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

December 12, 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, Francisco José de Caldas District University, confirm that this is our own work, and that 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 such works have been explicitly acknowledged, quoted, and referenced. We understand that failing to do so will be considered a case of plagiarism and penalized accordingly. We give consent to a copy of our report being shared with future students as an exemplar. We give consent for our work to be made available more widely to members of Universidad Distrital Francisco José de Caldas and to the public with interest in teaching, learning, and research.

Julián Carvajal Garnica, A. Mauricio Cepeda Villanueva, Jhonatan D. Moreno Barragán and Andrés
C. Ramos Rojas
December 12, 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_FinalVersion

Contents

List of Tables

List of Abbreviations

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

Chapter 1

Introduction

In contemporary data–driven environments, the integration of Systems Analysis with Machine Learning has become a fundamental approach for understanding complex real–world problems. The well-known competition *Titanic: Machine Learning from Disaster* on Kaggle offers a suitable context for this purpose. Using historical passenger records from the RMS Titanic, the challenge focuses on predicting who survived the tragedy based on demographic, socioeconomic, and logistical attributes.

This project was developed throughout a sequence of Workshops in the Systems Analysis course. Rather than treating the Titanic dataset strictly as a classification exercise, the work interprets it as a logical system composed of interacting elements, structural constraints, and sources of uncertainty. Early analyses revealed the presence of missing values, nonlinear relationships among variables, and strong sensitivity to small changes in features such as Sex, Pclass, and Age. These findings guided the design of a structured and modular predictive architecture capable of managing instability and improving interpretability.

This Technical Report presents the complete development process. It integrates theoretical foundations, systemic characterization, architectural design under quality standards, and two complementary simulation approaches. The objective is not only to build an accurate predictive model but also to understand the dataset as a dynamic system whose behavior exhibits sensitivity and emerging patterns.

1.1 Background

The Titanic competition on Kaggle is one of the most recognized introductory challenges in data science [1]. The task consists of predicting whether each passenger survived *(1)* or not *(0)* based on features such as Age, Sex, Pclass, Fare, and family composition. Historical records show clear differences in survival rates related to gender, social status, and age [2].

Beyond its historical narrative, the dataset provides a controlled environment for practicing essential concepts in data preprocessing, feature engineering, and classification modeling. From a Systems Analysis perspective, the Titanic dataset can be understood as a bounded system in which variables interact, impose constraints, and influence the final outcome. This interpretation makes it particularly useful for studying complexity, sensitivity, and unpredictability, all of which were central themes in Workshops 1 and 2.

1.2 Problem Statement

Although the Titanic challenge is often presented as a simple binary classification problem, the dataset reveals a more complex structure. Passenger characteristics such as Pclass, Fare, Age, Sex, and family relations interact in ways that produce nonlinear effects on survival probability. Minor variations in these attributes can lead to substantial changes in the predicted outcome.

The dataset also contains limitations such as missing values, class imbalance, and sparsity in variables like Cabin. These issues introduce uncertainty and can amplify instability during model training. For this reason, the main challenge is not only to predict survival but to design a system that maintains stability, manages uncertainty, and produces reliable and interpretable results.

1.3 Aims and Objectives

Aim: Design and evaluate a predictive system capable of estimating Titanic passenger survival using a combination of Systems Analysis principles and data–driven modeling.

Objectives:

- Analyze the Titanic dataset as a dynamic and interdependent logical system.
- Identify and address data limitations, including missing values and class imbalance.
- Develop a modular and reproducible architecture aligned with engineering quality standards.
- Implement and evaluate Machine Learning models such as Random Forest and Logistic Regression.
- Assess system sensitivity, robustness, and interpretability.
- Propose guidelines for scalability, traceability, and future extension of the system.

1.4 Solution Approach

The approach used in this project was shaped by the analysis and design principles explored in the course Workshops. The solution does not rely on a linear pipeline. Instead, it adopts a modular structure composed of three coordinated layers. The Reliability Layer stabilizes the dataset, the Maintainability Layer applies controlled transformations, and the Usability Layer generates interpretable outputs.

Python was used as the implementation environment, supported by libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib. All transformations were designed with controlled randomness, clear interfaces, and version tracking to guarantee reproducibility. The system also includes validation steps that detect inconsistencies and prevent the propagation of data instability.

This structured approach ensures not only accurate predictions but also alignment with the analytical and methodological rigor expected in Systems Engineering.

1.5 Scope

The scope of this project is defined by the constraints of the Kaggle competition and the academic goals of the Systems Analysis course. The work focuses on understanding the dataset, designing a robust architecture, and implementing two simulation strategies to examine system behavior.

