Titanic

Machine learning from disaster

Faculty of Engineering, Systems Engineering

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

Context

Sinking of the Titanic on April 15, 1912

Context

↓

Sinking of the Titanic on April 15, 1912

The goal is to predict which passengers might survive and which might not. Different aspects need to be considered and whether this prediction is accurate. → **Binary system**

0110

CONTEXT

↓

Sinking of the Titanic on April 15, 1912

The goal is to predict which passengers might survive and which might not. Different aspects need to be considered and whether this prediction is accurate. ⟶ BINARY SYSTEM

▯ DOES NOT SURVIVE

Context

↓

Sinking of the Titanic on April 15, 1912

The goal is to predict which passengers might survive and which might not. Different aspects need to be considered and whether this prediction is accurate. → **Binary system**

1 SURVIVE

CONTEXT

Sinking of the Titanic on April 15, 1912

The goal is to predict which passengers might survive and which might not. Different aspects need to be considered and whether this prediction is accurate. → **BINARY SYSTEM**

1 SURVIVE

Functional predictive model based on information provided → **SELF-LEARNING**

Operation

Understanding how machine learning algorithms work

↓

Improve the algorithm

Operation

Understanding how machine learning algorithms work

↓

Improve the algorithm

Considering...

Factors indicated and be careful in handling information, improving the reliability of what is stated

Operation

Understanding how machine learning algorithms work

↓

Improve the algorithm

Considering...

Factors indicated and be careful in handling information, improving the reliability of what is stated

WARNING: The purpose is not only to predict who survived, it is to build a robust system that is adaptable to different contexts.

For this to work...

Identification of the elements

Expected entries

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# 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

For this to work...

Identification of the elements

Expected entries

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# 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

Expected departure

SURVIVE

DOES NOT SURVIVE

For this to work...

Identification of the elements

But it is not that simple, this brings with it a complexity and sensitivity that increases the difficulty.

For this to work...

Identification of the elements

But it is not that simple, this brings with it a complexity and sensitivity that increases the difficulty.

Missing data

↓

Passenger age

For this to work...

Identification of the elements

But it is not that simple, this brings with it a complexity and sensitivity that increases the difficulty.

Missing data

↓

Passenger age

Imbalance

There are more passengers who did not survive than those who did.

For this to work...

Identification of the elements

But it is not that simple, this brings with it a complexity and sensitivity that increases the difficulty.

Missing data

↓

Passenger age

Imbalance

There are more passengers who did not survive than those who did.

A small altered input data can change the accuracy

Additionally, we will consider **Chaos and Randomness**

Additionally, we will consider

Chaos and Randomness

There are unforeseen interactions between passengers

Survival depends on human factors

Additionally, we will consider **CHAOS AND RANDOMNESS**

There are unforeseen interactions between passengers

↓

Survival depends on human factors

All of these are factors that alter the prediction and the precision with which they intend to know the survival of the Titanic passengers.

Additionally, we will consider CHAOS AND RANDOMNESS

There are unforeseen interactions between passengers

↓

Survival depends on human factors

IMPLEMENTING A DESIGN

Robust

Reproducible

Intelligent

Reliable

Modular

Requirements

ID	Requirement	Description
FR-1	Data Ingestion Module	Load and validate input datasets (`train.csv`, `test.csv`, and `gender_submission.csv`).
FR-2	Preprocessing and Cleaning	Handle missing or null values in variables such as `Age`, `Cabin`, and `Embarked`.
FR-3	Feature Engineering	Apply transformations (e.g., one-hot encoding for categorical variables).
FR-4	Model Training	Train a supervised learning model (e.g., Random Forest) to predict `Survived`.
FR-5	Evaluation and Metrics	Compute model accuracy and generate submission file.
FR-6	Submission Output	Export predictions to `submission.csv` following Kaggle's structure.

