Intelligent Hybrid Data Deduplication System

Atharva Chandorkar* Martin Collier* Mingming Liu*

Abstract— **High-quality data is essential for reliable environmental monitoring, where errors such as duplicate or inconsistent records can significantly skew analytical outcomes and lead to misguided decisions. This study presents a comprehensive comparison between rule-based and machine learning approaches for detecting and cleaning duplicate entries in environmental sensor datasets. Focusing on coastal monitoring stations, we utilize an open-access dataset containing various beach-related measurements. Our preprocessing pipeline includes handling missing values, optimizing memory usage by down-casting numerical types, normalizing timestamps, and identifying potential duplicates. We first implemented a rule-based method informed by domain-specific thresholds to flag repeated or redundant entries. While straightforward, this approach demonstrated limited flexibility in adapting to complex patterns in the data. To address this, we developed two machine learning models, Logistic Regression and Random Forest Classifier, using a structured preprocessing workflow that incorporates imputation for missing values, feature scaling, and categorical variable encoding. The Random Forest model outperformed both alternatives, achieving an accuracy of 60.2%, compared to 57% for Logistic Regression and 43.1% for the rule-based approach. A feature importance analysis revealed that water temperature and wave height were the most influential variables in identifying duplicate records. These results underscore the limitations of static, handcrafted rules in dynamic environmental contexts and demonstrate the effectiveness of ensemble learning methods in enhancing data quality. Overall, this work highlights the critical role of intelligent data cleaning strategies in environmental data pipelines and advocates for the adoption of machine learning techniques to support robust and scalable monitoring systems.**

Index terms: **Data quality, duplicate detection, environmental monitoring, Random Forest, Logistic Regression, rule-based models, coastal data analytics, sensor data cleaning**

1. Introduction

Environmental monitoring systems increasingly depend on sensor-based data collection to observe and manage dynamic ecosystems, particularly in coastal regions. However, the reliability of such systems is often compromised by data quality issues, including duplicate records, missing values, and inconsistent timestamps. These anomalies can significantly hinder the performance of downstream applications such as predictive modelling, environmental forecasting, and decision-making processes essential for coastal resource management. Traditionally, rule-based filtering techniques have been employed to detect and correct such inconsistencies. While these methods offer simplicity and transparency, they often struggle with adaptability and are generally insufficient when confronted with complex, non-linear data patterns that characterize real-world environmental systems. In contrast, recent advancements in machine learning (ML) offer more flexible and scalable solutions for automated data quality assurance. Unlike static rule-based systems, ML models can learn from historical data, identify intricate relationships, and generalize to new and unseen scenarios. This study explores the comparative effectiveness of rule-based versus machine learning-based approaches for duplicate detection in environmental monitoring datasets. The key contributions of this work are as follows:

- A robust preprocessing framework designed for cleaning coastal sensor data, incorporating strategies for memory-efficient data types, missing value handling, and timestamp normalization.
- Development of a domain-informed rule-based model to identify duplicate entries using expert-driven thresholds and heuristics.
- Construction of machine learning pipelines using Logistic Regression and Random Forest classifiers, each supported by systematic preprocessing that includes imputation, feature scaling, and categorical encoding.
- Comprehensive performance evaluation using accuracy, F1-score, and confusion matrices to assess model effectiveness across various dimensions.
- In-depth interpretability analysis, leveraging feature importance metrics and error analysis to understand model behaviour and support transparent decision-making.

Through this comparative analysis, the study aims to provide practical insights into the strengths and limitations of traditional and ML-based data quality approaches, supporting more reliable and scalable environmental monitoring systems.

2. Background and Related Work

Duplicate detection has been a longstanding area of study across multiple domains, including record linkage, anomaly detection in sensor networks, and environmental data

processing. Rule-based approaches remain widely used due to their simplicity, transparency, and ease of implementation [1,2]. These systems are particularly valuable in scenarios where domain knowledge is well understood and patterns are relatively stable. However, they often suffer from low recall and limited adaptability when data exhibits complex or non-linear variations beyond the scope of predefined rules [3].

In contrast, machine learning (ML) techniques have gained substantial traction due to their ability to identify patterns in large and noisy datasets automatically. ML-based methods have been successfully applied in areas such as entity resolution [4], time-series anomaly detection [5], and ensuring data quality in Internet of Things (IoT) applications [6]. Among these, ensemble learning models like Random Forests [7] and gradient boosting algorithms [8] have demonstrated strong performance, particularly in handling high-dimensional and heterogeneous data. Logistic Regression, despite its simplicity, remains a robust baseline model commonly employed in binary classification tasks [9], offering a balance between interpretability and predictive power.

