UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Universidad Distrital Francisco José de Caldas

Systems engineering, faculty of Engineering

Technical Report: Titanic Machine Learning from Disaster

Julián Carvajal Garnica, A. Mauricio Cepeda Villanueva, Jhonatan D. Moreno Barragán, Andrés C. Ramos Rojas

Supervisor: Carlos A. Sierra Virgüez

October 25, 2025

Declaration

We, Julián Carvajal Garnica, A. Mauricio Cepeda Villanueva, Jhonatan D. Moreno Barragán and Andrés C. Ramos Rojas of Systems engineering, faculty of Engineering, University of Reading, confirm that this is my own work, and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalized accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of UoR and public with interest in teaching, learning and research.

Julián Carvajal Garnica, A. Mauricio Cepeda Villanueva, Jhonatan D. Moreno Barragán and Andrés C. Ramos Rojas
October 25, 2025

Abstract

The Titanic Machine Learning from Disaster initiative investigates the application of systems analysis and data science concepts to simulate and forecast the survival of passengers on the Titanic. This technical report outlines the analytical and design principles formed during the Workshops, emphasizing the creation of a strong predictive system utilizing the Kaggle competition dataset. The system incorporates machine learning methods into an organized workflow, covering data ingestion, preprocessing, model training, and evaluation.

This study aims to analyze the Titanic dataset as a dynamic system formed by interrelated variables—like passenger class, age, gender, and family ties—whose nonlinear interactions affect survival results. This method reveals complexity, sensitivity, and randomness as crucial elements influencing the dependability of predictions. Python and Scikit-learn serve as the primary tools for execution, utilizing a modular design that guarantees reproducibility, scalability, and transparency in data management.

The document outlines the analytical framework, technical specifications, and design factors that direct the model's development, providing a basis for subsequent phases of system execution and performance enhancement

Keywords: Machine Learning, Systems Analysis, Titanic Dataset, Predictive Modeling, Data Science, Python, Kaggle

GitHub repository https://github.com/MaurooC12/Titanic

Contents

List of Tables iv

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Problem statement . . . 1
- 1.3 Aims and objectives . . . 2
- 1.4 Solution approach . . . 2
- 1.5 Scope . . . 2
  - 1.5.1 Included in scope: . . . 2
- 1.6 Assumptions . . . 3
- 1.7 Limitations . . . 3
- 1.8 Organization of the report . . . 3

2 Literature Review 4
- 2.1 Binary Classification and the Titanic Challenge Context . . . 4
  - 2.1.1 Ensemble Algorithms and Classification Methods . . . 4
  - 2.1.2 Critique and Systemic Focus . . . 4
  - 2.1.3 Summary . . . 4

3 Methodology 6
  - 3.0.1 Research Design . . . 6
  - 3.0.2 Data Collection and Description . . . 6
  - 3.0.3 Data Preprocessing . . . 7
  - 3.0.4 System Architecture . . . 7
  - 3.0.5 Tools and Environment . . . 7
  - 3.0.6 Summary . . . 8

References 9

List of Tables

3.1 Dataset Description 6

List of Abbreviations

AI	Artificial Intelligence
CSV	Comma-Separated Values
ML	Machine Learning
RF	Random Forest
LR	Logistic Regression
FR	Functional Requirement
NFR	Non-Functional Requirement
API	Application Programming Interface
CPU	Central Processing Unit
F1-Score	Harmonic Mean of Precision and Recall

Chapter 1

Introduction

In the age of data-informed choices, merging systems analysis with machine learning has become essential for comprehending intricate issues via computational frameworks. The Titanic: Machine Learning from Disaster challenge on Kaggle offers a perfect setting for implementing these concepts. Through the examination of passenger information from the ill-fated journey of the RMS Titanic, the initiative seeks to forecast survival outcomes based on personal traits like age, class, and gender.

This initiative, created during the Systems Analysis course, considers the Titanic dataset not just as a forecasting task but as an active system made up of interconnected variables, limitations, and feedback loops. During the Workshops, the project transformed from a systemic comprehension of the dataset into a structured design that can produce strong predictions. This report outlines that process, integrating theoretical analysis, system modeling, and technical design into a unified framework for comprehending and developing a predictive system

1.1 Background

The Titanic: Machine Learning from Disaster competition, held on Kaggle, is one of the most acknowledged beginner-level challenges in data science. The goal involves forecasting the survival of individuals on the RMS Titanic, which sadly sank on April 15, 1912, while on its first journey. From 2,224 passengers and crew, over 1,500 lost their lives, highlighting considerable differences in survival rates based on gender, age, and social class.

