1. BUSINESS UNDERSTANDING

The Titanic ship sea accident caused the world much and is one of the worst marine accidents that was witnessed, its occurrence left non healing wounds for those who lost their loved ones and to the human race at large it is something that is hard to forget. Many researches and studies to date have been done to determine the survivors and those who met their demise. To uncover this mystery, we are tasked to build a model to classify whether one survived the accident or not. This model will help much in solving future occurrences of such cases.

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 3 datasets provided by the company, which are on this drive : “https://drive.google.com/drive/folders/1xjnMNkKIqYBy-uo8OWBbFvummseFxUYo”

Data Description

In this project we have three data sets available to work on, which are on the below link ":https://drive.google.com/drive/folders/1xjnMNkKIqYBy-uo8OWBbFvummseFxUYo"

Verifying Data Quality

There were a number of missing values in our data set, a number of outliers were present..

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
   - Drop unnecessary variables from our data frame
   - Used capping technique to deal with outliers

   DATA Preprocessing for MACHINE LEARNING
   1. Encoding of categorical variables
   2. Defined our target and features
   3. Trained our Models

CONCLUSION AND RECOMMENDATION

- KNN is the least performing model with accuracy classification of 97% which is slightly lower than those of other models that give accuracy of 100% correct classification.
- Nlp is good as it is faster to implement and is best for large data sets, Logistics regression model will also give the best result and ist more simple to implement.