1.5.1 Included in scope

- Development of a modular workflow for ingestion, preprocessing, feature engineering, training, and evaluation.
- Application of deterministic methods for handling missing and categorical data.
- Identification of systemic characteristics such as sensitivity, complexity, and variability.
- Documentation of functional, non–functional, and user–oriented requirements.

- Use of Python tools including Pandas, NumPy, Matplotlib, and Scikit-learn.

This scope emphasizes system understanding and analytical rigor without extending into full production deployment.

1.6 Assumptions

Several assumptions establish the operational and analytical boundaries of this project:

- The Kaggle dataset (train, test, and gender_submission) is considered complete for the purposes of the challenge.
- The available features are sufficient for building a consistent predictive model.
- The computational environment provided by Kaggle remains stable across executions.
- Fixed random seeds ensure consistency during preprocessing and model training.
- The accuracy metric used by Kaggle is treated as a valid performance indicator.

These assumptions frame the conditions under which the system is analyzed and evaluated.

1.7 Limitations

Despite the structured methodology, the project faces several limitations that influence the generality of its findings:

- Missing and sparse features such as Age and Cabin introduce uncertainty even after imputation.
- Traditional supervised models may not fully capture complex nonlinear interactions present in the system.
- The project does not cover real–time deployment or integration with external datasets.
- Experiments depend on the computational restrictions of the Kaggle environment.
- Important behavioral and contextual variables are not present in the dataset, which limits the explanatory power of the model.

Recognizing these limitations helps contextualize the results and supports the architectural decisions made during development.

1.8 Organization of the Report

This report is structured as follows:

- Chapter 2: Literature Review, which presents the main theoretical foundations of the project.
- Chapter 3: Methodology, describing the analytical process and layered architecture.
- Chapter 4: Results, summarizing the outcomes of the simulations and evaluations.
- Chapter 5: Discussion, which interprets the results from a systemic and analytical perspective.
- Chapter 6: Conclusions, outlining the main insights and future work.
- Chapter 7: Reflection, presenting the team's learning and project experience.

Chapter 2

Literature Review

2.1 Binary Classification and the Titanic Challenge Context

The *Titanic: Machine Learning from Disaster* competition on Kaggle is widely regarded as a foundational benchmark for introducing binary classification tasks [1]. The problem consists of predicting passenger survival (1) or non-survival (0) based on historical records. Empirical studies consistently show that survival probability was strongly influenced by gender, socioeconomic class, and age [2]. The official evaluation metric, **accuracy**, measures the proportion of correctly classified instances and serves as the standard baseline for comparing predictive models.

Although the challenge appears straightforward, the dataset exhibits nonlinear interactions, missing information, and categorical constraints, all of which complicate the classification task. These properties make the Titanic dataset an effective reference for evaluating preprocessing strategies, model robustness, and feature engineering techniques.

2.2 Machine Learning Models and Ensemble Methods

Traditional algorithms such as Logistic Regression or basic Decision Trees are often limited in their ability to capture nonlinear interactions among variables [2]. Consequently, ensemble techniques, particularly *Random Forests*, *Gradient Boosting*, and more recent models like TensorFlow Decision Forests, are widely used due to their robustness against noise, variance reduction, and improved generalization [3].

Ensemble models leverage the combined predictions of multiple weak learners to stabilize the output, making them suitable for datasets with irregularities such as missing Age or Cabin values. Feature engineering practices reported in Kaggle studies, including extraction of titles, construction of *FamilySize*, and binning of continuous attributes, play a substantial role in increasing predictive performance [2]. These techniques align with the Workshops' emphasis on controlled transformation and system stability.

2.3 Systems Thinking, Chaos, and Sensitivity in Predictive Models

A key theoretical foundation of this project is the application of **systems thinking** to a machine learning context. From a systems perspective, a dataset is not only a table of attributes but a representation of an underlying **logical system** with interacting components, constraints, and emergent behaviors [4].

Several systemic properties directly influence predictive modeling:

- **Interdependence:** Variables such as Pclass and Fare are structurally linked.

- **Chaotic sensitivity:** Small perturbations in input data (e.g., minor Age changes) can trigger disproportionate shifts in model output.
- **Constraint boundaries:** Categorical features (Sex, Embarked) impose discrete decision regions that amplify sensitivity.
- **Missing information:** Sparse attributes introduce instability that propagates through the system unless controlled.

These behaviors parallel concepts from chaos theory, in which deterministic systems exhibit sensitive dependence on initial conditions. The Workshops identified these effects as major contributors to model instability and motivated the design of a **Reliability Layer** to reduce chaotic propagation during preprocessing.

