

Breast Cancer Analysis Using Machine Learning

line 1: 2nd Given Name Surname
line 2: dept. name of organization
(of Affiliation)
line 3: name of organization
(of Affiliation)
line 4: City, Country
line 5: email address or ORCID

Abstract— Breast Cancer Metabric disease is one of the most important cancers with the highest death rate. Breast cancer is occurring day by day in most women due to hormonal imbalances. In the research analysis, we are going to perform machine learning techniques such as classification techniques to predict the diagnosis of breast cancer. Exploratory data analysis is one of the important factors to understand the data features and data information. We have implemented the logistic regression, xg boost, naïve Bayes algorithm and according to this, an excellent performance score is achieved by logistic regression approx. 72%. By implementing the algorithms, we can diagnose breast cancer at an early stage to prevent women from dying.

I. INTRODUCTION

Breast cancer is one of the most common diseases present in women all over the world. According to the research analysis, 41760 women died from breast cancer only. A benign tumor is responsible for minor changes in the structure of the breast but not so harmful. Radio genomic technique provides 2 types of features, radiological and genetic which is helpful for proper diagnosis. Metabric breast cancer is a significant malignancy with staggering mortality rates. Its prevalence among women continues to rise, primarily attributed to hormonal fluctuations. Our study aims to utilize machine learning, specifically classification techniques, to forecast breast cancer diagnoses. Exploratory data analysis stands as a crucial step, enabling us to grasp the intricacies of data features and information.

II. ANALYSIS PERFORMED

Data analysis is a very important technique to predict certain diagnoses and provide proper treatment.

A. Collection of Data:

In the predictive analysis, evaluation of survival in women diagnosed with breast cancer to find the robustness result by using machine learning. To perform the analysis, we have imported Python libraries such as pandas, numpy, matplotlib, and seaborn. To run this code, we use the Jupyter Notebook software for creating the ipynb files. If we don't have to use Jupyter Notebook then we can use Google Colab which is open source where we just need to upload the dataset CSV file. We can read CSV files, excel files, or even text files as well.

```
In [264]: #import libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Figure 1: Python Libraries

We have read the breast cancer metabric dataset which is collected from the Molecular Taxonomy of Breast Cancer International Consortium database and present in Kaggle as well.

Out[265]:

	Patient ID	Age at Diagnosis	Type of Breast Surgery	Cancer Type	Cancer Type Detailed	Cellularity	Chemotherapy	Pam50 + Claudin-low subtype	Cohort	ER status measured by IHC	Overall Survival Status
0	MB-0000	75.65	Mastectomy	Breast Cancer	Breast Invasive Ductal Carcinoma	NaN	No	claudin-low	1.0	Positive ...	Living N
1	MB-0002	43.19	Breast Conserving	Breast Cancer	Breast Invasive Ductal Carcinoma	High	No	LumA	1.0	Positive ...	Living F
2	MB-0005	48.87	Mastectomy	Breast Cancer	Breast Invasive Ductal Carcinoma	High	Yes	LumB	1.0	Positive ...	Deceased F
3	MB-0008	47.68	Mastectomy	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Moderate	Yes	LumB	1.0	Positive ...	Living F
4	MB-0009	78.97	Mastectomy	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	High	Yes	LumB	1.0	Positive ...	Deceased F

Figure 2: Breast Cancer Dataset

There are various attributes present in the dataset to represent the patient's vital status based on the symptoms. So, the data features are ['Patient ID', 'Age at Diagnosis', 'Type of Breast Surgery', 'Cancer Type', 'Cancer Type Detailed', 'Cellularity', 'Chemotherapy', 'Pam50 + Claudin-low subtype', 'Cohort', 'ER status measured by IHC', 'ER Status', 'Neoplasm Histologic Grade', 'HER2 status measured by SNP6', 'HER2 Status', 'Tumor Other Histologic Subtype', 'Hormone Therapy', 'Inferred Menopausal State', 'Integrative Cluster', 'Primary Tumor Laterality', 'Lymph nodes examined positive', 'Mutation Count', 'Nottingham prognostic index', 'Oncotree Code', 'Overall Survival (Months)', 'Overall Survival Status', 'PR Status', 'Radio Therapy', 'Relapse Free Status (Month

s)', 'Relapse Free Status', 'Sex', '3-Gene classifier subtype', 'Tumor Size', 'Tumor Stage', 'Patient's Vital Status']

```
Out[266]: Index(['Patient ID', 'Age at Diagnosis', 'Type of Breast Surgery',
              'Cancer Type', 'Cancer Type Detailed', 'Cellularity', 'Chemotherapy',
              'Pam50 + Claudin-low subtype', 'Cohort', 'ER status measured by IHC',
              'ER Status', 'Neoplasm Histologic Grade',
              'HER2 status measured by SNP6', 'HER2 Status',
              'Tumor Other Histologic Subtype', 'Hormone Therapy',
              'Inferred Menopausal State', 'Integrative Cluster',
              'Primary Tumor Laterality', 'Lymph nodes examined positive',
              'Mutation Count', 'Nottingham prognostic index', 'Oncotree Code',
              'Overall Survival (Months)', 'Overall Survival Status', 'PR Status',
              'Radio Therapy', 'Relapse Free Status (Months)', 'Relapse Free Status',
              'Sex', '3-Gene classifier subtype', 'Tumor Size', 'Tumor Stage',
              'Patient's Vital Status'],
              dtype=object)
```

Figure 3: Breast Cancer Attributes

The number of rows and columns present in the breast cancer dataset are 2509 and 34 respectively.

```
In [267]: #rows and columns
          breast.shape

Out[267]: (2509, 34)
```

Figure 4: Rows and Columns in the Data

When we collect data, it will be in a structured format but of any type. Here we also check the data type of the column so it is also of integer, object, and float type.

