Comprehensive and Detailed Project Report for Breast Cancer Survival Prediction

1. Project Overview

The Breast Cancer Survival Prediction project aims to analyze patient data to predict survival outcomes. This involves extensive data exploration, preprocessing, feature engineering, and model training to build and optimize machine learning models. The primary focus is to provide an accurate and efficient survival prediction system.

---

2. Dataset Analysis

2.1 Dataset Summary

- **Dataset Size**: 4024 rows × 16 columns.
- **Features**:
  - 5 numeric columns (`Age`, `Tumor Size`, `Regional Node Examined`, `Regional Node Positive`, `Survival Months`).
  - 11 categorical columns (`Race`, `Marital Status`, `T Stage`, etc.).
- **Target Variable**: `Status` (Binary: `1` for Alive, `0` for Dead).

2.2 Data Characteristics

- **Missing Values**: 1 missing value in `A Stage` column.
- **Unique Values per Feature**: Example:
  - `Age`: 40 unique values.
  - `Grade`: Includes an outlier value `anaplastic; Grade IV`.

2.3 Statistical Overview

- Age:
  - Mean: 53.97, Range: 30–69.
- Tumor Size:
  - Mean: 30.47, Range: 1–140.
- Survival Months:
  - Mean: 71.30, Range: 1–107.

2.4 Data Imbalance

- `Status` column: Requires balancing due to potential skewness in survival outcomes.

---

3. Data Preprocessing and Feature Engineering

3.1 Data Cleaning

- Replaced `Grade` outlier (`anaplastic; Grade IV`) with `4`.
- Handled missing data:
  - Rows with missing values were dropped.

3.2 Feature Engineering

- **Label Encoding**:
  - Converted categorical features to numeric using `LabelEncoder`.
- **Normalization**:
  - Scaled numeric features using `MinMaxScaler` to the range [0, 1].
- **Principal Component Analysis (PCA)**:
  - Reduced dimensions to 5 principal components, explaining significant variance.

---

4. Exploratory Data Analysis

4.1 Correlation Analysis

- Generated heatmaps for feature correlation.
- Observed reduced multicollinearity after preprocessing.

4.2 Visualization

- Created boxplots to detect outliers.
- Pairplots to explore relationships between numerical features.

---

5. Model Training and Evaluation

5.1 Models Used

1. **Linear Regression**:
    - **Train R^2 Score**: 0.1181.
    - **Test R^2 Score**: 0.1127.
    - **Mean Squared Error**: 0.1149.
2. **Logistic Regression**:
    - **L1 Regularization** (Lasso): Accuracy = 85.33% (Train), 86.33% (Test).
    - **L2 Regularization** (Ridge): Accuracy = 85.33% (Train), 86.33% (Test).
3. **Support Vector Machine (SVM)**:
    - Default Accuracy: 86.21% (Test).
    - Optimized Accuracy: 86.21% (Test).
4. **Random Forest**:
    - Default Accuracy: 86.46% (Test).
    - Optimized Accuracy: 87.08% (Test).

5.2 Model Optimization

- **Hyperparameter Tuning**:
    - Used `GridSearchCV` for SVM and Random Forest.
    - SVM Best Parameters: `{C: 1, gamma: 'scale', kernel: 'rbf'}`.
    - Random Forest Best Parameters: `{max_depth: 30, max_features: 'sqrt', n_estimators: 100, min_samples_split: 10}`.

5.3 Confusion Matrices

- Analyzed precision and recall via confusion matrices for all models.

---

6. Results and Observations

1. **Best Model**:
    - Optimized Random Forest achieved the highest test accuracy: **87.08%**.
2. **Feature Importance**:
    - Random Forest revealed important predictors like `Age`, `Tumor Size`, and `Regional Node Positive`.

---

7. Deployment

- **Saved Models**:
  - Linear Regression: `linear_regression_model.joblib`.
  - Lasso Logistic Regression: `logistic_regression_lasso_model.joblib`.
  - Ridge Logistic Regression: `logistic_regression_ridge_model.joblib`.
  - Support Vector Classifier: `svm_model.joblib`.
  - Optimized Random Forest: `optimized_rf_model.joblib`.
- **Usage**:
  - Models can be loaded using `joblib` for predictions on new data.

---

8. Challenges and Future Improvements

1. **Data Limitations**:
   - Address class imbalance to improve generalization.
2. **Feature Exploration**:
   - Investigate non-linear relationships using advanced techniques like XGBoost.
3. **Deployment Pipeline**:
   - Develop a Flask or FastAPI-based web service for real-time predictions.

---

9. Conclusion

This project successfully implemented machine learning techniques for breast cancer survival prediction. While the optimized Random Forest model demonstrates high accuracy, future improvements in feature engineering and data augmentation could further enhance predictive performance.