# CST8390 Assignment 3

Due: March 21, 2020 at 11:59 PM Sharp!!!

(Late submissions will not be accepted)

---

**Goal**: The goal of this lab is to explore and analyze Titanic dataset and perform classification using Decision Trees.

**References:**


1. http://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html
2. https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8
3. https://www.kaggle.com/c/titanic
4. http://csis.pace.edu/~ctappert/srd2014/d3.pdf
5. https://titanicfacts.net/titanic-survivors/


**Steps:**

1. Explore and analyse Titanic dataset given with this assignment (both train & test sets provided). Read the pages given in references. You have to include a brief description (what is this dataset about, what is the purpose of analysing it etc. - 10 lines) about the dataset. **Also**, you have to provide a table with **all** attribute names and their description.
   **(5 marks)**
2. 
   a. Identify and record **relevant** attributes to perform a classification on this dataset. Remove irrelevant attributes in the dataset.
   b. Create a new attribute to represent age group
      (If age is not given, keep it is as NK,
      if age <12, then Child;
      else if age < 18, then Teen;
      else if age < 45, then Adult;
      else if age ≤ 68, then Senior Adult,
      else if age >68, then Elderly).
   c. Create a new column "numRelatives" that adds number of siblings, spouse, parents and children. Now, create a column "Relatives" that has the rule:
      if numRelatives is 0, then Alone (which means travelling alone)
      else if numRelatives < 2, then Low
      else if numRelatives < 5 then Average
      else High

      Delete numRelatives column and keep only Relatives (think why?).

d. Save the new file as Titanic_train_processed.csv. Provide a screenshot of this file (header and a few rows should be visible.) (Hint: you will have only 6 attributes in total, including the class attribute). **(4 marks)**

3. Load data into Weka. Double check the type of your attributes. If they are not as expected, apply filters to convert them to the right types. (for example, class attribute should be nominal. Similarly, any attribute that is of nominal nature should be converted to nominal.)

4. Paste a screen shot of (a) distribution of the class attribute and (b) for the attribute Relatives by selecting class attribute from the dropdown list for visualization. Include screen shots as Q4a and Q4b. **(2 marks)**

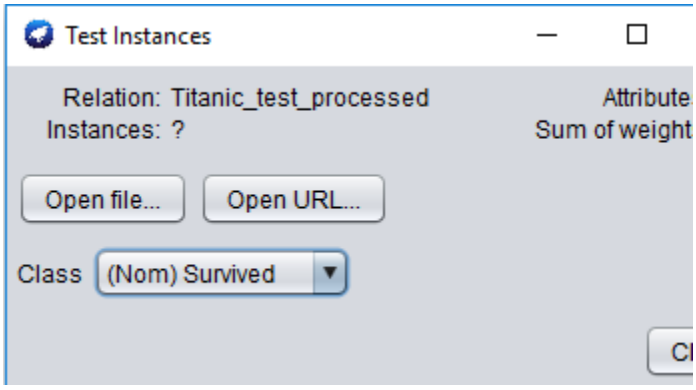
5. Save the file as titanic_train_processed.arff. Include a screenshot of the file as Q5. (Again, header and a few instances should be visible.) **(2 marks)**

6. Prepare the test set in the same way that you prepared training set. Test file should have the same format as the train set. In addition to all the other preprocessing steps, create a column for Survived, with '?' as the value. Save it as an arff file. Then use the same header from training file in the test file too (only change relation name).

7. Now, perform classification using Decision Trees with 10-fold cross validation. Copy your confusion matrix and include it in your submission document as Q7. **(2 marks)**

8. Visualize tree and paste the tree as Q8. **(2 marks)**

9. Now open **another** explorer and open your test file. This is just to ensure that the test file is in the right format to be used for testing. If there is an issue in opening the file, you need to make changes in the test file. Once test file is opened in the new explorer window, close the window.
10. Now, in the first explorer window, set the test set for testing. Click on Supplied Test set and set the test set.




11. Run Decision trees for the test set. As there is no actual Survived information, you will not get a valid confusion matrix. Right click on the execution and visualize classifier errors. Save your file from there as res.arff. Include a screenshot of this file as Q11. **(2 marks)**

12. Your new file will have a new column named “predicted Survived”. Fill in the following information:
    a. Total instances in the test file:
    b. Number of persons predicted to survive (1):
    c. Number of persons predicted not to survive (0):
    d. Percentage of predicted survival:

**(4 marks)**

13. From reference 5, check the actual information of the incident. Give an explanation on how your predicted results matches with the actual incident. List a few reasons why you think that your answer is different from the actual results. You need to compare results in detail based on various features. This is a 5 marks question, so a **detailed analysis and comparison of results** expected. **(5 Marks)**

***Note:*** *Make sure that you have selected relevant attributes. Otherwise, the analysis will be completely wrong.*

## Submission Details: **(2 marks)**

This is a partner assignment. Report should have a cover page with the names (Last name, first name) and student numbers. You need to paste all screenshots in the report. Now, create a zipped folder named:

<LastNameFirstStudent>_<FirstNameFirstStudent>_<LastnameSecondStudent>_<FirstNameSecondStudent>.zip with the

- Report in professional style,
- **processed** train and test arff files,
- model files from step 7 and
- res.arff file from step 11.

Upload the zipped folder to Brightspace.

## Marks:

This assignment will have a total of 30 marks. There will be **negative** marks if you miss explanation for any of the steps. Every step/question should be answered with explanation. **Prepare your assignment in a professional report style.**