MACHINE LEARNING PROJECT

TOPIC: BREAST CANCER PREDICTION

ROADMAP

UNDERSTANDING THE PROBLEM STATEMENT

- To Predict whether the given patient has Benign or Malignant Breast Cancer.
- What algorithms can be used to solve this objective ?

Machine Learning Models Used

- Logistic Regression
- Naïve Bayes
- Decision Tree
- Random Forest
- Support Vector Machine

Data Description

- 13 Columns
- **Independent Variables:-** 12,
- **Dependent Variable:-**1
- **Dependent Variable Type:** Binary and categorical

ATTRIBUTES

Independent Variables were:

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

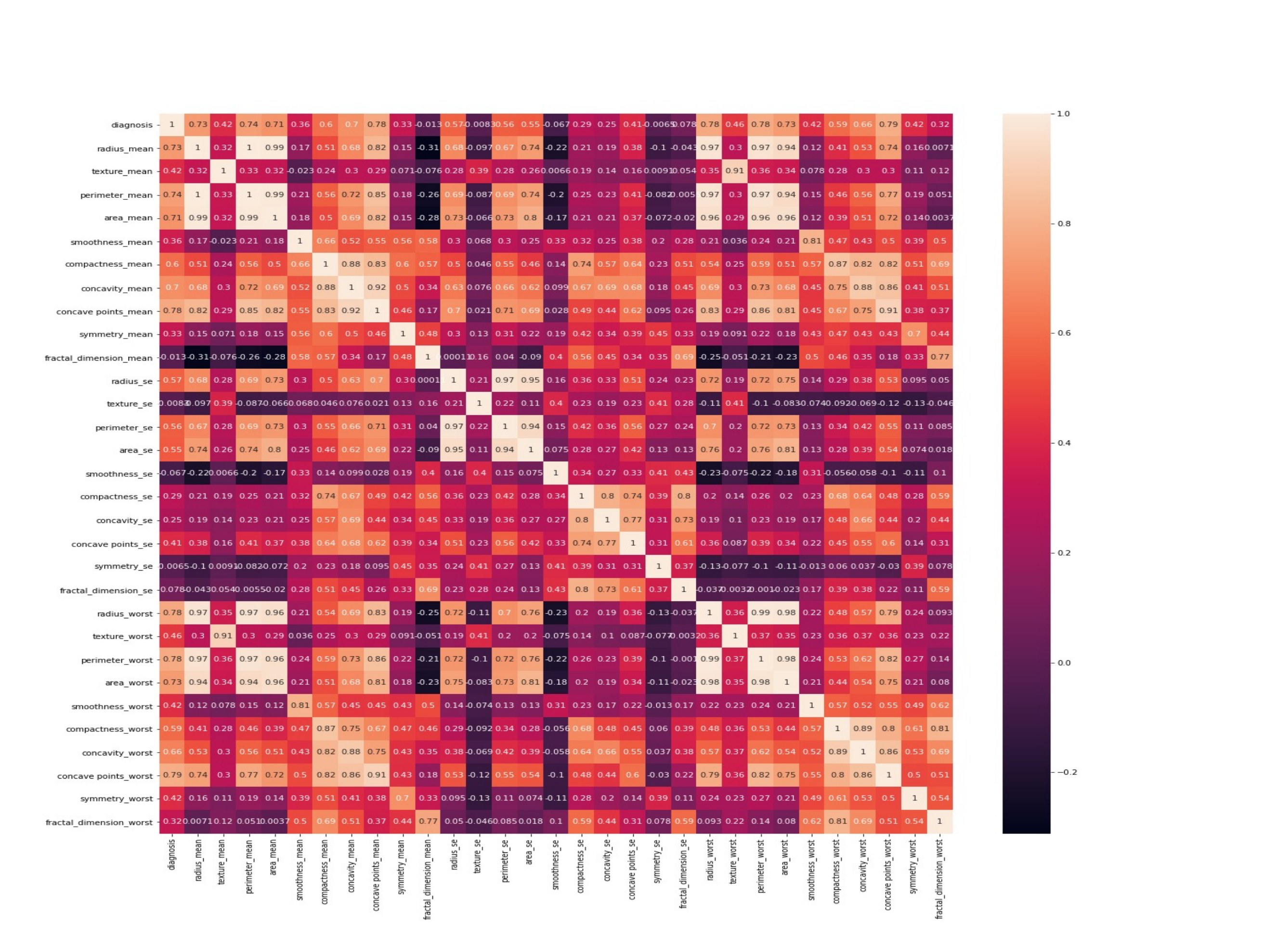
Dependent Variable was:

> Diagnosis

CORRELATION MATRIX

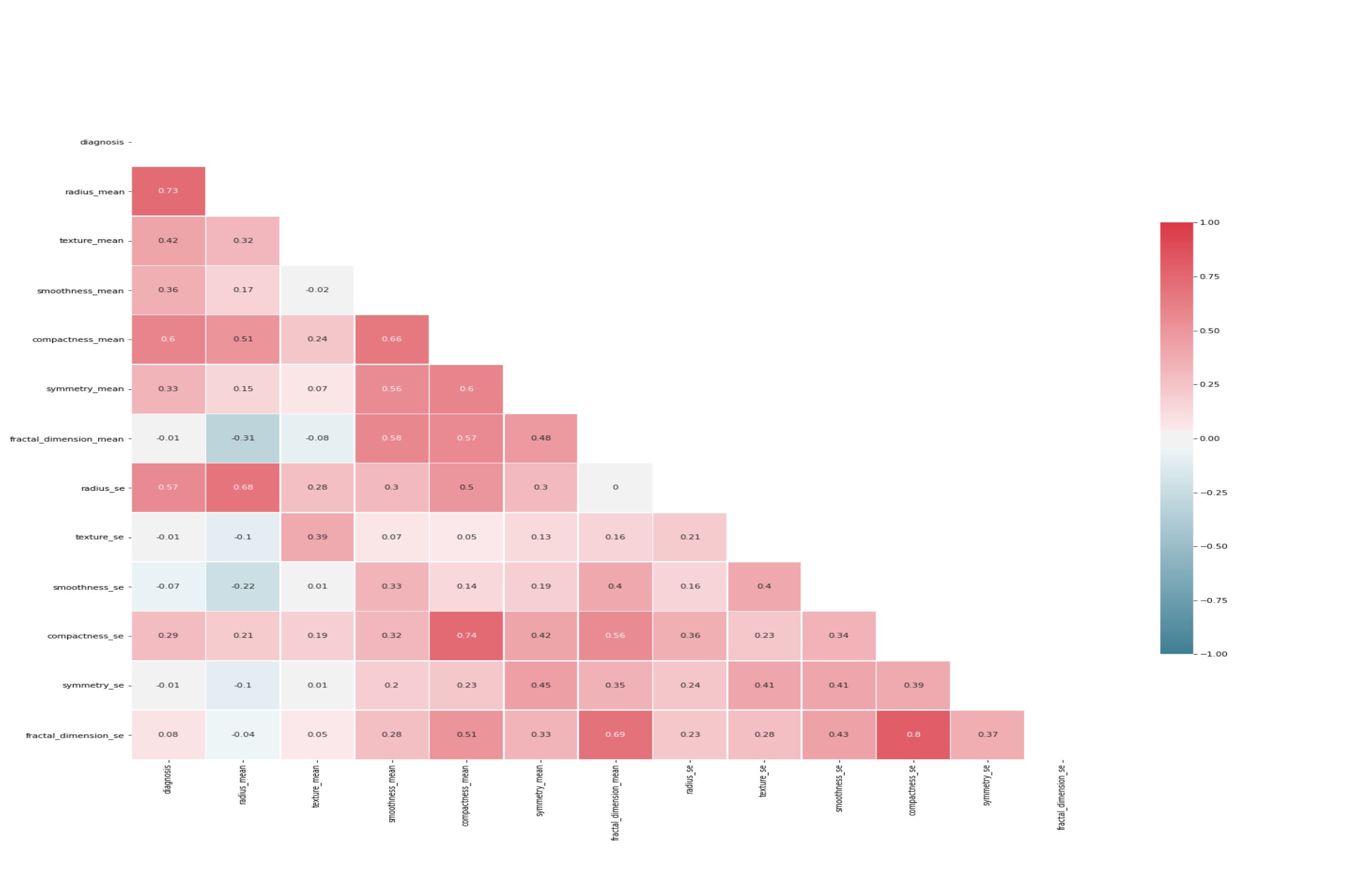
Number Of Variables in the data set : 31

As you can see there is high multicollinearity between variable ,hence removing variable having greater than .96 correlation.



FEATURE SELECTION

- Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.
- The variables which we have created for feature selection are :
- ' diagnosis'
- 'radius_mean',
- 'texture_mean'
- 'smoothness_mean'
- 'compactness_mean'
- 'symmetry_mean'
- 'fractal_dimension_mean'
- 'radius_se', 'texture_se'
- 'smoothness_se' 'compactness_se'
- 'symmetry_se'
- 'fractal_dimension_se'



LOGISTIC REGRESSION

- Logistic regression is a statistical method for predicting binary/multiple classes. The outcome or target variable is binary in nature.
- Logistic Regression Equation :-

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

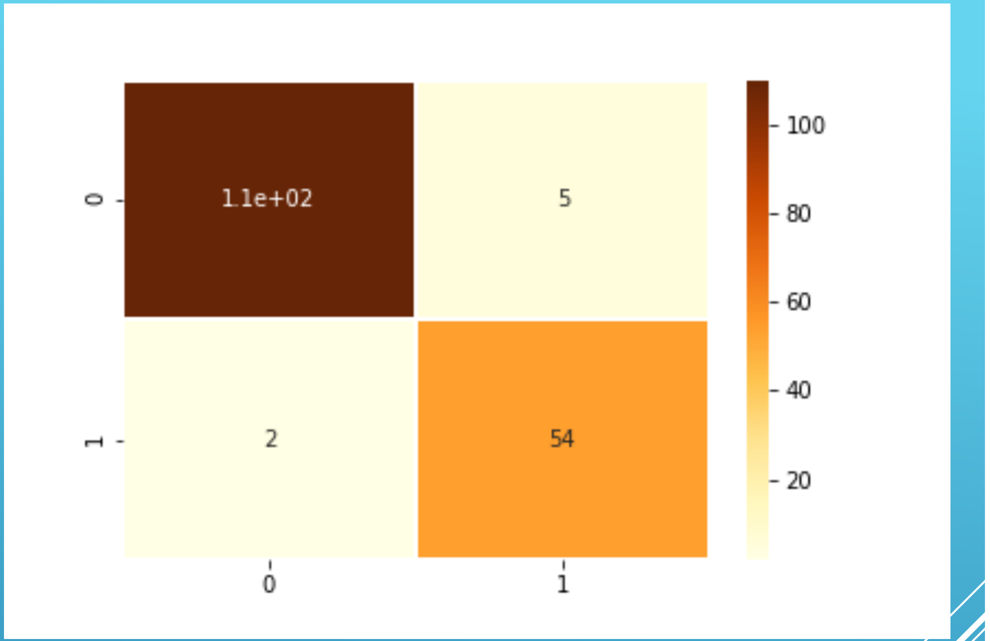
- $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