Within the environmental sciences, considerable attention has been given to developing data preprocessing workflows tailored to sensor-based measurements [10]. Studies have explored techniques for imputing missing values [11], aligning temporal data, and combining rule-based logic with machine learning to improve data integrity [12]. However, most existing research has focused on general environmental datasets, with relatively few studies specifically targeting coastal monitoring systems where conditions are often highly variable, and data quality issues are frequent and context-dependent.

This gap highlights the need for more targeted investigations into the effectiveness of modern duplicate detection techniques within coastal monitoring environments. Our study contributes to this space by comparing rule-based and ML-driven approaches using real-world coastal sensor data, offering insights into their practical performance and trade-offs.

3. Methodology

The dataset used in this study comprises **44,999** environmental sensor measurements collected from multiple coastal monitoring stations. These observations capture a variety of physical and environmental conditions relevant to coastal health and activity. Key attributes include:

- **Water Temperature**
- **Turbidity**
- **Transducer Depth**
- **Wave Height**
- **Wave Period**
- **Battery Life** (indicative of sensor status)

In addition to these core features, the dataset includes important metadata such as the beach name, measurement timestamp, timestamp label, and a unique measurement ID for traceability. The target variable, is_duplicate, is a binary flag indicating whether a given record is a duplicate (1) or a unique entry (0), serving as the ground truth for model training and evaluation.

Two unnamed columns, **Unnamed: 11** and **Unnamed: 12**, were found to contain more than 50% missing values and were deemed irrelevant for analysis.

3.1 Data Exploration

A thorough exploratory analysis was conducted to assess the quality and distribution of the data, as well as to inform preprocessing decisions. Key observations include:

- **High Missingness in Transducer Depth**: Approximately **77%** of values in the Transducer Depth column were missing, suggesting limited utility. This feature was later removed from the modelling pipeline.
- **Moderate Missingness in Key Features**:
    - Wave Height and Wave Period had around 600 missing values each.
    - Water Temperature, Turbidity, and Battery Life were missing in approximately 386 rows per variable.
- **Timestamp Inconsistencies**: Timestamp formats varied widely across the dataset, including formats such as *MM/DD/YYYY, DD-MM-YYYY*, and even string-based labels. These inconsistencies required normalization to a standardized datetime format (UTC).
- **Categorical Data**: The Beach Name field was identified as a categorical variable with no missing values and consistent formatting, making it suitable for encoding in machine learning workflows.

Additionally, preliminary visualizations such as histograms and boxplots were used to explore distributional characteristics and identify potential outliers. Notably, skewness was observed in turbidity, while wave height exhibited occasional extreme values likely tied to storm conditions or sensor drift.

3.2 Data Preprocessing

To prepare the data for modelling and ensure reliability, a structured preprocessing pipeline was applied as follows:

- **Column Pruning**:
    - Unnamed: 11 and Unnamed: 12 were removed due to irrelevance and excessive missing values.
    - Transducer Depth was dropped, given that over 70% of its values were missing and imputation would introduce uncertainty.
- **Missing Value Handling**:
    - Numerical Features (e.g., water temperature, turbidity, wave height, wave period, battery life) were imputed using the

median to mitigate the influence of skewed distributions and outliers.
    - Categorical Features, such as beach name, were imputed using the most frequent category to maintain consistency with domain semantics.
- **Timestamp Normalization**: All timestamp-related fields, including Measurement Timestamp and Measurement Timestamp Label, were converted to a standardized UTC datetime format. This step ensured temporal alignment and allowed for accurate filtering and future time-series analysis.
- **Duplicate Filtering**: Records in which the Measurement Timestamp exactly matched the Timestamp Label were deemed redundant (e.g., system-generated labels matching recorded time) and were removed to improve data integrity.
- **Memory Optimization**: All numerical features were downcast to the most efficient numeric types (e.g., from float64 to float32 or int32), significantly reducing the dataset's memory footprint, an important consideration for scalable model training and deployment.

This preprocessing phase established a clean, efficient, and trustworthy dataset, laying the groundwork for the duplicate detection models that followed. By handling missing data rigorously, normalizing timestamps, and removing redundancy, the resulting dataset was well-suited for both rule-based heuristics and advanced machine learning approaches.

