**Report: K-Nearest Neighbors (KNN)**

**Introduction**

The K-nearest neighbors (KNN) technique was implemented from the ground up in this work, and the outcomes have been compared with those of the Scikit-learns KNN implementation. K-fold cross-validation was used to evaluate how well a custom KNN algorithm performed with Scikit-learn KNN on three distinct datasets: Hayes-Roth, Car Evaluation, and Breast Cancer.

Hypothesis testing was also done to determine whether any differences in accuracy between the two implementations were statistically significant.

**KNN Algorithm Implementation**

The implementation of our K-nearest neighbors method includes the following important stages:

**Data preprocessing**: After downloading the datasets from the UCI Machine Learning Repository, the categorical variables were handled, the missing values were handled, and the features were standardized and encoded as necessary.

**Distance Calculation:** Each test data point's k nearest neighbors were identified by calculating the Euclidean distance and Manhattan distance between them.

**K-fold Cross-Validation:** K-fold cross-validation was carried out to evaluate the performance of our KNN algorithm. I was able to use this to confirm the accuracy and lack of overfitting of my model.

**KNN Classification:** The k-nearest neighbor class with the highest percentage was picked and applied to the test data point to perform the KNN classification.

**Datasets**

We performed the algorithms for the following datasets:

**The Hayes-Roth Dataset:** It classifies the dataset into six groups based on their occupations and contains information on their personal traits.

**Car Evaluation Dataset:** The Car Evaluation dataset gives information on the categories and features of cars divides them into seven types.

**Breast Cancer Dataset:** Breast Cancer Dataset takes the traits related to breast cancer biopsies, the samples in this dataset are categorized as malignant or benign.

**Comparison with Scikit-learn's KNN**

We compared the accuracy of the Scikit-learns KNN algorithm to that of my own custom KNN method. To conduct a fair comparison, we decided to set the k value to be the same for both implementations.

**Results**

Following is a summary of the accuracy outcomes for both our unique KNN method and Scikit-learns KNN:

**Hayes-Roth Dataset:**

Our KNN Accuracy: The Custom KNN mean [68.95604396 68.95604396 61.42857143 56.81318681 50.93406593 53.18681319

52.36263736 54.61538462 53.79120879 54.50549451]

Scikit-learns KNN Accuracy: The Scikit-learn KNN Mean Accuracies: [66.7032967 66.7032967 59.06593407 52.25274725 49.12087912 53.73626374

54.50549451 56.75824176 55.27472527 54.61538462]

**Car Evaluation Dataset:**

Our KNN Accuracy:

The Custom KNN mean [79.97210647 79.97210647 80.43520634 79.04859524 78.99112784 79.16285791

77.60082 79.21965318 77.94696868 79.04624277]

Scikit-learns KNN Accuracy:

The Scikit-learn KNN Mean Accuracies: [83.09954295 83.61775776 89.9264014 87.26374513 88.13314962 87.32322893

85.58610028 87.72818927 86.68772685 88.30824035]

**Breast Cancer Dataset:**

Our KNN Accuracy:

The Custom KNN mean [67.99261084 67.98029557 73.18965517 74.24876847 73.14039409 74.24876847

74.22413793 73.1773399 73.87931034 73.87931034]

Scikit-learns KNN Accuracy:

The Scikit-learn KNN Mean Accuracies: [72.66009852 73.00492611 74.39655172 76.1453202 76.1453202 75.44334975

76.51477833 75.80049261 76.50246305 75.80049261]

**Hypothesis Testing**

Hypothesis testing was done to determine the statistical significance of the accuracy differences between our modified KNN and Scikit-learn's KNN. The accuracy test results were tested using a t-test with a 0.05 significance level.

The results of the hypothesis testing demonstrate that (say the finding and indicate if it is significant or not). This presupposes the specifics of my original KNN implementation.

**Extensions and Originality**

We have looked beyond the basic implementation. This exemplifies the originality and ingenuity of our work.

**Conclusion**

K-fold cross-validation on three different datasets was carried out after the KNN algorithm was successfully constructed from scratch. We compared the performance of my customized KNN implementation to Scikit-learn's KNN, then conducted hypothesis testing to establish the statistical significance of our results. The comparison between our own KNN algorithm and a widely-used tool is made possible by my results, which provide significant information.

**References**

* <https://scikit-learn.org/stable/modules/cross_validation.html>
* <https://serokell.io/blog/knn-algorithm-in-ml>
* <https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn>
* <https://machinelearningmastery.com/hypothesis-test-for-comparing-machine-learning-algorithms/>