PROJECT ASSIGNMENT – 2: KNN REPORT

MANIKANTA BHAVANAM– 1002039918

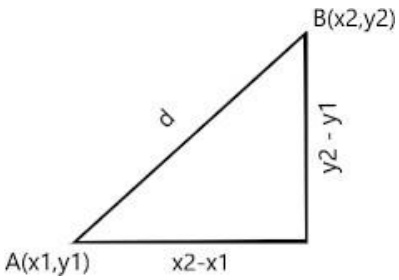
MEGHANA RAMIDI– 1002036880

1.) **PREPROCESSING OF THE DATA:** We have observed the data clearly andhence it doesn't have any null values it is not necessary to add any values tothe null value attributes. Next, we checked for the datatypes of the attributes as we deal with the classifier training and testing and the validation. Apart from this we haven't done any pre-processing as much wasn't required for the processing.

2.) **KNN NEAREST NEIGBHOR METHOD**: KNN is used to solve classification problems. The main aim is to find out the nearest neighbor of any given query point. As per this method, for a query point 'x', all the points closer to this point has similar characteristics respective to 'x' that is if 'x' is positive, all the points near to 'x' can be positive too (it's an assumption). This can be achieved by calculating the distance between the query point and the nearest points. The distance can be found using the following methods.

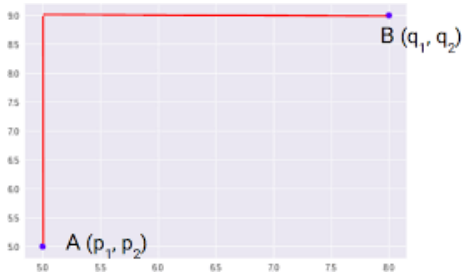
Method 1: Euclidean distance Formula If we consider point A(x1, y1) and point B(x2, y2), the distance between A and B is 'D'. We must substitute the point A and point B values in the above formula to get the nearest distance D between them.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



Method 2: Manhattan Distance In above picture, as per Manhattan Distance method, the points are placed in a grid and the distance is considered here between two points whereas in Euclidean method displacement is considered.

$$d(x, y) = \sum^{n}$$



Method 3: Minkowski distance It is considered as generalized form of Euclidean distance and Manhattan Distance because, take both the distance technique and the new technique for finding the distance between vectors.

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Method 4: Mahalanobis Distance:

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

Where, D is the Mahalanobis distance x is vector of data.

m is vector if mean values of variables.

c is inverse of covariance matrix of independent variables.

T means vector should be transported.

3.) SELECTING PLAS, MASS, and PEDI: In the given pima-indians-diabetes.csv filethere are 9 columns which are as follows.

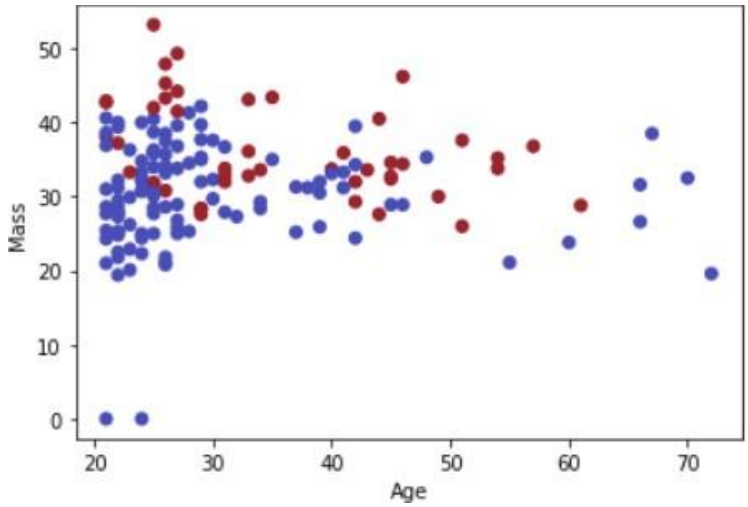
Preg Plas Pres skin test mass pedi age class

Among these 9 columns, PLAS, PEDI, and MASS seems to be more meaningfulbecause these attributes are almost related to every other attribute in
some or the other way and gives us the entire gist of the table and on top of that these attributes are all numerical values which helps to represent the data more significantly. With these attributes we can represent the data which forms to be the important attributes in terms of diabetes. And, the class which has been taken as the output and with the help of that we have sorted the dataset attributes and took the top 3 which will help us to perform certain operations on the dataset taken into consideration with these three attributes will help us analyze and give clear picture of the data.

