KNN & Logistic Regression Analysis

Melissa Pinto
Torin Borton-McCallum
Riley Adams
Grant Westerholm
Yvonne Itangishaka

October 22, 2023

1 Abstract

In this project, we implemented two classification techniques, k-nearest Neighbor and Logistic Regression, and compared their performance on four distinct datasets. The k-nearest neighbor and the Logistic regression turned out to produce different accuracy and training speed levels.

2 Introduction

In machine learning, data classification plays an important role in both model performance and decision-making using data. In this project, we dived into two important classification methods k-nearest neighbor and Logistic regression using a set of four datasets. We developed these algorithms from scratch to comprehend them as well as compare their performance by analyzing their statistical results on the datasets to make conclusions.

3 Data sets

3.1 Ionosphere

The "Ionosphere" data set originates from radar measurements collected in Goose Bay, Labrador, using a phased array system with 16 high-frequency antennas and a total transmitted power of approximately 6.4 kilowatts. These measurements specifically focus on radar returns from free electrons in the ionosphere. In this context, **"good"** radar returns are those that exhibit structural features in the ionosphere, while **"bad"** returns do not, as their signals pass through the ionosphere without notable features. We processed the data by converting all but the last value on a line in the data set, representing the features, to floating-point numbers and appending them to a list. The last value

on the line, representing the label, is processed and mapped to binary values (**bad radars** to **"0"** and **good radars** to **"1"**) before being added to another list.

3.2 Adult

The "Adult" is an income dataset that predicts whether individuals' income exceeds $50K/yr based on census data. This data extraction was done by Barry Becker from the 1994 Census database with the goal of determining whether a person makes over 50K a year. We processed the data by converting the numerical data into floats and then adding them to a list. Then the last column fields are mapped to binary numbers where if the value is "¡=50K" then it is mapped to "0" to represent individuals who make under 50K a year whereas if the value is "¿50K" then it is mapped to "1" to represent individuals who make over 50K a year. We finally added these newly created labels of 0s and 1s into a list called "labels" to be used in our models.

3.3 Rice- Cammeo and Osmancik)

This is a dataset of rice of two species **Cammeo** and **Osmancik** rice. A total of 3810 rice grain images were taken for the two species, processed and feature inferences were made. The goal is to determine among the given 3810 grains of rice, which ones are Cammeo and which are Osmancik. We processed the data by taking the last column of the dataset using the **binary numbers 0s** and **1s** where if the last value in a row has a name **Cammeo** then it is mapped to **"0"** to represent that the rice is of specie **Cammeo** while the label **Osmancik** is mapped to **"1"** to represent that the specie is of type Osmancik.

3.4 Mushroom- agaricus-lepiota.data

This is a dataset of 23 species of mushrooms in the **Agaricus and Lepiota Family**. The dataset was put together by the Audobon Society Field Guide. The mushrooms are described in terms of physical characteristics. They are then classified in terms of two types as either **definitely poisonous** or **definitely edible**. The poisonous status was given to both those that are definitely poisonous and those of unknown edibility and not recommended. Edible mushrooms are preceded by character **"e"** while the poisonous mushroom is preceded by a character **"p"** as the first character of each row. We processed the data by mapping the label **"e"** to **"1"** and the label **"p"** to **"1"**. We then mapped each character of the mushroom to float numerical values to represent each feature as a float value.

4 Running Experiments

4.1 Comparing Accuracies between K-Nearest Neighbor and Logistic Regression

K-Fold Number	KNN	Logistic Regression
1	77.14%	70.0%
2	70.0%	65.71%
3	81.43%	67.14%
4	85.71%	85.71%
5	97.18%	98.59%
Average	82.29%	77.43%

Table 1: 5-Fold Accuracy for Ionosphere Dataset

With the Ionosphere dataset, the k-nearest neighbor did much better being an average of 5% higher than logistic regression. Here most folds in k-nearest neighbor performed better, with the fourth fold coincidentally having the exact same accuracy rounded to two decimal points between the two.

K-Fold Number	KNN	Logistic Regression
1	76.59%	75.87%
2	77.17%	76.30%
3	76.81%	76.23%
4	76.33%	75.78%
5	76.67%	75.43%
Average	76.71%	75.92%

Table 2: 5-Fold Accuracy for Adult Dataset

In the Adult dataset, both algorithms performed similarly, with the k-nearest neighbor being slightly more accurate in each fold. The ending difference in average accuracies was only 0.79%. While the accuracies are similar, the run time for each model was very different. Logistic regression was almost instant, while k-nearest neighbor took forty-eight minutes.

