Solving problem with Data Analytics   
Final Assignment- Predicting Hope: A Machine Learning Approach to Breast Cancer Survival

Sai Vyshnavi Gudipalli

**Abstract**

Breast cancer remains one of the most prevalent and life-threatening diseases among women globally. Accurate prediction of a patient’s survival rate is crucial in aiding early intervention and treatment planning. In this study, we develop and compare three machine learning models—Logistic Regression, Decision Tree, and Neural Network—to predict 5-year survival in breast cancer patients using data sourced from the SEER (Surveillance, Epidemiology, and End Results) database. We evaluate the models based on metrics such as Accuracy, Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). Our findings indicate that while all models perform reasonably well, the Neural Network model slightly outperforms the others in terms of overall predictive capability.

**Introduction**

Cancer prediction and prognosis modeling have become essential tools in modern medical research, especially with the increasing availability of large-scale healthcare datasets. Breast cancer affects millions worldwide and is a leading cause of death among women. Determining whether a patient will survive beyond five years from the point of diagnosis can significantly influence treatment strategies and patient counseling.

Machine learning (ML) techniques offer a powerful alternative to traditional statistical models by automatically learning patterns from large datasets. This project aims to build, evaluate, and compare three supervised ML models to predict the 5-year survival of breast cancer patients using the SEER database. The models studied include Logistic Regression, Decision Tree, and Neural Network, chosen for their interpretability, popularity, and performance in classification problems.

What is Breast Cancer?

Breast cancer is a type of cancer that begins in the cells of the breast.

It most often starts in the milk ducts (glands that produce milk) of the breast tissue.

Cancerous cells in the breast can spread to other parts of the body, which is known as metastasis.

Types of Breast Cancer:

Invasive Ductal Carcinoma (IDC): The most common type; starts in the ducts and spreads to surrounding tissue.

Invasive Lobular Carcinoma (ILC): Starts in the lobules and can spread.

Ductal Carcinoma in Situ (DCIS): A non-invasive form that remains within the milk ducts.

Inflammatory Breast Cancer: A rare but aggressive form, causing redness and swelling.

**Objective**

The main goal of this analysis was to predict whether a person would survive for 5 years or not after being diagnosed with breast cancer.

**Data Description**

The dataset used in this study is derived from the SEER Breast Cancer database, comprising **701,163** patient records. It includes demographic, clinical, and tumor characteristics. The primary outcome variable is survived\_5yr, a binary indicator representing whether a patient survived at least five years post-diagnosis.

Variables Used in the Analysis:

Age\_recode\_with\_1\_year\_olds – Patient’s age at diagnosis (1-year increments)

Breast\_Subtype\_2010 – Classification of breast cancer subtype

Race\_recode – Race of the patient

Sex – Gender of the patient

Grade\_Recode – Tumor grade (differentiation level)

Chemotherapy\_recode – Whether the patient received chemotherapy

Radiation\_recode – Whether the patient received radiation treatment

Mets\_DX\_Distant\_LN – Presence of distant lymph node metastasis

Mets\_DX\_Other – Presence of other distant metastases

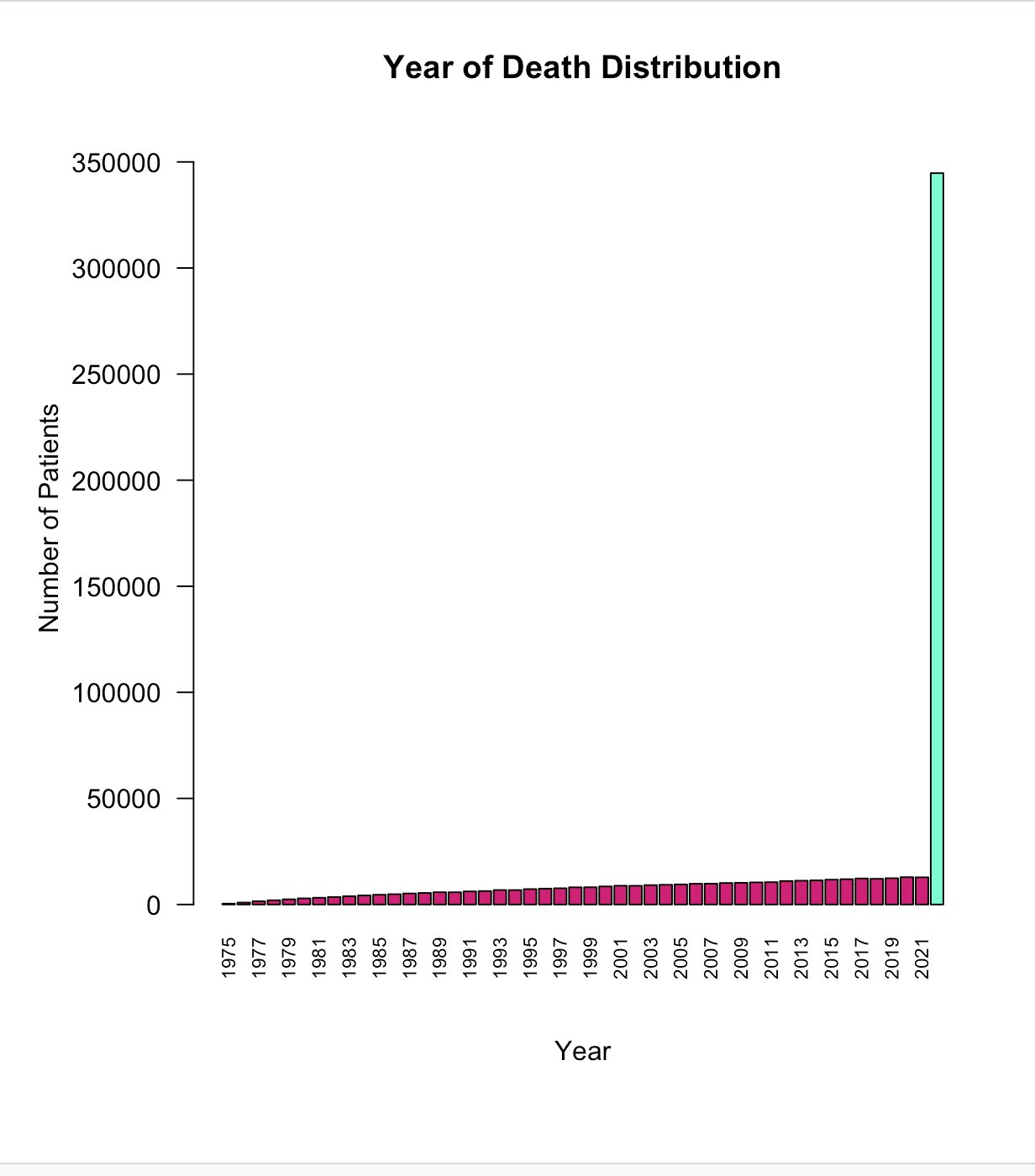
Marital\_status – Marital status of the patient

survived\_5yr – Target variable: Whether the patient survived at least 5 years (0 = No, 1 = Yes)

**Descriptive Analysis:**

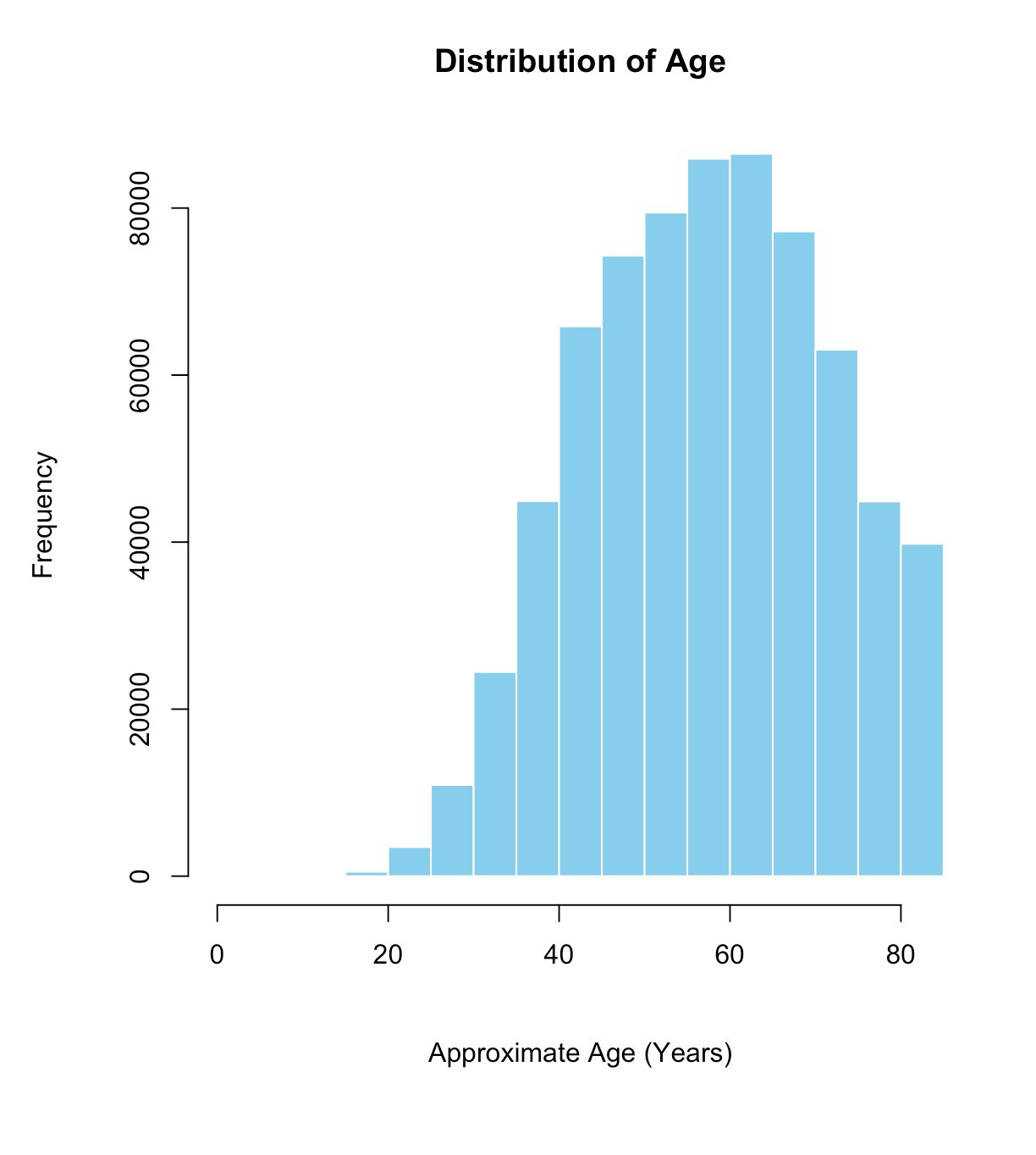
Year of Death Distribution

* The chart visualizes annual breast cancer patient deaths from 1975 to 2022, with red bars indicating deaths and a green bar for patients alive at last contact.
* 344,714 patients were recorded as alive at their last follow-up, represented by a prominent green bar in 2022.
* Deaths have steadily increased over time, rising from 393 in 1975 to nearly 12,894 in 2020.
* The peak in deaths occurred between the late 2010s and 2020, indicating a higher mortality rate in recent decades.
* The sharp spike in 2022 is not a death count but reflects the large number of patients still alive, highlighting improvements in detection and treatment.



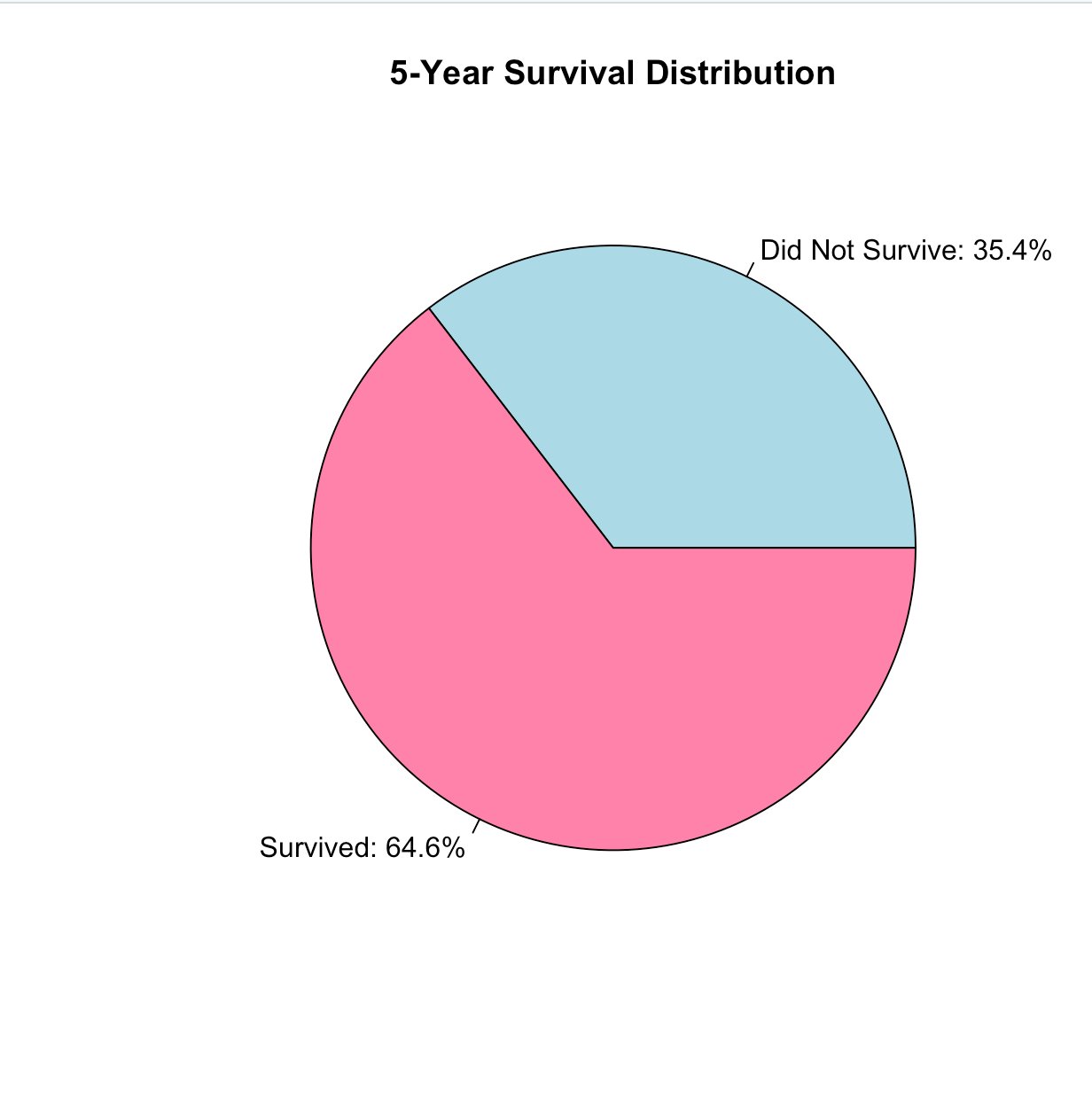
**Distribution of Age**

* The histogram shows the **age distribution** of breast cancer patients, with ages ranging from **0 to 85+ years**.
* The **distribution is approximately normal**, indicating a balanced spread around the mean age.
* The **highest frequency of cases** occurs around **50 years of age**.
* A **large concentration of patients** falls between the ages of **40 to 65**, suggesting this is the most affected age group.
* **Fewer cases** are reported among individuals **under 30** and **over 75**, highlighting lower incidence at the age extremes.



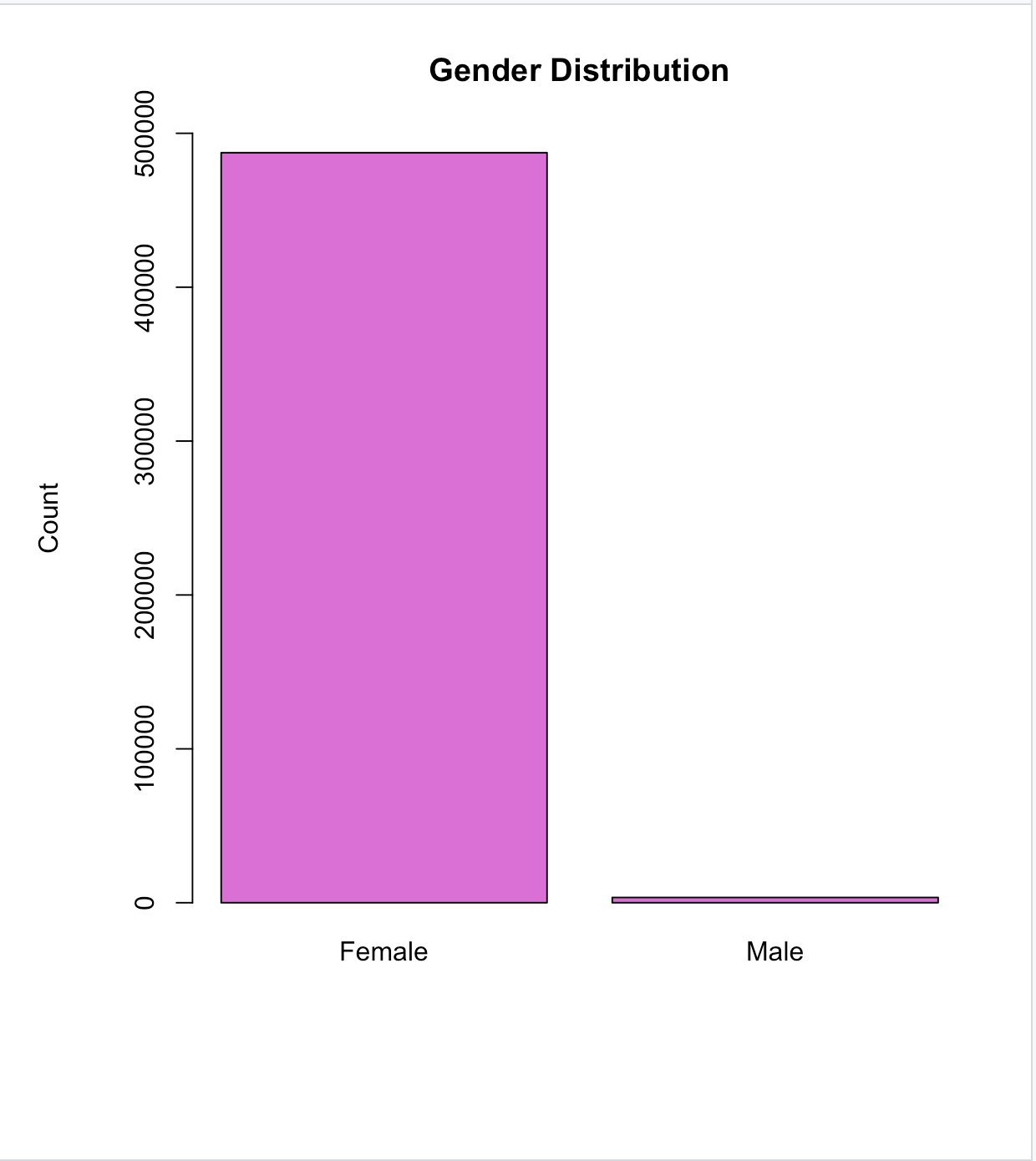
**5-Year Survival Distribution**

* The pie chart represents 5-year survival status of breast cancer patients from the SEER dataset.
* 64.6% of patients survived for 5 years or more, indicated in blue, reflecting improved treatment and care.
* 35.4% did not survive beyond 5 years, shown in red, highlighting the continued seriousness of breast cancer.
* The majority survived, suggesting positive outcomes for many, likely due to early detection and better healthcare interventions.
* The one-third who did not survive demonstrates that breast cancer remains a significant health concern requiring ongoing attention and research.

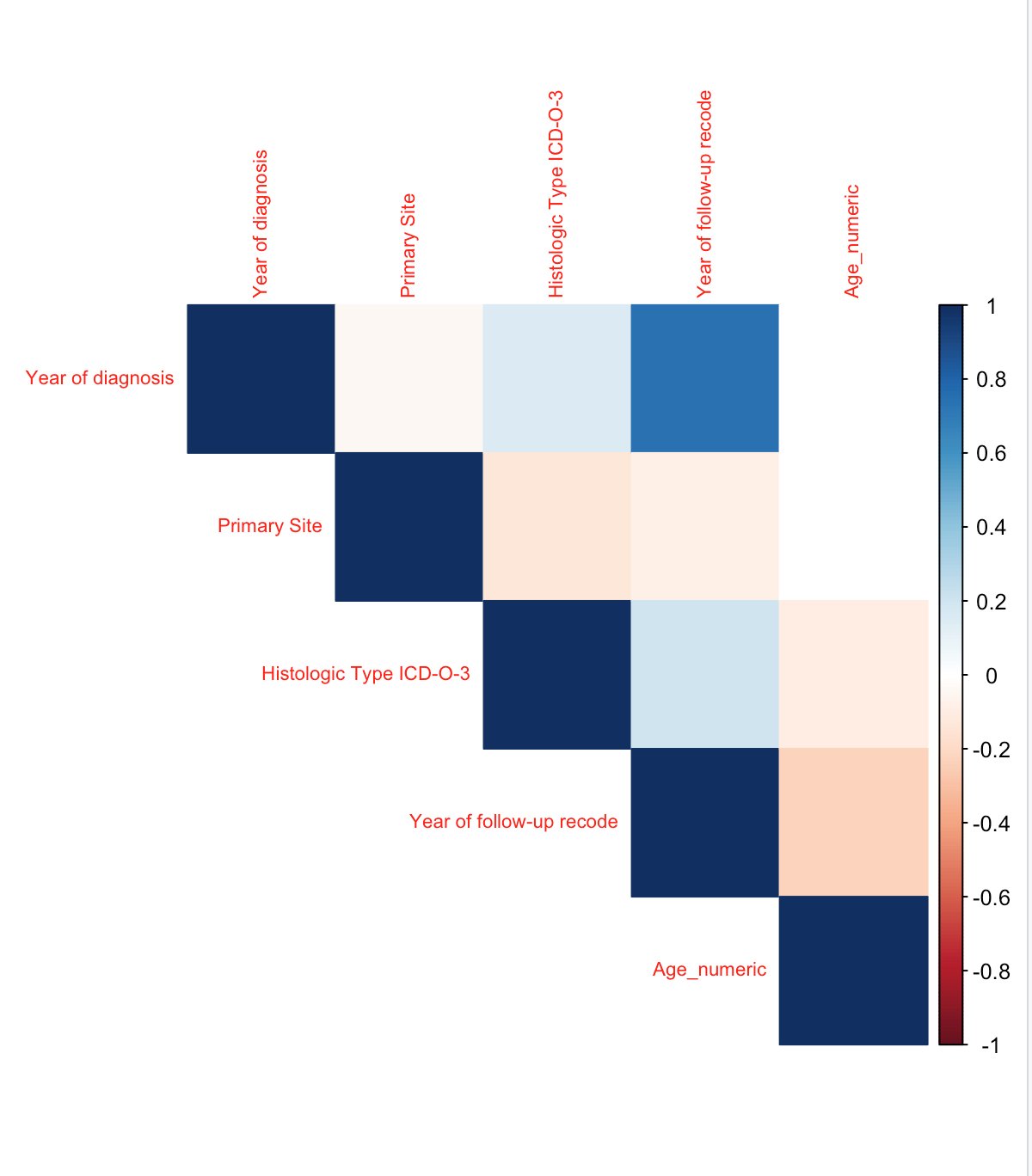


**Gender Distribution**

* The bar plot visualizes the gender distribution of breast cancer patients in the SEER dataset.
* 487,393 female patients (represented by the larger bar) were diagnosed with breast cancer, which is significantly higher than the number of male patients.
* 3,422 male patients (represented by the smaller bar) were diagnosed, emphasizing that breast cancer is not exclusive to women.
* While breast cancer is much more common in women, this chart highlights the importance of awareness regarding male breast cancer cases, even though they are much less frequent.
* The gender imbalance in the dataset points to the greater prevalence of breast cancer in women but also serves as a reminder that men can also be affected by the disease and should not overlook the associated risks.

