Cancer Classification

https://github.com/kymayodeji/cancer_classification

Objective

The objective of the analysis is to determine an optimal classification prediction model to use to detect breast cancer using the results of various fine needle aspirations (FNA) of patient breast masses. By predicting whether a tissue sample of a mass malignant or benign will allow the stakeholders (physicians and patients), to determine the next course of medical intervention if needed.

Data

Popular Breast Cancer Prediction dataset is from the University of Wisconsin Hospital: (https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data)https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. The dataset features are computed from digitized images of Fine Needle Aspirate (FNA) of patient breast masses. Dataset includes a primary key ID and a Diagnosis column with values of B (benign) or M (malignment). Each observation has real-valued feature columns describing various characteristics of the cell nuclei along with the Mean, Standard Error and "worst" or largest (mean of the three largest values) of these features computed for each image. Since the dataset is unbalanced (significantly more Benign diagnosis then Malignant), we will determine which classification model would produce the best accuracy results when used on the training and test dataset.

Exploratory Data Analysis, Data Cleaning and Data Pre-Processing

The dataset had a size of 569 rows and 33 columns and was well cleaned with only one column being dropped due to the amount of null values. The diagnosis column was the only non-numeric column and only contained one of two values (B for benign, M for malignant). The thirty feature columns were all of type float64 and varied in scale. in A few of the columns (perimeter and area) were very strongly correlated with their respective characteristics (mean, standard error, worst value) and even more columns were strongly correlated with the target diagnosis column.

Methods

We used multiple classifier models (Logistic Regression, k-Nearest Neighbors, SVM, and Random Forest models) that where trained on the same stratified train-test split (80%/20%) of the dataset. Each classification model was tuned using GridSearch to find the optimal hyperparameters to train, fit and test the model. The multiple metrics (accuracy, recall, precision and f1 score) were used to compare the training and testing errors for the different models used.

Findings and Insights Summary

From the comparison results, multiple models (Logistic Regression, SVM and Random Forest) achieved the highest testing accuracy of 97.3684% compared to the kNN model (96.4%). Since the Breast Cancer detection data set is unbalanced (63% benign vs 37% malignant), I needed to consider more than just the accuracy, as the goal is to minimize the number of cancer mis-diagnosis. Using the per label accuracy metrics of the models, I concluded that the Logistic Regression model was the best model to use since it correctly identified

100% of the malignant tumors in the test set compared to the SVM and kNN models (95% and 90% respectively).

Next Steps

In the above summary, the comparison of the models was done using all thirty (30) of the numeric features of the dataset. Since the dataset used metrics for three (3) types or measures (mean, standard error, worst result), we could also consider using a features set of only the means measures (case 1) or the strongly correlated features (case 2) to see if the accuracy and recall/sensitivity metrics can be improved upon. Another next step is to obtain and use different patient datasets of fine needle aspirations of breast cancer tissue to determine if the testing results are consistent with the Wisconsin dataset.