Data Cleaning:

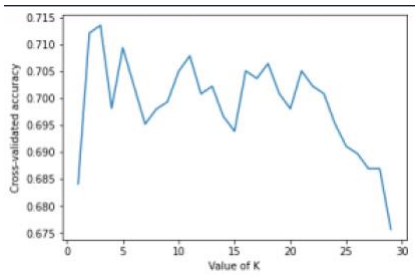
I decided to go back to the titanic dataset for this homework assignment. I tried my best to improve my data cleaning from the previous assignments. I only dropped a total of 4 variables because they had a large number of missing variables, such as Name, Ticket, Cabin, and Passenger Id. I then turned all of the variables into integers, so the data is easier to work with. Then turn the categorical variables Sex and Pclass into numerical variables. There were only a few missing variables for Embarked, so I filled them in with the most common value of 'S'. I then filled in the rest of the missing values for the remaining variables with their mean. This way we end with no null values, and only numerical values.

Question 1:

After the data was cleaned, I split my data into training (80%) and test (20%) sets. For my performance metric I will be using the accuracy score to determine the precision of my models. I am using accuracy score because the class distributions in this data set are more similar than they are imbalanced.

Question 2:

Next, I moved on to training a logistic regression model using k-fold cross validation from scratch. I first split the training set into n parts and then repeatedly trained on n-1 parts. I decided to test my values for folding between 1 and 30, to see which value of k was the most successful. Based on the graph, it looks like k=3 gives us the best value for this model. So, then when using k=3 for the evaluation on my logistic regression model, I ended up getting an accuracy score of 80.06% on the training dataset.



Question 3:

I then implemented a grid search from scratch to test a few different hyperparameters to see which one worked the best for my model. I tested to see which C parameter and if the l1 or l2 parameter would be the best. I came to find out that the best value for C was 10, and the l2 penalty had a better time fitting my model than l1. These two parameters together had the highest accuracy score, which was 79.64%. Once I found what the best two parameters were, included them in my logistic regression model. With this, I achieved an accuracy score of 80.76%.

Question 4:

The best model out of the two was the grid model. It had a slightly better accuracy score than the k-fold model. I also checked my work and implemented scikit-learn to see how my results from scratch compared to the API. The results were fairly similar, yet the API performed better. This was important for me to test, just to make sure I was in the right ballpark with my procedures built from scratch. I tested the grid search on my test data, and it came out with an accuracy score of 82.68%. This was a good improvement from the training data accuracy score. If I were to spend more time on the grid search, I would have tested more hyperparameters to see if they could have improved my model.