QMBU 450: Selected Topics in Quantitative Methods

Homework – 03 Report

Name: Abdul Rahman Hamadeh

Brief Introduction:

This homework consists of a classification problem: A data set is given to do analysis on and find a good way to pre-process the data and fit it into a model with optimized hyperparameters based on the data to perform classification on.

Data Processing:

Binary labels: the labels were converted into binary representation: 1 for True and 0 for False to make dealing with the labels easier.

Feature Selection: The feature selection was done using CHI2. The features with the highest scores (impact on the data) were chosen using a pre-defined threshold. This resulted in selecting 7 features out of the given 32 features.

One-hot Encoding: the data points are converted into one-hot encoded data using the `OneHotEncoder(sparse=True)`. The sparse flag was set to `True` since each feature has many options.

Models:

The classification was done twice with two separate models, the K Nearest Neighbor Classifier and the Logistic Regression Classifier.

Hyperparameter Optimization:

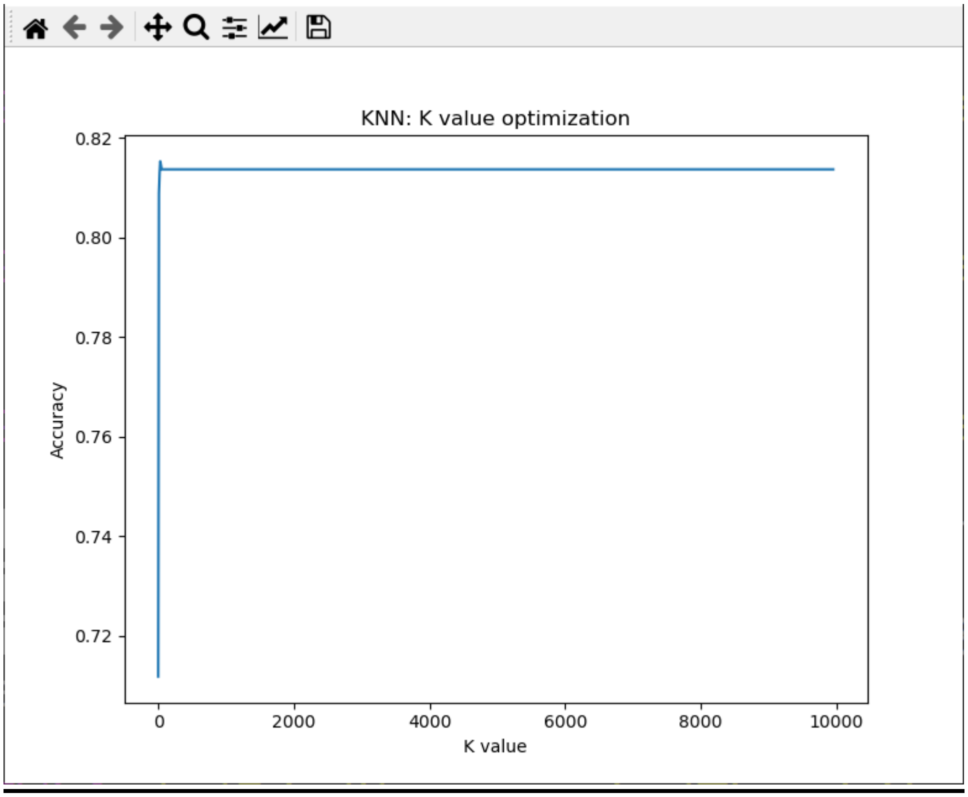
K-Nearest Neighbor:

The K value was optimized by trying many different K values and training the model with them and find the K value that yields the highest accuracy.

