# BDA MINI-PROJECT

1. **Title of Mini Project:**
   **Predicting Breast Cancer with the help of Logistic Regression**

2. **Team members:**
   a. Indira Pimpalkhare (PB17)
   b. Priya Bannur (PB23)
   c. Khushboo Agarwal (PB34)

3. **Introduction and Motivation**
   a. Breast cancer is cancer that develops in breast cells. The uncontrolled cancer cells often invade other healthy breast tissue and can travel to the lymph nodes under the arms.
   b. Breast cancer (BC) is the most common cancer in women, affecting about 10% of all women at some stages of their life. In recent years, the incidence rate keeps increasing.
   c. Early prediction of breast cancer is one of the most crucial works in the follow-up process. Data mining methods can help to reduce the number of false-positive and false-negative decisions.

4. **Literature Survey**
   a. Breast Cancer Prediction via Machine Learning (2019) [1]
      i. Used k-Nearest-Neighbours (KNN) technique. KNN is a technique that is used for the classification of data in machine learning. It will perform classification by finding the nearest and similar data points within the corresponding dataset, and it will perform a pre-trained guess depending on that classifications.
      ii. In this study, the breast cancer Wisconsin (Diagnostic) Data set collected from an online data mining repository of the University of California (UCI).
      iii. They used a hybrid of Gradient boosting, Random Forest, SVM and KNN.
      iv. The split percentage was 66%.
      v. Accuracy: Around 70%

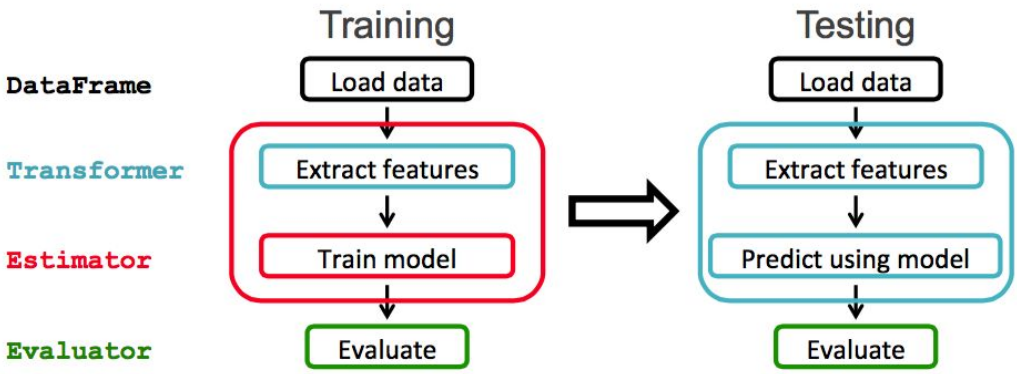
b. An IOT-based predictive system based on machine learning to successfully diagnose people with breast cancer [2]
    i. The dataset Wisconsin Diagnostic Breast Cancer (WDBC) available at the UCI machine learning repository was used.
    ii. Algorithms used- REF, SVM, preprocessing techniques.
    iii. REF is a feature selection algorithm that fits a model and removes the irrelevant feature or features until the specified number of features is reached.
    iv. The dataset was divided into 70% for training the classifier and 30% for the validation of the classifier.
    v. Accuracy- Around 99%

5. **Dataset Source:**
    a. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
    b. Name: Breast Cancer Wisconsin (Diagnostic) Data Set
    c. About the dataset:
        i. ID number (1)
        ii. Diagnosis (M = malignant, B = benign) (2)
        iii. (3-32) Attributes - Ten real-valued features are computed for each cell nucleus:
    d. The **mean**, **standard error** and **"worst" or largest** (mean of the three largest values) - computed for each image
    e. Total of 30 features.
    f. Class distribution: 357 benign, 212 malignant

