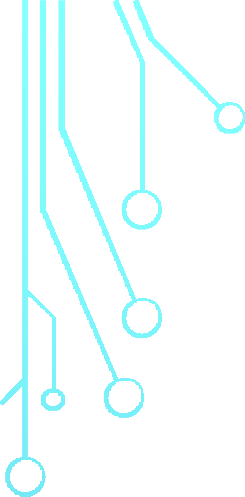


BIA| BOSTON INSTITUTE OF ANALYTICS

BREAST CANCER RISK PREDICTION



.



AGENDA:

* **Objective:**

**.**

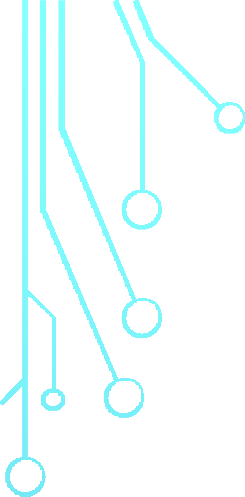
* **Data Understanding:**
* **Data Preprocessing**
* **Exploratory Data Analysis (EDA):**
* **Model Building & Evaluation**
* **Visualization of Results:**



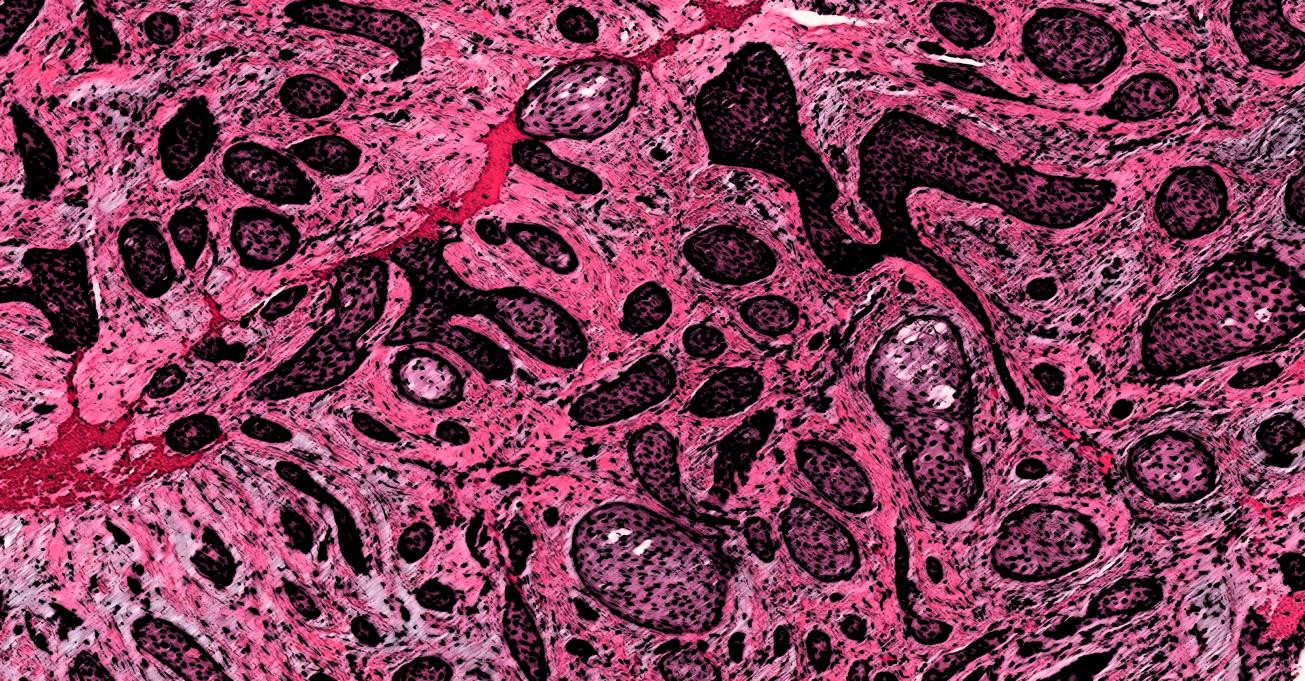
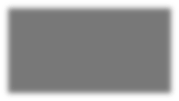
* **Insights & Recommendations:**





