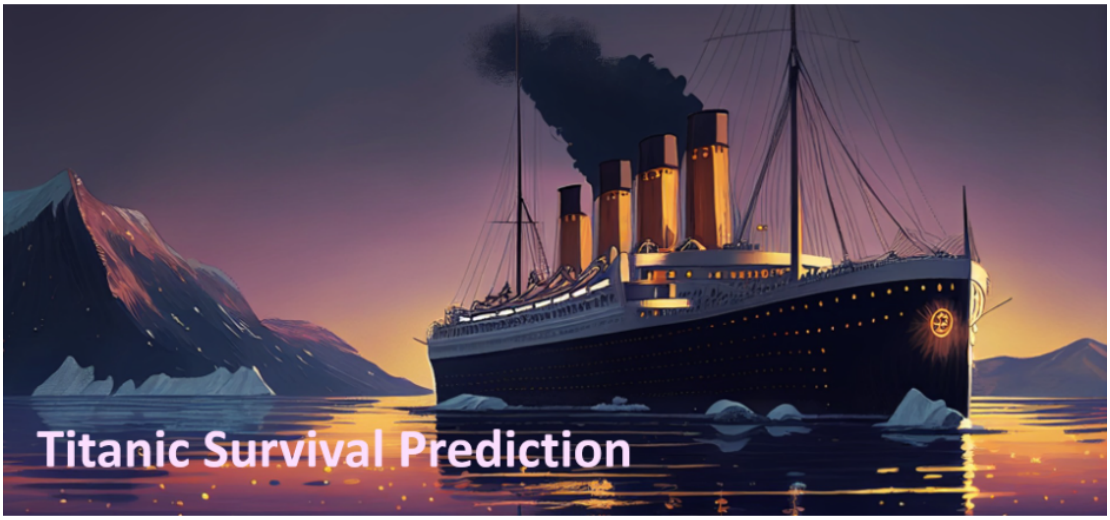
Assignment 1: Titanic - Machine Learning from Disaster

Deadline: 11:59PM CST(China Standard Time), May 17, 2025

May 11, 2025



This project aims to build a machine learning model that predicts the survival of passengers on the Titanic. The dataset used for this project is the famous Titanic dataset, which contains information about the passengers such as their age, gender, ticket class, fare, and whether they survived or not. The goal is to analyze the given data, prepare it for training, and then design and train your own classifiers to predict the survival outcome for new, unseen data.

1 Homework Grade Policy

Submission Format: Please format your homework **in English** using one of the two options below. Failing to do so will result in a minimum deduction of **10 points** from your homework grade:

1. **One PDF** generated by Jupyter Notebook that includes your code, complete output results, plotted figures, and written answers to analytical questions.
2. **One PDF + One PY file.** The `.py` file must contain all your code and be fully runnable. The PDF should include screenshots of your output results, plotted figures, along with your written answers to analytical questions.

Late Policy: Homework submitted within 72 hours after the deadline will receive 80% of the earned score. Submissions made more than three days after the deadline without a prior extension request will receive a score of 0. The request must be made before the deadline of the homework.

2 Dataset Overview

Train.csv will contain the details of a subset of the passengers on board (891 to be exact) and importantly, will reveal whether they survived or not, also known as the "ground truth". The dataset includes the following variables:

- **Survival**: Survival (0 = No, 1 = Yes)
- **Pclass**: Ticket class (1 = Upper, 2 = Middle, 3 = Lower)
- **Sex**: Sex
- **Age**: Age in years
- **SibSp**: Number of siblings/spouses aboard the Titanic
- **Parch**: Number of parents/children aboard the Titanic
- **Ticket**: Ticket number
- **Fare**: Passenger fare
- **Cabin**: Cabin number
- **Embarked**: Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

3 Data Preparation

Complete the following data cleaning tasks. Clearly document each step, including any code and explanations, in your submission.

1. Read and Visualize Data
   (a) **5pts:** Read the `train.csv` file. (Example)
   (b) **5pts:** Print out the first 5 samples. (Example)
2. Handling Missing Data
   (a) **5pts:** Print out the count of missing values in each column. (Example)
   (b) **5pts:** Impute missing entries with zero. (Example)
3. Handling numerical and categorical values.
   (a) **5pts:** Identify and write down which columns are numerical and which are categorical. (Example)
   (b) **5pts:** Convert the categorical columns into one-hot encoded columns. Print out the first 5 samples to verify. (Example)
4. Dataset Splitting
   (a) **5pts:** Divide the cleaned dataset into training features (`X_train`), training targets (`y_train`), testing features (`X_test`), and test targets (`y_test`).(Example)
   (b) **5pts:** Print out each set's shape. (Example)
5. Addressing Imbalanced Labels for Training Set (`X_train` and `y_train`)
   (a) **5pts:** How many samples are there for survival and non-survival? (Example)
   (b) **5pts:** Use oversampling technique to handle imbalanced classes. (Example)

4 Model Training and Validation

Complete the following model training tasks. Clearly document each step, including any code and explanations, in your submission.

1. **10pts:** Training a k-nearest neighbors classifier using the training dataset you have from prev section. (Example)
2. **10pts:** Print out your model's prediction accuracy on the training dataset.
3. **10pts:** Print out your model's prediction accuracy on the test dataset.
4. **10pts:** Plot the training/validation accuracy with respect to different values of k. The range of k should be [1, 100]. Save the plotted to your pdf submission file.

5 Analytical Questions

1. **5pts:** Which feature do you believe is the most important for the model's performance? Justify your answer with evidence, such as data analysis, visualizations, or feature importance scores.
2. **5pts:** For the `"Age"` feature, what alternative methods could be used to handle missing values instead of filling them with 0? Explain how these methods could improve the model's performance.

6 Optional: Try different models

Besides the providing example, try any of the following methods. The guide can be found on sklearn's website at https://scikit-learn.org/stable/supervised_learning.html.

- Decision Tree
- Random Forest
- Support Vector Machine Classifier

Specifically, report training accuracy, validation Accuracy, parameter you set (if any).