

Breast Cancer Diagnosis Prediction using K-Nearest Neighbors (KNN) and K-Means Clustering

It is my implementation to the Breast Cancer Diagnosis Prediction. The aim was to categorize the cases of cancer as Malignant (M) and Benign (B) with the help of KNN algorithm and also use K-Means clustering in order to monitor the manner in which the data is self grouping. I attempted to make it all simple and do everything as per the instructions offered in the assignment.

What I Did

1. Data Preprocessing

Translated the column of diagnoses:

M - 1, B - 0

Deleted the id and Unnamed columns because they are not useful in prediction.

Min-Max Normalization of the features.

Divide the data into two parts, 80 percent training and 20 percent testing.

2. K-Means Clustering

Used $k = 2$ clusters

Rejoined the cluster labels to the dataset to observe the way the algorithm clusters malignant and benign cases.

3. KNN Classification

Trained the model of KNN with 5 neighbors.

Evaluated it using:

- Accuracy
- Precision
- Recall
- F1-score
- Results

Here are my scores:

Metric Score

Accuracy	0.9649
Precision	0.9534
Recall	0.9534
F1-score	0.9534

These are standard values to this dataset and in most cases the model would be able to discriminate between a malignant and benign case.

K-Means Cluster Plot

The two seed features used to divide the data into 2 clusters.

How to Run

Open google colab notebook.

Upload the dataset.csv file.

Run the cells in order.

All that should just work without additional installation.

Final Thoughts

KNN model performed well and the clustering was also sensible. This assignment allowed me to gain experience in data preprocessing, clustering, classification, and evaluation on a complete workflow of ML processes.