**"SURVIVAL AND RISK PREDICTION IN BREAST CANCER: A DATA-DRIVEN APPROACH WITH METABRIC"**









# *"*BREAST CANCER PROGNOSIS PREDICTION USING METABRIC CLINICAL AND GENOMIC DATA*"*

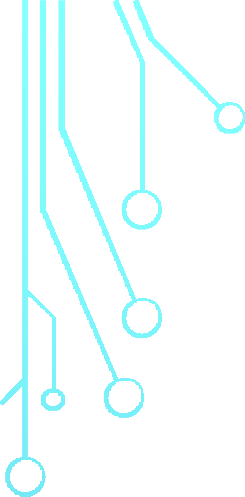
* + **Bullet (Main Points)**
  + **Bold or dark font for key agenda items.**
  + **Example:**
    - **Objective: Predict survival and relapse using patient clinical and genomic data.**
  + **Bullet (Details)**
  + **Slightly smaller font, lighter shade.**



* + **Example:**
    - **Includes factors such as age, tumor stage, ER/PR/HER2 status, and treatment history.**





INTRODUCTION:

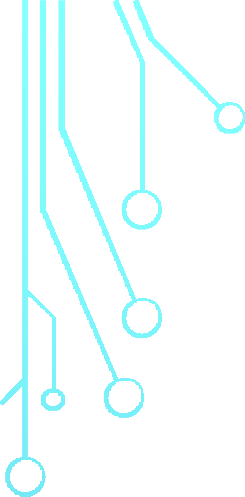
* **Breast cancer is one of the most prevalent cancers worldwide and a leading cause of cancer-related deaths among women. Early detection and accurate prognosis are critical for improving survival rates and guiding effective treatment strategies.**
* **The METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset provides a rich collection of clinical, pathological, and genomic information from nearly 2,000 breast cancer patients.**
* **Key Highlights:**
* **Combination of clinical and molecular data.**



* **Survival analysis with statistical and machine learning approaches.**
* **Insightful visualizations for better understanding of breast cancer patterns.**





OBJECTIVE 🎯

The main objective of this project is to **analyze and predict breast cancer prognosis**

using the METABRIC dataset by combining clinical and genomic information.

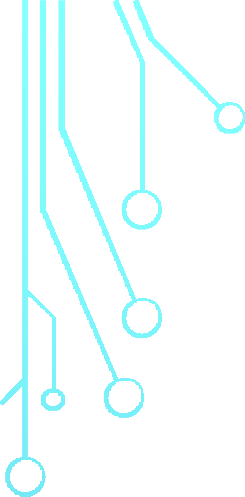
* **Specific Goals:**
* **Understand** patient demographics, tumor characteristics, and treatment patterns.
* **Identify** key features influencing survival and relapse.
* **Build** predictive models for survival status using machine learning techniques.



* **Visualize** patterns and correlations in the dataset for better interpretability





**DATA OVERVIEW**

**The METABRIC dataset (Molecular Taxonomy of Breast Cancer International Consortium) contains detailed information for nearly 2,000 breast cancer patients, including:**