6. **Application domain:** Healthcare

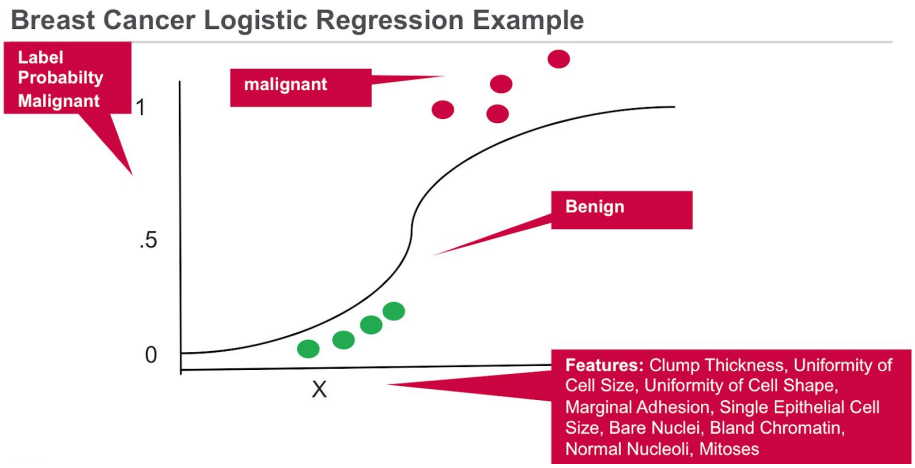
7. **Front end:**
    a. Python

8. **Back end:**
    a. MongoDB
    b. Python
    c. Apache Spark (PySpark)




**Machine Learning using Apache Spark**

9. **Proposed machine learning techniques to be used:**
    a. Logistic Regression