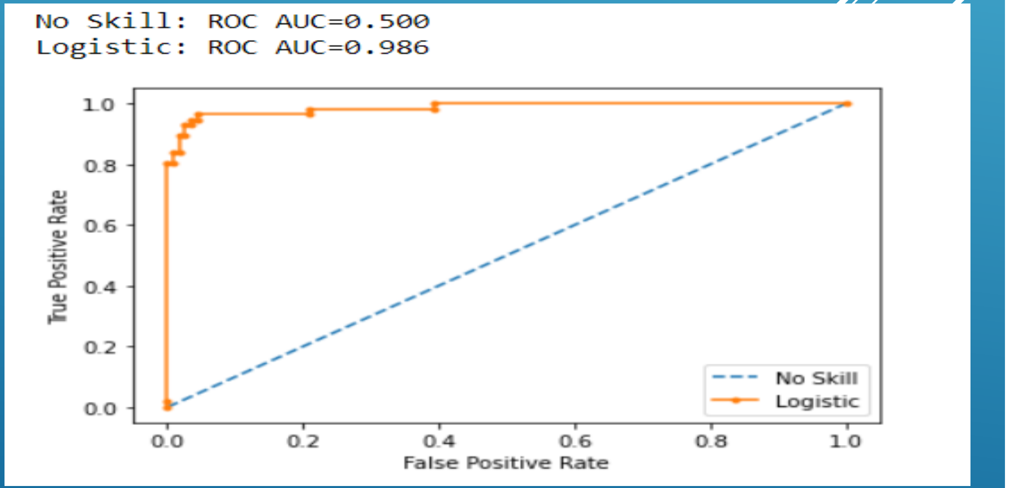
- TO CHECK THE PERFORMANCE (ROCR CURVE):

The threshold probability was found to be 0.986.

Sensitivity of model was about 95.6% i.e. the True positive rate.



- Accuracy: 0.9590643274853801

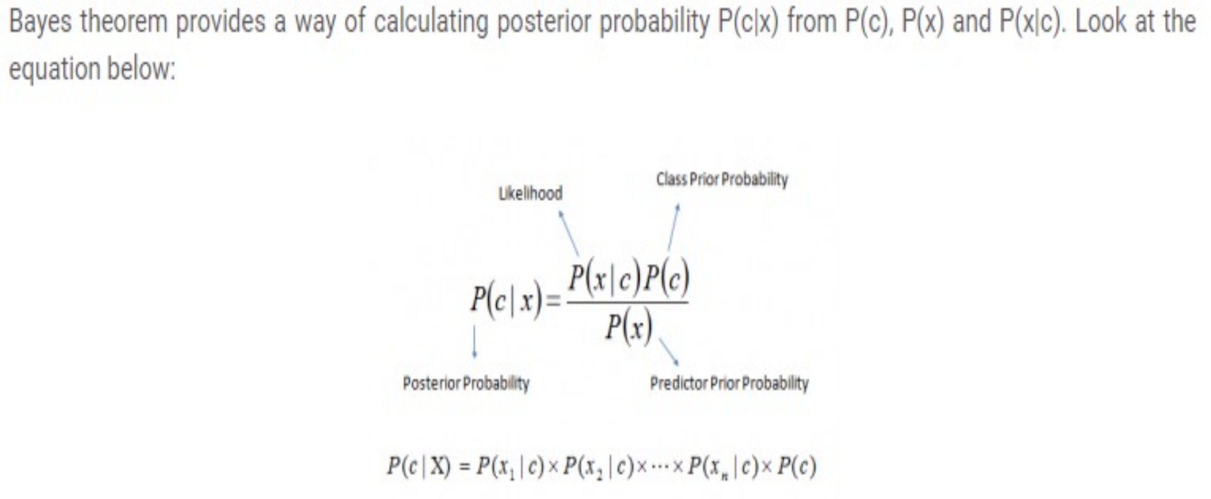


```
No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.986
```

NAIVE BAYES

- A classifier is a machine learning model that is used to discriminate different objects based on certain features.
- Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature

