Assignment 2: K-NN Algorithm

- **Members (GR-G, 65)**
  - **Name:** Surajit Kundu, Roll No: 21MM91R09
  - **Name:** Ankur Kumar Jaiswal, Roll No: 21MM62R08

1. Implement a k-NN classifier and measure the classification accuracy on the test instances. Classification accuracy is defined as the percentage of the total number of correctly classified instances to the total number of test instances. Use a train-to-test split ratio of 80:20.

Response :

- A k-nearest neighbors classifier is built for the analysis of movie reviewing using the Euclidean distance matrix. The value of *K* is taken *5* randomly. The total number of test instances is *30977*. Correctly classified instances are *16416* as shown in Table 1.

Accuracy = Total Correctly Classified Instances/Total Instances

= 16416/30977

= 0.529941

We get 52.99% accuracy of the classification.

The dataset contains *156060* records. We get *154885* instances after data preprocessing and cleansing. The dataset is split into an 80:20 ratio for training and testing(train *123908*, test *30977*).

Sentiment	Correctly Classified	Wrongly Classified
0 (negative)	165	1230
1(somewhat negative)	1093	4461
2 (neutral)	13878	1852

3(somewhat positive)	1139	5336
4 (positive)	141	1682
	16416	**14561**

Table 1: Sentiment wise classification statistics

☐ **Process of building the KNN:**

- Read the dataset (training.tsv) using the pandas library
- Select “Phrase” and “Sentiment” from the dataset
- Data PreProcess:
  - Convert phrase to lowercase
  - Remove the stop words from the phrase
  - Remove the numbers from the phrase
  - Remove the special characters from the phrase
  - Converting the phrases to a matrix of TF-IDF features
- Split the dataset in an 80/20 ratio
- Find the distance from each test input to all train data points.
- Take the *K* nearest neighbor’s point and select the mode of class.

2. Vary the value of k (depending on the number of classes) with three different similarity/distance measures such as a) cosine similarity, b) Euclidean distance, and c) Manhattan distance and evaluate the performance of your classifier on each of them independently. Compare their performances and analyse the results.
Response :

We change the value of *K* from *2* to *45* in the interval of *3* and analyze the accuracy of the KNN classifier.

☐ **Cosine similarity**

We calculate cosine similarity by measuring the angle between the vectors projected in multidimensional space. We have varied the value of *K* from *2* to *45* in the interval of *3*. Initially, the accuracy was *0.5043* when *K = 2*. Then it increases till the value of *K* is *11*. The maximum accuracy was *0.54930*.

Maximum Accuracy: 0.54930, K = 11

☐ **Euclidean distance**

We calculate Euclidean distance by measuring the shortest distance between two data points. We have varied the value of *K* from *2* to *45* in the interval of *3*. Initially, the accuracy was *0.4823* when *K = 2*. Then it increases rapidly till the value of *K* is *8*. The maximum accuracy was *0.53557* at *K* = 8.

Maximum Accuracy: 0.53557, K = 8

☐ **Manhattan distance**

We calculate the Manhattan distance by taking the absolute difference between the data points. We have varied the value of *K* from *2* to *45* in the interval of *3*. Initially, the accuracy was *0.4735* when *K = 2*. Then it increases rapidly till the value of *K* is *8*. The maximum accuracy was *0.53209* at *K* = 8.

Maximum Accuracy: 0.53209, K = 8

☐ **Performance analysis comparisons**

We have plotted the accuracy with three different distance matrices while varying the value of *K*. The accuracy is high when the range of *K* is between *5* to *11* as shown in *Figure 1*. Then the accuracy drops step-wise and converges to a minimum.

