Project Report

**Introduction:**

* *Overview*

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 correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling.

Classification methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

* *Existing Problem*

The dataset used in this story is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA.

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)

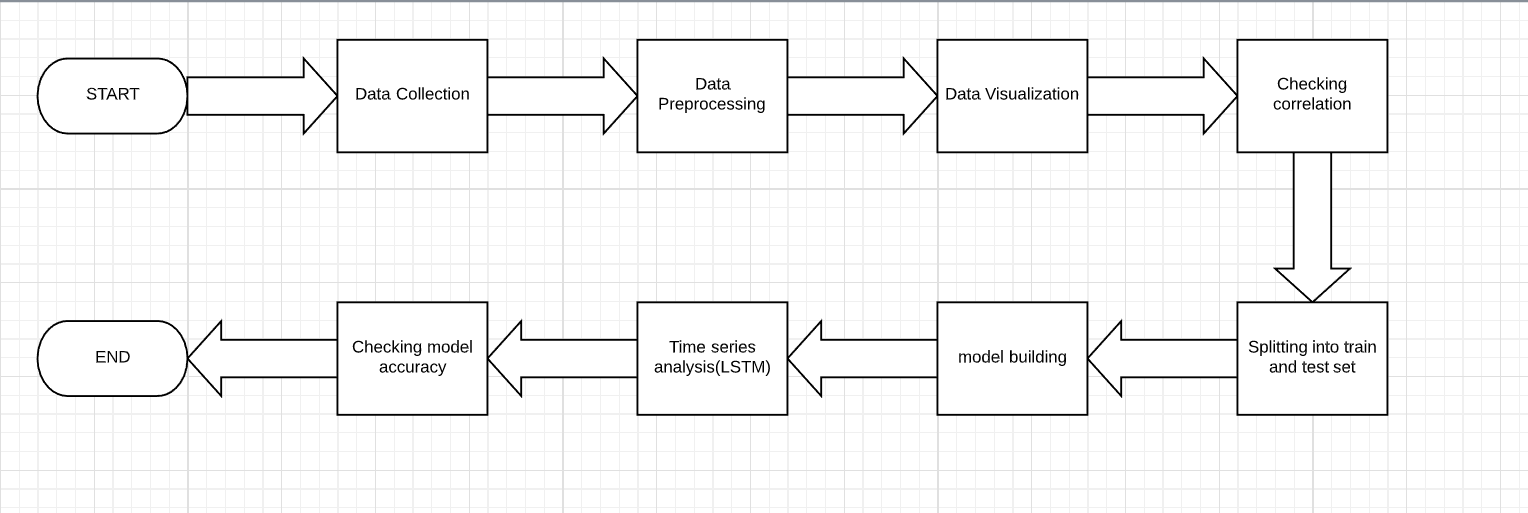
The mean, standard error and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

* *Objectives*

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 achieve this i have used machine learning classification methods to fit a function that can predict the discrete class of new input.

