Rhea Samuel
COE 379L
rss3488

Project 01: Breast Cancer Classifier

Data Preparation

To prepare the dataset, I began by thoroughly examining the data, checking data types, and identifying which variables required conversion. Most variables were of type object, except for the **deg-malig** variable, which was an integer. I first addressed missing values that I found in **tumor-size** and **inv-nodes** with the mode of each column. Additionally, I removed any duplicate entries to ensure data consistency. Once all inconsistencies were resolved, I applied one-hot encoding to transform the categorical data into a binary format for easier processing. Finally, I split the dataset into a **70/30** training and testing set using **train_test_split**, preparing it for model training and evaluation.

Insights from Data Preparation

During the data preparation process, I made several interesting observations about the dataset. I discovered that the entire dataset was composed of categorical variables, allowing for a straightforward application of one-hot encoding across all features. To gain a deeper understanding of the variables, I performed univariate analysis using specific types of visualizations: **histplots**, **countplots,** and **boxplots**. While all three provide similar insights, they differ in how they represent the distribution of data. From this analysis, I discovered several key patterns. For example, **menopause** is more commonly categorized as **premeno** compared to **lt40** or **ge40**. I also noticed that **tumor-size** tends to fall within the **30-34** range. Additionally, the dataset contains more **nonrecurrence** cases than **recurrence** cases.

Model Training Procedure

Within my project, I used three different classification machine learning models:

1. K-Nearest Neighbors
2. K-Nearest Neighbor Classifier using Grid search CV
3. Linear classification

I utilized the sklearn library to train these models. The K-Nearest Neighbor algorithm classifies data points based on the majority label of their nearest neighbors. Similarly, the KNN algorithm with GridSearchCV is an improvement of the regular KNN model, as it optimizes the value of k using Grid Search and cross-validation. Lastly, the Linear Classification model makes predictions by identifying a decision boundary that best separates the classes in the data. For this, I used the Perceptron algorithm.

Model Performance

In regards to the model performance, we can see the following results:

Testing Data:

	Accuracy	Recall	Precision	F1
KNN	58 %	19 %	28 %	23 %
KNN (GSCV)	67 %	11 %	50 %	18 %
Linear Classifier	53 %	51 %	35 %	42 %

Training Data:

	Accuracy	Recall	Precision	F1
KNN	76 %	49 %	68 %	57 %
KNN (GSCV)	70 %	13 %	69 %	22 %
Linear Classifier	58 %	63 %	40 %	49 %

Based on the data, KNN performs well in terms of accuracy and precision on the training data but not as well with recall. KNN with GridSearchCV slightly improves accuracy, but its recall remains very low. On the other hand, the Linear Classifier achieves the highest recall, meaning it is better at minimizing false negatives, though its overall accuracy is lower than KNN. This shows that while KNN is better at correctly predicting negative cases, the Linear Classifier is preferred when recall is the priority.

Model Confidence and Limitations

For this project, I have an average amount of confidence in the models based on the data and the observed outcomes. Although the KNN model shows excellent accuracy on the training data, its low recall makes it difficult to trust that it will generalize effectively. KNN with Grid Search CV is not much better as only the accuracy slightly increased. Linear classification does improve in the recall portion, however, the accuracy goes down in comparison to the other two models. For this specific project, recall is the most important metric, as improving recall will help minimize false negatives, which is crucial for accurately identifying breast cancer cases. Therefore, I believe the linear classification model is the best model for this case.

Alternative Approach

Another approach I attempted was to convert specific range-based variables—**tumor-size, inv-nodes, and age**—into integer values by taking the average of their ranges, instead of turning all variables into categorical data. For instance, a range like 30-39 was replaced with 34.5 to represent its midpoint.

Testing Data:

	Accuracy	Recall	Precision	F1
KNN	56 %	24 %	29 %	26 %
KNN (GSCV)	68 %	8 %	60 %	14 %
Linear Classifier	68 %	8 %	60 %	14 %

Training Data:

	Accuracy	Recall	Precision	F1
KNN	80 %	57 %	76 %	65 %
KNN (GSCV)	71 %	14 %	75 %	24 %
Linear Classifier	69 %	8 %	58 %	15 %

Comparing the two approaches, we can see that the first approach resulted in the Linear Classifier achieving the highest recall (51%), making it the best at identifying recurrence cases. The second approach led to a significant drop in recall (8%) for the Linear Classifier, making it much less effective. While KNN models performed well in accuracy, they consistently struggled with recall in both approaches, meaning they missed many positive cases. Since recall is the most important metric for this problem, the first approach is the better choice.