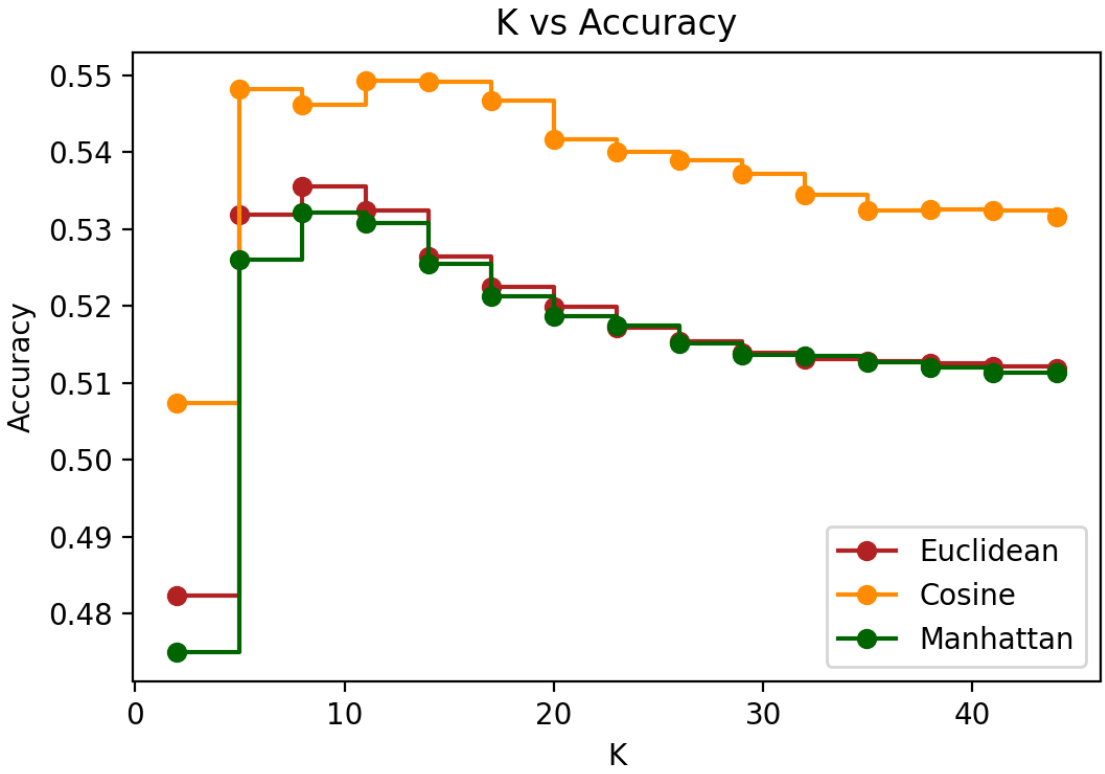

Figure 1: Accuracy comparison with *K* values

We can analyze that the graph for euclidean and manhattan distance is almost the same. For euclidean and manhattan distance, as we increase the value of *K*, the accuracy decreases. We are getting the maximum accuracy of almost ~ *53.3* % at *K = 8*. Similarly, for cosine similarity, as we are increasing the value of *K*, the accuracy decreases, and we have achieved the maximum accuracy of *54.93 %* at *k = 11*.

- **Comparison of performance and analysis of results**

If we compare the performance of different distance measures, we can conclude that the cosine similarity is having the best accuracy among all. It also means that it has the maximum correctly classified instances too. And a generalize trend can be analyzed from the results that accuracy is maximum in a certain range of *K* i.e. *K = 5 to K = 11.* After that, it decreases continuously and converges to a minimum.

3. Plot your results in different graphs (x-axis: k, y-axis: accuracy) for all the three metrics. What trends can be observed from the graphs?

Response :

When we plot the graph between accuracy and k using euclidean as the distance measure, initially, the accuracy is *48.23* % for *K = 2*. A rapid increase in the accuracy can be seen in the graph after *K = 5*. It will achieve the maximum accuracy of *53.557* % at *K = 8.*

