**Report**

https://github.com/stevensct/Breast-Cancer-Prediction-Analysis.git

**Introduction**

Breast cancer remains a significant public health concern, necessitating the development of effective diagnostic tools. This project aims to analyse the characteristics of breast cell nuclei to distinguish between malignant and benign cases. By utilizing machine learning models and dimensionality reduction techniques, we aim to uncover hidden patterns within cell nucleus characteristics to reveal possible variations in cell nucleus characteristics relevant for diagnosis and treatment.

**Methods**

**Imported Modules**

To carry out this analysis, several Python libraries were used, including:

* numpy: For numerical processes.
* pandas: For data manipulation and analysis.
* os: For interacting with the operating system.
* seaborn: For statistical data visualization.
* matplotlib: For creating plots and visualizations.
* scikit-learn: For machine learning algorithms.
* statsmodels: For statistical modeling.

**Data Exploration and Pre-Processing**

The dataset used in this analysis is the Breast Cancer Wisconsin (Diagnostic) Data Set. It contains 569 instances and 32 attributes, including one ID column, one diagnosis column, and 30 real-valued input features. These features are derived from digitized images of fine needle aspirates (FNA) of breast masses, characterizing the cell nuclei present in the images. The diagnosis column categorizes the tumours as either malignant (M) or benign (B). The class distribution includes 357 benign cases and 212 malignant cases, with no missing values reported.

The dataset was uploaded and thoroughly explored. Irrelevant columns were removed to ensure the focus remained on features directly contributing to the classification task. Data cleaning and normalization steps were also performed to prepare the dataset for analysis.

**Correlation Analysis**

A correlation analysis was performed to examine relationships between cell nucleus features and their relevance to tumour diagnosis.

**Principal Component Analysis (PCA)**

PCA was applied to reduce dimensionality and address multicollinearity. This technique transforms high-dimensional data into a lower-dimensional space by identifying principal components that capture the most significant variations. Using the elbow method, five principal components were selected, collectively explaining 100% of the variance.

**Machine Learning Models**

Several classification models were implemented and evaluated, including:

* Logistic Regression: A statistical model predicting binary outcomes based on input features.
* Decision Trees: A flowchart-like structure for decision-making and classification.
* K-Nearest Neighbours (KNN): A non-parametric algorithm classifying data points based on their neighbours.
* Random Forest: An ensemble learning method combining multiple decision trees for improved accuracy and robustness.

**Results**

**Correlation Analysis**

The heatmap analysis highlighted features such as size, shape, texture, compactness, and concavity as strongly associated with malignancy, while smoothness and fractal dimension had minimal impact. There was significant overlap among predictors, so dimensionality reduction was needed to improve model performance.

**PCA Results**

Dimensionality reduction through PCA revealed that five principal components were sufficient to capture the entire variance in the data.

**Model Performance**

All models demonstrated high accuracy, achieving 96% across metrics such as accuracy, precision, recall, and F1 score. Logistic Regression and Random Forest models slightly outperformed others by reducing false negatives, a critical factor for early detection and effective treatment of breast cancer.

**Summary**

This analysis successfully identified key features influencing breast cancer diagnosis and demonstrated the effectiveness of machine learning models in classification tasks. The application of PCA improved model performance by addressing multicollinearity. These findings underscore the potential of integrating machine learning techniques into clinical workflows to enhance diagnostic accuracy and patient outcomes.