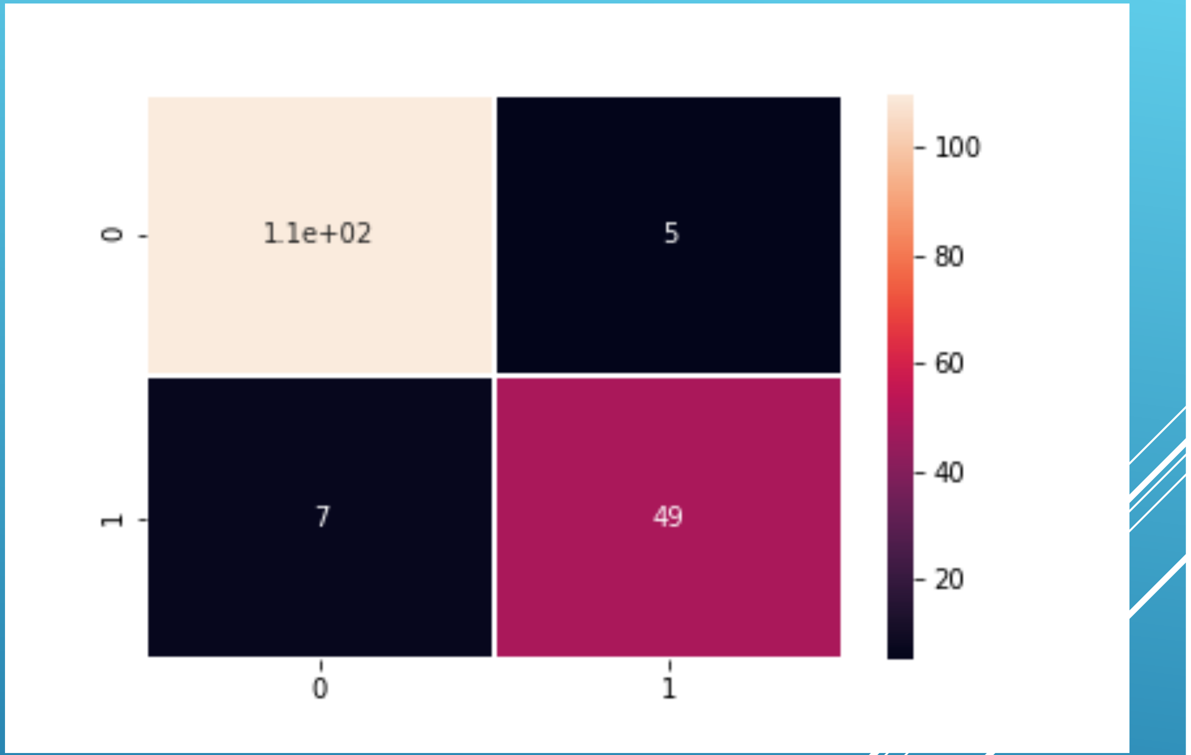
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$



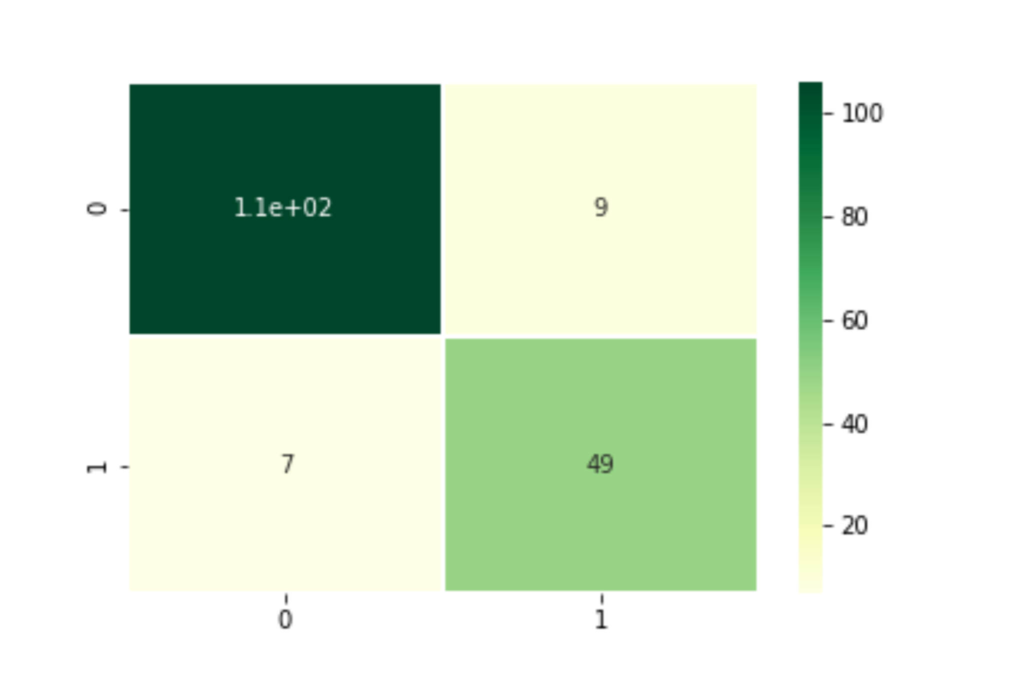
Accuracy -> 0.9298245614035088

DECISION TREE

Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter.

- Decision Tree consists of :
- Nodes : Test for the value of a certain attribute.
- Edges/ Branch : Correspond to the outcome of a test and connect to the next node or leaf.
- Leaf nodes : Terminal nodes that predict the outcome (represent class labels or class distribution).