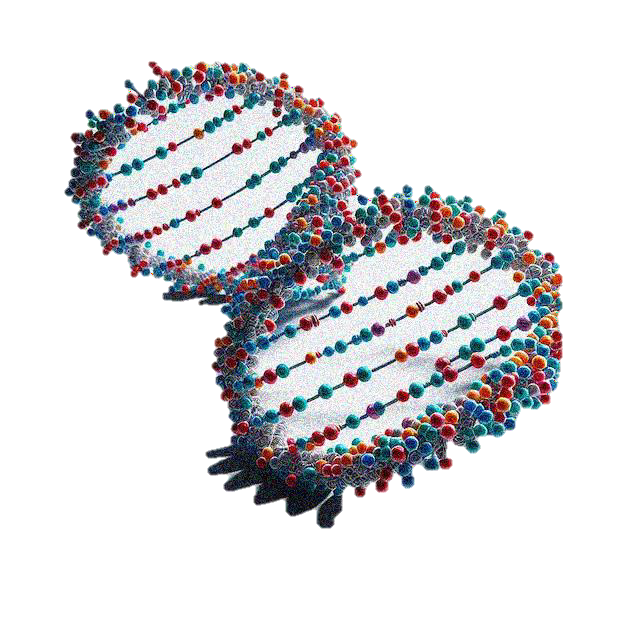
**Clinical Features:**

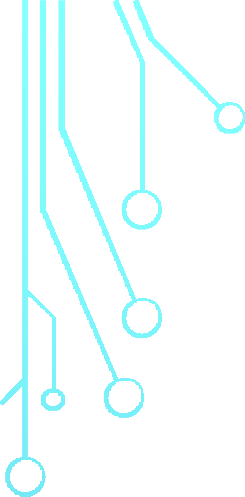
* **Age at Diagnosis & Tumor Size & Stage**
* **Lymph Nodes Examined Positive & Histologic Grade**



* **Hormone Receptor Status (ER, PR, HER2)**
* **Treatment Details (Chemotherapy, Radiotherapy, Hormone Therapy)**
* **Menopausal State**





**GENOMIC FEATURES**

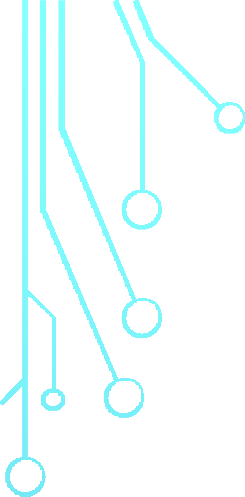
* **Mutation Count**
* **Integrative Cluster Classification**
* **3-Gene Classifier Subtype**
* **PAM50 + Claudin-low Subtype**
* **Outcome Variables:**
* **Overall Survival (Months)**
* **Overall Survival Status (Alive/Deceased)**
* **Relapse Free Status (Months)**
* **Relapse Free Status (Yes/No)**