Table 1: Functional Requirements

ID	Requirement	Description
NFR-1	Performance	Process datasets (1,300 records) in less than 5 seconds.
NFR-2	Scalability	Allow easy addition of features without refactoring.
NFR-3	Reproducibility	Use fixed random seeds and documented dependencies.
NFR-4	Maintainability	Modular architecture separating stages.
NFR-5	Usability	Provide clear workflow and outputs.
NFR-6	Reliability	Handle corrupted input files gracefully.

Table 2: Non-Functional Requirements

Requirements define the functions the system must perform and how it must behave to do so efficiently, reliably, and reproducibly, ensuring stable and well-structured performance.

Requirements

ID	Requirement	Description
FR-1	Data Ingestion Module	Load and validate input datasets (train.csv, test.csv, and gender_submission.csv).
FR-2	Preprocessing and Cleaning	Handle missing or null values in variables such as Age, Cabin, and Embarked.
FR-3	Feature Engineering	Apply transformations (e.g., one-hot encoding for categorical variables).
FR-4	Model Training	Train a supervised learning model (e.g., Random Forest) to predict Survived.
FR-5	Evaluation and Metrics	Compute model accuracy and generate submission file.
FR-6	Submission Output	Export predictions to submission.csv following Kaggle's structure.

Table 1: Functional Requirements

Specific actions that the system must perform.

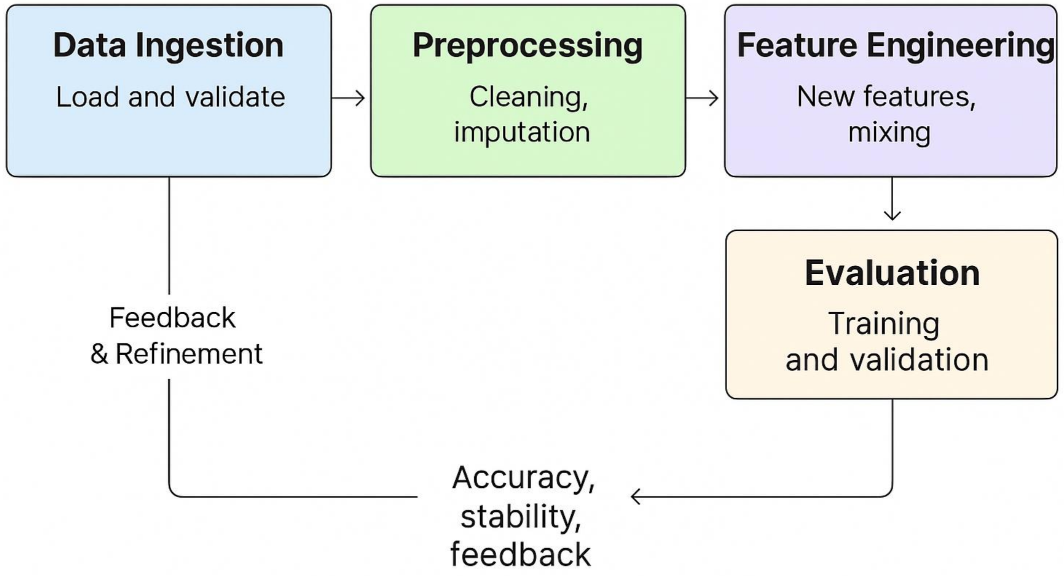
Requirements

These requirements ensure that the system is efficient, reliable, easy to maintain and stable.

ID	Requirement	Description
NFR-1	Performance	Process datasets (1,300 records) in less than 5 seconds.
NFR-2	Scalability	Allow easy addition of features without refactoring.
NFR-3	Reproducibility	Use fixed random seeds and documented dependencies.
NFR-4	Maintainability	Modular architecture separating stages.
NFR-5	Usability	Provide clear workflow and outputs.
NFR-6	Reliability	Handle corrupted input files gracefully.

