Assignment 1

Introduction

The study analyzes K-Nearest Neighbors (KNN) classification algorithm performance through simulations that use artificial data. The main purpose focused on assessing how varying values influence the accuracy levels of the predictive model. KNN operates as a prevalent supervised learning method which utilizes neighbor proximity to determine new data point classifications by resident class membership.

Methodology

Dataset Generation:
The make_classification function in the sklearn.datasets module produced a fabricated dataset for the analysis. The generated dataset contains 150 samples alongside 2 informative features while dividing into 3 distinct groups.

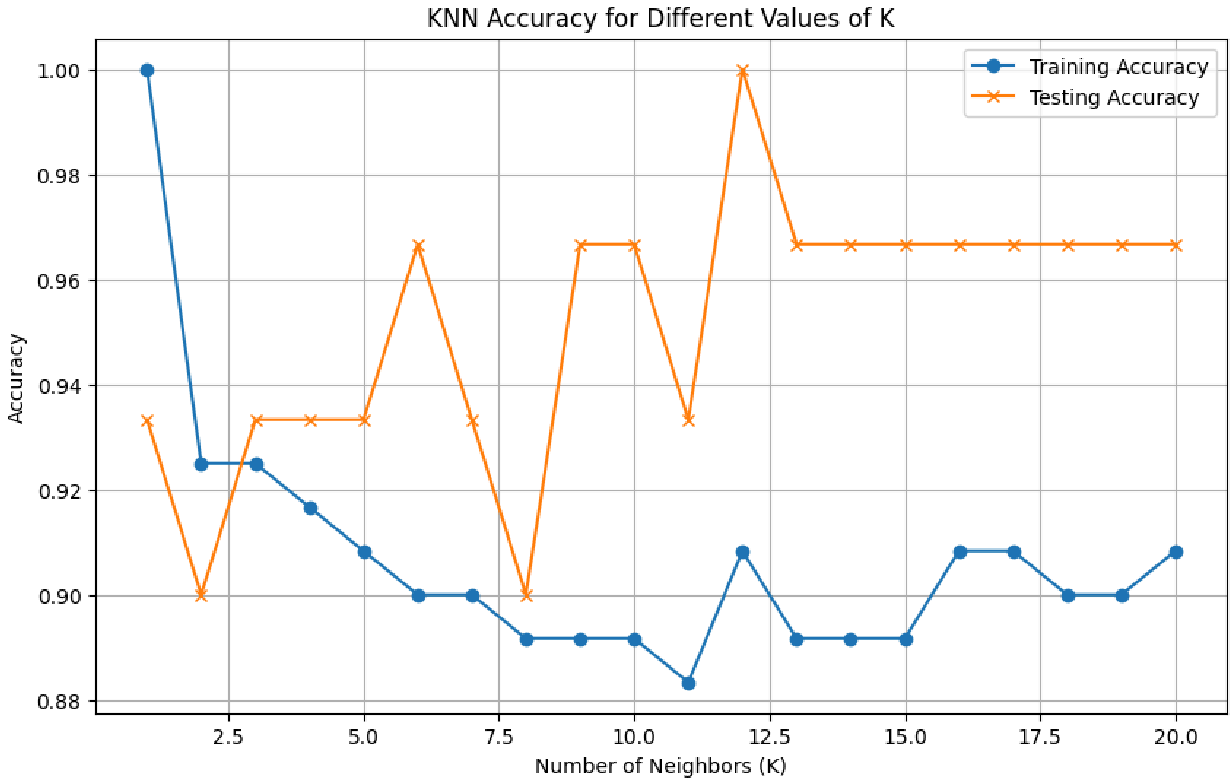
Data Splitting:
The dataset served as a foundation for training and testing purposes through train_test_split method application to guarantee unbiased results.

KNN Implementation:
KNN classifiers from sklearn.neighbors underwent training on the dataset utilizing values that increased from 1 to 20. Each value went through implementation and testing steps while the accuracies were noted down.

Evaluation:
The main evaluation measure focused on accuracy that represented the ratio between instances correctly classified among the total instances.

