

Integrating Data Smell Identification And Mitigation Techniques: A Methodical Approach To Data Quality

Snigdha Ramesh¹, Shilpa Mahalle², Neha Chopade³

¹⁻³ Masters of Computer Application, SIES College of Management Studies, Navi-Mumbai, India

Abstract - Artificial intelligence (AI) system reliability mainly depends on the input data quality. However, subtle issues known as data smells can significantly degrade model performance and reliability. However small issues like data smells can deteriorate the performance and reliability of the model. This study inspects the impact of six categories of data smells - Believability, Consistency, Distribution, Redundancy, Understandability, and Miscellaneous - and evaluate their impact on five models: Logistic Regression, Decision Trees, Random Forests, XGBoost, and a Multilayer Perceptron (MLP), using the UCI Adult Income dataset. Distribution and Redundancy smells proved most harmful, reducing F1-scores by up to 75% for Logistic Regression and significantly affecting MLP, while ensemble methods showed higher strength. Consistency and other smells had smaller effects on accuracy but caused concerns for interpretability, ethics, and maintainability. This study introduces a Data Smell Severity Index to rank smell types with respect to impact, detection difficulty, occurrence, outline mitigation strategies and future directions like real-time validation and automated detection. Overall, this work provides an empirical foundation for addressing data smells and strengthening the credibility of AI systems.

Keywords: Data Smells, Logistic Regression, Decision Trees, Random Forests, XGBoost, Multi-Layer Perceptron.

I. INTRODUCTION

The data quality on which Artificial Intelligence (AI) and Machine Learning (ML) systems are trained on determines their dependability and performance in today's data-driven world. Conventional data validation methods successfully detect possible problems like null entries, missing values, and type inconsistencies. However, they frequently miss more subtle inconsistencies known as "data smells patterns" in the respective data, which may overlook errors resulting in more serious issues like unclear values, inconsistent formats, or implausible entries.

Data smells point out possible challenges quite early with data quality that can damage AI performance, introduce technical debt, or result in biased and unreliable predictions, similar to code smells in software engineering. There is currently no unified framework to rank the severity of these smells, evaluate their impact, or offer systematic detection and mitigation strategies, despite recent research classifying them into groups such as believability, consistency, and understandability.

This study suggests a methodical way to recognise, contrast, and deal with data smells. It quantifies the effects of different data smell categories on AI models and examines their underlying causes and effects. This research provides practical insights and a thorough methodology for enhancing data quality assurance in AI pipelines through empirical evaluation on real-world datasets and machine learning models.

II. BACKGROUND & RELATED WORK

The impact of data quality and technical debt on AI systems is examined at the beginning of this section in II-A, with a focus on the subtle but enduring effects of data smells on long-term performance. A review of the fundamental theories and empirical research on data smells, including their categorisation, detection techniques, and present limitations, is given in Sections II-B and II-C. By reviewing current methods for quality assessment in large-scale datasets and data validation in production settings, Section II-D expands on this conversation. Section II-E highlights important research gaps in current studies on data smells. To address these limitations, this work proposes a clear and comparative framework that allows for systematic analysis, prioritization, and reduction of data smells in machine learning datasets.

A. Data Quality and Technical Debt in AI Systems

The quality of the data used for training has a direct impact on how reliable AI systems are. Conventional validation techniques are good at spotting obvious problems like missing or incorrect entries, but more subtle problems with data quality—known as "data smells"—often go unnoticed. Over time, these latent anomalies impair model performance, scalability, and fairness even though they may follow expected formats. These hidden issues can lead to technical debt in machine learning (ML), where unresolved data errors build up and impair predictive accuracy, as mentioned by Sculley et al. [1]. The fact that data preparation takes up a large amount of the machine learning lifecycle rather than model development is further highlighted by recent research in data-centric AI. According to Whang [7], real-world datasets are often noisy, biased, inconsistent, or incomplete, which poses systemic risks to machine learning pipelines if improperly handled. The importance of spotting hidden problems with data quality early in the data pipeline is highlighted by this move toward data-centric AI. Similar to this, Fischer et al. [8] address the difficulties in AI system engineering and stress the importance of dataset quality assurance, visualization-based sanity checks, and data validation in preventing downstream errors in ML systems. Their research demonstrates how undetected data issues can spread throughout the whole machine learning lifecycle and ultimately show up as technical debt. Golendukhina et al. [11], who looked into the occurrence and causes of data smells in a business travel data processing pipeline, offer more empirical evidence about data smells in real-world data processing environments. The significance of keeping an eye on the full data lifecycle was highlighted by their findings, which showed that data smells can come from both raw datasets and intermediate transformation and enrichment stages within the data pipeline. Furthermore, datasets gathered from various sources may have intrinsic distributional differences that impact model generalization, according to earlier research on dataset bias. Such dataset biases can have a substantial impact on machine learning model performance when systems are implemented in novel environments, as Torralba and Efros [14] showed.

B. Data Smell Taxonomy and Empirical Exploration

The conceptual underpinnings of the study of data smells were established by Foidl et al. [2]. They offered a structured definition and divided data smells into three categories—believability, understandability, and

consistency—to fill a vacuum in the literature on data quality. By presenting data smells as context-independent markers of bad data practices, their method—which was founded on a thorough literature review and conceptual modeling—made a substantial contribution even though it lacked empirical validation.

Recupito et al. [3] carried out a comprehensive empirical analysis spanning 19 AI datasets, building on this work. Using rule-based approaches, machine learning techniques, and schema validation tools, they added three more categories—redundant value, distribution, and miscellaneous smells— and evaluated how common these smells were. Although their study provided new empirical insights, it did not assess how data smells directly affect the performance of AI models, which leaves a significant gap regarding downstream effects.

