Subject: Machine learning

Mini-Project Report for Breast cancer diagnosis Wisconsin dataset

Problem Statement:
Breast cancer diagnosis is a critical task in healthcare, requiring early and accurate identification of malignant tumors to improve patient outcomes. This project focuses on building a classification model to predict whether a breast tumor is benign or malignant based on various diagnostic features. By applying machine learning algorithms, we aim to support healthcare professionals in making efficient preliminary diagnoses, reducing the need for invasive procedures.

Dataset:
The dataset is the Breast Cancer Diagnosis Wisconsin dataset from the UCI Machine Learning Repository. It includes **569 samples** with **30 continuous features** derived from cell nucleus images, such as:

- **Radius** (mean of distances from the center to points on the perimeter) •

Texture (standard deviation of gray-scale values)

- **Perimeter**
- **Area**
- **Smoothness** (local variation in radius lengths)
- **Compactness**, **Concavity**, **Symmetry**, among others. The target variable is binary, where **1** represents a **malignant** tumor, and **0** represents a **benign** tumor. The dataset was preprocessed by filling missing values and normalizing feature values.

Methodology:

1. **Data Preprocessing**:
    - **Handling Missing Values**: Replaced any missing values with mean/mode as appropriate.
    - **Normalization**: Standardized feature values to a similar scale, critical for algorithms like KNN and Logistic Regression, which are sensitive to feature scales.
2. **Feature Selection**: Used correlation analysis and domain knowledge to ensure relevant features were used, enhancing model interpretability and reducing computation.
3. **Train-Test Split**: Split the data into **80% training** and **20% testing** sets to evaluate model performance.

4. **Model Training and Evaluation**: Trained each classification model on the training set, then used the testing set to evaluate performance, focusing on accuracy, precision, recall, and F1-score for a comprehensive assessment.

Selected Techniques:

1. **Logistic Regression (LR)**: Efficient for binary classification, LR models the probability of a tumor being malignant, providing quick, interpretable results.
2. **Naive Bayes (NB)**: This probabilistic model is based on Bayes' theorem. It assumes feature independence and is fast, though sometimes limited in performance due to its simple assumptions.
3. **K-Nearest Neighbors (KNN)**: A non-parametric method that classifies samples based on the closest K neighbors. It performs well with normalized data and smaller datasets.
4. **Decision Tree (DT)**: Uses a tree-like structure to make decisions, making it highly interpretable and capable of handling non-linear relationships.
5. **Random Forest (RF)**: An ensemble technique combining multiple decision trees, which improves accuracy and robustness by reducing overfitting compared to single decision trees.
6. **K-means Clustering**: An unsupervised method used here as a benchmark. After clustering, we mapped clusters to labels for comparison, although it inherently lacks the label-based learning of other methods.

Results and Conclusion:

The table below shows the performance of each classification model based on metrics like recall, precision, F1-score, and accuracy in csv file

Conclusion:

The Random Forest model achieved the highest performance, with **96% accuracy and an F1- score of 0.96**, making it the most suitable for breast cancer diagnosis among the algorithms evaluated. Logistic Regression also performed well, proving to be a reliable, interpretable model. K-means, used as an unsupervised benchmark, demonstrated lower performance, highlighting the importance of supervised learning for this task. Overall, Random Forest and Logistic Regression are recommended for accurate and robust predictions in breast cancer diagnosis.