PROJECT 1

Who would survive the Titanic?

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this project, you should complete the analysis of what sorts of people were likely to survive. In particular, you should apply data analytics to predict which passengers survived the tragedy.

You should use predictive models from each of the following categories:

- Linear Regression
- Decision Trees
- Nearest Neighbors
- Clustering

You can use titanic_train.csv data for training and validating each of your models. You should use titanic_heldout.csv for testing the accuracy of your prediction algorithm.

Your objective is to improve the accuracy of the classification. Bonus points will be given to the student(s) with the best prediction algorithms.

For this project, you can work with at most one other student. In case of a joint project, only one student should submit.

What to submit:

- Report describing the methods that you have used and how to run your code. For each method, your report should show the confusion matrix and the accuracy.
- Source code
- Files titanic_predict_LinearRegression.csv, titanic_predict_DecisionTree.csv, titanic_predict_NearestNeighbors, titanic_predict_Clustering.csv, one file for each of the algorithms. Each file should have only one column indicating the prediction made by the algorithm (1: survived, 0: did not survive)

Titanic Data

VARIABLE DESCRIPTIONS

Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survival	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare (British pound)
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

Pclass is a proxy for socio-economic status
1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
If the Age is estimated, it is in the form xx.5

Fare is in Pre-1970 British Pounds