**KENNESAW STATE UNIVERSITY CS7267 – Machine Learning**

|  |
| --- |
| Fall 2022 Assignment 2 |
| September 20, 2022 |

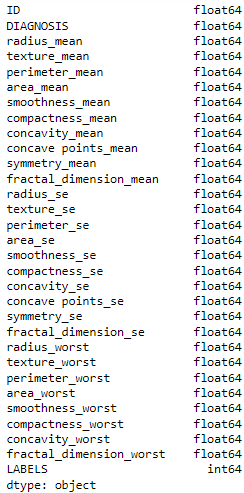
Student Name: .......................................................................................................

Student Number: .....................................................................................................

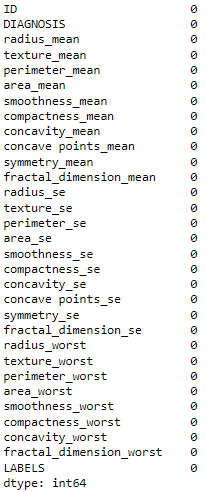
**Abstraction\_** Breast cancer is a condition in which the breast cells proliferate out of control. Breast cancer comes in several forms. Which breast cells develop into cancer determines the type of breast cancer. The k-nearest neighbor’s algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to producing classifications or predictions about the grouping of a single data point. Although it can be applied to classification or regression issues, it is commonly employed as a classification algorithm because it relies on the idea that comparable points can be discovered close to one another. Although this work describes the implementation of the KNN algorithm for the classification purpose using the data set [wdbc.data.mb.csv](http://localhost:8888/edit/mledge%20iiot%20data%20set/wdbc.data.mb.csv) available on the website UCI Machine learning data sets. The research aims to find the classification accuracy for the value of k=1,3,5,7,9 and implement a basic KNN algorithm to classify the given data set. The algorithm gives the accuracy for different k values as 91%, 97%, 97%, 96%, and 96%.

# Results and Discussions:

As the KNN algorithm gives the best accuracy for the proposed data set. The data set contains 31 attributes and 612 observations(tuples). The attributes are having the data types float64 **Figure 1**.



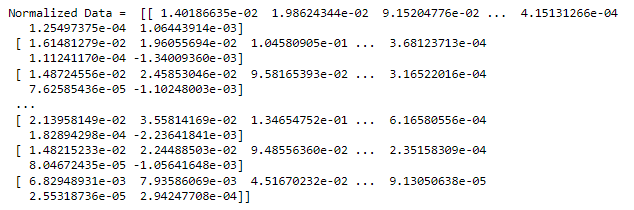
**Figure 1 Attributes and Data Types**



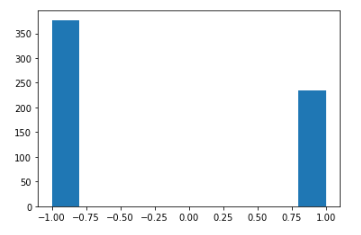
**Figure 2 Process data set**

**Figure 2** describes that the dataset does not contains any null or Nan values. Hence, we can process our analysis further.

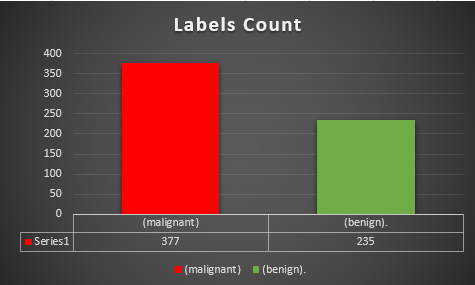
Data organization in a database is done by normalization. This entails building tables and linking those tables together in accordance with principles intended to safeguard the data and increase the database's adaptability by removing duplication and inconsistent reliance. Hence the proposed model classifier KNN performs best when the data is normalized. **Figure 3**