**Theoritical Analysis:**

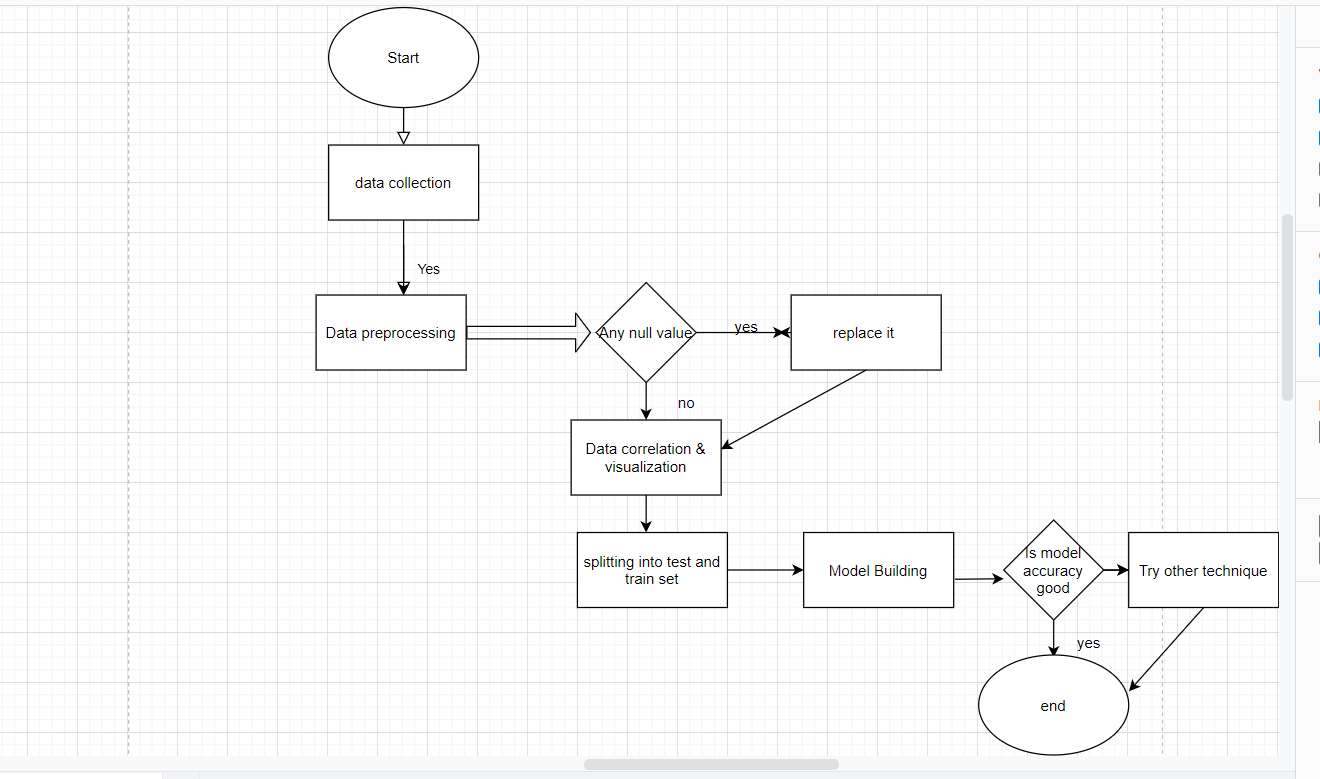
3.1 Block Diagram:



* *Flowchart:*

A flowchart is a type of [diagram](https://en.wikipedia.org/wiki/Diagram) that represents a [workflow](https://en.wikipedia.org/wiki/Workflow) or [process](https://en.wikipedia.org/wiki/Process). A flowchart can also be defined as a diagrammatic representation of an [algorithm](https://en.wikipedia.org/wiki/Algorithm), a step-by-step approach to solving a task.

The flowchart shows the steps as boxes of various kinds, and their order by connecting the boxes with arrows. This diagrammatic representation illustrates a solution model to a given [problem](https://en.wikipedia.org/wiki/Problem_solving). Flowcharts are used in analysing, designing, documenting or managing a process or program in various fields.