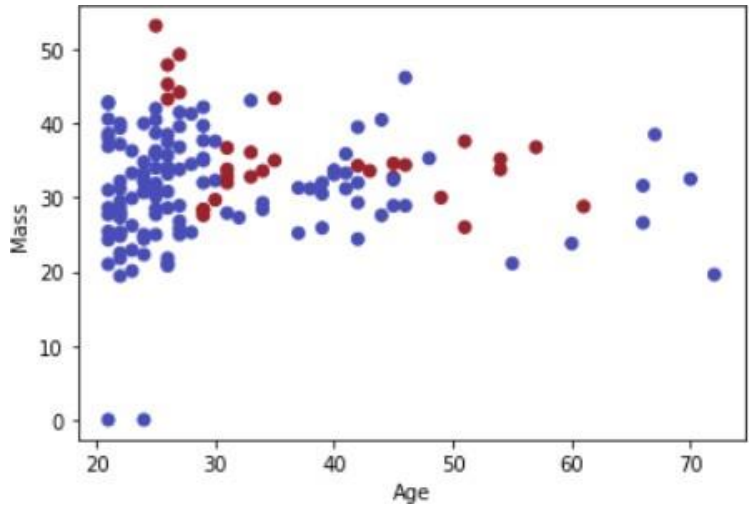
4. Visualizations of the classifier in a 2D projection, with three different K values and Comparisons:

We have considered 3 different values of K and trying to display get the visualizations and classifications below. In this report we are demonstrating visualizations for each case and their confusion matrices and classification reports. We used scatter plot for the visualization between the variable's age, survival and fare and classifying each point to a class variable. The 3 different values of k are 3,6,23. In the below projections, the x axis is representing AGE and y axis is representing the FARE. The red points in the graph represents the Fare and blue points indicates the age of the dataset.
K = 3:

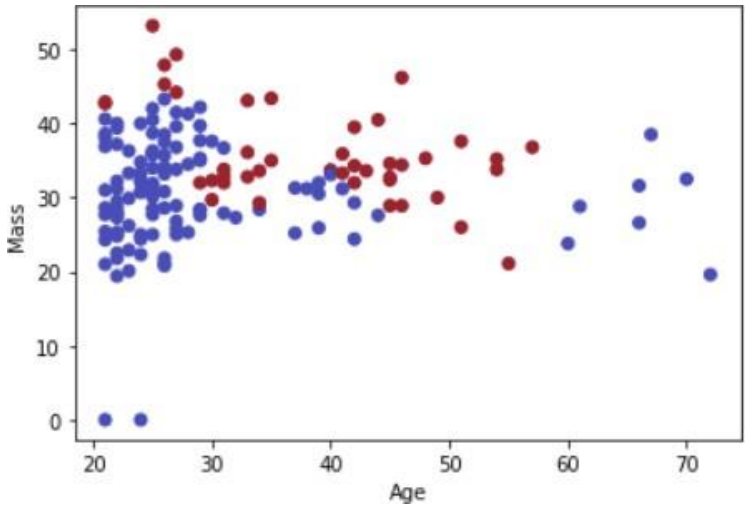
2D Projection for K=3



2D Projection for K=6



For K=23

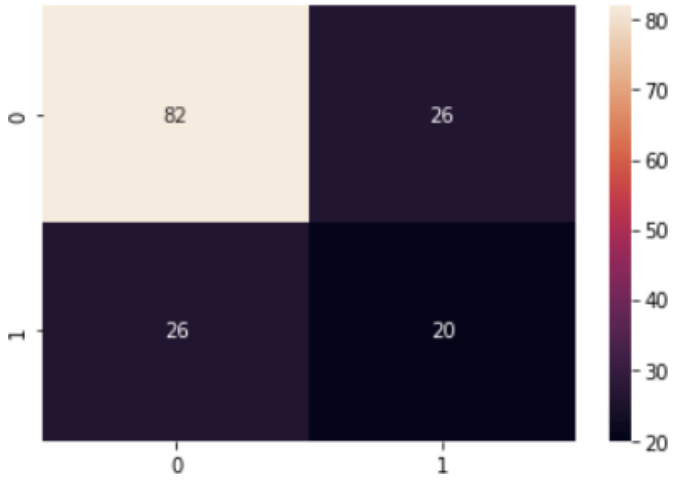


5. Interpreting and comparing the results for K=3, 6, and 12 is:

confusion matrix for k=3

[[82 26]
[26 20]]

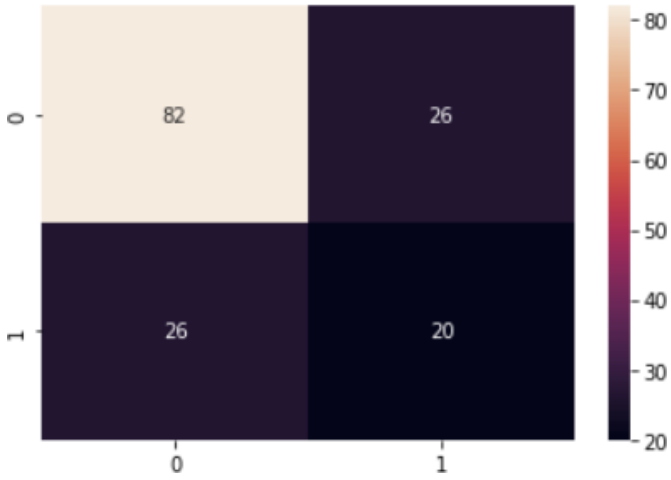
AxesSubplot(0.125,0.125;0.62x0.755)



confusion matrix for k=6

[[90 18]
[33 13]]

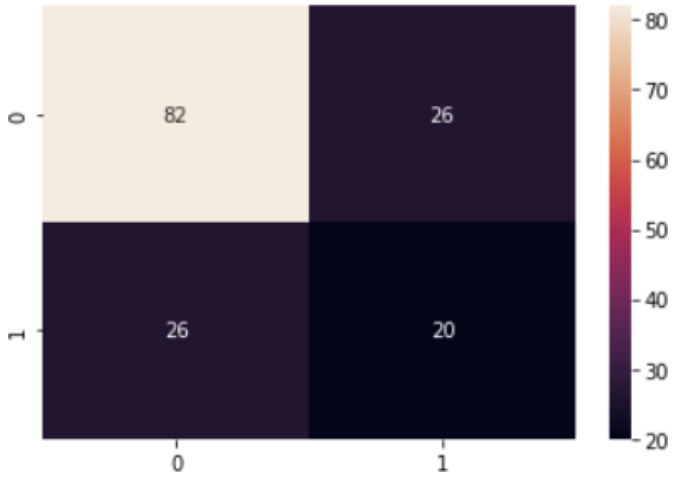
AxesSubplot(0.125,0.125;0.62x0.755)



confusion matrix for k=23

[[86 22]
[25 21]]

AxesSubplot(0.125,0.125;0.62x0.755)



classification report for k=3

	precision	recall	f1-score	support
0	0.76	0.76	0.76	108
1	0.43	0.43	0.43	46
accuracy			0.66	154
macro avg	0.60	0.60	0.60	154
weighted avg	0.66	0.66	0.66	154

classification report for k=6

	precision	recall	f1-score	support
0	0.73	0.83	0.78	108
1	0.42	0.28	0.34	46
accuracy			0.67	154
macro avg	0.58	0.56	0.56	154
weighted avg	0.64	0.67	0.65	154

classification report for k=23

	precision	recall	f1-score	support
0	0.77	0.80	0.79	108
1	0.49	0.46	0.47	46
accuracy			0.69	154
macro avg	0.63	0.63	0.63	154
weighted avg	0.69	0.69	0.69	154

PLOT OF ALL K VALUES AND ACCURACY: