# DOCUMENTATION

## Overview

This assignment consists of Classification and Clustering done Via kNN and K-Means respectfully. It later tests its efficiency via F1/accuracy/recall/precision for kNN and Davies-Bouldin Index for Clustering.

## Sorting Data

**Mammographic Mass for k-NN**

```
[786 rows x 6 columns] (After cleaning and normalization)
```

http://archive.ics.uci.edu/ml/datasets/mammographic+mass

**Messidor.data for Clustering**

```
[813 rows x 10 columns] (After cleaning and normalization)
```

https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set

- The data was first converted into panda and cleaned to normalize it between 0-1 so that one column does not have the most significant impact and making it more accurate.
- Outliers were removed using the quartile method.
- None values replaced with mean of the columns.

## k-NN Classification

k-NN Classification is used to identify the class of a new input from the dataset.

Shape 1,2,3,4 are used as classes here.

- The Mammographic Mass dataset is divided in a 70% for training set and 30% test set division.
- The Training set is initiated first and is given the classes identified inside the data
- The Test set then runs through the training set data. Picks the closest K choices via euclidean distance inside the find_closest_knn_point.
- The K indexes are stored and used in the find_test_class function. It guesses the class for the test set and stores it.

## Evaluation Metrics

NOTE: The mammographic mass had a precision and f1 etc score of around 80% But it was binary classification using the severity as the binary classes.
so I used shape for the classification which was not originally meant to be a class so my overall evaluation metrics reduced. It can easily be fixed with a new dataset but I would have to change my code and everything runs fine as it (the functions and estimations are correct) Is so I didn't do so.

Precision answers what proportion of predicted Positives is truly Positive?

Recall answers what proportion of actual Positives is correctly classified?

F1 = mixture of both as some datasets have more precision and some have more recall. It gives more weightage to the lower value.

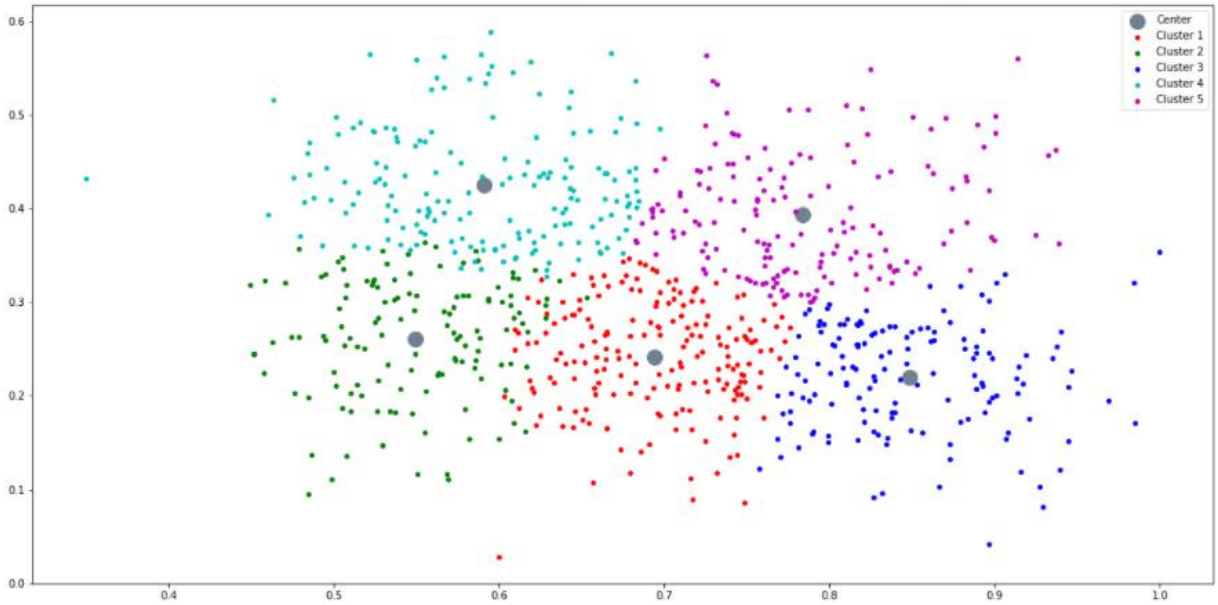
- Confusion Matrix of 4x4 made (according to the number of classes in the classification)
- Precision found via True and Predicted / (All values in the predicted column)
- Recall found via True and Predicted / (All values in the predicted row)
- F1 = 2*((precision*recall)/(precision+recall))
- Macro values = Average value of all classes individual values.
- Micro Values = adding all of them together and has equal precision recall f1.

## K-Means Clustering

In Clustering the aim is to separate groups with similar traits and assign them into clusters.

- K-means taken as input. Initial centroid made by picking a random value from the dataset for each k centroid. Initial cluster made from the centroid.
- Cluster given positions according to the nearest centroid in find_closest_centroids using Euclidean distance.
- Set iterations made in which after each iteration the center is computed again and is given the position of the average position of each cluster.
- Divided the data according to clusters.

## Plot




- Features selected according to the which dataset column was the most different.
- EU_Dist
- Diameter _ picked as features.

## Davies-Bouldin Index

The smaller the DB index the more different the values for each cluster are and the better the quality of the DB index.

- Distance measure must be same as the one used in clustering.
- P = 2 as features are 2
- Some Variations because less iterations are taken for less computation time and random points are taken.