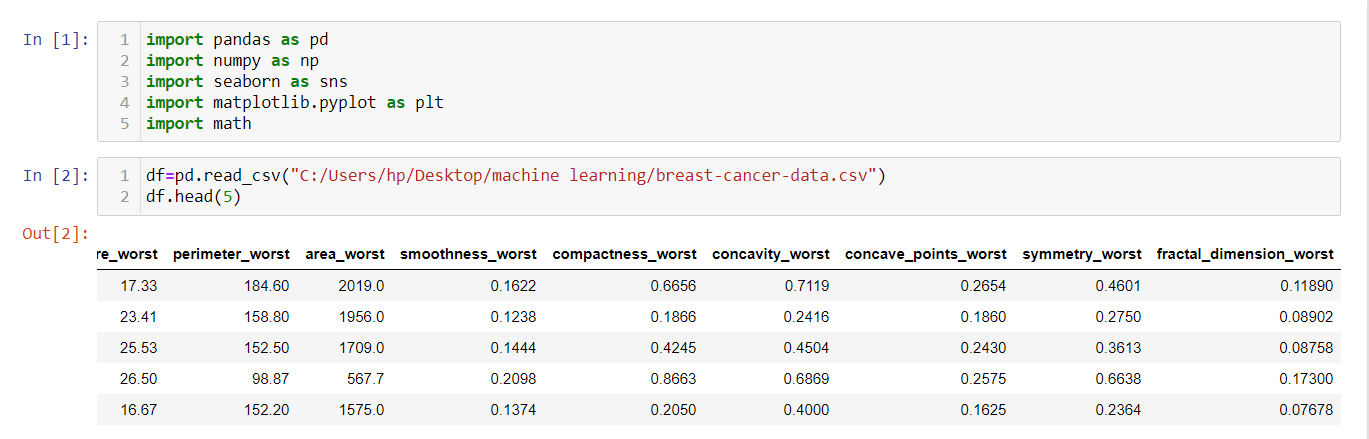
**Problem Solving Approach:**

1. Data Collection:

I collect the breast cancer prediction dataset from Kaggle.

2.Data Pre-processing:

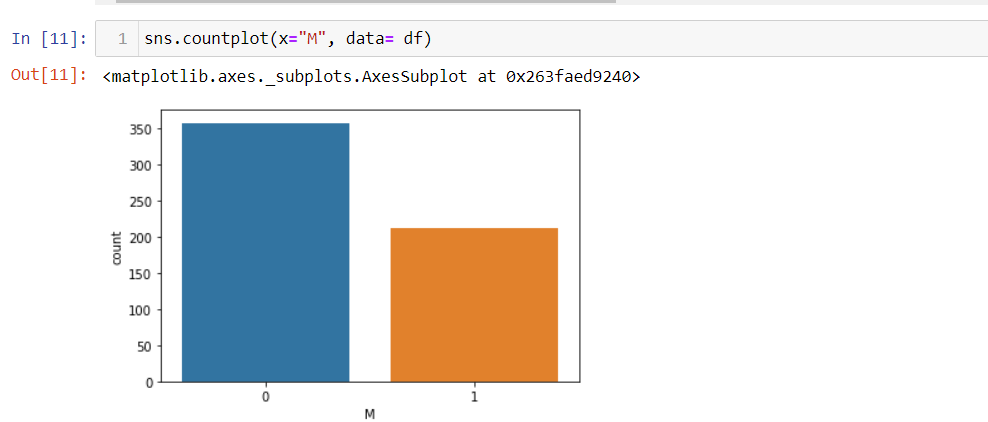
We can find the dimensions of the data set using the panda dataset ‘shape’ attribute. We can observe that the data set contain 569 rows and 32 columns. ‘Diagnosis’ is the column which we are going to predict , which says if the cancer is M = malignant or B = benign. 1 means the cancer is malignant and 0 means benign.



3.Data Visualization:

Visualization of data is an imperative aspect of data science. It helps to understand data and also to explain the data to another person. Python has several interesting visualization libraries such as Matplotlib, Seaborn etc.

Here , we see the count of malignant cell and benign cell.



