ADDIS ABABA INSTITUTE OF TECHNOLOGY

DONE BY: NAHOM SENAY
ID: GSR/4848/17
TO: DR. FANTAHUN
DATE: DECEMBER 2/2024 G.C

Problem Definition

Breast Cancer is one of the latent and pervasive ailments caused on women. Hence, accurate prediction models that are capable of describing the problem are crucial. In this document it is tried to achieve a good success rate by using a binary classifier model. The model to be developed ideally tries to choose either an instance of a dataset is malignant or benign. The former means that an instance has breast cancer while the latter means that an instance fed for prediction is benign.

Goal

The goal is to achieve an accuracy of greater than 70%. If the accuracy of the model to be built falls in this range we can say that the model is reliable.

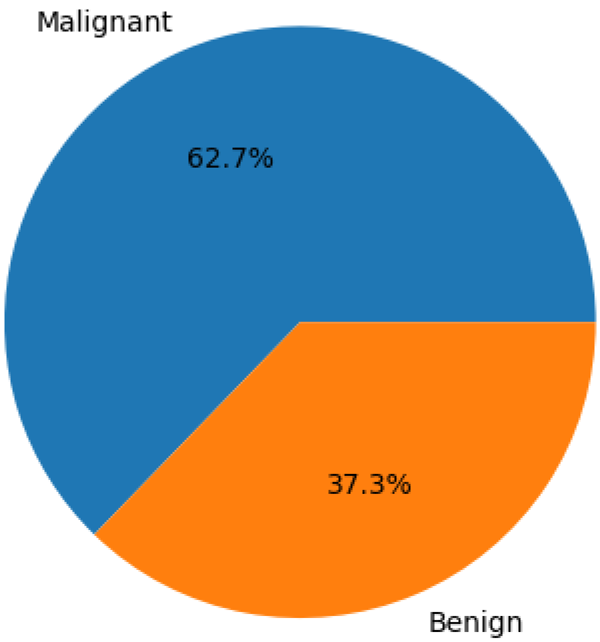
Data Collection

The dataset used in the model training process is found from kaggle. I have used the famous Wisconsin breast cancer dataset. The link for the dataset can be found [here]. It has 32 independent variables which can be tuned to be fed to the model and produce a binary classified output. The data set has 569 instances in the dataset.

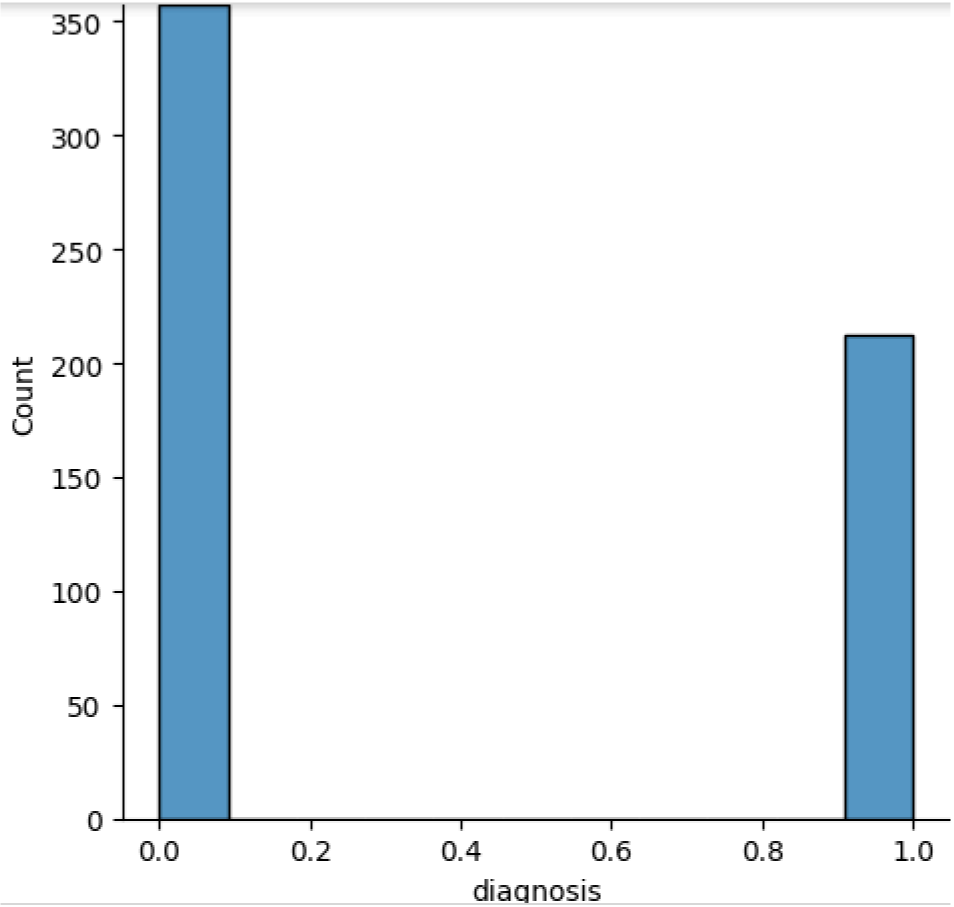
Data Exploration and Preparation

The dataset that was found has a problem of having not a Number data for a column. That column was dropped using the pandas library. This and the id column has been dropped so that they are irrelevant to the model. The former was removed since it can’t be used for fitting a model. And the latter was dropped because it can skew the model.