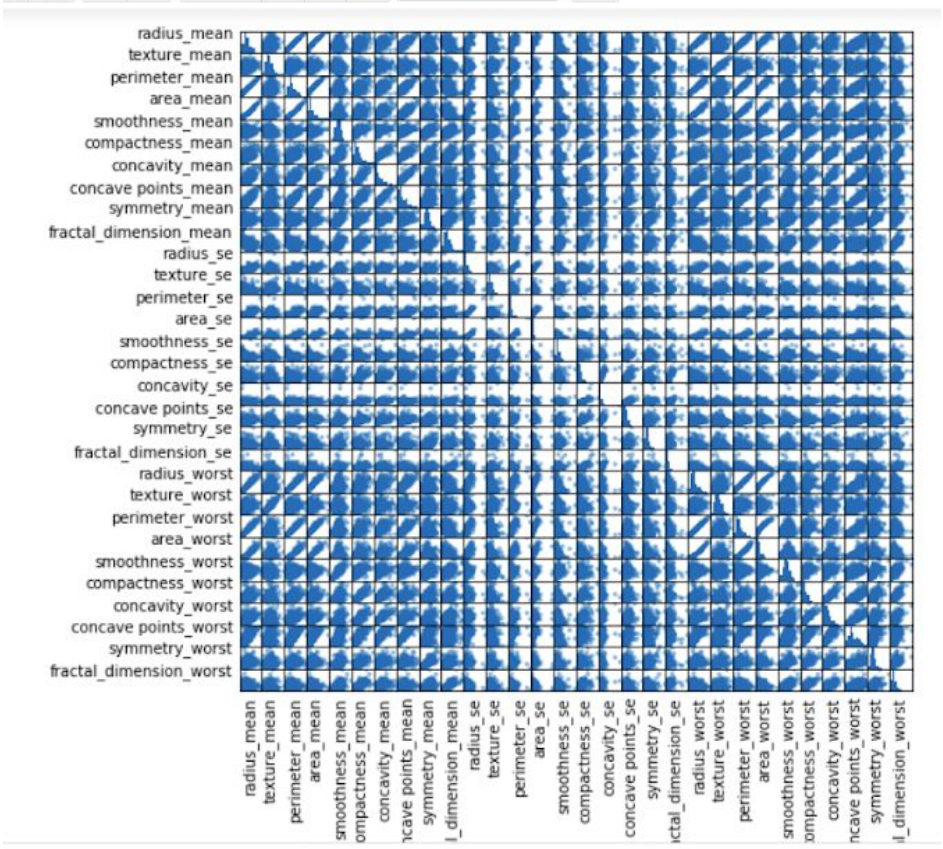


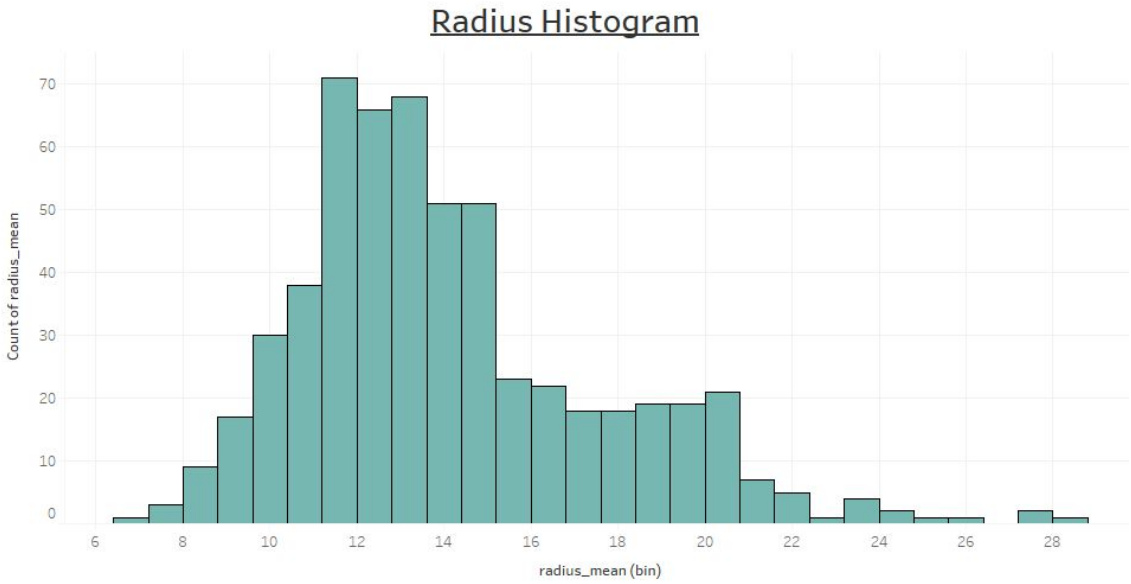

10. **Big Data Visualization tool to be used:**
    a. Tableau
    b. MatPlotLib

## 11. Results

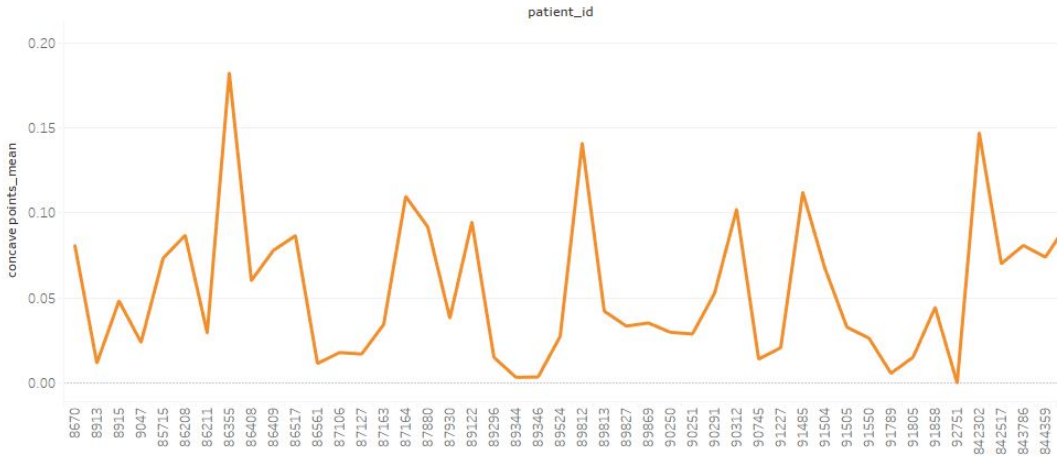
a. Visualizations

**Scatter Matrix for all attributes**



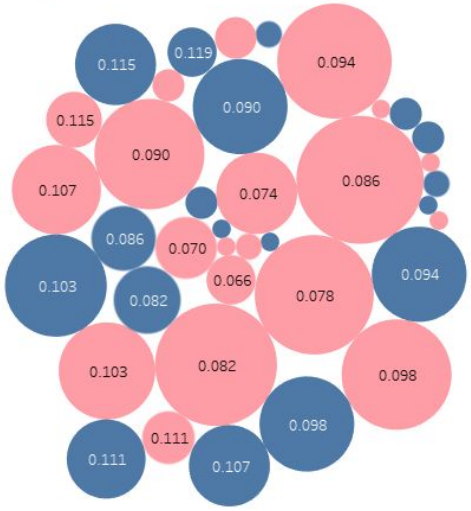



