Breast Cancer Wisconsin (Diagnostic) Data Set

Sanjana Senthilkumar
UC Riverside
ssent013@ucr.edu

Kishan Sivakumar
UC Riverside
ksiva011@ucr.edu

Akash S M
UC Riverside
asund016@ucr.edu

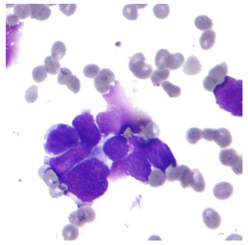
Deepak Urs G V
UC Riverside
dgage002@ucr.edu

Shadhrush Swaroop
UC Riverside
sswar010@ucr.edu

Puneet Singhania
UC Riverside
psing088@ucr.edu

Introduction

- The Breast Cancer Wisconsin (Diagnostic) Data Set is a dataset that contains information on breast cancer tumors, including measurements obtained from digitized images of fine needle aspirates.



Problem Definition

- The first problem is a binary classification task, where the goal is to classify a given breast mass as malignant or benign based on its numerical features.
- The second problem is a clustering task, where the goal is to group similar breast masses into coherent clusters based on their numerical features.

Related Work

- The dataset has been widely used in various machine learning approaches, including support vector machines, neural networks, decision trees, and clustering algorithms.
- Several studies have focused on improving model performance by utilizing feature selection and engineering techniques such as PCA, LDA, and wavelet transforms.

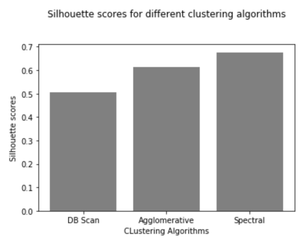
Proposed Method

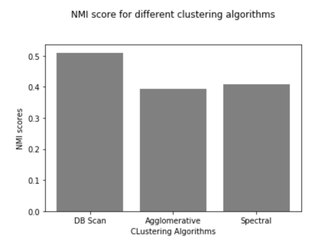
- To solve the classification and clustering problems in the Breast Cancer Wisconsin (Diagnostic) Kaggle competition, a proposed method involves utilizing six machine learning algorithms.

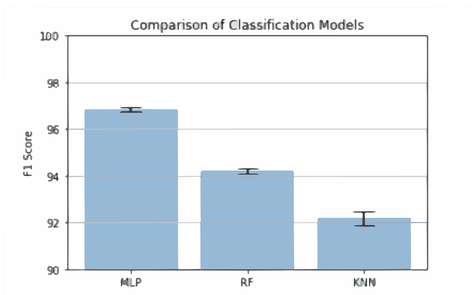
➢ Random Forest classifier
➢ Multi Layer Perceptron
➢ K-nearest neighbors classifier
➢ DBSCAN clustering
➢ Spectral clustering
➢ Agglomerative Clustering with Single Linkage.

- The selection of these algorithms aims to offer a range of approaches to address the classification and clustering problems, with the goal of achieving high accuracy and performance.

Results







Conclusions

- We could see that the classification models did better than the clustering models.
- The performance of all the clustering models were close to each other with silhouette scores between 0.5 - 0.6.
- The F1 scores of the classification models are between 0.88 to 0.98.
- In addition, the implementation correctness provided in the project corroborates the above results.

UC RIVERSIDE