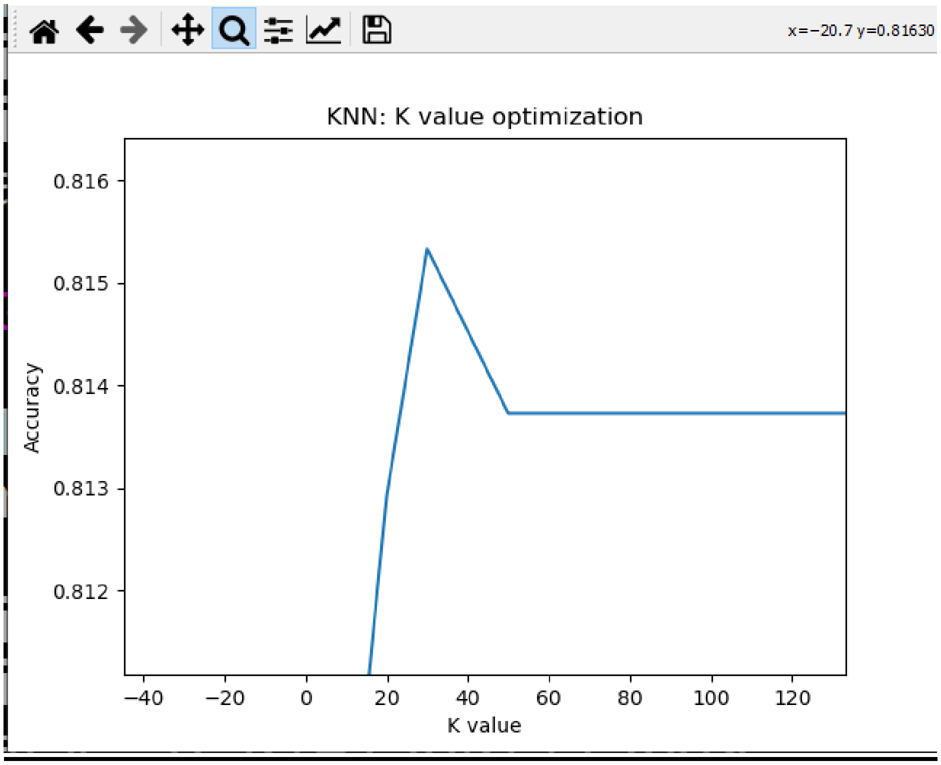
Logistic Regression:

Maximum Iteration: The maximum iteration number was also optimized by trying multiple maximum iteration values and picking the one that yields the highest accuracy.

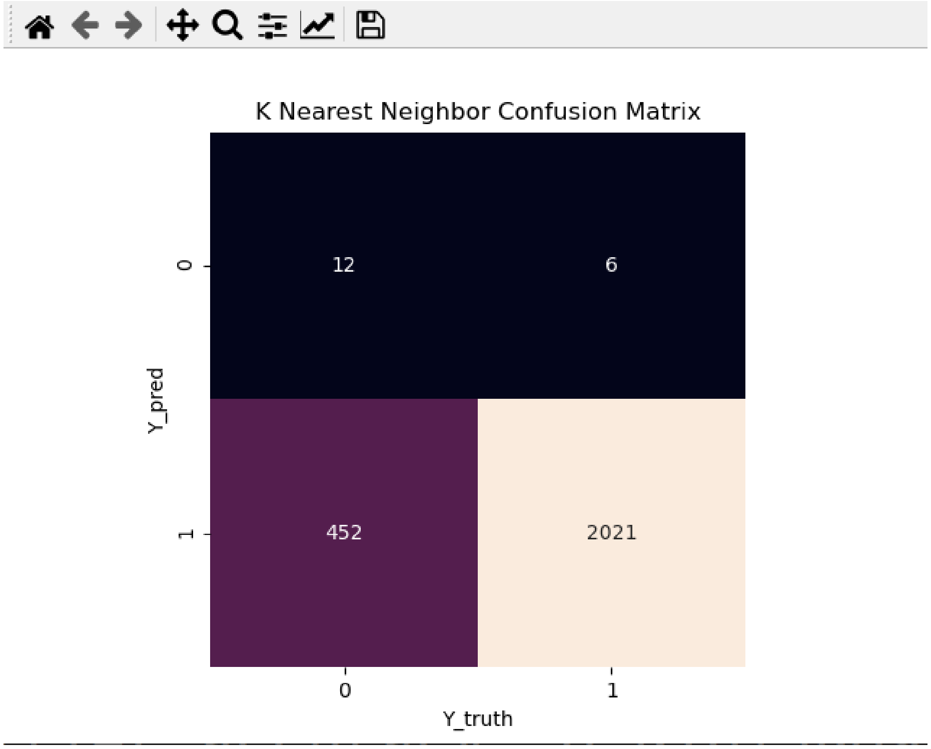
Graphs and figures:



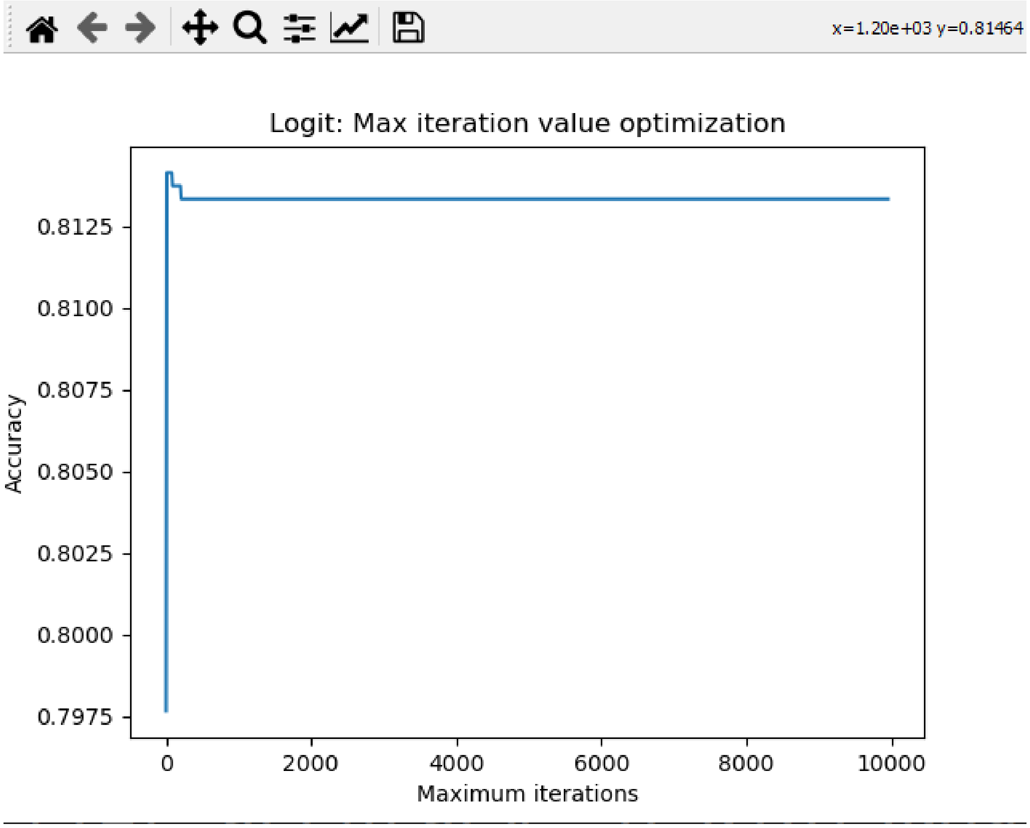
K Nearest Neighbor: K Optimization Graph



