**Breast Cancer 10-Year Mortality Risk Prediction**

**Problem Statement:**

This project aims to predict the 10-year mortality risk for breast cancer patients using the METABRIC dataset. By leveraging clinical and genomic data, the project employs data exploration, survival analysis, feature engineering, and machine learning techniques to build a predictive model for 10-year mortality.

**Objectives**

* Conduct Exploratory Data Analysis (EDA) to understand feature distributions and relationships.
* Implement Kaplan-Meier and Cox Proportional Hazards models for survival analysis.
* Apply various classification models to predict 10-year mortality.
* Evaluate model performance using appropriate metrics.

**1. Data Understanding**

**Dataset Overview**

The METABRIC dataset contains clinical and survival data for breast cancer patients, with 34 columns such as:

* **Age at Diagnosis**: Age of patient at the time of diagnosis.
* **Tumor Size**: Size of the tumor measured in millimeters.
* **Lymph Nodes Examined Positive**: Count of lymph nodes with cancer cells.
* **Nottingham Prognostic Index**: A prognostic index based on tumor size, lymph node status, and grade.
* **Overall Survival (Months)** and **Overall Survival Status**: Patient’s survival time and status (Deceased/Alive).
* **Target Variable**: Derived as a binary indicator of 10-year mortality, based on overall survival data.

**2. Data Preparation and Exploratory Data Analysis (EDA)**

**2.1 Data Cleaning**

* **Missing Values**: Missing values were identified and handled using median imputation for numerical features and mode for categorical features.
* **Encoding**: Categorical variables were label-encoded to prepare them for modeling.
* **Scaling**: Numerical features were scaled using StandardScaler to ensure uniformity in the modeling process.

**2.2 Exploratory Data Analysis (EDA)**

EDA provided insights into the distributions, outliers, and relationships among features:

* **Distribution Analysis**: Histograms illustrated the spread of numerical features such as **Age at Diagnosis** and **Tumor Size**.
* **Outlier Detection**: Boxplots identified outliers, especially in features like **Tumor Size**, which may influence predictions.
* **Correlation Matrix**: A heatmap showed correlations, informing feature selection and highlighting any highly correlated variables.

**3. Survival Analysis**

* **Kaplan-Meier Survival Curve**: This curve provided an overview of patient survival probability over time. The 10-year survival threshold was a critical factor in defining the target variable.
* **Cox Proportional Hazards Model**: This model helped assess the impact of various features (e.g., **Age**, **Tumor Size**) on patient survival time, revealing which variables significantly contribute to survival outcomes.

**4. Feature Engineering**

* **Target Variable Creation**: The target variable, **10\_Year\_Mortality**, was generated based on the survival data, defined as 1 if the patient’s overall survival was less than or equal to 120 months and 0 otherwise.
* **Feature Selection**: Important features were selected based on domain knowledge and the Cox model’s insights to improve model interpretability and performance.

**5. Model Selection and Training**

**Logistic Regression**

* **Model Choice**: Logistic Regression was chosen due to its interpretability and suitability for binary classification tasks.
* **Hyperparameter Tuning**: A GridSearchCV approach optimized the model’s C parameter to balance bias and variance.

**Model Training and Evaluation**

The Logistic Regression model was trained on the processed dataset, with evaluation metrics calculated on the entire dataset to assess model performance.

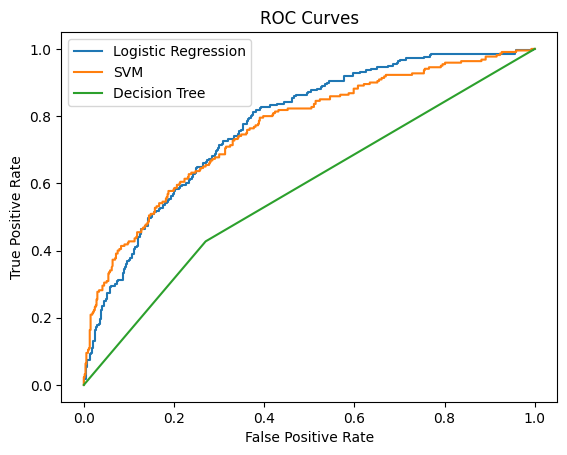
**6. Evaluation and Results**

**Performance Metrics**

* **Accuracy**: Measures the proportion of correct predictions.
* **Precision, Recall, F1 Score**: Evaluates the model’s performance in identifying true positives and balancing false positives and false negatives.
* **ROC-AUC Score**: Indicates the model’s ability to distinguish between classes, providing insight into the model’s predictive power.

**ROC Curve**

The ROC curve visualizes the model’s sensitivity vs. specificity, helping assess its trade-offs in predicting mortality risk.



**Example Predictions**

For each patient, the model predicts:

* **10\_Year\_Mortality Prediction**: Binary prediction (1 for deceased within 10 years, 0 otherwise).
* **Prediction Probability**: Probability estimate of the patient’s 10-year mortality risk, aiding in interpreting confidence levels of predictions.

**7. Conclusion and Future Work**

This analysis developed an interpretable, reliable model for predicting 10-year mortality risk in breast cancer patients using logistic regression, based on clinical and genomic features. Potential future steps include exploring other survival modeling techniques, optimizing additional features, and validating on external datasets.