Other issues that needed tweaking was that the target variable called the diagnosis column wasn’t represented numerically. It had two values, the first being malignant represented as ‘M’ and benign represented as ‘B’. Hence, these were mapped as 1 and 0 respectively. The distribution of the dataset according to the target values can be represented in the below code base:



The above diagram represents the malignant-benign distribution among the available dataset for the model training to proceed. It shows that breast cancer positive data instances are greater than negative ones. Hence, the data set is imbalanced which implies that it should be altered to combat model bias.



The above graph shows that the target variable called diagnosis has around 350 entries of malignant and 200 entries of benign.

Algorithm Selection

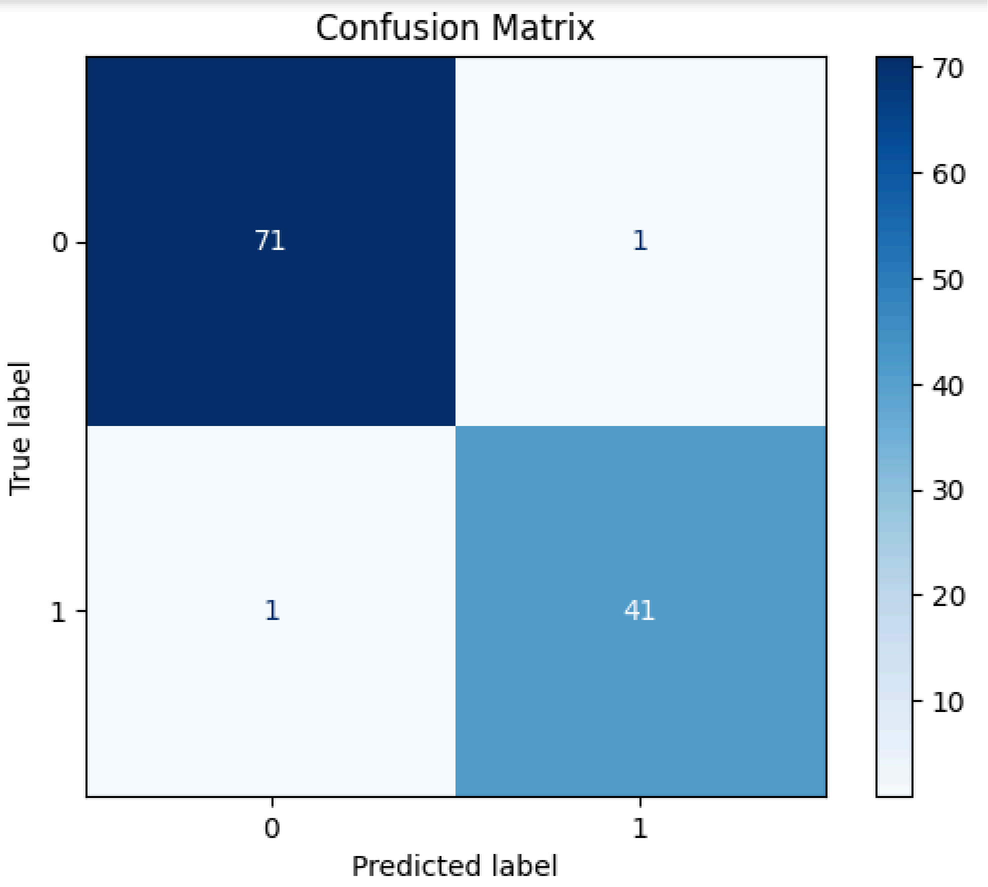
The data set used for the model development has both target and input variables. Hence, the model used in this case is a supervised learning model. Since the dataset is relatively small, logistic regression has been chosen from available alternatives.

Model development and Training

The logistic regression model was coupled with principal component analysis for feature extraction. This was to prevent the curse of dimensionality. As the number of columns increases it would be harder to generalize on a given dataset without overfitting. Furthermore, standard scaler was used to prevent bias to a certain output because some values were greater than others from the available columns.

Model Evaluation and Hyperparameter Tuning

The model had scored 98.2% on the test data with the following confusion matrix that has resulted:



This indicates that from the test data set it had scored 71 True positive, 1 False Positive, 41 True Negative and 1 False negative.

Model Testing and Deployment
The model was saved and deployed using streamlit and can be accessed at this link.