experiment 1:*** *Report the effect of varying K in [KNN_Normal] on Test data. Choose K values from [1, 3, 5, 10, 20]. Plot Percentage Accuracy vs K. Find the best value of the hyperparameter K. Further, plot the confusion matrix for the best K.*

Accuracy for k=1 and k=3 is same . So optimal K can be K=3 and confusion matrix is same for K=3 and K=1

```
K = 1, Accuracy = 97.78%
K = 3, Accuracy = 97.78%
K = 5, Accuracy = 95.56%
K = 10, Accuracy = 95.56%
K = 20, Accuracy = 95.56%
Best K: 1,3
```

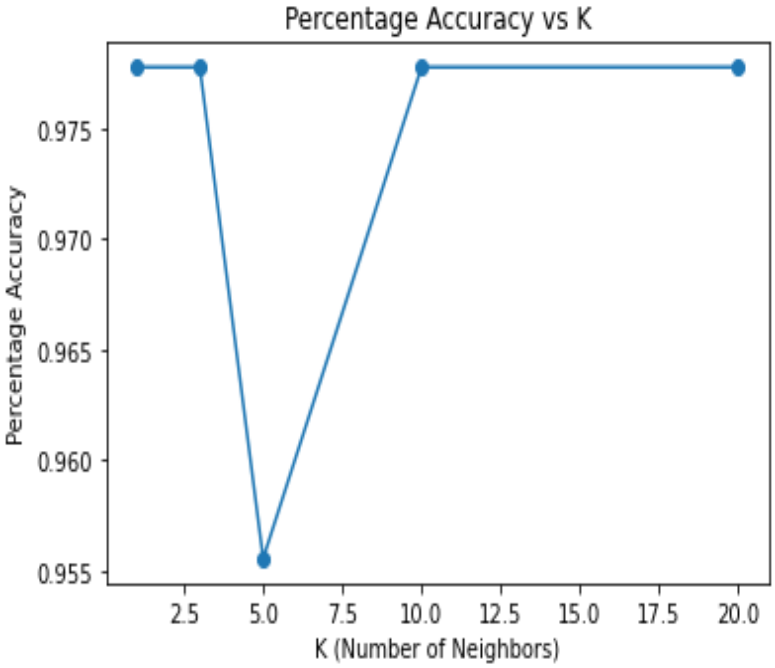


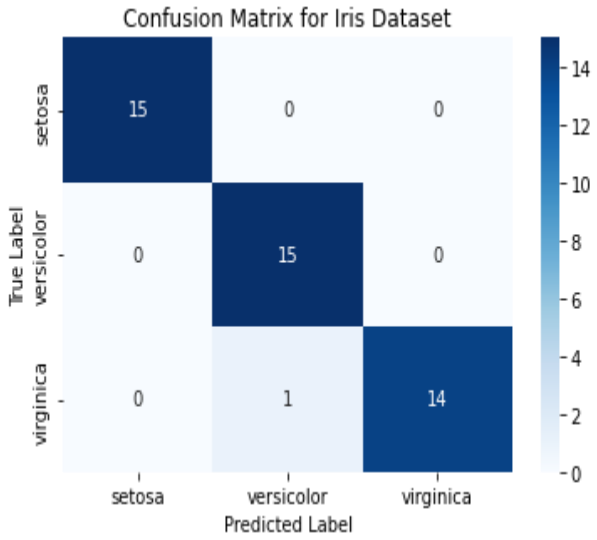


2. ***Experiment 2:*** *Report the effect of varying K in [KNN_Weighted] on Test data. Choose K values from [1, 3, 5, 10, 20]. Plot Percentage Accuracy vs K. Find the best value of the hyperparameter K. Further, plot the confusion matrix for the best K.*

```
K = 1, Accuracy = 97.78%
K = 3, Accuracy = 97.78%
K = 5, Accuracy = 95.56%
K = 10, Accuracy = 97.78%
K = 20, Accuracy = 97.78%

Best K value: 1,3,10,20
```





3. ***Experiment 3:*** *Add noise to only a fraction of the training data: Consider 10% of the training data for noise addition*. Choose a normal distribution with zero mean and standard deviation 1.0. Next, evaluate the new data using [KNN_Normal] and [KNN_Weighted] employing the optimal K found in the earlier experiments (Experiments 1 and 2). How do the performances vary as compared to that of the noiseless case (i.e., Experiments 1 and 2)?*

Choosing **K=3 as optimal** for both as K=1 can be **Sensitive to noise points** and higher values of K like 10,20 can cause **Neighbourhood to include points from other classes.** Lower values of k can have high variance, but low bias, and larger values of k may lead to high bias and lower variance. The choice of

k will largely depend on the input data as data with more outliers or noise will likely perform better with higher values of k.

