COE379L_Prog1

What did you do to prepare the data?

First, I loaded the dataset into the machine and I prepared the data by first checking the file type of the given data set. Then, I read the csv file using pandas. After reading the csv file, I check the the shape, size, and data types of the raw data. After that, I check if there are duplicate rows and remove them. Finally, I fill in the missing or invalid values by using the mode method. Finally, I perform one-hot encoding on categorical variables such as age group, breast_quad_..., etc. Then to prepare the data for the model, I split the dataset into training (70%) and testing (30%) sets, ensuring the class distribution was maintained .

What insights did you get from your data preparation?

After preparing the data, I noticed that the target variable (class) is imbalanced with significantly more no-recurrence cases as compared to recurrence cases. This suggests that recall may be a more critical metric than accuracy and splitting the data into training and testing data needs to be more carefully handled. I also noticed that tumor size and inv-nodes were strong predictors of recurrence, where larger tumors meant a higher risk. I also noticed the most missing values where in node-caps, which were removed to avoid biases in training.

What procedure did you use to train the model?

To train the model, I split the data using stratified split (70% train, 30% test). Then I perform classification using the K-Nearest Neighbor Classifier, K-Nearest Neighbor Classifier using Grid search CV, and Linear classification. For the K-Nearest Neighbor Classifier, I used KNeighborsClassifier(n_neighbors=5) as a baseline model. Then,using GridSearchCV, I find the best k value in the range [1, 20], for this was 13. Finally, for the linear classification, I LogisticRegression(max_iter=1000) to handle the classification task. I use logistic regression instead of LinearRegression() because LinearRegression is for continuous output, while LogisticRegression is designed for binary/multiclass classification.

How does the model perform to predict the class?

KNN (k=3): Performed reasonably well but was sensitive to data scaling.
KNN (with Grid Search [1-21]): Improved performance by selecting an optimal k value, but still had challenges in cases with overlapping classes.
Logistic Regression: Provided a simpler model, however, its performance was also competitive against KNN.
In general, the accuracy was generally high, but can be misleading due to class imbalance. Then, the recall was more important in this case since missing a recurrence is costly. Finally, the F1-score provided a balanced view between precision and recall.

How confident are you in the model?

I am moderately confident in my model because the model accuracy is greater than 50% and the f1 score is greater than 75%. However, improvements could be made:

- The dataset is small and imbalanced, so the model might not generalize well.
- Alternative models such as Support Vector Machines and Random Forests could be explored for better generalization.