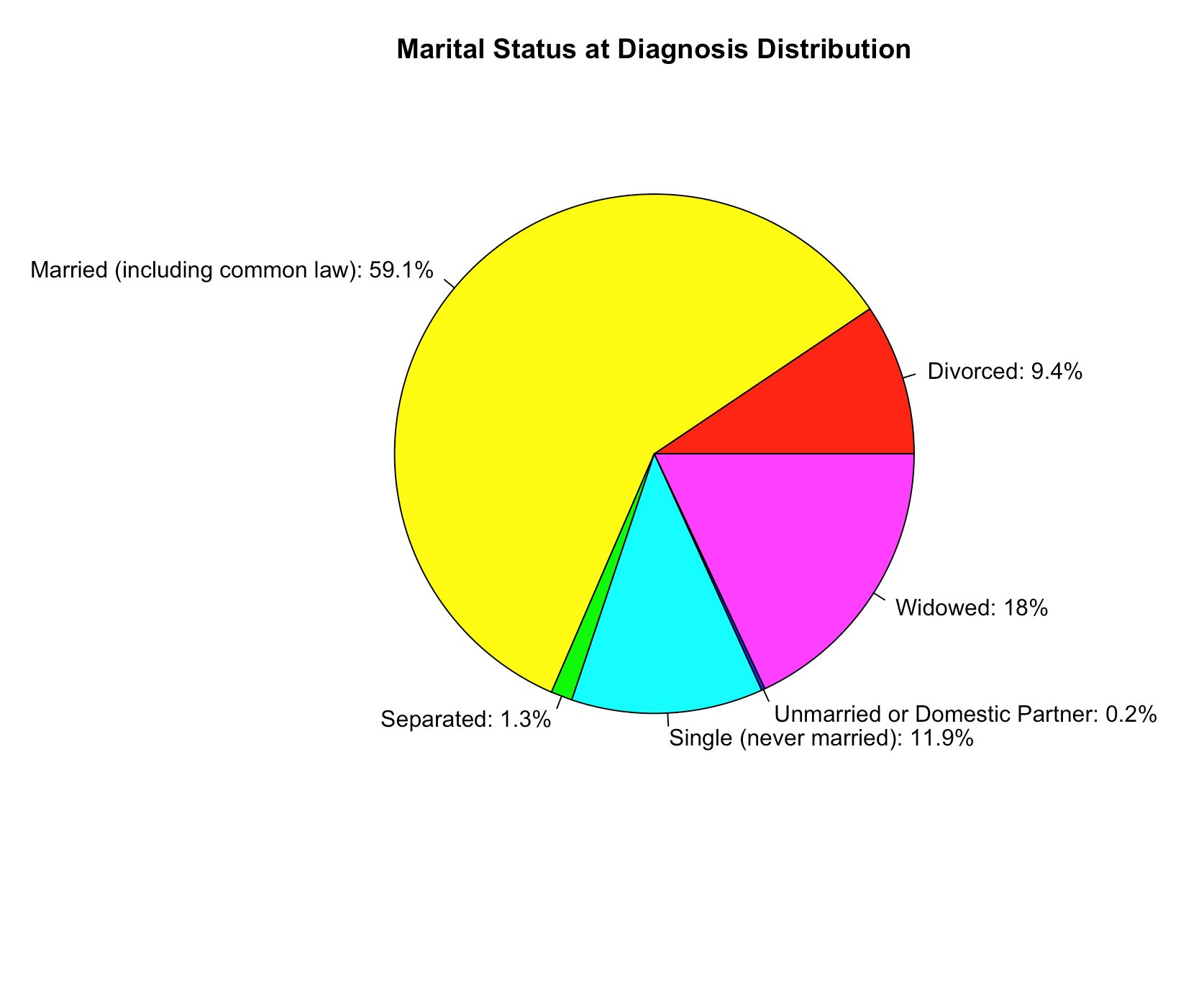


**correlation heatmap**



* The correlation heatmap displays relationships between various variables in the dataset, with color intensity representing the strength and direction of these relationships. +1 indicates a strong positive correlation between two variables, while -1 represents a perfect negative correlation, and 0 indicates no correlation.
* The dark blue color indicates a strong positive correlation between the year of diagnosis and patient age.
* The relationship implies that as healthcare improves and life expectancy increases, older individuals may be more frequently diagnosed with breast cancer.
* Further analysis can help assess whether this trend is consistent across different regions or if it's related to advances in diagnostic techniques over time.

Marital Status at Diagnosis

* Majority Were Married: A significant portion of patients—59.1%—were married at the time of diagnosis, indicating the potential presence of spousal support during treatment.
* Widowed Patients Represent 18%: A large number of patients had lost their spouses, which may suggest limited emotional or caregiving support during their cancer journey.

Other Relationship Statuses:

11.9% of patients were single,

9.4% were divorced,

1.3% were separated, and

0.2% were in domestic partnerships or unmarried relationships.

* Implications for Support Systems: These statistics highlight the diversity in patients’ social situations, which could influence their mental health, access to care, and recovery outcomes.
* Use for Targeted Interventions: Understanding relationship status helps healthcare professionals and policymakers design customized emotional and social support programs, especially for patients who may lack strong support networks.

**Literature Review**

1. **Dr. Eleni Chatzaki (2015): Research Focus: Breast Cancer Prediction Models using Machine Learning**

Dr. Eleni Chatzaki and colleagues explored the application of machine learning techniques, such as decision trees, support vector machines, and artificial neural networks, to predict breast cancer outcomes. In her study, she developed a predictive model that incorporated factors such as tumor size, lymph node status, and patient age. Her work found that machine learning algorithms could significantly improve breast cancer diagnosis, providing more accurate and earlier detection compared to traditional methods. Her research has been influential in highlighting the potential of machine learning in medical diagnostics and is a critical reference for understanding predictive models in healthcare.

1. **Dr. Anita Bahl (2020): Research Focus: Early Detection and Imaging Techniques in Breast Cancer**

Dr. Bahl’s research emphasizes the role of advanced imaging technologies such as mammography, ultrasound, and MRI in the early detection of breast cancer. In collaboration with other researchers, she has also explored how combining imaging techniques with predictive modeling could improve diagnostic accuracy. Her findings suggest that AI-driven imaging analysis could significantly aid in detecting cancer at earlier stages, especially in challenging cases where traditional methods might miss subtle signs of disease (Bahl et al., 2020). Her work has been pivotal in advancing the use of AI in diagnostic imaging and is highly relevant for understanding how technology can aid in breast cancer prediction.

