Breast Cancer Classification and Clustering using

Machine Learning Approach

**Introduction:**

Breast cancer is a major concern for public health in the United States, affecting a considerable number of women every year. Statistics demonstrate that breast cancer is among the primary causes of death in women, and the incidence of new cases is expected to increase in the near future.

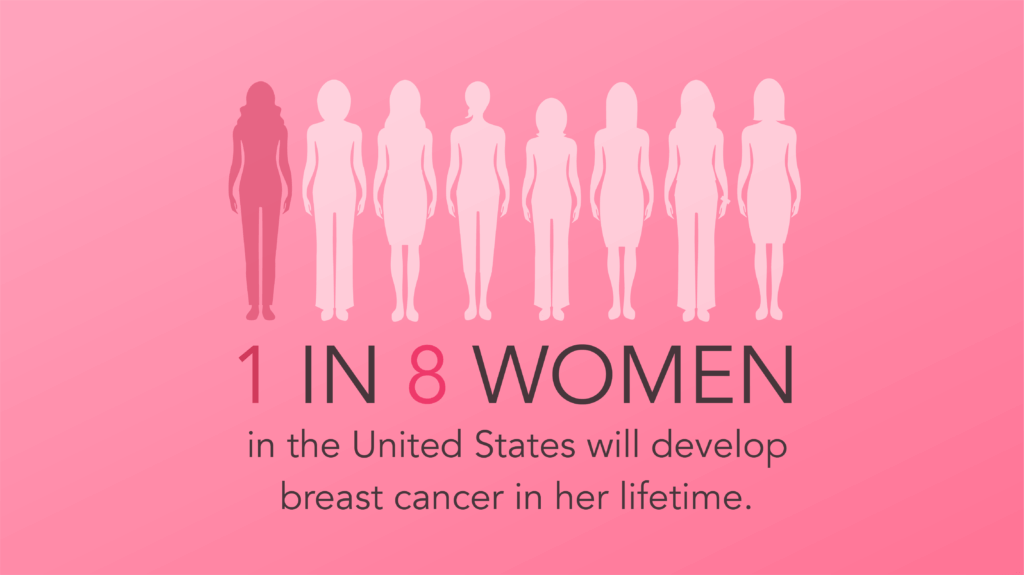


Fig 1: The Research motivation for Breast cancer detection using the ML approach (\*WHO)

Given that one in every eight women in the US could develop breast cancer at some point in their lives, it is essential to continue exploring the factors that cause it, the risks involved, and potential treatments. Despite progress in early detection and treatment, breast cancer still causes the loss of thousands of lives annually, while many survivors confront prolonged physical and emotional challenges.

The high number of breast cancer fatalities in 2021, reaching 43,600, highlights the pressing need for better therapies and interventions. These numbers are more than just statistics, as they represent a tragic loss for families, friends, and communities that have been affected by this illness.

Moreover, the projected 2.5 million invasive breast cancer diagnoses in 2022 underscore the necessity of ongoing research to enhance our comprehension of the disease, recognize novel risk factors, and develop more effective therapies.

On a positive note, the fact that 3.8 million people survived breast cancer in the US in 2021 is an indication of progress in the field. However, survivors face continuing challenges, including the possibility of recurrence and long-term adverse effects from treatments. As a result, further research is critical to improving the quality of life of breast cancer survivors and enhancing their long-term outcomes. In this project, we will use a machine learning approach to find early detection, forecasting, and fining pattern of the 2 types of breast cancer (malignant, and benign) and will do the exploratory data analysis using classification and clustering algorithm.

**Methodology:**

1. Access and Load data

2. Preprocess

3. Derive Features

4. Model Training

5. Model Tuning

6. Result

Fig 2: Generic Machine Learning workflow.

**The reason we should use machine learning:**

Machine learning can solve problems that traditional classification and clustering methods cannot due to its ability to detect intricate patterns and relationships in data that traditional methods may miss. Machine learning algorithms can learn from data, attributes, and data correlation and enhance their performance over time, while traditional techniques are based on predefined rules and assumptions that may not be flexible enough to handle complex data structures or dynamic patterns.

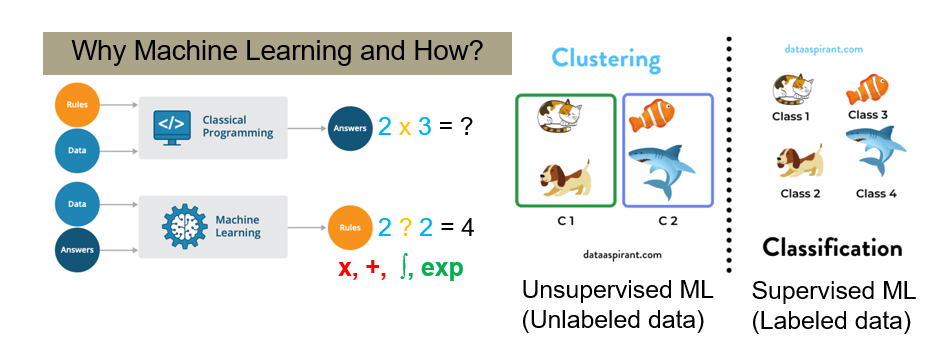


Fig: 3. Classic example of how we could find the data correlation and solve the problem using ML

For example, in image recognition tasks, machine learning algorithms can identify relevant features like edges, shapes, and textures in a large dataset of images to classify new images with high accuracy, even without prior exposure. On the other hand, traditional image recognition methods rely on hand-crafted features and heuristics that may not work effectively on new images or datasets.

In conclusion, machine learning can solve problems that traditional classification and clustering techniques cannot because of its superior ability to detect and learn intricate patterns and relationships in data.

**The typical workflow for machine learning (ML) involves several stages:**

Initially, relevant data is collected from diverse sources such as databases, APIs, or web scraping, and is subsequently processed and cleaned by removing duplicates, handling missing values, and transforming it into a usable format. After this, feature engineering is conducted, where the most appropriate data features are chosen and transformed for the purpose of training the ML model. This includes tasks such as feature scaling, normalization, and one-hot encoding. Following this, an appropriate ML model is chosen and trained by feeding the prepared data into an algorithm such as linear regression, decision trees, or neural networks. Once the model is trained, it is evaluated to assess its accuracy and effectiveness, which may include using metrics such as precision, recall, and F1-score, and performing cross-validation to prevent overfitting. If necessary, adjustments may be made to the feature selection, algorithm, or hyperparameters. After this, the model can be deployed into a production environment to make predictions on new data by integrating it into an existing system or creating a new application. Finally, the ML model must be monitored and maintained over time to ensure it continues to perform effectively by periodically retraining the model as new data becomes available and updating its algorithms or hyperparameters to improve its accuracy and effectiveness.