Table 2: Non-Functional Requirements

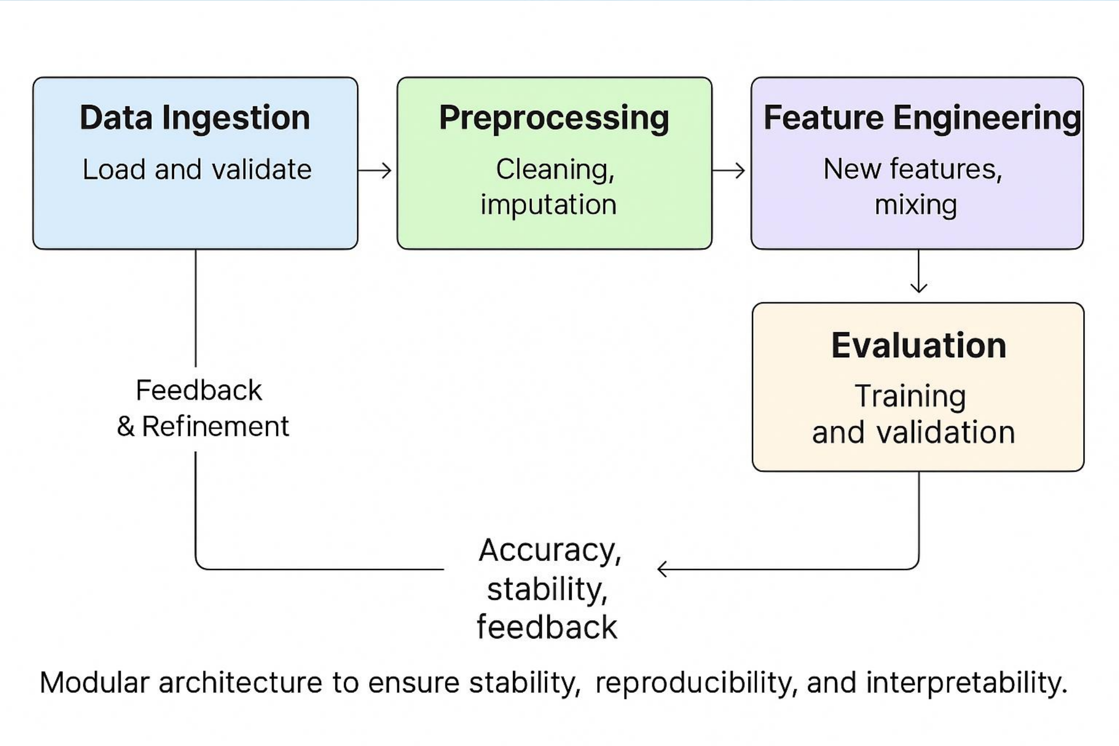
Architecture



Modular architecture to ensure stability, reproducibility, and interpretability.

