Project 2: Breast Cancer Prediction

Pranjal Adhikari

Introduction

In this project, the main objective was to explore a breast cancer data set and utilize various different Python libraries, such as *pandas* and *scikit-learn*, to predict the recurrence of cases using different models. The data set itself contains various different attributes of breast cancer cases, such as tumor size, the age of the patient, and menopausal status. The *pandas* library was first used to process and prepare the data, along with dealing with missing/invalid values that were present. Next, *scikit-learn* was used to train three different models: KNN, decision tree, and logistic regression. These models were used to predict the recurrence of breast cancer given the different attributes of the cases. Lastly, each model's accuracy and performance was determined in these predictions.

Data Preparation

The data set used in this project (found here) contains 10 attributes of 286 different breast cancer cases, including:

1. class: no-recurrence and recurrence events
2. age: binned age for sample population
3. menopause: menopausal status of patient at the time of study
4. tumor-size: size of tumor binned (mm)
5. inv-nodes: invasive nodes binned
6. node-caps: node capsule
7. deg-malig: degree of malignancy
8. breast: left, right
9. breast-quad: left-up, left-low, right-up, right-low, central
10. irradiat: irradiation

Before the model could be trained on the data set, the data set had to be prepared for optimization. First, the data type of each column was determined to verify the values within each column were set to the appropriate data type. In this data set all attributes, except tumor-size (being type int), were type object, which was then converted to type category. Additionally, there

were missing values in some cases for the attributes node-caps and breast-quad. These missing values were filled in with the most frequently occurring value within the respective attribute. Next, the newly set category type attributes were then transferred to create binary vector columns utilizing the one-hot encoding technique. This data type conversion and utilization of one-hot encoding allowed for efficiency in the model training seen later in the project.

After preparing the data, visualization techniques were applied to understand the attributes in the data set. The visualization gave insight that the most frequent tumor size seen in the data set fell in the range 30-34mm, the age range that had the highest number of breast cancer cases was between 50-59 years old, and the left breast had a higher number of cases than the right breast.

Training the Model

First, two new variables were created to split the attributes. The class_recurrence-events attribute was saved as the y variable (target, dependent variable) and other attributes saved as the X variable (independent variable). This was done so the models could be trained to predict the breast cancer recurrence using the case attributes. Next, the data contained within the two variables were split into training and testing sets using *scikit-learn*, with 30% allocated to the testing set and 70% allocated to the training set. After the allocation, the training data was fit to the KNN, decision tree, and logistic regression models. For the KNN classification, the hyperparameter *k* was selected using cross-validation and searching for the most optimal value to be used. In this case, a value of 10 was selected for *k*. The other two models did not utilize any such hyperparameter.

Conclusion

To evaluate all the models' performance in predicting breast cancer recurrence, the accuracy, recall, precision, and f1-score was calculated using *scikit-learn*. The following are the results for the test data set for each model.

Model	**Accuracy**	**Recall**	**Precision**	**F1-Score**
KNN	0.74	0.23	0.75	0.35
Decision Tree	0.62	0.31	0.35	0.33

Logistic Regression	0.72	0.31	0.57	0.40

The KNN and logistic regression model had similar accuracy scores at 0.74 and 0.72, respectively, with the decision tree model being the lowest at 0.62. With these values, it can be stated that the KNN model is the best in predicting the recurrence of breast cancer among patients. However, the same model also had the lowest recall score among all the models at 0.23, while the other two scored 0.31 each. A low recall score means that there are more false negatives, in this case meaning the model predicted incorrectly that breast cancer will not recur that the other models. This incorrect prediction could cause further issues with a patient not receiving the treatment they need for the recurrence of breast cancer. Thus, the recall score is the most important performance metric to optimize, in my opinion.

References

Class website and materials: https://coe-379l-sp24.readthedocs.io/en/latest/index.html