Big Data Science Homework 5 Report

Professor Anasse Bari

Spring 2024

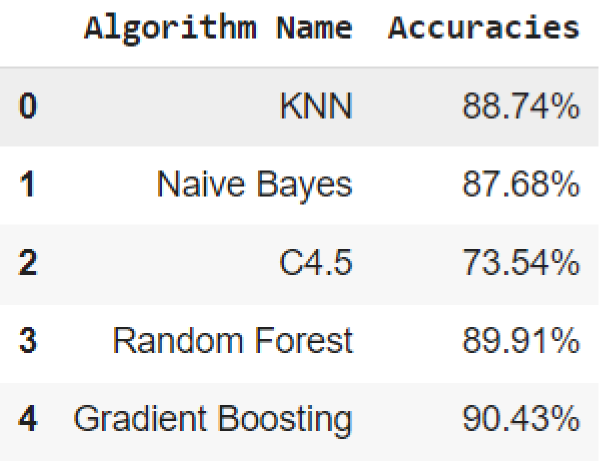
Ryan So and Samuel Moerman

(a) Feature engineering and preprocessing done

We preprocessed our data by removing outliers from the data, imputing missing values, and standardizing/normalizing all of the data. This cleaned our data, but with 14 features, we decided to reduce redundancy and circumvent overfitting by performing feature selection. We reduced the feature space to the 10 best features, and also converted all values to numerical values. After conducting feature engineering and preprocessing, we proceeded with data modeling on our data.

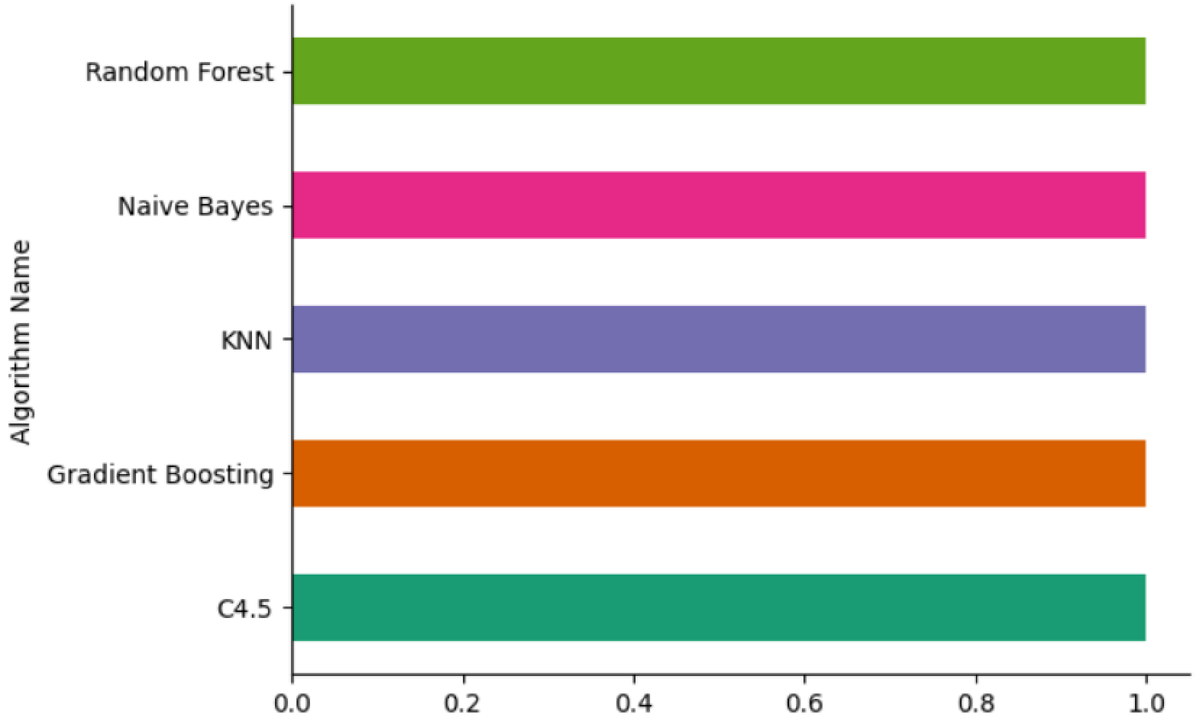
(b) Results from all the models and the feature selection/ranking in the form of a table

We trained and tested five different classification models, those being KNN, Naive Bayes, C4.5 Decision Trees, Random Forest, and Gradient Boosting. Below are our achieved accuracy scores.



	Algorithm Name	Accuracies
0	KNN	88.74%
1	Naive Bayes	87.68%
2	C4.5	73.54%
3	Random Forest	89.91%
4	Gradient Boosting	90.43%

As you can see, we achieved stellar results with a minimum accuracy of 73.54% (C4.5 Decision Trees), a maximum accuracy of 90.43% (Gradient Boosting), and four of the five classifiers achieving accuracies of greater than 87%.



(c) Results from the hyperparameter search

We conducted hyperparameter tuning on the Random Forest and Gradient Boosting classifiers with Grid Search. For Random Forest, we tuned the following hyperparameters: n_estimators, criterion, max_depth, min_samples_split. For Gradient Boosting, we tuned the following hyperparameters: n_estimators, learning_rate, and max_depth. The ideal combinations achieved were as follows:

- Random Forest: {'criterion': 'entropy', 'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 100}
- Gradient Boosting: {'learning_rate': 0.1, 'max_depth': 1, 'n_estimators': 200}

(d) Conclusions

We were certainly able to achieve our initial goal of developing a predictive model that can accurately classify patients into survivability outcomes based on a set of features derived from the SEER database. Each of our models performed relatively well with a minimum accuracy score of 73.54, maximum accuracy score of 90.43%, and four of the five algorithms having an accuracy higher than 87%. Not only did we successfully implement four high performing models, but we also successfully implemented hyperparameter tuning and visualization methods. Overall, we are proud of our results.