COE379L Project 1 Report

Ayushi Sapru

as98489

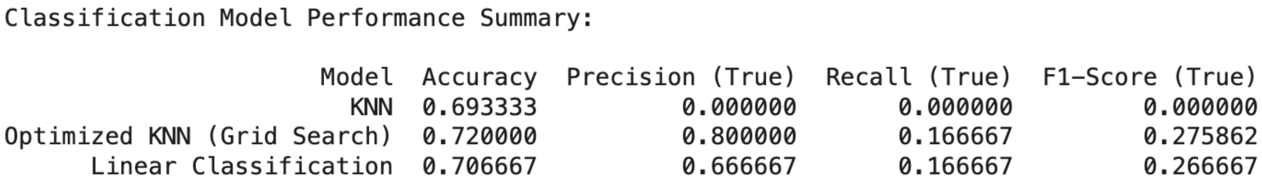
Breast cancer recurrence prediction is an important task for medical diagnoses as it allows for early intervention and planning treatments. This project focuses on analyzing a dataset containing patient records and applying machine learning models to classify recurrence events. The objective is to assess different classification models, K-Nearest Neighbors, Optimized K-Nearest Neighbors using Grid Search CV, and Linear Classification, to determine which one gives the most accurate results.

To prepare the dataset for classification, I first handled missing values by identifying placeholders like ? and *, replacing them with NaN, and imputing categorical columns using mode. This guaranteed consistency and prevented data loss that could impact model training. Next, categorical variables were transformed using one-hot encoding, allowing the machine learning models to process them effectively. The target variable, which represents recurrence events, was encoded into binary values to help with classification. Additionally, the dataset was split into 80% training and 20% testing while maintaining proportions. This step was vital to ensure that the training data accurately reflected the overall distribution of recurrence cases. Finally, exploratory data analysis was conducted, involving histograms and count plots to gain insights into recurrence, tumor size and location, and level of malignancy.

From the data preparation process, I observed a significant imbalance, with a larger proportion of patients classified as non-recurrent cases compared to recurrent ones. This imbalance could potentially bias the model towards predicting non-recurrence more frequently. The tumor size distribution showed that certain size ranges were more common, with a right-skewed distribution, which indicates more smaller tumors. The analysis of breast quadrants showed that tumors were more prevalent in the left-lower quadrant. The degree of malignancy variable displayed that most cases were concentrated at lower severity levels, potentially affecting how the model differentiates between classes. These insights guided the choice of evaluation metrics, leading to a focus on recall to ensure recurrence cases were not overlooked.

I trained three classification models: K-Nearest Neighbors (KNN), Optimized KNN using Grid Search CV, and Linear Classification. First, I implemented a baseline KNN model where k = 5 to establish initial performance metrics. Since the value of k can significantly impact KNN's effectiveness, I applied Grid Search CV to find the optimal k value, which was 17. Next, recognizing that recall is important in this application, I performed another Grid Search CV optimization, this time prioritizing recall instead of accuracy. Finally, I trained a Logistic Regression model as a linear classifier; I set a high iteration limit (max_iter = 1000). Throughout training, I maintained a consistent data split for a fair comparison.

The Optimized KNN for Recall where k = 3 provided the best balance between recall and precision, making it the most suitable model for detecting recurrence cases. The standard KNN model where k = 5 had moderate accuracy but had low recall, which meant it failed to detect many recurrence cases. Grid Search CV improved accuracy, but recall stayed low, suggesting that prioritizing accuracy alone led to a model biased towards the majority. Linear Classification showed promising results at first glance, offering balanced metrics but still underperforming in recall compared to the recall-optimized KNN. The classification report indicated that false negatives were reduced when optimizing for recall, which is critical in medical predictions where missing a recurrence case could have severe consequences.



```
Classification Model Performance Summary:

                     Model  Accuracy  Precision (True)  Recall (True)  F1-Score (True)
                       KNN  0.693333          0.000000       0.000000         0.000000
Optimized KNN (Grid Search)  0.720000          0.800000       0.166667         0.275862
     Linear Classification  0.706667          0.666667       0.166667         0.266667
```

While the model provided relatively decent classification performance, the presence of class imbalance suggests that further improvements could be made. The recall-optimized KNN was the best-performing model for recurrence detection, but its precision remains lower than ideal. This means that while it identifies more recurrence cases, it may also produce more false positives. If deployed in a real-world setting, other techniques could be used to balance the dataset and further improve recall. Additionally, an approach combining KNN and Linear Classification might yield better results. Overall, while the model is fairly reliable, there is room for improvement, particularly in recall without having to yield precision.