Project- 4

Breast Cancer Prediction

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society.

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling.

Data Preparation

The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector

Attribute Information:

1. ID number 2) Diagnosis (M = malignant, B = benign) 3–32)

Ten real-valued features are computed for each cell nucleus:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter² / area — 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension ("coastline approximation" — 1)

Objective:-

This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achive this I have use two machine learning algorithms first one is Linear Regression method and the second one is Random Forest Method.

Phase-2 Data Exploration

We can observe that the data set contain 569 rows and 32 columns. '*Diagnosis*' is the column which we are going to predict, which says if the cancer is M = malignant or B = benign. 1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labelled as B (benign) and 212 as M (malignant)

As the first column of data is id of the patient which is not required, so we dropped it.

We have also converted the diagnosis data column into factor data type.

Missing or Null Data points

We checked for any missing value in dataset which is null there is no missing value in the data set, which is clear from the function **sum(is.na(x))** and the plot by the function **vis_mis(data).**

We also visualise the dataset based on Benign and Malignant.

From the correlation plot we analysed that the attribute radius_mean, texture_mean, perimeter_mean, area_mean, radius_se, perimeter_se, area_se, radius_worst, texture_worst, perimeter_worst, area_worst, are the important characteristics in deciding the cancer.

Splitting of Dataset

Splitting of dataset is done with the help of caret package into 80 % training and 20% test dataset.

Model Selection

It is already advised in the instruction to use Regression Analysis and Random Forest Method for the Model preparations.

Linear Regression Result

Linear regression model tells that the radius_mean, perimeter_mean and texture worst have the 99% of confidence interval in the cancer prediction.

This model gives Residual deviance of 81.812 and AIC of 105.81, which is pretty good.

Model performance evaluation gives the accuracy of 96.4% at the probability limit of 0.3 %.

Random Forest

Random forest does the iteration with 500 trees with 5 set of data with each tree. It gives the accuracy of 100% with 95% of confidence level

Important variable

varImpPlot function plot the important variable according to their priority, which gives area_worst top priority and texture_se least priority