International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 1076-1083 www.ijaem.net ISSN: 2395-5252

Breast Cancer Prediction (ML based mobile application which detects weather a person is suffering from Breast cancer or not by having Chest x-ray images as an input)

Radhika
Dronacharya College of Engineering,Farukhnagar, Gurugram, Haryana

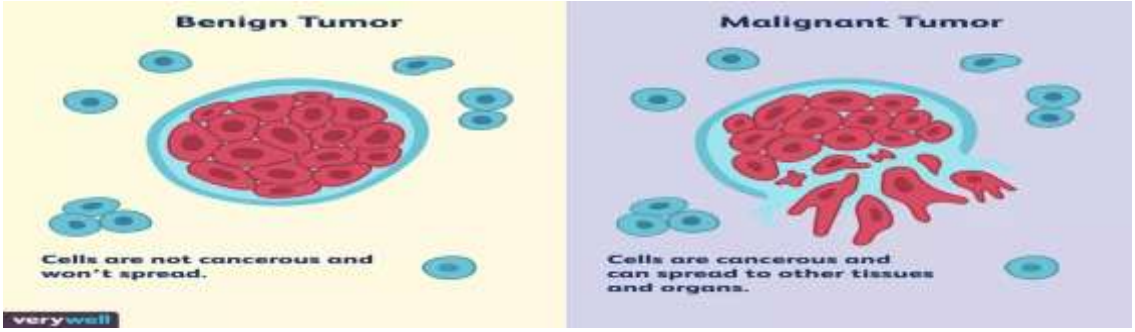
Submitted: 10-09-2021 Revised: 19-09-2021 Accepted: 23-09-2021

ABSTRACT
Breast cancer is the most diagnosed cancer among women worldwide, accounting for 1 in 4 cancer cases. It is the most frequent cancer amongst both sexes and is the leading cause of death from cancer in women. The estimated 2.3 million new cases indicate that one in every 8 cancers diagnosed in 2020 is breast cancer. In 2020, there were an estimated 684,996 deaths from breast cancer, with a disproportionate number of these deaths occurring in low-resource settings.The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients and reduces the death rates.
The doctors do not identify each and every breast cancer patient. That's the reason Machine Learning Engineer / Data Scientist comes into the picture because they have knowledge of maths and computational power.
KEYWORDS : Breast Cancer Prediction, Logistic Regression, Efficient Mobile Apllication, Python, Numpy, Pandas, MatPlotLib, Sklearn

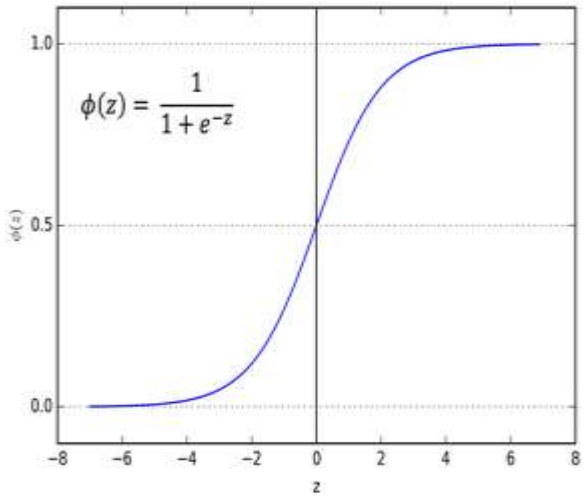
Goal of the ML project
We have extracted features of breast cancer patient cells and normal person cells.The aim of this analysis is to use Logistic Regression to classify the data into two classes of diagnosis— Malignant & Benign. The evaluation metrics used are accuracy, ROC, confusion matrix, precision-recall.