```
268]: Patient ID      object
      Age at Diagnosis float64
      Type of Breast Surgery object
      Cancer Type      object
      Cancer Type Detailed object
      Cellularity       object
      Chemotherapy      object
      Pam50 + Claudin-low subtype object
      Cohort            float64
      ER status measured by IHC object
      ER Status         object
      Neoplasm Histologic Grade float64
      HER2 status measured by SNP6 object
      HER2 Status       object
      Tumor Other Histologic Subtype object
      Hormone Therapy   object
      Inferred Menopausal State object
      Integrative Cluster object
      Primary Tumor Laterality object
      Lymph nodes examined positive float64
      Mutation Count    float64
      Nottingham prognostic index float64
      Oncotree Code     object
      Overall Survival (Months) float64
      Overall Survival Status object
      PR Status         object
      Radio Therapy     object
      Relapse Free Status (Months) float64
      Relapse Free Status object
      Sex              object
      3-Gene classifier subtype object
      Tumor Size       float64
      Tumor Stage      float64
      Patient's vital Status object
      dtype: object
```

Figure 5: Data Type of Columns

Data Information is also very important to understand so here we use the info function which provides some details such as data type, total patient entries, non-null count, and total attribute counts.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2509 entries, 0 to 2508
Data columns (total 34 columns):
#   Column                                     Non-Null Count  Dtype
---  --
0   Patient ID                               2509 non-null   object
1   Age at Diagnosis                         2498 non-null   float64
2   Type of Breast Surgery                   1955 non-null   object
3   Cancer Type                             2509 non-null   object
4   Cancer Type Detailed                     2509 non-null   object
5   Cellularity                             1917 non-null   object
6   Chemotherapy                             1980 non-null   object
7   Pam50 + Claudin-low subtype              1980 non-null   object
8   Cohort                                  2498 non-null   float64
9   ER status measured by IHC                2426 non-null   object
10  ER Status                               2469 non-null   object
11  Neoplasm Histologic Grade                 2388 non-null   float64
12  HER2 status measured by SNP6              1980 non-null   object
13  HER2 Status                             1980 non-null   object
14  Tumor Other Histologic Subtype            2374 non-null   object
15  Hormone Therapy                          1980 non-null   object
16  Inferred Menopausal State                1980 non-null   object
17  Integrative Cluster                      1980 non-null   object
18  Primary Tumor Laterality                 1870 non-null   object
19  Lymph nodes examined positive             2243 non-null   float64
20  Mutation Count                           2357 non-null   float64
21  Nottingham prognostic index               2287 non-null   float64
22  Oncotree Code                            2509 non-null   object
23  Overall Survival (Months)                 1981 non-null   float64
24  Overall Survival Status                   1981 non-null   object
25  PR Status                               1980 non-null   object
26  Radio Therapy                           1980 non-null   object
27  Relapse Free Status (Months)              2388 non-null   float64
28  Relapse Free Status                     2488 non-null   object
29  Sex                                       2509 non-null   object
30  3-Gene classifier subtype                 1764 non-null   object
31  Tumor Size                              2360 non-null   float64
32  Tumor Stage                             1788 non-null   float64
33  Patient's Vital Status                   1980 non-null   object
dtypes: float64(18), object(24)
memory usage: 666.6+ KB
```

Figure 6: Information of Data

The breast dataset contains multiple types of values to show the symptoms and type of breast cancer. Every column contains various unique values and also displays the count of unique values which may be helpful at the time of data cleaning.

	Unique	unique
Patient ID	2509	[MB-0000, MB-0002, MB-0005, MB-0008, MB-0008, ...]
Age at Diagnosis	1843	[75.85, 43.19, 48.87, 47.88, 78.97, 78.77, 58...
Type of Breast Surgery	2	[Mastectomy, Breast Conserving, nan]
Cancer Type	2	[Breast Cancer, Breast Sarcoma]
Cancer Type Detailed	8	[Breast Invasive Ductal Carcinoma, Breast Mixe...
Cellularity	3	[nan, High, Moderate, Low]
Chemotherapy	2	[No, Yes, nan]
Pam50 + Claudin-low subtype	7	[claudin-low, LumA, LumB, Normal, nan, Her2, B...
Cohort	9	[1.0, 2.0, 3.0, 5.0, 4.0, 9.0, 7.0, 6.0, nan, ...]
ER status measured by IHC	2	[Positive, Negative, nan]
ER Status	2	[Positive, Negative, nan]
Neoplasm Histologic Grade	3	[3.0, 2.0, 1.0, nan]
HER2 status measured by SNP6	4	[Neutral, Loss, nan, Gain, Undefined]
HER2 Status	2	[Negative, nan, Positive]
Tumor Other Histologic Subtype	8	[Ductal/INST, Mixed, Lobular, Tubular/ cribrifo...
Hormone Therapy	2	[Yes, No, nan]
Inferred Menopausal State	2	[Post, Pre, nan]
Integrative Cluster	11	[4ER+, 3, 0, 7, 4ER-, nan, 5, 8, 10, 1, 2, 6]
Primary Tumor Laterality	2	[Right, Left, nan]
Lymph nodes examined positive	32	[10.0, 0.0, 1.0, 3.0, 8.0, nan, 11.0, 24.0, 4...
Mutation Count	32	[nan, 2.0, 1.0, 4.0, 5.0, 3.0, 7.0, 8.0, 11.0...
Nottingham prognostic index	436	[8.044, 4.02, 4.03, 4.05, 6.08, 4.082, 6.13, 4...
Oncotree Code	8	[IDC, MDLC, ILC, BRCA, IMMC, PBS, BREAST, MBC]
Overall Survival (Months)	1743	[140.5, 84.83333333, 183.7, 184.93333333000004...
Overall Survival Status	2	[Living, Deceased, nan]
PR Status	2	[Negative, Positive, nan]
Radio Therapy	2	[Yes, No, nan]

Figure 7: Data Uniqueness

B. Exploratory Data Analysis:

Exploratory Data analysis is the step that provides the summary of data and plots the chart to understand the data information. In this process, we use the describe function to find the summary of data. We calculate the mean, median, minimum, and maximum, 25%, 50%, and 5% of the numerical data.