**Line Graph** showing concave points of each patient:

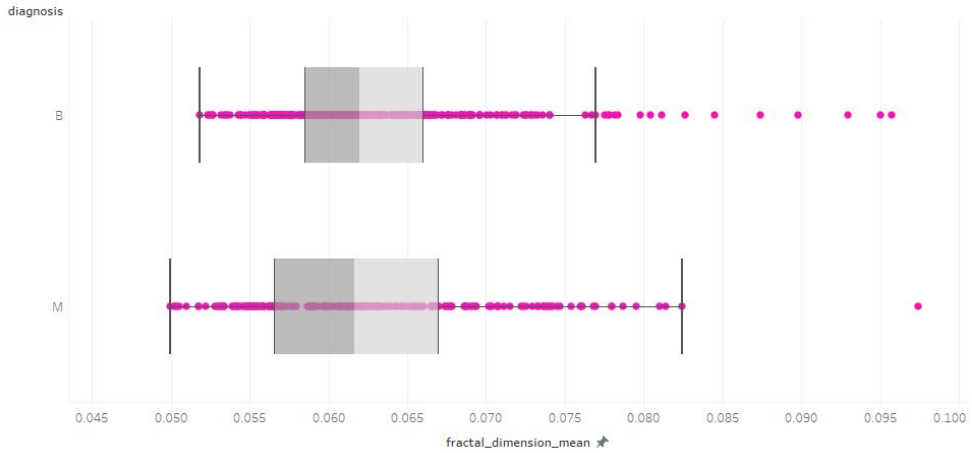



## Bubble Chart

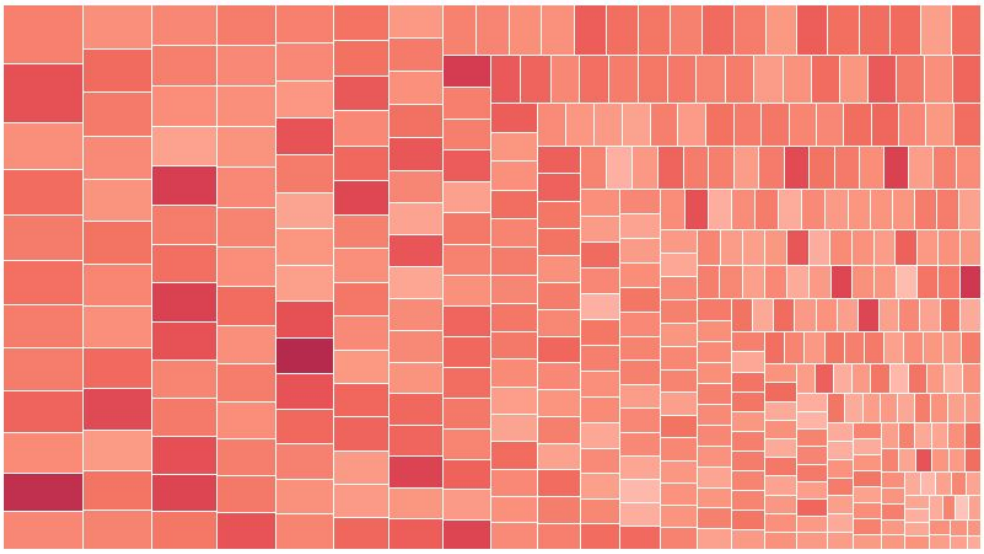
(colour:- **Malignant**, Benign; size:-no. of people diagnosed, label:- smoothness_index )




