# 

# **Detection of breast cancer**

Introduction

Cancer is a condition when cells proliferate uncontrollably. A woman's chance of dying from breast cancer is 1 in 39, or roughly 2.5%. In the United States, about 42,000 women die from breast cancer each year, out of the 264,000 who are diagnosed. One of the most prevalent malignancies in women worldwide is breast cancer. According to estimates, 30% of newly diagnosed malignancies in women in the United States in 2020 will be breast cancer. A higher survival percentage and more effective treatment choices are both possible with early identification of breast cancer. We will talk about different machine learning models and how well they perform when determining if a tumor is benign or malignant in this session.

Aim

93% of women with breast cancer who are diagnosed early will survive for the first five years. Mammogram screening is used to find the tumor, but it is not completely accurate. These statistics suggest using machine learning to diagnose breast cancer as the conventional method is not very reliable.

Our goal in this project is to examine the widely available Wisconsin Breast Cancer dataset and assess how well various machine learning models perform at identifying tumor malignancy. We can evaluate how well these models perform at correctly classifying tumors as benign or malignant by studying the dataset and using a variety of classification algorithms. This data can help medical personnel make educated decisions about patient care and lead to the creation of more precise diagnostic tools.

Data Collection:

I gathered the dataset from Kaggle

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

Total no of rows 569 and total no of columns is 32 Columns are as follows ID

Diagnosis

Radius\_mean

Texture\_mean

Perimeter\_mean

Area\_mean

Smoothness\_mean

Compactness\_mean

Concavity\_mean

Concave points\_mean

Symmetry\_mean

Fractal\_dimension\_mean

Radius\_se

Perimeter\_se

Area\_se

Smoothness\_se

Compactness\_se

Concavity\_se

Concave points\_se

Symmetry\_se

Fractal dimension\_se

Radius\_worst

Texture\_worst

Perimeter\_worst

Area\_worst

Smoothness\_worst

Compactness\_worst

Concavity\_worst

Concave points\_worst

Symmetry\_worst

Fractal dimension\_worst

Ten real value features are computed for each nucleus a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

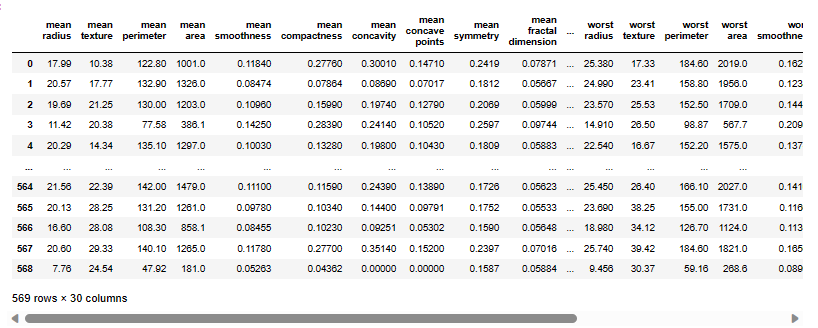
i) symmetry

j) fractal dimension ("coastline approximation" - 1)

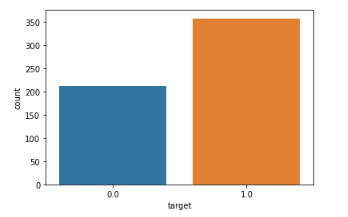
There are 569 breast cancer patients represented in this dataset. The digital photographs of the breast tissue for each patient have been used to extract 30 characteristics. The radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension of the image's cells are only a few examples of these attributes.

The diagnosis of the patient, which can either be "M" (for malignant) or "B" (for benign), is the target variable in this dataset. 212 patients had malignant cancer diagnoses in this sample, while 357 patients had benign diagnoses.

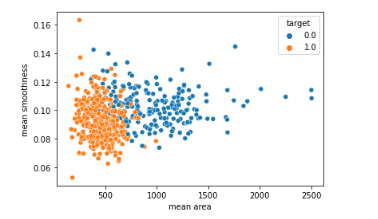
Data Set



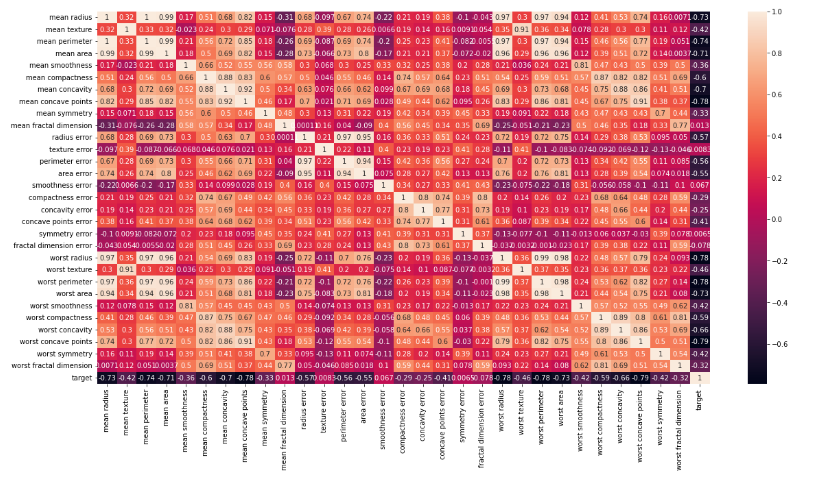
Category Distribution



scatter plot diagram for mean area and mean smoothness



Attribute Correlations:

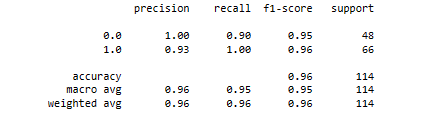


## **Machine Learning Algorithm Used:**

Logistic Regression:

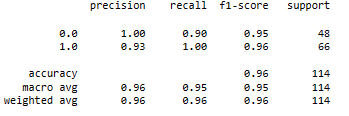
Logistic regression is one of the most widely used machine learning methods for detecting breast cancer. A binary classification approach called logistic regression models the likelihood of a binary response variable. By estimating the coefficients of a linear equation that connects the input features to the output variable, it operates.

The machine learning classification report provides insights about precision, recall and f1-score.



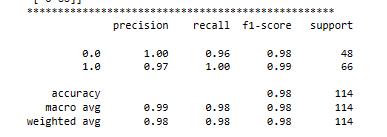
KNN:

This project uses KNN classifier, KNN is a simple, non-parametric algorithm for classification. It identifies the k nearest training data points based on a distance metric and assigns the majority class label to the new observation. It works well for large datasets and nonlinear decision boundaries and is often used as a baseline for complex models.



Gradient Boosting:

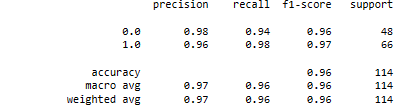
Another well-liked approach for classification issues is gradient boosting. It is an ensemble algorithm that creates a strong learner by fusing a number of weak learners. Iteratively fitting models to the residuals of earlier models is how Gradient Boosting operates.



Random Forest Classifier:

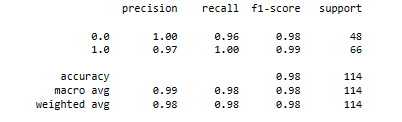
A well-liked machine learning method called Random Forest Classifier is utilized for both classification and regression problems. It is an ensemble learning technique that mixes various decision trees to provide predictions that are more accurate. In a random forest, each decision tree is built using a different subset of the training data, and the combined forecasts of all the individual trees are used to make the final prediction.

The Random Forest Classifier is renowned for its proficiency with high-dimensional datasets as well as its resistance to outliers and data noise. Additionally, it excels on unbalanced datasets like the Wisconsin Cancer dataset, where the proportion of benign samples is significantly larger than that of malignant samples.

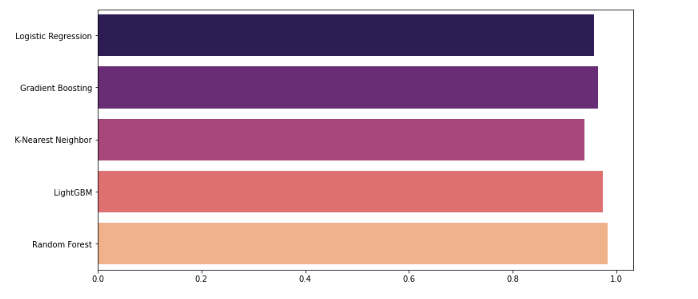


Light GBM:

A gradient boosting algorithm known as Light GBM uses a leaf-wise growth approach to shorten training time and improve model accuracy.



Here we are comparing the efficiencies of all the machine learning models and finding out which one has the highest efficiency.



### **Challenges:**

* Understanding the data
* Feature selection
* Modeling, finding the best model to improve the accuracy of the prediction.
* hyperparameter tuning

Conclusion:

Through this effort, we hope to shed light on the performance traits of several machine learning models for categorizing breast cancer. To identify the advantages and disadvantages of each model, we will evaluate each one's accuracy, precision, recall, and other evaluation criteria. We'll also contrast our findings with those of earlier research and look into how these models might be used in actual-world scenarios. By correctly identifying breast cancers as benign or malignant, we can help medical professionals make defensible choices about additional diagnostic procedures, therapeutic strategies, and patient care. The ultimate goal is to increase the precision and effectiveness of breast cancer diagnosis, which will improve patient outcomes and maybe save lives.