PA 2 KNN Report

Sai Charan Thannir - 1001635048

Jahnavi Nuthalapati - 1001827251

1.Nearest Neighbors method:

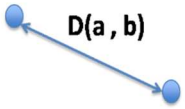
The nearest neighbor method is used to classify a data point based on the classification of its neighbors. To classify a given point, we calculate the distance between all the points in the given plane or proximity using different metrics such as Euclidean distance, Manhattan distance, Minkowski distance and Mahalanobis distance. Using one of these distances, we calculate the distance of the given point from all the other points and based on the K value (which tells the number neighbors to be considered) we select K neighbors which are closest in the distance to the given data point. The given data point is classified based on its nearest neighbors.

For example if k=3 and for the three nearest neighbors of a data point we have class labels point1:A, point2:B, point3:A then we classify the given data point as it belongs to class label A as the majority of its neighbors are classified as A.

Formulae for distance metrics:

Euclidean Distance:

$$D(a,b) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$



Manhattan Distance:

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Minkowski Distance:

$$D(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^{1/p} \right)^{p}.$$

Mahalanobis Distance:

$$D^2 = (x - m)^T \cdot C^{-1} \cdot (x - m)$$

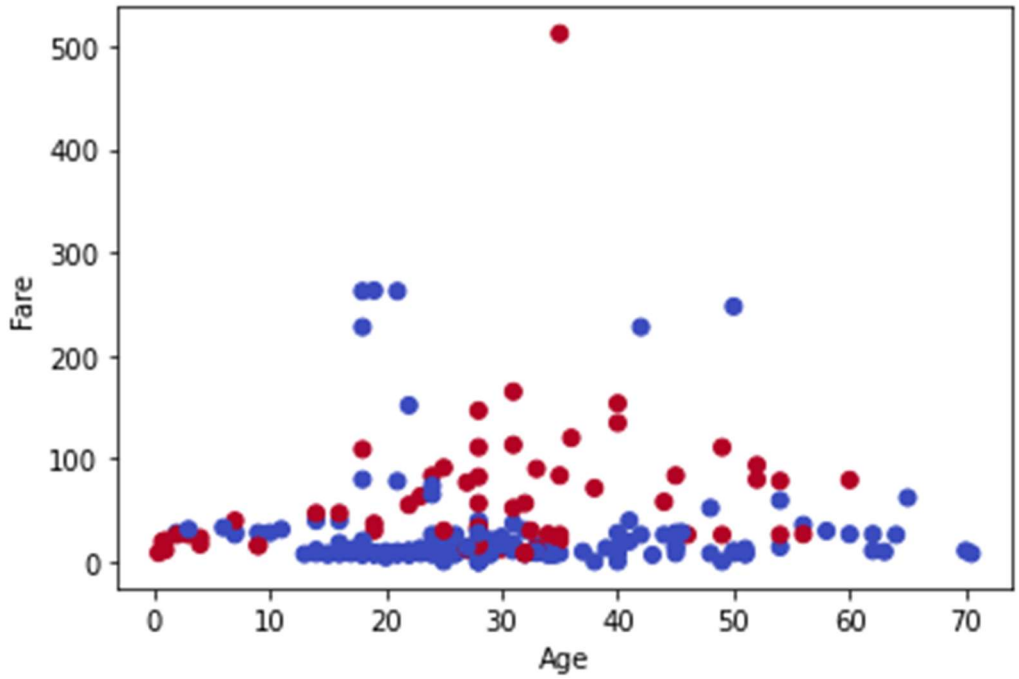
2.Criteria for selecting the three attributes:

In the given dataset there are eight meaningful attributes which are survival, class, sex, age, sibsp, parch, fare, embarked. To get the meaningful patterns we need to select either survival, age or fare. If we select other attributes, we fail to get meaningful results. So, for this reason we have selected survival, age and fare as our three attributes to get meaningful interpretations and meaningful results.