4.Null Point Handling:

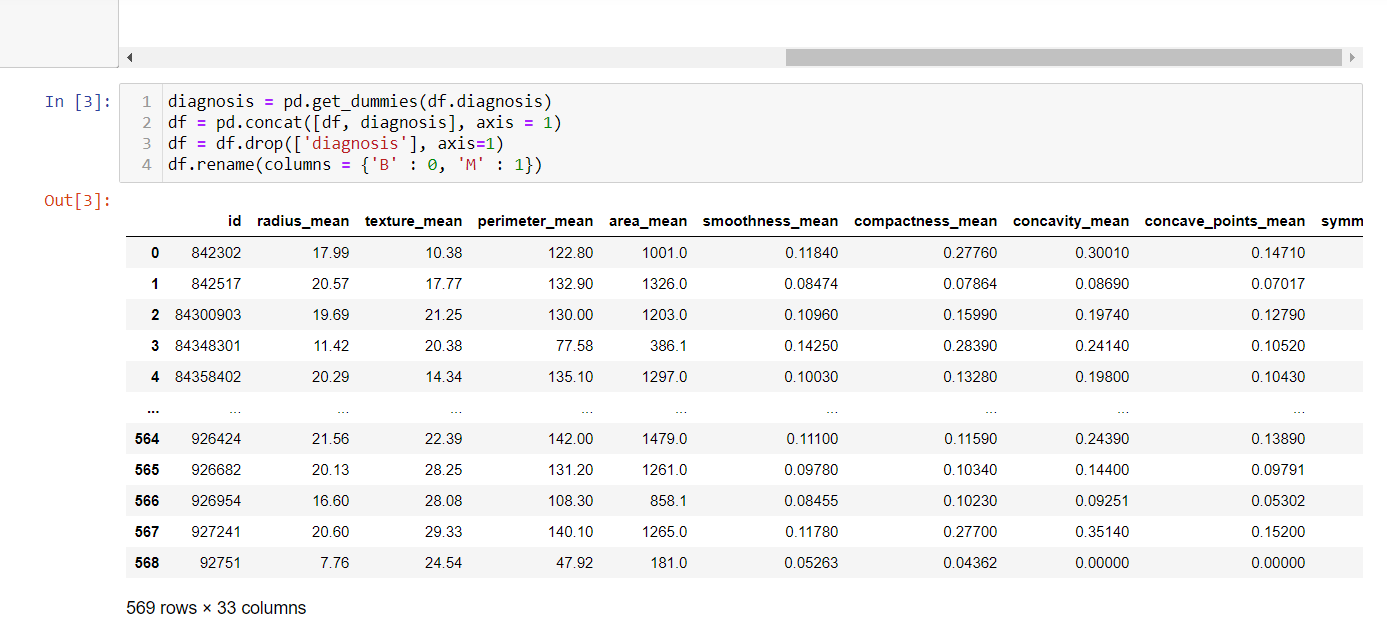
We can find any missing or null data points of the data set (if there is any) using the following pandas function. In this dataset there is no null value in any attribute.



5.One Hot Encoding:

Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set.

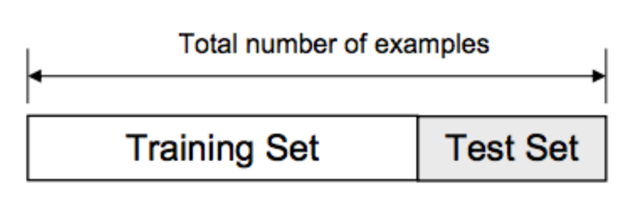
We will use One Hot Encoder to label the categorical data. Label Encoder is the part of Scikit Learn library in Python and used to convert categorical data into numbers, for which our predictive models can better understand.

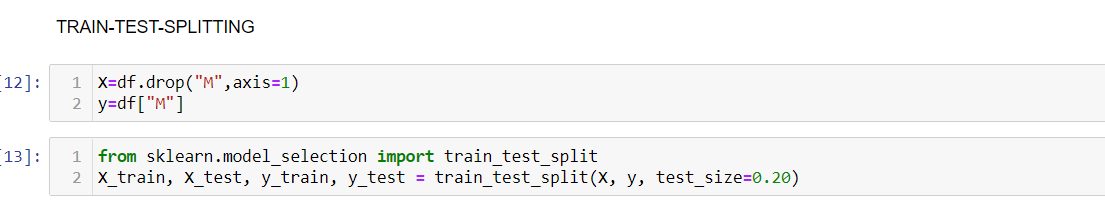


6.Spliting Dataset:

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model’s prediction on this subset.