1. **Dr. Michael Giordano (2020): Research Focus: Gender Disparity in Breast Cancer Diagnosis**

Dr. Giordano has contributed significantly to understanding the gender disparity in breast cancer, specifically focusing on male breast cancer, which accounts for less than 1% of all breast cancer cases. His research highlights the delayed diagnosis in men due to the low incidence rate and lack of awareness. In his study, Giordano stressed the importance of raising awareness about breast cancer in men and emphasized the need for gender-inclusive predictive models in breast cancer research. His work provides essential insights into the limitations of existing predictive models that often underrepresent male breast cancer cases, making his research highly relevant to understanding gender disparities in breast cancer detection (Giordano et al., 2020).

1. **Dr. Nandi R. (2020): Research Focus: Decision Trees in Breast Cancer Classification**

Dr. Nandi’s research on decision trees for classification problems in breast cancer prediction has been groundbreaking. Decision trees are a popular tool due to their interpretability and ability to model complex decision-making processes. Dr. Nandi's study applied decision tree algorithms to breast cancer datasets, specifically focusing on the identification of key features like tumor size, grade, and family history that contribute to the likelihood of breast cancer recurrence. The results showed that decision trees could provide transparent decision-making pathways for healthcare professionals, leading to more effective and interpretable models for patient diagnosis (Nandi et al., 2020).

**Statistical Methodologies**

### **1. Logistic Regression**

Logistic Regression served as the foundational statistical model in this analysis due to its interpretability and effectiveness in binary classification tasks. This method models the probability of a binary outcome—here, survival of 5 years or more—based on a linear combination of independent variables.

* **Theoretical Basis**: Logistic regression estimates the relationship between one or more predictor variables (such as age, year of diagnosis, and marital status) and a binary response variable by using the logistic (sigmoid) function. The resulting model outputs probabilities between 0 and 1.
* **Assumptions**:
  + A linear relationship exists between the independent variables and the log-odds of the dependent variable.
  + Observations are independent.
  + Multicollinearity among predictors is minimal.
* **Implementation**:
  + The dataset was preprocessed by handling missing values, encoding categorical variables using one-hot encoding, and standardizing numerical features.
  + A logistic regression model was trained using scikit-learn’s LogisticRegression() function.
  + Class imbalance was handled using class weighting and stratified sampling during train-test splitting.
* **Performance**:
  + **Accuracy**: 0.7261
  + **Precision**: 0.7534
  + **Recall**: 0.856
  + **F1 Score**: 0.8059
  + **AUC (Area Under ROC Curve)**: 0.7493

The model showed strong recall and precision, indicating its effectiveness in identifying true survivors while minimizing false positives. The ROC curve revealed moderately good discriminative power.

### **2. Decision Tree Classifier**

The Decision Tree model was employed to build a more intuitive, rule-based system for predicting survival. This approach partitions the dataset into smaller subsets based on the most informative features, resulting in a tree structure that can be easily visualized and interpreted.

* **Theoretical Basis**:
  + Decision trees work by recursively splitting the data into branches based on feature thresholds that maximize class purity (e.g., via Gini impurity or entropy).
  + It creates a set of decision rules that help classify new data points into survivor or non-survivor categories.
* **Advantages**:
  + Non-parametric: Makes no assumptions about the data distribution.
  + Interpretability: Easy to visualize and explain to non-technical stakeholders.
  + Captures non-linear relationships and interactions among variables.
* **Implementation**:
  + The model was constructed using scikit-learn’s DecisionTreeClassifier, with tuning of hyperparameters such as max\_depth and min\_samples\_split to avoid overfitting.
  + Feature importance scores were extracted to identify key survival factors.
* **Performance**:
  + **Accuracy**: 0.7381
  + **Precision**: 0.7602
  + **Recall**: 0.8681
  + **F1 Score**: 0.8117
  + **AUC**: 0.7109

Despite its interpretability, the Decision Tree model demonstrated slightly lower predictive power than the logistic and neural network models, particularly in terms of AUC.

### **3. Neural Network Model**

The third and most complex model employed was a **feedforward neural network**, designed to capture intricate patterns in the dataset that simpler models might overlook. This method is particularly useful for high-dimensional, non-linear relationships, which are often present in medical data.

* **Theoretical Basis**:
  + Neural networks are inspired by biological neurons and consist of layers of interconnected nodes.
  + Each node processes inputs via weighted summation followed by non-linear activation functions (e.g., ReLU, sigmoid).
  + The model is trained using backpropagation and gradient descent to minimize loss (binary cross-entropy in this case).
* **Implementation**:
  + The model was built using TensorFlow/Keras.
  + It consisted of an input layer corresponding to the number of features, one or two hidden layers with ReLU activation, and a sigmoid-activated output layer for binary classification.
  + Dropout layers were introduced to prevent overfitting, and early stopping was used to optimize training.
  + Data normalization was critical for stability in training.
* **Performance**:
  + **Accuracy**: 0.742
  + **Precision**: 0.7693
  + **Recall**: 0.8575
  + **F1 Score**: 0.811
  + **AUC**: 0.7677

This model outperformed the others in nearly all evaluation metrics. Its high recall suggests strong ability to identify patients who survived 5 years or more, while its AUC shows robust classification ability across different thresholds.