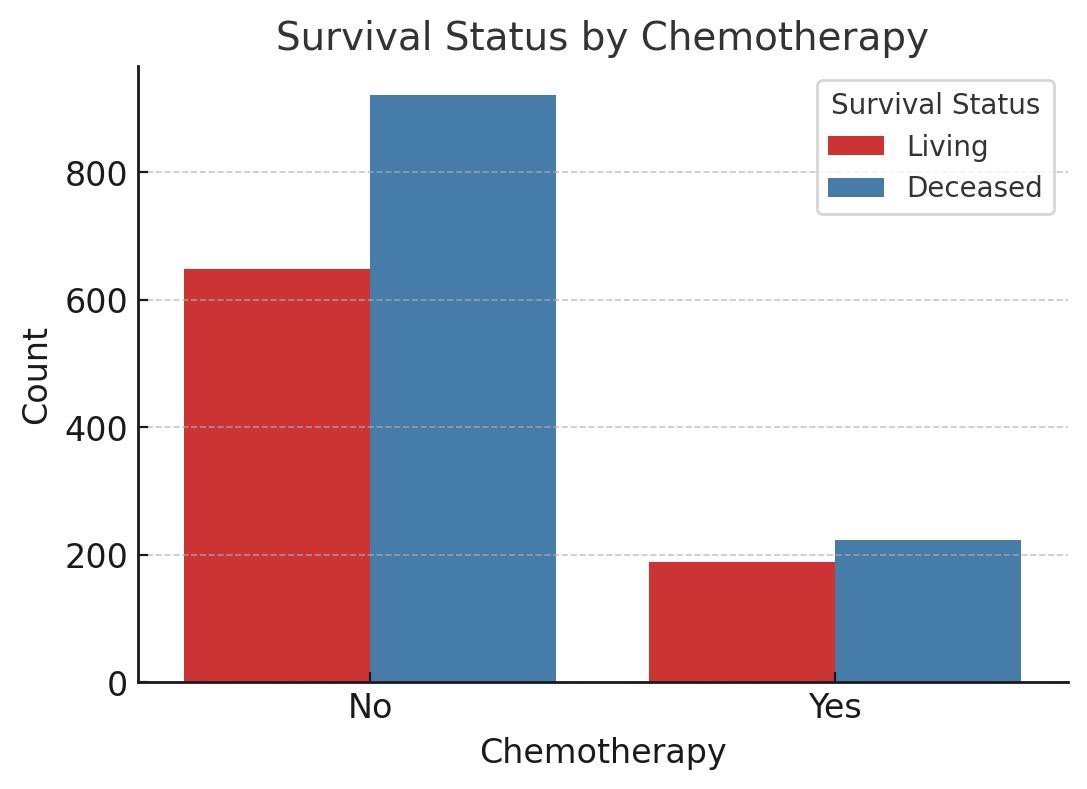
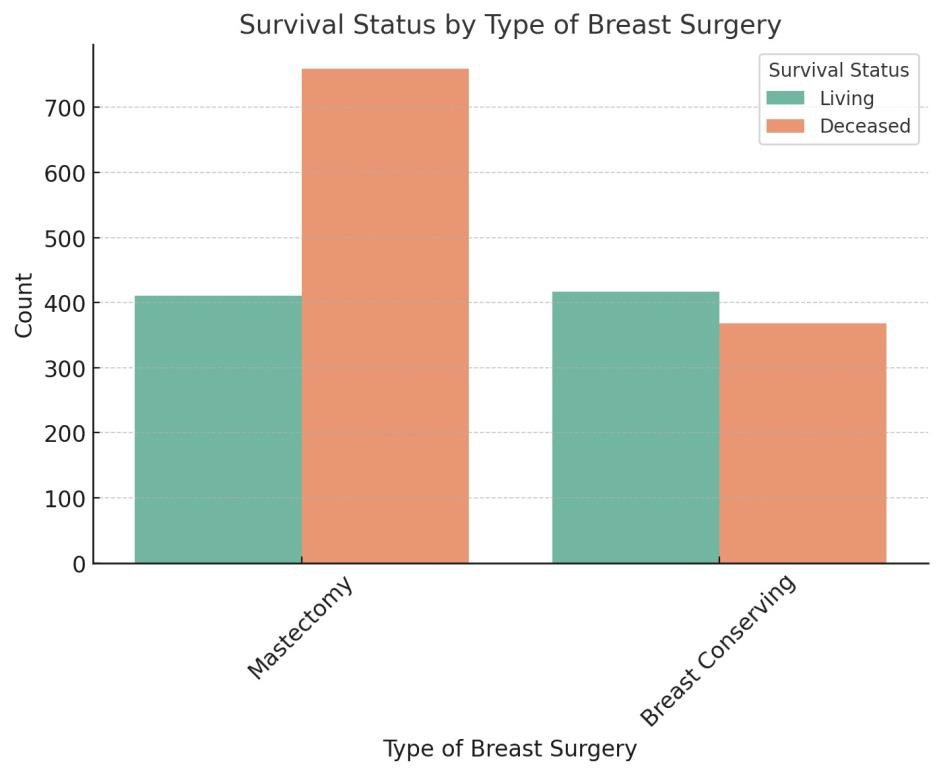
* **Dataset Size:**
* **Patients: ~1,980**
* **Features: 30+**
* **Target Variables: Survival & Relapse outcomes**





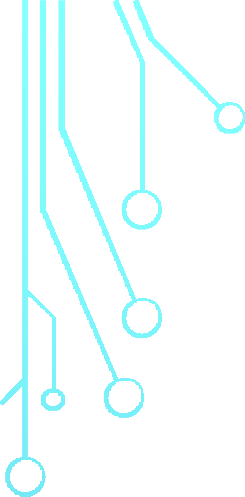
**THESE TWO BAR CHARTS ARE COMPARING OVERALL SURVIVAL STATUS WITH:**

1. **TYPE OF BREAST SURGERY**
2. **CHEMOTHERAPY**.







EXPLORATORY DATA ANALYSIS (EDA)

* 1. **DATA OVERVIEW**
     + **The dataset from the METABRIC study contains rich patient-level breast cancer**

information, combining clinical, pathological, and genetic features.

