Systems Analysis and Design
Workshop No. 1
Systems Engineering Analysis
Psychopathy Prediction Competition on Twitter

Juan David Bejarano Cristancho 20232020056

Raúl Andrés Diaz Losada 20232020058

Juan David Avila 20232020154

David Sanchez Acero 20232020049

Systems Engineering Program

Francisco José de Caldas District University

September 27, 2025

1. Competency Description

The selected Kaggle competition, "Predicting Psychopathy Based on Twitter Usage," poses a complex machine learning challenge that exemplifies the intersection between social media analysis, psychological assessment, and predictive modeling systems. Organized by the Online Privacy Foundation in 2012, this competition sought to determine the feasibility of predicting levels of psychopathy in individuals solely based on their Twitter usage patterns and linguistic characteristics.

1.1. Competition Objectives and Structure

The main objective is to identify Twitter users with high psychopathy scores, defined as those with a score of 2 standard deviations above the mean (1.98), which represents approximately the 3 % of the population sample. This creates a highly unbalanced classification problem, characteristic of many real machine learning systems.

The competition provided participants with:

- An anonymized dataset of 2,927 Twitter users from 80 countries.
- 337 variables derived from Twitter data, usage patterns, and linguistic analysis.
- Self-reported psychopathy scores based on the scale developed by Professor Del Paulhus of the University of British Columbia.
- Over 3 million tweets for analysis.

1.2. Data Set Structure and Restrictions

The dataset represents a sophisticated data engineering system with multiple layers of complexity:

1. **Privacy Restrictions**: All identifiable information has been removed to protect user privacy.
2. **Feature Engineering**: The 337 variables were derived from raw Twitter data through complex linguistic and behavioral analysis.
3. **Unbalanced Distribution**:Only 3 % of users were classified as high psychopaths, creating significant modeling challenges.
4. **Temporal Dynamics**: The data captures usage patterns over time, introducing temporal dependencies.

2. Systems Analysis Report

2.1. Systemic Analysis: Elements and Relationships

The psychopathy prediction system on Twitter can be broken down into several interconnected subsystems, as illustrated in Figure 1:

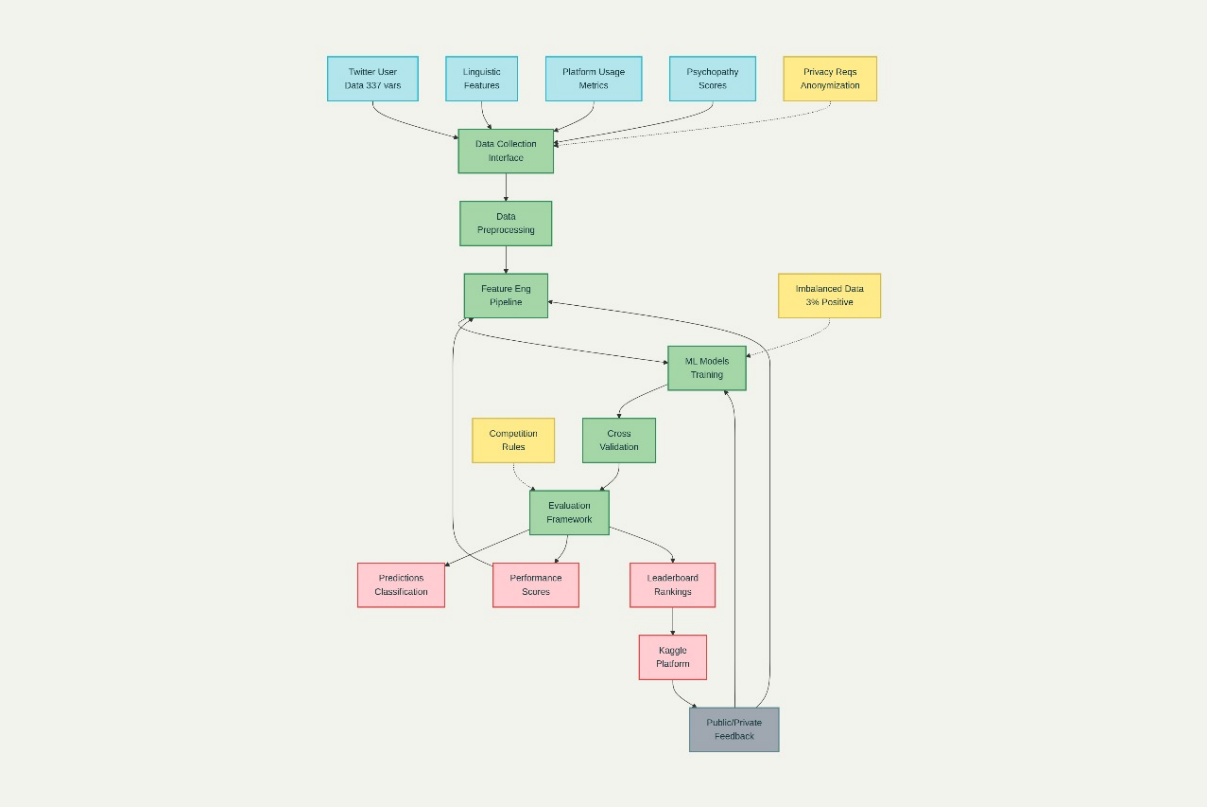

Figura 1: System architecture of the Psychopathy Prediction Competition on Twitter

2.1.1. Fundamental Elements of the System

1. Data Collection Subsystem

- Integration with the Twitter API for user data extraction
- Psychological assessment instruments (Paulhus scale)
- Anonymization and privacy filtering processes
- Quality assurance and validation mechanisms

2. Feature Engineering Pipeline

- Linguistic analysis engines (sentiment, word frequency, syntactic patterns))
- Calculation of behavioral metrics (posting frequency, social connections))
- Platform usage analysis (Klout scores, interaction patterns))
- Algorithms for temporal pattern extraction

3. Machine Learning Framework

- Infrastructure for model training
- Cross-validation systems
- Modules for hyperparameter optimization
- Capacidades para aprendizaje en conjunto (ensemble)

4. Assessment and Competency Platform

- Kaggle submission system
- Average accuracy scoring mechanism
- Dual leaderboard systems (public/private))
- Performance monitoring and analysis

2.1.2. Relationships between Elements

The system exhibits complex interdependencies:

1. **Data Flows**: Raw Twitter data is processed in the pipeline to generate structured inputs for the models.
2. **Feedback**: Model performance metrics influence the selection and design of features
3. **Control**: Competition rules and privacy constraints govern all operations
4. **Temporal Relationships**: Historical usage patterns feed into model training

2.2. Complexity and Sensitivity Analysis

2.2.1. Critical Sensitivity Points

The system shows high sensitivity to several critical parameters:

1. Sensitivity in Feature Selection

- Small changes in linguistic weights can significantly impact model performance
- The correlation between variables introduces multicollinearity issues

2. Sensitivity in Data Quality

- Missing or corrupted data can generate cascading effects in the pipeline
- Anonymization may remove key predictive signals
- Self-assessment biases in psychopathy scores create uncertainties in the data

3. Sensitivity in Model Architecture

- The selection of hyperparameters has exponential effects on performance
- The cross-validation strategy dramatically impacts evaluation
- Ensemble schemes create nonlinear performance landscapes

2.2.2. Sources of Complexity

The complexity of the system arises from multiple sources:

1. **Dimensional Complexity**: 337 features generate high-dimensional spaces with exponential search complexity
2. **Temporal Complexity**: Sequential usage patterns introduce dependencies
3. **Linguistic Complexity**: Natural language processing introduces semantic and syntactic ambiguity
4. **Behavioral Complexity**: Human patterns have nonlinear and context-dependent dynamics

2.3. Chaos and Randomness: Nonlinear Dynamics

2.3.1. Observed Chaotic Aspects

The competition system exhibits several traits of chaotic and complex adaptive systems:

1. Sensitivity to Initial Conditions

- Small variations in preprocessing lead to very different model results
- Random seeds generate divergent trajectories
- The selection of folds for cross-validation introduces unpredictable variance

2. Emergent Behaviors

- Competitive dynamics among participants generate strategies and emergent meta-learning
- Feedback loops from the public leaderboard cause collective overfitting
- Significant feature patterns emerge that were not explicitly designed

3. Nonlinear Feedback Loops

The system includes several feedback mechanisms with nonlinear behavior:

- **Performance Loop**: Leaderboard scores → Model adjustments → (Non-proportional) changes in performance
- **Engineering Loop**: Performance → Feature modification → Changes in data representation
- **Competitive Loop**: Participant strategies → Leaderboard dynamics → Strategic evolution

2.3.2. Application of Chaos Theory

Several concepts of chaos apply to this system:

1. Attractors and Phase Space

- Performance trajectories in the hyperparameter space may exhibit strange attractors
- Local optima create zones where algorithms become trapped
- The state of the system evolves in a phase space defined by parameters and metrics

2. Bifurcation Points

- Thresholds in regularization parameters cause abrupt changes in the model
- Cutoffs in feature selection produce discontinuous performance landscapes
- The transition from the public to the private leaderboard represents a bifurcation of the system

3. Fractal Properties

- Error surfaces in high-dimensional spaces exhibit self-similar structures
- Linguistic patterns on Twitter present fractal-like hierarchies
- Performance optimization paths display recursive patterns

2.3.3. Unpredictable Elements

Several aspects of the system are resistant to prediction:

1. **Human Variability**: Users exhibit linguistic evolution and inconsistent posting patterns
2. **Social Media Dynamics**: Platform changes and trends alter data distributions over time
3. **Noise in Psychological Assessments**: Self-reported scores possess inherent uncertainties
4. **Competitive Interactions**: Participant strategies generate game-like dynamics with unpredictable outcomes

2.4. Systems Engineering Perspective

2.4.1. Requirements Engineering

The system requirements can be categorized from a systems engineering perspective:
Functional Requirements:

- Accurately predict psychopathy scores with measurable precision
- Process 337 input variables for 2,927 users
- Handle highly imbalanced classification (3 % positive class)
- Support real-time model evaluation and ranking

Non-Functional Requirements:

- Preserve privacy through anonymization
- Computational scalability for multiple participant submissions
- Reproducibility and transparency in evaluation
- Robustness against overfitting and manipulation strategies

2.4.2. Architectural Considerations

The system architecture follows several engineering principles:

1. **Modularity**: Clear separation between data processing, modeling, and evaluation
2. **Scalability**: Distributed capacity for processing multiple submissions
3. **Reliability**: Redundant evaluation mechanisms (public/private leaderboards)
4. **Maintainability**: Standardized data interfaces and formats

2.4.3. System Lifecycle

The system presents differentiated phases in its lifecycle:

- **Conception**: Problem definition and data collection planning
- **Development**: Construction of the feature engineering pipeline
- **Deployment**: Platform launch and participant registration
- **Operation**: Active competition period and continuous evaluation
- **Decommissioning**: Final evaluation and publication of results

3. Conclusion

3.1. Main Findings

The competition represents a sophisticated socio-technical system that highlights several important principles of systems engineering:

1. **Emergent Complexity**: System behaviors arise from the interaction of its components
2. **Nonlinear Dynamics**: Small changes in input parameters can produce large variations in outcomes
3. **Adaptive Behavior**: The competitive environment creates a co-evolutionary dynamic between participants and the evaluation system
4. **Multiscale Interactions**: The system operates from the level of individual linguistic features to population-level patterns

3.2. System Strengths

1. Robust Evaluation Framework

- The dual leaderboard system prevents overfitting
- The average precision metric adequately handles imbalance
- The cross-validation requirement encourages model generalization

2. Privacy-Preserving Design

- Anonymization protects participant privacy
- Derived features maintain predictive capacity without identifiable information
- The ethical framework aligns with responsible AI principles

3. Scientific Rigor

- Validated psychological instruments (Paulhus scale)
- Large-scale dataset provides statistical power
- Reproducible methodology enables scientific validation

3.3. System Weaknesses

1. Data Distribution Challenges

- Extreme imbalance hinders modeling
- Self-assessment bias compromises ground truth quality
- Temporal collection may introduce biases

2. Feature Engineering Limitations

- High dimensionality increases the risk of overfitting
- Derived features may lose relevant context
- Linguistic analysis may overlook cultural and demographic variations

3. Generalization Concerns

- Twitter-specific patterns may not generalize to other platforms
- Temporal data may not remain valid
- Cultural and linguistic biases may limit applicability

3.4. Implications for Complex Systems

This analysis reveals key implications for complex data science systems:

1. **Chaos-Aware Design**: Systems must accommodate and leverage chaotic dynamics
2. **Adaptive Evaluation**: The evaluation framework must evolve to prevent manipulation
3. **Multiscale Validation**: Validation must occur across multiple scales and temporal horizons
4. **Ethical Integration**: Privacy and ethics must be integrated from the architectural design stage

The competition demonstrates that complex machine learning systems exhibit characteristics of adaptive chaotic systems. Understanding these dynamics from a systems engineering perspective provides insights for designing, implementing, and evaluating similar socio-technical systems in the future.

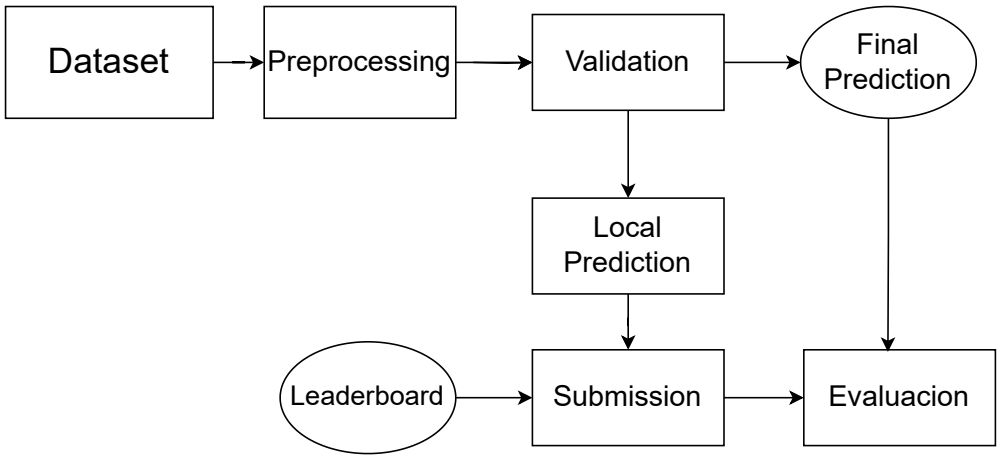

Figura 2: Data flows

4. Data Flow

As shown in Figure 2, the process begins with the *dataset*, followed by the preprocessing phase, predictions, and finally the evaluation in the *leaderboard*.

