Project group members:

1. Saurabh Jadhav (A20491831)
2. Rutika Dhoka (A20501969)
3. Utkarsha Malegaonkar (A20493621)

Project Area:

Breast Cancer Prediction

1 Project proposal

1.1 Introduction

Breast cancer is considered one of the most common cancers in women caused by various clinical, lifestyle, social, and economic factors. can be predicted. The identification of a breast cancer tumor at an early stage is critical in the therapy procedure. We will analyze the data and come up with best model that can be used for predicting breast cancer.

1.2 Research Goal

1.2.1 Description

Predicting the breast cancer is a challenging problem and considering right set of key consideration points is very important to predict breast cancer in its early stage by analysing the collected data. Our objective is to create a predictive model which will predict with good accuracy (big ROC AUC value) the malignant and benign tumors correctly.

1.2.2 Specific Questions:

- What impact does the points such concave.points_worst, concavity_mean, concavity_worst, perimeter_worst, area_worst have on the analysis?
- What factors have a strong effect on the prediction?

1.2.3 Our prediction

Our goal is to analyze and predict the breast cancer correctly for the given dataset. We are going to compare different models such as Random Forest, Gradient Boosting Machine (GBM), Light Gradient Boosting Machine (lightGBM) and XGBoost through PCA analysis and then after comparing these models using cross validation decide the best model.

1.3 Methodology

Our methodology consists of three main parts:

1.3.1 Data Preparation:

- Dataset preparation for next steps.
- Dataset loading and split the datasets into test and training dataset.

1.3.2 Compare various data models:

- Predict the malignant and benign tumors based on the four models Random Forest, Gradient Boosting Machine (GBM), Light Gradient Boosting Machine (lightGBM) and XGBoost.
- Use cross validation to decide the best model.

1.3.3 Training and testing the model

- After deciding the model, we will train the model

Then test the model with test dataset

2 Project Outline:

2.1 Literature Review and related work:

In medical diagnosis, the prediction of a disease acts as an important core in analyzing the medical images. The unwanted cell growth in any part of the organ is known as tumor. The tumor may be benign or malignant. Malignant tumor is considered to be the most dangerous tissue. So, the early diagnosis of the disease helps to prevent the cancer. In women, breast cancer is treated as the most significant issue. There are various researchers studied about the prediction of breast cancer. Machine learning (ML) technics are an effective way to classify data especially in medical field, where those methods are widely used in diagnosis and decision making.

We are going to analyze various regression models and selecting the best performer. We also referred to other related work on the breast cancer prediction. Some of these are the ones listed below:

1. N. Fatima, L. Liu, S. Hong and H. Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis," in *IEEE Access*, vol. 8, pp. 150360-150376, 2020, doi: 10.1109/ACCESS.2020.3016715.

2. U. Pratap and S. Chhabra, "Breast Cancer Prediction using Different Machine Learning Algorithms," *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2021, pp. 451-454, doi: 10.1109/ICAC3N53548.2021.9725688.

3. M. S. Yarabarla, L. K. Ravi and A. Sivasangari, "Breast Cancer Prediction via Machine Learning," *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019, pp. 121-124, doi: 10.1109/ICOEI.2019.8862533.

2.2 Dataset:

The Breast Cancer Prediction Dataset is a large dataset of various patients in Wisconsin. It has 33 columns containing various attributes that affect the breast cancer prediction. It has been collected across various hospitals in Wisconsin.

Dataset Used: Breast Cancer Wisconsin (Diagnostic) Data Set

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

The details of the dataset are listed below :

#	Attribute	Description
Feature Name	Description	Data Type
id	ID number	numeric
diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)	string, dual value
radius_mean	mean of distances from center to points on the perimeter	numeric
texture_mean	standard deviation of gray-scale values	numeric
perimeter_mean	mean size of the core tumor	numeric
smoothness_mean	mean of local variation in radius lengths	numeric
compactness_mean	mean of perimeter^2 / area - 1.0	numeric
concavity_mean	mean of severity of concave portions of the contour	numeric
concave points_mean	mean for number of concave portions of the contour	numeric

2.3 Data Processing and Pipeline:

- Input: Loading the training and testing data into R using the read.csv function
- Data Cleaning: We will remove the address as a variable as we won't be using it in the predictive model. We will also remove the null values if any and replace them with relevant value.

- Transformation: We will split the training data set further into a training dataset and test dataset with a ratio of 80:20 respectively. These splits will be used to train the different algorithms and test their accuracy. The best model will then be used to predict the results on the test dataset provided.

2.4 Model Selection:

We have planned to use the following regression techniques:

1. We will use four models:
    a) RandomForest(RF)
    b) Gradient Boosting Machine(GBM)
    c) Light Gradient Boosting Machine(light GBM)
    d) XGBoost

2. Classification/regression approaches.
    a) Random ForestClassifier
    b) Logistic Regression Model
    c) Decision TreeClassifier
    d)Support Vector Machine
    e)K-Means Clustering

Programming Languages: R

Software Packages: RStudio

R Libraries: ggplot, geom_col, randomforest, mltools, glmnet, tidyverse, gbm

Project Management and Version Control: GitHub