Objective:

The purpose of this work is to assess the efficiency for the classification algorithm of the focused study K-Nearest Neighbors (KNN). The performance of the classifier is evaluated, while varying k which means understanding the effect of choosing an appropriate k on the MCD. A heatmap is employed in order to represent the accuracy within range of k parameters from 1 through 10.

Methodology:

1. Data Generation:

Using the make\_blobs function from the sklearn, a synthetic dataset was created so that the output could be volunteering by manipulating the random state variable to five times of the number of features of the dataset. datasets module.

Dataset Characteristics:

The dataset includes 310 samples which are grouped into three clusters with the points (2, 4), (6, 6) and (9, 1).

2. Data Splitting:

The entire data was then divided into the features matrix and the target vector Y. The entire database was then partitioned into test and training set, with the test set comprising 20% of the data using the train\_test\_split function.

The KNN model was constructed with the training data with the testing data used for evaluating the efficiency of the model.

3. KNN Model Implementation:

The KNN classifier was initially set with value of k = 6 neighbors for testing purposes.

The model was built and developed utilizing the training data and the model was tested for the test data.

In this case, the accuracy of the predictions, made by the models was evaluated using the accuracy\_score function from the sklearn. metrics package.

4. Accuracy Testing for Different k Values:Accuracy Testing for Different k Values:

In order to check the impact of the change of k, the accuracy of the classifier was computed by taking k as 1, 2, …, 10.

A heatmap was created in order to represent all these accuracy scores based on different k values.

Results:

Classifier Accuracy with k=6:

Classification performance of KNN classifier was average with a test accuracy of: 0. 98 when k=6. This implies that with the use of 6 nearest neighbors, the accuracy of classification of test samples was at 98%.

Accuracy Across Different k Values:Accuracy Across Different k Values:

The accuracy level did not decrease at all throughout the experiments; it kept at approximately 0. 98 for all k values ranging from 1 to 10. This proves that the dataset is rather small and the model works fine irrespective of the number of neighbors used.

As we can see in the heatmap below, the accuracy scores fluctuated a little but they were almost consistently high no matter what value of k was used. This is in line with the stated hypothesis implying that since the dataset is well organized, then variation of k has little influence on the performance of the model ahead.

A blue bar graph with white text

Description automatically generated

Discussion:

High Accuracy Across All k Values: It was however observed that, the efficiency of the KNN classifier hardly varied, averaging about 98%, for all the different values of ‘k’. This means that indeed the clusters that exist in the dataset are distinguishable hence it becomes easier for the classifier to make right predictions on the input values.

Model Stability: As for the k value, it is often generalized that setting k too low would result in overfitting the model while setting it high would lead to underfitting of the model. However, in this case, the stable accuracy raises some issues claiming that the dataset does not pose a difficult classification task of the model.

Visualization: The use of heatmap is preferred because it gives a direct and easy to understand understanding of how the classifier performs in different ‘k’ values. The heatmap’s entire silhouette shows dark blue color, which indicates that they achieved a high level of accuracy.

Conclusion:

When the KNN classifier was applied to this synthetic data set the accuracy was 0. 98 for all values of k =1 – 10. This means that the dataset used in this work is basic and the data elements are neat for the classification process. In future tests one can use more complicated test data sets with classes which are intersected to investigate the influence of changes in k more effectively.