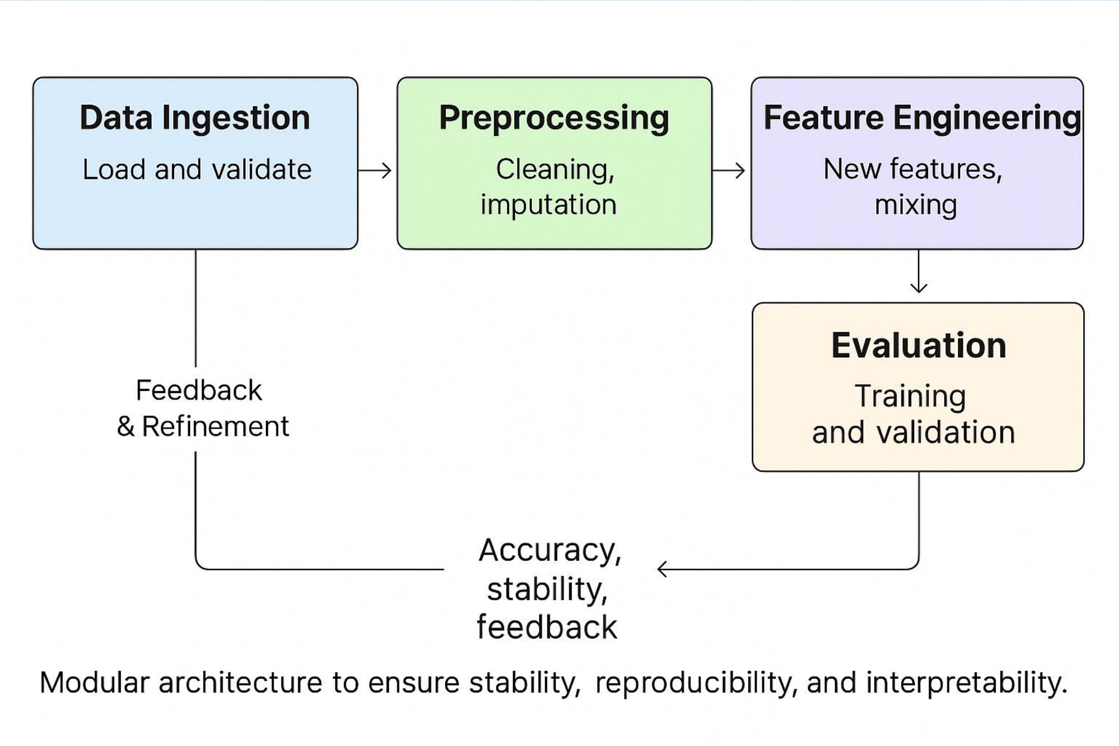
This, managing uncertainty and variability, on the Titanic

Architecture



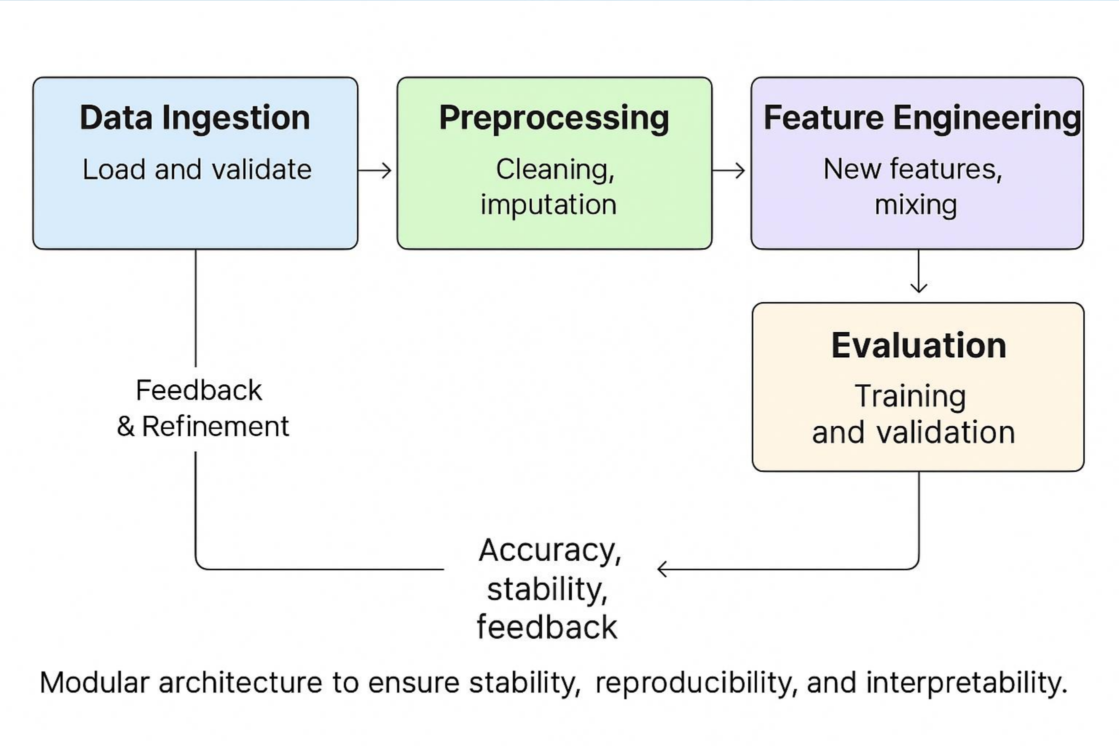
Data Ingestion, where input data is loaded and validated, Preprocessing which cleans, corrects, and transforms information

