Part 3: Project 1 Lab Report

Data Preparation

To make the dataset ready for analysis, I carried out several preprocessing steps to ensure it was clean and suitable for training models. First, I checked for and removed any duplicate entries to preserve data accuracy. Next, I addressed missing values by filling them with the most frequently occurring values in each column. Specifically, categorical fields that included special characters like ? or * were considered missing data and replaced as needed.

For numerical attributes such as age, tumor size, and invasive nodes, I converted range-based categories into their midpoint values. This adjustment ensured the data was structured appropriately for model input. Additionally, categorical features—including menopause, node-caps, breast, breast-quad, and irradiat—were transformed using one-hot encoding, making them easier for machine learning models to process.

To maintain a balanced class distribution during model training, I used a stratified sampling approach, reserving 80% of the data for training and 20% for testing. This helped prevent biased learning, especially since cases without recurrence were more common than those with recurrence. To ensure consistency across numerical features, I applied a standardization technique that adjusted their mean to 0 and standard deviation to 1. This step was particularly important for models that rely on distance calculations, such as KNN.

Insights from Data Preparation

During the data preprocessing stage, I uncovered several important patterns within the dataset. One notable finding was the imbalance in class distribution, which directly influenced model predictions. This skew could cause models to favor the more common class (no-recurrence-events), potentially reducing accuracy for the less frequent category.

I also noticed that the majority of patients were between 40 and 60 years old, highlighting this age range as a key factor in predicting recurrence. The distribution of tumor sizes revealed the presence of outliers, suggesting that larger tumors might be associated with a higher likelihood of recurrence.

Applying one-hot encoding expanded the number of features, but this tradeoff was necessary to accurately represent categorical information in a format suitable for machine learning. Lastly, standardizing numerical features significantly improved the effectiveness of KNN, as unscaled data would have led to inaccurate distance calculations and reduced model performance.

Model Training Procedure

Three classification models were trained and evaluated:

- **K-Nearest Neighbors (KNN) with K=5**: This served as a baseline model to measure initial performance without optimization.
- **KNN with Grid Search Cross-Validation**: GridSearchCV was applied to identify the optimal K value, improving model performance by tuning hyperparameters.

- **Logistic Regression**: A simple but effective linear model to serve as an additional baseline comparison.

The dataset was divided into training and test sets using a stratified approach to preserve class distribution. Before training, feature scaling was applied to enhance model performance. To evaluate each model thoroughly, I measured its accuracy, recall, precision, and F1-score, ensuring a well-rounded assessment of its effectiveness.

Model Performance

Among the trained models, KNN showed improved performance when optimized using GridSearchCV to find the best K-value. Logistic Regression delivered similar results, making both models viable options for the task. The performance of each model was evaluated as follows:

- **K-Nearest Neighbors (K=5):**
    - Accuracy: 64%
    - Recall: 12.5%
    - F1-Score: 18.2%
    - The recall for detecting recurrence cases was low, meaning that several cases were missed.
- **Optimized KNN (Best K=19):**
    - Accuracy: 68%
    - Recall: 0%
    - F1-Score: 0%

- Despite an improvement in overall accuracy, the model completely failed to identify recurrence cases.

- **Logistic Regression:**
  - Accuracy: 65%
  - Recall: 20.8%
  - F1-Score: 27.8%
  - This model provided a better balance between recall and accuracy compared to KNN.

Confidence in the Model

Although the models generated reasonable predictions, there is potential for improvement. The class imbalance impacted recall, resulting in some recurrence cases being misclassified. Using a more advanced classifier, such as Random Forest or a Support Vector Machine with a non-linear kernel, could enhance performance. Additional refinements could include:

- **Addressing Class Imbalance:** Collecting more recurrence-event cases or using techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset.
- **Exploring More Advanced Models:** Testing classifiers such as XGBoost or Neural Networks to enhance predictive accuracy.

- **Feature Selection and Engineering:** Identifying the most relevant features or creating new engineered features to improve model effectiveness.
- **Hyperparameter Tuning:** Further optimizing models, particularly for Logistic Regression and SVM, to achieve better results.

Conclusion

The models developed in this study offer a solid starting point for predicting breast cancer recurrence. While KNN and Logistic Regression delivered acceptable results, further enhancements could be achieved by leveraging more advanced classification techniques and addressing class imbalance more effectively. Since recall is crucial in medical applications, future improvements should focus on increasing sensitivity to recurrence cases while maintaining overall predictive reliability.