Project 01

This report provides a comprehensive overview of the exploratory data analysis and classification modeling performed on the dataset. The primary object was to predict the recurrence of breast cancer in patients, given the dataset from the class git repository. The process included analyzing the shape and the size of the raw data, cleaning and handling missing values, featuring transformation, selecting appropriate model, and lastly evaluating the model's overall performance. Understanding the recurrence pattern is crucial for effective treatment planning and improving patient outcomes to cure breast cancer efficiently and more effectively in the future.

Data Preparation:

Loading and Cleaning the Data

- The dataset was loaded from a link-available class git repository and initially inspected for its structure, including its shape, size, and data types.
- Duplicate rows were identified and removed to maintain data integrity and avoid bias in model performance.
- Invalid values such as '?' and '*' were replaced with 'NaN,' making it easier to handle missing data.
- Missing values were imputed using mode imputation for categorical variables to ensure consistency across the dataset.
- Categorical variables were transformed using one-hot encoding, which converts text-based features into numerical format for machine learning models.
- The dataset was split into training (80%) and testing (20%) sets, ensuring class balance through stratified sampling.

Insights from Data Preparation

- The dataset contained missing or invalid values that required careful handling to prevent model bias.
- Certain categorical features had a high number of unique values, necessitating one-hot encoding for proper model processing.
- Class imbalance was observed, meaning a simple accuracy metric would not provide an adequate measure of model performance.
- Feature distributions revealed that some variables had strong correlations with the target variable, which could be used for model interpretability.

Model Training and Evaluation:

Procedure Used to Train the Model

To determine the best classification model for predicting recurrence, the following steps were taken:

- K-Nearest Neighbors (KNN): Implemented as a baseline classifier to assess the relationship between feature proximity and recurrence.

- KNN with GridSearchCV: Applied hyperparameter tuning using cross-validation to optimize n_neighbors parameter and enhance predictive performance.
- Logistic Regression: Used as a linear classification benchmark to evaluate the dataset's suitability for linear separability.
- Performance metrics: The models were assessed using precision, recall. F1-score, and accuracy to ensure a holistic evaluation. Since class imbalance was present, recall was given more importance to minimize false negatives.
    - Recall (sensitivity) is the most critical since false negatives a costly in medical conditions like cancer. If a model incorrectly classifies a patient as non-recurrent when they actually have a recurrence, they may not receive necessary treatment, which can be life-threatening. Also, a model with high recall ensures that most actual recurrence cases are identified, even if it means some false positives.

Model Performance

- The KNN model provided an initial classification framework but exhibited sensitivity to the choice of n_neighbors.
- GridSearchCV optimization improved KNN's performance by selecting the best hyperparameter value, leading to more reliable predictions.
- The Logistic Regression model performed competitively and was useful for interpreting feature contributions.
- Recall was prioritized in evaluation, as false negatives could lead to missed recurrence predictions, which is particularly risky in a medical setting.

Confidence in the Model:

- The models, particularly KNN with GridSearchCV and Logistic Regression, demonstrated the dataset's suitability for classification.
- Recall and F1-score were emphasized over accuracy due to the class imbalance and the potential consequences of false negatives.
- Future iterations may benefit from feature engineering, deeper hyperparameter tuning, and alternative sampling techniques to address class imbalance.

This study successfully applied exploratory data analysis, data preprocessing, and classification modeling to predict medical recurrence. The prioritization of recall ensured that the model focused on correctly identifying recurrence cases. While the models performed reasonably well, future work should explore more advanced techniques to enhance predictive accuracy. By refining feature selection, testing alternative models, and addressing class imbalance, we can further improve the reliability of predictions, ultimately benefiting medical decision-making and patient care.