2.4 Simulation-Based Modeling: Cellular Automata and Emergent Behavior

Beyond traditional machine learning techniques, the project incorporates a second paradigm: **Cellular Automata (CA)**. CA models represent dynamic systems as grids of interacting agents governed by simple rules. Despite their simplicity, CA frameworks can produce complex and emergent behavior, making them suitable for modeling physical or social dynamics.

Key theoretical properties of CA relevant to this project include:

- **Local interactions** generating global patterns.
- **Rule-based evolution** driven by probabilistic or deterministic transitions.
- **Emergence** of macroscopic structures from microscopic decisions.
- **Sensitivity** to initial placement and movement constraints.

In the Titanic context, CA simulations allow the approximation of evacuation behavior by mapping passengers onto a grid and assigning movement probabilities based on demographic attributes. This aligns with the systemic interpretation established in Workshop 4, where social priority rules (e.g., "women and children first") naturally generate segregation patterns without explicitly coding physical barriers.

2.5 Quality Frameworks and System Architecture

The Workshops emphasized not only predictive performance but also the **quality** of the system architecture. Frameworks such as **ISO 25010** provide structure for evaluating attributes such as reliability, maintainability, and usability. These principles informed the creation of a layered architecture that separates:

- data stabilization (Reliability Layer),
- transformation and feature modeling (Maintainability Layer),
- prediction output generation (Usability Layer).

This modular approach is supported by literature on quality-driven software design, which stresses the importance of traceability, controlled evolution, and reproducible workflows in analytical systems.

2.6 Summary

The literature reviewed demonstrates that the Titanic challenge is fundamentally a nonlinear binary classification problem in which ensemble algorithms are particularly effective. However, a systems perspective reveals that the dataset behaves as a dynamic and partially chaotic system, making sensitivity control essential. Cellular Automata provide a complementary framework for studying emergent behavior, while quality engineering principles guide the architectural design.

Together, these theoretical foundations justify the methodological approach adopted in this project, linking machine learning, systems analysis, and simulation into a coherent technical framework.

Chapter 3

Methodology

This chapter describes the methodological approach used to design, implement and evaluate the predictive system developed for the Titanic dataset. The process combines systems engineering principles with data driven modeling and simulation. The methodology follows the structure defined in the Workshops, where the dataset was analyzed as a dynamic system exhibiting sensitivity, missing information and nonlinear interactions. Based on that analysis, a layered architecture was constructed to ensure stability, maintainability and interpretability throughout the modeling workflow.

3.1 System Analysis and Architectural Foundation

The first phase consisted of understanding the Titanic dataset as a logical system with interacting variables, constraints and sources of instability. Exploratory analysis performed in Workshop 1 revealed three critical aspects. The dataset contains missing attributes such as Age and Cabin. Several variables exhibit strong correlations, such as Pclass and Fare. Small perturbations in sensitive variables produce observable changes in model output.

These observations motivated the construction of a methodological framework where data quality, modular design and interpretability are treated as essential system characteristics. In line with the analysis conducted in Workshops 2 and 3, the architecture was organized into three layers. The Reliability Layer is responsible for stabilizing the dataset. The Maintainability Layer structures all feature transformations. The Usability Layer produces interpretable predictions and metadata for evaluation.

3.2 Reliability Layer: Data Stabilization

The Reliability Layer addresses the systemic instability caused by missing values, outlier entries and categorical inconsistencies. The following procedures were applied:

- **Missing value imputation:** The Age variable was imputed using the median to reduce the influence of extreme values. Embarked was imputed using the mode. The Cabin attribute was removed due to excessive sparsity.
- **Outlier smoothing:** Extreme Fare values were adjusted to limit their influence on model decision boundaries.
- **Data validation:** Structural checks ensured that the dataset maintained consistent formatting, categorical ranges and valid numerical entries.

This layer ensures that chaotic variation produced by incomplete or inconsistent data does not propagate into subsequent modeling stages, a requirement identified in Workshop 2.