* + - **Structure: A mix of categorical variables (e.g., ER Status, Tumor Stage, Type of Surgery) and numerical variables (e.g., Age at Diagnosis, Tumor Size, Mutation Count).**
    - **Cleaning: Missing values are addressed using imputation techniques. Identifiers**

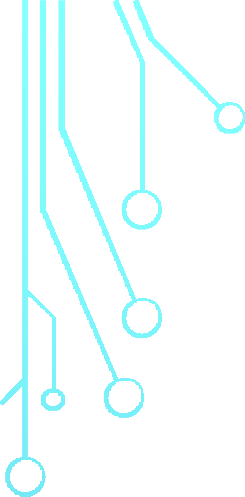


like *Patient ID* are excluded from modeling to avoid data leakage.

* + - **Goal: Understand the data distribution, detect anomalies, and uncover patterns linked to survival outcomes.**





* 1. **UNIVARIATE ANALYSIS**
* **Numerical Features**
* **Histograms & Boxplots** for: Age at Diagnosis, Tumor Size, Lymph Nodes Examined Positive, Mutation Count, Overall Survival (Months), Relapse Free Status (Months).
* Insights into skewness, spread, and presence of outliers.
* **Categorical Features**
* **Bar Charts** showing frequency for: Type of Breast Surgery, ER Status,



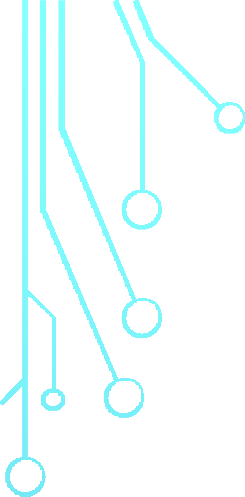
Chemotherapy, Hormone Therapy, Integrative Cluster, Tumor Stage.

* Highlights class imbalance and dominant categories.





# BIVARIATE ANALYSIS



•

**Grouped Bar Charts** to compare *Overall Survival Status* with: Type of Breast Surgery, ER Status, Chemotherapy, Tumor Stage.

* **Boxplots** comparing survival months across survival status groups to visualize outcome disparities.
* **Correlation Heatmap** for numerical features to

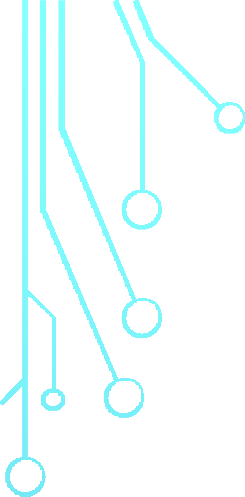
spot strong associations and potential multicollinearity.



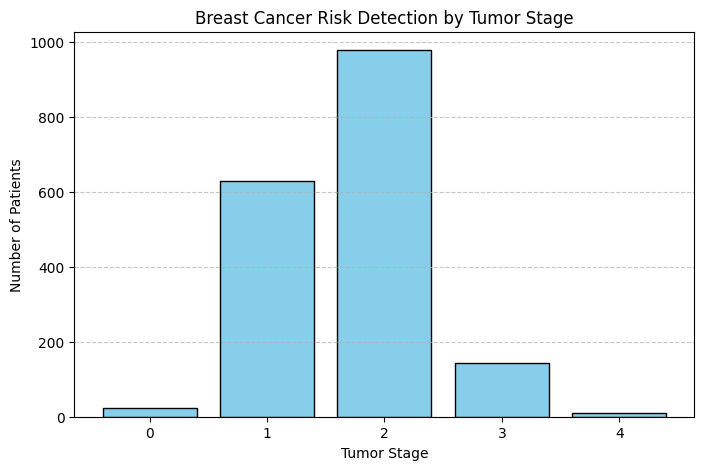






COUNT PATIENTS BY TUMOR STAGE (DROPPING MISSING VALUES)

Patient Count by Tumor Stage

**Purpose**

Tumor stage describes the extent to which cancer has spread in the body at the time of diagnosis. It is one of the most important prognostic factors in breast cancer, guiding treatment plans and helping predict patient outcomes.

By counting the number of patients in each tumor stage, we can:

Understand the stage distribution within our dataset.



