Homework 1
COMP 379
Brian Nguyen
9/13/2021

This program was an attempt at creating an algorithm that predicted whether a passenger survived on the Titanic based on their demographic and manifest features or not, analyzing both the *train.csv* for accuracy and *test.csv* datasets for predictions.

Studying the training dataset, I decided to model my algorithm around the *Sex*, *Pclass*, *Age*, *SibSp*, and *Parch* features of each passenger because I believed they held the more analytical potential for explaining how a person with a select combination of those features might have behaved during the actual disaster, as compared to the *Ticket*, *Fare*, *Cabin*, and *Embarked* features. The algorithm itself resembled a weighted scoring system, in which between 0-4 points would be allocated for each selected feature to a passenger's total "survivability score", depending on the influence that feature had on a passenger's survivability. The passenger's predicted survival would be based on whether the passenger's total score was greater or equal to a threshold of half of the maximum score possible (10 points) or not.

To determine the weight of each chosen feature in the model, I calculated the survivability rate of a passenger based on independent assumption values from the training dataset (see *Table 1*). For example, calculating the survivability rate of only females yielded a rate of 74.20%. By playing around with the thresholds a bit and supporting them with logic I thought was reasonable, I was able to determine the assumption values with the highest independent survivability rates regarding their respective features. An example of such reasoning would be that passengers with children, parents, siblings, and/or other relatives would probably have higher priority for being saved, as keeping families together is typically seen to be of higher importance than individuals with no other dependents when evacuating mass groups of people. With these survivability rates calculated, I ranked them from highest to lowest and then assigned each feature a proportion of the maximum score possible reflecting their perceived importance to a passenger's survivability. Perhaps the most interesting assumption I calculated was for age. At age 18, a passenger had about a 50% survivability rate. However, for each year younger than 18, the survivability rate increased by 1-4%, and for every year older than 18, the survivability rate decreased by 1-4%. As such, I decided that a passenger 18 years old or younger had a better than chance of survival than a passenger older than 18, which is reflected in the algorithm by allocating a point to the minors.

Initially, this algorithm yielded about a 70% accuracy rate out of 891 passengers in the training dataset. Upon further tweaking of assumption values and point allocations, I was content with a final algorithm that yielded a 78.45% accuracy rate. The algorithm also predicted that 30.86% (129 passengers) of the 418 passengers in the testing dataset survived, a result that I was not able to verify for accuracy due to the lack of actual survival data for those passengers.

The following table details the possible point allocation and independent survivability rate (based on the training dataset) for each feature's assumption values analyzed in my algorithm. For all other assumptions not listed in the table, they were allocated zero points, as they were perceived to have no positive influence on a passenger's survivability odds.

Table 1: Chosen Features and Assumption Values

Feature	Assumption Value	Point Allocation	Survivability Rate
Sex	female	4	0.7420
Pclass	1	3	0.6296
Pclass	2	2	0.4728
Age	<= 18	1	0.5036
SibSp	>= 1	1	0.4664
Parch	>= 1	1	0.5117