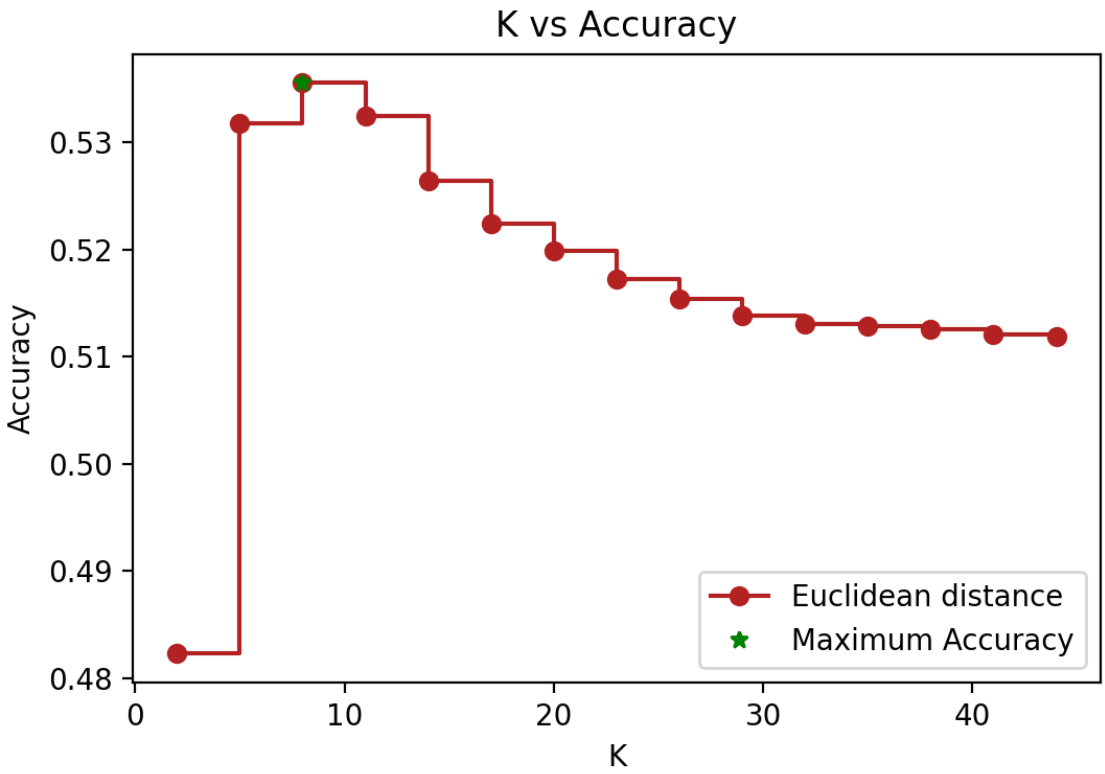

Figure 2: Accuracy vs K using Euclidean distance

When we plot the graph between accuracy and k using cosine similarity as the distance measure, initially, the accuracy is *50.43* % for *K = 2*. A rapid increase in the accuracy can be seen in the graph after *K = 5*. It will achieve the maximum accuracy of *53.557* % at *K = 11*.