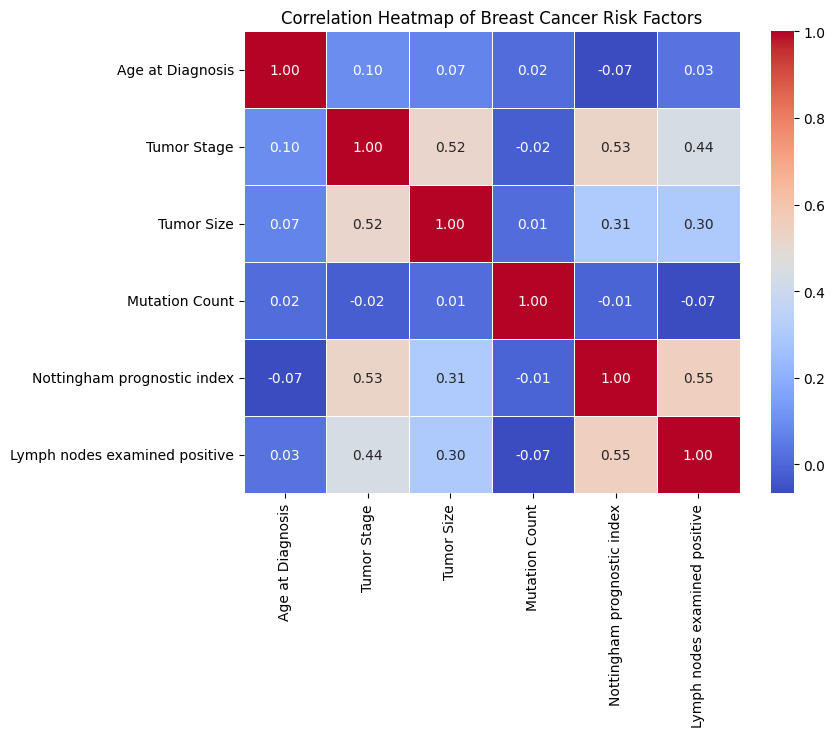
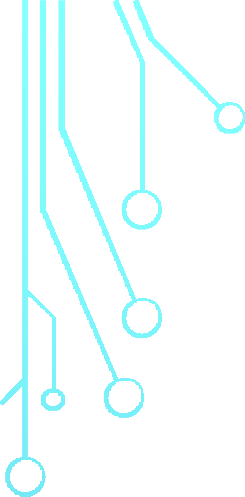
Identify whether the data is skewed toward early or

late-stage cancers.

Detect possible data imbalances that may affect model training.





CORRELATION HEATMAP OF BREAST CANCER RISK FACTORS

**Purpose**

A correlation heatmap helps identify relationships between numerical features in the dataset.

In the context of breast cancer risk factors, it

can reveal:

Which clinical or genetic factors are most strongly related to survival, relapse, or tumor size.

Multicollinearity between features that may



affect model performance.

Potential predictive variables for building classification models.





# MODEL PREDICTIONS:

|  |  |
| --- | --- |
| **Predicted Class** | **Probability of Deceased** |
| 1 | 1.00 |
| 1 | 1.00 |
| 0 | 0.01 |
| 1 | 0.99 |
| 1 | 1.00 |
| 1 | 1.00 |
| 1 | 1.00 |
| 1 | 0.96 |
| 1 | 0.99 |
| 0 | 0.01 |

**Predicted Class** —

* **1 likely means "Deceased"**
* **0 likely means "Alive“.**

**Probability of Deceased---**

*  **Values close to 1.00 means the model is very confident that the patient has deceased…**

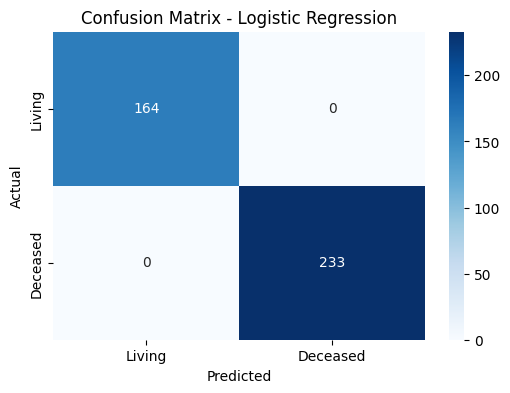
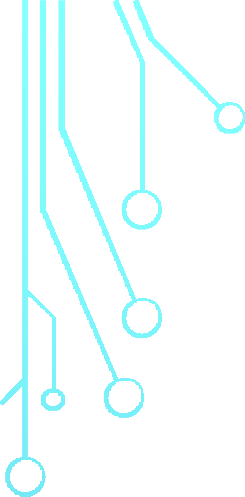


* **Values close to 0.00 means the model us very confident that the patient is alive….**







CONFUSION MATRIX LOGISTIC REGRESSIOIN

The image presents a **confusion matrix** for a **Logistic Regression** model predicting patient survival status ("Living" or "Deceased").