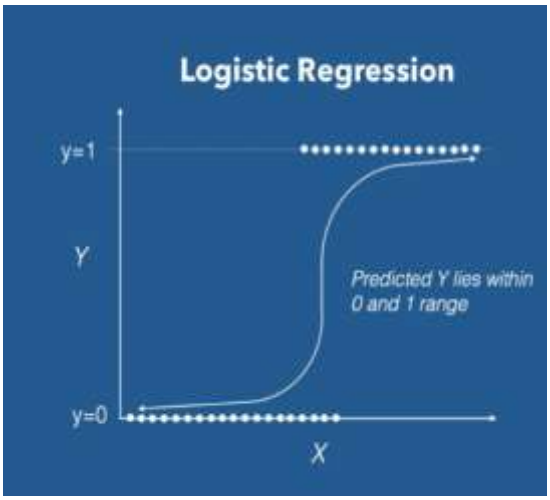
Algorithmused: Logistic regression

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. Here, binary classification is used so logistic regression is most suitable algorithm.In simple words, the dependent variable is binary in nature having data coded as either 1 or 0.



International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 1076-1083 www.ijaem.net ISSN: 2395-5252

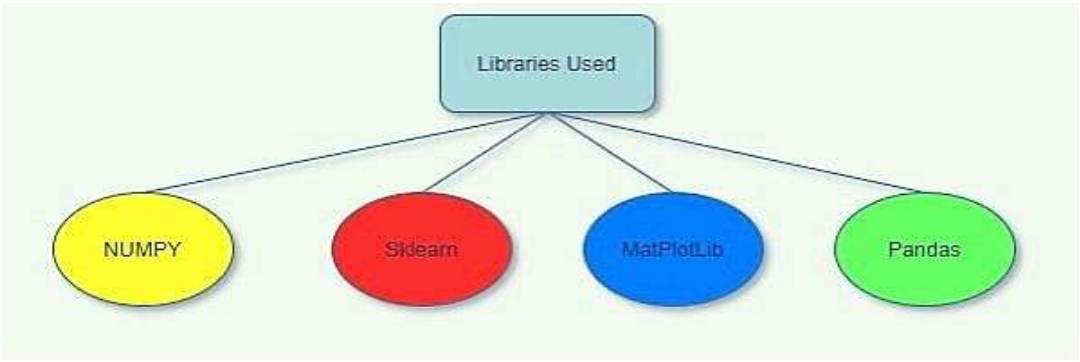




Programming language used : Python

Libraries used : Numpy , Sklearn , MatPlotLib, Pandas.

- **Numpy:** NumPy is a library for supporting work with large multi-dimensional arrays, with many mathematical functions. NumPy is faster for mathematical functions.



- **Sklearn:** The Sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification,