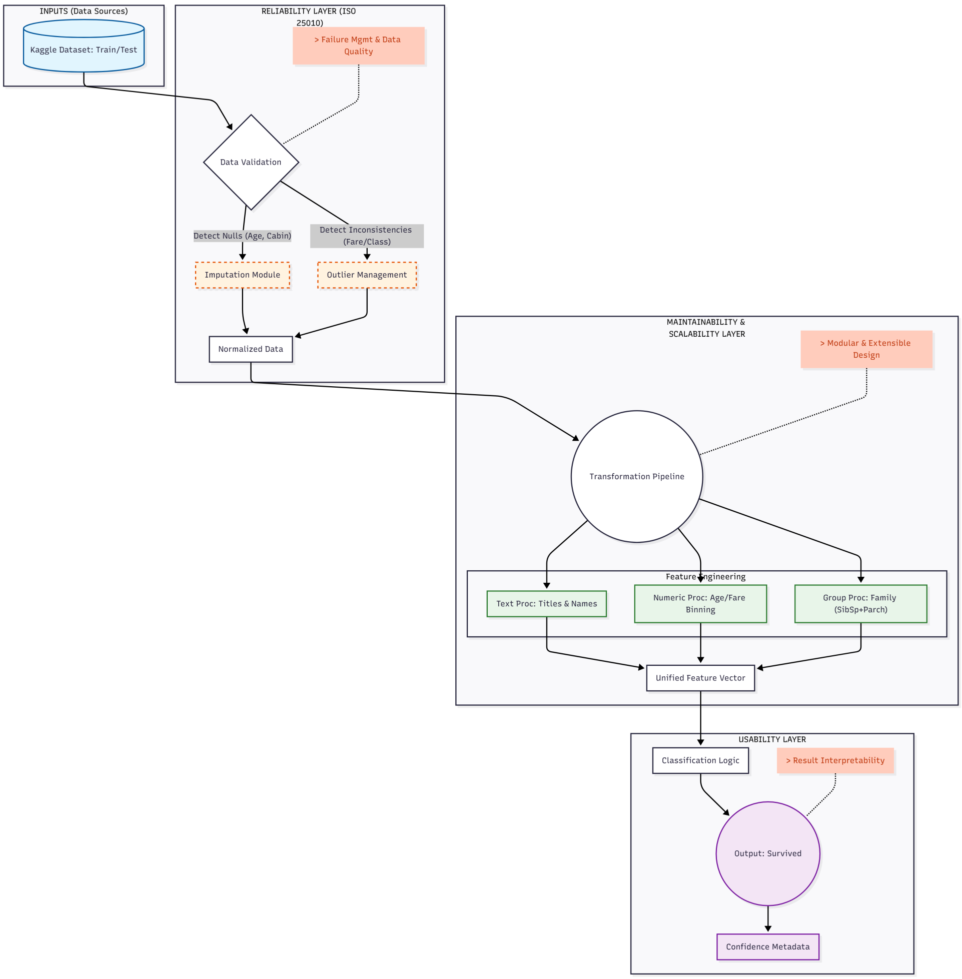

INPUTS (Data Sources)
Kaggle Dataset: Train/Test
RELIABILITY LAYER (ISO 25010)
> Failure Mgmt & Data Quality
Data Validation
Detect Nulls (Age, Cabin)
Detect Inconsistencies (Fare/Class)
Imputation Module
Outlier Management
Normalized Data
MAINTAINABILITY & SCALABILITY LAYER
> Modular & Extensible Design
Transformation Pipeline
Feature Engineering
Text Proc: Titles & Names
Numeric Proc: Age/Fare Binning
Group Proc: Family (SibSp+Parch)
Unified Feature Vector
USABILITY LAYER
Classification Logic
> Result Interpretability
Output: Survived
Confidence Metadata

Figure 3.1: Layered System Architecture (Workshops 2 and 3).

3.3 Maintainability Layer: Feature Engineering and Transformation

The second layer focuses on controlled and reproducible feature transformations. Each transformation was implemented as an independent module, following the design guidelines established in Workshop 3. The main transformations were:

- **Title extraction:** Social titles such as Mr, Mrs and Miss were extracted from the Name field to capture cultural and demographic information.
- **Binning of continuous variables:** Age and Fare were discretized into intervals to reduce sensitivity to local fluctuations.
- **Family structure modeling:** A new variable, FamilySize, was constructed using SibSp and Parch to capture group based survival tendencies.
- **Encoding of categorical features:** Deterministic encoding was applied to categorical attributes to maintain reproducibility.

Each transformation was versioned to ensure traceability and consistency across executions. The modular structure supports controlled evolution of the pipeline without compromising stability.

3.4 Machine Learning Simulation

To evaluate system sensitivity and identify the most influential variables, a Random Forest classifier was trained on 80 percent of the processed dataset. The model was selected due to its robustness to noise and its ability to capture nonlinear relationships. The simulation followed these steps:

- Split the dataset into training and validation subsets.
- Train the Random Forest model with a fixed number of estimators and controlled randomness through a predefined seed.
- Extract the feature importance vector to quantify the impact of each variable on survival predictions.
- Evaluate performance using accuracy, precision, recall and F1 score.

