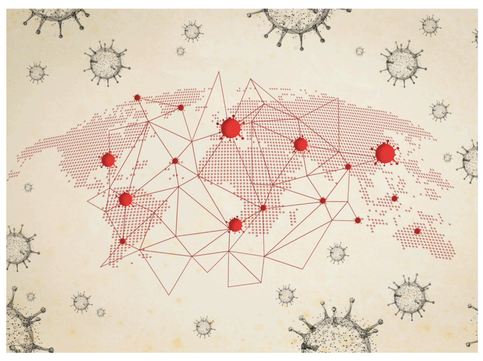
COVID-19 GLOBAL FORECASTING SYSTEM: A MODULAR AND REPRODUCIBLE DATA PIPELINE

DANIEL CASTRO, SANTIAGO VARGAS, DAVID SANCHEZ, DILAN TRIANA

UNIVERSIDAD DISTRITAL
FRANCISCO IOSÉ DE CALDAS

INTRODUCTION

The **COVID-19 pandemic** revealed the need for reliable and reproducible forecasting systems.
Most models focused only on **prediction accuracy**, ignoring reproducibility and system design.
The Kaggle "COVID-19 Global Forecasting" competition provided a standard dataset to test data-driven approaches.
This project applies **systems engineering principles** to design and implement a modular pipeline for **predicting confirmed cases and fatalities over time**.



GOAL

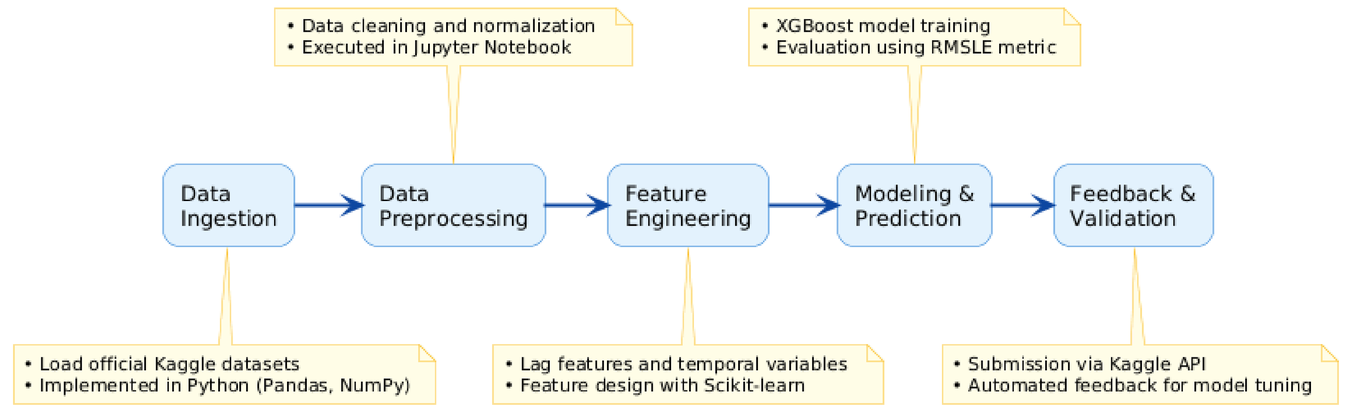
Design and implement a **reproducible, modular, and adaptive forecasting system**
for COVID-19 case prediction, integrating data ingestion, preprocessing,
model training, and **evaluation** under a controlled feedback loop.

RESULTS

Module	Description
Data Ingestion	Loading and verification of Kaggle datasets (train.csv and test.csv).
Preprocessing	Cleaning, normalization, and standardization of input data.
Modeling	Application of predictive algorithms such as Random Forest and XGBoost.
Evaluation	Validation of model outputs using conceptual metrics like RMSLE.
Feedback	Iterative refinement of parameters and pipeline consistency checks.

The results correspond to the **structural analysis and modular design** of the forecasting system.
Through the workshops, a five-stage architecture was defined, integrating **data ingestion**, **preprocessing, modeling, evaluation, and feedback.**
This configuration ensures **consistent data flow, traceability, and adaptability** for future improvements.
The analysis demonstrates the application of **systems engineering principles** to create a **reproducible and scalable predictive framework.**

PROPOSED SOLUTION



The proposed system uses a **five-stage modular pipeline** for **COVID-19 forecasting**. Data from **Kaggle** are ingested, **cleaned**, and transformed into **model-ready features.** Predictions are generated using **XGBoost**, evaluated with **RMSLE**, and refined through an **automated feedback loop** via the **Kaggle API**.

EXPERIMENTS

Instead of numerical experiments, the project focused on analyzing different system configurations to assess **integration and consistency.** The evaluation compared three conceptual setups: a **baseline version** representing a simple regression approach, an **improved version** incorporating data cleaning and lag features, and an **adaptive version** that included a **feedback mechanism** through the **Kaggle API.** This analysis validated the **logical flow and interoperability** of all modules in the proposed pipeline.

CONCLUSIONS

The proposed system fulfills its objective of applying **systems engineering principles** to the design of a **modular and reproducible forecasting framework.**
The structured analysis confirmed the **coherence between the modules** and the **feasibility of maintaining data integrity, traceability, and continuous improvement** through feedback.
This **modular design ensures reproducibility** and **easy future updates** as new data become available.

Kaggle. (2020). COVID-19 Global Forecasting (Week 1–5) [Dataset]. Kaggle.
Scikit-learn Developers. (2023). Scikit-learn: Machine learning in Python [Software documentation]. Scikit-learn.
Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.