Apart from its historical significance, the Titanic dataset has established itself as a common reference point for examining classification methods, data preprocessing techniques, and model assessment in machine learning. It offers a regulated setting for implementing analytical techniques while highlighting the interconnectedness of variables. Within Systems Analysis, this competition serves as an exemplary case study for exploring complexity, sensitivity, and chaos in a structured predictive framework

1.2 Problem statement

Initially, the Titanic problem seems like a straightforward binary classification challenge: forecasting survival (1) or demise (0). Yet, beneath this straightforwardness exists a network of interconnections among factors like class (Pclass), fare, age, gender, and familial ties (SibSp, Parch). These connections generate nonlinear impacts that render the system sensitive to slight variations in input information.

Additionally, absent data, small sample size, and class imbalance contribute to uncertainty in the predictive process. Therefore, the primary obstacle is to create a model that not only attains precision but also addresses uncertainty and sensitivity—characteristics of intricate systems

1.3 Aims and objectives

Aims: To develop a predictive system that estimates passenger survival on the Titanic through systemic analysis and data-driven modeling.

Objectives:

- Analyze the Titanic dataset as a dynamic and interdependent system.
- Identify and manage data limitations such as missing or unbalanced features.
- Design a modular and reproducible workflow for model development and evaluation.
- Apply appropriate machine learning algorithms (e.g., Random Forest, Logistic Regression).
- Evaluate system sensitivity and robustness based on predictive performance.
- Establish technical and systemic guidelines for scalability and interpretability.

1.4 Solution approach

The suggested approach combines systems thinking with the design of machine learning. The method starts with a systematic examination of inputs, outputs, and limits, then progresses to the development of a modular structure that distinguishes the primary processes: data ingestion, preprocessing, feature engineering, model training, and assessment.

Python acts as the core for execution, employing libraries like Pandas, NumPy, Scikit-learn, and Matplotlib. Every module functions autonomously while still connecting with others via clearly established interfaces, embodying the concepts of modularity and maintainability. Fixed seeds are used to introduce controlled randomness for reproducibility, and mechanisms for monitoring are implemented to identify anomalies or model drift.

This design not only facilitates precise predictions but also reflects the analytical and engineering discipline anticipated in system-focused problem solving

1.5 Scope

The parameters of the Titanic Machine Learning from Disaster project are determined by the limits of the competition and the goals set within the Systems Analysis course. The research centers on examining, creating, and partially executing a predictive system that can classify survival outcomes utilizing the Kaggle dataset.

1.5.1 Included in scope:

- Creation of a modular system for data intake, preprocessing, feature development, model development, and assessment.
- Handling missing and categorical data using imputation and encoding methods.
- Recognition of sensitivity, intricacy, and unpredictability within the system.
- Recording of functional, non-functional, and user-focused requirements.

  Application of Python libraries including Pandas, NumPy, Matplotlib, and Scikit-learn.

This defined scope ensures that the project remains focused on analytical and design objectives, emphasizing system understanding and technical feasibility over production-level implementation.

1.6 Assumptions

This study's evolution hinges on multiple assumptions that establish a systematic framework for analysis and design:

- Data Completeness Assumption: (train.csv, test.csv, and *gender_submission.csv*). are presumed to embody the complete and final information for this challenge, necessitating no external data incorporation.
- Assumption of Feature Relevance: The existing passenger characteristics (e.g., Sex, Age, Pclass, Fare, Embarked) are adequate for developing a dependable predictive model.
- Assumption of Environmental Stability: The Kaggle computational environment is uniform, guaranteeing that code runs and outcomes can be replicated in different sessions.
- Assumption of Randomness Control: Initializing with a random seed will reduce random fluctuations throughout model training.
- Assumption of Evaluation Consistency: The accuracy metric on the Kaggle leaderboard accurately represents the predictive model's performance.

These premises define the circumstances in which the system can be evaluated and executed, aiding in the preservation of consistency throughout analytical and technical phases

1.7 Limitations

Although it employs a systematic approach, the project encounters specific limitations that affect its results and applicability.