The confusion matrix is a 2×2 table showing

the model’s classification performance.

**Interpretation:**

**Rows:** Represent the **actual labels** (ground

truth).

**Columns:** Represent the **predicted labels**



from the model.

|  |  |  |
| --- | --- | --- |
|  | **Predicted : Living** | **Predicted : Deceased** |
| Actual Living | 164 (True Positives) | 0 (False Negative) |
| Actual Deceased | 0 (False Positive) | 233 (True Negative) |





# MODEL PERFORMANCE – BREAST CANCER RISK PREDICTION

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Interpretation** |
| **Accuracy** | XX% | Percentage of total correct predictions out of all test cases. |
| **Precision (Deceased)** | XX% | Of all patients predicted as deceased, how many were actually deceased. |
| **Recall (Deceased)** | XX% | Of all actual deceased patients, how many were correctly predicted. |
| **F1 Score (Deceased)** | XX | Harmonic mean of precision and recall — balances both metrics. |

Overview:

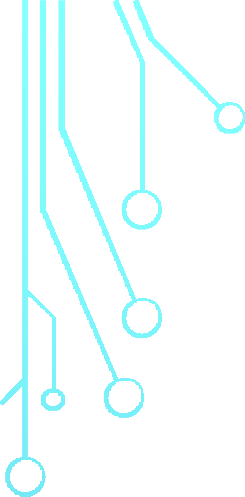
The trained **Logistic Regression** model was evaluated on unseen test data (20% split). The model predicts the probability of patient survival status ("Alive" or "Deceased") based on clinical and genomic features.

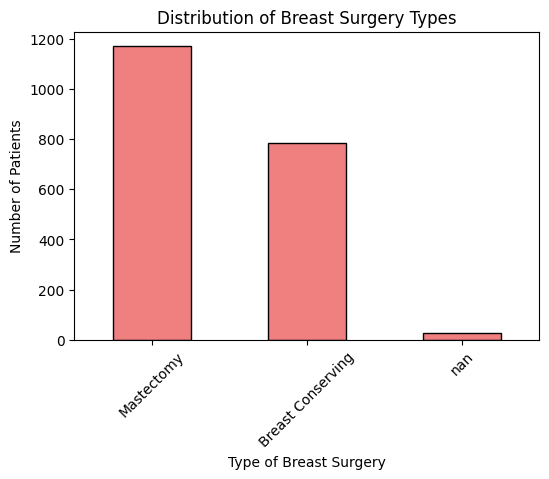


|  |  |  |
| --- | --- | --- |
|  | **Predicted Deceased** | **Predicted Alive** |
| **Actual Deceased** | TP = XX | FN = XX |
| **Actual Alive** | FP = XX | TN = XX |