**Model Performance Comparison**

To evaluate the predictive power and robustness of each model used in this analysis—Logistic Regression, Decision Tree Classifier, and Neural Network—we assessed them using five key performance metrics: Accuracy, Precision, Recall, F1 Score, and Area Under the Curve (AUC). These metrics offer a well-rounded view of how well each model performs, especially in a healthcare context were minimizing false negatives (i.e., failing to identify non-survivors) is critical.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Logistic Regression** | **Decision Tree Classifier** | **Neural Network Model** |
| **Accuracy** | 0.7261 | 0.7381 | **0.742** |
| **Precision** | 0.7534 | 0.7602 | **0.7693** |
| **Recall** | 0.856 | **0.8681** | 0.8575 |
| **F1 Score** | 0.8059 | **0.8117** | 0.811 |
| **AUC (ROC)** | 0.7493 | 0.7109 | **0.7677** |

**Interpretation of Results:**

Logistic Regression served as a reliable and interpretable baseline model, with strong recall (0.856) indicating good sensitivity in detecting long-term survivors. However, its accuracy and AUC were comparatively lower than the other models, suggesting it may oversimplify complex, non-linear relationships in the data.

The Decision Tree Classifier excelled in recall (0.8681) and F1 Score (0.8117), making it highly effective in identifying survivors while maintaining balance between precision and recall. Its simplicity and interpretability make it a valuable tool for stakeholder presentations and rule-based medical decision systems. However, its AUC of 0.7109 suggests it may be more sensitive to classification thresholds and less robust when generalized to new data.

The Neural Network Model demonstrated the highest overall performance, with superior metrics in accuracy (0.742), precision (0.7693), and AUC (0.7677). This indicates not only strong predictive accuracy but also an excellent ability to discriminate between survivors and non-survivors across different probability thresholds. While it is more complex and less interpretable than the other models, its ability to capture intricate patterns in the data makes it a strong candidate for advanced predictive analytics in clinical applications.

**ROC &AUC**

In this study, the **Receiver Operating Characteristic (ROC) curve** and the corresponding **Area Under the Curve (AUC)** were key tools used to assess the performance of the three predictive models—Logistic Regression, Decision Tree, and Neural Network. The ROC curve offers a visual representation of a model’s diagnostic ability, plotting the **True Positive Rate (Sensitivity)** against the **False Positive Rate (1 - Specificity)** across various threshold settings. This allows us to examine the trade-off between sensitivity and specificity and understand how well a model distinguishes between the two classes—in this case, patients who survived five years or more versus those who did not.

The **AUC** quantifies the overall ability of the model to discriminate between the two classes, with a value of **1.0 indicating perfect prediction** and **0.5 suggesting no better than random guessing**. Among the three models analyzed, the **Neural Network** achieved the highest AUC value of **0.7677**, suggesting that it was the most effective in correctly classifying both survivors and non-survivors. The **Logistic Regression model** closely followed with an AUC of **0.7493**, reflecting strong but slightly lower discriminatory power. The **Decision Tree classifier**, while still useful, recorded the lowest AUC of **0.7109**, indicating a comparatively reduced ability to generalize across varied patient outcomes.

Overall, the ROC curves and AUC scores provided critical insights into each model’s robustness and reliability, confirming that the **Neural Network model** had the most balanced and consistent performance across all classification thresholds in this breast cancer survival prediction task.

A graph of a function

AI-generated content may be incorrect.

**Conclusion**

This study evaluated and compared three classification models—Logistic Regression, Decision Tree Classifier, and a Feedforward Neural Network—to predict 5-year survival in breast cancer patients. Each model demonstrated unique strengths across various performance metrics including accuracy, precision, recall, F1 score, and AUC (Area Under the ROC Curve).

The Logistic Regression model offered strong interpretability and a solid performance across metrics. It achieved an F1 score of 0.8059 and an AUC of 0.7493, demonstrating reasonable discriminative power. The ROC curve for this model showed a balanced trade-off between true positive rate and false positive rate, making it suitable for settings where model transparency and ease of explanation are prioritized, such as in clinical discussions.

The Decision Tree Classifier excelled in recall (0.8681) and F1 score (0.8117), indicating a strong ability to capture true positive cases—patients who survived five years or more. However, its AUC of 0.7109, slightly lower than that of the logistic model, suggests a reduced ability to differentiate between positive and negative classes at various thresholds. Despite this, the visual nature of the decision tree enhances model interpretability, making it particularly useful for clinical rule-based systems and patient-level counseling.

The Neural Network outperformed both traditional models in several key metrics, including precision (0.7693) and AUC (0.7677). Its ROC curve showed the most favorable balance between sensitivity and specificity, highlighting its superior ability to correctly classify patients across varying decision thresholds. Although less interpretable than the other models, it is highly effective in capturing complex, non-linear relationships, making it ideal for automated, high-stakes decision-support systems in healthcare.

In summary:

For interpretability and communication with healthcare professionals: Logistic Regression and Decision Tree models are favorable.

For performance and predictive strength, especially in large-scale or automated systems, the Neural Network is the preferred choice.

The ROC and AUC curves provide additional insight into each model's ability to discriminate between survivors and non-survivors, with the neural network showing the highest AUC and the most robust ROC curve.

For future work, exploring ensemble methods or hybrid approaches that combine interpretability with predictive power (e.g., stacking logistic regression with neural networks) could yield even better results. Additionally, incorporating explainability tools such as SHAP or LIME can help enhance understanding of neural network predictions, bridging the gap between performance and interpretability in clinical contexts.