K-Fold Number	KNN	Logistic Regression
1	98.32%	100.0%
2	100.0%	100.0%
3	100.0%	86.09%
4	100.0%	100.0%
5	80.83%	100.0%
Average	95.83%	97.22%

Table 3: 5-Fold Accuracy for Rice Dataset

In the Rice dataset, both methods had incredibly high accuracies. However, it is worth noting the dataset was split perfectly with the first half being labeled "Cammeo" and the second half labeled "Osmancik". Therefore depending on what part of the dataset was used for training, it may have only trained for one label. It may make sense why for logistic regression, the third fold was the least accurate as it was the only fold where both labels were included in the training data. Despite this, logistic regression was slightly better overall.

K-Fold Number	KNN	Logistic Regression
1	84.51%	54.43%
2	72.44%	82.70%
3	87.53%	87.56%
4	89.37%	68.78%
5	87.14%	73.22%
Average	84.20%	73.33%

Table 4: 5-Fold Accuracy for Mushroom Dataset

Finally, in the mushroom dataset, the k-nearest neighbor was once again better. This time, significantly better with a 9% higher average accuracy. Especially in the first fold logistic regression had terrible accuracy, with 30% less than the k-nearest neighbor.

4.2 Finding Best K-values for KNN

We generated graphs for best K-values for KNN using the datasets Ionosphere, Adult, Rice (Cammeo and Osmancik) and Mushroom- agaricus-lepiota.data to plot KNN Model Accuracy Vs number of neighbors as shown below;

