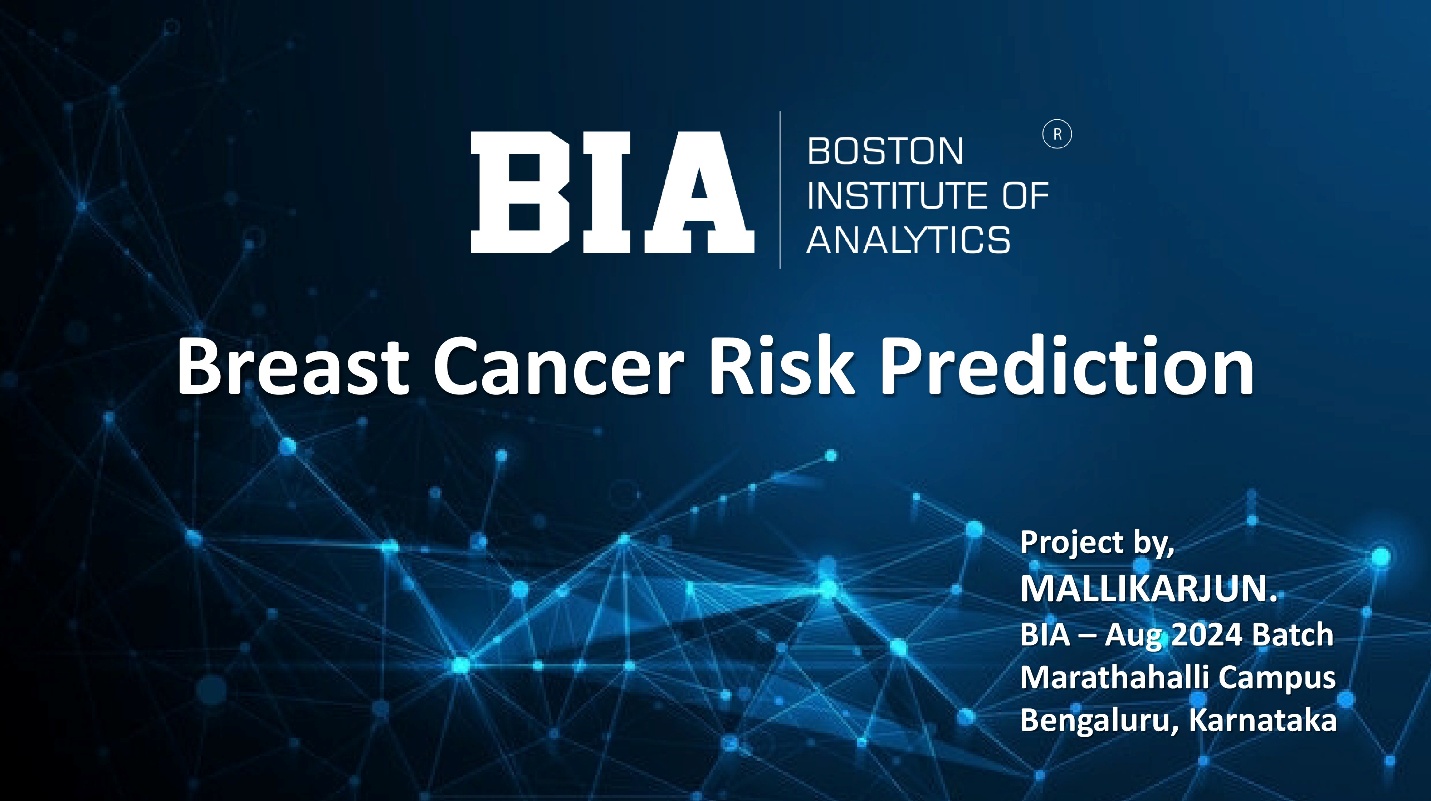
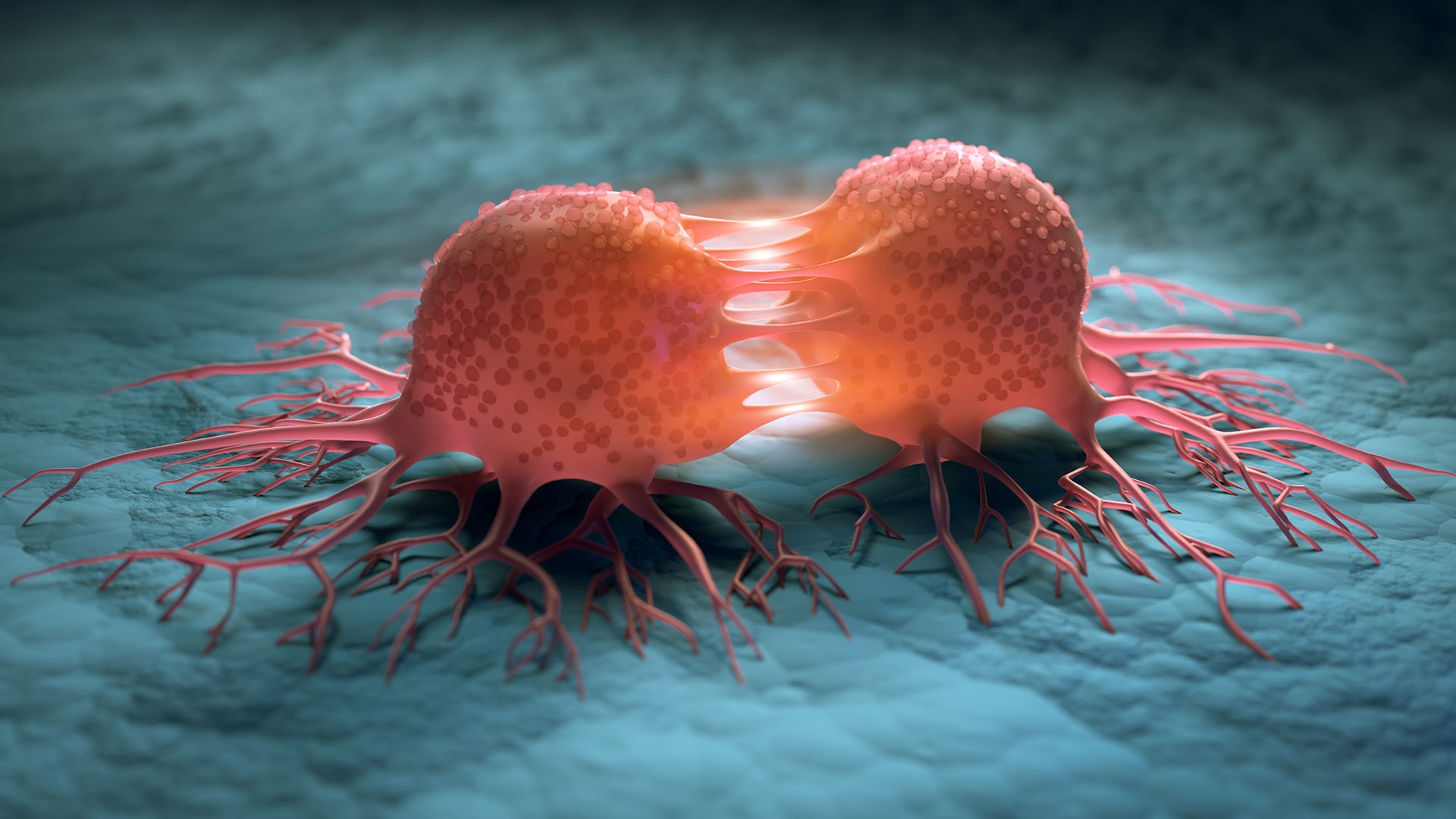
Capstone Project Final Report