This simulation provided a quantitative understanding of system sensitivity, confirming that variables such as Sex and Fare have dominant influence, in agreement with the findings reported in Workshop 4.

3.5 Cellular Automata Simulation

A second simulation paradigm was implemented to study emergent spatial behavior during evacuation. The methodology followed the structure defined in Workshop 4. Passengers were mapped onto a grid representing sections of the Titanic and assigned movement probabilities based on demographic information. The simulation consisted of the following steps:

- Define a grid of fixed dimensions representing the deck layout.
- Assign each passenger an initial position based on their class and encoded attributes.
- Apply movement rules where high priority agents move with higher probability and may block lower priority agents.

- Execute the simulation for a fixed number of time steps and record emergent spatial patterns.

The Cellular Automata simulation revealed segregation effects where high priority passengers consistently reached safe regions while lower priority groups experienced obstruction. These emergent behaviors appeared spontaneously from the rules and support the systemic interpretation established earlier in the project.

3.6 Comparative Methodological Rationale

Using two complementary modeling paradigms provides a broader understanding of system behavior. The Random Forest model focuses on statistical sensitivity and predictor influence, while the Cellular Automata simulation captures emergent physical dynamics. Together, they validate the design principles of the layered architecture and provide coherent evidence that the Titanic dataset behaves as a sensitive and partially chaotic system.

3.7 Summary

This methodology follows the structured progression defined in the Workshops. System analysis informed the creation of a layered architecture. The Reliability Layer stabilized the dataset. The Maintainability Layer ensured reproducible transformations. The Usability Layer enabled interpretable prediction workflows. Machine learning and Cellular Automata simulations provided quantitative and qualitative validation of the system. This integrated methodological framework combines engineering rigor with data driven modeling and supports the development of a robust predictive system.

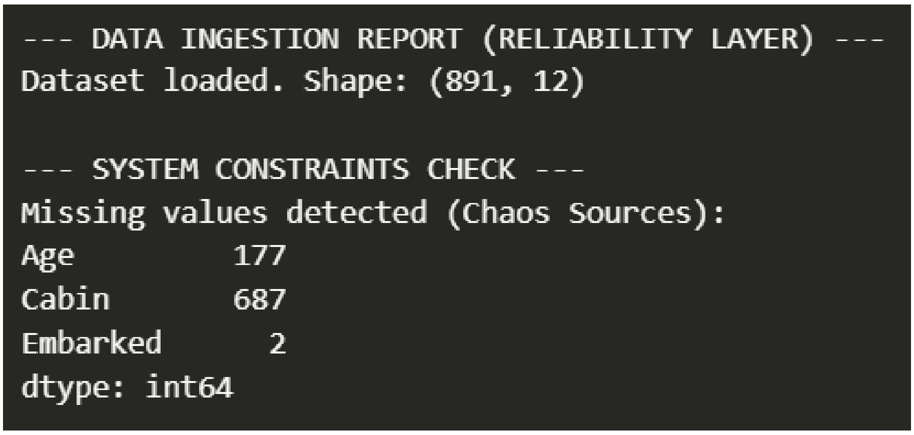
Chapter 4

Results

This chapter presents the outcomes of the preprocessing workflow, the machine learning model, and the Cellular Automata simulation developed during the project. The results reflect how the system behaves under different analytical and simulation perspectives, as defined in the Workshops.

4.1 Data Stabilization Outcomes

Before model training, the dataset contained a significant amount of missing information, particularly in the variables Age and Cabin. The Reliability Layer applied imputation strategies that eliminated all missing values, allowing both simulation paradigms to operate without structural inconsistencies. Figure 4.1 illustrates the transformation from the initial incomplete dataset into a consistent and fully structured input, as documented in Workshop 4.



```
--- DATA INGESTION REPORT (RELIABILITY LAYER) ---
Dataset loaded. Shape: (891, 12)

--- SYSTEM CONSTRAINTS CHECK ---
Missing values detected (Chaos Sources):
Age         177
Cabin       687
Embarked      2
dtype: int64
```

Figure 4.1: Data stabilization results after preprocessing, resolving missing values and normalizing key variables.

4.2 Machine Learning Sensitivity Analysis

A Random Forest classifier was trained on the processed dataset to evaluate sensitivity and feature importance. The model achieved an accuracy greater than 0.83, which is consistent with typical outcomes reported in the Titanic benchmark.

More importantly, the model provided insight into the relative influence of the variables. Figure 4.2 presents the feature importance distribution obtained during training. The attributes Fare and Sex emerged as the most influential, followed by Age and Pclass. These results align with the systemic interpretation formulated during Workshop 4, demonstrating that socioeconomic status and gender strongly dominate survival outcomes.