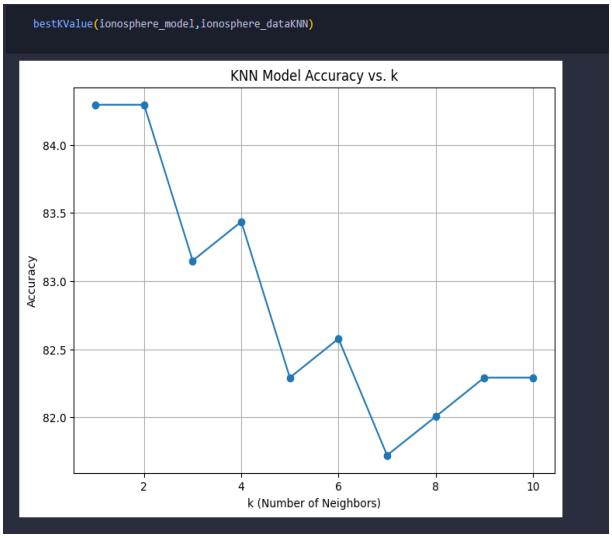

Figure 1: Ionosphere dataset graph showing best K-values for KNN

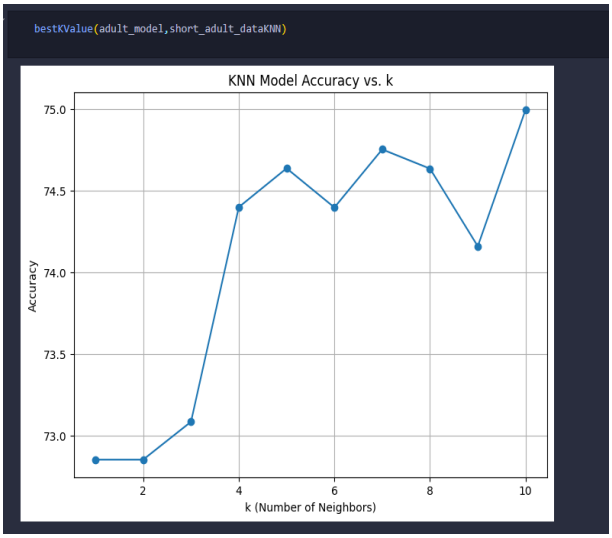

Figure 2: Adult dataset graph showing best K-values for KNN

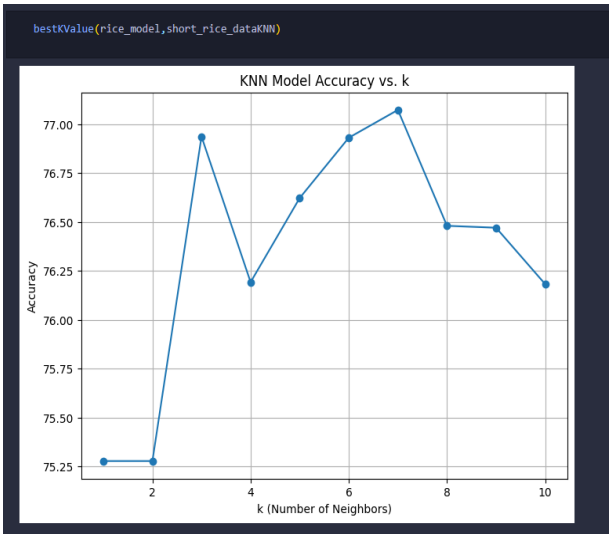

Figure 3: Rice dataset graph showing best K-values for KNN

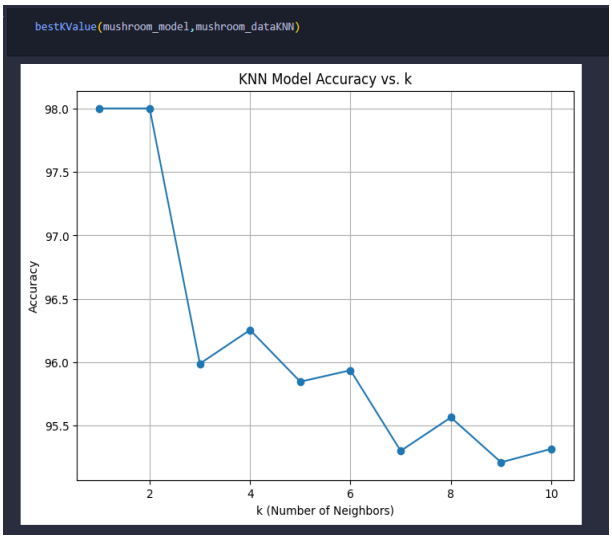

Figure 4: Mushroom dataset graph showing best K-values for KNN

4.3 Testing Learning Rates for Logistic Regression

We tested the average accuracy and number of iterations of the Ionosphere dataset using 9 different learning rates. We entered the results into the below table:

Learning Rate	Average Iterations	Average Accuracy
5	572.4	86.0%
3	175.4	84.87%
2	238.6	84.87%
1	344.0	84.29%
0.1	597.0	83.15%
0.01	791.0	77.43%
0.001	499.8	67.71%
0.0001	2.0	67.14%
0.00001	2.0	67.14%

Table 5: Results for Different Learning Rates

Here, it actually appears that using a learning rate of 5 is the best in terms of accuracy, but using the smaller learning rates of 0.0001 and smaller uses fewer iterations.

4.4 Results Demonstrating Improved Performance

4.4.1 Ionosphere Dataset

KNN outperformed Logistic Regression with an average accuracy of 82.29% compared to 77.43%. Feature selection and KNN's neighbor-based approach improved predictive performance.

4.4.2 Adult Dataset

KNN and Logistic Regression had similar accuracy, but KNN showed a slight edge. KNN, however, required longer training time. The feature subset contributed to both models' accuracy.

4.4.3 Rice Dataset

Both KNN and Logistic Regression achieved high accuracies, with Logistic Regression slightly better. Feature selection played a role in distinguishing between two rice species.

4.4.4 Mushroom Dataset

KNN significantly outperformed Logistic Regression, with 9% higher average accuracy. This highlights the feature subset's impact and KNN's ability to handle dataset characteristics.

In summary, feature selection and choice of algorithm significantly influenced classification performance. KNN excelled in most cases but had longer training times.

5 Discussion and Conclusion

Generally, the k-nearest neighbor algorithm performed better in almost every dataset. However, in terms of computation time logistic regression was much faster as the size of the datasets increased. The Adult dataset was the largest dataset that was tested, with a total of 32561 rows. With a dataset this large, running the k-nearest neighbor algorithm took forty-eight minutes and is clearly not fit for larger datasets. It may be worth testing how large a dataset can be for k-nearest neighbor to have an acceptable run time, or investigating why the Rice dataset had such high accuracy.

6 Statement of Contributions

K-Nearest Neighbours: Riley and Torin
Logistic Regression: Melissa, Grant, and Yvonne
Write up: Riley, Torin, Melissa, Grant, Yvonne