**Q 2.3(d)**

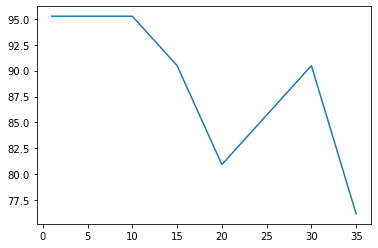
**Iris Dataset**

Accuracy of the basic model

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Accuracy | F-score | AUC |
| Test set | 100 | 1.0 | 1.0 |

The basic model with default parameters had an perfect prediction accuracy on the test set and a 95% accuracy on the training and validation sets.

K Parameter Tuning on Validation Set

****

The highest accuracy occurs at k = {1, 5, 10} but we selected k = 10 as the optimal k because in KNN, a small value of k will make our model highly susceptible to noise which will cause high variations in the performance of the model on different sets of unobserved data and a high value of k will lead to higher computational cost. Also 5 wasn’t chosen because there are 3 target labels and chances of collision (equal prediction probability) will be higher if k = 5 than if k = 10.

**Heart Disease Dataset**

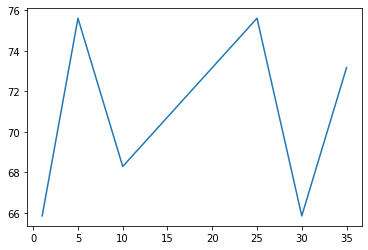
**Before One-Hot Encoding**

Accuracy of the basic model

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Accuracy | F-score | AUC |
| Test set | 76.19 | 0.76 | 0.74 |

The basic model with default parameters had an accuracy of 76.19% on the test set which is not considered good enough and can well be improved on.

K Parameter Tuning on Validation Set

****

Before one-hot encoding the categorical features, highest accuracy occurs at k = {5, 25}

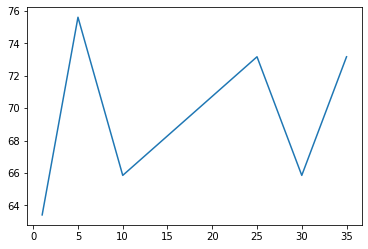
**After One-Hot Encoding**

In one-hot encoding, the integer encoded variable is removed from the dataset and a new binary variable is added for every unique integer value in the feature. This is done to ensure that the categorical data is not ranked when model building.

Accuracy of the basic model

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Accuracy | F-score | AUC |
| Test set | 73.81 | 0.74 | 0.74 |

K Parameter Tuning on Validation Set



The highest accuracy occurs at k = 5 but we selected k = 25 as the optimal k because in KNN, a small value of k will make our model highly susceptible to noise which will cause high variations in the performance of the model on different sets of unobserved data. Also, k = 5 led to overfitting of the model parameters to the training set.