Architecture



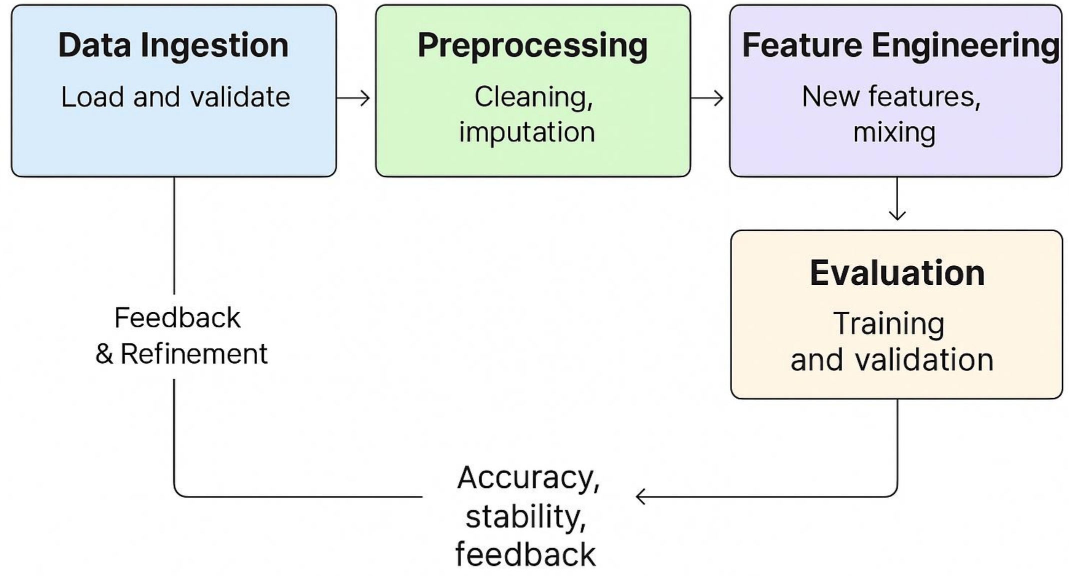
Feature Engineering, where new variables are created to improve the model's predictive capacity. Model Training, which trains and fine-tunes machine learning algorithms

Architecture



and Evaluation, which measures performance and generates feedback to optimize the process, ensuring a stable and reproducible workflow.

Architecture



Modular architecture to ensure stability, reproducibility, and interpretability.