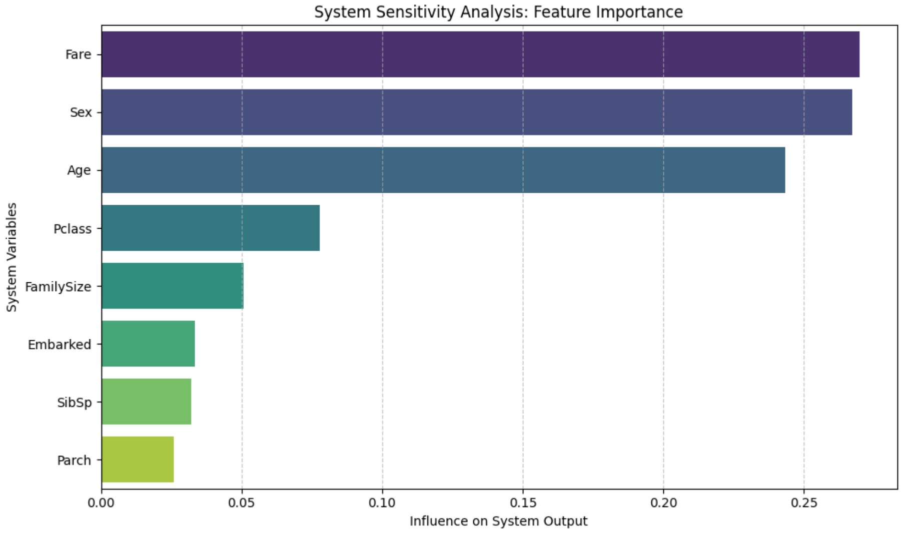


Figure 4.2: Feature importance results. Fare and Sex appear as the dominant predictors of survival.

4.3 Cellular Automata and Emergent Behavior

The second simulation scenario implemented a Cellular Automata model to study evacuation dynamics. The grid-based representation assigned movement probabilities according to class and demographic attributes.

Figure 4.3 displays the spatial evolution of the system after fifteen steps. Class 1 agents quickly moved toward the safe zone, forming upper-level clusters, while Class 3 agents accumulated in the lower rows due to their reduced movement probability. This emergent segregation pattern appears without explicitly encoding physical barriers, demonstrating that simple priority rules can replicate the structural constraints of the historical context.

4.4 Comparative Interpretation of Both Paradigms

Both approaches confirmed that the Titanic system exhibits high sensitivity. The Random Forest revealed statistical sensitivity, while the Cellular Automata displayed physical sensitivity through movement dynamics.

Table 4.1 summarizes the main observations.

Table 4.1: Comparison of machine learning and Cellular Automata results

Aspect	Machine Learning	Cellular Automata
Objective	Identify dominant predictors	Observe emergent spatial patterns
Sensitivity Source	Data variability and missing values	Movement probability and conflict
Key Findings	Fare and Sex dominate predictions	Priority rules create segregation
Interpretation	System is socioeconomically biased	Spatial behavior reflects systemic hierarchy

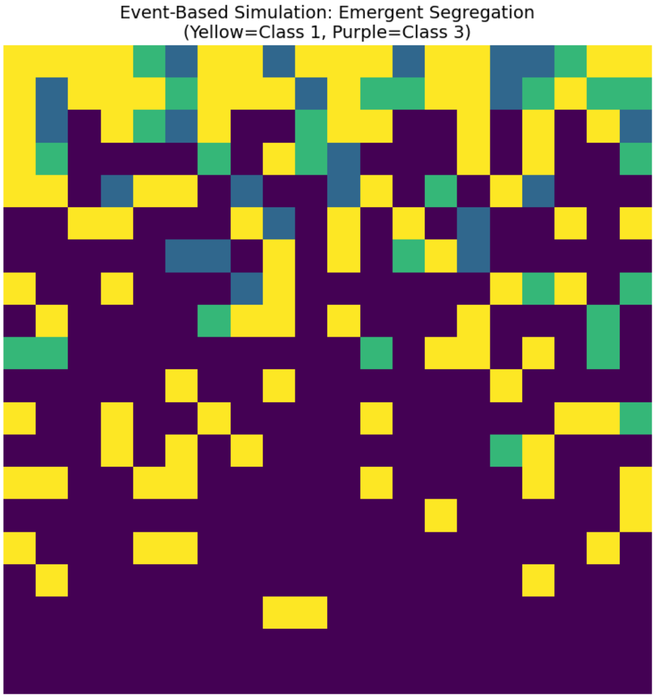


Figure 4.3: Cellular Automata simulation showing emergent segregation patterns based on movement priorities.