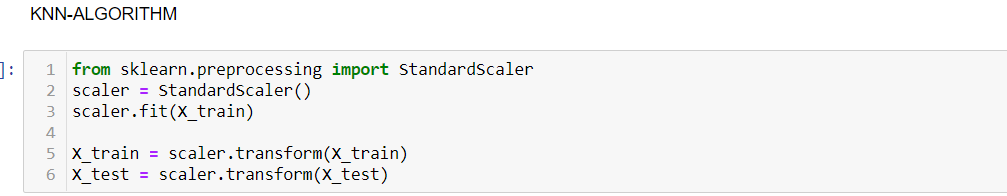
I will do this using Scikit-Learn library in Python using the train\_test\_split method.





7.Feature Scaling:

I use StandardScaler method from Scikit-Learn library.

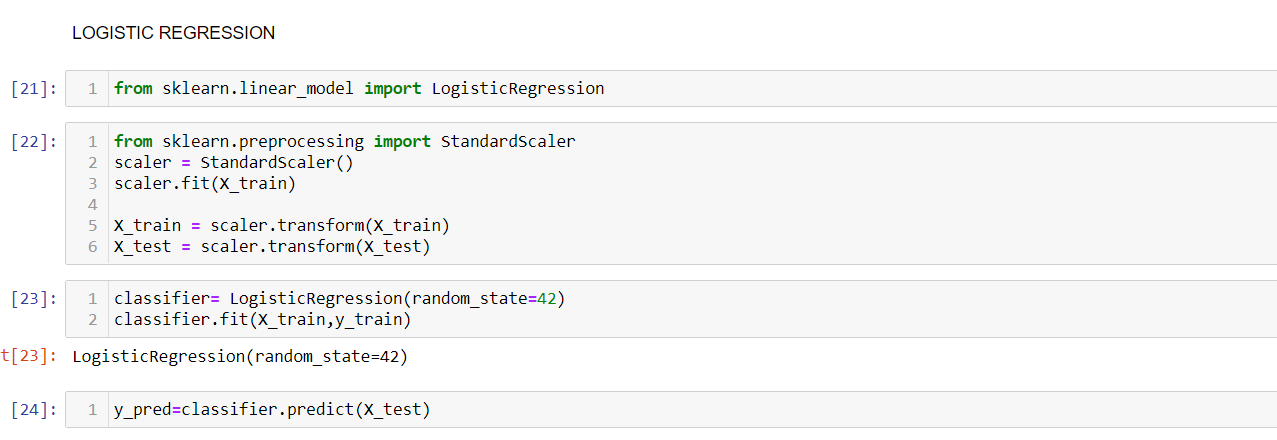


8.Model Building:

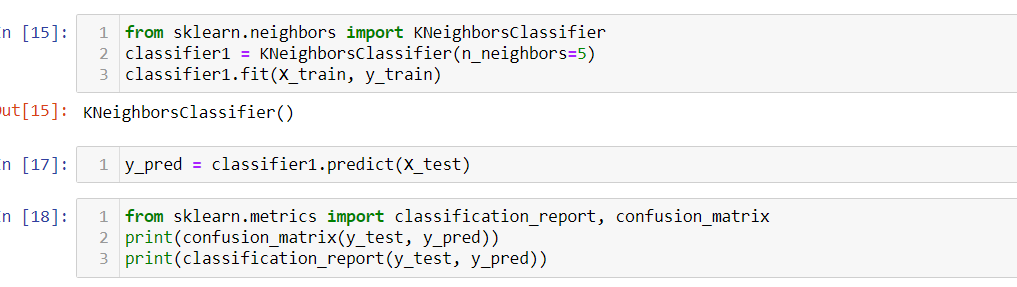
This is the most exciting phase in Applying Machine Learning to any Dataset. It is also known as Algorithm selection for Predicting the best results.Usually Data Scientists use different kinds of Machine Learning algorithms to the large data sets.

