# TW4: Decision Tree Classification

Submit a summary of your learning to Canvas. Your document should include:

* Full names of your team members who work on the assignment.
  + Xiaomei Xie
  + Lili Hao
* URL links to the notebook of each student on GitHub repo.
  + <https://github.com/xiaomeiX/TW4-DT>
  + <https://github.com/lhaoSeattleu/TW4-DT>
* A summary of what you learned from the teamwork assignment.

In supervised algorithms, we use the Gini Impurity or Entropy to split nodes since data labels are known. In TW4, we use “entropy” and “Gini” two method for our decision tree. Both Gini Impurity and Entropy are criteria to split a node in a decision tree.

Two metrics for choosing how to split a tree. Gini measurement is the probability of a random sample being classified incorrectly if we randomly pick a label according to the distribution in a branch.

Entropy is a measurement of information (or rather lack thereof). You calculate the information gain by making a split. Which is the difference in entripies. This measures how you reduce the uncertainty about the label.

#### **change the following parameters and observe the changes.**

Change max-tree-depth would slightly change the accuracy result. If a Decision Tree is overfitting the training set, is it a good idea to try decreasing maximum tree depth, which called regualiztion.

**Part 2:**

#### 1. Look through an example list at: [(Links to an external site.)](http://localhost:8888/notebooks/A-CPSC4310/TW/week2/06_decision_trees.ipynb#1.-Look-through-an-example-list-at:)

[https://www.kaggle.com/dmilla/introduction-to-decision-trees-titanic-dataset (Links to an external site.)](https://www.kaggle.com/dmilla/introduction-to-decision-trees-titanic-dataset)

#### 2. Discuss about different ways to handle the following types of data for decision tree classification. [(Links to an external site.)](http://localhost:8888/notebooks/A-CPSC4310/TW/week2/06_decision_trees.ipynb#2.-Discuss-about-different-ways-to-handle-the-following-types-of-data-for-decision-tree-classification.)

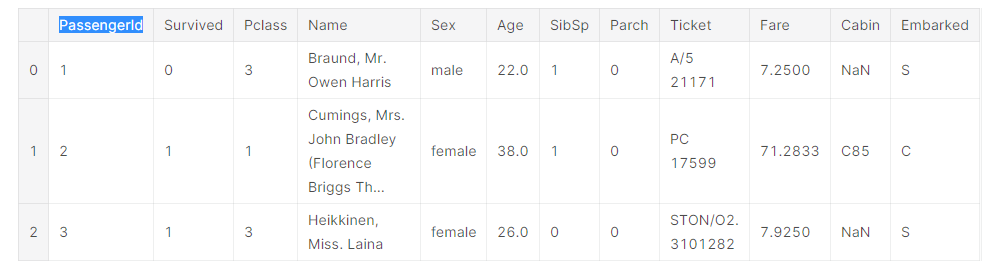
* text data (strings): in the case a dataset includes non-numerical data continuous data like age, weight, income, etc.

\*\*\* Notes: This will be also a reference for our next problem-solving assignment.

## different ways to handle the following types of data for decision tree classification: -- Lili

In the article “Introduction to Decision Trees (Titanic dataset)”, the dataset dealt with the missing data, transformed text data into numerical values, and created new potential meaningful features based on existing features. So, further analysis, such as Pearson Correlation, heatmap, and decision tree can be construct based on the cleaned dataset.

The original dataset is as below:



## Impute the missing data:

The missing data first is filled with different reasonable values.



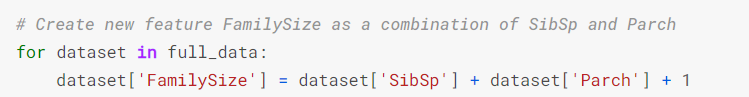
## Transformation:

Then, the text data is transformed into numeric values to represent different class:

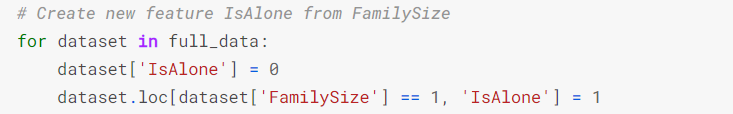
* A new feature, “title” is created based on name.



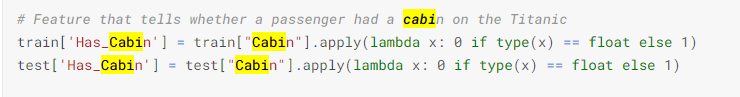
* A new feature, “FamilySize”, is created to combine SibSp and Parch.



A new feature, “isAlone” is further created based on Familysize

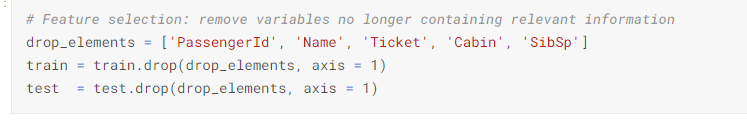


* A new feature, “Has\_Cabin” is created as a Boolean value from the text data.

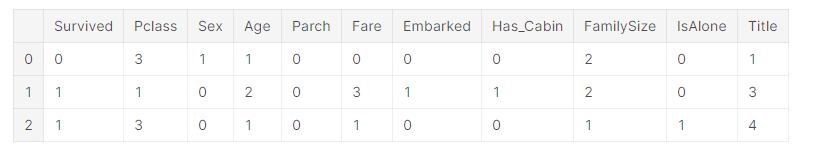


* Sex, title, embarked, fare, and age are further categorized into classes.



Then, unnecessary columns are dropped.

* PClass: this stays the same.
* Survived: This class is in binary format. It indicates whether the passenger is survived or not. No additional formatting is necessary for this field.



After all this, the dataset is much cleaner than before, with only numerical values and potential meaningful features. So, on top of those values, we can do further analysis, such as Pearson Correlation and heatmap.

Decision Tree:

*For this dataset, Title and Sex are both attribute in the dataset. But Title* is therefore more useful than *Sex.* Because *Title* implicitly includes information about *Sex* in most cases. T*he author uses Gini decision tree* learning algorithms to find the best split for each node of the tree. As result, Sex gets negative result. *Title* feature is slightly better at reducing the Gini Impurity than *Sex*. The most “Informative” attribute is the one maximized Gain.

For the decision trees, the 'max\_depth' parameter determines the maximum number of attributes the model is going to use for each prediction. When the max\_depth value is changed the Average Accuracy will slightly different. We can use max\_depth to avoid the over-fitting our model, which called regulization.

***Submission(s)***

*Each student should make individual submissions.*

* **Part 1**:
  + Push an updated notebook file to his/her/their Git repo.
    - **You do not need to submit any notebook files to Canvas**.
    - I will visit your Github to check the file.
* **Part 2**:
  + Submit a summary of your learning to Canvas. Your document should include:
    - Full names of your team members who work on the assignment.
    - URL links to the notebook of each student on GitHub repo.
    - A summary of what you learned from the teamwork assignment.