After adding noise , the results are

```
K = 3, Accuracy for KNN_Normal= 24.44%
K = 3, Accuracy for KNN_Weighted= 22.22%
```

We can see that in case of noisy data the accuracy dropped from 97.78 to 24 and 22 percentages for Normal and Weighted KNN respectively. The performance has worsened.

On taking higher values of K, we get

```
K = 10, Accuracy for KNN_Normal= 31.11%
K = 10, Accuracy for KNN_Weighted= 37.78%

K = 15, Accuracy for KNN_Normal= 31.11%
K = 15, Accuracy for KNN_Weighted= 35.56%

K = 20, Accuracy for KNN_Normal= 28.89%
K = 20, Accuracy for KNN_Weighted= 33.33%
```

We can see the accuracy improved on increasing K from 3 to 10,15,20 in case of noisy data.

Robustness to Noise:

KNN_Normal can be sensitive to noise in the training dataset, especially when the noise is present in a significant portion of the data. Outliers or erroneous data points can affect the decision boundaries and lead to misclassifications.

KNN_Weighted, which considers the inverse of distances as weights, can be more robust to noise compared to KNN_Normal. By assigning lower weights to distant neighbors, KNN_Weighted tends to downplay the influence of noisy points that are far away from the query point.

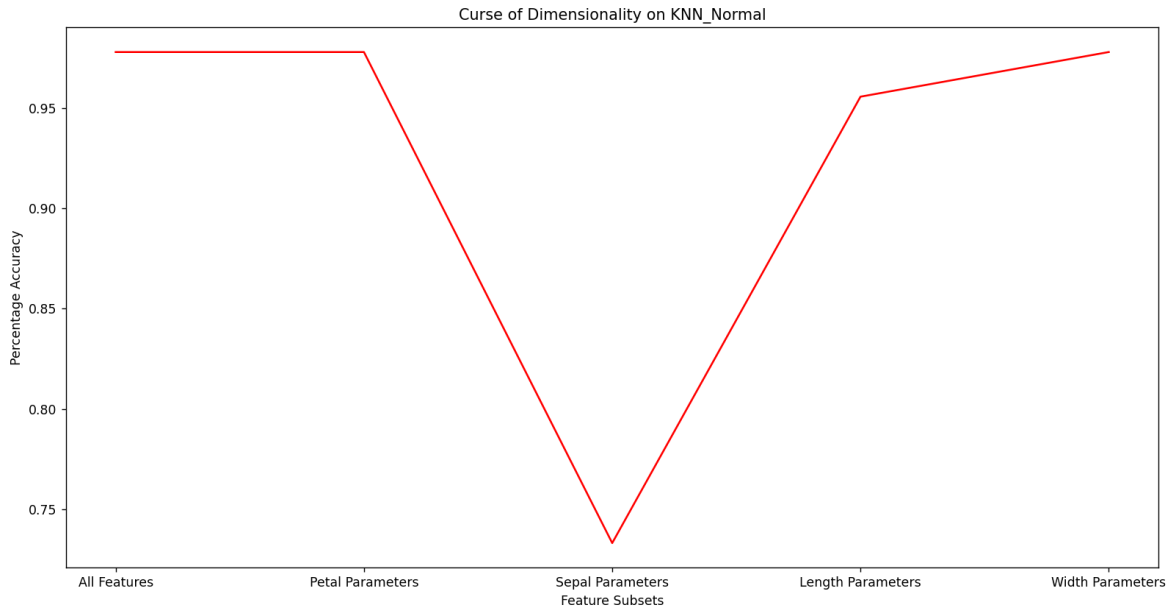
4. ***Experiment 4:*** *For the case of [KNN_Normal], study the effect of the curse of dimensionality.*

Using the optimal K obtained (in Experiment 1), consider (a) All four inputs (sepal length, sepal width, petal length, petal width), (b) Only petal parameters (i.e., petal length and petal width), (c) Only sepal parameters, (d) Only length parameters (sepal length and petal length) and (e) Only width parameters. Analyse whether the curse of dimensionality is imparted by petal parameters, sepal parameters, length parameters and width parameters.

```
All Features: Accuracy = 97.78%
Petal Parameters: Accuracy = 97.78%
Sepal Parameters: Accuracy = 73.33%
Length Parameters: Accuracy = 95.56%
Width Parameters: Accuracy = 97.78%
```

I have chosen Optimal K value =3

We can see the accuracy drops severely if we take only the sepal length and width as features.



Thus, the curse of dimensionality is imparted due to **sepal parameters.**

On comparing length and width parameters, we can see that the accuracy is less if we consider only petal and sepal lengths , so length parameters also impart curse of dimensionality on comparison to width parameters.