Samuel Amoabeng
COE379L

PROJECT 2 REPORT

PART 1

DATA PREPARATION

Data preparation is made using different steps in data processing, first, all the libraries needed to perform this process. Installation of these libraries is essential for this stage, it gives all the necessary packages for the preparation stage. The installation helps in calculating statistical descriptions and plots for the virtualization of distributed key variables to identify patterns and potential outliers for the model. Further dataset classification using three different supervised learning techniques. Classifications used for the training are K-Nearest Neighbor, Decision Trees, and Logistic Regression.

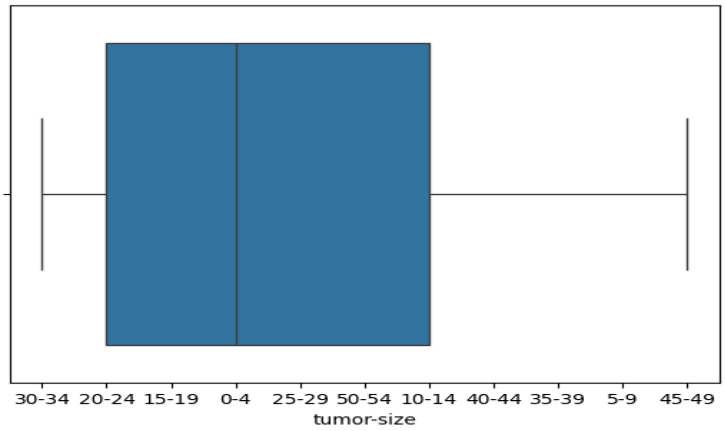
Cleaning the dataset is one of the integral processes by checking missing, and handling duplicates. It is also important to check the relationships, distribution of the data, and what part of the data is needed for the modeling. Creating new features or modifying the existing dataset to improve modeling performance. This can be done by converting categorical variables into formats that best fit the model, for this dataset process, deg-malig is converted from an integer to an object. The dataset had no missing data or invalid values out of the 286 counts. However, there was only one unspecified data that is insignificant and will not change the data process.

DATA INSIGHTS

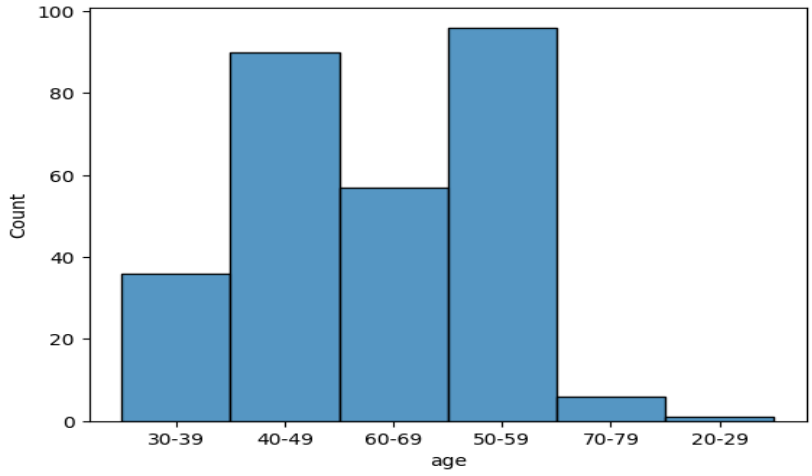
The data set contains information on 286 rows with 10 columns. Each row represents a patient, and each column represents a feature or attribute of a patient, the first patient is from the "no-recurrence-events" class, aged 30-39, in premenopausal status, with a tumor size of 30-34,

0-2 involved lymph nodes, no node caps present, degree of malignancy 3, left breast affected, in the left lower quadrant, and did not receive radiation therapy. The dataset mostly contains objects with one integer, deg-malig columns datatype is an integer, which needed to be converted to an object.

DATA VIRTUALIZATION



The box plot shows the split of tumor-size, the dataset is skewed to the right because most of the dataset follows a trend to a specific age range.



