A Systems Engineering Approach

Johan S. Beltrán Merchán, Edison D. Álvarez Varela, Yader I. Quiroga Torres, Julián D. Celis Giraldo

Systems Engineering Department
Universidad Distrital Francisco José de Caldas

1. Introduction: The Systems Problem

Forecasting daily traffic for approximately **145,000 Wikipedia articles** is a large-scale Systems Engineering challenge characterized by:

- **Massive Scale:** Requires high computational efficiency (**Scalability**).
- **Chaos Factors:** High volatility due to viral or unpredictable events.
- **Heterogeneity:** No single model fits all series.

The architecture must be adaptive and robust to minimize the **SMAPE** metric.

2. Goal: Adaptive Forecasting Architecture

Research Question: How can a scalable and maintainable system architecture minimize SMAPE across heterogeneous time series using **Systems Engineering Principles**?

Expected Product: A **Modular Monolith** based on a **Hierarchical Ensemble** that dynamically selects the best forecasting model per article.

Performance Metric

Symmetric Mean Absolute Percentage Error (SMAPE):

$$SMAPE = \frac{100}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

3. Proposed Solution: Architecture and Patterns

A **Modular Monolith** with clear separation of concerns, anchored by two design patterns:

System A: Data Flow Integrity

- **Pattern:** Chain of Responsibility.
- **Function:** Defines a linear 9-module pipeline (*Ingestion* $\rightarrow$ *Feedback*) ensuring traceability and data consistency.

System B: Adaptive Forecasting

- **Principle:** Equifinality (multiple valid pathways to the goal).
- **Implementation:** Hierarchical Ensemble with the Strategy Pattern.

Hierarchical Ensemble Breakdown

- **Level 1 (Strategies):** Base models (**ARIMA**, **Prophet**, **LSTM**).
- **Level 2 (Meta-Model):** Analyzes metadata (*language*, *volatility*) and dynamically selects the optimal model.

Scalability: Achieved via parallel processing with `Joblib`, distributing models across multiple CPU cores.

Architectural Blueprint (Data Flow)

Insert a high-resolution diagram of the 9-module pipeline and ensemble structure here.

4. Validation and Testing Philosophy

Rigorous testing ensures robustness and maintainability:

- **Unit Tests:** Ensure deterministic behavior and prevent data leakage.
- **Integration Tests:** Validate the Chain of Responsibility pipeline.
- **Acceptance Testing:** Evaluate overall SMAPE performance.

5. Results Projected: Granular Analysis

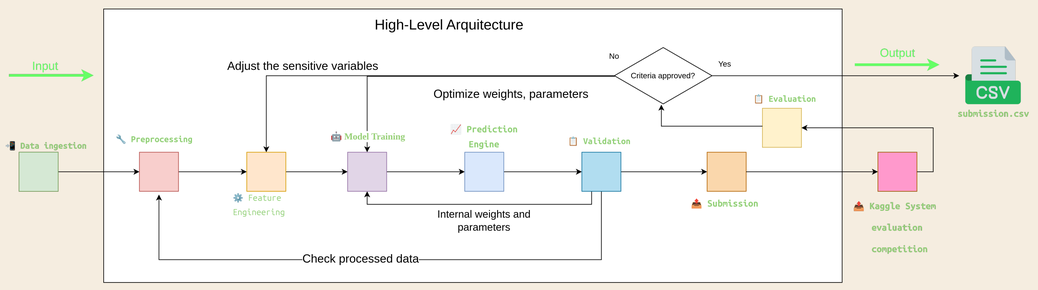
The **Evaluation Module** produces a **Stratified Post-Prediction Analysis** feeding insights back into the system.

This analysis refines feedback loops for volatile or underperforming subgroups.

6. Conclusion and Future Work

The architecture provides a robust, scalable, and adaptive solution for chaotic web traffic forecasting.

- Chain of Responsibility ensures data integrity.
- Hierarchical Ensemble ensures adaptive model selection.



Conceptual Archivtecture: Inputs, Core Processes, and Quality Loop

Future Work: Integrate advanced models (**Neural Networks**, **Random Forests**) as new strategies within the Ensemble framework to enhance modularity and predictive accuracy.

Acknowledgments and References

The team thanks Professor Carlos Andrés Sierra Virgüez for his guidance.

1. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, 1948.
2. Additional references to be added.