Out[278]:

	Age at Diagnosis	Cohort	Neoplasm Histologic Grade	Lymph nodes examined positive	Mutation Count	Nottingham prognostic index	Overall Survival (Months)
count	2498.000000	2498.000000	2388.000000	2243.000000	2357.000000	2287.000000	1981.000000
mean	60.420300	2.900320	2.412080	1.950513	5.578702	4.028787	126.244271
std	13.032997	1.982216	0.848383	4.017774	3.987987	1.189082	78.111772
min	21.930000	1.000000	1.000000	0.000000	1.000000	1.000000	0.000000
25%	50.920000	1.000000	2.000000	0.000000	3.000000	3.048000	60.888887
50%	61.110000	3.000000	3.000000	0.000000	5.000000	4.044000	116.488887
75%	70.000000	4.000000	3.000000	2.000000	7.000000	5.040000	185.133333
max	98.290000	9.000000	3.000000	45.000000	80.000000	7.200000	355.200000

Figure 8: Summary of Breast Data

We have plotted the count plot of the columns to show the count of values. We have displayed the value count of the patient's vital status.

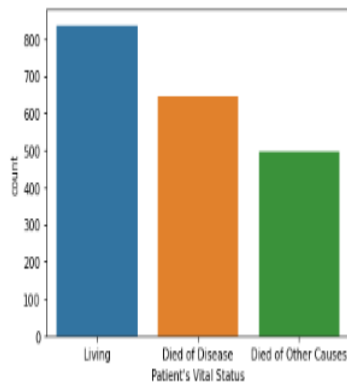


Figure 9: Count Plot of Patient Vital Status

We have displayed the value count of the Oncotree Code and according to the graph, the IDC value is present maximum.

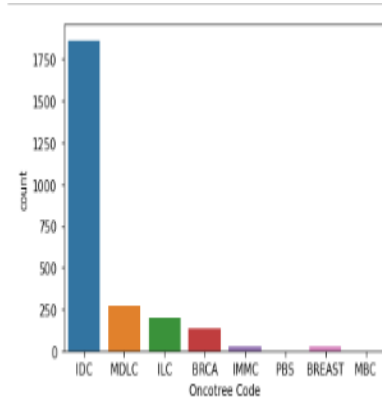


Figure 10: Count Plot of Oncotree Code

We have displayed the value count of the Tumor's other histologic subtype.

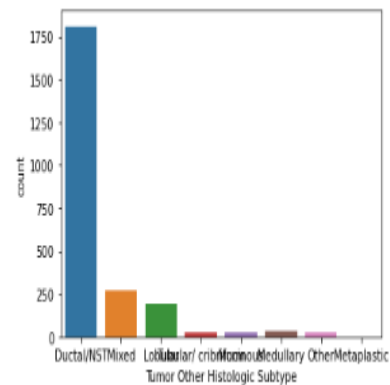


Figure 11: Count Plot of Tumor Other Histologic Subtype

We have displayed the value count of the chemotherapy.

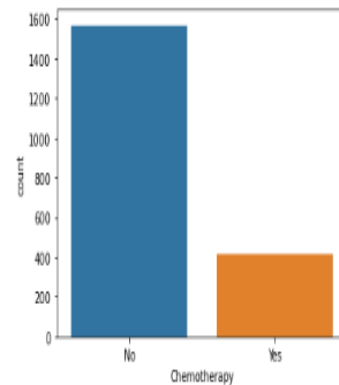


Figure 12: Count Plot of Chemotherapy

We have displayed the value count of the type of breast surgery. Mastectomy count is present more as compared to breast-conserving.

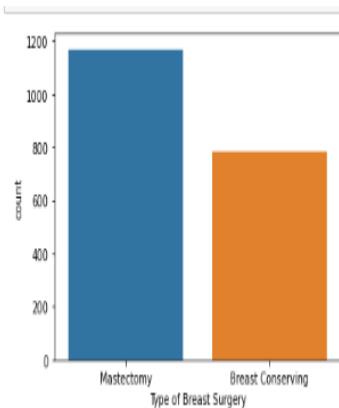


Figure 13: Count Plot of Type of Breast Surgery

We have plotted the histogram on the age of the diagnosis.

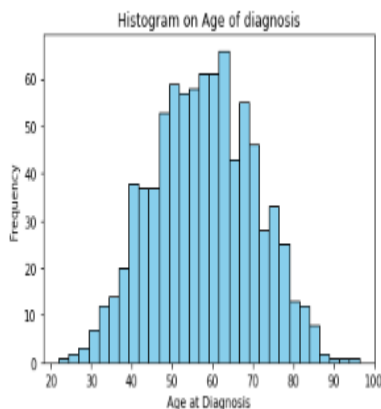


Figure 14: Histogram on Age of Diagnosis

We have plotted the histogram on the overall survival months.

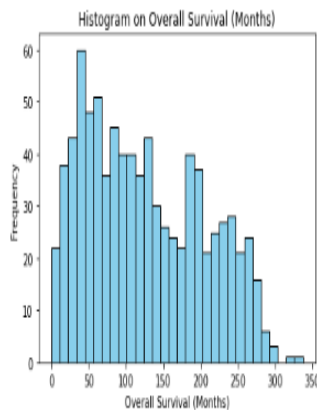


Figure 15: Histogram on Overall Survival Months

C. Data Cleaning:

Data Cleaning is a special technique to clean the data in which we remove duplicate records, missing records, data transformation, and data preprocessing. We have checked the missing records using isnull function and there are various missing records present in the data.

```
5]: Patient ID      0
Age at Diagnosis   11
Type of Breast Surgery 554
Cancer Type        0
Cancer Type Detailed 0
Cellularity         592
Chemotherapy        529
Pam50 + Claudin-low subtype 529
Cohort              11
ER status measured by IHC 83
ER Status           40
Neoplasm Histologic Grade 121
HER2 status measured by SNP6 529
HER2 Status         529
Tumor Other Histologic Subtype 135
Hormone Therapy     529
Inferred Menopausal State 529
Integrative Cluster 529
Primary Tumor Laterality 639
Lymph nodes examined positive 266
Mutation Count      152
Nottingham prognostic index 222
Oncotree Code        0
Overall Survival (Months) 528
Overall Survival Status 528
PR Status            529
Radio Therapy        529
Relapse Free Status (Months) 121
Relapse Free Status  21
Sex                  0
3-Gene classifier subtype 745
Tumor Size           149
Tumor Stage          721
Patient's Vital Status 529
dtype: int64
```

