Laboratory 4

Variant 3, Group 13

By Saumya Shah and Anas Tagui

Dataset & Task

The Iris dataset contains 150 samples of iris flowers from three species (setosa, versicolor, virginica), with 50 samples for each class. Each sample has four measurements: sepal length, sepal width, petal length, and petal width. The task is to create a model to classify an iris flower into one of the three classes based on its measurements.

Data Preparation

Undersampling was considered as a possible technique to negate any class imbalances. However, upon inspection of the dataset, the classes were balanced (50 samples for each class), so we didn't do this.

Imputation was considered to fill in any missing values in the dataset. However, there were no missing values, so this was unnecessary.

The data were graphed using box plots and checked for outliers, but no significant outliers were found in the four measurements, so no removal of outliers was necessary.

Since one of the models used (KNN) is a distance-based algorithm, it may be sensitive to features with larger scales, since they could disproportionately affect the distance calculations, outweighing the impact of features with smaller values. Thus, standard scaling (using SciKit's `Standard_Scaler`) was used to normalise the scales of all of the input features.

Dimensionality reduction techniques (like PCA) were considered. However, since there are only 4 features, it is not necessary, and may lead to a loss of information.

Model Choice

We selected K-Nearest Neighbours (KNN) and Logistic Regression classifiers to compare a distance-based algorithm against a linear model. This comparison is valuable as KNN's performance is heavily influenced by the scale and distance metrics of features, while Logistic Regression, though linear, benefits from scaling for efficient gradient descent optimisation.

Other models, such as decision trees and random forests were considered, but there was the concern that deeper trees could overfit and lead to a lack of generalisability compared to simpler, more robust models.

Results

The accuracy metric was selected to evaluate the models. This is because the classes are balanced, so the accuracy represents the model's overall performance over all classes. If the dataset were imbalanced, then it may have made more sense to use metrics such as F1-score.

KNN:

Value of K	**Test Accuracy**
1	93.33%
3	96.67%
5	100%
7	100%
10	100%

Logistic Regression:

Number of Iterations	**Test Accuracy**
2	73.33%
3	86.67%
5	96.67%
7	96.67%
10	100%

For KNN, the key hyperparameter tuned was the number of neighbours (k). For logistic regression, the number of iterations for the optimisation algorithm was tuned. Multiple hyperparameters were tested for each model and documented above.

As evident from these results, the tuning process revealed that specific hyperparameter settings were sufficient to achieve perfect classification on the unseen testing data. Specifically, a KNN model configured with 5 neighbours and a logistic regression model trained for 10 iterations both successfully obtained 100 per cent testing accuracy.

This performance suggests that for the Iris dataset's relatively simple structure and the clear separability of its classes, both a distance-based approach and a linear approach are sufficient to perfectly discriminate the classes in the test data.