Part 3: Report

Data Preparation

I applied several preprocessing steps to ensure the dataset was clean and well prepared for training the model. First, any missing values were handled - replacing the special characters (?, *) which NaN, in order to make sure that all data points could be properly accounted for and processed. Then all numerical columns were (such as age, tumor-size, and inv-nodes) modified by having their missing values filled with the median of the values in the column to prevent biases introduced by extreme values or outliers.

Next, I converted any categorical range-based values (e.g., 30-39 for age) to their midpoint values in order to transform them into a numerical representation rather than a string that I could then use for model inputs. Categorical variables such as menopause, node-caps, breast, breast-quad, and irradiat were one-hot encoded, making them into binary variables and allowing the machine learning model to interpret them effectively.

In order to ensure proper class distribution during model training, the dataset was split using stratified sampling, allocating 80% of the data for training and 20% for testing. This helped to maintain the proportion of no-recurrence-events and recurrence-events cases, preventing biased learning. Furthermore, I applied feature standardization using StandardScaler, which scales numerical data to have a mean of 0 and a standard deviation of 1. This was particularly needed for distance-based models like KNN, which are sensitive to differences in feature magnitudes.

Insights from Data Preparation

During data preprocessing, I learned several important insights into the state of the data. These included the class imbalance of the dataset, as there were more cases of no-recurrence-events than recurrence-events. This imbalance had the potential to affect the model's performance, particularly for models that rely on majority class prediction.

The age distribution indicated to me that most patients were between 40-60 years old, making this an essential factor in aiding recurrence predictions. The tumor size distribution showed the presence of outliers, with some patients having significantly larger tumors, which may influence recurrence probabilities.

Implementing one-hot encoding increased the number of features but this inefficiency was worth it for the improved interpretability, which allowed for categorical variables to be represented in a way that the model could process. Standardizing the features also played a significant role in enhancing KNN's performance, as any unscaled features would lead to incorrect distance calculations, thus resulting poor model performance metrics.

Model Training Procedure

Three classification models were trained and evaluated:

- K-Nearest Neighbors (KNN) with K=5: This was used as a baseline model to measure initial performance without hyperparameter tuning.
- KNN with Grid Search Cross-Validation: To optimize the KNN model, GridSearchCV was used to determine the best K value. The search identified K=22 as the optimal number of neighbors.
- Linear Classifier (Support Vector Machine - SVM): This model was trained using a kernel='linear' configuration, making it well-suited for linearly separable data.
- Logistic Regression: Included as an additional baseline model.

Each model's performance was evaluated using accuracy, precision, recall, and F1-score, ensuring a comprehensive assessment of their effectiveness.

Model Performance

Among the four models, SVM performed the best, achieving an accuracy of 71.4%, which outperformed KNN with K=5 (66.2%), KNN with the optimized K=19 (68.8%), and Logistic Regression (64.9%).

The SVM classifier not only had the highest accuracy but also exhibited superior recall, making it the most reliable model. KNN with Grid Search improved over the default KNN model, proving the importance of tuning hyperparameters, but it still underperformed compared to SVM.

In a medical dataset like this, the recall is the most critical metric. A low recall score means that the model is failing to identify cases of recurrence, which can have severe health consequences. Since missing a recurrence prediction could result in delayed treatment, recall must be prioritized over precision or accuracy. The recall values for the models were:

- KNN (K=5): Recall = 0.662
- KNN (Best K=22): Recall = 0.688
- SVM (Linear): Recall = 0.714
- Logistic Regression: Recall = 0.649

Confidence in the Model

I am reasonably confident in this final iteration of the model, however there is still room for improvement, as the SVM model demonstrated strong predictive power with an accuracy of 71.4%. I have considered several enhancement for future iterations that are listed below:

- Addressing Class Imbalance: Collecting more data, particularly for recurrence-events, could improve model generalizability.

- Exploring More Advanced Models: Other classifiers such as Random Forest, XGBoost, or Neural Networks could be tested for better performance.
- Feature Selection and Engineering: Identifying the most relevant features or creating new engineered features could enhance predictive accuracy.
- Hyperparameter Tuning for SVM: Additional tuning of the SVM model (e.g., adjusting the regularization parameter C) might yield further performance improvements.

Given the medical nature of this dataset, false negatives (missed recurrence cases) must be minimized to avoid severe health consequences. Therefore, recall should remain the top priority in future improvements. While the SVM model currently offers the best balance between accuracy and recall, further refinements are necessary to enhance predictive performance and clinical applicability.