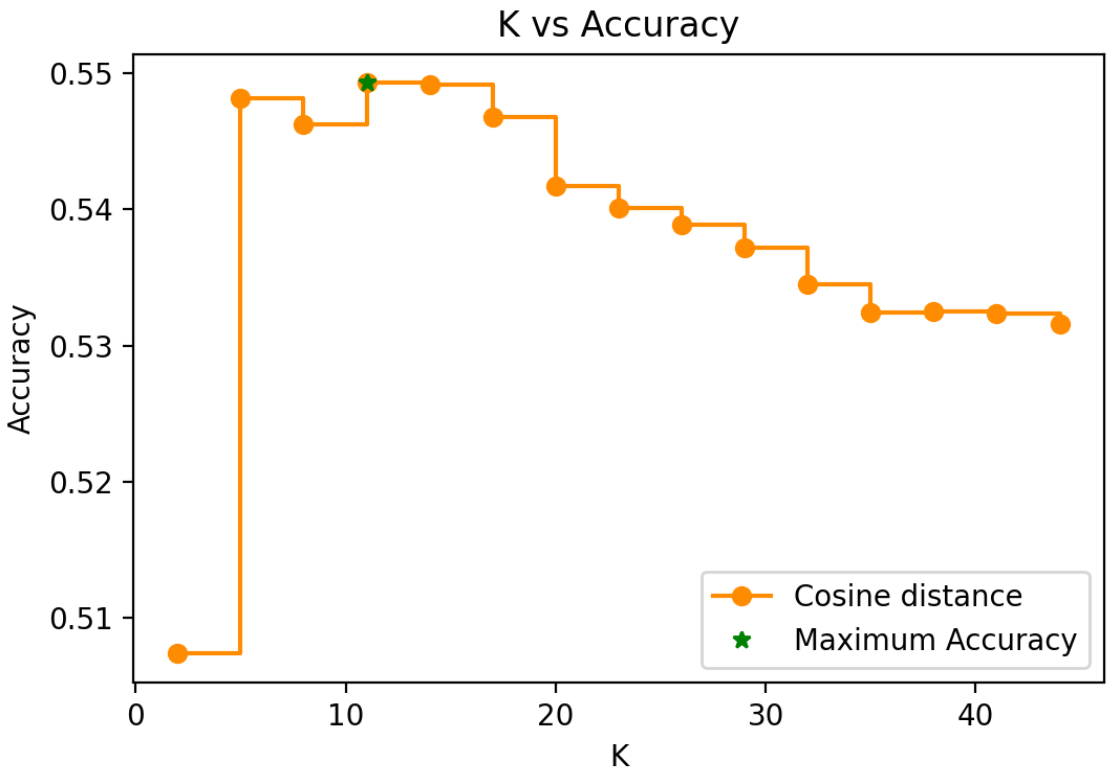

Figure 3: Accuracy vs K using Cosine similarity

When we plot the graph between accuracy and k using Manhattan as the distance measure, initially, the accuracy is *47.35* % for *K =2*. A rapid increase in the accuracy can be seen in the graph after *K = 5*. It will achieve the maximum accuracy of 53.209 % at *K = 8*.

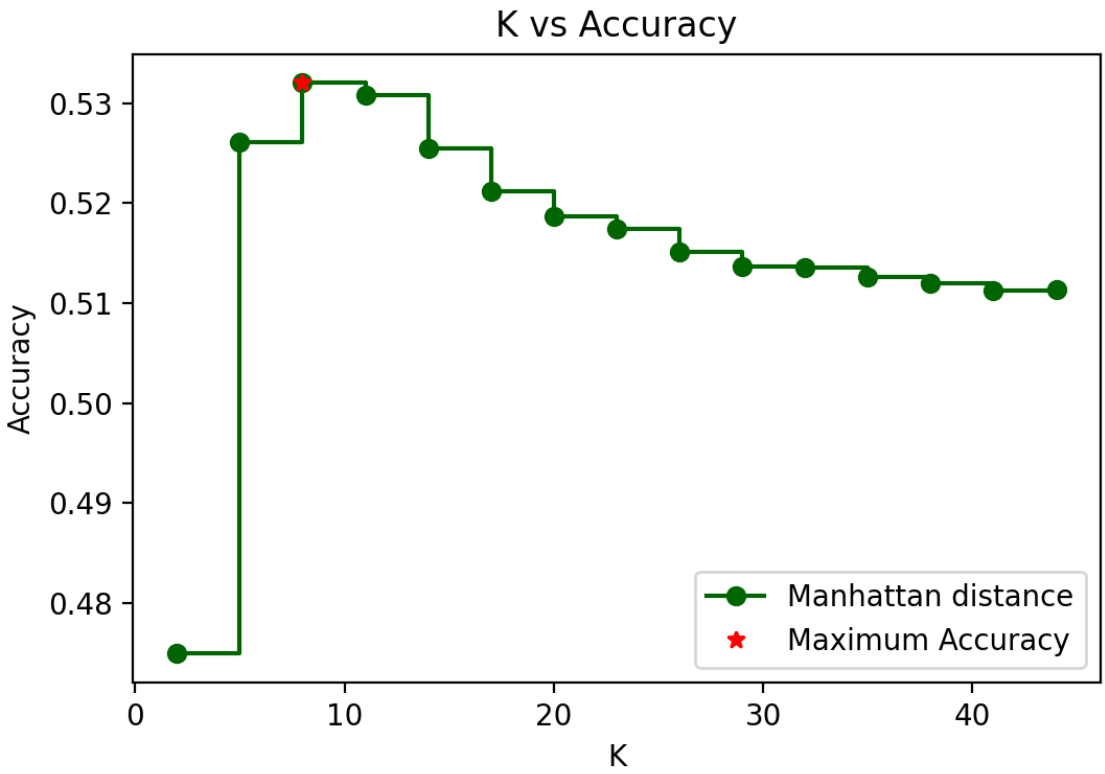

Figure 4: Accuracy vs K using Manhattan distance

☐ A **trend** can be **observed** from the **graph** that initially when we increase the value of *K*, the accuracy starts rising and reaches its peak(maximum) at a specific value of *K.* Then, after that, it starts decreasing with the increasing value of K and converges to the minimum.

4. A brief report explaining the procedure and the results.

Response :

☐ Please refer to the document "Procedure and Results.pdf " for procedure and results.