IML REPORT LAB -3

B21ME047

PRANJAL VERMA

Q.1

(a) We used panda library to import the csv file.

(b) The algorithm for noise removal involves removing all rows in which “Insulin” exceeds Mean + 2* (standard deviation) ,since 95% of values lies inside ‘mean+2standard deviation’.

(c) We used `train_test_split` library to perform splitting of datasets in ratio of 70:20:10.

Q.2

(a) We used `KNeighborsClassifier` file from `sklearn.neighbors` library.

We used Minowski distance as metric.

(b) We varied the k value from 4 to 10.

```
Accuracy --> k =  (4) 70.12987012987013
Accuracy --> k =  (5) 71.42857142857143
Accuracy --> k =  (6) 70.77922077922078
Accuracy --> k =  (7) 74.02597402597402
Accuracy --> k =  (8) 73.37662337662337
Accuracy --> k =  (9) 74.02597402597402
Accuracy --> k =  (10) 74.02597402597402
```

k value is maximum for 7,9 and 10.

(c) In K-Nearest Neighbors there is no learning required as the model stores the entire dataset and classifies data points based on the points that are similar to it. It makes predictions based on the training data only.

Ways to calculate the distance in KNN:

- Manhattan Method
- Euclidean Method
- Minkowski Method
- mahalanobis distance

Implementing K-Nearest Neighbors from Scratch in Python

Step 1. Figure out an appropriate distance metric to calculate the distance between the data points.

Step 2. Store the distance in an array and sort it according to the ascending order of their distances (preserving the index i.e. can use NumPy argsort method).

Step 3. Select the first K elements in the sorted list.

Step 4. Perform the majority Voting and the class with the maximum number of occurrences will be assigned as the new class for the data point to be classified.

Our predict function requires a Training dataset, True Labels, Datapoints to classify, and the number of nearest neighbor (K) as the input arguments.

Q.3

(a) I calculated standard deviation for all the features and found that features **Pregnancies , SkinThickness** have maximum standard deviation.

I made new dataset with these existing features.

(b) We imported `GaussianNB` file from `sklearn.naive_bayes` library.

We used bin size = 4. It resulted in lower accuracy compared to K-NN classifier.