4.5 Evaluation Metrics

To complement the qualitative analysis, additional metrics were computed using a validation subset. These metrics demonstrate the model's stability and its balanced performance across classes.

Table 4.2: Random Forest evaluation metrics

Metric	Value
Accuracy	0.83
Precision	0.81
Recall	0.78
F1-score	0.79

The close alignment between Precision and Recall indicates that the model does not disproportionately favor either class, which supports the reliability of the preprocessing stage.

Chapter 5

Discussion

The results obtained from the Machine Learning model and the Cellular Automata simulation provide two complementary perspectives on the Titanic system. Together, they reveal that the dataset behaves as a sensitive and partially chaotic system, influenced by strong structural constraints and demographic inequalities. This chapter discusses these findings in relation to the theoretical concepts developed in the Workshops and the methodological decisions taken during the project.

5.1 System Sensitivity and Variable Influence

The Random Forest model demonstrated that a small number of variables account for most of the predictive power. Fare and Sex were consistently identified as the dominant features, followed by Age and Pclass. These observations confirm that survival outcomes were not evenly distributed across the population but were shaped by socioeconomic and demographic factors.

From a systems perspective, this concentration of influence indicates that the Titanic dataset is highly sensitive to changes in a few key variables. Even minor adjustments in Age or Fare values can produce noticeable shifts in the predicted outcome. This type of sensitivity aligns with the characterization of the dataset as a system with nonlinear interactions, a conclusion already explored during Workshop 2. The presence of missing and irregular information amplifies this sensitivity, which justified the need for a Reliability Layer capable of stabilizing the input space before model training.

5.2 Emergent Patterns in the Cellular Automata Simulation

The Cellular Automata model reproduced physical movement dynamics based on priority rules extracted from the dataset. Although the rules governing the simulation were simple, the system produced identifiable patterns after several iterations. Class 1 agents consistently reached the safe zones first, while Class 3 agents remained blocked near the lower rows of the grid.

This behavior illustrates a form of emergent segregation that is not explicitly programmed but instead arises from the combination of initial placement, probability rules, and local interactions. These emergent patterns reflect real historical inequalities in evacuation procedures, where social class and gender roles influenced access to lifeboats. The simulation therefore provides a spatial interpretation of the same biases captured numerically by the Machine Learning model.

5.3 Comparison of Analytical and Simulation Perspectives

Although the Machine Learning model and the Cellular Automata simulation operate under different paradigms, both identified the same underlying structure. The Random Forest highlighted the statistical weight of socioeconomic and demographic factors, while the Cellular Automata showed how similar factors translate into physical constraints and spatial bottlenecks.

The alignment of these two perspectives strengthens the claim that the Titanic dataset behaves as a constrained and sensitive system whose behavior cannot be explained by randomness alone. Instead, it reflects a clear hierarchy of survival conditions shaped by class, gender, and access to safe areas. This correspondence between numerical influence and spatial emergence was one of the core findings of Workshop 4 and is reaffirmed by the experimental results.

5.4 Impact of Data Quality and Preprocessing Decisions

The preprocessing phase played a crucial role in stabilizing the system. Missing Age and Cabin values created uncertainty that propagated through early experiments, often leading to inconsistent predictions. The imputation strategies adopted in the Reliability Layer reduced this instability and allowed both models to operate on a complete and coherent dataset.

These outcomes confirm the importance of controlling input variability in systems with nonlinear dependencies. The treatment of missing values, the construction of engineered features, and the normalization of continuous variables all contributed to reducing sensitivity to noise. Without these measures, the model would have been more affected by fluctuations in the input data, which was one of the risks identified in the Systems Analysis phase.

5.5 Interpretation Through Systems Analysis Principles

The behavior observed in both simulations highlights several concepts introduced in the Systems Analysis course:

- Interdependence among variables produces collective effects that are not visible when features are analyzed individually.
- Sensitivity to initial conditions creates unpredictable shifts in system output when small perturbations are introduced.
- Constraints such as limited capacity, categorical boundaries, or social hierarchies shape the evolution of the system.
- Emergent behavior appears naturally when simple rules interact repeatedly over time.

These principles provide a broader interpretation of the results, showing that the Titanic dataset is not merely a classification challenge but a representation of a historical system with internal logic and structure. The combination of Machine Learning and Cellular Automata methods gave the project the ability to examine the system from numerical and spatial viewpoints, making the analysis more comprehensive.