Results and Analysis

KNN Accuracy for Different Values of K

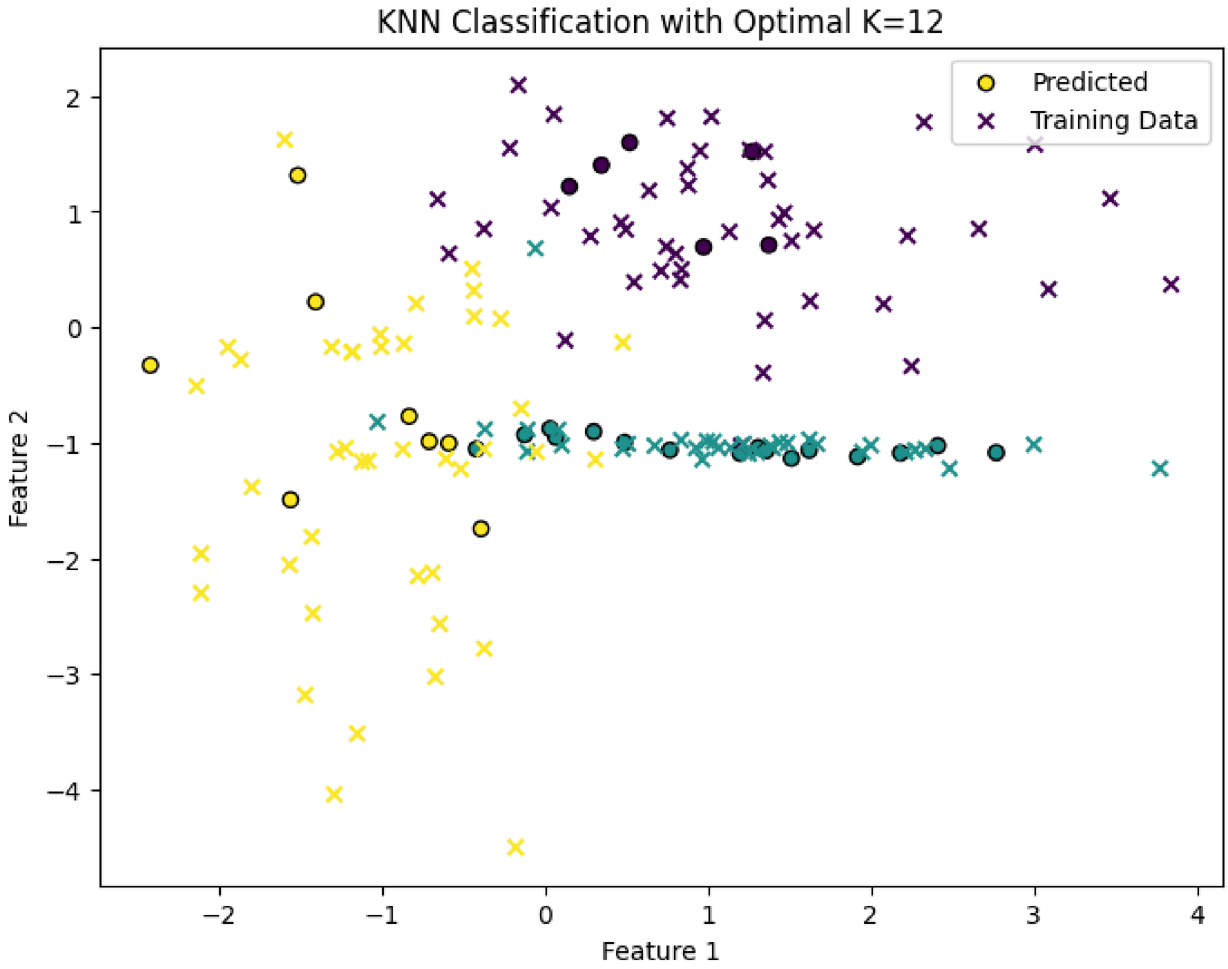


Description : The graph depicts training and testing accuracy values based on various values ranging from 1 to 20.

Analysis:

- For K = 1, the training accuracy is maximum (1.0), which suggests that the model perfectly classified the training data, but the testing accuracy is notably lower to indicate the overfitting.
- With the increase in the value of K, the training accuracy goes down gradually, which is obvious because larger values of K smooth the decision boundary.
- Testing accuracy increases with the increase of up to a certain point, which reflects a better generalization and reduced overfitting. Testing accuracy peaks around K = 12 at 1.0, indicating this as the optimal.
- Beyond the optimal K, testing accuracy stabilizes, while training accuracy continues to decrease slightly, reflecting a balance between bias and variance.

KNN Classification with Optimal K



Description : This scatter plot presents data from the classification outcomes based on the optimal value discovered in the previous graph.

Analysis:

- **Training Data**: 'x' markers are the original labels of the training set.
- **Predicted Test Data**: 'o' markers plot the test set predictions by the classifier.
- The different color clusters indicate successful classification. The fact that the between-cluster overlap is minimal suggests that this model identifies the class boundaries well.
- The visualization will show that the majority of predicted test points align well with the distribution of training data, demonstrating that the model generalizes well to unseen data.

- A few misclassified points, if present, are seen as mismatched colors, showing possible regions of class overlap or ambiguous data points.

Conclusion

The performance of a KNN classifier depends tremendously on the chosen value of K. Optimal K is experimentally identified by balancing the error rate for the training set with that over test sets. This would mean that most machine learning hyperparameters must be further tuned in order to achieve the best predictability. Such a study might be explored further by checking with other metrics while applying k-folds cross-validation for stronger model evaluation.

References

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
2. Cover, T., & Hart, P. (1967). *Nearest neighbor pattern classification*. IEEE Transactions on Information Theory, 13(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964
3. Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.
4. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.