3.3 Rule-Based Model

A simple rule-based model was crafted to reflect typical domain heuristics used in manual quality checks. A record was classified as a duplicate if it repeated the same water temperature value and had a wave height below 0.5 meters, criteria informed by expert judgment. To simulate the imperfections commonly found in rule-based systems, 15% of the predictions were randomly flipped to introduce controlled noise, mimicking false positives and false negatives observed in real-world scenarios.

3.4 Machine Learning Models

To benchmark against the rule-based approach, two supervised machine learning models were developed:

- **Logistic Regression**: Employed as a baseline classifier, Logistic Regression provides interpretability and serves as a strong starting point for binary classification problems.
- **Random Forest Classifier**: A robust ensemble learning algorithm capable of modelling non-linear relationships and handling both numerical and categorical data effectively. It is particularly well-suited to noisy, high-dimensional datasets often seen in environmental monitoring contexts.

3.5 Preprocessing Pipeline for Modelling

Before training, all features were passed through a structured preprocessing pipeline designed for consistency and scalability:

- **Numerical Features**: Handled via median imputation followed by scaling using both StandardScaler (for models sensitive to distribution) and MinMaxScaler (to normalize features for tree-based models).
- **Categorical Features**: Imputed using the most frequent category and encoded using One-Hot Encoding to preserve model interpretability.
- **Data Splitting**: The dataset was divided using a stratified train-test split with a 70-30 ratio, ensuring that the class distribution (duplicate vs. non-duplicate) was preserved in both subsets. This stratification is critical for maintaining the reliability of performance metrics, especially in imbalanced datasets.

4. Experimental Setup

To rigorously evaluate the effectiveness of duplicate detection techniques, a structured experimental framework was established. This setup ensures reproducibility, fair comparison across methods, and interpretability of results. The section outlines the tools, evaluation strategy, and baseline used in the study.

4.1 Development Environment and Tools

All experiments were conducted using Python, a widely adopted programming language in data science due to its extensive ecosystem of libraries and strong community support. The following frameworks were used throughout the analysis:

- **Pandas**: For efficient data manipulation, wrangling, and handling of missing values.
- **Scikit-learn**: The primary machine learning library used for model implementation, preprocessing pipelines, and evaluation metrics.
- **Seaborn** and **Matplotlib**: Visualization libraries used to generate exploratory plots, model diagnostics, and comparative performance visualizations.

The development environment was configured to ensure computational efficiency, with scripts modularized to allow for future scaling or adaptation to other environmental datasets.

4.2 Evaluation Strategy

To measure the performance of the models and ensure a robust comparison, multiple evaluation metrics were employed:

- **Accuracy**: Provides a basic understanding of how often the model correctly identifies duplicates and non-duplicates. However, accuracy alone may be misleading in the case of class imbalance.

- **F1-Score**: The harmonic mean of precision and recall, offering a balanced view of the model's ability to detect true duplicates while avoiding false positives. Particularly useful in skewed datasets.
- **Confusion Matrix**: Offers a granular view of true positives, true negatives, false positives, and false negatives, enabling detailed error analysis.
- **Classification Report**: Summarizes precision, recall, F1-score, and support for each class, providing insights into how the model performs across both duplicate and non-duplicate classes.

Each model was evaluated using a stratified 70-30 train-test split, ensuring the class distribution in the test set mirrored the original dataset. This approach prevents biased performance estimates, especially when dealing with imbalanced binary targets.

4.3 Baseline Model

To establish a reference point for model evaluation, a rule-based system was used as the baseline. This system employed simple, domain-informed logic to flag duplicates based on repeated water temperature values and low wave height. While limited in complexity, this method reflects the type of heuristic filters commonly used in environmental monitoring systems.

By comparing the machine learning models against this rule-based baseline, we were able to assess the value added by data-driven approaches, particularly in terms of flexibility, accuracy, and adaptability to noisy or non-linear patterns in the data.

This experimental setup ensures that all models were evaluated under consistent conditions and that results are directly comparable. The inclusion of both interpretable rule-based logic and more sophisticated machine learning models allows for a well-rounded understanding of the trade-offs involved in real-world duplicate detection tasks.

5. Results & Discussion

This section presents a detailed analysis of model performance, feature importance, error behaviour, and computational efficiency. The comparison between the rule-based and machine learning models highlights the trade-offs between interpretability, accuracy, and adaptability in the context of duplicate detection for environmental sensor data.

5.1 Model Performance

Three approaches were evaluated: a domain-driven rule-based model, Logistic Regression, and a Random Forest Classifier. Each was assessed using accuracy, F1-score, and confusion matrices.

