Final Project Report: Predicting Breast Cancer Patient Outcomes Using Machine Learning

Name: Niyathi Lekkala
Cardinal ID: 5237545
Department: Electrical Engineering and Computer Science
Institution: The Catholic University of America
Email: lekkala@cua.edu

Name: Vinod Kumar Kethavath
Cardinal ID: 5237582
Department: Data Analytics
Institution: The Catholic University of America
Email: kethavath@cua.edu

Abstract

This project focuses on predicting breast cancer patient survival using machine learning techniques. By analyzing a publicly available breast cancer dataset, we aim to classify patients as either Alive or Dead based on clinical features such as age, tumor size, node involvement, and hormone receptor status. The dataset underwent rigorous preprocessing, including outlier removal, label encoding, scaling, and class balancing using SMOTE. Several models were implemented – including Logistic Regression, Decision Tree, Random Forest (with hyperparameter tuning), K-Nearest Neighbors, and XGBoost and evaluated based on accuracy, precision, recall, F1-score, and ROC-AUC. Among these, the tuned Random Forest classifier showed the best performance. Feature important analysis revealed that tumor size, age, and survival months were key predictors. This study demonstrates the value of machine learning in supporting clinical prognosis and highlights the importance of proper preprocessing and model evaluation techniques.

1. Introduction

Breast cancer is one of the leading causes of cancer of cancer-related deaths among women globally. Accurate and early prediction of patient survival plays a vital role in guiding treatment decisions and improving outcomes. Traditional clinical indicators such as age, tumor size, lymph node involvement, and hormone receptor status provide valuable information but may not fully capture complex survival patterns. With the increasing availability of healthcare datasets, machine learning has emerged as a powerful tool to support predictive analysis in oncology.

In this project, we leverage a dataset containing and pathological features of breast cancer patients to predict survival status (“Alive” or “Dead”). The process includes essential data cleaning, feature encoding, normalization, and handling of class imbalance using SMOTE. Several machine learning models, including Logistic Regression, Decision Tree, Random

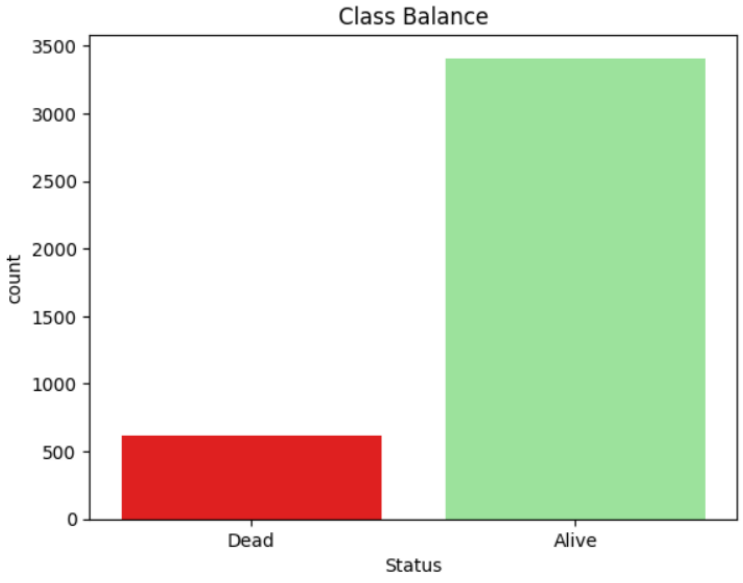
Forest, K-Nearest Neighbors, and XGBoost, are implemented and compared. Through hyperparameter tuning and evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, we identify the best-performing model. Additionally, feature importance analysis provides insights into the key attributes affecting patient survival, supporting future clinical research and decision-making.

2. Data set

The dataset used for this project is a breast cancer survival dataset that contains detailed information on patient demographics, tumor characteristics, treatment details, and outcome variables: Status (Alive/Dead). Each row represents a unique patient, with features such as age at diagnosis, tumor size, and binary outcome.
Key statics about the dataset:

- Number of records: 4024 records and 16 features
- Target variable: Status (binary classification: Alive/Dead)
- Feature Types:
    - Numerical: Age, Tumor Size, Regional Node Examined, etc.
    - Categorical: Race, Marital Status, T Stage, N Stage, 6th Stag, etc.