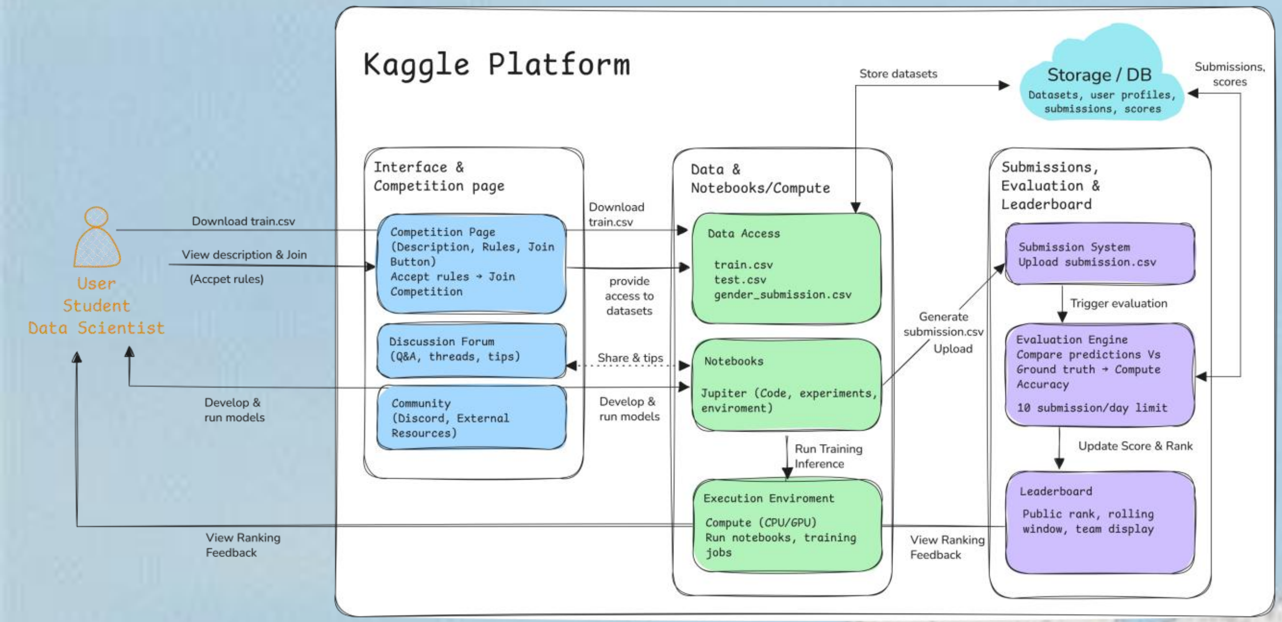
Through...

Include new variables, improving the analysis	&	Analyze and adjust the model to unpredictable behavior