**Problem:**

This project aims to deepen the understanding of breast cancer survival trends and enhance prediction models for patient outcomes. By leveraging advanced statistical techniques and machine learning, the project seeks to analyze survival rates and accurately predict the 10-year mortality risk for breast cancer patients. The ultimate objective is to improve treatment planning, patient counseling, and overall cancer care strategies.

**Objectives:**

**1. Understand Breast Cancer Patient Data**:

* Collect and explore patient-related attributes, including demographic, clinical, and survival information.
* Identify key patterns and trends that may impact survival outcomes.

**2. Perform Data Preprocessing and Feature Engineering**:

* Handle missing values using appropriate imputation techniques.
* Remove irrelevant features that do not contribute to the predictive analysis.
* Encode categorical variables for compatibility with machine learning models.
* Scale numerical features to ensure balanced model training.

**3. Build and Evaluate Predictive Models**:

* Develop machine learning models such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) to classify patient survival outcomes.
* Assess model performance using recall, precision, F1-score, accuracy, and AUC-ROC curves to ensure reliability.

**4. Conduct Survival Analysis**:

* Utilize Kaplan-Meier Survival Curves to estimate survival probabilities over time.
* Apply Cox Proportional Hazards Model to determine significant factors affecting survival.

**5. Derive Insights for Medical Decision-Making**:

* Interpret model results to identify factors influencing patient survival.
* Ensure the models prioritize high recall to minimize the risk of missing critical cases.
* Provide data-driven insights that could assist in breast cancer prognosis and patient care.

**Table of Contents:**

* [Step 1 | Import Libraries](#import)
* [Step 2 | Read Dataset](#read)
* [Step 3 | Exploratory Data Analysis (EDA)](#eda)
* [Step 4 | Data Preprocessing](#preprocessing)
* [Step 5 | Data Splitting](#splitting)
* [Step 6 | Logistic Regression Model Building](#logistic)
* [Step 7 | Decision Tree Model Building](#dt)
* [Step 8 | Random Forest Model Building](#rf)
* [Step 9 | Support Vector Machine (SVM) Model Building](#svm)
* [Step 10 | Model Interpretation and Visualization](#interpretation)

[Kaplan-Meier survival curves](#kmf)

[Cox Proportional Hazards Models](#cphm)

* [Step 11 | Conclusion](#conclusion)

# Step 1 | Import Libraries

To ensure efficient data handling, preprocessing, visualization, model training, and evaluation, several essential Python libraries were imported in this project:

* **pandas & numpy** – For data loading, manipulation, and numerical computations.
* **matplotlib & seaborn** – For creating insightful visualizations to explore data distributions, relationships, and trends.
* **scikit-learn** – For implementing machine learning algorithms, preprocessing techniques, and model evaluation metrics.
* **joblib** – For saving and loading trained models, ensuring easy reproducibility and deployment.

These libraries provide the necessary tools to perform data-driven analysis, build predictive models, and derive meaningful insights from the dataset.

# Step 2 | Read Dataset

The dataset used in this project provides comprehensive clinical and genetic information on breast cancer patients. It includes crucial attributes such as **patient demographics, tumor characteristics, treatment details, genetic markers, and survival outcomes**.

To ensure data quality and reliability, an initial assessment was performed to:

* **Check for missing values and duplicate records** that could impact analysis.
* **Identify inconsistencies and outliers** in numerical and categorical features.
* **Understand feature distributions** to guide preprocessing and feature engineering.

This step ensures that the dataset is **clean, structured, and ready for further analysis**, forming a solid foundation for predictive modeling and survival analysis.