- Data Constraints: The dataset contains gaps and incomplete data, especially regarding variables like Age, Cabin, and Embarked. Imputation methods can reduce but not completely remove uncertainty.
- Model Limitations: This research emphasizes traditional supervised learning methods (e.g., Random Forest, Logistic Regression), which might not effectively capture intricate nonlinear interactions compared to more advanced algorithms.
- Scope Limitations: The project does not encompass real-time deployment or the enhancement of external data, restricting its practical use beyond the Kaggle setting.
- Computational Limitations: The training and testing of models occur within Kaggle's computational capabilities, limiting scalability and the variety of experiments.
- Unobserved Variables: Elements affecting survival—like individual actions, position on the vessel, or sheer luck—are not included in the dataset, resulting in intrinsic uncertainty.

Recognizing these constraints promotes openness and offers an accurate understanding of the model's effectiveness and evaluative results. Consequently, the system architecture prioritizes resilience, modular design, and clarity within the existing data and resource limitations

1.8 Organization of the report

This report is organized into several chapters to provide a comprehensive overview of the project:

- Chapter 2: Literature review, .
- Chapter 3: Methodology, detailing the design approach.

Chapter 2

Literature Review

2.1 Binary Classification and the Titanic Challenge Context

The *Titanic Challenge*, hosted on Kaggle, serves as one of the most recognized benchmarks in data science [1]. Its primary goal is to predict whether a passenger survived (1) or not (0) based on passenger information, which defines a **binary classification** task. Historical evidence and data analysis confirm that survival was not purely determined by chance but strongly influenced by factors such as **gender** (Sex), **age**, and **social class** (Pclass) [2]. The standard evaluation metric for this competition is **accuracy**, representing the percentage of correctly predicted outcomes [1].

2.1.1 Ensemble Algorithms and Classification Methods

Although the problem appears simple, it involves a network of variables (e.g., class, fare, age) that interact nonlinearly, making the system highly sensitive to changes in data. As highlighted in multiple Kaggle studies, simple models such as Logistic Regression or Decision Trees are often insufficient to capture these complex relationships [2]. To address this, ensemble-based methods like *Random Forests* and *Gradient Boosted Trees* are commonly employed, as they improve predictive accuracy by combining the outputs of multiple weak learners [3]. Additionally, **feature engineering** techniques—such as creating the *FamilySize* variable or extracting titles from names—play a crucial role in enhancing model performance [2]. Recent implementations, including TensorFlow Decision Forests, demonstrate that robust ensemble frameworks can further improve generalization in structured datasets like the Titanic [3].

2.1.2 Critique and Systemic Focus

Many solutions within the competition focus exclusively on maximizing accuracy, overlooking a key aspect of engineering: **systemic analysis**. According to the system analysis framework applied in the present project, a dataset should not be treated merely as a collection of numerical values, but as a **dynamic system** composed of interrelated variables and feedback relationships [4]. A systemic perspective highlights the importance of considering sensitivity, complexity, and randomness in predictive modeling—factors often neglected in data-driven projects [5]. Thus, the challenge extends beyond prediction accuracy to include the capacity to design a model that is both reliable and resilient to uncertainty [4].

2.1.3 Summary

This chapter establishes that the Titanic competition is fundamentally a binary classification problem in which ensemble algorithms provide the most effective approach for handling nonlinear dependencies [1, 3]. The review confirms that variables such as gender, age, and class are the most influential predictors of survival [2]. Finally, it reinforces that combining *Machine Learning* techniques with

systems analysis principles is essential to achieving a robust, interpretable, and sustainable predictive system [4, 5]. listings xcolor

Chapter 3

Methodology

3.0.1 Research Design

This study follows a quantitative and experimental research design, focusing on the use of structured data and computational modeling to identify patterns related to passenger survival. All experiments were conducted within the Kaggle competition environment, which provides the necessary datasets (`train.csv`, `test.csv`, and `gender_submission.csv`) and a standardized infrastructure for model evaluation.

A pipeline-based architecture was adopted, enabling data to flow sequentially through interconnected yet independent modules. This modular structure simplifies testing and enhances reproducibility, ensuring that each component can be refined without affecting the others. All implementation and testing were performed using Python within Kaggle Notebooks.

3.0.2 Data Collection and Description

The datasets used in this research were obtained directly from the official Kaggle repository. No external data sources were incorporated to maintain comparability and integrity across submissions. Each dataset contains information about passengers aboard the Titanic, including demographic, socioeconomic, and familial details.

Table 3.1: Dataset Description

File Name	Description
`train.csv`	Passenger features with survival labels (0 = no, 1 = yes).
`test.csv`	Passenger features without labels, used for prediction.
`gender_submission.csv`	Reference file indicating expected submission format.

