**Name:** Mahesh Gaya

**DESCRIPTION**

We pulled the data from the titanicFull.csv. There are about 891 records in that file. We also separated the header from the data. We found out that there were some of the values that were missing. In the next paragraphs, we will discuss how we cleaned up the data, constructed the distance function, implemented the prediction functions, and finally plotted the results in a graph.

*(ref: exercise 4, 5)*

Some of the ages were missing. To resolve that, we first calculated the average age of all the people in the table. Then, we replaced the missing values with that average.

*(ref: exercise 6)*

We also fill out some of the missing values in the Embark table. To do that, we first found out how many people survived, or not, per each category. Given that the records that had missing embark values did survive, we picked the category that had most survived people. That category was Cabin.

*(ref: exercise 7)*

Now that we did not have any missing values, we went ahead and convert every number into floating point to enable us to perform mathematical calculations on them.

*(ref: exercise 8)*

We wanted to deal with some more attributes in an efficient way. We converted the sex column into numbers. 0 stands for “male” and 1 stands for “female”.

*(ref: exercise 9)*

Now that we had all the attributes that we needed, and that none of them had missing value, we implemented the distance function that took in the following five attributes into consideration: pclass, age, sibsp, parch, and fare.

*(ref: exercise 10)*

In order to have a better prediction, we shuffled the data, and then we split ten percent of the data into test data and the rest into training data.

*(ref: normalize data)*

We also normalized the data so that we could compare the regular data with the normalized data.

*(ref: exercise 15)*

Then it was time to plot the graph. We calculated the weighted kNN and the unweighted kNN with both the regular data and the normalized data when k was 1, 3, 5, 10, 25, 50, 75, 100, 150, 175, 180, 190, and 200. *(see graph below)*



**RESULTS:**

1. We found out that
2. Regular data outperformed the weighted kNN algorithm when k was 50.
3. Regular data outperformed the unweighted kNN algorithm when k was 10.
4. Normalized data outperformed the weighted kNN algorithm when k was 75.
5. Normalized data outperformed the unweighted kNN algorithm when k was 10.
6. The best variation is non-normalized data with unweighted kNN” algorithm.

**CONCLUSION:**

The unweighted data seemed to have a better performance when there are less kNN. This is because the weighted data adds an extra constraint to the algorithm.

The accuracy is about 0.6 and 0.7. It could have been better, if we changed the way we replaced the missing data and if we changed the way we replaced sex and embark values.