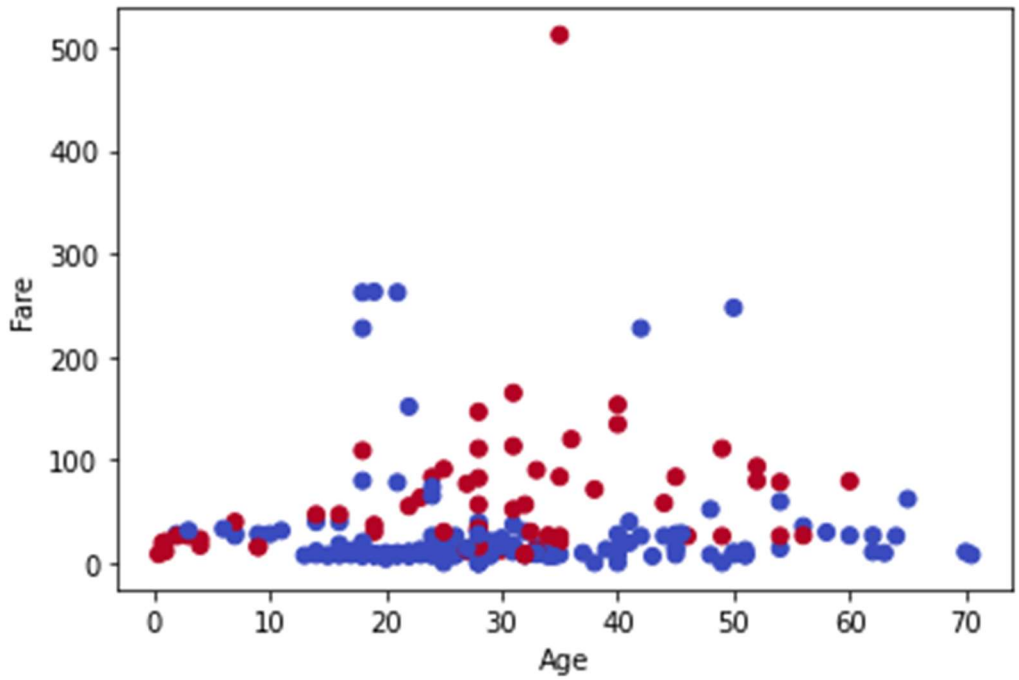
3. Visualizations of the classifier in a 2D projection, for all three different number of neighbors:

We have considered three different values for the K which are 1,3 and 7. Here are the visulaizations for each case and their confusion matrices and classification reports. From these visualizations we can determine which case has more accuracy. When plotting the graph, we are visualizing the relationship between sepal length and sepal width for each case. We used scatter plot for the visualization between the variable's survival, age and fare and classifying each point to a class variable. The Blue points represent the 'class - 0'which represents "age". The Red Point in the graph represents the 'class-2' which represents the target variable "fare".