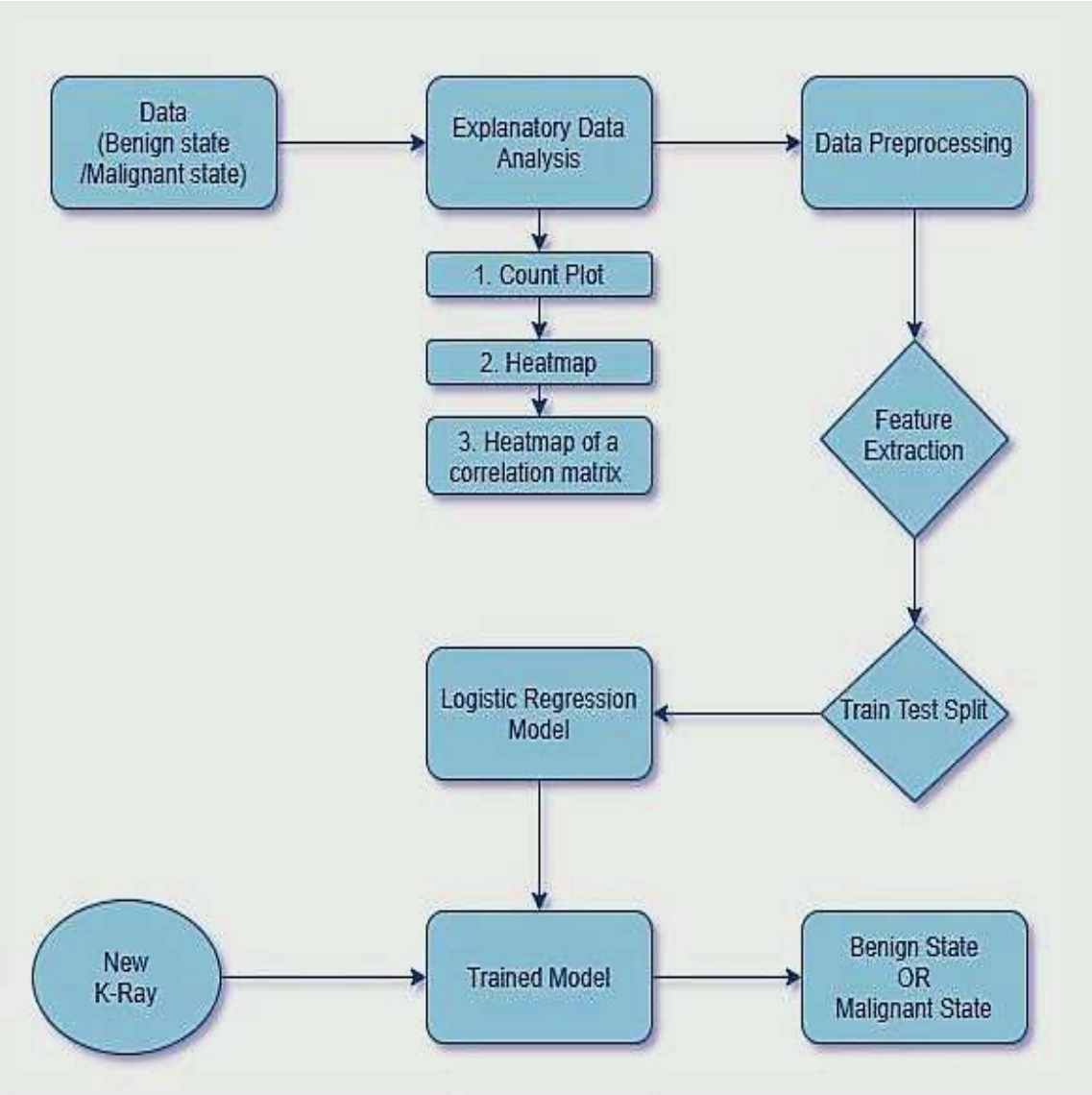
International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 1076-1083 www.ijaem.net ISSN: 2395-5252

regression, clustering and dimensionality reduction.

- **Pandas:** to import data , Pandas is a library that allows manipulation of large arrays of data. Data may be indexed and manipulated based on index.
- **MatPlotLib** : for plotting graphs.

METHODOLOGY

Dataset :



The Wisconsin Breast Cancer (Diagnostic) dataset has been extracted from the UCI Machine Learning Repository. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

- Class distribution: **357** benign, **212** malignant
- Number of instances: **569**; Number of attributes: **32**

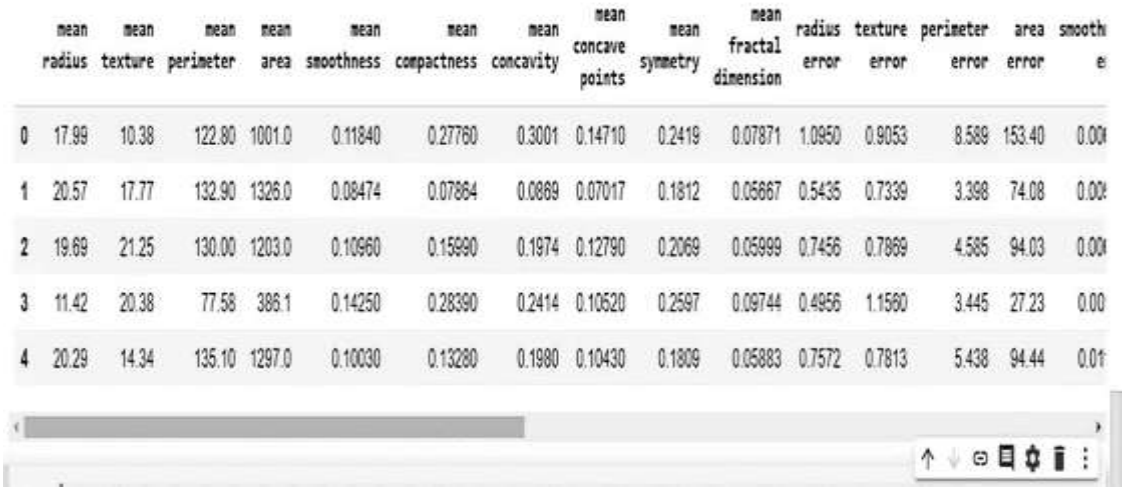
Attributes:

Thirty real-valued features are computed for each cellnucleus. Some of them are:

- radius (mean of distances from center to points on the perimeter)

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 1076-1083 www.ijaem.net ISSN: 2395-5252

- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter² / 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)



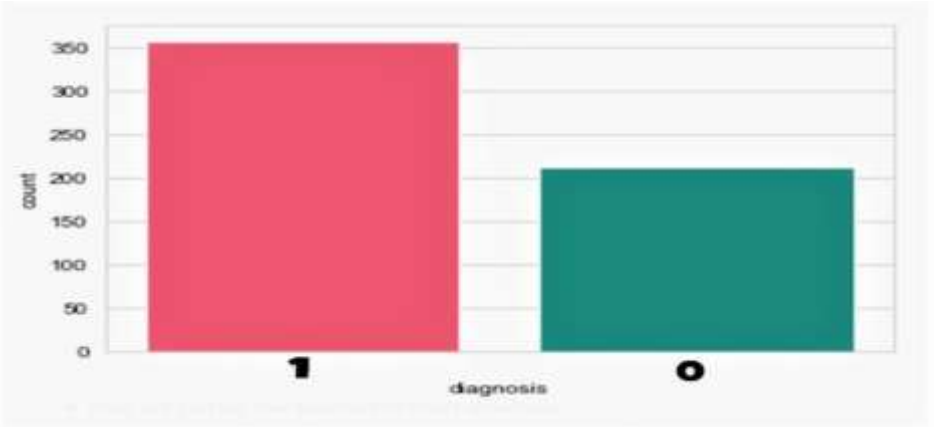
	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	radius error	texture error	perimeter error	area error	smooth e
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	1.0950	0.9053	8.589	153.40	0.00
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.00
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	0.4956	1.1560	3.445	27.23	0.00
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01

Exploratory Data Analysis :
Exploratory Data Analysis (EDA) answers the "What are we dealing with?" question. EDA is where we try to understand our data first. We want to gain insights before messing around with it.

Visualizations are a great way to do this.