Activate Window

Figure 16: Checking Missing Values

We have dropped all missing data from the main dataset and stored them in a breast variable.

```
In [16]: breast = breast.dropna()

In [17]: breast.shape

Out[17]: (1092, 34)
```

Figure 17: Data Cleaning

We checked the patient's vital status containing 3 types of values so we selected data for living and dead of disease.

```
In [18]: #remove data for Died of Other Causes
breast = breast[breast["Patient's Vital Status"].isin(['Living', 'Died of Diseases'])
breast.shape

Out[18]: (854, 34)
```

Figure 18: Remove Unwanted Data

D. Data Preprocessing

Data Preprocessing is a technique that is used to transform the data in the proper format. Machine learning always tries to work on

data type conversion. Here we performed the label encoding to encode the data from an object or string to an integer.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 854 entries, 1 to 1743
Data columns (total 34 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                -
0   Patient ID                            854 non-null    object
1   Age at Diagnosis                      854 non-null    float64
2   Type of Breast Surgery               854 non-null    int32
3   Cancer Type                          854 non-null    int32
4   Cancer Type Detailed                 854 non-null    int32
5   Cellularity                          854 non-null    int32
6   Chemotherapy                        854 non-null    int32
7   PAM50 + Claudin-low subtype         854 non-null    int32
8   Cohort                               854 non-null    float64
9   ER status measured by IHC            854 non-null    int32
10  ER Status                            854 non-null    int32
11  Neoplasm Histologic Grade            854 non-null    float64
12  HER2 status measured by SNP6         854 non-null    int32
13  HER2 Status                          854 non-null    int32
14  Tumor Other Histologic Subtype       854 non-null    int32
15  Hormone Therapy                      854 non-null    int32
16  Inferred Menopausal State            854 non-null    int32
17  Integrative Cluster                  854 non-null    int32
18  Primary Tumor Laterality             854 non-null    int32
19  Lymph nodes examined positive        854 non-null    float64
20  Mutation Count                       854 non-null    float64
21  Nottingham prognostic index          854 non-null    float64
22  Oncotree Code                        854 non-null    int32
23  Overall Survival (Months)            854 non-null    float64
24  Overall Survival Status              854 non-null    int32
25  PR Status                            854 non-null    int32
26  Radio Therapy                        854 non-null    int32
27  Relapse Free Status (Months)         854 non-null    float64
28  Relapse Free Status                  854 non-null    int32
29  Sex                                  854 non-null    int32
30  3-Genes classifier subtype           854 non-null    int32
31  Tumor Size                           854 non-null    float64
32  Tumor Stage                          854 non-null    float64
33  Patient's Vital Status               854 non-null    int32
dtypes: float64(10), int32(23), object(1)
memory usage: 156.8+ KB
```

Figure 19: Data After Label Encoding

We have plotted the heat map to show the correlation matrix and understand that some of the features are not useful.

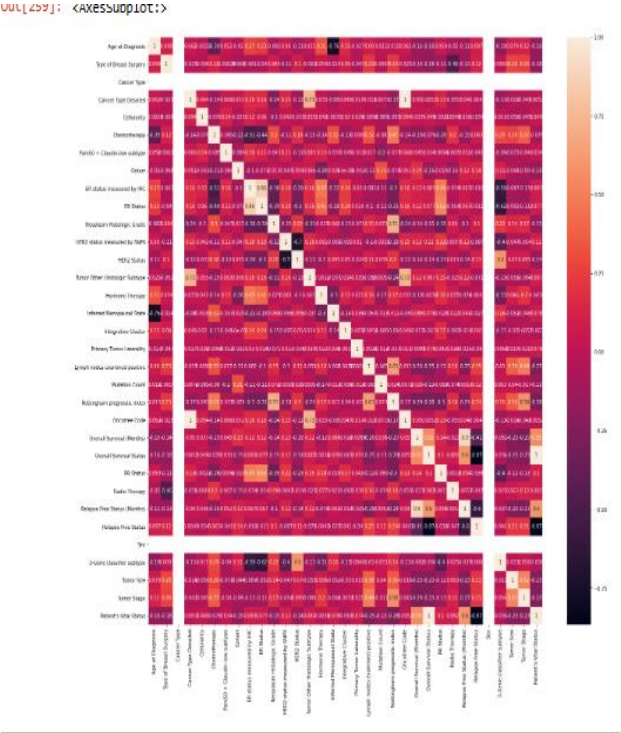


Figure 20: Heat Map

We have dropped the unuseful features from the dataset. Dropped columns are 'Patient ID,' Inferred Menopausal State, 'ER status measured by IHC', 'ER Status', 'Relapse Free Status', 'HER2 status

measured by SNP6', 'HER2 Status', 'Cancer Type', 'Sex', 'PR Status', '3-Genes classifier subtype', 'Relapse Free Status (Months)'