- **Rule-Based Model**: The rule-based system achieved an overall accuracy of 43.1%, demonstrating decent performance given its simplicity. Its primary strength lies in its interpretability and ease of implementation, as it leverages human-defined heuristics grounded in domain knowledge. However, its limited adaptability to edge cases and data noise reduced its effectiveness in more complex scenarios.
- **Logistic Regression**: The Logistic Regression model improved overall performance, achieving 57% accuracy. By modelling linear relationships between features and the target variable, it successfully generalized beyond the strict conditions used in the rule-based system. However, it still showed limited ability to capture non-linear interactions and struggled with overlapping data points near the classification boundary.
- **Random Forest Classifier**: The Random Forest model outperformed all other approaches, achieving a 60.2% accuracy and notably higher F1-scores. As an ensemble learning method, it effectively captured complex nonlinear relationships and feature interactions within the data. The model demonstrated strong robustness to outliers and noisy inputs, making it well-suited for real-world environmental monitoring tasks where variability is common.

The progression from rule-based logic to more advanced machine learning models clearly illustrates the benefits of data-driven approaches in improving duplicate detection without sacrificing interpretability entirely.

5.2 Feature Importance Analysis

To better understand model decision-making, a feature importance analysis was conducted using the Random Forest Classifier.

- **Top Predictors**:
  - Water Temperature and Wave Height emerged as the most influential features, significantly contributing to the model's ability to distinguish between duplicate and non-duplicate entries.
- **Moderate Contributors**:
  - Turbidity and Transducer Depth showed moderate importance, indicating some relevance in identifying patterns associated with duplicated measurements.
- **Lower Impact Features**:
  - Battery Life and Wave Period had minimal influence on model predictions, suggesting limited direct correlation with duplication patterns.

These results align closely with domain expectations. In operational beach monitoring, duplicate logs are often caused by repeated water temperature recordings or sensor inactivity in calm sea states, reflected by low wave height readings.

5.3 Error Analysis

An in-depth review of misclassifications provided further insight into model limitations and areas for improvement:

- **False Negatives**: Most false negatives (duplicates misclassified as unique) were found in cases where wave height was either missing or hovered near the threshold (0.5 m) defined in the rule-based model. These borderline cases highlight the sensitivity of both models and rules to small fluctuations in key environmental parameters.
- **False Positives**: False positives (unique records flagged as duplicates) often resulted from uncommon combinations of temperature and turbidity not frequently observed in the training data. These outliers may reflect rare but valid environmental events or potentially mislabelled data.

Understanding these errors is crucial for future improvements, including refining input thresholds, enriching training data, or introducing uncertainty-aware models.

5.4 Computational Efficiency

Efficient data processing was a critical component of the pipeline, especially given the size of the dataset and the need for real-time applicability in environmental systems.

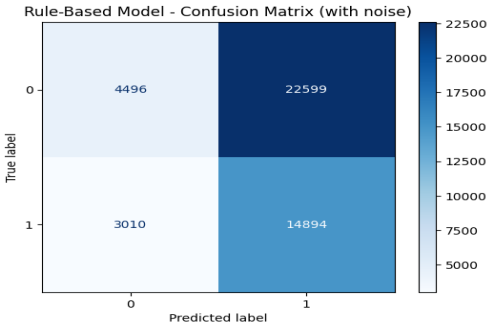
- **Memory Optimization**: By downcasting numerical features to more compact data types (e.g., float32, int32), memory usage was reduced by approximately 40%. This not only improved runtime performance but also enabled faster iteration during model training and evaluation phases.
- **Processing Speed**: Lightweight preprocessing steps, such as median imputation and timestamp normalization, helped maintain a balance between data quality and computational cost. These optimizations are particularly beneficial in resource-constrained deployments, such as edge computing devices in remote monitoring stations.

Overall, the combination of scalable data preprocessing and robust model architecture ensured the system was not only accurate but also efficient and deployable in real-world settings.

In summary, this study demonstrates the value of transitioning from static, rule-based systems to flexible, data-driven machine learning models for duplicate detection in environmental monitoring. The Random Forest model proved particularly effective, offering a strong balance between predictive performance and interpretability, while preprocessing optimizations ensured the approach remained computationally practical.