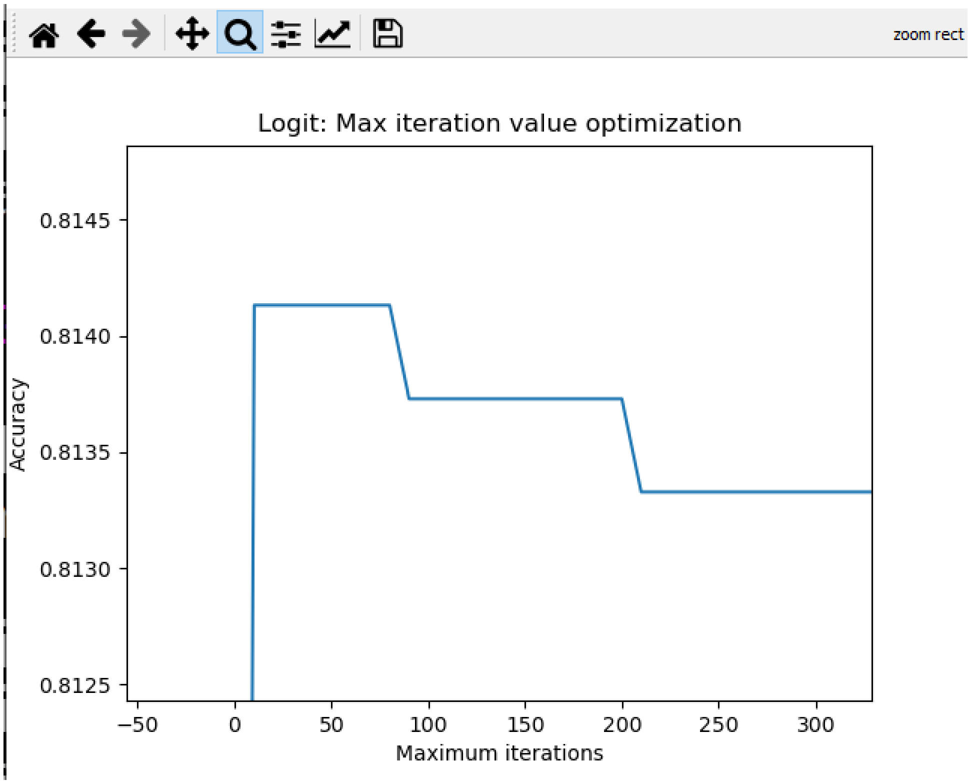
K Nearest Neighbor: K Optimization Graph (Zoomed in)



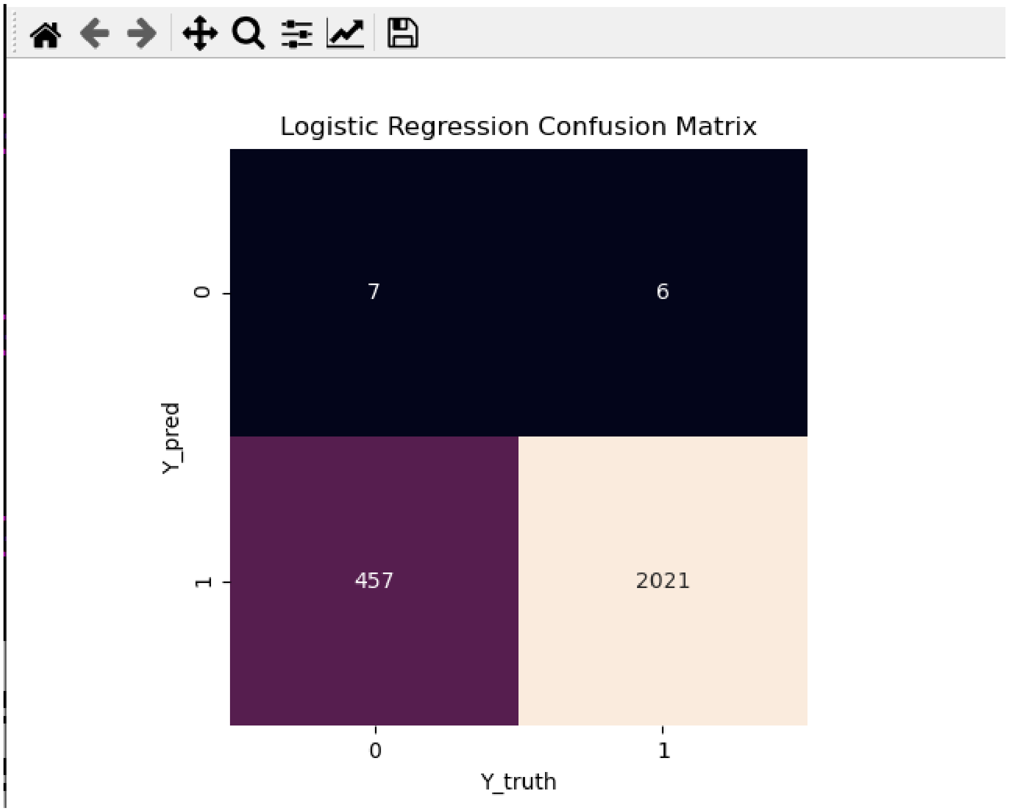
K Nearest Neighbor: Confusion Matrix



Logistic Regression: Maximum Iteration Value Optimization



Logistic Regression: Maximum Iteration Value Optimization (Zoomed in)



Logistic Regression: Confusion Matrix