```
Data columns (total 22 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                -
0   Age at Diagnosis                      854 non-null    float64
1   Type of Breast Surgery               854 non-null    int32
2   Cancer Type Detailed                 854 non-null    int32
3   Cellularity                          854 non-null    int32
4   Chemotherapy                        854 non-null    int32
5   PAM50 + Claudin-low subtype         854 non-null    int32
6   Cohort                               854 non-null    float64
7   Neoplasm Histologic Grade            854 non-null    float64
8   Tumor Other Histologic Subtype       854 non-null    int32
9   Hormone Therapy                      854 non-null    int32
10  Integrative Cluster                  854 non-null    int32
11  Primary Tumor Laterality             854 non-null    int32
12  Lymph nodes examined positive        854 non-null    float64
13  Mutation Count                       854 non-null    float64
14  Nottingham prognostic index          854 non-null    float64
15  Oncotree Code                        854 non-null    int32
16  Overall Survival (Months)            854 non-null    float64
17  Overall Survival Status              854 non-null    int32
18  Radio Therapy                        854 non-null    int32
19  Tumor Size                           854 non-null    float64
20  Tumor Stage                          854 non-null    float64
21  Patient's Vital Status               854 non-null    int32
dtypes: float64(9), int32(13)
memory usage: 110.1 KB
```

Figure 21: Drop Unwanted Columns

E. Data Visualization:

Data Visualization is a technique that we use to plot the graphs using some tools or programming language. Here we use the Python language to perform the data visualization and the libraries we used are seaborn, matplotlib, and plotly. We have designed the line chart to show the variation between the variables.

We've created a line chart that illustrates the relationship between breast cancer status and age at diagnosis.

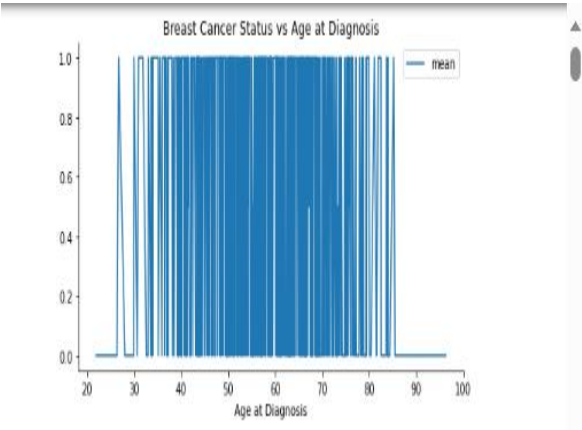


Figure 22: Line Chart between Breast Cancer Status and Age

We've created a line chart that illustrates the relationship between breast cancer status and types of breast surgery.

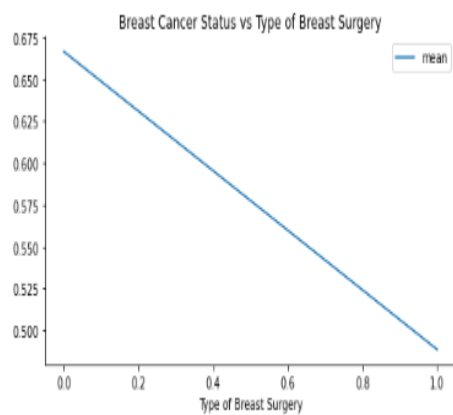


Figure 23: Line Chart between Breast Cancer Status and type of breast surgery

We've created a line chart that illustrates the relationship between breast cancer status and cancer type detailed.

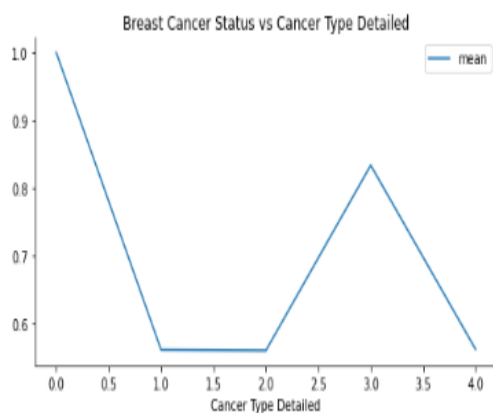


Figure 24: Line Chart between Breast Cancer Status and Cancer Type detail