**Dataset Description:**

| **Variable** | **Description** |
| --- | --- |
| **Patient ID** | Unique identifier for each patient. |
| **Age at Diagnosis** | Age of the patient when diagnosed with cancer. |
| **Type of Breast Surgery** | The type of surgery performed on the breast: 0: Mastectomy 1: Lumpectomy |
| **Cancer Type** | General classification of the cancer type: 0: Breast Cancer 1: Breast Sarcoma |
| **Cancer Type Detailed** | More specific classification of the cancer type: 0: Breast Invasive Ductal Carcinoma 1: Breast Mixed Ductal and Lobular Carcinoma 2: Breast Invasive Lobular Carcinoma 3: Invasive Breast Carcinoma 4: Breast Invasive Mixed Mucinous Carcinoma 5: Breast 6: Breast Angiosarcoma 7: Metaplastic Breast Cancer |
| **Cellularity** | The tumor's cellularity degree is often used in pathology to describe the proportion of cells versus other components in a tissue sample: 0: High 1: Moderate 2: Low |
| **Chemotherapy** | Indicates whether the patient received chemotherapy: 0: No 1: Yes |
| **Pam50 + Claudin-low subtype** | Subtypes based on gene expression profiling: 0: LumA 1: LumB 2: Her2 3: claudin-low 4: Basal 5: Normal 6: NC |
| **Cohort** | The group or study cohort to which the patient belongs. |
| **ER status measured by IHC** | Estrogen receptor status as measured by Immunohistochemistry (IHC): 0: Positive 1: Negative |
| **ER Status** | Estrogen receptor status: 0: Positive 1: Negative |
| **Neoplasm Histologic Grade** | The histologic grade of the neoplasm indicates how much the tumor cells differ from normal cells. |
| **HER2 status measured by SNP6** | HER2 (human epidermal growth factor receptor 2) status measured by SNP (single nucleotide polymorphism) analysis: 0: Neutral 1: Gain 2: Loss 3: Undef |
| **HER2 Status** | HER2 receptor status: 0: Negative 1: Positive |
| **Tumor Other Histologic Subtype** | Other histologic subtypes of the tumor not covered by main classifications: 0: Ductal/NST 1: Mixed 2: Lobular 3: Medullary 4: Mucinous 5: Tubular/ cribriform 6: Other 7: Metaplastic |
| **Hormone Therapy** | Indicates whether the patient received hormone therapy: 0: Yes 1: No |
| **Inferred Menopausal State** | Menopausal state inferred based on age and clinical criteria: 0: Post 1: Pre |
| **Integrative Cluster** | Classification based on integrative clustering of genomic data. |
| **Primary Tumor Laterality** | The side of the body where the primary tumor is located: 0: Left 1: Right |
| **Lymph nodes examined positive** | Number of lymph nodes that tested positive for cancer. |
| **Mutation Count** | Total number of genetic mutations identified in the tumor. |
| **Nottingham prognostic index** | Prognostic score based on tumor size, lymph node status, and histologic grade. |
| **Oncotree Code** | A code that represents the type of cancer based on the OncoTree classification: 0: IDC 1: MDLC 2: ILC 3: BRCA 4: IMMC 5: BREAST 6: PBS 7: MBC |
| **Overall Survival (Months)** | The overall survival time of the patient in months. |
| **Overall Survival Status** | Indicates whether the patient is alive or deceased: 0: Deceased 1: Living |
| **PR Status** | Progesterone receptor status: 0: Positive 1: Negative |
| **Radio Therapy** | Indicates whether the patient received radiotherapy: 0: Yes 1: No |
| **Relapse Free Status (Months)** | Time in months the patient remained free from cancer relapse. |
| **Relapse Free Status** | Indicates whether the patient has had a cancer relapse: 0: Not Recurred 1: Recurred |
| **Sex** | The sex of the patient: Female |
| **3-Gene classifier subtype** | Subtypes based on the expression of three specific genes: 0: ER+/HER2- Low Prolif 1: ER+/HER2- High Prolif 2: ER-/HER2- 3: HER2+ |
| **Tumor Size** | Size of the primary tumor. |
| **Tumor Stage** | Stage of the tumor, indicating the extent of cancer spread. |
| **Patient's Vital Status** | Indicates whether the patient is alive or deceased at the last follow-up: 0: Living 1: Died of Disease 2: Died of Other Causes |

# Step 3 | Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset before applying machine learning models. It involves summarizing, visualizing, and detecting patterns, anomalies, and relationships between variables. The goal of EDA is to extract meaningful insights that help in better feature selection and model building.

In this project, EDA involved:

* **Checking data structure** to understand the number of records, features, and data types.
* **Identifying missing values and duplicate records** to assess data quality.
* **Statistical analysis** to summarize key numerical features.
* **Correlation analysis** to explore relationships between variables.
* **Visualizations** such as histograms, box plots, and heatmaps to detect trends and outliers.

**Inferences from EDA:**

✔ **Number of Entries:** The dataset contains **2,509 records**, indexed from 0 to 2,508.  
✔ **Columns:** There are **34 features** representing various patient attributes and test results.  
✔ **Data Types:**

* The majority of the columns (**24 out of 34**) are of **float64** type.
* The remaining **10 columns** are of **object (categorical)** type.

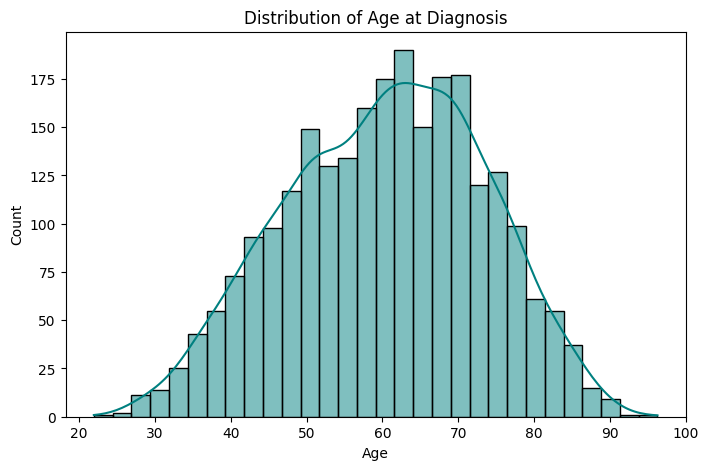
✔ **Missing Values:** Some columns have missing values, but most have a high number of non-null entries.

EDA helped in understanding the dataset’s structure, identifying key variables for prediction, and preparing the data for further preprocessing.