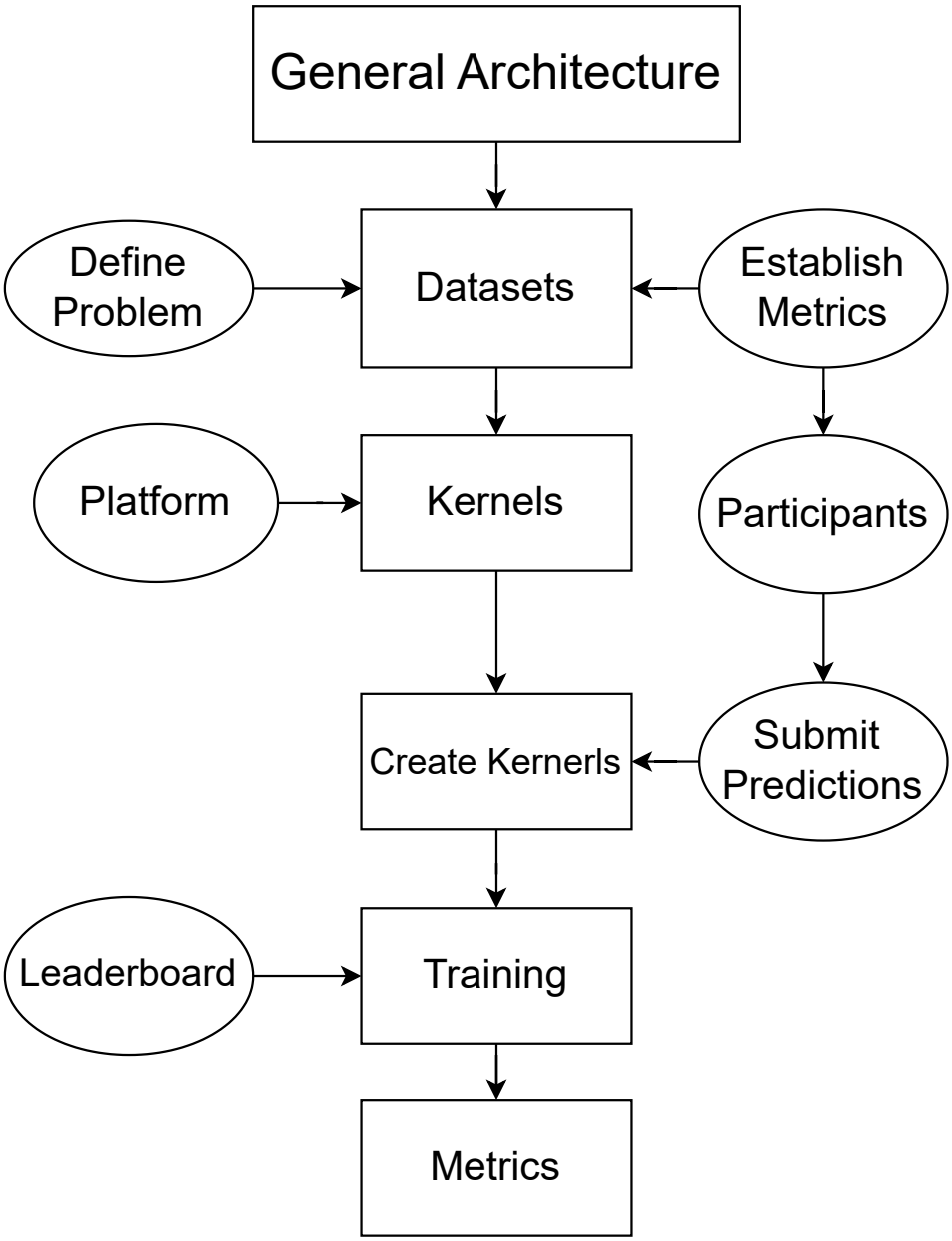

Figura 3: General Architecture

5. General Architecture

As shown in Figure 3,the structure of the competition includes the organizers, who define the problem, provide the *datasets*, and establish the metrics. Participants create *kernels*, train models, and submit predictions to the platform, which manages the *leaderboard* and the final evaluations.

6. References

1. Online Privacy Foundation. (2012). Psychopathy Prediction Based on Twitter Usage. Kaggle Competition.

2. Paulhus, D.L. & Jones, D.N. (2011). Introducing a short measure of the Dark Triad. Poster presented at the meeting of the Society for Personality and Social Psychology, San Antonio.

3. Gosling, S.D., Rentfrow, P.J. & Swann, W.B., Jr. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37, 504-528.

4. Kossiakoff, A., et al. (2011). Systems Engineering Principles and Practice. Wiley.

5. Razavi, S., et al. (2021). The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. Environmental Modelling & Software, 137, 104954.

6. Borgonovo, E. & Plischke, E. (2016). Sensitivity analysis: A review of recent advances. European Journal of Operational Research, 248(3), 869-887.