Project 2: Breast Cancer Recurrence Prediction with Classification Models

This project aims to use various classification machine learning models trained on a dataset that contains information about breast cancer patients to predict potential recurrence given additional information about the patient.

Before the machine learning models could be implemented, the dataset needed to be preprocessed.The original dataset consisted of 286 patients, containing data on whether the cancer recurred, the patient's age, the patient's menopausal status, the tumor size, the number of invasive nodes, the presence of node capsules, the degree of malignancy, which breast had the tumor, the location of the tumor in the breast, and whether the tumor was irradiated as part of the treatment. The node capsule and tumor location columns had missing values, so these were filled by the mode value of a subset of rows which contained matching column values with the rows that had missing values. Once this was completed, univariate analysis of the data showed that all of the columns were categorical and a majority were strings, so those columns were converted to categorical values via one-hot encoding. The only columns left untouched were the cancer recurrence to keep it as the target variable and the degree of malignancy because it was discrete numeric data (using one-hot encoding here could potentially cause a loss of "distance" between different values). No data columns were dropped, and the cancer recurrence was chosen as the target variable with all other columns becoming the independent features. Finally, the data was shuffled and split into training and test datasets in a reproducible manner, with the training dataset consisting of 70% of the original dataset and the test dataset making up the other 30%. While doing this, care was taken to maintain the proportions of the target variable classes in each subset.

Four classification machine learning models were used to analyze the preprocessed dataset: K-Nearest Neighbors, Logistic Regression, Decision Tree, and Random Forest. While the Logistic Regression and Decision Tree methods had no hyperparameters to optimize, the K-Nearest Neighbors and Random Forest methods did have hyperparameters. To properly optimize these, the scoring function was chosen to maximize the precision (minimize false positives) of the "no-recurrence-events" class. The reasoning behind this was that the possibility of incorrectly diagnosing someone as having no risk of recurrent breast cancer would be more detrimental than the possibility of incorrectly diagnosing someone as at risk of recurrent breast cancer. This scoring function also acts to maximize the recall (minimize false negatives) of the "recurrence-events" class to an extent, but the "no-recurrence-events" perspective was chosen here because it is the larger class (approximately 70% of the dataset).

The important values for each model's performance can be found in the table below:

	Test Precision	Training Precision	Test Accuracy	Training Accuracy
K-Nearest Neighbors	0.75	0.85	0.67	0.80
Logistic Regression	0.72	0.80	0.70	0.79
Decision Tree	0.73	1.00	0.64	0.98
Random Forest	0.77	0.98	0.76	0.98

Table 1: Performance Metrics for Analyzed Classification Models

Here, precision of "no-recurrence-events" for the test dataset was primarily used to determine overall performance due to previously mentioned reasoning, while the model accuracy was compared between the test and training datasets to find potential overfitting. We can see that the Decision Tree model falls into this overfitting trap; despite having the highest high precision and

accuracy values for the training dataset, its accuracy on the test dataset is the lowest of the four models. On the other hand, the Random Forest dataset performed the best overall, managing to fit the training dataset extremely well while still maintaining the highest precision and accuracy scores on the test dataset. K-Nearest Neighbors and Decision Tree performed similarly on the test dataset, with the former being marginally more precise and the latter being marginally more accurate. Based on these results, I would primarily recommend that the Random Forest model be used for this dataset, with K-Nearest Neighbors and Logistic Regression being good potential alternatives.