Breast Cancer Detection Using Machine Learning - Final Project Report

Project Overview:

Breast cancer is one of the most common and life-threatening diseases affecting women globally. Early and accurate detection is crucial for effective treatment and improved survival rates. This project aims to develop a robust machine learning model for classifying tumors as malignant or benign using diagnostic features derived from digitized images of fine needle aspirates (FNAs) of breast masses.

Dataset Summary:

- **Source**: Breast Cancer Wisconsin (Diagnostic) Dataset
- **Instances**: 569 samples
- **Features**: 32 (including ID and diagnosis label)
- **Target Variable**: diagnosis (M = malignant, B = benign)

Business Case:

To develop an automated system that assists medical professionals in diagnosing breast cancer more quickly and accurately by leveraging machine learning techniques. This helps in improving the speed and precision of diagnosis, leading to early intervention and better patient outcomes.

Data Processing Workflow:

1. Uploading the Dataset

- Imported the CSV file containing the dataset into the working environment.
- Checked the file for proper formatting and loaded it using pandas.read_csv().

2. Data Understanding

- Reviewed all 32 columns, focusing on feature definitions and their relevance to cancer diagnosis.
- Dropped non-informative columns like id.
- Converted the target variable diagnosis into binary format (M=1, B=0).

3. Exploratory Data Analysis (EDA)

- **Descriptive Statistics**: Used to describe and summarize the central tendency and spread of each feature.

- **Univariate Analysis**: Used histplot and countplot to explore the distribution of individual features and class imbalance.
- **Multivariate Analysis**: Generated a correlation matrix and heatmap to identify relationships between features.

4. Handling Multicollinearity

- Removed features with high correlation ($r > 0.9$ or $r < -0.9$) to reduce dimensionality and avoid redundancy.
- Dropped 10 features including perimeter_mean, area_mean, concave points_mean, perimeter_se, and others.

5. Data Cleaning and Preprocessing

- Identified corrupted values (e.g., 0 in features like concavity_mean, which should not be zero).
- Replaced corrupted values with the median of respective columns to maintain data integrity.

6. Outlier Handling

- Applied **IQR-based Capping**:
  - Calculated Q1, Q3, and IQR.
  - Capped values outside [Q1 - 1.5*IQR, Q3 + 1.5*IQR] using np.clip().
  - This technique ensures that extreme outliers do not distort model training.

7. Feature Scaling

- Used StandardScaler to normalize all numeric features so that they have a mean of 0 and a standard deviation of 1.

Model Development & Evaluation:

Models Applied:

1. **Logistic Regression**
2. **K-Nearest Neighbors (KNN)**

Train-Test Split:

- Split the dataset into 75% training and 25% testing subsets using train_test_split().

Evaluation Metrics:

Metric	Logistic Regression	KNN
Accuracy	95%	90%
Precision	94%	93%
Recall	94%	81%
F1-Score	94%	87%

Best Model:

- **Logistic Regression** was selected as the best-performing model.
- It showed superior metrics in accuracy, precision, recall, and F1-score.
- Offers interpretability, making it ideal for real-world medical applications.

Final Conclusion:

This project presents a comprehensive pipeline for breast cancer detection using diagnostic features. By applying rigorous EDA, cleaning, preprocessing, and model evaluation techniques, we were able to achieve:

- High model accuracy (95%) using Logistic Regression
- Robust handling of multicollinearity and outliers
- Effective preprocessing, including median imputation and feature scaling

Logistic Regression stood out due to its performance and simplicity. It is thus recommended as the best model for practical deployment in clinical settings. This project highlights how machine learning can significantly enhance diagnostic workflows, ultimately contributing to better patient care.