We have different types of classification algorithms in Machine Learning :-

1. Logistic Regression



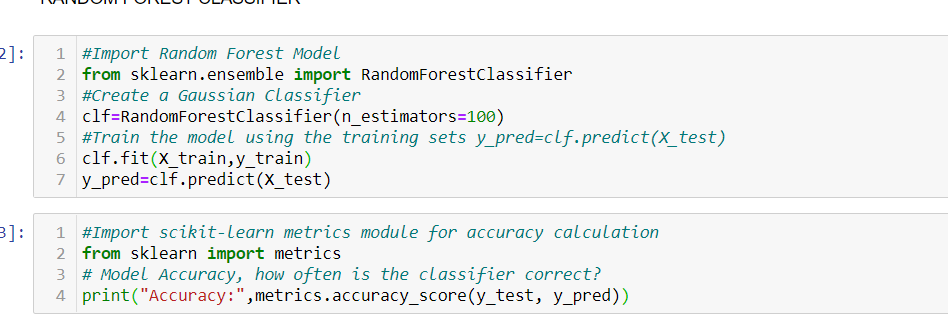
2.K- Nearest Neighbor



3. Decision Tree Algorithm



4. Random Forest Classification



Here I use sklearn library to import all the methods of classification algorithms.

**Results:**

After applying the different classification models, we have got below accuracies with different models:

1. Logistic Regression — 0.96

2. K- Nearest Neighbor — 0.958

3. Decision Tree Algorithm — 0.95

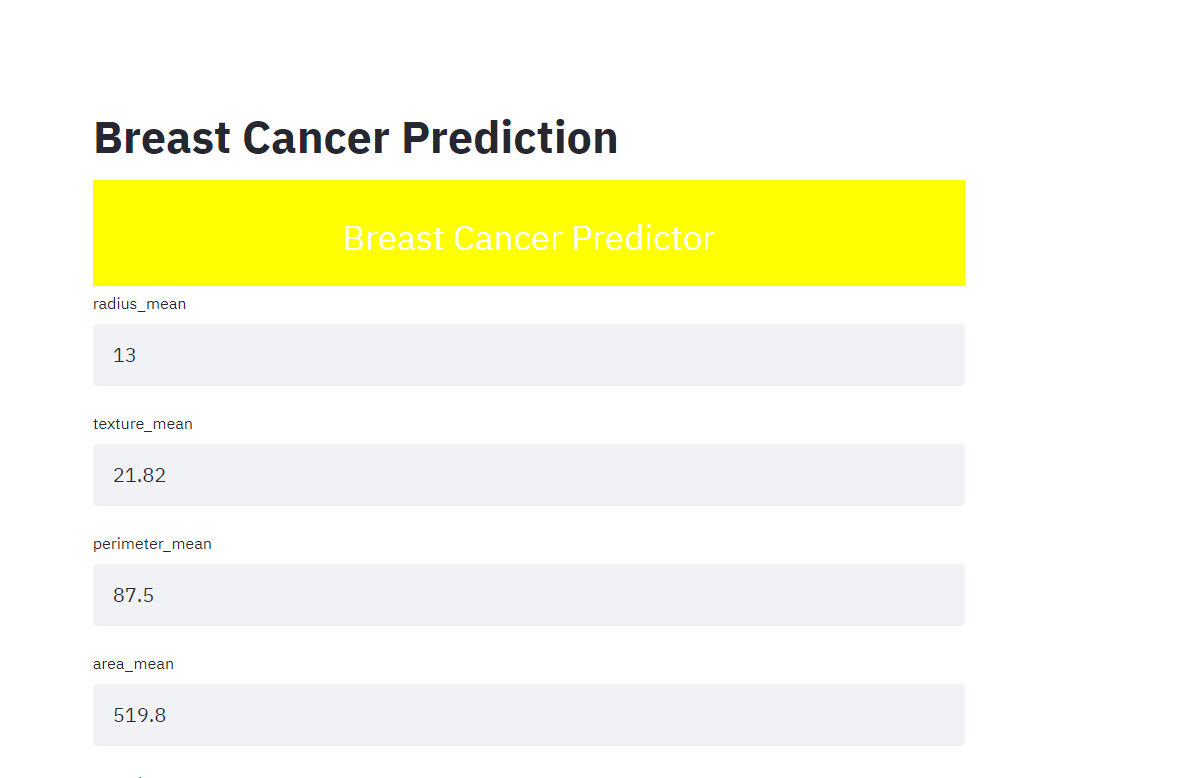
4. Random Forest Classification — 0.94

**Conclusion:**

So finally I have built the classification model and I can see that Logistic Regression Classification algorithm gives the best results for our dataset. Well its not always applicable to every dataset. To choose the model we always need to analyse our dataset and then apply our machine learning model.