1. **Access the data:**

**Data source: Breast Cancer Wisconsin (Diagnostic) Data Set (1995)**

<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

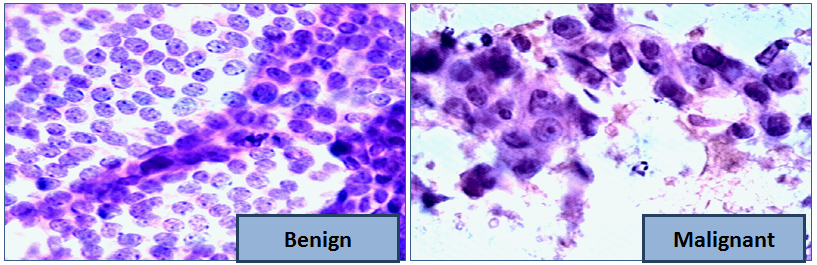


Fig 4: The lab-based sample of 2 breast cancer (\*google.com)

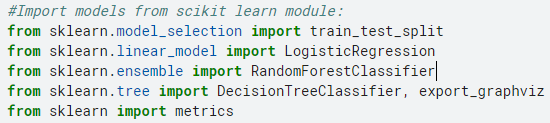
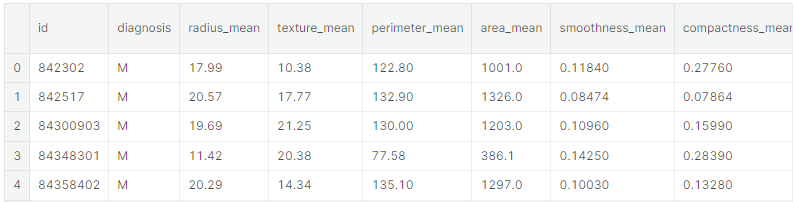
Size- 569 Cancer Samples

Class/ Data Labeling: 2 classes, Benign (B), and Malignant (M)

Class Distribution: Benign (62.74%), and Malignant (37.26 %)

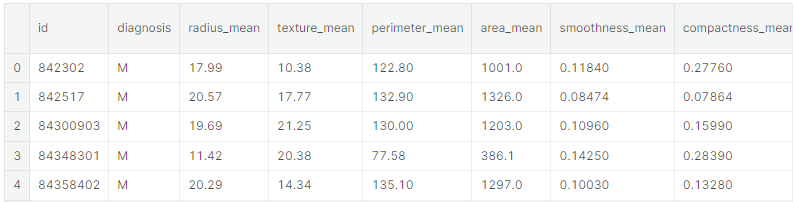
Data Attributes: 32, clean text (string), category (numeric)

**Load the packages and access the data:**

****

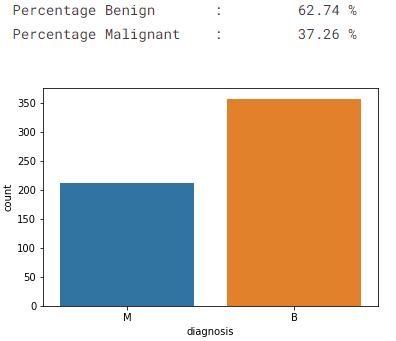
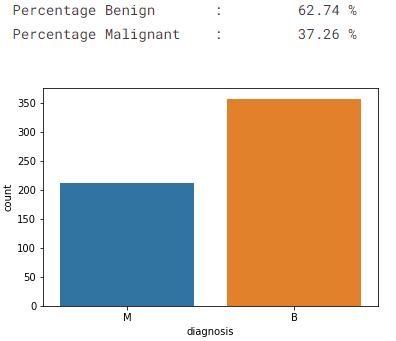


1. **Preprocess the data:**
2. **Drop nulls, and redundant data and normalize the data**