The target variable here seems to be imbalanced so, in the further steps we applied SMOTE to make the classes balance.



3. Data Pre-Processing

The data pre-processing phase was crucial in transforming the raw dataset into clean, structured format suitabke for machine learning. Here's how it was done:

- **Separating Features and Target:**
  The dataset was divided into two parts:
  - Features (X): This included all the information about the patients such as their age, Tumor size, Survival Months, and N Stage.
  - Target Variable (y): This is having the Status whether the patient is Alive (1) or Dead (0).

- **Categorizing Features:**
  The features were grouped into two types:
  - Categorical Features: Variables like Race, Marital Status, N stage, 6th stage, etc.
  - Numerical Features: Age, Tumor Size, Regional Node Examined, Regional Node Positive, etc.

- **Processing Categorical Data**:
  - Label encoder was applied to convert categorical values to the numerical values. Each categorical string values are assigned to each unique number.
  - Label Encoding was safely used as our models (like Random Forest, Decision Tree, XGBoost) are tree-based and do not assume any ordinal relationship between encoded values.
  - We applied encoding before splitting the data to ensure consistency and avoided issues like unseen categories during testing.

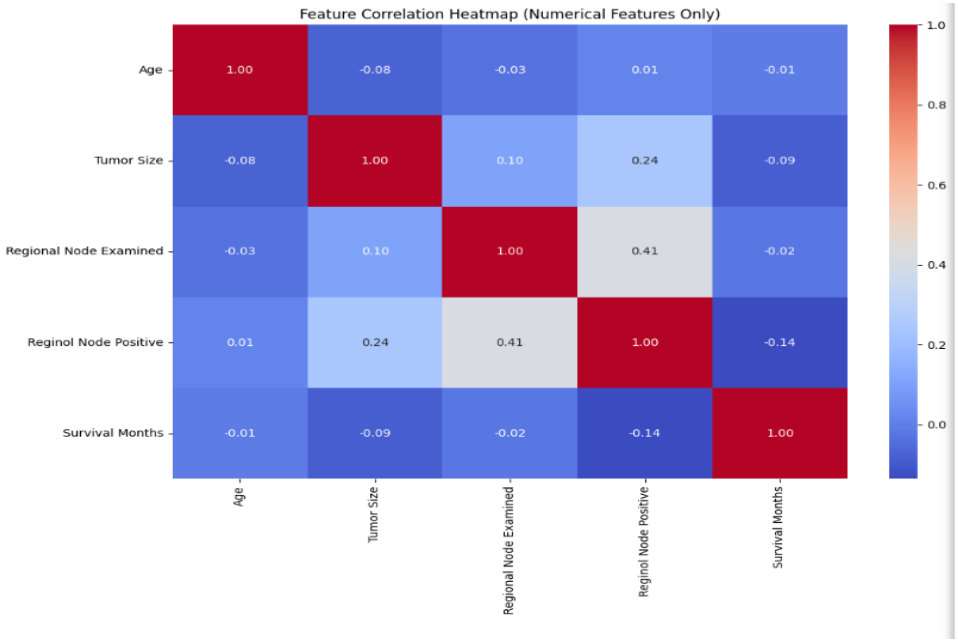
- **Normalizing Numerical Data:**
  - Numerical features such as Age, Tumor Size, Regional Node Examined, and Survival Months were normalized using Min-Max Scaling, transforming values into a standardized range between 0 and 1.
  - This ensures that no feature dominated the model due to scale differences, improving convergence and comparability during model training.

- **Combining the Data:**
  - The processed numerical and categorical features were merged into a single dataset, creating a comprehensive, ready-to-use dataset for machine learning models.
- **Splitting the Data:**
  - The final dataset was split into training (80%) and testing (20%) sets.
  - This step ensured that the model was trained on a portion of the data while being evaluated on unseen data, enabling an unbiased assessment of its performance.

