Q 2.4 and 2.7

**Iris Dataset**

**Using the best-found parameter (k = 10)**

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Accuracy | F-score | AUC |
| Test set | 95.24 | 0.95 | 1.0 |
| Training set | 95.24 | 0.95 | 0.99 |
| Validation set | 95.24 | 0.95 | 1.0 |

**Normalization**

The goal of normalization is to make every datapoint have the same scale so each feature is equally important. The normalization technique used is z-score normalization which modifies the data to have a mean of 0 and a standard deviation of 1. This also helps in handling outliers in the dataset. It subtracts the mean of the feature from every observation and then scales it by the standard deviation.

Z-score normalization was chosen over MinMax normalization because the former handles outliers better unlike the latter. Also, if new observations are not in the range of the preselected min and max then the observation will be scaled to an out-of-range value.

**Weighted KNN**

In KNN algorithm, weights can be assigned to neighbors based on their distance from the observation. The aim is to use a kernel function to assign weights to the neighbors so that more weight or importance is given to neighbors closer to the observation and less weights are given to neighbors farther away from the observation.

Where represent the assigned weight.

In this project, we tuned the model with uniform and distance-based weights to see the effect on the performance of the model. We also tuned the distance metric used in finding the neighbors. The two distance metrics we used in tunning are Manhattan distance and Euclidean distance and all tuning was done on the validation set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Weight | Distance Metric | Accuracy | F-Score | AUC |
| Uniform | Manhattan | 85.71 | 0.85 | 0.99 |
| Uniform | Euclidean | 80.95 | 0.80 | 0.98 |
| Distance | Manhattan | 95.24 | 0.95 | 1.0 |
| Distance | Euclidean | 95.24 | 0.95 | 0.9 |

All the tuned parameters produced the same accuracy, F-score and AUC on the validation set.

We opted to use the weighted KNN with the Euclidean distance metric because the mean values for the different species for some features are really close and the model and the model will perform better if higher importance is given to the closest neighbors. Also, Euclidean distance metric was chosen as the input features are similar in type and this will give the shortest possible distance between two points.

**Evaluating the improved model (k = 10, weights = distance, metric = Euclidean)**

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Accuracy | F-score | AUC |
| Test set | 95.24 | 0.95 | 1.0 |

The improved model had a test set accuracy of 95.24% which shows that the model performs well on new observations.

We were also interested in the training set evaluations of the model to ensure it doesn’t underfit or overfit.

How well did the model learn?

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Accuracy | F-Score | AUC |
| Training set | 100 | 1.0 | 1.0 |
| Validation set | 95.24 | 0.95 | 0.99 |

From the table above, the improved model had a training accuracy of 100%.

**Conclusion**

Comparing the training, validation and test evaluations of the model, it can be concluded that the model performs well in predicting the specie of an iris flower based on the petal length, petal width, sepal length and sepal width.

**Heart Disease Dataset**

**Using the best-found parameter (k = 25)**

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Accuracy | F-score | AUC |
| Test set | 66.67 | 0.66 | 0.65 |
| Training set | 65.04 | 0.65 | 0.65 |
| Validation set | 73.17 | 0.74 | 0.74 |

**Normalization**

Normalization was done because all other variables are categorical and have been one hot encoded thus reducing them to multiple columns of their binary representation and thalach being a numerical variable with a much higher scale could offset the model parameters.

Again, Z-score normalization was chosen over MinMax normalization because the former handles outliers better unlike the latter. Also, if new observations are not in the range of the preselected min and max then the observation will be scaled to an out-of-range value.

**Weighted KNN**

In this project, we tuned the model with uniform and distance-based weights to see the effect on the performance of the model. We also tuned the distance metric used in finding the neighbors. The two distance metrics we used in tuning are Manhattan distance and Euclidean distance and all tuning was done on the validation set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Weight | Distance Metric | Accuracy | F-Score | AUC |
| Uniform | Manhattan | 85.37 | 0.85 | 0.82 |
| Uniform | Euclidean | 82.93 | 0.83 | 0.81 |
| Distance | Manhattan | 78.05 | 0.79 | 0.77 |
| Distance | Euclidean | 80.49 | 0.81 | 0.79 |

For the final model we used uniform weights and Euclidean distance because ­­­­­­­­­­­­most of the features are categorical and Euclidean distance is more appropriate for categorical features compared to Manhattan distance.

**Evaluating the improved model (k = 25, weights = uniform, metric = Euclidean)**

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Accuracy | F-score | AUC |
| Test set | 92.86 | 0.93 | 0.92 |

The built model has a high prediction accuracy on the test set.

We were also interested in the training set evaluations of the model to ensure it doesn’t underfit or overfit the training set data.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Accuracy | F-score | AUC |
| Training set | 83.74 | 0.81 | 0.84 |
| Validation set | 82.93 | 0.83 | 0.81 |

**Conclusion**

Comparing the training, validation and test evaluations of the model, it can be concluded that the model performs well in predicting the heart disease status of an individual using cp, slope, thalach, thal and ca. Although, other classification algorithms can be explored to get a reduced generalization gap.