# README - METABRIC Breast Cancer Survival Analysis

This document provides an overview and guide for the METABRIC Breast Cancer Survival Analysis project. The analysis integrates clinical variables, gene expression, and mutation data from the METABRIC dataset to explore mortality prediction in breast cancer patients using machine learning and statistical techniques.

## 1. Project Overview

This project aims to identify genomic and clinical predictors of breast cancer-specific mortality using the publicly available METABRIC dataset. We utilized data preprocessing techniques such as imputation and normalization, dimensionality reduction using Principal Component Analysis (PCA), and modeling using Logistic Regression, Random Forest, and XGBoost. The results were visualized through various plots including violin plots, heatmaps, and volcano plots. Class imbalance was handled with SMOTE to ensure fair model training.

## 2. Repository Structure

The repository is organized as follows:

/project\_root/  
├── README.docx # Overview of the project (this document)  
├── code/  
│ ├── preprocessing.R # Scripts for data loading and cleaning  
│ ├── modeling.R # Model training and evaluation scripts  
│ └── visualization.R # Code for all plots and visual summaries  
├── data/  
│ └── METABRIC\_RNA\_Mutation.csv # Primary dataset used in this project  
├── results/  
│ ├── confusion\_matrices.txt # Confusion matrices from models  
│ └── auc\_scores.csv # AUC scores for all model evaluations  
└── plots/  
 ├── Figure\_1.png to Figure\_20.png # Saved figures used in the report

## 3. Methodologies Used

- Data Harmonization: Missing value treatment via median/mode imputation.  
- Dimensionality Reduction: PCA performed on standardized expression data.  
- Modeling: Built classification models using Logistic Regression, Random Forest, and XGBoost.  
- Class Imbalance: Addressed using SMOTE (Synthetic Minority Oversampling Technique).  
- Visualization: Generated 20 figures including bar plots, box plots, ROC curves, violin plots, volcano and UpSet plots.

## 4. Dataset Information

The METABRIC dataset used in this study contains gene expression, mutation, and clinical profiles of approximately 1980 breast cancer patients. It is publicly available on Kaggle:  
https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric

## 5. Execution Environment

The code was executed in R (version 4.3.3), using packages such as tidyverse, ranger, xgboost, pROC, FactoMineR, DMwR, pheatmap, ComplexUpset, and caret.

## 6. Author

Prepared by: Deepak Sunkavalli  
Saint Louis University, MS Bioinformatics  
May 2025