**Machine Learning**

**Assignment – 1**

1. Randomly generate 120 values of x in the range [0,1]. Let them be x1, x2, · · · , x120
2. **Classification**:

i. Label the first 80 points {x1, · · · , x80} as follows. If xi ≤ 0.5 then xi ∈ Class1, else (if xi > 0.5) xi ∈ Class2 for i = 1, 2, · · · , 80.

ii. Classify the remaining points, that is x81, · · · , x120 using kNNC. Do this for k = 1, k = 3, k = 4, k = 5, k = 40, k = 80.

iii. Compute classification accuracy.

1. **SOLUTION**

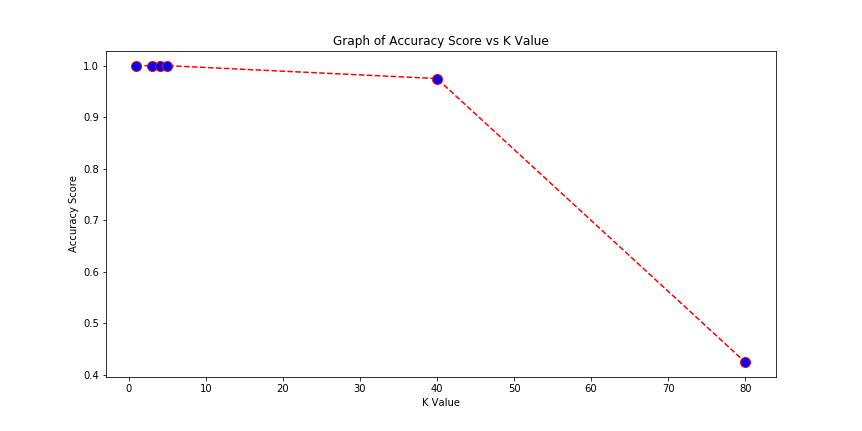
**CODE:**

|  |
| --- |
| Please find the code for Classification submitted in main branch as file: [KNNC\_Clasification\_impl\_1.py](https://github.com/vijaykumar-mishra/ML-Implementations/blob/main/KNNC_Clasification_impl_1.py)   * Classification using k= 5 Nearest Neighbours is shown in result * Classification using k = 1, k = 3, k = 4, k = 5, k = 40, k = 80. * Output is plotted to show accuracy vs K values |

**RESULT:**

|  |
| --- |
| accuracy\_score : 1.0 -> Result when K=5  [1.0, 1.0, 1.0, 1.0, 0.975, 0.425] -> Accuracy score for k = 1, k = 3, k = 4, k = 5, k = 40, k = 80 |

**PLOT:**

****

**INFERENCE/ANALYSIS:**

|  |
| --- |
| * **Accuracy** is the ratio of correctly **predicted data points** over **total number of test samples** * The program uses a **random number generator** and hence each run produces slightly **different outputs**. * **Accuracy is the least for k=80**, where the number of neighbours considered is as high as training sample count. This shows that the model is considering **too many points** to classify or a case of overfitting. * When **k<=5**, the accuracy values are **nearly similar** and shows that a **suitable k** value exists in this range for the given data set. * **When k=1**, the given classification model is showing high accuracy. Since the **test set** and **training set** are generated using the **same mathematical equation** (xi<=0.5∈ Class1, else Class2), they have the **same characteristics,** and the test samples seem to resemble the training sample and are able to fit well when k=1. * However, in **real world**, the data samples may have no direct mathematical relation for classification. In such scenarios, accuracy or error vs k values is an important factor to determine good k values. |