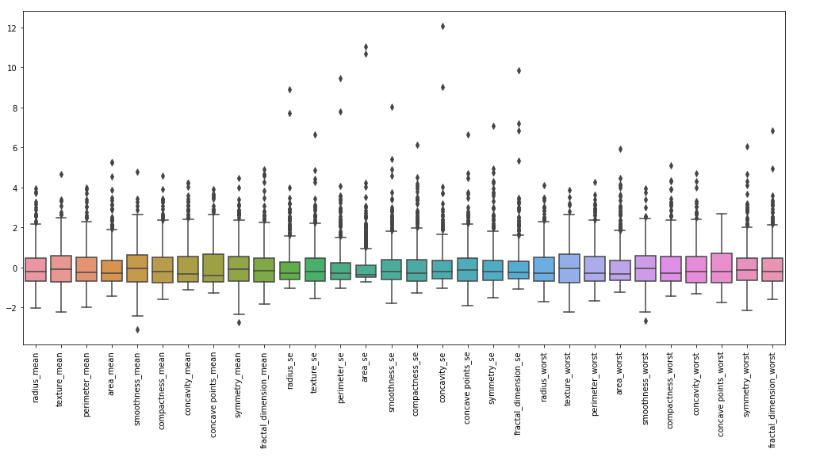




1. **Check the data distribution:**



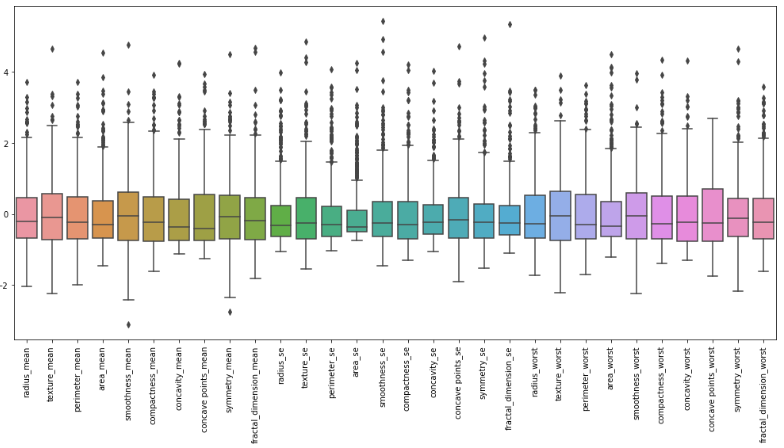
1. **Remove outliers:**



Outliers

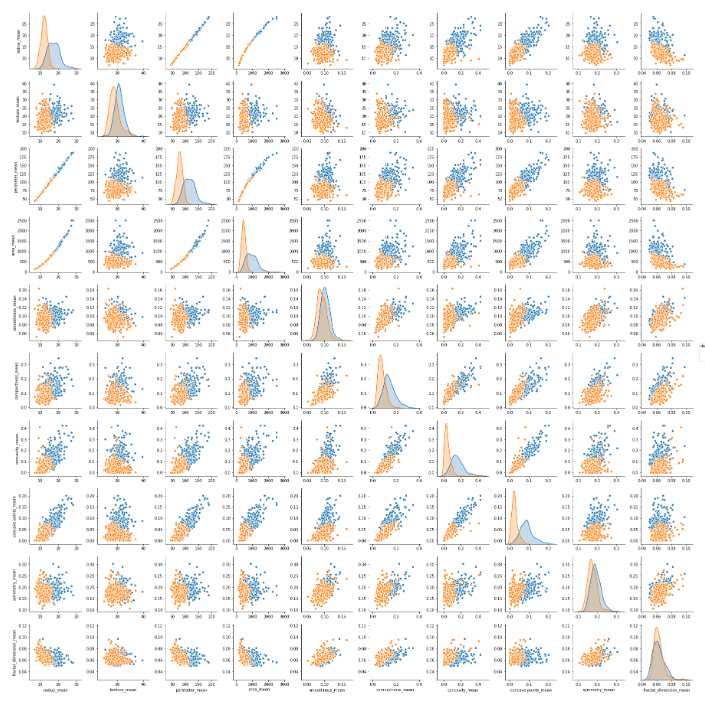
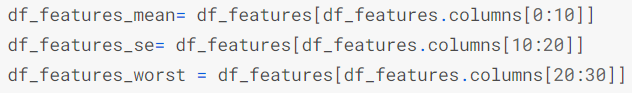


Remove Outliers



No Outliers

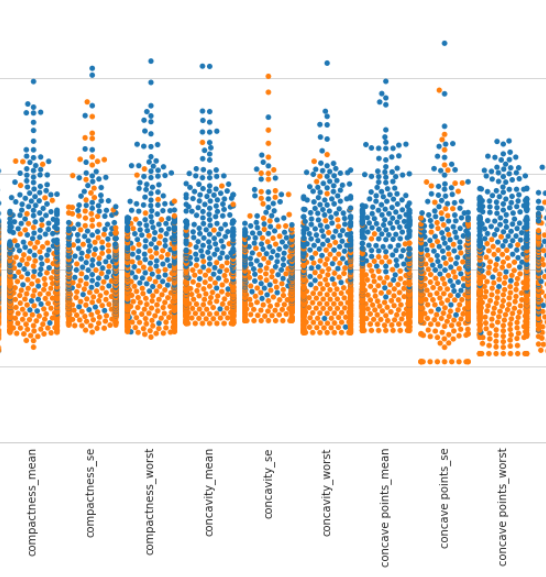
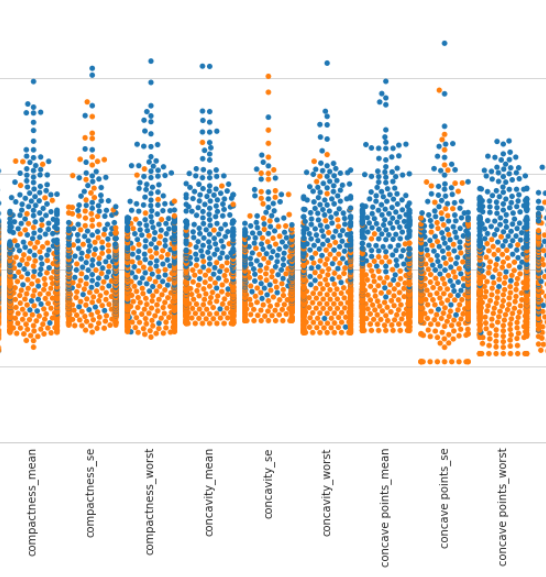
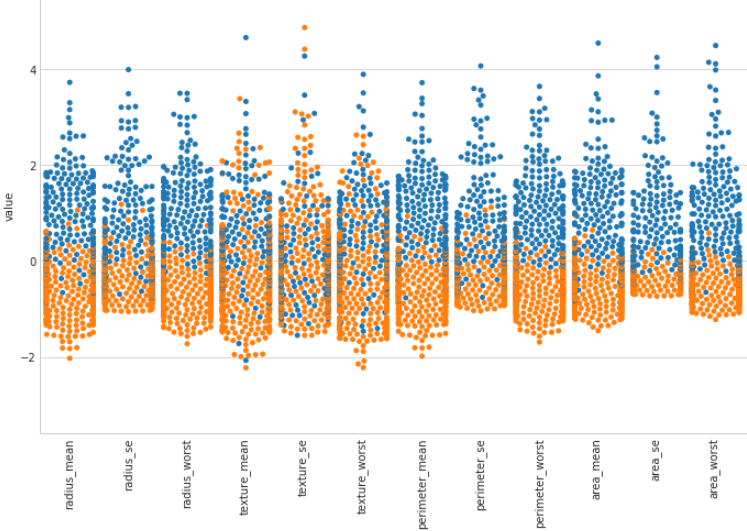
1. **Feature Engineering:**
2. **Check the data correlation:**



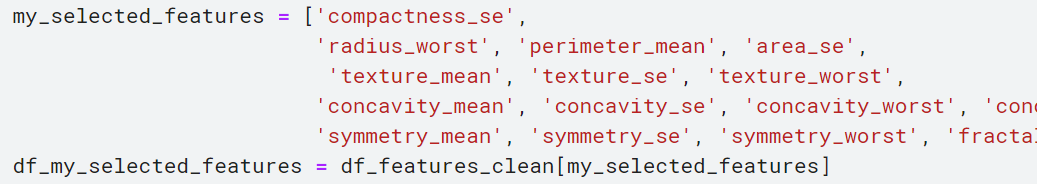
Pair plots

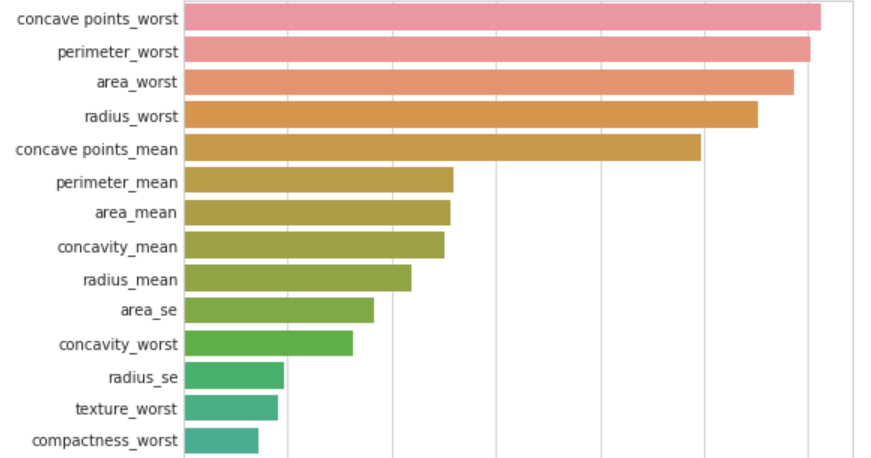
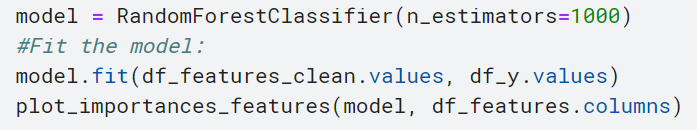
* **Outcome: Radius, area, perimeter and compactness, concave points, and concavity are highly Correlated**
* **Statistical distribution: Radius, area, perimeter and compactness, concave points, and concavity are highly Correlated**

1. **Choose the Correct Feature:**



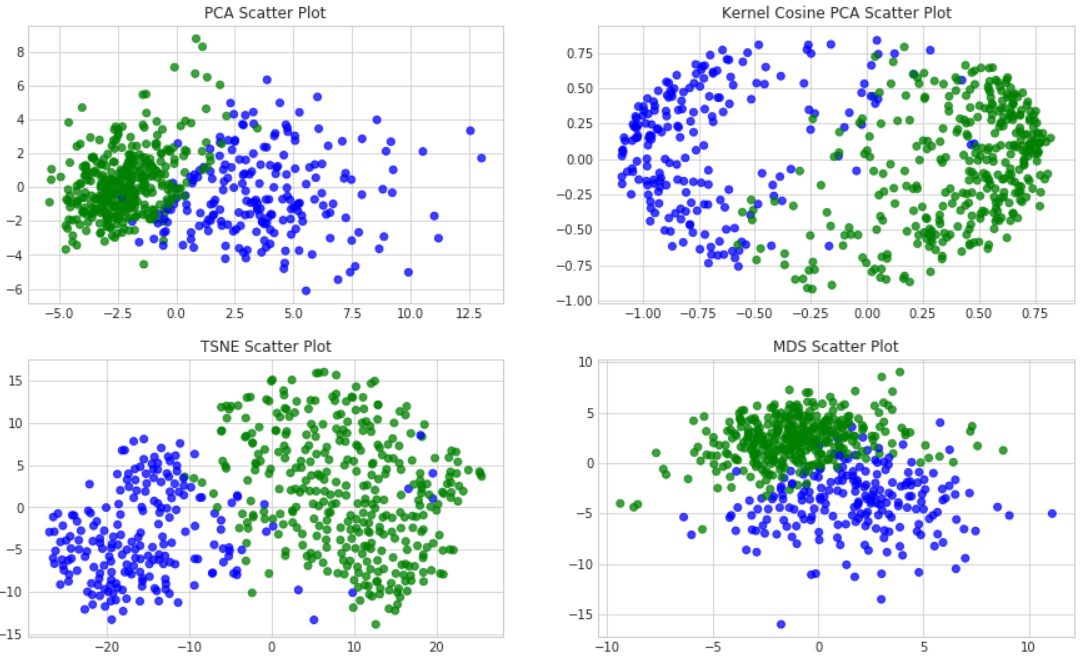
**We should select the features that are most discriminating (swarm plot) among the most correlated features (correlation plot)**

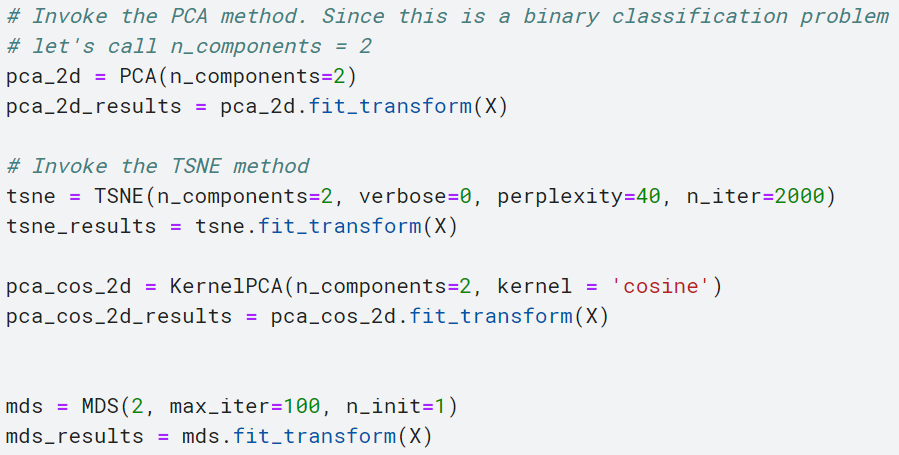
****

1. **Selected Feature Ranking:**
2. **Dimensionality reduction:**

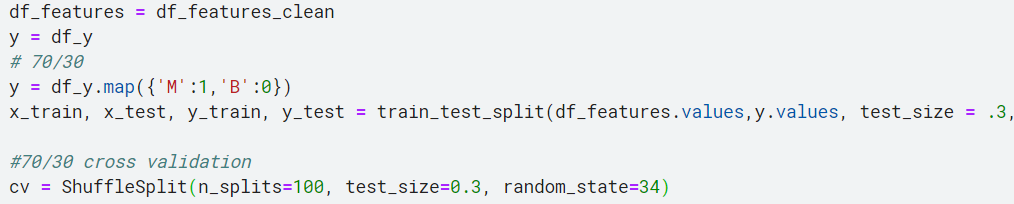
The reason behind the dimensionality reduction:

* Too many variables
* Too complex visualization problems
* Decrease efficiency by including variables that have no effect on the analysis
* Make data interpretation difficult
* Understanding how to simplify the dataset, without losing relevant information 🡪 PCA, TSNE, MDS algorithm.





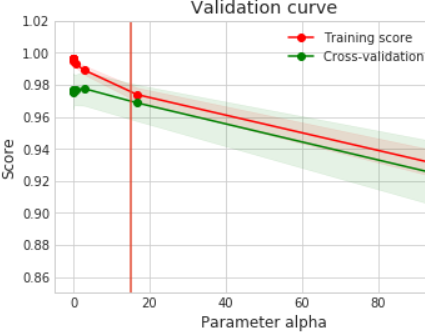
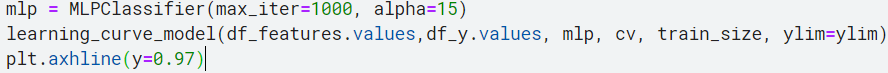
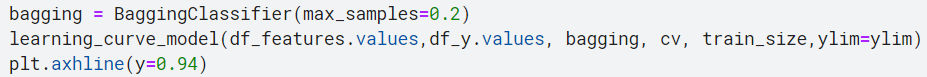
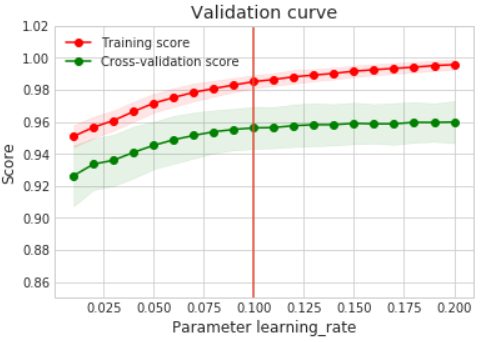
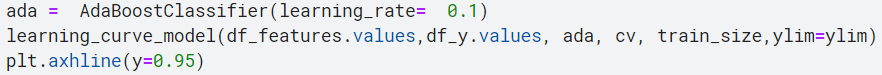
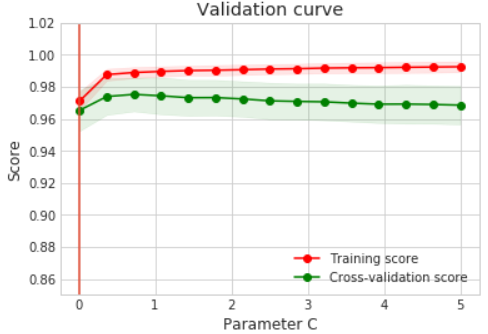
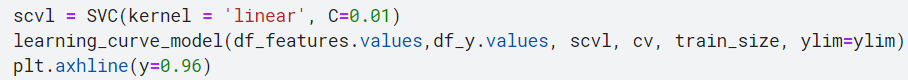
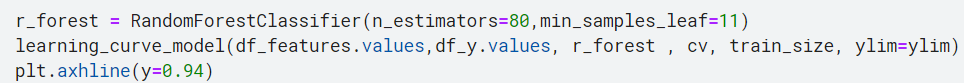
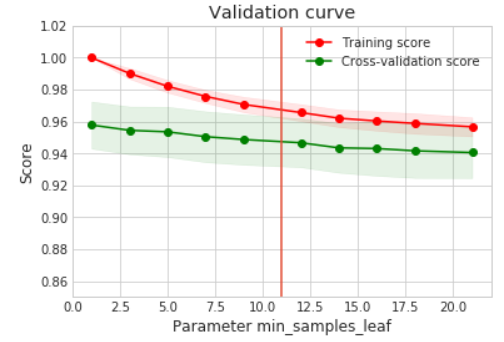
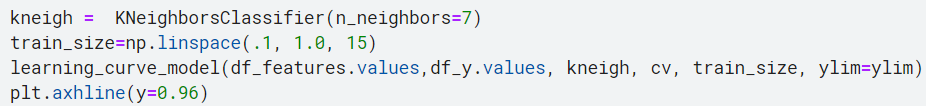
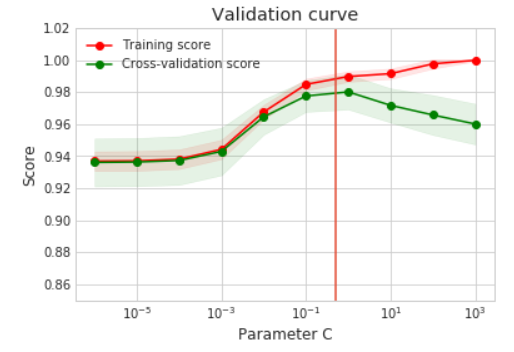
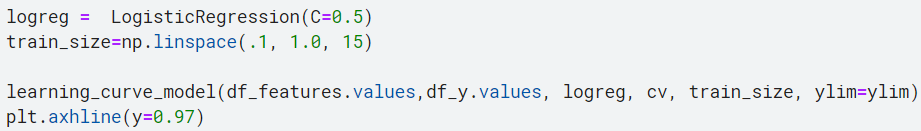
1. **Train, Validation, and Test split:**



Train: Validation: Test: 40%: 30%: 30%

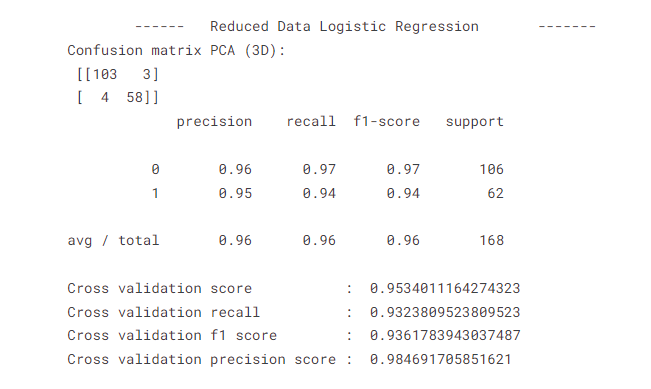
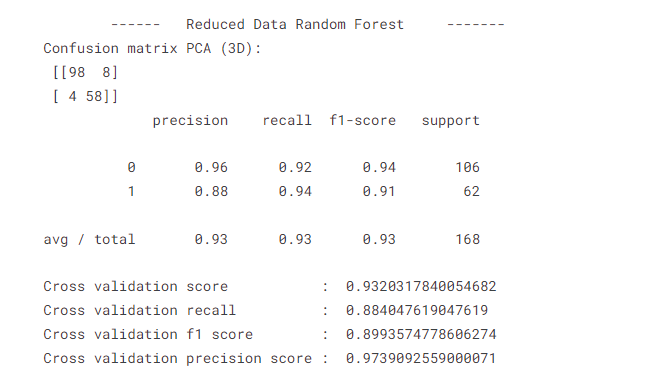
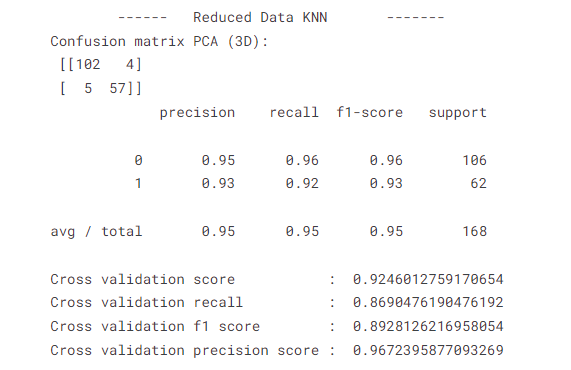
1. **Train the model**

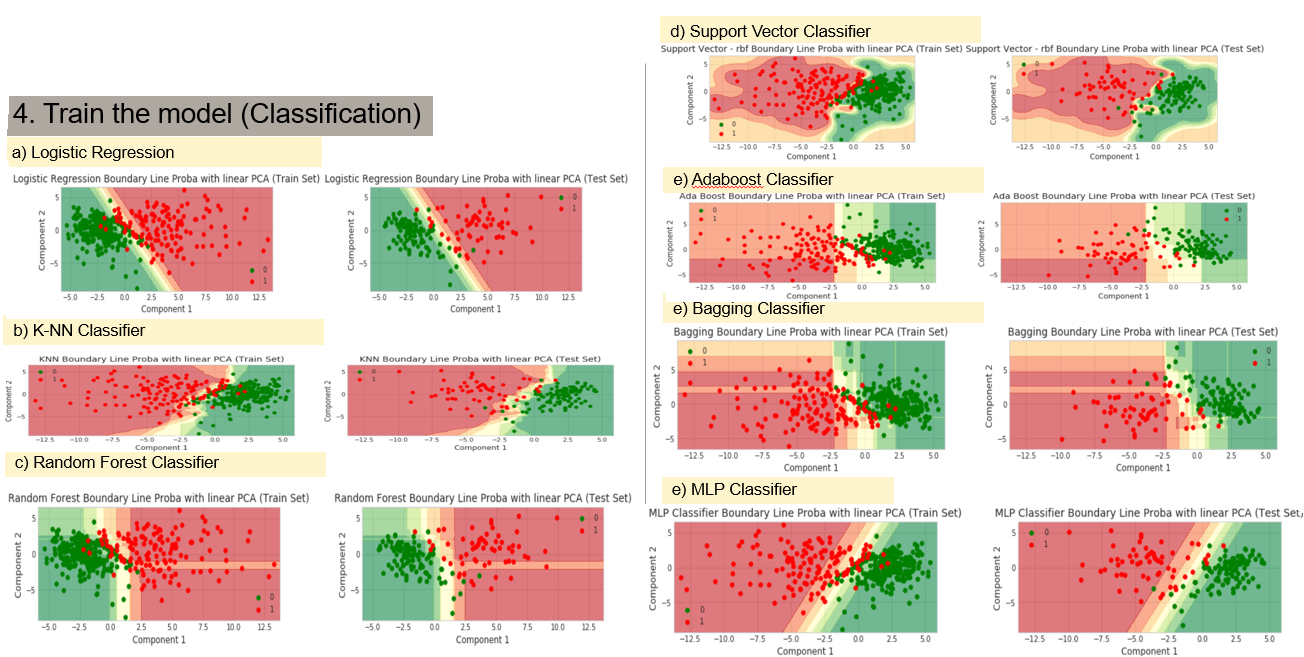
**We will train 7 classification ML models (LR, KNN, Random Forest, SVM, Adaboost, Bagging, MLP etc.)**



**Result analysis of Classification algorithm:**

**(For display purposes we just mentioned the best few of them but we have all 7 classification results. To see those please find those on the code)**

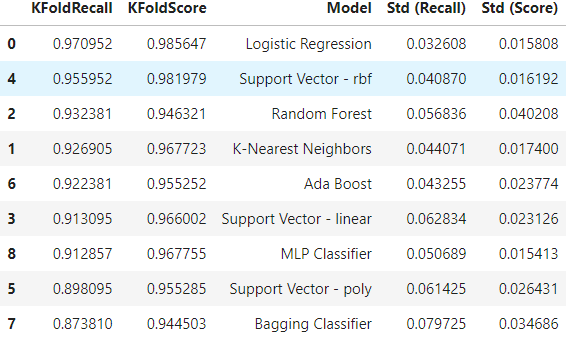




**Fig: 5 Describe all class boundaries from ML train (7 algorithms)**

1. **Evaluate the best model**

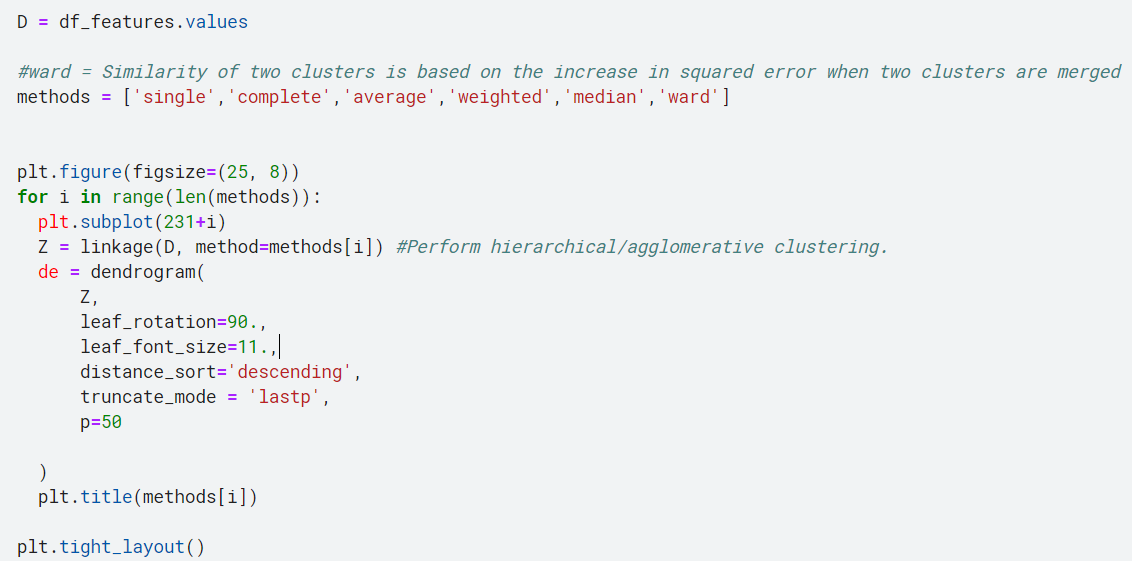
**Based on precision, recall, F1 score, and sensitivity Linear regression is the best classifier**

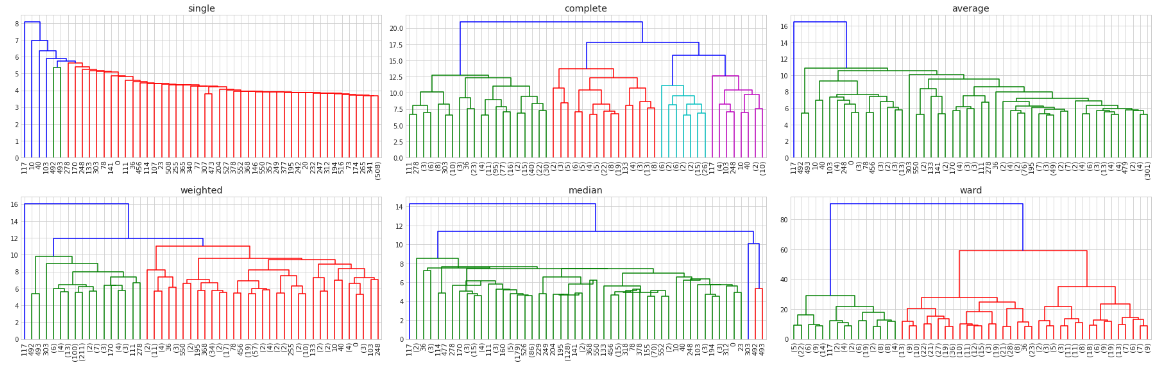
****

1. **Train the model**

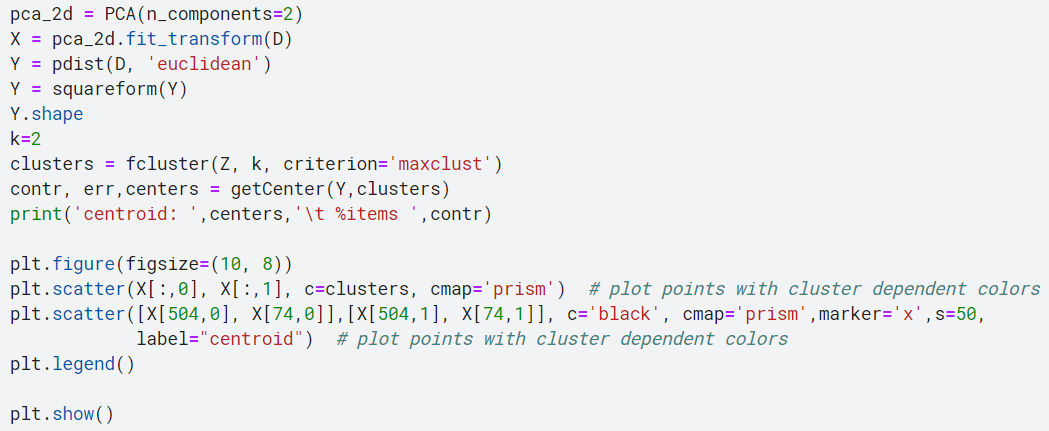
**We will train 2 clustering ML models (Hierarchical and K-means clustering etc.)**

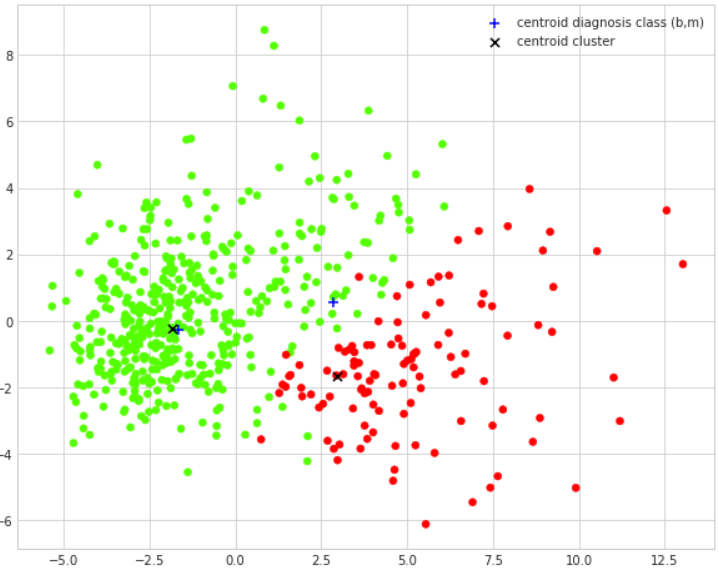
1. **Hierarchical clustering:**
2. **Find the shape of the clusters: Ward linkage method**

****

****

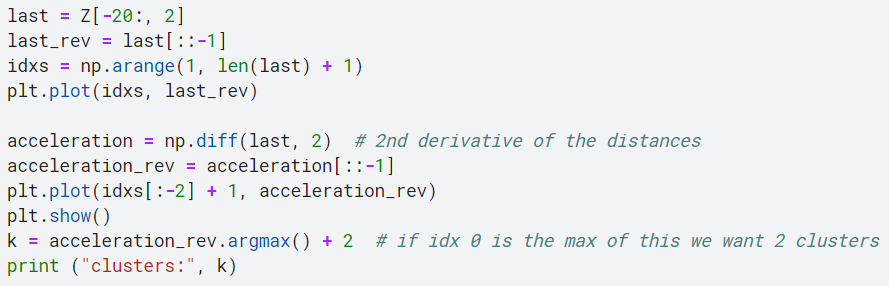
**We have found that a complete shape will fit best for our data (complete tree types)**

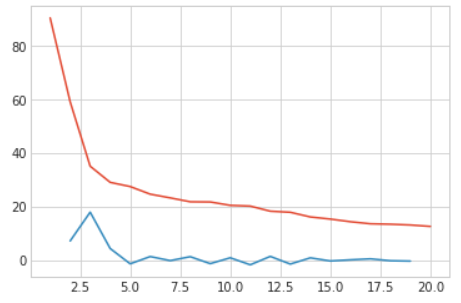
**2) Apply hierarchical clustering**

****

**Fig 6: Best number of cluster K= 2for complete hierarchical shape (The ML assumes it as 2 classes: a) Benign, 2) Malignant)**

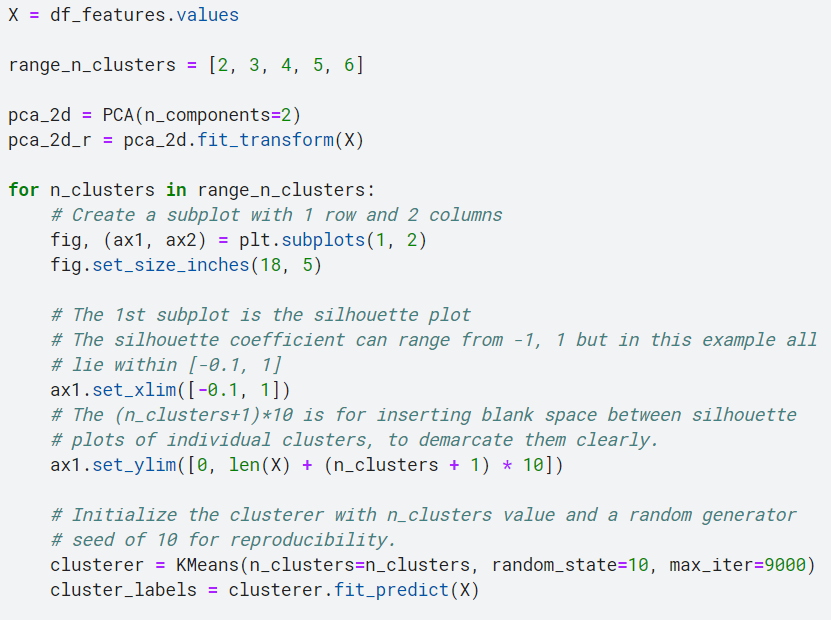
1. **K-means clustering**
2. **How many K we should choose: We need to apply the Elbow method**

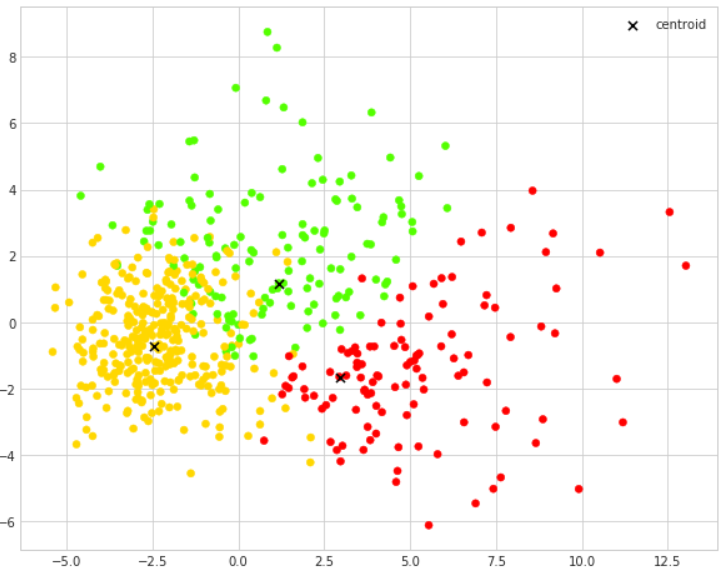
****

****

**So, K should be 3 because acceleration got a peak after 2.5**

****

****

****

**Fig 7: Best number of cluster K= 3 for complete K-means (The ML assumes it as 3 classes: a) Benign, 2) Malignant) 3) exception this two cancer.**

**Conclusion and Summary:**

* **Explained the exploratory data analysis and data distribution**
* **Showed how to choose the best feature, feature ranking**
* **Explained the dimensionality reduction – PCA approach**
* **Analyze 7 classification and 2 Clustering algorithms**
  + - **Accuracy wise Logistic Regression is the winner (98.5%)**
    - **Clustering shape and feature wise: K-means is the winner (3-means)**

Overall, the application of machine learning in breast cancer classification and clustering has demonstrated its effectiveness. By utilizing a range of ML algorithms, experts have been successful in precisely classifying breast cancer tumors and detecting significant features that can aid in cancer diagnosis and treatment. Additionally, the use of clustering algorithms has enabled the identification of unique subtypes of breast cancer that may necessitate diverse treatment strategies. Despite the need for further research in this area, the advancements made thus far exhibit the capacity of machine learning to enhance our comprehension and management of breast cancer.

**References:**

1. https://www.kaggle.com/code/dreamslab/eda-classification-and-clustering-diagnostic
2. https://www.kaggle.com/code/yoghurtpatil/clustering-and-classification-on-student-data
3. https://www.kaggle.com/code/pratik1120/penguin-dataset-eda-classification-and-clustering