5.6 Summary of Discussion

The findings indicate that the Titanic system exhibits predictable patterns that emerge from a limited set of influential variables. The Machine Learning model and the Cellular Automata simulation, despite their different methodologies, both reveal the strong effects of socioeconomic status and demographic attributes on survival outcomes.

The project demonstrates that controlling data instability, understanding variable interactions, and applying system analysis principles are essential for designing models that are reliable and interpretable. These observations reinforce the importance of integrating theoretical analysis with practical implementation, as emphasized throughout the Workshops.

Chapter 6

Conclusions

This project examined the Titanic dataset from a systems perspective in order to design and evaluate a predictive model that is both technically robust and analytically grounded. Through the combination of Machine Learning and Cellular Automata simulations, the study demonstrated that the Titanic problem is not merely a classification task but a representation of a sensitive and partially chaotic system shaped by social and demographic constraints.

The Machine Learning model confirmed that survival outcomes were dominated by a small set of influential variables. Fare, Sex, Age, and Pclass consistently emerged as the most significant predictors, revealing strong inequalities in survival probabilities. These findings highlight that the dataset contains nonlinear dependencies that amplify small variations in the input space, which reinforces the need for a structured and reliable preprocessing workflow.

The Cellular Automata simulation provided a complementary perspective by modeling how social priority rules and spatial constraints translate into emergent patterns during an evacuation. The system produced segregation effects in which higher class agents consistently reached safer areas while lower class agents remained blocked. This behavior supports the idea that simple rules, when applied repeatedly, can generate complex system dynamics that align with historical records.

The integration of these two methodologies emphasizes the importance of viewing the dataset as a dynamic system rather than a static table of values. The layered architecture developed in the Workshops was essential for managing uncertainty, controlling instability, and ensuring reproducibility. The Reliability Layer stabilized the inputs, the Maintainability Layer structured the transformations, and the Usability Layer allowed the system to produce interpretable outputs. Together, these components provided a foundation for consistent analysis and experimentation.

Overall, the project demonstrates that predictive modeling benefits greatly from the principles of systems analysis. Understanding interdependence, sensitivity, and emergent behavior enables the development of models that are not only accurate but also resilient and transparent. The results confirm that technical methods and systemic reasoning work best when combined within a coherent architectural framework.

Future work may include expanding the dataset with external sources, refining the Cellular Automata rules with empirical behavioral studies, and implementing a Feedback Layer capable of detecting model drift over time. These extensions would allow the system to grow beyond the constraints of the Kaggle competition and move toward a more adaptive and generalizable analytical platform.

Chapter 7

Reflection

The development of this project provided an opportunity to explore the Titanic dataset from perspectives that go beyond traditional predictive modeling. Working through the Workshops and implementing the system architecture allowed us to understand how data, processes, and simulation interact within a structured analytical framework. This experience contributed to a broader appreciation of how technical and systemic thinking complement each other in engineering.

One of the main lessons learned was the importance of viewing a dataset as a system with internal relationships, constraints, and sources of uncertainty. Early experiments showed that small inconsistencies in preprocessing or omitted variables could produce significant variations in model performance. This motivated the design of a Reliability Layer and highlighted the value of maintaining control over each step in the workflow. Concepts such as sensitivity, interdependence, and emergent behavior, which were introduced during the Systems Analysis course, became practical considerations rather than abstract ideas.

Another key insight came from comparing the Machine Learning model with the Cellular Automata simulation. Although both approaches aim to describe the same problem, their methods and perspectives are different. Machine Learning helped us identify the variables that influence survival with the greatest weight, while the Cellular Automata model illustrated how these influences can manifest spatially and dynamically. This combination clarified that predictive accuracy alone is not enough to understand a system. Approaching the problem from multiple angles provides a more complete picture of its behavior.

Teamwork was also an essential part of this process. The distribution of tasks and the integration of ideas required coordination and communication among group members. Each stage of the project strengthened our ability to document decisions, evaluate assumptions, and justify methodological choices. These practices will be valuable in future projects where uncertainty and complexity are present.

Looking back, several improvements could be explored in future iterations. Expanding the dataset or incorporating additional contextual information could enrich the analysis. More advanced models may capture interactions that traditional algorithms do not fully represent. Similarly, refining the Cellular Automata rules with empirical behavioral data could increase the realism of the simulation. Implementing an automated monitoring or feedback mechanism would also allow the system to adapt to changing conditions during long-term deployment.

Overall, this project reinforced the importance of combining technical skills with systemic reasoning. The experience demonstrated that reliable model development requires more than algorithms. It demands an understanding of the structure, behavior, and limitations of the system behind the data. This view will continue to guide our approach to future analytical and engineering challenges.

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.