Histo. above indicate the age ranges from the dataset. The most count is between 50-59 and the lowest from 20-29

PART 2

CLASSIFICATION USING SUPERVISED LEARNING TECHNIQUES

PROCEDURE FOR TRAINING MODEL

The dataset is split into training and testing, with the representation of 70% to 30%. Using one-hot encoding, it shows the shape of both X and y training and testing of the dataset.

```
X_train shape: (200, 9)
X_test shape: (86, 9)
y_train shape: (200,)
y_test shape: (86,)
```

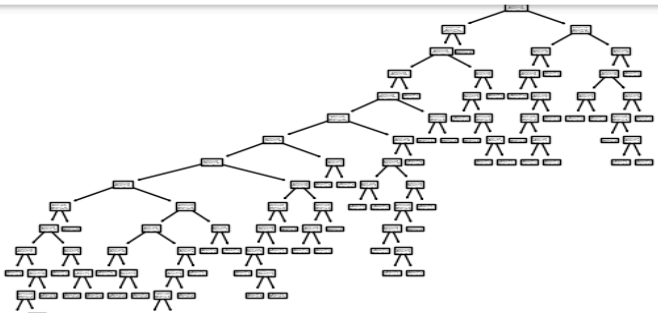
K-Nearest Neighbor Classification: Shows just the accuracy of both the train and test of the modeled dataset. With the help of chatGPT, training, and testing were combined that helped to encode categorical variables using OneHotEncoder used to convert categorical variables into numerical format.

```
Accuracy of knn on test data is: 0.79
Accuracy of knn on train data is: 0.82
```

Decision Tree Classification:

```
Accuracy: 0.6744186046511628
Performance on TEST
*******************
                        precision    recall  f1-score   support

no-recurrence-events       0.74      0.83      0.78        60
   recurrence-events       0.44      0.31      0.36        26

            accuracy                           0.67        86
           macro avg       0.59      0.57      0.57        86
        weighted avg       0.65      0.67      0.65        86

Performance on TRAIN
********************
                        precision    recall  f1-score   support

no-recurrence-events       0.96      1.00      0.98       141
   recurrence-events       1.00      0.90      0.95        59

            accuracy                           0.97       200
           macro avg       0.98      0.95      0.96       200
        weighted avg       0.97      0.97      0.97       200
```

```
[20]: tree.plot_tree(model)
```



The above decision tree report indicates low accuracy, high error rates, or poor performance metrics compared to what is expected, in the tree above is complex which can be the cause of overfitting.

Logistic Regression:

```
Performance on TEST
*******************
                     precision    recall  f1-score   support

no-recurrence-events       0.78      0.85      0.81        61
   recurrence-events       0.53      0.40      0.45        25

            accuracy                           0.72        86
           macro avg       0.65      0.63      0.63        86
        weighted avg       0.70      0.72      0.71        86

Performance on TRAIN
********************
                     precision    recall  f1-score   support

no-recurrence-events       0.80      0.93      0.86       140
   recurrence-events       0.73      0.45      0.56        60

            accuracy                           0.79       200
           macro avg       0.76      0.69      0.71       200
        weighted avg       0.78      0.79      0.77       200
```

The recommended model that will be best for the dataset is logistic regression as compared to other classifications, with high accuracy.

For the test data: Precision for the positive class (recurrence-events) ranges from 0.44 to 0.53, indicating a moderate ability to avoid false positives. Recall for the positive class ranges from 0.31 to 0.40, indicating a moderate ability to avoid false negatives.

For the train data: Precision for the positive class ranges from 0.73 to 0.80, indicating a moderate to good ability to avoid false positives. Recall for the positive class ranges from 0.45 to 0.93, indicating a moderate to high ability to avoid false negatives.

The standard model performant metric that is most important to optimize is recall because it is the part of the model that needs to be improved.