Breast Cancer Prediction Using ML

1. Project Overview

This project aims to develop a predictive model for breast cancer diagnosis using features derived from digitized images of fine needle aspirates (FNA) of breast masses. The goal is to accurately classify tumors as malignant or benign based on the characteristics of cell nuclei present in the images. The dataset for this project is sourced from the UCI Machine Learning Repository and the UW CS FTP server.

2. Data Preprocessing

2.1. Dataset Exploration and Cleaning

- **Loading the Dataset:**
    1. The UCI Machine Learning Repository obtained the dataset, comprising 569 instances and 30 features.
    2. Features include mean, standard error, and worst values for attributes like radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

- **Checking for Duplicate Rows and Columns:**
    1. **Duplicate Rows:** No duplicate rows were found.
    2. **Duplicate Columns:** No duplicate columns were identified.

- **Handling Missing Values:**
  There were all the values in the dataset.

- **Correlation Analysis:**
    1. A correlation matrix was computed to identify highly correlated features.
    2. Features highly correlated with each other or the target variable (diagnosis) were noted for further consideration.

2.2. Feature Engineering

- **Standardization:**
  All features were standardized using `StandardScaler` to ensure they have a mean of 0 and a standard deviation of 1. Standardization is essential for algorithms sensitive to feature scaling, such as SVM and logistic regression.

- **Principal Component Analysis (PCA):**
    1. PCA was applied to reduce the dimensionality of the dataset while retaining at least 90% of the variance.
    2. This resulted in selecting a subset of principal components that captured the majority of the variance, simplifying the model and improving computational efficiency.

3. Exploratory Data Analysis (EDA)

- **Distribution of Classes:**
  The dataset contains 357 benign and 212 malignant cases, showing an imbalance towards benign cases.

- **Visualization:**
  1. **Histograms:** Displayed the distribution of each feature, revealing their range and skewness.
  2. **Boxplots:** Highlighted outliers and the spread of feature values.
  3. **Scatter Plots:** Showed the relationship between pairs of features, helping to identify patterns and correlations.
  4. **Correlation Heatmap:** Visualized the correlation matrix, indicating multicollinearity among some features.

- **Key Findings:**
  1. Features like mean radius, mean texture, and mean perimeter showed significant differences between benign and malignant cases.
  2. High correlations were found between features like mean radius and mean perimeter, suggesting redundancy.

4. Model Implementation

4.1 Model Selection

Several classification models were evaluated:

- **Logistic Regression:**
  A baseline model that provides interpretability and handles binary classification well.

- **K-Nearest Neighbors (KNN):**
  A non-parametric method used for classification, where the majority class determines the output among the k-nearest neighbors.

- **Support Vector Machine (SVM):**
  Effective in high-dimensional spaces and cases where the number of dimensions exceeds the number of samples.

- **Random Forest:**
  An ensemble method that uses multiple decision trees to improve classification accuracy.

- **Gradient Boosting:**
  An iterative method that builds trees sequentially to minimize the loss function.

4.2. Performance Metrics

Model performance was evaluated using:

- **Accuracy:** The proportion of correctly classified instances among the total instances.
- **Precision:** The ratio of true positive predictions to the total predicted positives.
- **Recall:** The ratio of true positive predictions to all actual positives.
- **F1 Score:** The harmonic mean of precision and recall, providing a balance between the two metrics.

4.3. Hyperparameter Tuning

- **Grid Search Cross-Validation (CV):**
    1. Employed 3-fold cross-validation to tune hyperparameters for each model.
    2. Parameters such as the number of neighbors (KNN), C parameter (SVM), and number of trees (Random Forest) were optimized.

4.4. Ensemble Methods

- **Stacking:**
  Combined multiple base learners (e.g., logistic regression, SVM, Random Forest) using a meta-learner (e.g., logistic regression) to improve predictions.

- **Voting:**
  Combined the predictions of several models (e.g., logistic regression, KNN, SVM) through majority voting, where the final prediction is based on the majority class predicted by individual models.

4.5. Final Model

The stacking model was selected as the final model due to its superior performance.

5. Model Performance

5.1. Final Model Metrics

- **Accuracy:** 0.99123
- **Precision:** 1.0
- **Recall:** 0.97619
- **F1 Score:** 0.98795

5.2. Confusion Matrix

The confusion matrix for the final model showed:

- **True Positives (TP):** High
- **True Negatives (TN):** High
- **False Positives (FP):** Low

- **False Negatives (FN):** Low

5.3. ROC Curve

The ROC curve for the final model demonstrated a high area under the curve (AUC), indicating excellent model performance. The AUC score was close to 1, reflecting the model's ability to distinguish between malignant and benign cases effectively.

6. Challenges Faced

1. **Class Imbalance:**
   The dataset had more benign cases than malignant ones, which could lead to biased model predictions. Techniques such as oversampling the minority class (malignant cases) and using balanced class weights were employed to mitigate this issue.

2. **Feature Correlation:**
   A high correlation between features necessitated the use of PCA to reduce dimensionality and multicollinearity. This step was crucial in simplifying the model and improving its generalization.

3. **Hyperparameter Tuning:**
   Extensive hyperparameter tuning was required to optimize model performance. This process was computationally intensive and time-consuming, necessitating the use of efficient cross-validation techniques.

7. Conclusion

The project successfully developed a highly accurate model for breast cancer prediction using machine learning. The final stacking model achieved excellent performance metrics, demonstrating its ability to accurately classify tumors as malignant or benign. Further improvements could include exploring additional ensemble methods and incorporating more advanced feature selection techniques to enhance model performance and robustness.