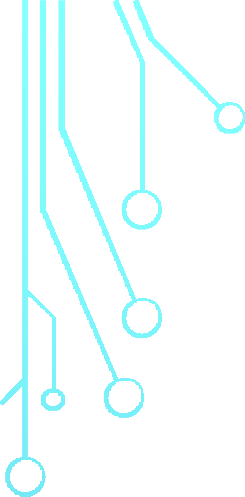


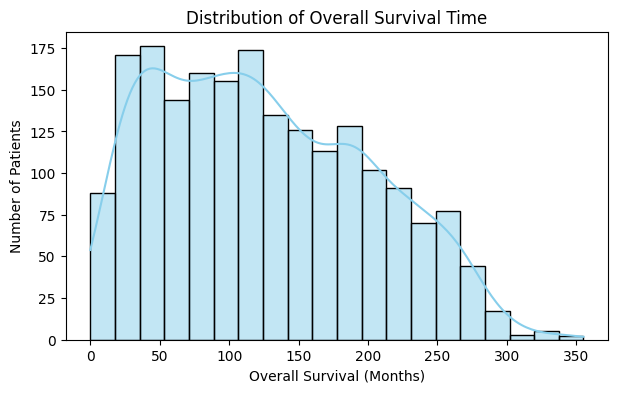
TYPES OF BREAST SURGERY





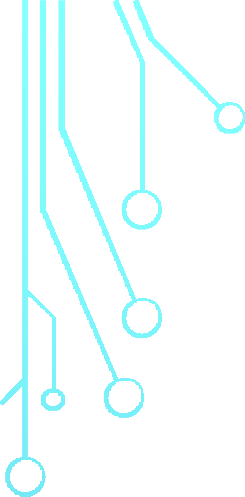


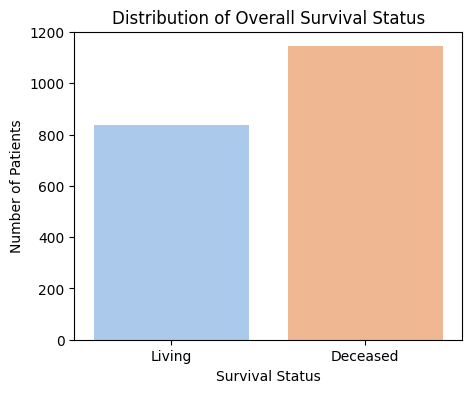
OVERALL SURVIVAL TIME OF A PATIENT IN MONTHS





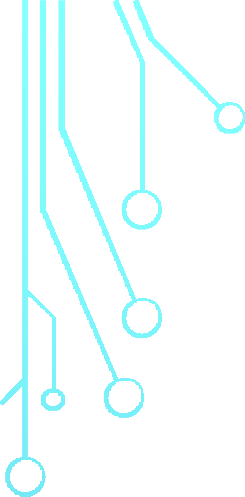


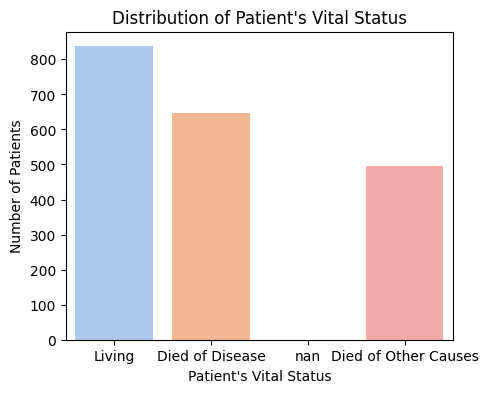
SURVIVAL STATUS – DECEASED OR ALIVE





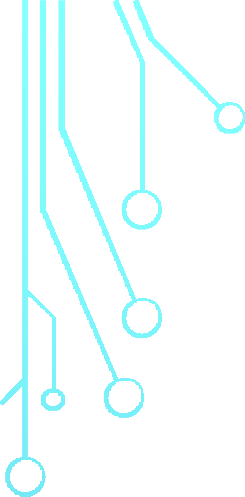


LAST FOLLOW UP – THE PATIENT IS DECEASED OR ALIVE







CONCLUSION

* The study of Breast Cancer Risk Prediction is one of the important phenomena to bring mitigation of mental pains of people who suffer from this fatal disease.
* Risk prediction may or may not be positive, still some expectation would rise in one’s

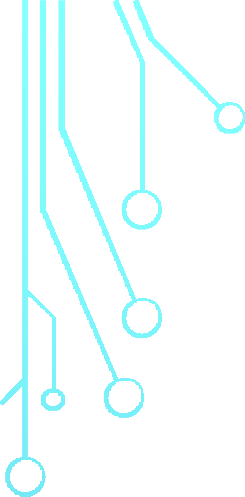
mind towards positive direction unless it brings the truth.

* I have shown in my study the different phases of prediction which is definitely helpful.



* To conclude, it may be pertinent to say that Early Prediction of Breast Cancer may

delay the final conclusion of human’s life.





Thank You