In the decision tree, the root and internal nodes contain attribute test conditions to separate records that have different characteristics. All the terminal node is assigned a class label Yes or No

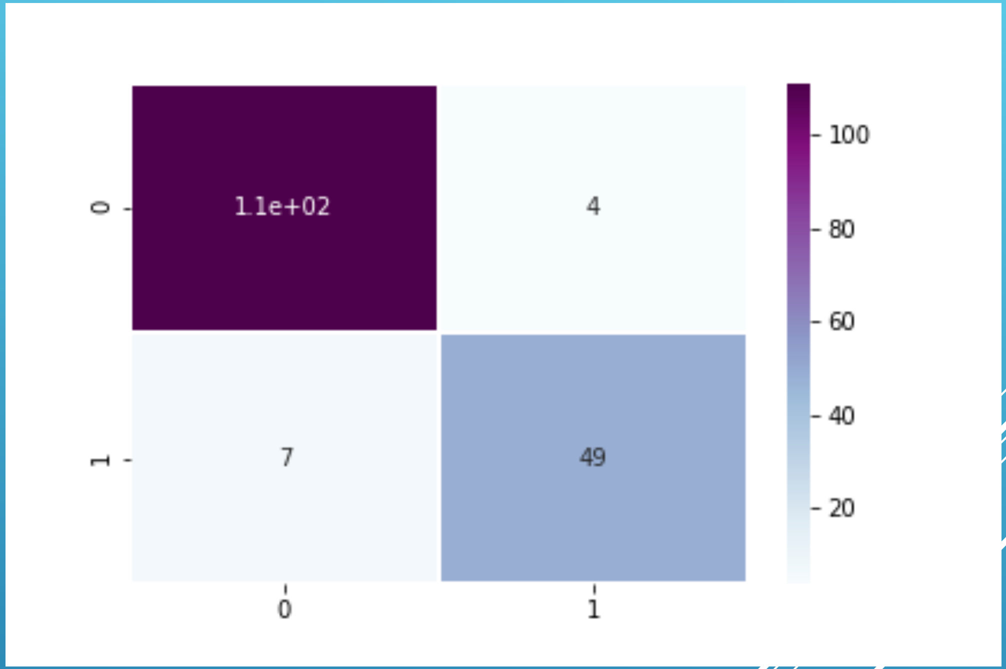


- Accuracy:0.9064327485380117

RANDOM FOREST

- **Random forests** or **random decision forests** are **an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.**

- **Random forests correct for decision trees' habit of overfitting to their training set.**

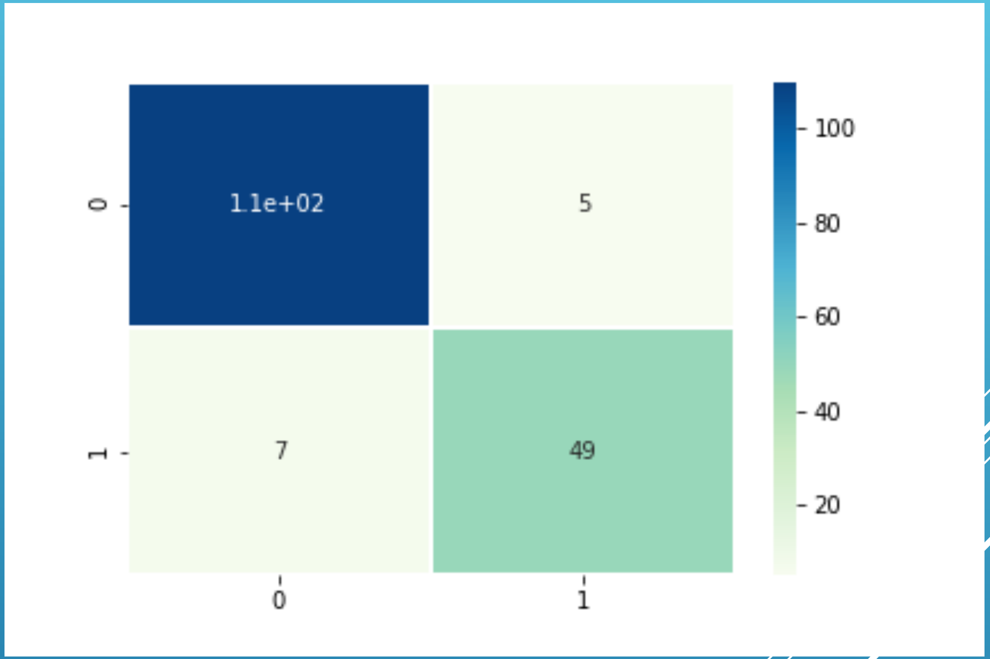


Accuracy:0.935672514619883

SUPPORT VECTOR MACHINES (SVM)

- Supervised learning algorithms try to predict a target (dependent variable) using features (independent variables).
- Depending on the characteristics of target variable, it can be a classification (discrete target variable) or a regression (continuous target variable) task. Prediction is done with a mapping function which maps independent variables to dependent variable.

- The Kernel we used in our model was 'Linear'.