We've created a line chart that illustrates the relationship between breast cancer status and cellularity.

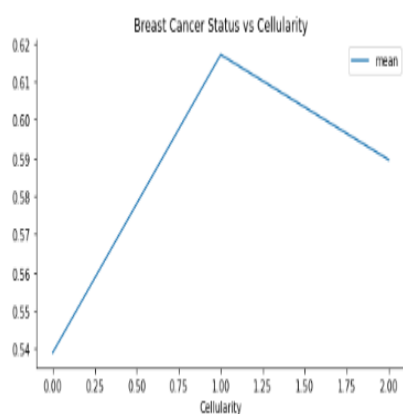


Figure 25: Line Chart between Breast Cancer Status and Cellularity

We've created a line chart that illustrates the relationship between breast cancer status and chemotherapy.

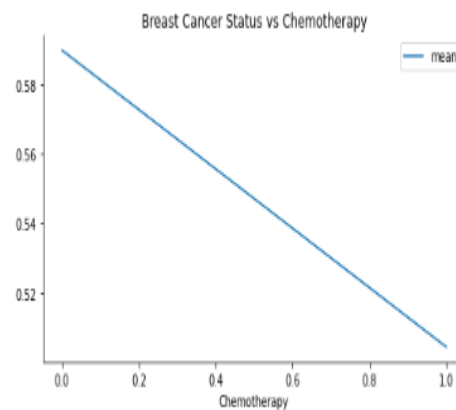


Figure 26: Line Chart between Breast Cancer Status and Chemotherapy

We've created a line chart that illustrates the relationship between breast cancer status and pam50 and claudin-low subtype.

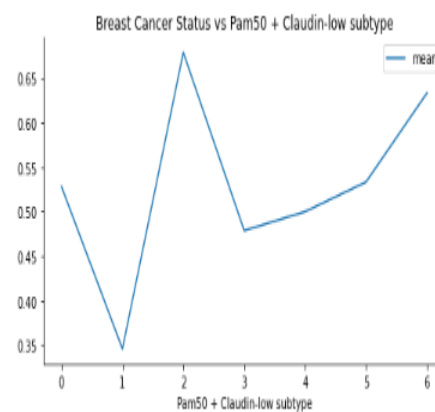


Figure 27: Line Chart between Breast Cancer Status and Pam 50

We've created a line chart that illustrates the relationship between breast cancer status and cohort.

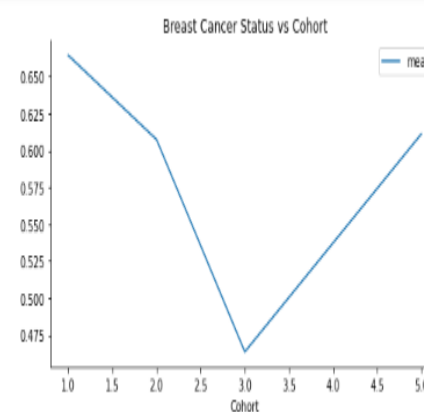


Figure 28: Line Chart between Breast Cancer Status and Cohort

We've created a line chart that illustrates the relationship between breast cancer status and neoplasm histologic grade.

We've created a line chart that illustrates the relationship between breast cancer status and neoplasm histologic grade.

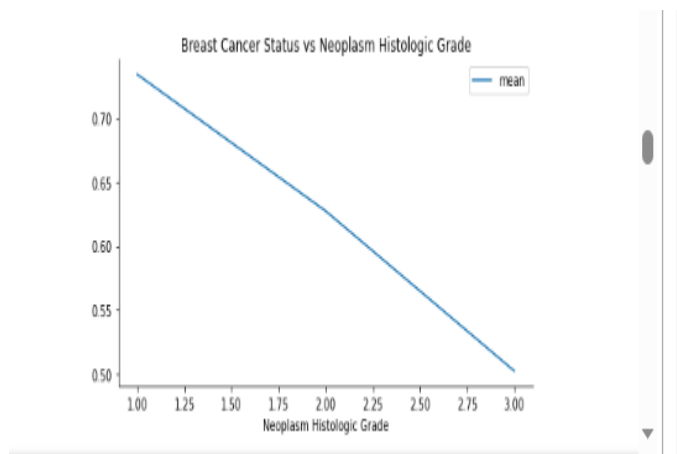


Figure 29: Line Chart between Breast Cancer Status and Neoplasm Grade

We've created a line chart that illustrates the relationship between breast cancer status and tumor another histologic subtype.

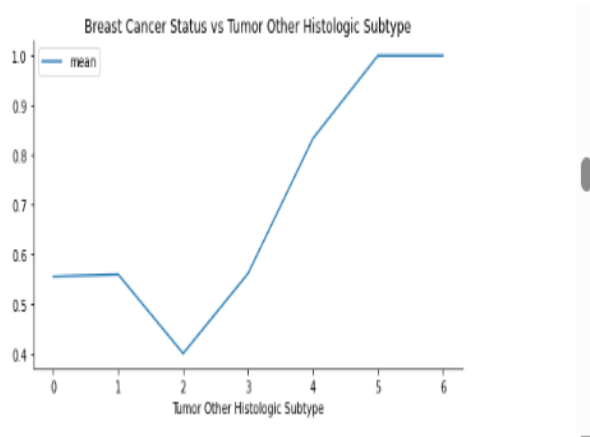


Figure 30: Line Chart between Breast Cancer Status and Tumor

We've created a line chart that illustrates the relationship between breast cancer status and hormone therapy.

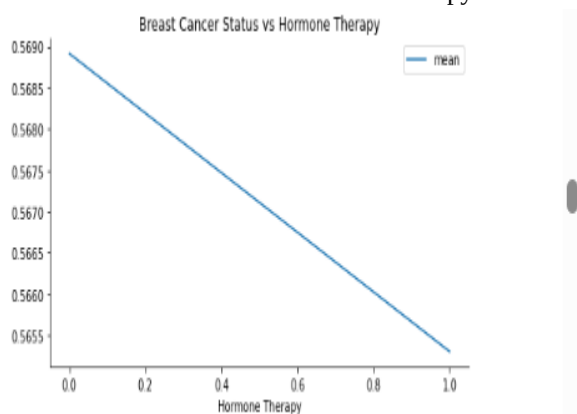


Figure 31: Line Chart between Breast Cancer Status and Hormone Therapy

We've created a line chart that illustrates the relationship between breast cancer status and integrative cluster.

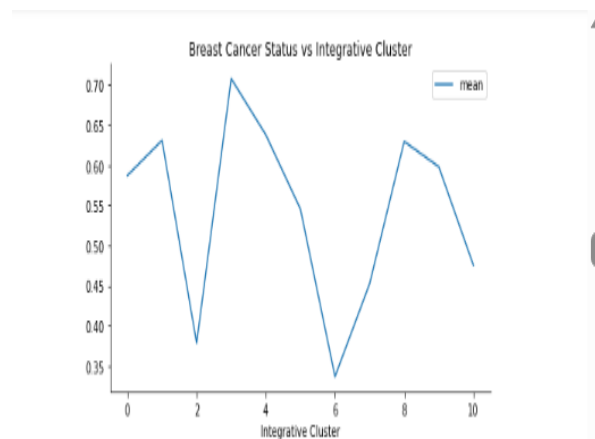


Figure 32: Line Chart between Breast Cancer Status and Integrative Cluster

We've created a line chart that illustrates the relationship between breast cancer status and primary tumor laterality.

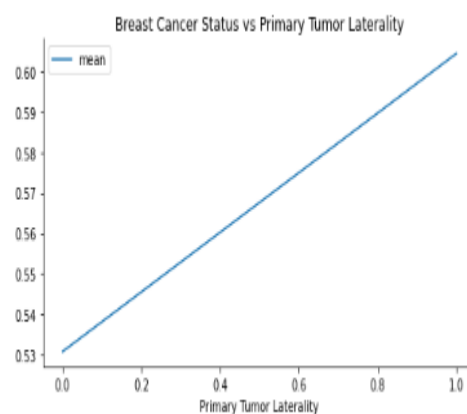


Figure 33: Line Chart between Breast Cancer Status and Primary Tumor

We've created a line chart that illustrates the relationship between breast cancer status and lymph nodes examined positive.

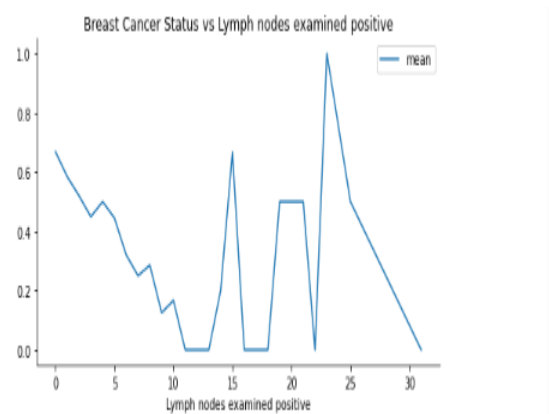


Figure 34: Line Chart between Breast Cancer Status and Lymph Nodes

We've created a line chart that illustrates the relationship between breast cancer status and mutation count.

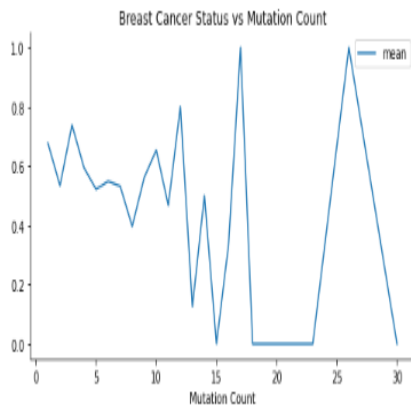