**Box Plot** for outlier detection of the *fractal-dimension*:



## Tree Map (darker colour=more symmetry, greater size=more compactness)



## Scatter Plot : Area vs Texture (Benign+, Malignant+)

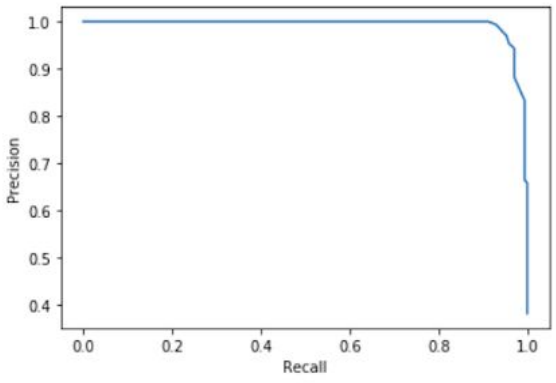
b. Results

**Label and the corresponding predictions**

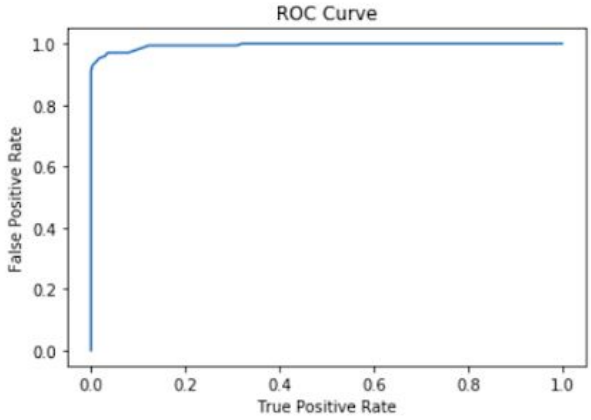
```
In [43]: 1 predict_test=model.transform(test)
         2 predict_test.select("label","prediction").show()
```

```
+-----+----------+
|label|prediction|
+-----+----------+
|  1.0|       1.0|
|  1.0|       1.0|
|  0.0|       0.0|
|  1.0|       1.0|
|  1.0|       1.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       1.0|
|  1.0|       1.0|
|  1.0|       1.0|
|  1.0|       1.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       1.0|
|  1.0|       1.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
+-----+----------+
only showing top 20 rows
```

**Precision vs. Recall**




**ROC Curve for Training Set**




```
Training set areaUnderROC: 0.9950039968025579
```

**The area under ROC curve for Test set**

```
In [73]: from pyspark.ml.evaluation import BinaryClassificationEvaluator
         evaluator=BinaryClassificationEvaluator(rawPredictionCol='rawPrediction',labelCol='label')
         predict_test.select("label","rawPrediction","prediction","probability").show(5)
         print("The area under ROC for train set is {}".format(evaluator.evaluate(predict_train)))
         print("The area under ROC for test set is {}".format(evaluator.evaluate(predict_test)))
```

```
+-----+--------------------+----------+--------------------+
|label|       rawPrediction|prediction|         probability|
+-----+--------------------+----------+--------------------+
|  1.0|[-9.1119950415597...|       1.0|[1.10322201416982...|
|  1.0|[-1.3667099321543...|       1.0|[0.20315192649587...|
|  1.0|[-5.7255420138824...|       1.0|[0.00325098161966...|
|  1.0|[-7.8215833886522...|       1.0|[4.00825537245302...|
|  1.0|[-5.5248774023552...|       1.0|[0.00397052943089...|
+-----+--------------------+----------+--------------------+
only showing top 5 rows

The area under ROC for train set is 0.9952010871873272
The area under ROC for test set is 0.997925925925926
```

## 12. References


a. M. S. Yarabarla, L. K. Ravi and A. Sivasangari, "Breast Cancer Prediction via Machine Learning," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 121-124, doi: 10.1109/ICOEI.2019.8862533.

b. Memon, M. H., Li, J. P., Haq, A. U., Memon, M. H., & Zhou, W. (2019). *Breast Cancer Detection in the IOT Health Environment Using Modified Recursive Feature Selection. Wireless Communications and Mobile Computing, 2019, 1–19.* doi:10.1155/2019/5176705