Part 1: Data Preprocessing and Visualization

To prepare the data, I first looked at the shape and size of the data. The size of the data was originally 2860 with a shape of 286 rows and 10 columns. The datatypes all seemed to coincide with what I expected to be listed for each feature. There were a couple of features such as tumor size and age that I, at first, expected to be of type "integers" or "float". However, I realized that they were split into ranges such as "30-34", so I felt it was better to leave the type as "object". I did change the values of "no-recurrence-events*"* to be 0 and "recurrence-events" to be 1 because when I was trying to improve the recall of the K-Nearest Neighbor model (which I will talk about in the next section), it gave me an error due to the values of *class*, so I figured it would be easier to make them 0s and 1s. Because I did this, I also had to change the data type of *class* to integer. When looking at the data, I noticed that there were 2 features that contained an invalid value of "?". These features were *node-caps* and *breast-quad*. Since the datatype of these 2 features were objects and not integers or floats, I couldn't use the mean or median to replace the invalid values, so I replaced them with the mode.

Next, I visualized 4 features using histograms: *age*, *menopause*, *class*, and *irradiat*. The histogram of age showed exactly what I was expecting which was that a majority of breast cancer patients are around 40-60 years old. The histogram of class showed that a majority of the breast cancer patients did not have recurrence events. The histogram of irradiation showed that a majority of the patients did not undergo irradiation procedures. This made me wonder if there was a correlation between irradiation and whether or not a patient would have a recurrence event. The last histogram I looked at was a histogram of the menopausal status of the patient at the time of the study. The histogram showed that a majority of patients were premenopausal. This was a bit surprising to me as I know menopause typically starts around ages 40-65. Since I saw in a previous histogram that most patients were around 40-60 years old, I would have expected the majority of patients to have been either in menopause or postmenopausal.

Lastly, I performed one-hot encoding on the categorical variables. I left the *class* variable as is since I knew it would be needed for training and building the model. I also left the *deg-malig* variable as is because the values were integers, and performing one-hot encoding on numbers can be tricky and misleading. I performed one-hot encoding on the rest of the variables: *age, menopause, tumor-size, inv-nodes, node-caps, breast, breast-quad,* and *irradiat.* I felt that the values of all of these variables could be considered categories, so I one-hot encoded all of them to help the model understand the data better.

Part 2: Building and Assessing Models

To split the data, I defined x to be all of the variables except for *class* and y to just be the variable *class.* I used 70% of the data for training and 30% for testing. The techniques I used to train the model were the Decision Tree model, the K-Nearest Neighbors model, and the Logistic Regression model. The standard model performance matrix that I believe is most important to optimize is recall. I think it is more dangerous to predict a patient won't have recurrence if they actually will, so false negatives are more important to reduce. Improving recall means reducing false negatives.

The Decision Tree model had an accuracy of 63% on the test data and 97% on the training data. It also achieved 31% recall and 36% precision on the test data. This means that this model has slightly more false positives than false negatives, but they are still about the same. It is not too much of a difference.

The K-Nearest Neighbors model had an accuracy of 70% on the test data and 74% on the training data. It achieved a recall of 4% and precision of 50%. This means that the model has a lot more false negatives than false positives, since the recall score is drastically lower than the precision score. Because we are more concerned in reducing false negatives in this case, I found the KNeighbors model with the best recall. This model had a 58% accuracy on the test data and 97% accuracy on the training data. It achieved 23% recall and 27% precision on the test data. This model is the least accurate out of all of the models, but in comparison with the K-Nearest model that has 70% accuracy, this model has a much better recall score in comparison to the 4%.

The Logistic Regression model had an accuracy of 72% on the test data and 76% on the training data. It achieved 31% recall and 57% precision on the test data. This means that the model has fewer false positives than false negatives. The recall score of this model is the same as the Decision Tree model, but the precision score is much better for this model.

Overall, the Logistic Regression model had the highest accuracy of all the models as well as the lowest recall and precision score. Because of this, the Logistic Regression model is the model I would recommend to be used for this dataset.