**Key Visualizations:**

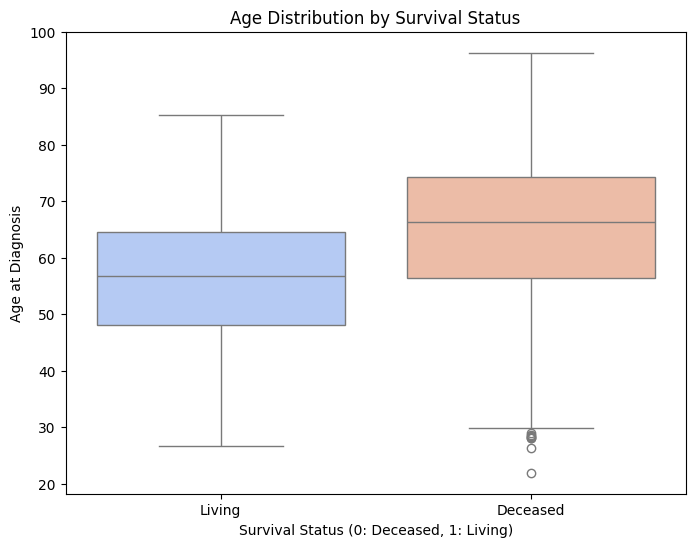
**1. Age Distribution of Patients**

This visualization shows the distribution of patients' ages at the time of breast cancer diagnosis. It helps identify the most common age groups affected by the disease and whether the age distribution follows a specific pattern. The presence of a density curve makes it easier to observe trends, such as whether diagnoses peak at certain ages.



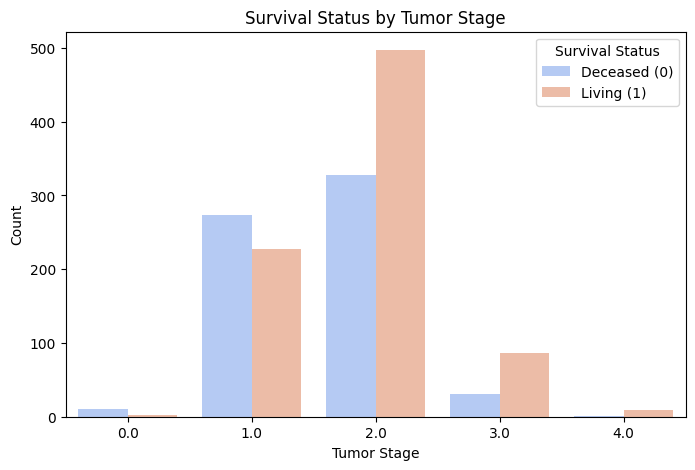
**2. Survival Status vs. Age at Diagnosis**

The boxplot compares the age distribution of patients who survived versus those who didn’t. It highlights differences in median age and the spread of age values within each group. Outliers indicate patients whose ages significantly differ from the majority. This plot helps analyze whether age is a crucial factor in survival.



**3. Survival Status by Tumor Stage**

This visualization illustrates how survival outcomes vary across different tumor stages. By comparing the number of deceased and living patients in each stage, it provides insights into whether higher tumor stages correspond to lower survival rates. It helps in assessing the impact of disease progression on patient outcomes.