- Count Plot
- Heat Map
- Heatmap of a correlation matrix

Visualization #1: Count Plot

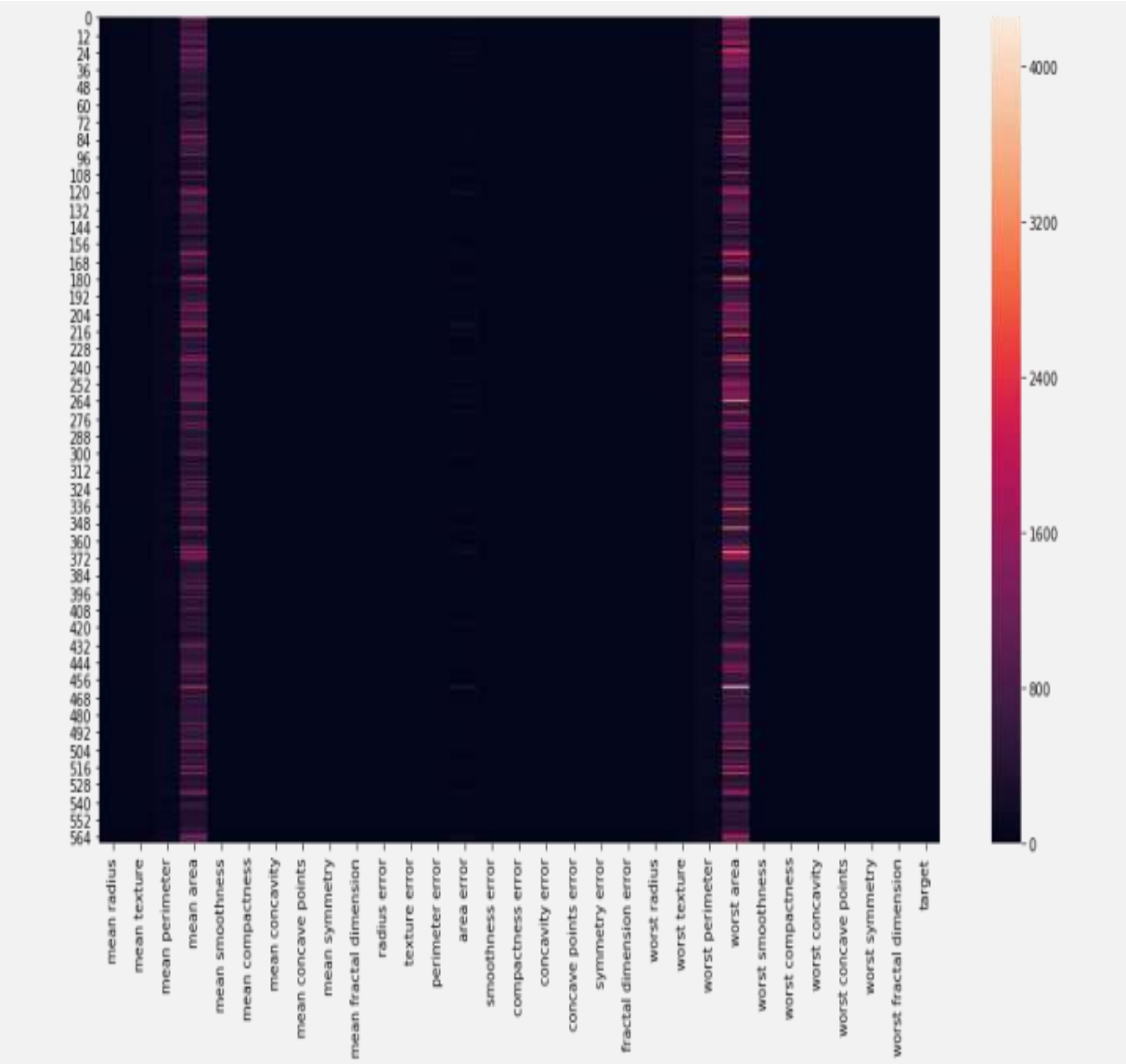


where:
0 indicates Malignant state (breast cancer)
1 indicates Benign state (no breast cancer)

Visualization #2: Heat Map

In the below heatmap we can see the variety of different feature's value. The value of feature **'mean area'** and **'worst area'** are greater than other and 'mean perimeter', 'area error', and 'worst perimeter' value slightly less but greater than remaining features.