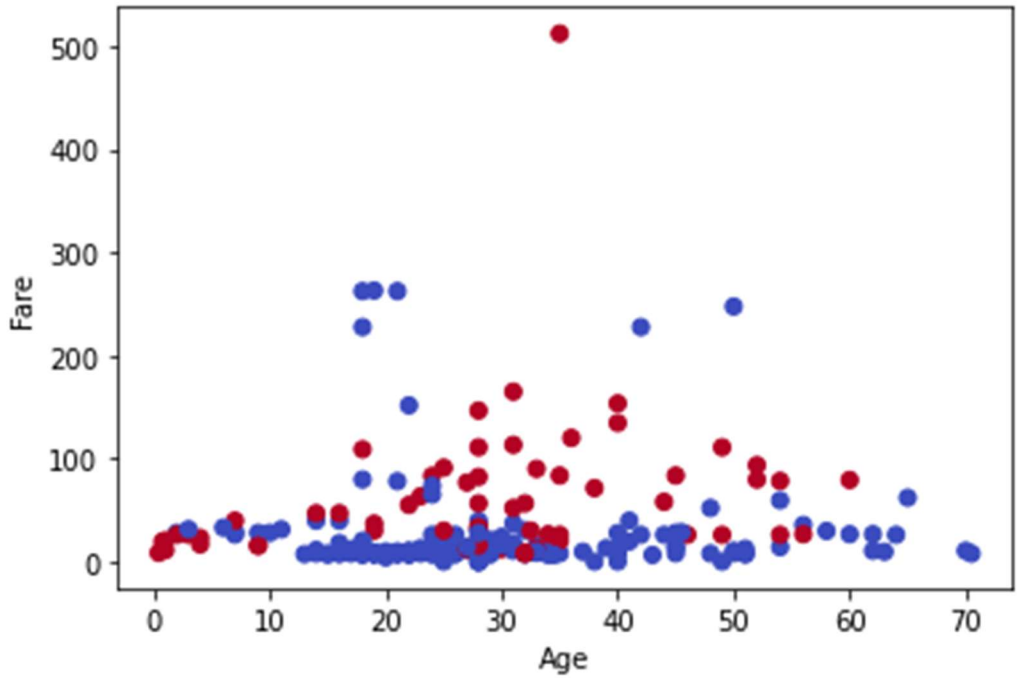
For K=1:


Fare
Age
0
100
200
300
400
500
0
10
20
30
40
50
60
70

For K=3:


Fare
Age
0
100
200
300
400
500
0
10
20
30
40
50
60
70

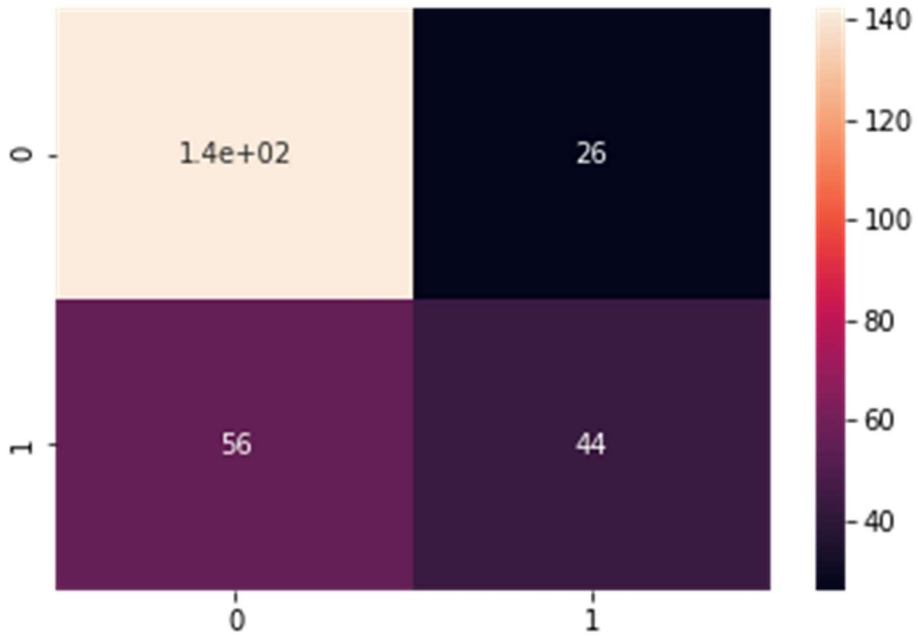
For K=7:



4. Interpretation and comparison of Results:

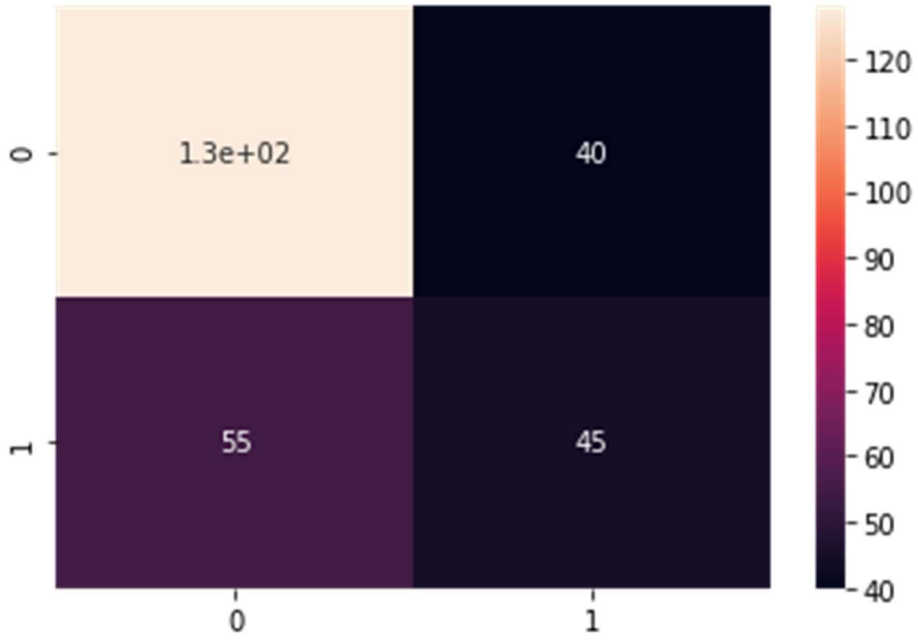
For K=1:

```
confusion matrix: [[142  26]
 [ 56  44]]
confusion matrix: AxesSubplot(0.125,0.125;0.62x0.755)
classification report:               precision    recall  f1-score   support

           0       0.72      0.85      0.78       168
           1       0.63      0.44      0.52       100

    accuracy                           0.69       268
   macro avg       0.67      0.64      0.65       268
weighted avg       0.68      0.69      0.68       268
```



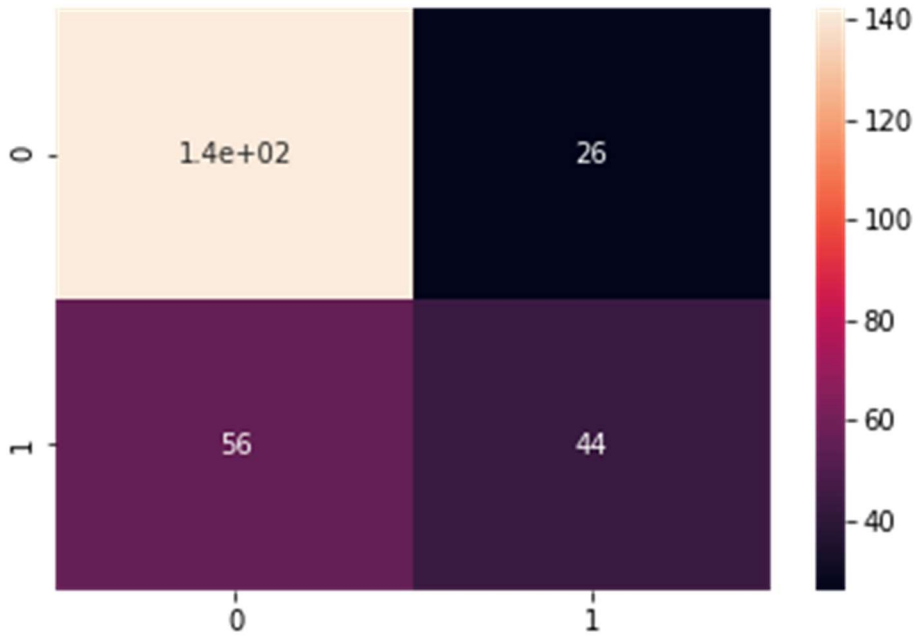
For K=3:

```
 confusion matrix: [[128  40]
  [ 55  45]]
classification report:              precision    recall  f1-score   support

           0       0.70      0.76      0.73       168
           1       0.53      0.45      0.49       100

    accuracy                           0.65       268
   macro avg       0.61      0.61      0.61       268
weighted avg       0.64      0.65      0.64       268

confusion matrix: AxesSubplot(0.125,0.125;0.62x0.755)
```