Accuracy:0.9298245614035088

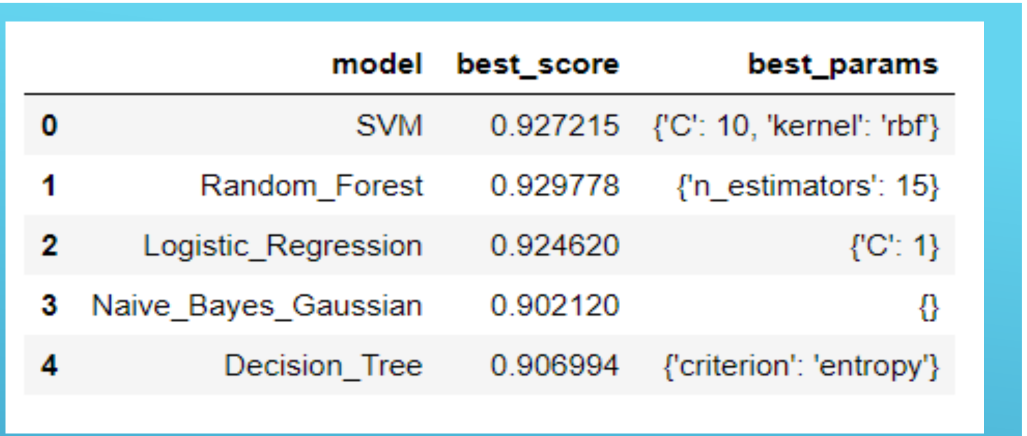
GRID SEARCH CV

- Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model.

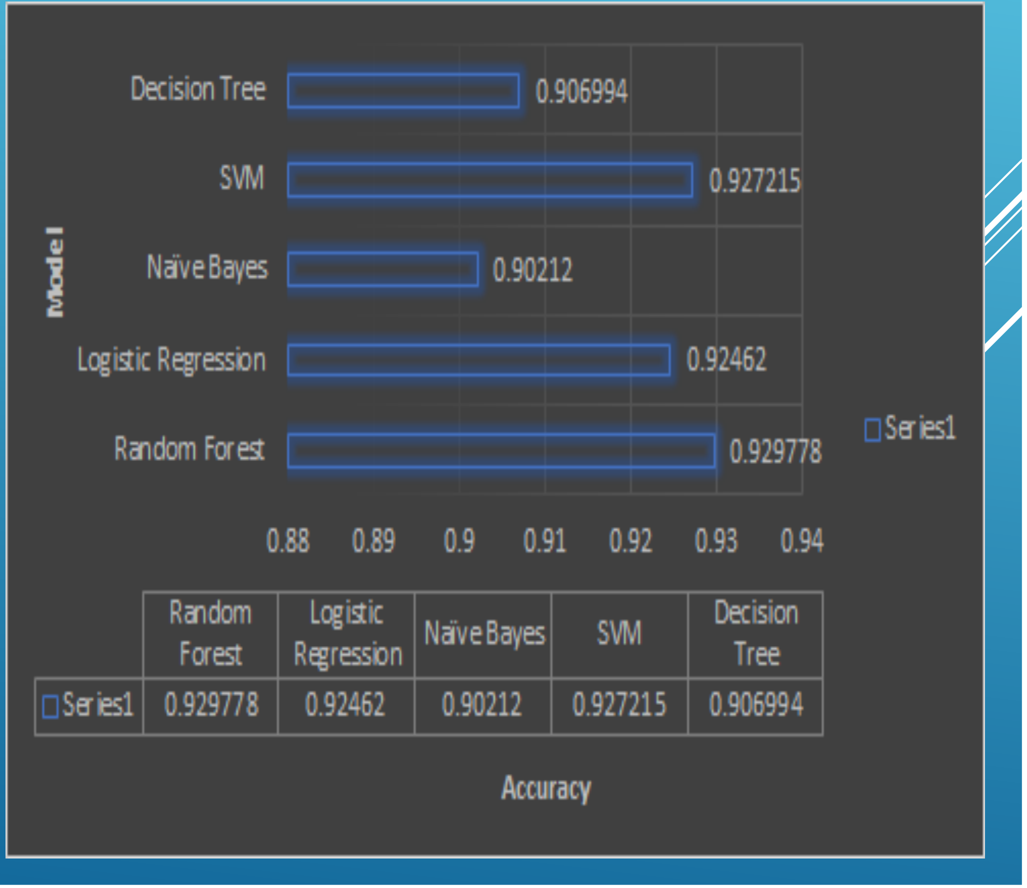
- What are hyperparameters?
- **Hyperparameters**, are the parameters that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

- The C and sigma hyperparameters for support vector machines.

- **Here, we can see that Random Forest is giving the highest accuracy score with 93% taking "n_estimators" = 15 for predicting breast cancer for a patient.**



	model	best_score	best_params
0	SVM	0.927215	{'C': 10, 'kernel': 'rbf'}
1	Random_Forest	0.929778	{'n_estimators': 15}
2	Logistic_Regression	0.924620	{'C': 1}
3	Naive_Bayes_Gaussian	0.902120	{}
4	Decision_Tree	0.906994	{'criterion': 'entropy'}



	Random Forest	Logistic Regression	Naïve Bayes	SVM	Decision Tree
□ Series1	0.929778	0.92462	0.90212	0.927215	0.906994

CONCLUSION

- Of all the models applied on the dataset, grid search CV suggests the best accuracy score of the Random Forest classifier.
- Even after the hyperparameter tuning of the various models, the best scores of each fell less of the accuracy score of the Random Forest .
- 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. mean values of texture, smoothness does not show a particular preference of one diagnosis over the other
- It performs better in the case of categorical input variables as compared to numerical variables. In the dataset most of the selected variables were categorical.