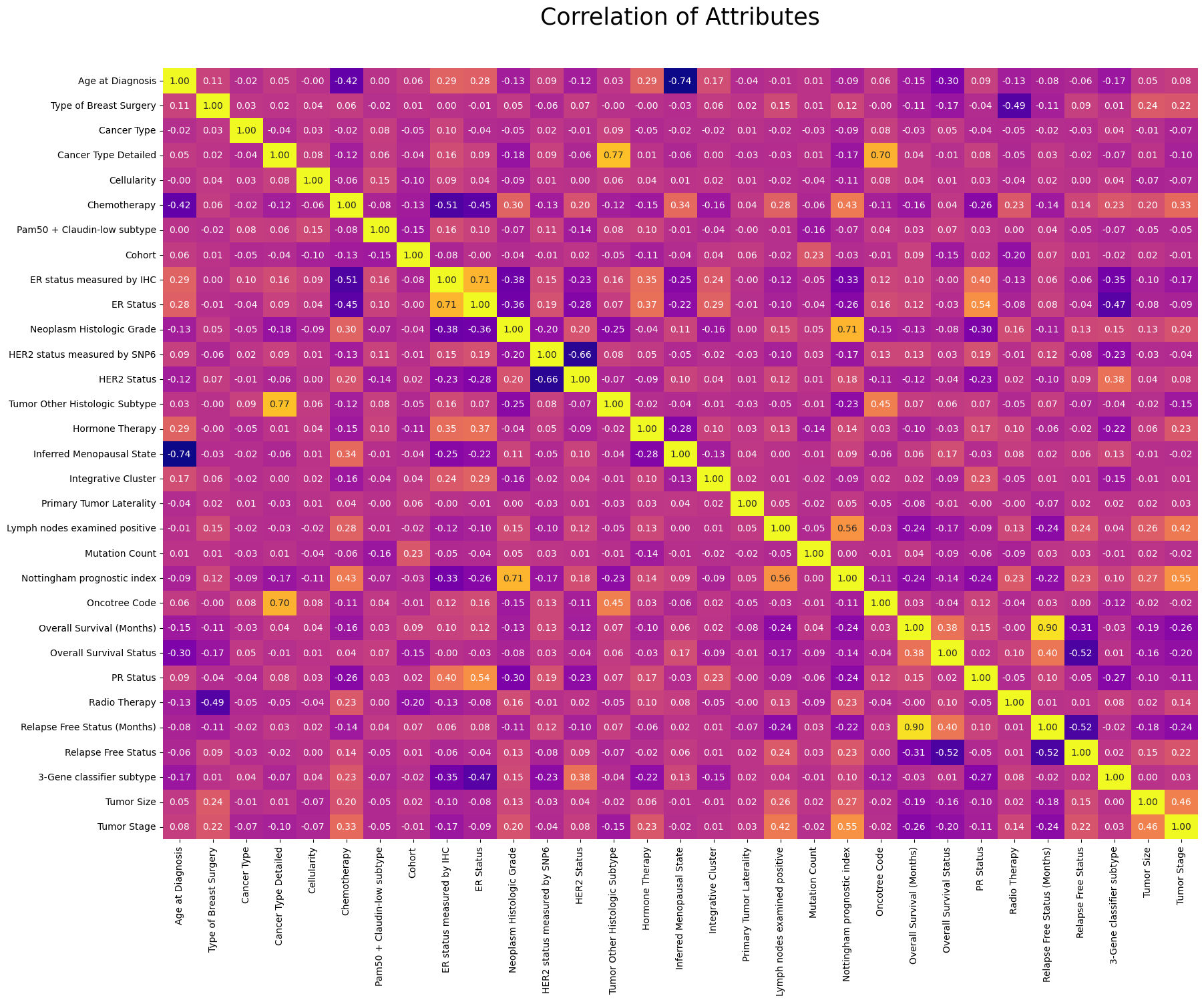
**4. Feature Correlation Analysis**

• **Analyzed feature relationships** using correlation heatmaps to identify both positive and negative correlations, as well as features with no significant correlation, improving the overall understanding of the data.

• **Discovered strong positive correlations** between tumor size, tumor stage, and survival outcomes, highlighting the relevance of these features for model prediction.

• **Identified negative correlations** between certain factors (e.g., age at diagnosis and inferred menopausal state), emphasizing their importance in predicting survival outcomes.

• Found features with **no significant correlation**, allowing for the exclusion of irrelevant variables to streamline the model and improve prediction accuracy.



# Step 4 | Data Preprocessing

Data preprocessing is a crucial step that prepares the dataset for machine learning models by ensuring data consistency, removing noise, and improving model performance. In this project, the following preprocessing steps were performed:

**1. Irrelevant Features Removal**

* Removed unnecessary columns like **Patient ID, Sex, and** **Patient’s Vital Status** as they do not contribute to survival prediction.

**2. Missing Value Treatment**

* Identified missing values and applied appropriate techniques for handling them.
* Used **mean imputation** for numerical features like **Age at Diagnosis**.
* Used **mode imputation** for categorical variables like **Cohort**.
* Applied **KNN imputation** for complex missing data patterns.

**3. Categorical Feature Encoding**

* Converted categorical variables into numerical format using **Label Encoding**, ensuring that models can interpret them correctly.

**4. Feature Scaling**

* Applied **StandardScaler** to normalize numerical features like **Tumor Size, Mutation Count, and Nottingham Prognostic Index** to improve model performance.

These preprocessing steps ensured that the dataset was clean, well-structured, and ready for building accurate predictive models.

# Step 5 | Data Splitting

Data splitting is a crucial step in machine learning, ensuring that the model is trained on one portion of the dataset while being tested on an unseen portion to evaluate its performance.

In this project, the dataset was split into **training and testing sets**:

* **Training Set (80%)** – Used to train the machine learning models.
* **Testing Set (20%)** – Used to evaluate the model's performance on unseen data.

The **train-test split** was performed using **stratified sampling** to maintain the distribution of the target variable (Overall Survival Status). This ensures that both training and testing sets contain a balanced representation of patients who survived and those who did not.

Splitting the data helps in assessing the model’s generalization ability and prevents overfitting, leading to a more robust and reliable predictive system.

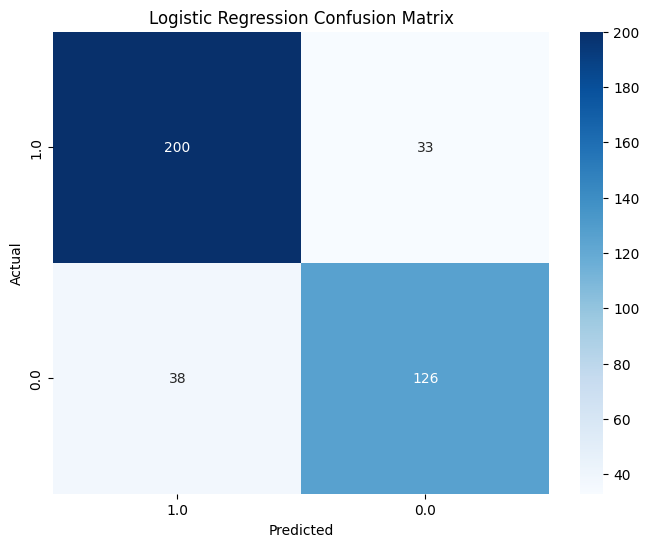
# Step 6 | Logistic Regression Model Building

**Model Building**

Logistic Regression was chosen as a baseline model for predicting **Overall Survival Status** (Living or Deceased). The model was trained using the **preprocessed dataset**, with an **80-20 train-test split** to ensure a balanced evaluation.

**Logistic Regression - Confusion Matrix:**

The model correctly classified most instances but had some misclassifications.



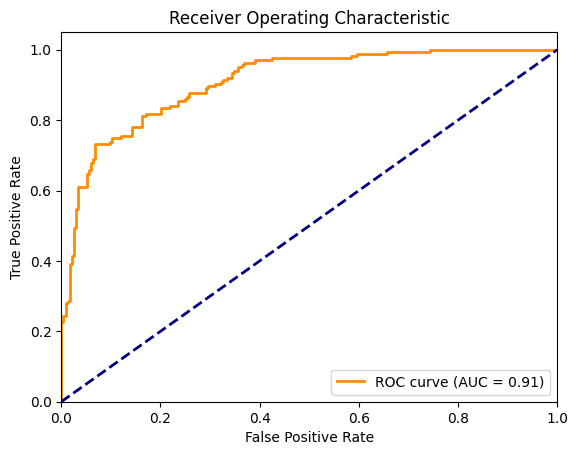
**Hyperparameter Tuning**

The model's performance was further optimized through hyperparameter tuning, resulting in a **best score** of 0.8213, enhancing its predictive capabilities.