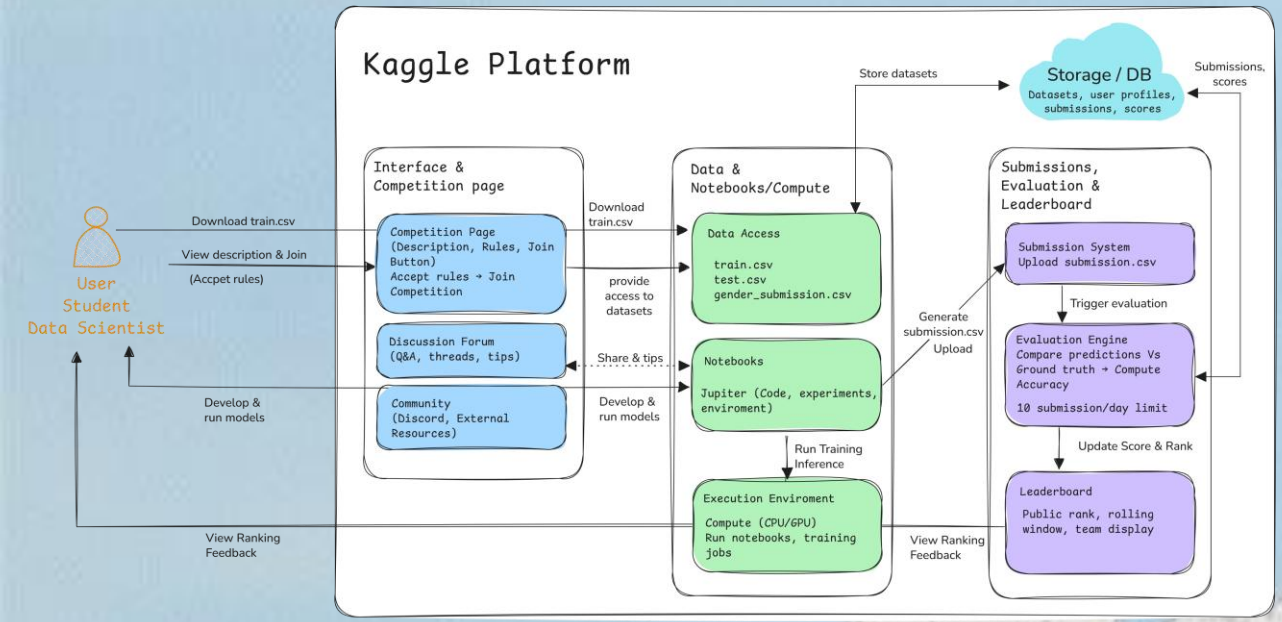
System analysis design

Architecture

Tools and techniques



ARCHITECTURE

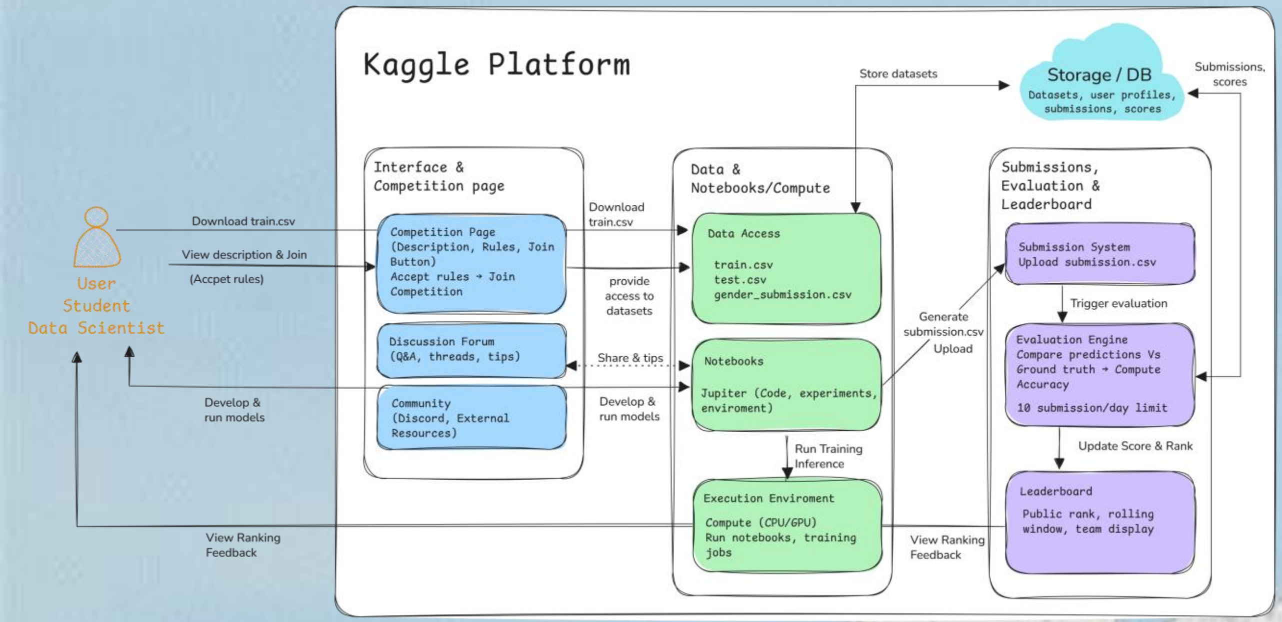


TOOLS AND TECHNIQUES

Using Python as a language

Libraries like Pandas/NumPy

ARCHITECTURE



TOOLS AND TECHNIQUES

Using Python as a language

Libraries like Pandas/NumPy

Pipeline structure

Works independently but connected

System analysis design

Still in progress...

BIBLIOGRAPHY

[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.

AÚN EN PROGRE

BI

[1] Kaggle, "Titanic: Machine … Available: https://www.kaggle.com/com…

[2] A. Cook, "Titanic Tutor… …line]. Available: https://www.kag…

[3] Gusthema, "Titanic Com… 2023. [Online]. Available: ht… titanic-competition-w-tensor…

[4] J. Carvajal Garnica, A. M. C… …Rojas, "Titanic: Machine Learning from D… …)," *Universidad Distrital Francisco José … …ering*, 2025.

[5] C. A. Sierra Virgüez, "Systems Analysis Cour… *…a Distrital Francisco José de Caldas, Faculty of Engineering*, 2025.