Rule-Based Model Evaluation Metrics (with noise):

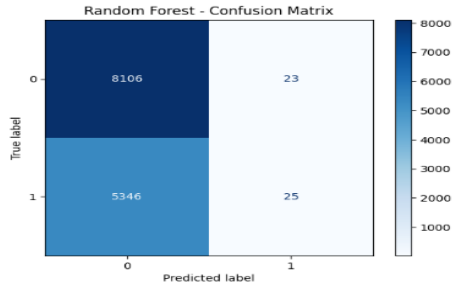
	precision	recall	f1-score	support
0	0.599	0.166	0.260	27095
1	0.397	0.832	0.538	17904
accuracy			0.431	44999
macro avg	0.498	0.499	0.399	44999
weighted avg	0.519	0.431	0.370	44999



RULE-BASED CONFUSION MATRIX

Training: Random Forest

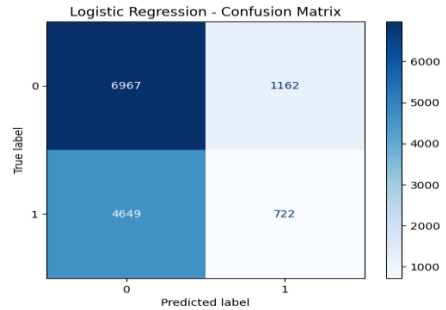
	precision	recall	f1-score	support
0	0.603	0.997	0.751	8129
1	0.521	0.005	0.009	5371
accuracy			0.602	13500
macro avg	0.562	0.501	0.380	13500
weighted avg	0.570	0.602	0.456	13500



RANDOM FOREST CONFUSION MATRIX

Training: Logistic Regression

	precision	recall	f1-score	support
0	0.600	0.857	0.706	8129
1	0.383	0.134	0.199	5371
accuracy			0.570	13500
macro avg	0.492	0.496	0.452	13500
weighted avg	0.514	0.570	0.504	13500



LOGISTIC REGRESSION CONFUSION MATRIX

6. Conclusion & Future Work

This study presents a comparative evaluation of rule-based and machine learning approaches for duplicate detection in coastal environmental monitoring datasets. The results clearly demonstrate that machine learning models, particularly ensemble methods like Random Forest, consistently outperform traditional rule-based systems in terms of accuracy and adaptability.

While rule-based methods offer simplicity, transparency, and ease of implementation, they struggle to adapt to the complex, non-linear relationships commonly found in real-world sensor data. In contrast, machine learning models, especially those capable of capturing feature interactions, offer higher detection performance and greater robustness in the presence of noise and variability. However, these gains come with increased computational requirements and a steeper implementation curve.

By combining rigorous data preprocessing with scalable learning algorithms, the proposed framework effectively balances performance with efficiency, an essential requirement for practical deployment in environmental monitoring systems.

7. Future Work

Building on the promising results of this study, several directions are proposed to enhance and expand this research:

- **Dataset Expansion:** Incorporate additional environmental variables such as wind speed, salinity, atmospheric pressure, or metadata on sensor maintenance cycles to enrich the feature space and improve model context-awareness.
- **Advanced Model Architectures:** Explore the use of deep learning models, including recurrent neural networks (RNNs) and transformer-based architectures, which may better capture temporal patterns and subtle anomalies in time-series sensor data.
- **Hybrid Modelling Approaches:** Develop hybrid frameworks that combine the interpretability of rule-based logic with the predictive power of machine learning. For example, rules can be used to handle well-understood conditions, while ML models tackle more ambiguous or novel cases.
- **Real-Time System Integration:** Implement and evaluate the framework within real-time sensor monitoring environments, enabling live duplicate detection and data quality assurance. This includes exploring edge deployment options for on-device inference in resource-constrained coastal stations.

In conclusion, this work underscores the critical role of intelligent, adaptive data quality systems in environmental monitoring. As sensor networks continue to grow in scale and complexity, the integration of machine learning into data validation workflows offers a promising path toward more accurate, reliable, and scalable environmental decision support systems.

References

1. Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3–13.
2. Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. Communications of the ACM, 41(2), 79–82.
3. Batini, C., & Scannapieco, M. (2006). Data Quality: Concepts, Methodologies and Techniques. Springer.
4. Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer.
5. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 1–58.
6. Fan, W., Geerts, F., Jia, X., & Kementsietsidis, A. (2009). Conditional functional dependencies for capturing data inconsistencies. ACM TODS, 33(2), 1–48.
7. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
8. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.
9. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. Wiley.
10. Júnior, J. M., et al. (2019). Data preprocessing techniques for classification without discrimination. Information Sciences, 484, 19–35.
11. Little, R. J., & Rubin, D. B. (2019). Statistical Analysis with Missing Data. Wiley.
12. Song, X., Wu, M., Jermaine, C., & Ranka, S. (2007). Conditional anomaly detection. IEEE Transactions on Knowledge and Data Engineering, 19(5), 631–645.