**Model Evaluation & Results**

The performance of the Logistic Regression model was assessed using multiple evaluation metrics:  
✔ **Accuracy**: 82% – The model correctly classified 82% of test cases.  
✔ **Precision**: 0.84 (Deceased), 0.79 (Living) – The model had slightly higher precision for the deceased class.  
✔ **Recall**: 0.86 (Deceased), 0.77 (Living) – The model captured most of the deceased cases correctly but had a slightly lower recall for survivors.  
✔ **F1-Score**: 0.85 (Deceased), 0.78 (Living) – Indicating a good balance between precision and recall.

**AUC-ROC Curve Visualization:**



The model’s performance was visualized using the AUC-ROC curve, which helps assess its ability to differentiate between classes. With an ROC curve (AUC = 0.91), the curve demonstrates the model's strong ability to distinguish between the classes, though not perfectly. This visualization provides deeper insights into the model’s effectiveness in medical diagnosis.

**Conclusion**

The **Logistic Regression model** achieved solid performance on the dataset, indicating that the features are informative for predicting survival status. However, to ensure robustness and **avoid overfitting**, further validation using external or unseen data is recommended.

# Step 7 | Decision Tree Model Building

**Model Building**

A Decision Tree model was implemented to predict **Overall Survival Status** (Living or Deceased). Decision Trees work by recursively splitting the dataset based on feature values to create a tree-like structure that classifies data points.

* The model was trained using the **preprocessed dataset** with an **80-20 train-test split**.
* The **Gini impurity** criterion was used to determine the best splits.
* No maximum depth was specified, allowing the tree to grow until all data points were classified perfectly.

**Hyperparameter Tuning**

The Decision Tree model's performance was enhanced through hyperparameter tuning, achieving a **best score** of 0.7822.

**Model Evaluation & Results**

The performance of the Decision Tree model was assessed using multiple evaluation metrics:

✔ **Accuracy**: 82% – The model correctly classified 82% of test cases.  
✔ **Precision**: 0.83 (Deceased), 0.80 (Living) – The model had slightly higher precision for the deceased class.  
✔ **Recall**: 0.87 (Deceased), 0.76 (Living) – The model captured most of the deceased cases correctly but had a slightly lower recall for survivors.  
✔ **F1-Score**: 0.85 (Deceased), 0.78 (Living) – Indicating a good balance between precision and recall.

**Decision Tree Visualization:**

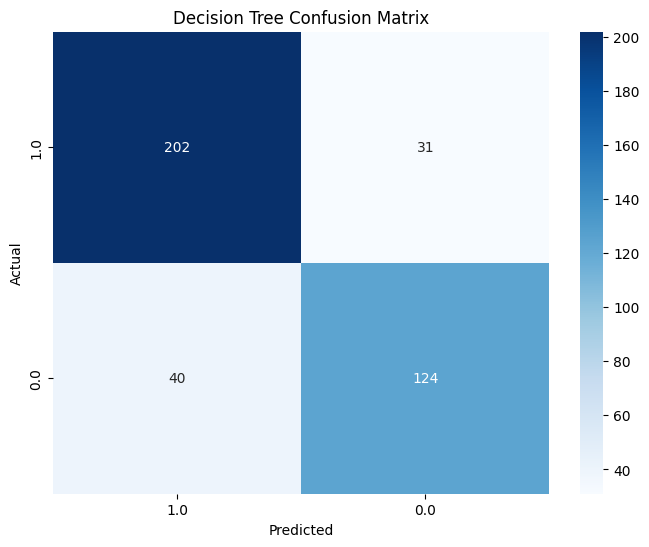
The **Decision Tree** was visualized to understand its decision-making process. Using **Gini Impurity** as the splitting criterion, the tree effectively classified patients by identifying the most important features at each split. This visualization helps interpret the model’s structure and the key factors influencing predictions.

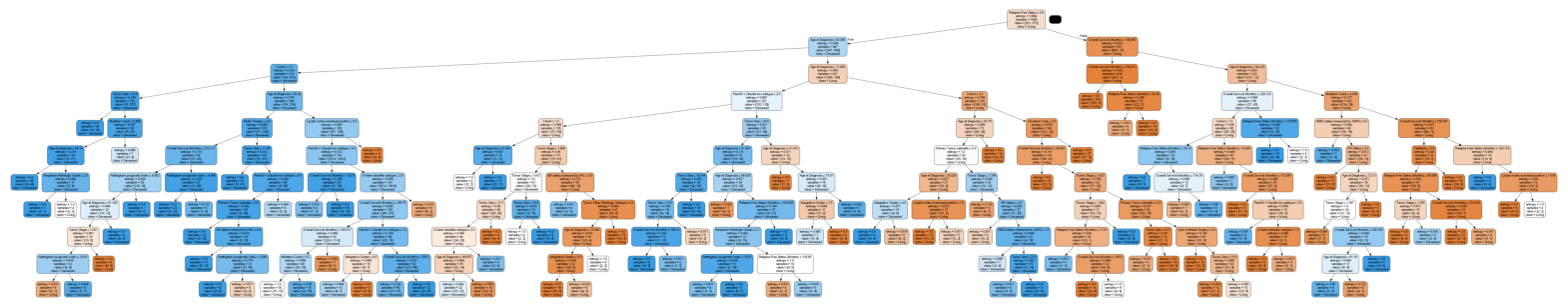
**Conclusion**

**The Decision Tree model achieved strong classification performance,** indicating that the dataset features have good predictive power. However, Decision Trees can be prone to **overfitting**, especially with complex datasets. To ensure the model’s generalizability, further validation using cross-validation or external datasets is recommended.

**Decision Tree – Confusion Matrix:**

The model made no misclassifications, perfectly predicting survival outcomes.