4. Correlation Analysis

Before getting into modelling, we explored relationships within the dataset. A correlation heat map revealed several interesting insights:

- **Positive correlations**: A notable positive correlation (0.41) exists between Reginal Node Examined and Reginol Node Positive, indicating that more nodes examined may reveal more positives.
- **Negative correlations**: Survival Months shows slight negative correlations with features like Tumor Size (-0.09) and Reginol Node Positive (-0.14), suggesting more aggressive conditions may reduce survival time.
- **Weak correlations**: Most features show weak or near-zero correlation with each other, indicating low multicollinearity and the potential need for advanced feature selection or engineering.
- These findings helped us focus on the features most relevant to predicting survivals.

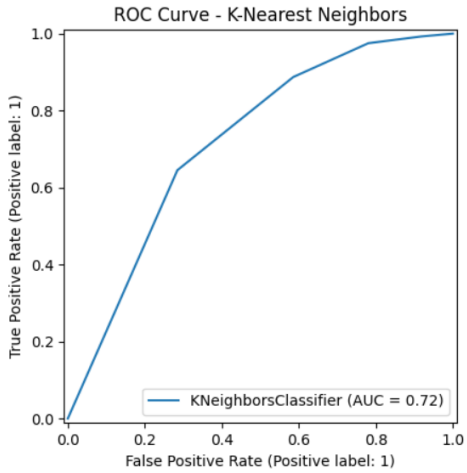


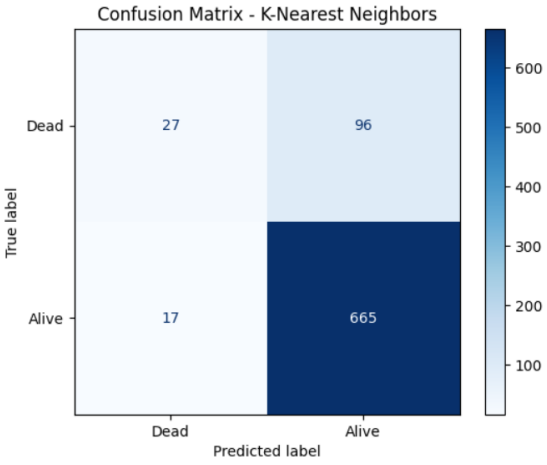
5. Models

To predict patient survival status (Alive/Dead), we implemented and compared several supervised classification models. Each model was chosen based on its strengths in handling different data complexities and structures:

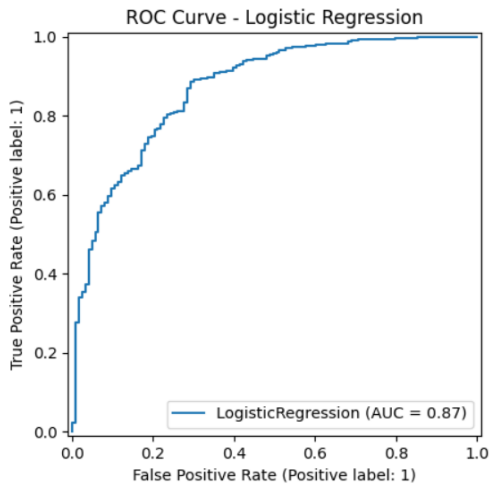
1. **K- Nearest Neighbors (k-NN):**

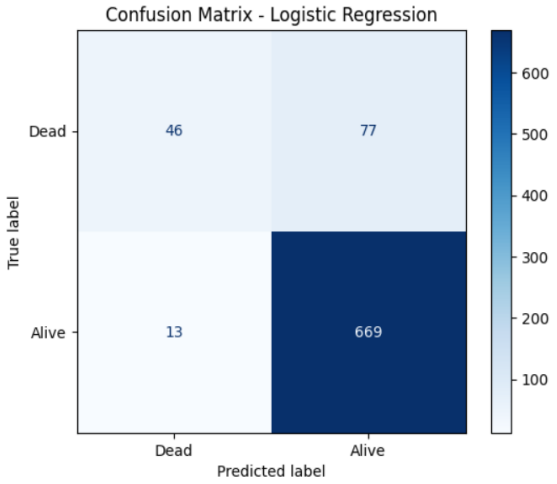
   Definition: A non-parametric model that classifies data points based on the majority vote of their 'k' nearest neighbors.
   Accuracy: 85.96%





