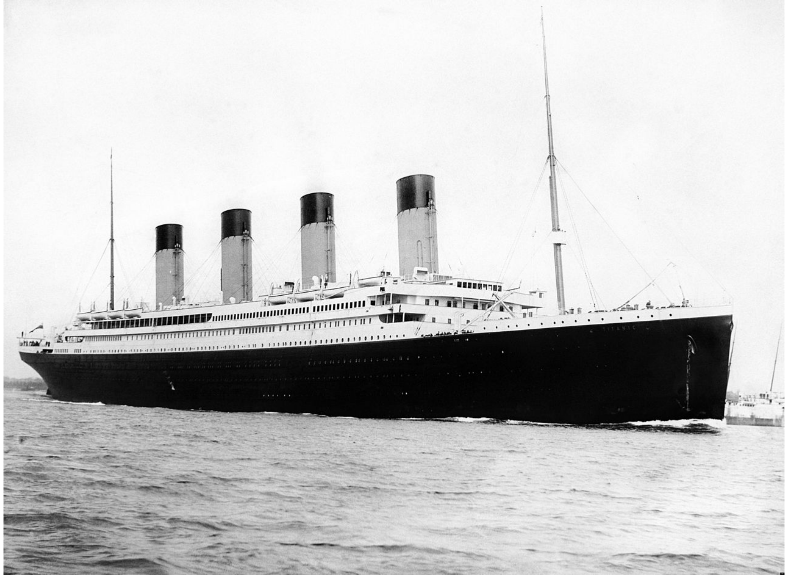
EE4484/IM4483 Artificial Intelligence and Data Mining

Continuous Assessment – Project (Option 1)

Due: Friday, 23 November 2018



RMS Titanic departing Southampton on 10 April 1912.
[Source: https://commons.wikimedia.org/wiki/File:RMS_Titanic_3.jpg]

On 15 April 1912, during her maiden voyage from Southampton to New York, RMS Titanic, the largest commercial ship afloat at that time, collided with an iceberg and sank to the bottom of the Atlantic Ocean. More than 1500 of the 2,224 passengers and crew on board perished in this infamous tragedy.

You are required to build a classifier to predict which passengers on board of Titanic would survive the tragedy. Two data sets – training set (“train.csv”) and test set (“test.csv”) – are made available. The attributes of the data are as follows:

- PassengerID – No. 1 – 1309
- Survived – 1=Yes, 0=No
- Pclass – Ticket class, 1 = 1st, 2 = 2nd, 3 = 3rd
- Name – Name of passenger
- Sex – Male, Female
- Age – Age in years
- SibSp – Number of siblings/spouses on board
- Parch – Number of parents/children on board
- Ticket – Ticket no.
- Fare – Passenger fare
- Cabin – Cabin no.
- Embarked – Port of embarkation, C = Cherbourg, Q = Queenstown, S = Southampton

The “ground truth” of the class of interest “Survived” for each sample is provided in the training set.

(a) Select at least one appropriate model (e.g., decision tree, neural network, support vector machine, etc.) to build your classifier(s) using only the data from the training set (“train.csv”). Your classifier(s) should take as input only the attributes available in the training set, excluding attribute “Survived”, which is the output of the classifier.

(b) Discuss how you handle attributes with missing values and what attributes are excluded from your classifier(s), and why.

(c) Discuss the model(s) you consider and the parameters/settings as well as your reasons of doing so.

(d) Compute the accuracy of your classifier(s) based on the training data.

(e) Comment on the results obtained. If you build more than one classifier, discuss and compare the results obtained from different classifiers.

(f) Apply the classifier(s) built in part (a) to the test set ("test.csv") which contains the data of 418 passengers not included in the training set. How many of these 418 passengers are classified as survivors from the Titanic tragedy? Among these survivors, how many of them were (i) female, (ii) below 18 years old, (iii) single? Based on the test set, passengers from which (iv) ticket class and (v) port of embarkment had the least chance of surviving the tragedy?

(g) [Optional] Submit the classification results of the test set to the Kaggle competition website at https://www.kaggle.com/c/titanic to obtain the accuracy of your classifier(s). Please note that you need to create an account at Kaggle and join the competition to submit your predictions for evaluation. Your submission must follow the format as provided in the sample submission file "submission.csv", and you can submit up to 10 submissions using your account at the Kaggle competition website per day.

(h) [Optional] Report your best results (Rank #, Team name, Score) as showed on the Kaggle Public LeaderBoard.

(i) [Optional] Does your best classifier obtain a better accuracy from the test set or the training set? Why? If you use more than one model/setting to build your classifier(s), which model/setting has a better generalization capability? Why?

Notes:

- Your project will be assessed based on the originality and effectiveness of your classifier(s), your justifications and discussions, your contributions, as well as the presentation and quality of your report.
- The training set ("train.csv"), test set ("test.csv"), and sample submission file ("submission.csv") can be downloaded from the content page of EE4483/IM4483 course website at NTULearn.
- You may use any suitable software tools/programmes for your work. Clearly state the software tools/programmes, functions, and parameters/settings used.
- If you do not have any idea/preference on which software tool/programme to use, try Orange, an interactive open-source software package for data visualization, machine learning and data mining, which is available for download at http://orange.biolab.si.
- If you couldn't obtain any meaningful results or answers to the questions above, you may describe what you have done and attach the relevant working, codes, or screenshots, if available.
- You may choose to work in pair with another student also taking the course. However, each of you must submit your own report, and the report must include at least two different classifiers and answers to all optional questions. Clearly specify in your report who is your project partner and your respective contributions to the project.
- You should clearly cite all the references and sources of information used.
- You are expected to uphold NTU Honour Code.
- Submit your project report in hardcopy to Mrs. Elaine Tan or Ms. Agnes Chia Peck Lan at S1-B1a-17 (Phone: 67905402, Email: elaine@ntu.edu.sg/eplchia@ntu.edu.sg) no later than the due date.