Fabiola Li Wu

Big Data Technology
Project 3

Chosen algorithms

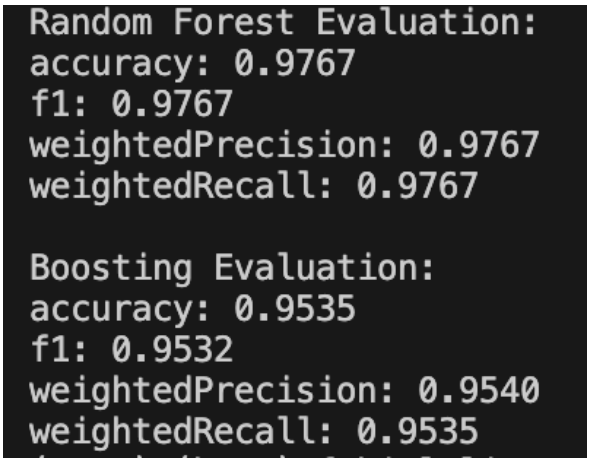
- Random Forest: I chose random forest because this algorithm has high accuracy, it does not have overfitting issues, it works well without tuning, it can handle multiple features, very robust and fast to train.
- Boosting: I chose boosting because ithas higher accuracy than random forest, it does not have a lot of bias, it can handle multiple features like random forest, you can customize loss function, and it gives you the features that influences on the decision made for the prediction.

Training procedure

1. Removed the ID column because I just thought that we dont need it.
2. Labeled M as 1 and B as 0.
3. Put all the features together.
4. Split the data into two parts, where 80% of teh data will be used for training and the remaining 20% for testing.

Testing results

Random Forest is slightly better on all metrics, but Boosting still over 95% which is already very accurate and effective but it can be improved with hyperparameter tuning.



Limitations

- I didnt do cross validation
- It might not handle errors
- not reusable code

Future improvements

- Having a cross validation
- Having Try and Catch blocks