Figure 35: Line Chart between Breast Cancer Status and Mutation Count

We've created a line chart that illustrates the relationship between breast cancer status and Nottingham prognostic index.

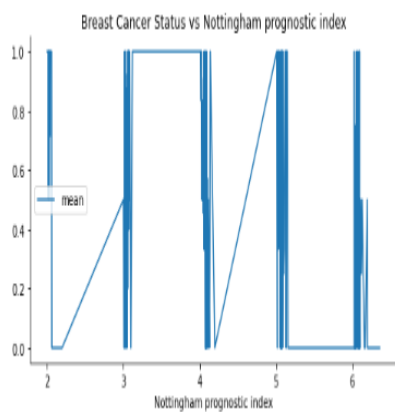


Figure 36: Line Chart between Breast Cancer Status and Prognostic index

We've created a line chart that illustrates the relationship between breast cancer status and oncotree code.

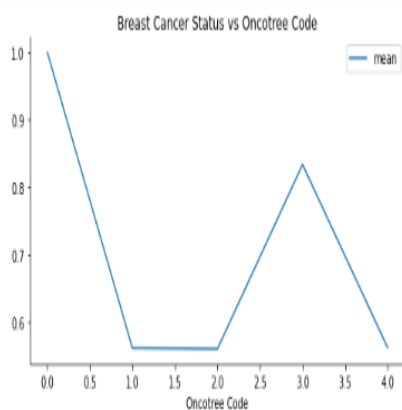


Figure 37: Line Chart between Breast Cancer Status and Oncotree Code

We've created a line chart that illustrates the relationship between breast cancer status and overall survival months.

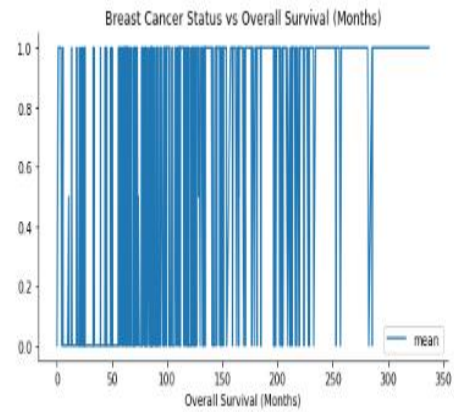


Figure 38: Line Chart between Breast Cancer Status and Overall Survival Months

We've created a line chart that illustrates the relationship between breast cancer status and overall survival status.

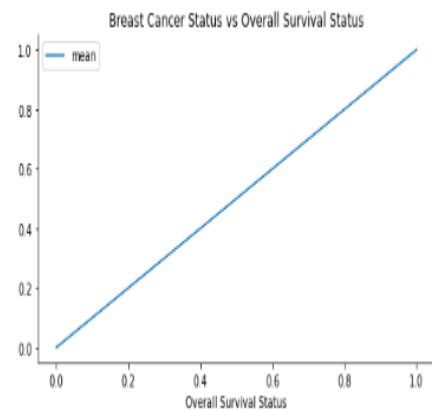


Figure 39: Line Chart between Breast Cancer Status and Overall Survival Status

We've created a line chart that illustrates the relationship between breast cancer status and radiotherapy.

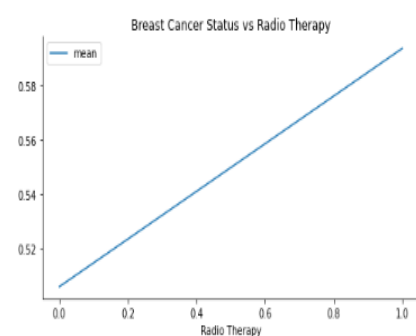


Figure 40: Line Chart between Breast Cancer Status and Radio Therapy

We've created a line chart that illustrates the relationship between breast cancer status and tumor size.

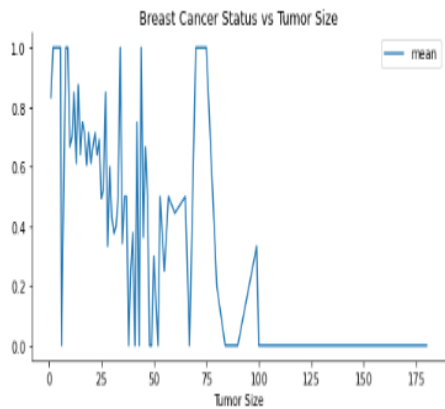


Figure 41: Line Chart between Breast Cancer Status and Tumor Size

We've created a line chart that illustrates the relationship between breast cancer status and tumor stage.

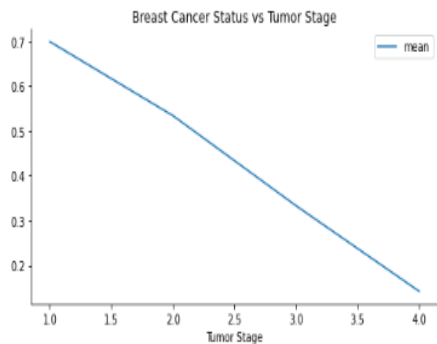


Figure 42: Line Chart between Breast Cancer Status and Tumor Stage

III. TECHNIQUES USED

After performing the above steps, we have split the data into 2 parts that is training and test data. Standardization is also performed to standardize the data to improve the performance of machine learning models. We have selected only some independent variables such as 'Age at Diagnosis', 'Neoplasm Histologic Grade', 'Lymph nodes examined positive', 'Tumor Size', 'Tumor Stage', 'Chemotherapy', 'Radio Therapy', 'Type of Breast Surgery', 'Hormone Therapy' and patient vital status is taking as dependent variable.

A. Logistic Regression:

Logistic Regression is a supervised learning algorithm that works on binary classification where the dependent variable has 2 outcomes only. It finds the interpretable coefficient which shows the impact of the features. It can handle large datasets very easily and efficiently. It uses the 2 types of regularization techniques: lasso and ridge.

Logistic Regression