OTHER APPICATIONS OF OUR MODEL

- Text Mining : Machine learning techniques are used to extract text information about typical symptoms of malignant breast cancer by checking medical records of patients. machine learning algorithm to find the most common symptoms of breast cancer that patients reported. They algorithm checked 103,564 sentences to identify pain, fatigue, and nausea as the most common cancer symptoms
- Segregation analysis can be done to provide evidence and identification of a single rare dominant allele carried by people leading to increased susceptibility to breast cancer
- It can used to identify risk to breast cancer based on demographic and lifestyles that an individual follow
- It can also be modified to predict the minimum the amount required for treatment and the time span so that critical decision can be taken accurately

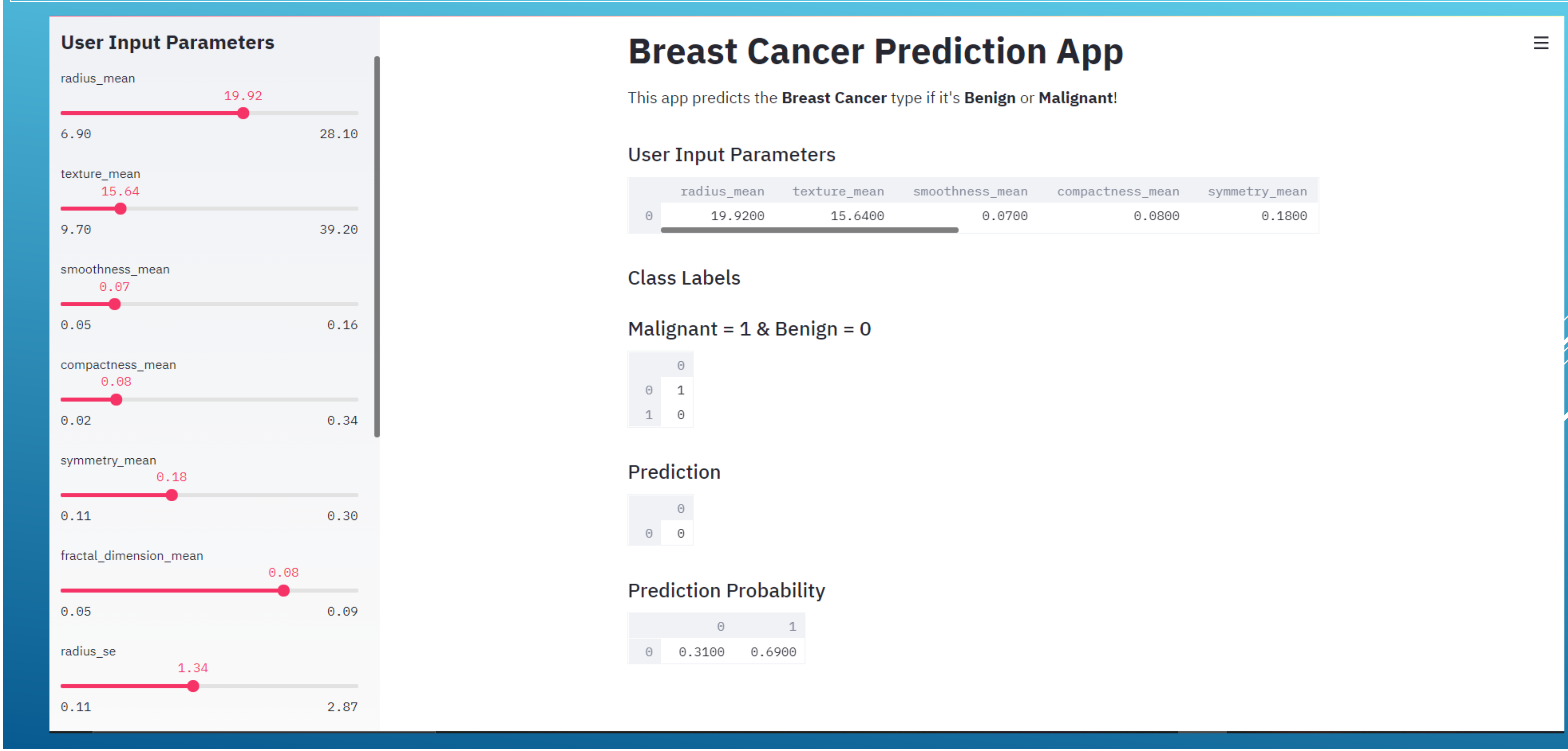
ADVANATAGES OF BREAST CANCER MODEL

- Easy model building less formal statistical knowledge required
- Capable of capturing interaction between predictors.
- Capable of capturing between predictors and outcome

DISADVANATGES OF BREAST CANCER MODEL

- Clinical intrepretation of model parameters is difficult
- Prone to overfitting due to the complexity of model structure.

