Project 1 Report – Breast Cancer Recurrence

Data Preparation:

1. Handling Missing Values: Replaced missing values represented by '?' and '*' with NaN. Replaced NaN or null values with mode of respective column.
2. Categorical Data Identification: Identified categorical columns in the dataset.
3. One-Hot Encoding: Applied one-hot encoding to categorical columns with more than two unique values.
4. Splitting Data: Split the data into features (X) and target (y), and then into training and testing sets.
5. Standardization: Standardized the features to have a mean of 0 and a standard deviation of 1.

Insights from Data Preparation:

1. Class Imbalance: Observed the distribution of the target variable to check for class imbalance.
2. Feature Importance: Identified which features are categorical and need encoding, ensuring that the model can handle them correctly.
3. Exploratory Data Analysis:
    a. Box Plots: To identify outliers and understand the distribution of numerical features.
    b. Histograms: To visualize the frequency distribution of numerical features and understand their shapes.
    c. Count Plots: To visualize the distribution of categorical features and identify any imbalances.

Model Training Procedure:

1. Model Initialization: Initialized the K-Nearest Neighbor (KNN), KNN with Grid Search CV (to optimize k value), and Logistic Regression models.
2. Model Training: Trained the models on the standardized training data. Fitted each model on standardized training data.
3. Evaluated Models: accuracy, recall, precision, and F1 score.
4. Optimized hyperparameters: Used Grid Search CV to find the best hyperparameters for the KNN model.

Model Performance:

1. K-Nearest Neighbor Classifier:

```
Accuracy: 0.584070796460177
Classification Report:               precision    recall  f1-score
support

           0       0.67      0.78      0.72        77
           1       0.26      0.17      0.20        36

    accuracy                           0.58       113
   macro avg       0.46      0.47      0.46       113
weighted avg       0.54      0.58      0.55       113

Confusion Matrix: [[60 17]
 [30  6]]
```

2. K-Nearest Neighbor Classifier with Grid Search CV:

```
Best Parameters: {'n_neighbors': 10}
Accuracy: 0.6283185840707964
Classification Report:               precision    recall  f1-score
support

           0       0.67      0.90      0.77        77
           1       0.20      0.06      0.09        36

    accuracy                           0.63       113
   macro avg       0.43      0.48      0.43       113
weighted avg       0.52      0.63      0.55       113

Confusion Matrix: [[69  8]
 [34  2]]
```

3. Logistic Regression:

```
Accuracy: 0.6283185840707964
Classification Report:               precision    recall  f1-score
support

           0       0.68      0.84      0.76        77
           1       0.33      0.17      0.22        36

    accuracy                           0.63       113
   macro avg       0.51      0.51      0.49       113
weighted avg       0.57      0.63      0.59       113

Confusion Matrix: [[65 12]
 [30  6]]
```

Model Confidence:

Moderate confidence in models. Model performance depends on data quality and class balance, so it could improve performance with feature selection, more data, or training on advanced models for accuracy and recall. This model should be validated further before deploying in a real-world scenario. The following were used to measure and evaluate the model's performance.

1. Evaluation Metrics: Evaluated the models using accuracy, precision, recall, and F1-score.
2. Confusion Matrix: Analyzed the confusion matrix to understand the model's performance in terms of true positives, true negatives, false positives, and false negatives.

Most Important Metric:

For predicting recurrence in medical data, recall might be the most important metric because it is crucial to identify as many true positive cases as possible (i.e., correctly identifying patients who will have a recurrence). Missing a recurrence (false negative) could have serious consequences. Thus, in the context of this project, recall seems to be the most important metric as of now.