

Decision tree (Tiatanic Survival)

ByeongKyu Park (byeonggyu.park)

Name: ByeongKyu Park

Programming Language: Python

Decision Tree:



Data Selection

* chose the Titanic dataset from Kaggle (https://www.kaggle.com/competitions/titanic/data), which is a popular dataset among data science enthusiasts.
* I was browsing famous datasets for a project that would allow for meaningful analysis of survival factors. The Titanic dataset stood out because it provides an opportunity to explore how characteristics like age, sex, and ticket class affected survival rates.

# Data Description

* **Survival**: Indicates whether a passenger survived or not. o Key: 0 = No, 1 = Yes
* **Pclass (Ticket class)**: A proxy for socio-economic status.
  + Key: 1 = 1st (Upper), 2 = 2nd (Middle), 3 = 3rd (Lower)
* **Sex**: Gender of the passenger.
* **Age**: Age in years.
* **SibSp**: Number of siblings or spouses aboard the Titanic.
* **Parch**: Number of parents or children aboard the Titanic.
* **Ticket**: Ticket number.
* **Fare**: Passenger fare.
* **Cabin**: Cabin number.
* **Embarked**: Port of Embarkation.
  + Key: C = Cherbourg, Q = Queenstown, S = Southampton

# Data Cleaning & Categorization

* **Setting Index**: **PassengerId** was used as the index for the dataset. This unique identifier is crucial for tracking individual passengers in the decision tree.

* **Dropping Columns**: Non-relevant and non-numeric attributes such as **Name**, **Ticket**, and **Cabin** were dropped. These fields contained too much text data or had many missing values.

* The **Sex** column was manually encoded to numeric values for simplicity and compatibility with the decision tree model. Male was mapped to 0, and female was mapped to 1.

* **Age Categorization (0,1,2,3)** o **Child (0-12 years)**: Includes ages from 0 up to but not including 12.
  + **Young Adult (12-18 years)**: Includes ages from 12 up to but not including 18. o **Adult (18-60 years)**: Includes ages from 18 up to but not including 60.
  + **Senior (60-100 years)**: Includes ages from 60 up to 100.

* **Fare Categorization (0,1,2,3)**

*(the upper limit is inclusive due to the lower bound ‘-1’)*

* + **Low (Fare <= 10)**: Includes fares greater than -1 and up to 10. o **Medium (10 < Fare <= 50)**: Includes fares greater than 10 and up to 50. o **High (50 < Fare <= 100)**: Includes fares greater than 50 and up to 100.
  + **Very High (Fare > 100)**: Includes fares greater than 100 and up to 600.

* The **Embarked** column was encoded into numerical categories based on the unique ports of embarkation.