Vrinda Pandey
Project 2
3/14/2025

Which techniques did you use to train the models? (1 point)

To train the model I:

- Divided the data into train (0.3) and test (0.7) sets
- Used sklearn for the kNN, random forest, decision tree, and adaboost models
- Checked metric reports and the confusion matrix for each model
- Ultimately the confusion matrix was the best visual indicator of problematic training
- Looking at misclassification with the confusion matrix and the other report metrics report, I worked on refining the models for better and more accurate outputs (which I will delved further into during model performance)

Explain any techniques used to optimize model performance? (1 point)

In the case of my project the kNN and random forest performed the poorest. For that reason I focused on optimizing these models. Since the KNN model is not a tree based model I worked on data processing that would best benefit the model. In the case of kNN this was employing a standard scalar for data standardization.

In the case of the random forest standardization and normalization techniques were not an option as it is a tree based model. Because of this I focused on catching and narrowing down the parameters that would improve the models performance. This was ultimately tweaking the number of trees along with the depth of the trees to increase the efficacy and avoid the possibility of overfitting.

Compare the performance of all models to predict the dependent variable? (1 point)

As a quick summary the model performance can briefly be described as:

kNN - Train accuracy 0.9 with a test accuracy of 0.79 with a recall of 0.68 and a f1 of 0.77

Decision Tree- Train accuracy of 1.0 with a test accuracy of 0.83 with a recall of 0.84 and a f1 of 0.84

Random Forest- Train accuracy of 0.94 with a test accuracy of 0.89 with a recall of 0.89 and a f1 of 0.89

AdaBoost- Train accuracy of 0.89 with a test accuracy of 0.88 with a recall of 0.88 and a f1 of 0.88

With these brief metrics it is quick to be able to see the models behavior in accurately predicting unseen data along with its approach of use of the training data to generate the model (overfitting/underfitting) along with their abilities to correctly classify positive cases.

Which model would you recommend to be used for this dataset (1 point)

For this problem of the four models I looked at I would recommend the Random forest. It has the most balanced training and testing accuracy and doesn’t show signs of overfitting like the decision tree model. Additionally it has the best ability to identify positive cases

For this dataset, which metric is more important, why? (1 point)

In this data set the recall is extremely important in understanding the model's accuracy. With the confusion matrices we are able to see how the accuracy is affected by the model's inability to correctly classify positive cases and or incorrectly classify negative cases. This is very important as simply looking at the accuracy doesn’t tell us about where the model is lacking and what approaches it may need for optimization. Additionally when the dataset is skewed in its prediction variable it may affect how the model classifies and reports the accuracy, but the recall tells us exactly in what way the model misclassified a data point.