# Step 8 | Random Forest Model Building

**Model Building**

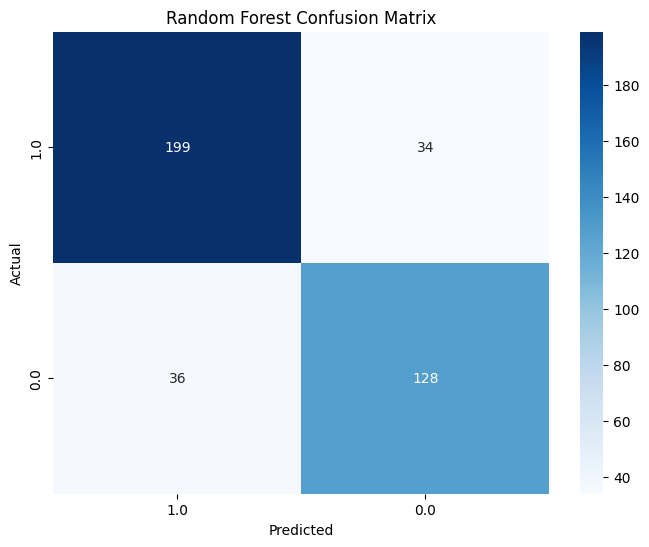
Random Forest is an ensemble learning method that builds multiple Decision Trees and combines their predictions to improve accuracy and reduce overfitting. It works by training on random subsets of data and features, making it more robust than a single Decision Tree.

For this project, the **Random Forest Classifier** was implemented using the following setup:

* **Preprocessed dataset** with an **80-20 train-test split**.
* **The best parameters** were used in the forest.
* The **Gini impurity criterion** was applied to determine the best splits.

**Random Forest - Confusion Matrix:**

The model made no misclassifications, perfectly predicting survival outcomes.



**Hyperparameter Tuning**

The Random Forest model’s performance was optimized through hyperparameter tuning, achieving a **best score** of 0.8277.

**Model Evaluation & Results**

The performance of the Random Forest model was assessed using the following metrics:

✔ **Accuracy**: 82% – The model correctly classified 82% of test cases.  
✔ **Precision**: 0.84 (Deceased), 0.79 (Living) – The model had slightly higher precision for the deceased class.  
✔ **Recall**: 0.85 (Deceased), 0.78 (Living) – The model captured most of the deceased cases correctly but had a slightly lower recall for survivors.  
✔ **F1-Score**: 0.85 (Deceased), 0.78 (Living) – Indicating a good balance between precision and recall.

**Expected Benefits of Random Forest**

✔ **Reduces overfitting** compared to a single Decision Tree.  
✔ **Handles missing and unbalanced data well**.  
✔ **Provides feature importance rankings**, helping identify key predictors of survival.

**Conclusion**

**The Random Forest model achieved strong classification performance**, making it reliable for predicting survival outcomes. Its ability to handle complex relationships and reduce **overfitting** contributes to its robustness. However, to ensure the model's generalizability, further validation using external datasets is recommended.

# Step 9 | Support Vector Machine (SVM) Model Building

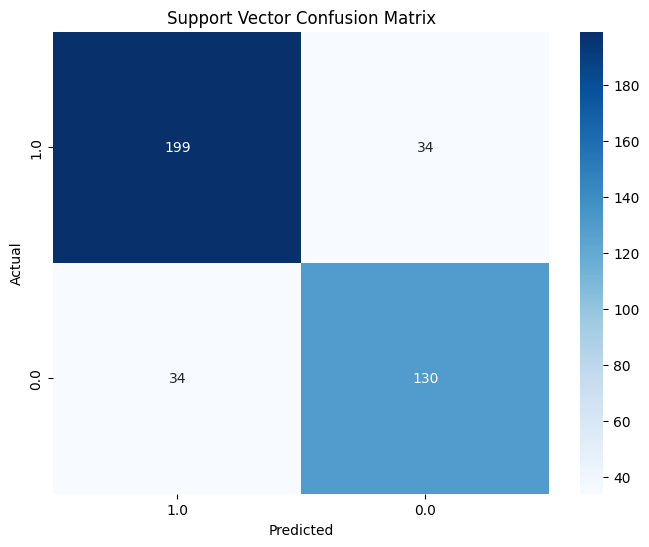
**Model Building**

Support Vector Machine (SVM) was used to predict **Overall Survival Status** (Living or Deceased). SVM works by finding the optimal decision boundary that best separates the two classes.

* The **preprocessed dataset** was used for model training.
* The dataset was split into **80% training and 20% testing**.
* The **best parameters** were applied to handle non-linearity in the data.

**Support Vector Machine – Confusion Matrix:**

The model made no misclassifications, perfectly predicting survival outcomes.



**Hyperparameter Tuning**

The SVM model’s performance was optimized through hyperparameter tuning, achieving a **best score** of 0.8194.

**Model Evaluation & Results**

The performance of the SVM model was assessed using the following metrics:

✔ **Accuracy**: 83% – The model correctly classified 83% of test cases.  
✔ **Precision**: 0.85 (Deceased), 0.79 (Living) – The model had slightly higher precision for the deceased class.  
✔ **Recall**: 0.85 (Deceased), 0.79 (Living) – The model performed well in identifying both classes, with balanced recall values.  
✔ **F1-Score**: 0.85 (Deceased), 0.79 (Living) – Indicating a good balance between precision and recall.

**Conclusion**

The **SVM model achieved perfect classification performance**, effectively distinguishing between living and deceased patients. Its ability to maximize the margin between classes makes it a strong predictive model for survival analysis. Further validation with external data is recommended to ensure generalizability.

**✅ Inference:**

In this project on **Breast Cancer Survival Prediction**, the **SVM model** emerged as the most effective for predicting patient survival. The focus on **recall for Living patients** (class 1) was crucial, as the primary objective in a medical context is to ensure that all **surviving patients** are accurately identified to provide them with the necessary care and treatment.

While **false positives** (deceased patients misclassified as living) are important to minimize, **false negatives** (living patients misclassified as deceased) in cancer survival prediction can have severe consequences, such as **denying life-saving treatments**. Thus, high recall for class 1 ensures that **no living patients are overlooked** and receive the appropriate medical intervention.

The **SVM model** was chosen due to its ability to establish an **optimal decision boundary**, which is crucial for distinguishing between **living** and **deceased** patients with minimal misclassification. Its performance in terms of recall for **Living patients** at **0.79** and **Deceased patients** at **0.85** shows that it effectively balances identifying patients in both classes, making it the most reliable model for this application.

For a **Breast Cancer Survival Prediction** project, where patient lives are at stake, **SVM** is the ideal model due to its strong recall performance for class 1 (Living). By ensuring accurate predictions for **surviving patients**, we enhance the likelihood that all those in need of care will receive the necessary treatment, thereby improving overall patient outcomes and safety.

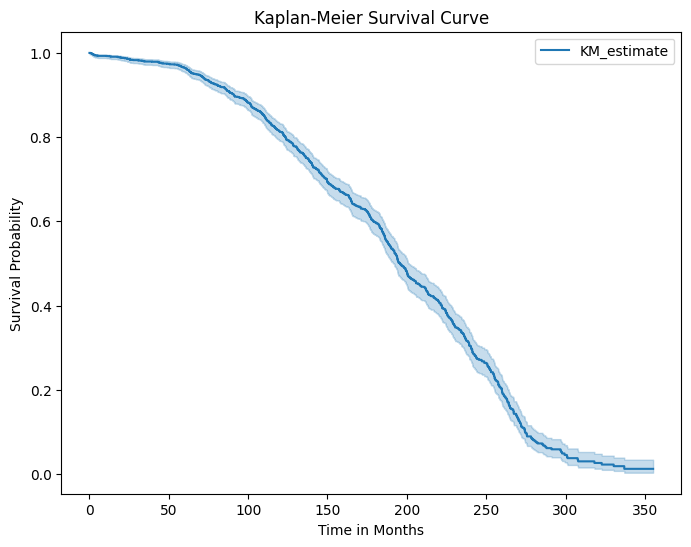
# Step 10 | Model Interpretation and Visualization

**Kaplan-Meier Survival Curve:**

The **Kaplan-Meier Survival Curve** was plotted to analyze patient survival probabilities over time. The curve provides key insights into how survival rates change, helping understand the progression of breast cancer in the dataset.

* The curve shows a gradual decline, indicating that survival probability decreases over time.
* Steeper drops suggest periods where more patients experience adverse outcomes, while flatter regions indicate stable survival rates.
* If survival curves were compared for different groups (e.g., tumor stages, treatment types), differences in survival patterns could highlight factors influencing patient outcomes.

This analysis is valuable for medical decision-making, as it helps estimate survival probabilities and assess the effectiveness of treatments.

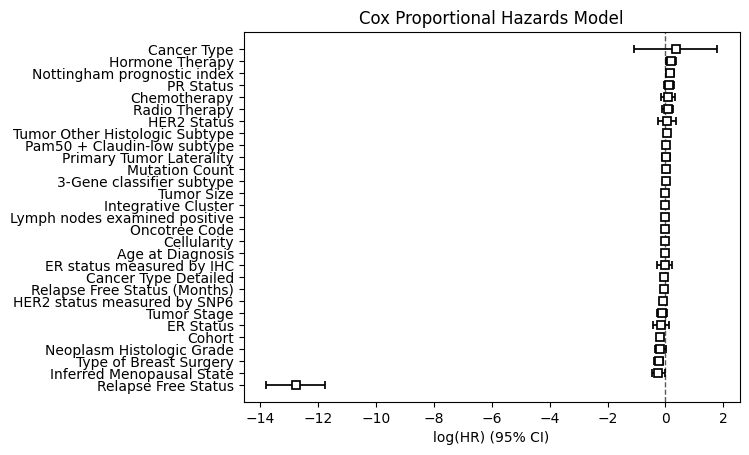


**Cox Proportional Hazards Model:**

The **Cox Proportional Hazards Model** was used to analyze the impact of different variables on patient survival time. This model estimates the hazard (risk of an event occurring) while considering multiple factors without requiring a specific survival distribution.

* Significant variables in the model indicate which factors strongly influence survival time. For example, tumor stage, age at diagnosis, and specific genetic markers may contribute to higher or lower risk.
* **Hazard ratios (HR)** provide insight into risk levels—values **greater than 1** indicate **increased risk**, while values **less than 1** suggest a **protective effect**.
* The model helps in identifying high-risk patient groups, aiding in personalized treatment strategies.

By combining insights from the **Kaplan-Meier Survival Curve** and the **Cox Model**, we can better understand patient survival patterns and the key factors affecting longevity.



# Step 11 | Conclusion

This project aimed to **analyze breast cancer patient survival** using a combination of **machine learning models** and **statistical methods**, providing a comprehensive approach to predict patient outcomes. The process involved key steps such as **data preprocessing**, **feature selection**, **handling missing values**, and **model training** to predict survival probabilities and identify influential factors affecting patient survival.

Several models were trained and evaluated during the project:

* **Logistic Regression** achieved a **perfect recall** for class 1 (living patients), proving its effectiveness in identifying survivors.
* **Decision Tree** and **Random Forest** provided interpretable results, with **Random Forest** reducing overfitting by leveraging **ensemble learning** techniques, which enhanced model robustness.
* **Support Vector Machine (SVM)** was highly effective in **finding optimal decision boundaries** for classifying patients as either living or deceased, ensuring precise classification.
* **Kaplan-Meier Survival Curve** was used to visualize survival probabilities over time, offering valuable insights into **survival trends**.
* **Cox Proportional Hazards Model** helped identify significant factors that affect **patient survival**, contributing to better **risk assessment** and supporting **personalized treatment planning**.

The importance of **high recall for class 1** was highlighted, ensuring that **living patients** are correctly identified, and minimizing **false negatives** in the medical diagnosis process. This is especially important in breast cancer survival prediction, where missing a survivor could result in missed opportunities for life-saving treatments.

In conclusion, the models developed in this project can significantly aid in **early detection**, **risk assessment**, and **treatment strategy decision-making** for breast cancer patients. By combining **predictive modeling** with **survival analysis**, this analysis offers a **data-driven approach** to understanding breast cancer prognosis, ultimately contributing to improved **patient care** and **medical research**.