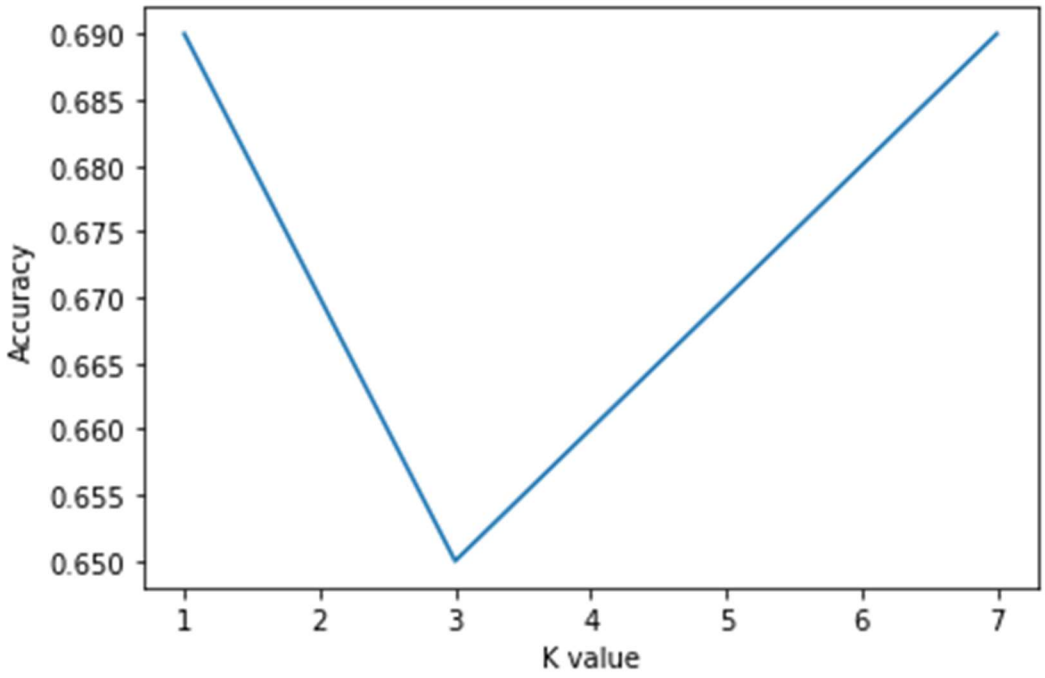
For K=7:

```
confusion matrix: [[142  26]
 [ 56  44]]
classification report:              precision    recall  f1-score   support

           0       0.72      0.85      0.78       168
           1       0.63      0.44      0.52       100

    accuracy                           0.69       268
   macro avg       0.67      0.64      0.65       268
weighted avg       0.68      0.69      0.68       268

confusion matrix: AxesSubplot(0.125,0.125;0.62x0.755)
```



Here we have the confusion matrices and classification reports for each of the three cases. Using these we can determine which case is has more accurate value. From the above-mentioned details, it is evident that when K=7 it is more accurate and less accurate when the K value is 3 and moderate for k=1.

Plot between K values and accuracy:



The above plot which is between three different considered K values and the obtained accuracy values for each of the case. It is evident that when K value is 1 accuracy is more which is 0.68 and when K value is 3 accuracy is reduced to 0.65 which is moderate and when the K value is & the accuracy is high which is 0.69.