Main Attributes:

- **Pclass:** Passenger's ticket class (1st, 2nd, or 3rd).
- **Sex:** Gender of the passenger.
- **Age:** Age in years.
- **SibSp / Parch:** Number of relatives aboard.
- **Fare:** Ticket fare.
- **Embarked:** Port of embarkation (C, Q, S).
- **Survived:** Target variable (1 = survived, 0 = not survived).

3.0.3 Data Preprocessing

Before training, the dataset underwent several preprocessing steps to ensure quality and consistency. These procedures included cleaning, transformation, and feature engineering.

1. **Data Cleaning:** Missing values in *Age* were imputed using the mean, while missing *Embarked* values were replaced with the mode. Irrelevant columns such as *Ticket* and *Cabin* were removed due to excessive null values.
2. **Feature Transformation:** Categorical variables (*Sex*, *Embarked*) were converted to numerical form using one-hot encoding. Continuous variables (*Age*, *Fare*) were standardized to reduce scale bias.
3. **Feature Engineering:** New variables were created to enhance model performance, such as $FamilySize = \text{SibSp} + \text{Parch} + 1$ and a binary variable *IsAlone*.
4. **Data Splitting:** The `train.csv` dataset was divided into 80% training and 20% validation subsets to evaluate model generalization.

3.0.4 System Architecture

The overall system architecture follows a modular pipeline that ensures a clear flow of data and tasks. Each module performs a distinct function within the predictive system:

1. **Data Ingestion Module:** Loads and validates the input datasets.
2. **Preprocessing Module:** Cleans data and handles missing or categorical values.
3. **Feature Engineering Module:** Creates new attributes that enhance prediction accuracy.
4. **Model Training Module:** Trains and validates the predictive model using machine learning algorithms.
5. **Evaluation & Output Module:** Computes accuracy metrics and exports the final submission file (`submission.csv`).

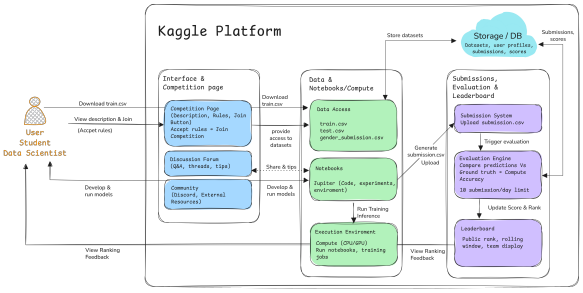

Figure 3.1: System Pipeline Architecture

This modular structure ensures maintainability, reproducibility, and adaptability for future improvements.

3.0.5 Tools and Environment

- **Programming Language:** Python 3.11
- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

- **Platform:** Kaggle Notebooks
- **Version Control:** GitHub repository for source code and documentation
- **Hardware:** Kaggle CPU-based compute environment

All analyses and implementations were executed in a controlled environment to ensure reproducibility across different sessions.

3.0.6 Summary

The methodology establishes a transparent, replicable process that connects data analysis, system design, and machine learning implementation. By integrating systemic principles with predictive modeling, the project builds a reliable framework for understanding complex interactions in survival prediction. The pipeline approach not only enhances modularity and accuracy but also aligns with the broader goals of systems engineering in data-driven contexts.

References

[1] Kaggle, "Titanic: Machine Learning from Disaster," *Kaggle Competitions*, 2025. [Online]. Available: https://www.kaggle.com/competitions/titanic

[2] A. Cook, "Titanic Tutorial: Machine Learning from Disaster," *Kaggle Notebooks*, 2019. [Online]. Available: https://www.kaggle.com/code/alexisbcook/titanic-tutorial

[3] Gusthema, "Titanic Competition w/ TensorFlow Decision Forests," *Kaggle Notebook*, 2023. [Online]. Available: https://www.kaggle.com/code/gusthema/titanic-competition-w-tensorflow-decision-forests. [Accessed: Oct. 25, 2025].

[4] J. Carvajal Garnica, A. M. Cepeda Villanueva, J. D. Moreno Barragán, and A. C. Ramos Rojas, "Titanic: Machine Learning from Disaster — System Analysis Project (Workshop 1 and 2)," *Universidad Distrital Francisco José de Caldas, Faculty of Engineering, Systems Engineering*, 2025.

[5] C. A. Sierra Virgüez, "Systems Analysis Course Guidelines," *Universidad Distrital Francisco José de Caldas, Faculty of Engineering*, 2025.