**Figure 3 Normalized Data set**

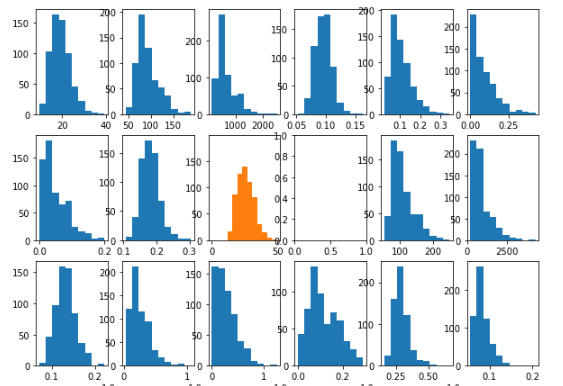


**Figure 4 Labels Count**



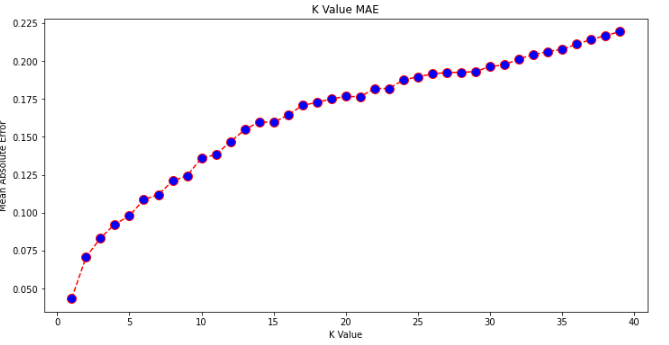
**Figure 5 LABELS COUNT**

**Figure 4** describes the DESTRIBUTION total number of available binary labels. Similarly, **Figure 5** describes the total number of labels (1, and -1).



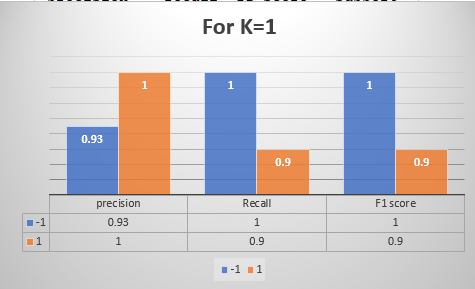
**Figure 6 Attributes distribution**

***Figure 6*** describes the distribution of the attributes in the data set.

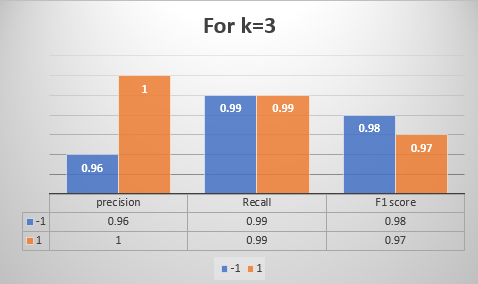


**Figure 7 Mean Absolute Error**

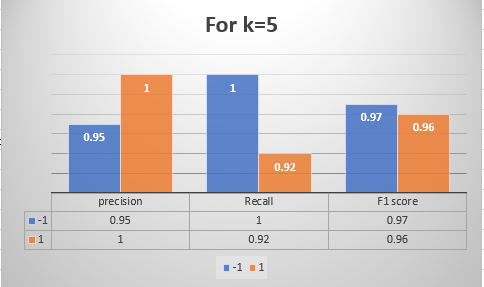
**Figure 7** shows the mean absolute error of the available data set. Similarly ***Figure 8***, **Figure 9**, **Figure 10**, **Figure 11**, **Figure 12** shows the metrics evaluation of the model classifier KNN. Hence the last figure **Figure 13** describes the overall accuracy of the model classifier for the values of K=1,3,5,7, and 9.



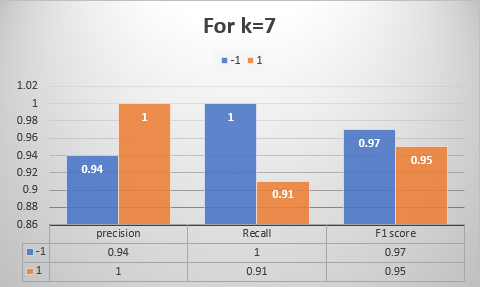
**Figure 8 Confusion Graph for K=1**



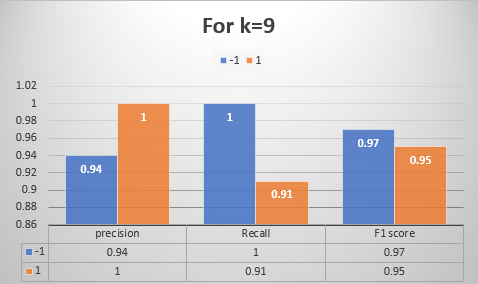
**Figure 9 Confusion Matrix for k=3**



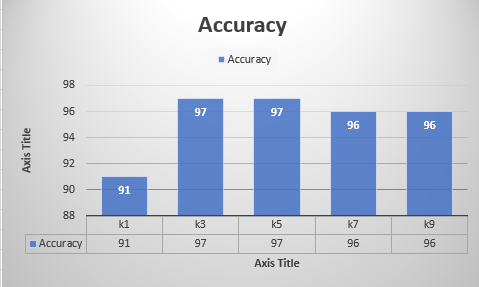
**Figure 10 Confusion Metrics for K=5**



**Figure 11 Confusion metrics for K=7**



**Figure 12 Confusion Metrics for K=7**



**Figure 13 Overall Accuracy**