1. BUSINESS UNDERSTANDING

Breast Cancer poses a major health challenge in the current society, it has taken away many lives which is becoming a risk to Human life span.The health specialists face challenges and may take much time in determining if the breast cancer detected on the patient is Malignant of Benign , thus the need for an automated and more effective means to detect this.Classification machine learning model will solve their challenges and reduce the time taken to give response to their patients.

Main Objective

To Train a classification machine learning model that classifies whether a breast mass is Malignant or Benign upon feeding certain data observation to the model..

Business Success Criteria

Reviewing the dataset to understand the variables we have , getting to know which is our target variable and which are the features.

Assessing the Situation

The datasets given are :
“https://drive.google.com/drive/folders/1vkQweADehrB0oq6FPVpi2OsP3gE3WJ6V”

Software used are : ***Github, Google Collaboratory,***

Assumption;

The data given is correct and up to date.

Constraints ;

There are no constraints identified

2.Data Understanding

Data Understanding Overview

For this research project, we are using the datasets provided by the company, which is:
“https://drive.google.com/drive/folders/1vkQweADehrB0oq6FPVpi2OsP3gE3WJ6V”

Data Description

In this project we have one data sets available to work on

Verifying Data Quality

Only one variable had missing values , that column was dropped from the data frame.

3.DATA PREPARATION

During our data preparation we followed the following steps :

1. **Loading of datasets csv** - this is our first step where we uploaded our datasets csv to the google collaboratory notebook and loaded the dataset and used th pandas python for extraction of data

2. **Data cleaning ;**
   After merging the dataset.csvs we performed a number of operations which are:
   - Checked for duplicates and missing values
   - Dealt with null values by dropping the column with missing values

   DATA Preprocessing for MACHINE LEARNING
   1. Encoding of our target variable
   2. Performing dimensionality reduction with LDA on features
   3. Training models with our data

CONCLUSION AND RECOMMENDATION

- ***The best performing model is SVM with 98% classification accuracy followed by NLP(natural language processing) and Logistic regression model with 96% classification accuracy followed by Random forest Classifier which has 95% Accuracy while KNN is the least performing with 94% Classification accuracy.***
- ***For this case, since we are working with a relatively small dataset SVM model is most prefered as it works well with data that is small and with very low noise in it, the presence of few outliers and missing values Makes SVM the best. I we***

were working with larger dataset with much noise in it Nlp or Logistic regression model could be the best option.