Analysis and Prediction of Breast Cancer Diagnoses

Project Overview
This project aims to enhance the accuracy of breast cancer diagnosis using machine learning techniques. By employing Support Vector Machine (SVM) algorithms on a dataset with reduced dimensionality through Principal Component Analysis (PCA), the project seeks to predict whether a tumor is benign or malignant based on several measured features. The objective is to provide a reliable diagnostic tool that can assist medical professionals in making informed decisions.

Data Source
The dataset utilized in this project is from a publicly available source containing features derived from digitized images of breast mass tissue samples. The dataset includes various attributes such as radius, texture, perimeter, and area of the tumor, alongside smoothness, compactness, and symmetry indices.
Data set used in this project :
https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset

Data Processing
- Preprocessing: The data was first examined for missing values, and none were found. This ensured that further analysis could proceed without the need for imputation.
- Normalization/Standardization: Prior to PCA, data standardization was applied to normalize the feature scales, ensuring that PCA accurately captures the variance.
- Dimensionality Reduction: PCA was implemented to reduce the dataset to three principal components, capturing the most significant variance in the data with fewer dimensions. This reduction was aimed at improving model performance and computational efficiency.

Exploratory Data Analysis (EDA)
- Statistical Summary: Descriptive statistics provided insights into the central tendencies and dispersion of the data.
- Visualization: scatter plots were used to understand the distributions and relationships of variables. A heatmap of the correlation matrix was also generated to identify multicollinearity.
- Outcomes: The EDA phase revealed key patterns and anomalies in the data, guiding the feature selection for model building.

Feature Engineering
No new features were created for this project as PCA was used to transform the feature space into principal components.

Model Building

- Model Selection: A Support Vector Machine (SVM) model was chosen for its effectiveness in binary classification problems, especially with a high-dimensional space.
- Training: The model was trained on the principal components of the training data, with the target variable being the diagnosis ('M' for malignant, 'B' for benign).
- Parameters: The initial SVM model used a linear kernel with default parameters.

Model Evaluation

- Performance Metrics: The model achieved an accuracy of approximately 97.37% on the test set. Precision, recall, and F1-score were also computed to assess the model's performance across different aspects:
- Precision: High precision indicates a low false positive rate.
- Recall: High recall indicates a low false negative rate.
- F1-Score: Harmonic mean of precision and recall, indicating balance between the two.
- Confusion Matrix: Provided detailed insight into the true positives, true negatives, false positives, and false negatives.

Conclusions and Recommendations

The SVM model demonstrated high accuracy and robustness in classifying breast cancer tumors based on PCA-reduced features. Future work could explore alternative modeling techniques like Random Forest or Gradient Boosting for comparative analysis. Additionally, further tuning of SVM parameters and cross-validation could enhance model reliability and generalizability.