NEXUS INTERNSHIP

Data science intern

Name: Manoj Prabakar B

Ph no: 9843172255

Email ID: manojrockerguy246@gmail.com

Project Phase 2: Breast Cancer Prediction

1. Data Preprocessing:

- Clean and preprocess the Breast Cancer Wisconsin (Diagnostic) dataset.

- Handle missing values, outliers, and any other inconsistencies in the data.

2. Feature Selection and Engineering:

- Identify relevant features for breast cancer prediction.

- Create new features or transformations that might enhance the predictive model's performance.

3. Machine Learning Model (SVM):

- Implement a Support Vector Machine (SVM) model for classifying tumors into malignant or benign.

- Train and evaluate the model on the Breast Cancer dataset.

4. Documentation:

- Document your data preprocessing, feature selection, and machine learning model implementation.

- Explain the model's performance metrics and any challenges faced during the

analysis.

1. Introduction:

This documentation provides a step-by-step explanation of data preprocessing, feature selection, and machine learning model implementation for the Breast Cancer Prediction dataset. I will also provide the model's performance metrics and the challenges faced during the analysis.

2. Data Preprocessing:

Data preprocessing involves cleaning and transforming raw data into a suitable format for analysis. The steps include loading data, handling missing values, encoding categorical variables, and scaling features.

Steps:

- Loading Data
- Understanding the Data
- Dropping Unnecessary Columns

- Encoding Categorical Data
  - The 'diagnosis' column is categorical and needs to be converted to numerical values for machine learning algorithms to process it.

- Splitting the Dataset
- Feature Scaling:
  1. Standardizing the features to have a mean of 0 and a standard deviation of 1.

- Exploratory Data Analysis (EDA)

  EDA involves visualizing the data to understand its structure and relationships between features.

  1. Heatmap of Correlations
  2. Scatter Plots for Correlated Features
  3. Count Plot of Target Variable

- Machine Learning Model Implementation

- Implementing and evaluating different machine learning models to classify the diagnosis.
    1. Support Vector Classifier
    2. Model Evaluation:
       -Evaluating the performance of the model using various metrics.

1. Accuracy Score

2. Precision and Recall

3. Classification Report

4. Confusion Matrix

5. ROC curve

Model Performance Metrics:

1. **Accuracy**: Measures the ratio of correctly predicted instances over the total instances.
2. **Precision**: The ratio of correctly predicted positive observations to the total predicted positives.
3. **Recall**: The ratio of correctly predicted positive observations to the all observations in actual class.
4. **F1-Score**: The weighted average of Precision and Recall.
5. **ROC AUC Score**: Measures the ability of the classifier to distinguish between classes.

Challenges Faced:

1. **Handling Imbalanced Data**: The dataset may have more benign cases than malignant cases, which can bias the model.
2. **Feature Scaling**: Ensuring that features are on a similar scale to improve the performance of the model.
3. **Hyperparameter Tuning**: Finding the optimal parameters for each model to achieve the best performance.
4. **Model Evaluation**: Selecting appropriate metrics to evaluate the performance of the model, especially in the context of medical diagnoses where precision and recall are critical.