PSDL Mini Project

1.Group Member Roll Numbers, C Numbers and Name:-

Group Members	Roll nos.	C Numbers
Disha Ingole	2352	C22020221352
Mrunal Jambenal	2359	C22020221359
Mukta Joshi	2360	C22020221360
Tanaya Khaire	2372	C22020221372

2. Problem Statement

To create a python Program to detect breast cancer using Machine learning

3. Keywords

Machine Learning using Python, test, train, predict, Logistic Regression, Breast Cancer dataset

4. Abstract

Attribute Information:

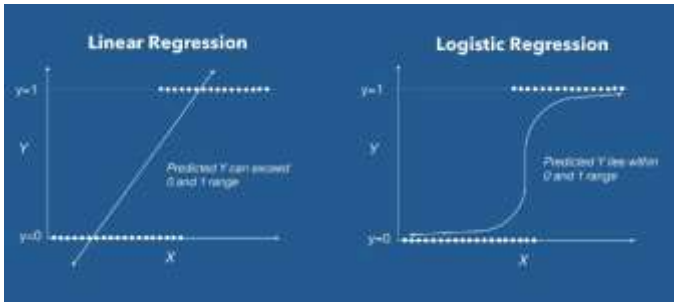
1) ID number

2) Diagnosis (M = malignant(cancerous growth), B = benign(non harmful tumour)

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

- a) radius (mean of distances from center to points on the perimeter)
- b) texture (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) smoothness (local variation in radius lengths)
- f) compactness (perimeter^2 / area - 1.0)
- g). concavity (severity of concave portions of the contour)
- h). concave points (number of concave portions of the contour)
- i). symmetry
- j). fractal dimension ("coastline approximation" - 1)

Logistic Regression:



The Logistic Regression formula aims to limit or constrain the Linear and/or Sigmoid output between a value of 0 and 1. The main reason is for interpretability purposes, i.e., we can read the value as a simple Probability; Meaning that if the value is greater than 0.5 class one would be predicted, otherwise, class 0 is predicted.

In Machine Learning we create models to predict the outcome of certain events. To measure if the model is good enough, we can use a method called Train/Test.

Train/Test

Train/Test is a method to measure the accuracy of your model.

It is called Train/Test because you split the the data set into two sets: a training set and a testing set.

For eg. - 80% for training, and 20% for testing.

You *train* the model using the training set.

Train the model means create the model.

You *test* the model using the testing set.

Test the model means test the accuracy of the model.

From Histogram:

1) mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in classification of the cancer. Larger values of these parameters tends to show a correlation with malignant tumors.
2) Mean values of texture, smoothness, symmetry or fractual dimension does not show a particular preference of one diagnosis over the other. In any of the histograms there are no noticeable large outliers that warrants further cleanup.

5. Module-wise description

1) Matplotlib

Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays.

2) NumPY

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

3) sklearn

Simple and efficient tools for data mining and data analysis. It features various classification, regression and clustering algorithms etc. Accessible to everybody and reusable in various contexts. Built on the top of NumPy, SciPy, and matplotlib.

- sklearn.model_selection.train_test_split(*test_size=None*, *train_size=None*)
  Split arrays or matrices into random train and test subsets.
- sklearn.metrics.accuracy_score(y_true, y_pred, sample_weight=None)
  Accuracy classification score.
- sklearn.linear_model.LogisticRegression()
  This class implements regularized logistic regression

6. Technology Selected and Technology features covered

What is machine learning?

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values**.**

Supervised Learning**:**

Supervised learning is a machine learning approach that's defined by its use of labelled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labelled inputs and outputs, the model can measure its accuracy and learn over time. Supervised learning can be separated into two types of problems when data mining: classification and regression Classification problems use an algorithm to accurately assign test data into specific categories, such as separating apples from oranges. Regression is another type of supervised learning method that uses an algorithm to understand the relationship between dependent and independent variables. Some popular regression algorithms are linear regression, logistic regression and polynomial regression.

7. References

Dataset : Breast Cancer Wisconsin (diagnostic) dataset

https://scikit-learn.org/

https://youtu.be/6P3HSOcCYPc

https://colab.research.google.com/drive/1xb03zuJQOPI69CpF-CMtF6_47JthvKtj