Statistical Learning

For this assignment we will work on the data files “student_performance_large_dataset.csv”, which can be found on Canvas. This dataset describes the learning habits and outcomes of 10.000 students.

1) Using decision trees, SVM models, and random forests build models (or boosting/BART) that predict the students’ performance, based on the attributes described in the datasets. You may use the grade or the final exam as your response depending on if you are interested in regression of classification. When you choose one of them to keep, remove the other from the dataset!

s

   To get full points, you need to prune the trees, discuss kernels and parameter optimization in the SVM model, and optimize the parameters in the ensemble models. You need to do cross validation and evaluate the performance of the models on the testing dataset. (10 points)

2) Using k-means and hierarchical clustering perform clustering on the full dataset (WITHOUT THE RESPONSE). How many clusters are optimal based on the results? Do the grades validate the clusters created? Does k-means and hierarchical clustering provide the same results?

   To get full points, you need to evaluate the clusters you created by assigning labels to them. You also need to compare the two clustering methods to each other. (10 points)