CST8390 Assignment 2

Due: Feb 18, 2023 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:

Data Understanding

Explore and analyse Titanic dataset given with this assignment (both train & test sets provided). Read the pages given in references. You must include a brief description (what is this dataset about, what is the purpose of analysing it etc. - 10 lines) about the dataset. **Also**, you must provide a table with **all** attribute names and their description.

Data Preparation

1. Identify and record **relevant** attributes to perform a classification on this dataset. Remove irrelevant attributes in the dataset.
2. Create a new attribute to represent age group
   (If age is not given, keep it is as NK,
   if age <2, then Baby;
   else if age < 12, then Child;
   else if age < 25, then Youth;
   else if age ≤ 60, then Adult,
   else if age >60, then Senior).
3. Create a new attribute "Relatives". To create this column, the total number of relatives that include siblings, spouse, parents and children should be calculated.
   if the number of relatives is 0, record it as "None",
   else if the number is less than 3, record it as "Few",
   else if the number is 3 or greater, record it as "Many".
4. Apply binning (do both equal width and equal frequency and use the best out of these. Record why you think one is better than the other) to the Fare attribute.
5. Perform all data preparation steps in CRISP-DM.

6. Save the new file as Titanic_train_processed.csv. Provide a screenshot of this file (header and a few rows should be visible.)

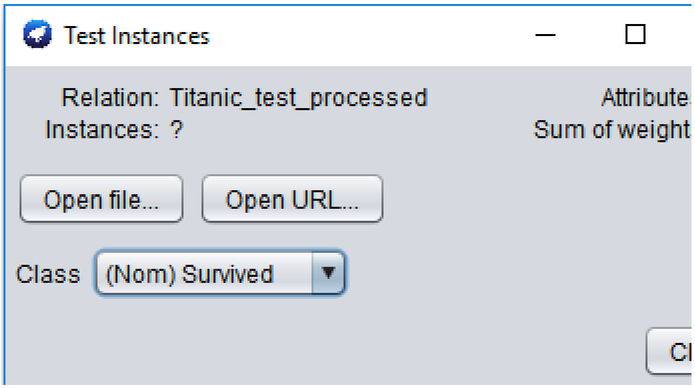
Load data into Weka. Double check the type of your attributes. If they are not as expected, apply filters to convert them to the right types. (For example, class attribute should be nominal. Also, as you will be using decision trees, all attributes should be converted to the right types accordingly.)

Paste a screenshot of (a) distribution of the class attribute and (b) for the attribute Age Group by selecting class attribute from the dropdown list for visualization. Include screen shots in the report. Save the file as titanic_train_processed.arff. Include a screenshot of the file. (Open the file in Notepad++. header and a few instances should be visible.)

Modeling & Evaluation

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). Now, perform classification using Decision Trees with 10-fold cross validation. Copy your confusion matrix and include it in your report. Visualize tree and paste the tree. Explain the meaning of the obtained tree.

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. Now, in the first explorer window, set the test set for testing. Click on Supplied Test set and set the test set.



Relation: Titanic_test_processed
Instances: ?

Class (Nom) Survived

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. 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:

Discussion of Results

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. Also, 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** with tables and charts expected.

Note: *Make sure that you have selected relevant attributes. Otherwise, the analysis will be completely wrong and if you select totally irrelevant attributes that do not have any effect on the survival, you will not get any marks.*

Submission Details:

This is a partner assignment. Report should have a cover page with the names (Last name, first name) and student numbers.

Create a small table and include who-did-what information in it. I want to see how you shared the workload in the group.

You must paste all screenshots in the report. The report should have table of contents, images, etc., sections titled Introduction, Data Understanding, Data Preparation, Modeling and Evaluation, Discussion of Results, Conclusion, References etc.

Font: Times New Roman. Font size: 12, justified

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 and
- res.arff file.

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.**