Workshops of the System analysis project

Julián Carvajal Garnica

20242020024

Andrés Mauricio Cepeda Villanueva

20242020010

Jhonatan David Moreno Barragan

20201020094

Andrés Camilo Ramos Rojas

20242020005

Part 1 of the System Analysis project

1. Competition Overview

The sinking of the Titanic on April 15, 1912, during its first voyage, is one of the most infamous shipwrecks in history. After hitting an iceberg, the lack of sufficient lifeboats led to 1,502 deaths out of 2,224 people onboard. Although luck played a role in survival, some groups had higher chances of surviving than others.

The Kaggle Competition "Titanic: Machine Learning from Disaster" challenges participants to predict which passengers survived the sinking of the Titanic.

Objectives:

- The main objective of this project is to develop a predictive model capable of determining, based on the provided passenger information, whether an individual aboard the Titanic would survive or not. The prediction is formulated as a binary classification problem, where the output variable takes the value 0 for passengers who did not survive and 1 for those who did.

- To understand how machine learning algorithms operate to derive a concrete and reliable solution to the survival prediction problem.

- Improve the predictive system after obtaining an initial solution by incorporating additional factors and refinements, thereby increasing the overall accuracy and robustness of the model.

Data structure:

- `train.csv` → passenger features + target Survived.

- `test.csv` → passenger features without target.

- `gender_submission.csv` → example of the expected submission format.

Features: Passenger class, sex, age, fare, cabin, port of departure, etc.

Target variable: Survived.

Metric: Accuracy (percentage of correct predictions).

Constraints: Submissions must include only PassengerId and Survived.

Example:

```
PassengerId,Survived
892,0
893,1
894,0
```

2. System Analysis Report

Systemic Analysis

Its goal is to estimate the survival of the passengers based on individual characteristics. To understand this system, we must identify its components, relationships, and boundaries.

System Elements:

- Input data:

    - PassengerId: Unique identifier

    - Pclass: Ticket class (1st, 2nd, 3rd)

    - Name: Full name

    - Sex: Gender

    - Age: Age in years

    - SibSp: Number of siblings/spouses aboard

    - Parch: Number of parents/children aboard

    - Ticket: Ticket number

    - Fare: Ticket fare

    - Cabin: Cabin number

    - Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

- Target output:

    - Survived: Binary label (0 = not survive, 1 = survived). Present in `train.csv`.

- Evaluation Metric: Accuracy score calculated by Kaggle on the test set.

- Constraints: Submission format (PassengerId, Survived), missing values, class imbalance.

- External Environment: Kaggle app, competition rules, leaderboard dynamics.

Relationships Between Elements: Variables do not act independently. For example:

- Pclass is related to Fare and Cabin, reflecting socioeconomic status.

- Sex and Age combined strongly influence survival probability.

- Embarked correlates with Pclass and Fare, revealing boarding patterns by class.

System Boundaries:

- Inside the system: Data, model, prediction, evaluation.

- Outside the system: Human decisions, real-life rescue dynamics, historical context.

Complexity & Sensitivity

The system exhibits several factors that increase its complexity and affect its sensitivity.

Complexity:

- Missing data: Variables such as Age, Cabin, and Embarked contain null values that must be handled.

- Feature interactions: Combinations like Sex with Age, or Pclass with Fare, produce nonlinear effects.

- Class imbalance: There are more passengers who did not survive than those who did.

Sensitivity:

- Small changes in input: Slight modifications to Age or Fare can shift predictions.

- Removal of key features: Dropping columns like Sex or Pclass reduces model accuracy.

- Limited dataset size: With few samples, complex models risk overfitting.

Chaos and Randomness

Although the system is built on structured data, elements of unpredictability emerge, reflecting aspects of chaos theory:

- Unforeseen interactions: Some passengers with favorable traits didn't survive, suggesting hidden variables or unrecorded decisions.

- Feedback loops: SibSp and Parch indicate family ties. These may create behavioral feedback.

- Nonlinear effects: The impact of variables like Age or Fare is not consistent.

- Real-world randomness: Survival depends on human factors, chances, and situational chaos.

Conclusions

The Titanic competition isn't just about predicting survival—it's a system packed with connections, gaps, and decisions that mirror human complexity. The data looks simple, but every variable interacts in surprising ways. There are patterns, but also chaos. Logic, but also randomness. Understanding this system means more than training a model—it's about thinking holistically, seeing beyond the numbers, and realizing that even in a controlled setup, uncertainty is always part of the game.
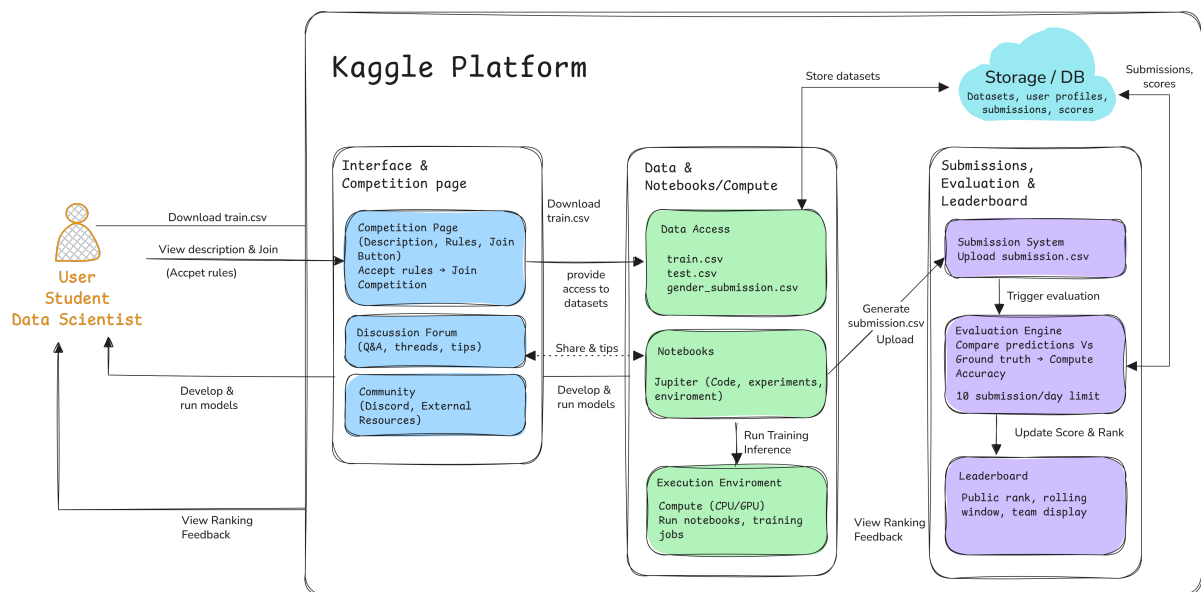
3. Visual Representation

Figure 1: shows the architecture of the Kaggle competition environment.

Figure 1 shows the architecture of the Kaggle competition environment. The competitor accesses the Competition Page to join and download datasets (`train.csv`, `test.csv`). Model development occurs in Notebooks (Kaggle Notebooks), using Compute resources to train and produce `submission.csv`. The Submission System accepts submissions, and the Evaluation Engine computes the accuracy against the hidden ground truth, updating the public Leaderboard. Social and knowledge-sharing channels (Discussion Forum, Discord, GitHub) provide community feedback and starter code. Storage/DB centralizes datasets, user profiles, submissions and scores.