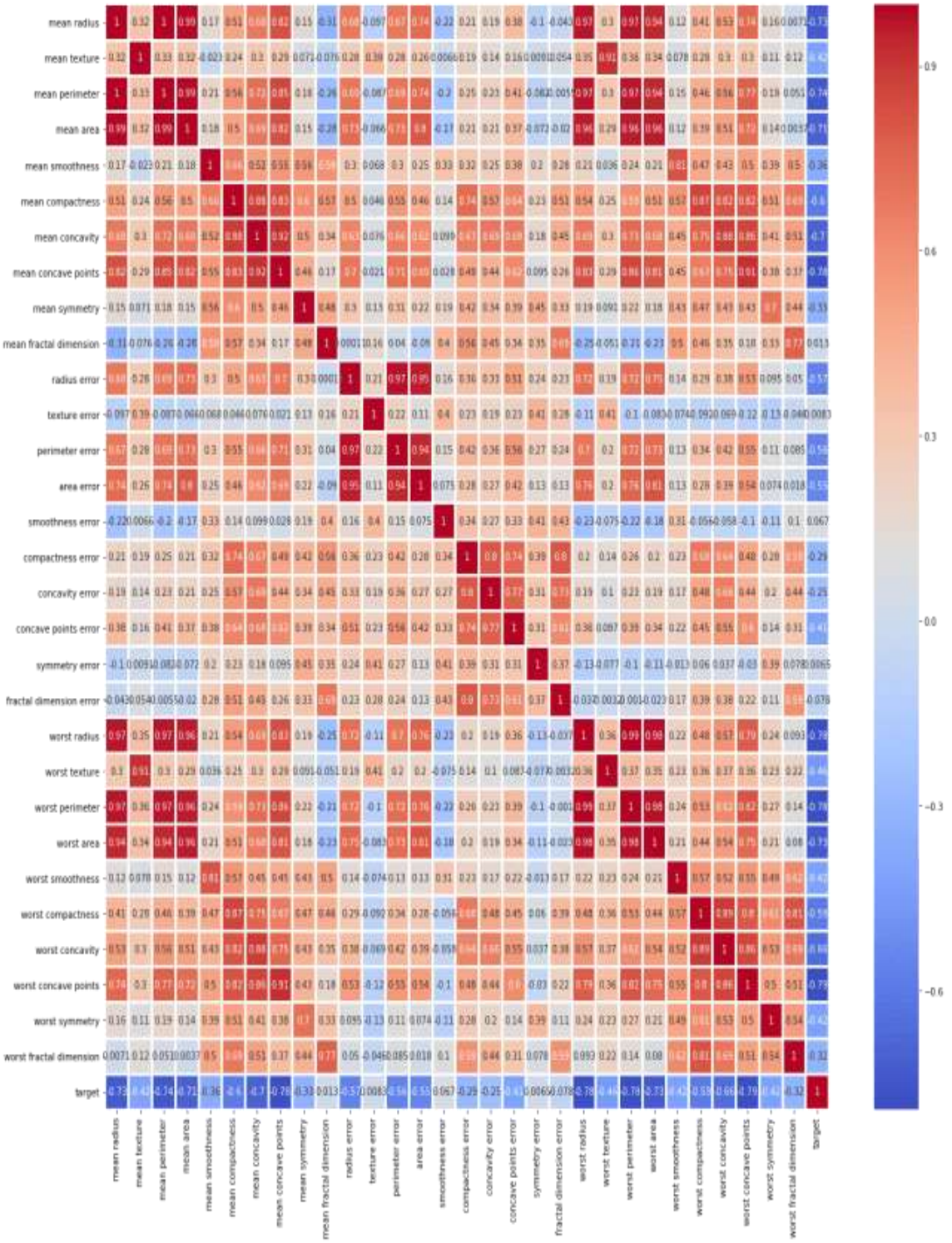
International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 1076-1083 www.ijaem.net ISSN: 2395-5252



Visualization #3:Heatmap of a correlation matrix

To find a correlation between each feature and target we visualize heatmap using the correlation matrix.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 1076-1083 www.ijaem.net ISSN: 2395-5252



Data Preprocessing :

The dataset provided is already clean and does not have any missing values. As a part of the preprocessing stage, the data is standardized using Standard Scaler library. It transforms the attributes to normal distribution.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 1076-1083 www.ijaem.net ISSN: 2395-5252

Split DataFrame in train and test

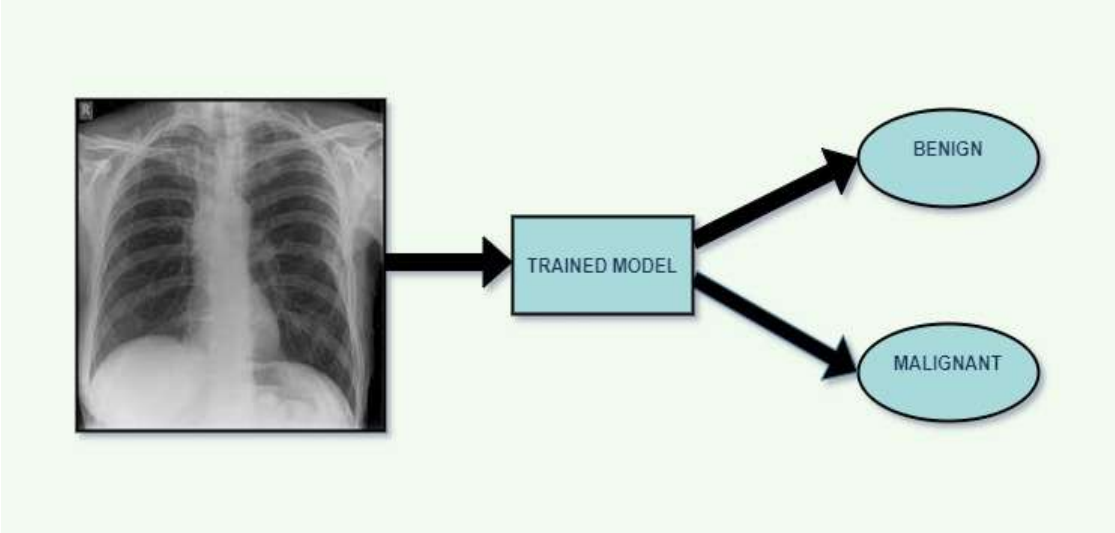
The Python code for training and testing the accuracy of model:

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X,Y)
print(Y.shape, Y_train.shape, Y_test.shape)
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.1)
print(Y.shape, Y_train.shape, Y_test.shape)
print(Y.mean(), Y_train.mean(),Y_test.mean())
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.1, stratify=Y)
print(Y.mean(), Y_train.mean(),Y_test.mean())
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.1, stratify=Y, random_state=1)
print(X_train.mean(), X_test.mean(), X.mean())
print(X_train)
```

Logistic Regression:

```
[33] 1 # import accuracy_score
     2 from sklearn.metrics import accuracy_score

[35] 1 prediction_on_training_data = classifier.predict(X_train)
     2 accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)

[36] 1 print('Accuracy on training data : ', accuracy_on_training_data)

Accuracy on training data :  0.951171875

[37] 1 # prediction on test_data
     2 prediction_on_test_data = classifier.predict(X_test)
     3 accuracy_on_test_data = accuracy_score(Y_test, prediction_on_test_data)

[38] 1 print('Accuracy on test data : ', accuracy_on_test_data)

Accuracy on test data :  0.9298245614035088
```

Now,it can be detected whether person have breast cancer or not.



FUTURE SCOPE

The rapid growth of mobile device development indicates an increasing trend line every year. According to, the use of mobile device is dominated by women. The time that women are spending on mobile device isapproximately 8% higher than men. These statistics is same for the rural area. It shows the higher opportunity to improve health services for women by using mobile device application.As we know, the rate of breast cancer is increasing rapidly.The doctors do not identify each and every breast cancer

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 1076-1083 www.ijaem.net ISSN: 2395-5252

patient.But it is possible using a machine learning model because it includes knowledge of mathematics and computational power.By Machine Learning based Mobile application it becomes easy to use for people.

My main focus is to be able to take this technology and use it in rural India where 70% of our population reside and empower people where, talking about breast cancer is a taboo.It will be available in English,Hindi and some regional languages.This app aims to address this issue by providing every conceivable information about every aspect of breasthealth making people better informed and better prepared to make informed decisions.A simple and user friendly application it mainly relates to three components — benign (non-cancer breast health issues),Malignant (breast cancer information) and common myths and facts about various aspects of the cancer.All one needs to do is to download the app. Once the user opens the app there will be the option to upload the X-ray image,then after visualizing it will give the information regarding breast cancer accordingly.

We need to reach a wider audience and these technological innovations help us in penetrating remote areas.For this we can collaborate with a pathology, they encounter a number of people having some kind of health issue everyday.

In rural areas , the information can be shared by organizing health camps by the pathologies.The early diagnosis is important because

- Spotting cancer early increases the chances of survival
- Diagnosing cancer before it has the chance to spread too far means that treatment is more likely to be successful
- You know your body best, so you can take test using the app with the help of your chest X-ray if something doesn't seem right.

CONCLUSION

Logistic regression model does not necessarily require data feature scaling of data,neither is it greatly affected by unbalanced data nor dependency among data setfeatures. Hence, for medium size data, logistic regression is a good probabilisticprediction model to employ for a binary classification problem, because of its simplicityand less time complexity; therefore, logistic regression model can be used for theprediction of breast cancer, which greatly help physicians to make proper and earlydiagnosis, which will go a long way in increasing the survivability rate of breast cancerpatients.

Breast cancer apps should be readily available.While several tools and apps exist for awareness, screening, and education about available treatments, very few relevant resources are available in app form that specifically address the needs of breast cancer survivors. With better awareness programs, screening and treatment options available,there needs to be a conscious shift towards providing more tools and resources for survivors, as their population and life-expectancy continue to grow.There is high opportunity to improve health services for women by using mobile device application.