FINALS: Simulate Dataset

**Title:** " Titanic - Machine Learning from Disaster"

## Description

This analysis delves into the Passenger data of the Titanic, aiming to uncover patterns and factors that influenced the chances of survival using a machine learning approach. By leveraging the Random Tree algorithm in WEKA, a powerful machine learning tool, we seek to create a predictive model that can accurately estimate the likelihood of survival based on the attributes

# STEP-BY-STEP PROCESS:

## Data Cleaning

## The Titanic dataset exhibits a considerable number of missing values, which can significantly impede the simulation process. Hence, prioritizing data cleaning is imperative to ensure the integrity and reliability of subsequent analyses. The ensuing procedures delineate the meticulous steps involved in this pivotal preparatory phase.

## For a classification dataset, numerical data is needed, meaning there are no descriptive or textual information in the dataset, such as "male," "female," "high," or "low," and so on.

## Age: If there are missing values in the age column, the average of the age column will be calculated. Subsequently, these missing values will be filled with the calculated average to ensure completeness of the data.

## Sexuality or Gender: This can be represented as binary values, typically denoted as 1 and 0. However, it's essential to note that the interpretation of these values can vary across datasets. For instance, in one dataset, 1 might correspond to female, while in another dataset, 1 might represent male.

## In Excel, you can use the IF function to efficiently clean data:

## Example: =IF(someCell = 'someValue', valueIfTrue, valueIfFalse)

## Example: =IF(A1 = 'male', 1, 0) - Here, if the gender is male, it will be assigned a value of 1 and vice versa.

## In some cases, if data cannot be cleaned effectively, the last resort would be to delete those rows.

## Regarding numerical data, the features need to be numerical, as these are the determining factors that affect the label. Cleaning methods are not limited to just averaging; mean, median, and mode can also be used.

## Duplicate data must be removed to eliminate redundancy, as it can affect the performance of the dataset.

## Removing irrelevant features is also part of data cleaning, as redundant or irrelevant data may not affect or have any implications on the model's performance.

## Optionally, the dataset may be split into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance.

## Furthermore, it's important to check for incorrect data types in columns and unexpected values depending on the desired or expected data type for each column.

## Data Simulation

Given the attributes, a simulated dataset was created using CSV. The dataset includes attributes such as Pclass, Sex, Age, Sibsp, Parch, Fare, Survived.

## Open WEKA and Load Simulated Data

* 1. Launch the WEKA application.
  2. Navigate to the "Explorer" tab.
  3. Click the "Open file" button or select "Open" from the "File" menu.
  4. Load the simulated dataset file, 'Traincleanse.CSV,' located in the specified directory (e.g., D:\dataset)

## Preprocess Data

* 1. Explore the dataset using the "Preprocess" tab to examine its structure and nominal attributes.
  2. Handle any preprocessing steps required for nominal data, such as handling missing values or ensuring proper attribute types.



**UNIVERSITY OF CALOOCAN CITY**

Congress Campus

***COMPUTER STUDIES DEPARTMENT***

## Select Algorithm

* 1. Choose the "Classify" analysis to discover association rules among nominal attributes.
  2. Utilize the Random Tree algorithm for predicting survival probabilities in the Titanic dataset.

## Model Building and Evaluation

a. Configure the Random Tree algorithm settings, with consideration for the supplied test set option.

b. Initiate the algorithm to build the decision tree and evaluate its performance using the provided test set.

## Insights

* 1. Identify the strong influence of attributes of the passengers to the survival likelihood in the Titanic Dataset.
  2. Gain insights into combinations of attributes such as class, age, sex, and fare that are associated with survival or non-survival outcomes in the Titanic dataset.

# INSIGHTS FROM THE CLASSIFICATION MODEL

Upon applying the Random Tree algorithm to the Titanic dataset simulation, detailed patterns among its attributes have emerged, shedding light on factors influencing survival likelihood. The analysis reveals several noteworthy findings:

1. **Passenger Class and Survival:** Passengers belonging to first-class have a significantly higher likelihood of survival compared to those in second or third class. This association is particularly pronounced when considering passengers in the age range of 20-40 years.
2. **Age and Survival:** Age plays a crucial role in survival outcomes. Children under the age of 10 exhibit a notably higher survival rate, especially when accompanied by female guardians. Conversely, elderly passengers aged 60 and above have a lower likelihood of survival, particularly if traveling alone.
3. **Gender:** Gender disparities in survival rates are evident, with females having a considerably higher chance of survival across all age groups. This association remains consistent irrespective of other factors such as passenger class or ticket fare.
4. **Ticket Fare and Survival:** There exists a correlation between ticket fare and survival probability. Passengers who paid higher fares, indicative of higher-class accommodations, demonstrate a higher likelihood of survival, particularly when combined with attributes such as gender and age.
5. **Siblings and Spouses:** Passengers traveling with family members, especially spouses or siblings, tend to have improved chances of survival. This association is more pronounced among female passengers, suggesting the presence of familial support networks during the crisis.
6. **Parent-Child Relationships (Parch) and Survival**: The analysis highlights the significance of parent-child relationships aboard the Titanic concerning survival outcomes. Passengers traveling with a moderate number of parents or children (e.g., 1 or 2) demonstrate a higher likelihood of survival compared to those traveling alone or with a larger family. This association suggests that having a small family unit onboard may have facilitated better coordination and assistance during the evacuation process.

**RECOMMENDATION**

Understanding these analysis simulations results offers invaluable insights for historians and for future researchers aiming to unravel the intricate dynamics of survival aboard the Titanic. To deepen the understanding further, the researcher recommends exploring alternative algorithms or adjusting model parameters. Such endeavors hold the promise of unearthing additional insights, thus enriching our comprehension of this historic event and its underlying factors.

**LEADER: 1.** Bernal, Christian Dave

# MEMBERS:

1. Arendayen , Ajhay
2. Custodion, Melvin
3. Galanaga, Desiree
4. Molina, Ricky James