PROTOTYPE OF PREDICTION MODEL ON WEB



User Input Parameters

radius_mean
19.92
6.90 — 28.10

texture_mean
15.64
9.70 — 39.20

smoothness_mean
0.07
0.05 — 0.16

compactness_mean
0.08
0.02 — 0.34

symmetry_mean
0.18
0.11 — 0.30

fractal_dimension_mean
0.08
0.05 — 0.09

radius_se
1.34
0.11 — 2.87

Breast Cancer Prediction App

This app predicts the **Breast Cancer** type if it's **Benign** or **Malignant**!

User Input Parameters

	radius_mean	texture_mean	smoothness_mean	compactness_mean	symmetry_mean
0	19.9200	15.6400	0.0700	0.0800	0.1800

Class Labels

Malignant = 1 & Benign = 0

	0
0	1
1	0

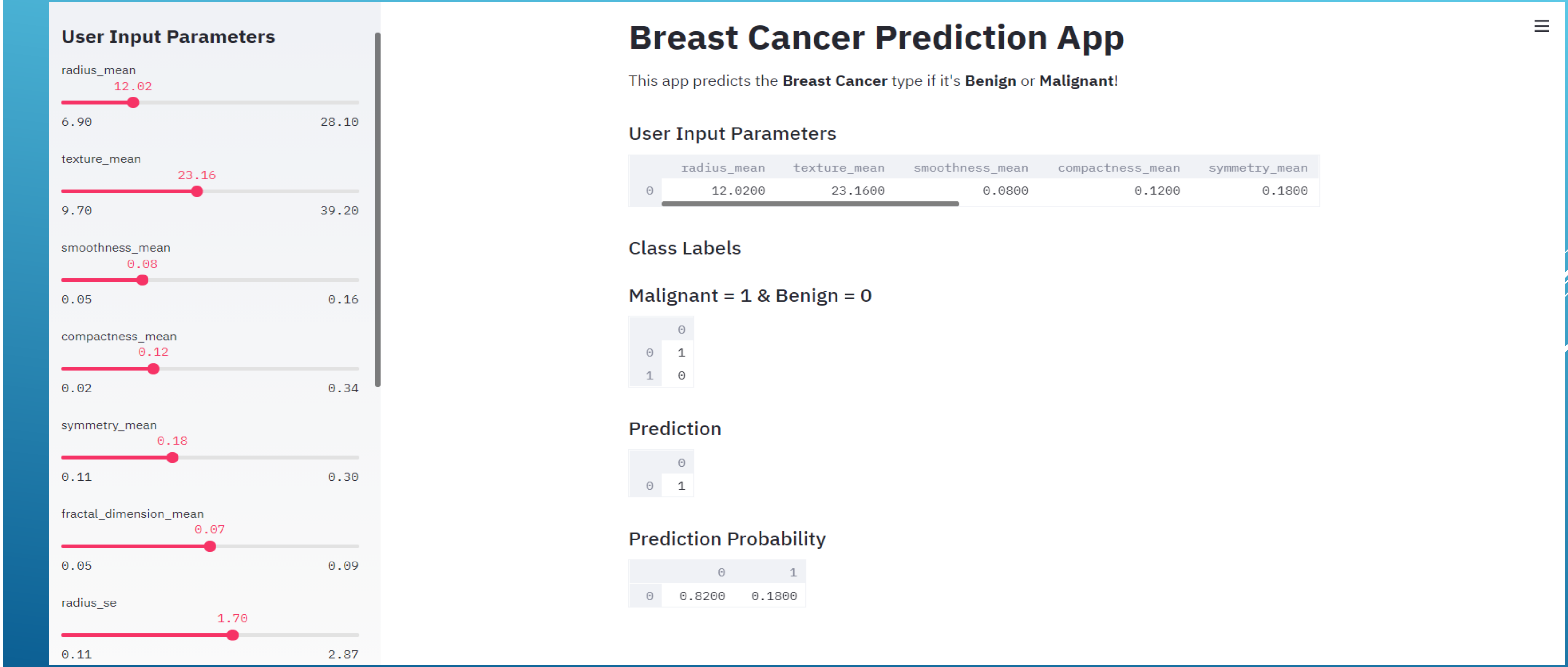
Prediction

	0
0	0

Prediction Probability

	0	1
0	0.3100	0.6900

PROTOTYPE OF PREDICTION MODEL ON WEB



User Input Parameters

radius_mean
12.02
6.90 28.10

texture_mean
23.16
9.70 39.20

smoothness_mean
0.08
0.05 0.16

compactness_mean
0.12
0.02 0.34

symmetry_mean
0.18
0.11 0.30

fractal_dimension_mean
0.07
0.05 0.09

radius_se
1.70
0.11 2.87

Breast Cancer Prediction App

This app predicts the **Breast Cancer** type if it's **Benign** or **Malignant**!

User Input Parameters

	radius_mean	texture_mean	smoothness_mean	compactness_mean	symmetry_mean
0	12.0200	23.1600	0.0800	0.1200	0.1800

Class Labels

Malignant = 1 & Benign = 0

	0
0	1
1	0

Prediction

	0
0	1

Prediction Probability

	0	1
0	0.8200	0.1800

THANK YOU