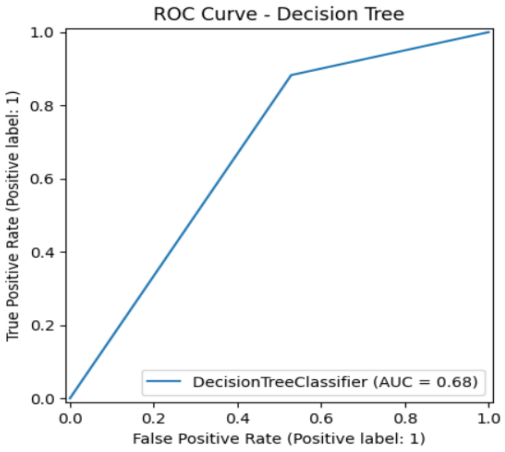
2. **Logistic Regression:**
   Definition: A linear model that estimates the probability of a binary outcome using a sigmoid function.
   Accuracy: 88.82%

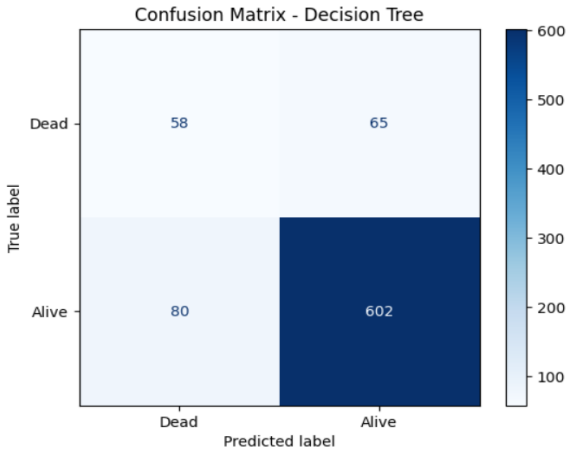




3. Decision Tree Classifier:

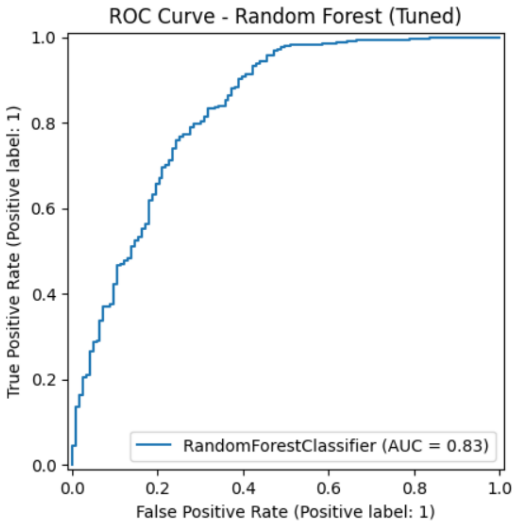
Definition: A tree-structured model that splits the data recursively based on feature importance using criteria like Gini impurity.
Accuracy: 81.99

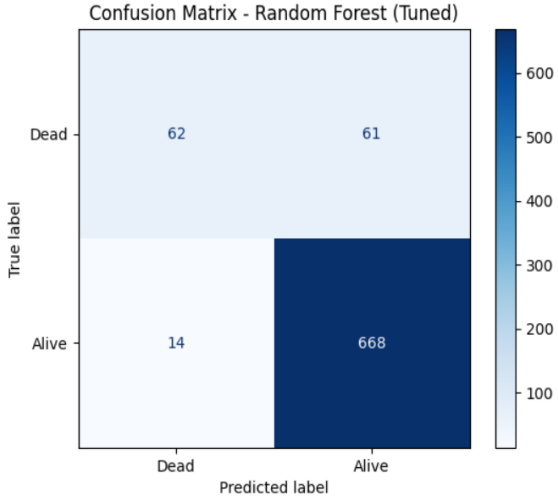




4. Random Forest (with Hyperparameter Tuning):

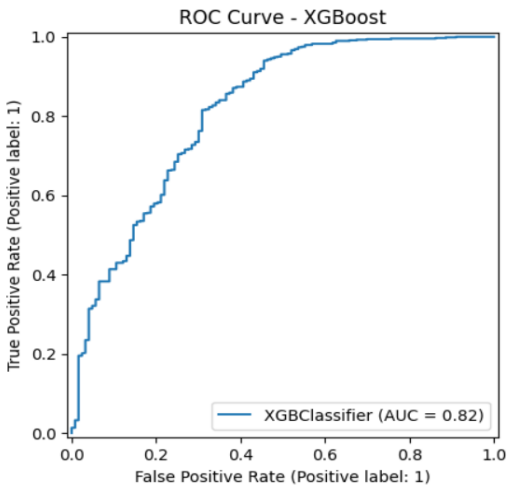
Definition: An ensemble of multiple decision trees built on different subsets of data and features, using bagging and majority voting.
Accuracy: 90.68

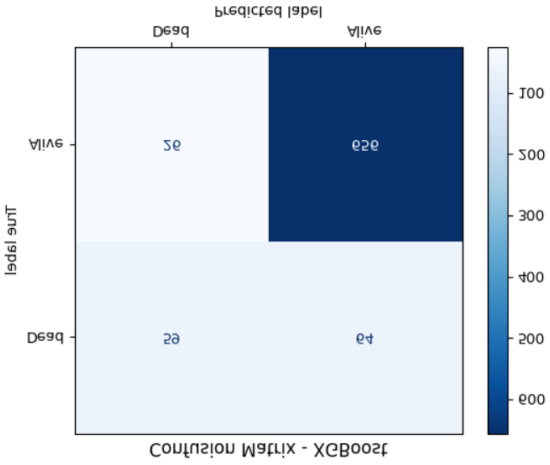




5. XGBoost Classifier:

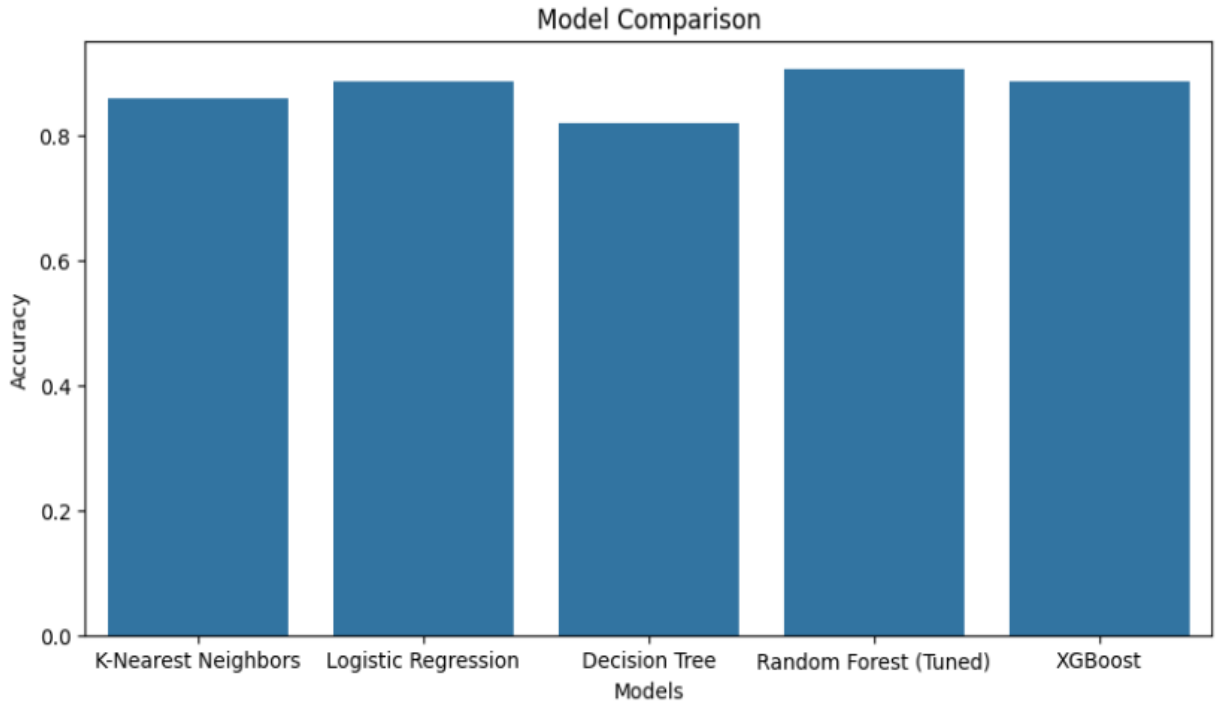
Definition: A gradient boosting-based ensemble model that builds trees sequentially, correcting errors from previous trees.
Accuracy: 88.82%





Model with High Accuracy: Random Forest

Random Forest achieved the highest accuracy of 90.68%. It was tuned using GridSearchCV to optimize parameters such as the number of estimators and maximum depth. This model demonstrated balanced precision and recall, supported by the class distribution correction using SMOTE, making it the most robust performer in this study.



6. Results and Discussion

We evaluated each model using metrics like accuracy, precision, recall, and F1-score. Random Forest demonstrated the best overall performance, achieving the 90.68% highest accuracy and F1-score of 0.94. This highlights its superior ability to make balanced and accurate predictions, especially after applying hyperparameter tuning and SMOTE.

Summary of Model Performance

	Accuracy	Precision	Recall	F1-Score	ROC-AUC
K-Nearest Neighbors	0.859627	0.873850	0.975073	0.921691	0.720913
Logistic Regression	0.888199	0.896783	0.980938	0.936975	0.870229
Decision Tree	0.819876	0.902549	0.882698	0.892513	0.677121
Random Forest (Tuned)	0.906832	0.916324	0.979472	0.946846	0.829495
XGBoost	0.888199	0.911111	0.961877	0.935806	0.815279

Key Takeaways:

- Random Forest (tuned) has achieved the best performance with 90.68% accuracy and high F1-score, making it the most effective model for predicting breast cancer survival.
- Data preprocessing involved label encoding, Min-Max scaling, and SMOTE to handle class imbalance, ensuring clean and balanced input for model training.
- Feature importance analysis highlighted key predictors like Survival Months, Age, and Tumor Size, aiding model interpretability.
- Multiple ML models (K-NN, Logistic Regression, Decision Tree, Random Forest(Tuned), XGBoost) were compared using metrics like precision, recall, F1, and ROC-AUC to ensure reliable evaluation.

7. Conclusion

In this project, we aimed to predict the survival status (Alive or Dead) of breast cancer patients using clinical features. After cleaning, encoding, balancing the dataset with SMOTE, and evaluating multiple machine learning models, we found the Random Forest (with hyperparameter tuning) to perform the best, achieveing an accuracy of around 91%.

Future Work:

1. Include more clinical or genomic features if available.
2. Apply advanced ensemble techniques like Gradient Boosting or LightGBM.
3. Perform longitudinal prediction using time-series modeling if data is collected over time.

References:

1. **Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T.** (2016). *Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis*. Procedia Computer Science, 83, 1064–1069.
2. **Delen, D., Walker, G., & Kadam, A.** (2005). *Predicting breast cancer survivability: A comparison of three data mining methods*. Artificial Intelligence in Medicine, 34(2), 113–127.
3. **Pedregosa, F., et al.** (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
4. Dataset Information: Breast Cancer Dataset (Kaggle). https://www.kaggle.com/datasets/reihanenamdari/breast-cancer