```
In [210]: from sklearn.linear_model import LogisticRegression
log = LogisticRegression(random_state=16)
log.fit(x_train, y_train)
y_pred = log.predict(x_test)

In [211]: from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, y_pred)
acc = (accuracy_score(y_test, y_pred))*100
print("Accuracy Score:-", acc)
print ("Confusion Matrix : \n", cm)

Accuracy Score:- 72.09302325581395
Confusion Matrix :
[[20 16]
 [ 8 42]]

In [212]: from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.71	0.56	0.63	36
1	0.72	0.84	0.78	50
accuracy			0.72	86
macro avg	0.72	0.70	0.70	86
weighted avg	0.72	0.72	0.71	86

Figure 43: Logistic Regression

B. XG Boost:

It is an ensemble learning method that combines 2 things that is multiple decision trees and gradient boosting whose aim is to combine weak learners and find a strong learner. It is highly robust in nature and tends to compute nonlinear relationships as well. It uses 2 regularization techniques, shrinkage, and pruning which increase the learning rate and help in reducing the overfitting respectively.

```
In [213]: from xgboost import XGBClassifier
xgb = XGBClassifier()
xgb.fit(x_train, y_train)
y_pred2 = xgb.predict(x_test)

C:\Users\SR\anaconda3\lib\site-packages\xgboost\sklearn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use_label_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num_class - 1].
warnings.warn(label_encoder_deprecation_msg, UserWarning)

[19:07:46] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release.1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old behavior.

In [214]: from sklearn.metrics import confusion_matrix, accuracy_score
cm2 = confusion_matrix(y_test, y_pred2)
acc2 = (accuracy_score(y_test, y_pred2))*100
print("Accuracy Score:-", acc2)
print ("Confusion Matrix : \n", cm2)

Accuracy Score:- 68.6046511627907
Confusion Matrix :
[[22 14]
 [13 37]]

In [215]: from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred2))
```

	precision	recall	f1-score	support
0	0.63	0.61	0.62	36
1	0.73	0.74	0.73	50
accuracy			0.69	86
macro avg	0.68	0.68	0.68	86
weighted avg	0.68	0.69	0.69	86

Figure 44: XG Boost

C. Naïve Bayes:

This algorithm is based on the Bayes algorithm which performs the assumption on feature importance. In this method, all features are conditionally independent which is sometimes dangerous but can simplify the model. It is a most efficient algorithm. It is efficient for noisy data because it ignores the dependencies between the features.

Naive Bayes

```
In [220]: from sklearn.naive_bayes import GaussianNB
          gnb = GaussianNB()
          gnb.fit(x_train, y_train)
          y_pred_NB = gnb.predict(x_test)
          accuracy_score(y_test, y_pred_NB)
```

```
Out[220]: 0.686046511627907
```

```
In [221]: from sklearn.metrics import classification_report
          print(classification_report(y_test, y_pred_NB))
```

	precision	recall	f1-score	support
0	0.71	0.42	0.53	36
1	0.68	0.88	0.77	50
accuracy			0.69	86
macro avg	0.70	0.65	0.65	86
weighted avg	0.69	0.69	0.67	86

Figure 45: Naive Bayes

IV. CONCLUSION

The use of artificial intelligence models plays an important role in predicting the survival rate of breast cancer. Big data datasets can be collected from anywhere. In this research work, we understand that breast cancer is increasing day by day and most of the time, it is found in women. We have collected the data from Kaggle which contains information on breast cancer for women only. We have implemented the exploratory data analysis and understand the feature variation and how it is upgrading and downgrading. We have implemented 3 machine learning models such as logistic regression, XG Boost, and Naïve Bayes. The accuracy score of logistic regression is 72%. The accuracy score of the XG boost is 68%. The accuracy score of Naïve Bayes is 68%.

V. FUTURE SCOPE

The future scope of breast cancer prediction is to merge the various risk factors to model the techniques that help to diagnose breast cancer at an early stage because, in most developing countries, treatment for most diseases is still not found on time so for this, we can use the deep learning and neural network methods to make the disease management effective.

REFERENCES

1. Wu, J. and Hicks, C., 2021. Breast cancer type classification using machine learning. *Journal of personalized medicine*, 11(2), p.61.

2. Ak, M.F., 2020, April. A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. In *Healthcare* (Vol. 8, No. 2, p. 111). MDPI.

3. Islam, M.M., Haque, M.R., Iqbal, H., Hasan, M.M., Hasan, M. and Kabir, M.N., 2020. Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, 1, pp.1-14.

4. Jasti, V.D.P., Zamani, A.S., Arumugam, K., Naved, M., Pallathadka, H., Sammy, F., Raghuvanshi, A. and Kaliyaperumal, K., 2022. Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis. *Security and communication networks*, 2022, pp.1-7.

5. Fatima, N., Liu, L., Hong, S. and Ahmed, H., 2020. Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, 8, pp.150360-150376.

6. Mohammed, S.A., Darrab, S., Noaman, S.A. and Saake, G., 2020. Analysis of breast cancer detection using different machine learning techniques. In *Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5* (pp. 108-117). Springer Singapore.

7. Rawal, R., 2020. Breast cancer prediction using machine learning. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 13(24), p.7.

8. Amethiya, Y., Pipariya, P., Patel, S. and Shah, M., 2022. Comparative analysis of breast cancer detection using machine learning and biosensors. *Intelligent Medicine*, 2(2), pp.69-81.

9. Rabiei, R., Ayyoubzadeh, S.M., Sohrabei, S., Esmaeili, M. and Atashi, A., 2022. Prediction of breast cancer using machine learning approaches. *Journal of biomedical physics & engineering*, 12(3), p.297.

10. Shamrat, F.J.M., Raihan, M.A., Rahman, A.S., Mahmud, I. and Akter, R., 2020. An analysis on breast disease prediction using machine learning approaches. *International Journal of Scientific & Technology Research*, 9(02), pp.2450-2455.