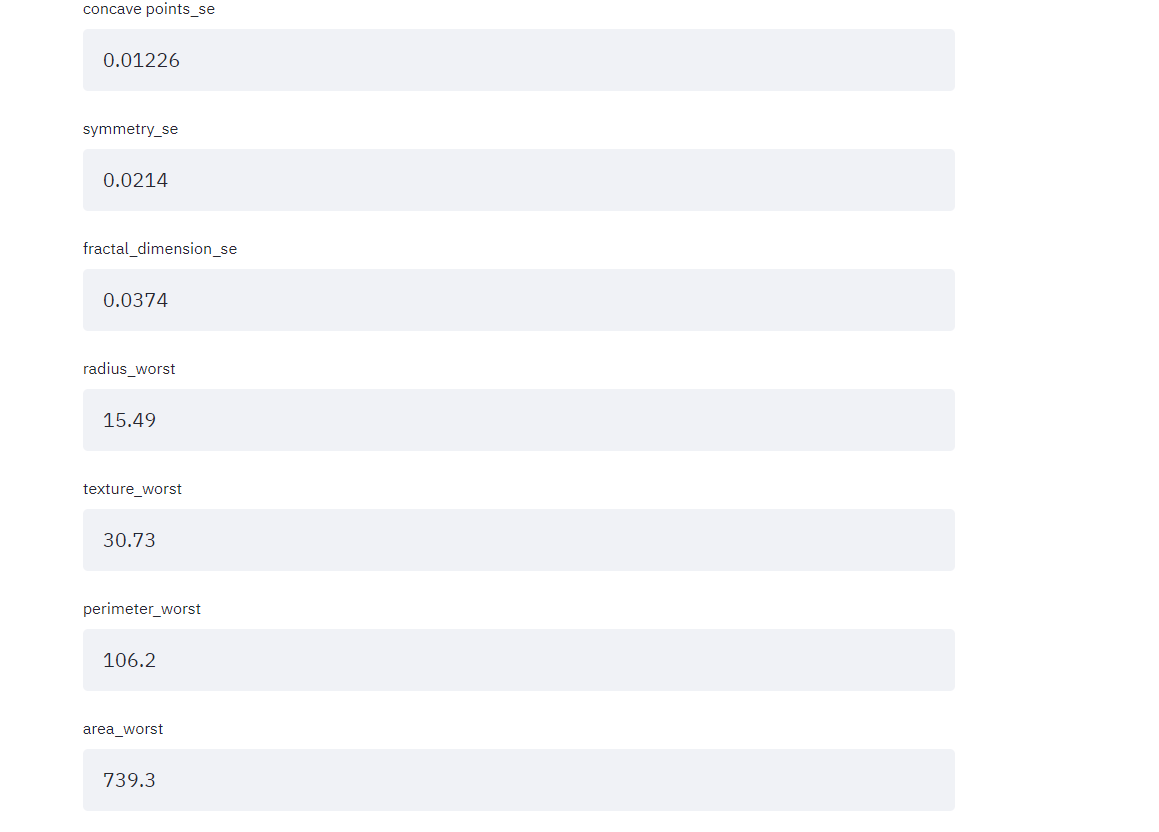
**Using Streamlit:**

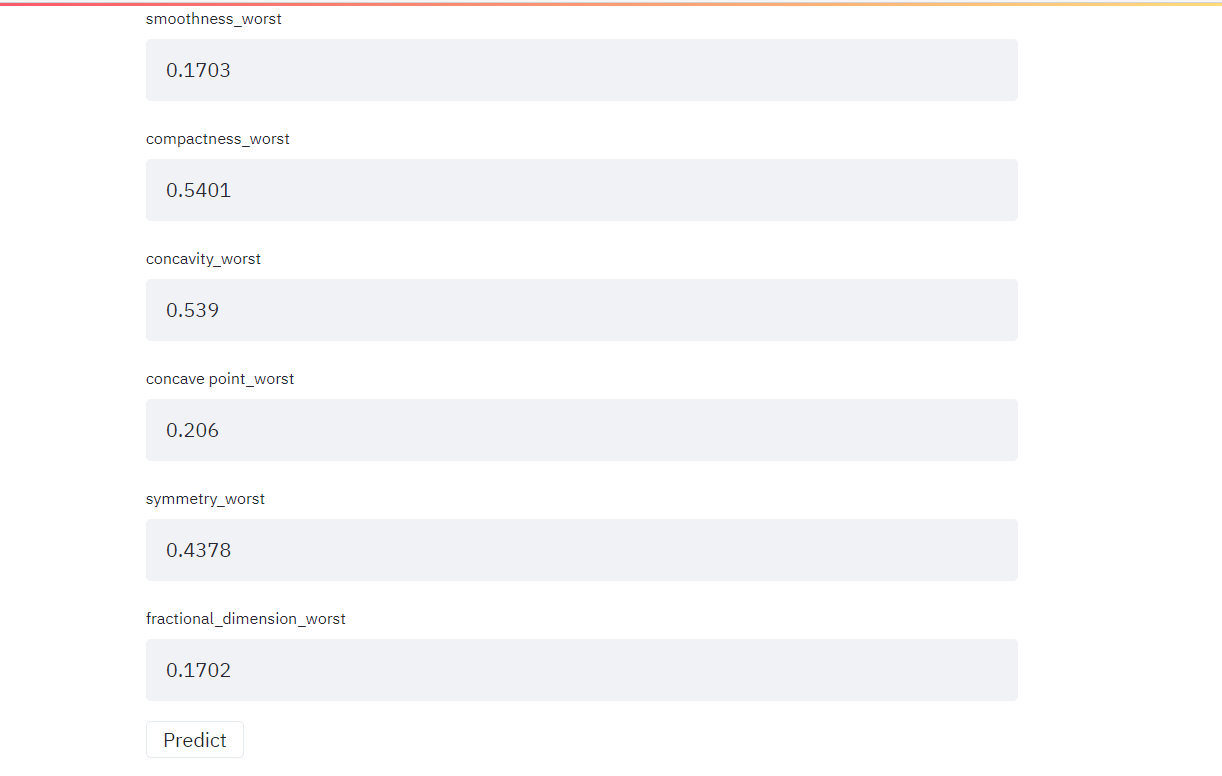
I use streamlit library to make this model user friendly. The streamlit app helps users to check if they have any chance of breast cancer or not.The screen shots are given below:











And finally the output is:

