BREAST CANCER DIAGNOSIS ANALYSIS

Breast cancer is one of the most prevalent cancers affecting women worldwide. Early detection and accurate diagnosis are crucial for effective treatment and improved outcomes. This report presents an exploratory data analysis (EDA) and principal component analysis (PCA) of the Breast Cancer dataset while comparing various types of machine learning algorithms for accurately diagnosing. The objective of the project was to build a machine learning model to predict whether a tumour is malignant or benign. Through EDA, various visualizations and summary statistics were utilized to gain insights into the dataset. PCA was employed to reduce the dimensionality of the data and uncover latent variables.

Dataset:

Link: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

The Breast Cancer dataset contains features computed from digitized images of breast mass samples, aiming to predict whether a tumour is malignant or benign. The dataset consists of 569 samples and 30 features, including various measurements of tumour characteristics such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave point, symmetry, and fractal dimension.

Exploratory Data Analysis (EDA):

- Investigated the distribution of the target variable 'diagnoses' to understand the class balance between malignant(M) and benign(B) cases.
- Explored the relationship between variables through correlation coefficient heatmap.
- Visualized the pairwise relationships between the target variable and selected numerical features to identify potential patterns or clusters.

Principal Component Analysis (PCA):

- Principal Component Analysis (PCA) is a dimensionality reduction technique commonly used to reduce the number of features in a dataset while preserving the most important information. It works by transforming the original features into a new set of orthogonal variables called principal components.
- The Scree plot helps in determining the optimal number of principal components to retain. In this case, the Scree plot indicates that retaining six principal components captures a significant amount of variance in the data.
- Finally, PCA transformation is performed on the original data, reducing its dimensionality to the selected number of principal components.

Model Building and Comparison:

1. **Logistic Regression:**
    - Trained a logistic regression model using the reduced feature set obtained from PCA.
    - Achieved an accuracy of 98.25% on the test set with precision, recall, and F1-score of 0.97, 1.00, and 0.99, respectively.

2. **K-Nearest Neighbors (KNN):**
    - Employed a KNN classifier with k=5 neighbours.
    - Attained an accuracy of 96.49% on the test set, demonstrating precision, recall, and F1-score of 0.96, 0.99, and 0.97, respectively.
3. **Random Forest:**

- Utilized a Random Forest classifier and yielded an accuracy of 95.61% on the test set, with precision, recall, and F1-score of 0.96, 0.97, and 0.97, respectively.

4. **Support Vector Machine (SVM):**
    - Trained an SVM classifier with probability estimation enabled and achieved an accuracy of 96.49% on the test set, demonstrating precision, recall, and F1-score of 0.96, 0.99, and 0.97, respectively.
5. **XGBoost:**
    - Employed an XGBoost classifier and achieved an accuracy of 96.49% on the test set, demonstrating precision, recall, and F1-score of 0.96, 0.99, and 0.97, respectively.

Conclusion:

- The evaluation of various classifiers demonstrates their effectiveness in accurately distinguishing between benign and malignant breast cancer tumour.
- Logistic regression exhibits the highest accuracy among the models evaluated in this study, making it a promising candidate for practical implementation in diagnosing breast cancer.