Furthermore, common data problems like duplication, class imbalance, and inadequate data coverage can seriously impair machine learning performance, according to studies on dataset quality evaluation. To systematically assess dataset reliability, Chen [9] suggests five data quality dimensions: comprehensiveness, correctness, variety, class imbalance, and duplication. Although these dimensions deal with quantifiable issues related to data quality, they do not specifically address the latent anomalies that data smells describe. Furthermore, for machine learning-based software systems, Foidl and Felderer [12] proposed a risk-based data validation approach that ranks validation tasks according to the likelihood and significance of low-quality features. Their approach considers data source reliability, feature importance, and potential data smells when determining validation strategies.

Systematic errors may arise during the stages of data collection, preprocessing, and model evaluation, according to research on bias in machine learning pipelines. Mikołajczyk and Grochowski [15] provide a taxonomy of more than forty potential bias sources within ML pipelines, emphasizing the importance of early detection and mitigation of data-related anomalies.

C. Schema-Based Validation in Production Systems

Breck et al. (2019) [4] developed a schema-driven validation framework that was integrated into Google's TFX system to address data quality issues in production ML pipelines. Their solution, which was implemented in more than 700 production pipelines and validated trillions of data records, concentrated on identifying schema mismatches, missing features, and data drift. This framework decreased debugging time and greatly increased model reliability. But rather than latent or semantic problems like data smells, it focused on structural integrity and overt anomalies. Further research on ML data validation highlights how difficult it is becoming to validate datasets in contemporary AI systems. The difficulties of putting strong data validation procedures in place in large-scale pipelines are covered by Agrawal [6], who also emphasizes the importance of data drift detection, statistical validation, and schema validation in spotting abnormalities prior to model training. Nevertheless, these methods mainly concentrate on explicit structural errors rather than subtle semantic inconsistencies, just like other validation

frameworks. Lifecycle management tools that support data and model monitoring throughout the machine learning development process have also been investigated in recent work. MLflow, a platform created to oversee machine learning experiments, monitor data dependencies, and guarantee reproducibility across ML pipelines, was first presented by Zaharia et al. [16]. Throughout the ML lifecycle, these systems help to improve data quality monitoring.

D. Quality Scoring and Metric Weighting in Big Data

To capture the relative importance of different quality dimensions, Elouataoui et al. [5] proposed an advanced big data quality assessment framework that uses weighted metrics. A real-world Twitter dataset was used to assess the framework's additional metrics, which included integrity, accessibility, security, and ease of manipulation. The study did not address latent issues like data smells or their potential impact on the development of AI models, even though the weighted scoring methodology improved the accuracy of quality measurement.

Similar to this, studies on deep learning datasets have discovered several persistent quality issues, such as adversarial data contamination, noisy features, and missing values. Manju [10] emphasizes that if these problems are not appropriately addressed, they can have a substantial impact on training stability and predictive reliability. These studies do not systematically address the identification and prioritization of data smells, even though they offer insightful information about dataset challenges. In agreeing with these large-scale surveys on fairness and bias mitigation, improving dataset quality before training is a major frameability point in development of models. Hort et al. On bias mitigation methods: [17] reviewed over 300 papers on the topic of different bias mitigation techniques and classified solutions in three categories: pre-process, in-process and post-process solutions to train fair machine learning systems.

In related studies, Siddique et al. [18] analyze bias mitigation techniques and demonstrate that ML models can be made more reliable and fairer by group-based data preprocessing methods like rebalancing, noise filtering and feature transformation.

E. Positioning of This Work

Unlike previous attempts, this study presents a cohesive and methodical framework that incorporates the following elements:

- (1) A comparative analysis of current data smell categories; and
- (2) An empirical evaluation that quantifies the effect of different smells on model performance by injecting them into benchmark dataset. This work fills important research gaps in smell prioritization and AI performance quantification by combining conceptual taxonomy and empirical validation, offering a scalable and useful solution for contemporary data pipelines.

III. PROBLEM DEFINITION

Large-scale datasets are crucial to the training and decision-making processes of contemporary AI systems. Conventional data validation techniques are good at detecting explicit problems like null entries, missing

values, and type mismatches, but they frequently miss latent, context-independent anomalies, also known as data smells. By adding technical debt, these minor problems—while not necessarily incorrect right away—can subtly impair model performance, equity, and long-term maintainability.

Even though different data smells have been defined and categorized by earlier research, there is currently no unified framework to compare smell types systematically, evaluate their practical impact, or prioritize mitigation strategies. Furthermore, many studies do not empirically assess how particular data smells affect important AI performance indicators like accuracy, fairness, or robustness; instead, they are either theoretical or restricted to surface-level quality metrics.

The creation of reliable, data-centric AI systems is hampered by the absence of a severity-aware and empirically validated methodology. A thorough, scalable framework that permits the methodical identification, comparative assessment, and prioritization of data smells based on their practical implications is therefore desperately needed.

IV. OBJECTIVE & SCOPE

A. Objective

Developing a cohesive and methodical framework for detection, categorization, and mitigation of data smells in AI-based systems is the main goal of this research. The study specifically seeks to address shortcomings in the body of current literature by:

- Creating a framework for comparative assessment of data smell categories according to their prevalence, severity, and underlying causes;
- Carrying out empirical tests to assess how various types of data smells impact machine learning models' capacity for generalization and prediction accuracy;
- Finding workable, category-specific mitigation techniques that are scalable and efficient;
- Closing the gap between theoretical definitions and real-world data smell management.

B. Scope

Six main categories of data smells are examined in this study, which focuses on the analysis of structured datasets: distribution, redundancy, consistency, understandability, believability, and miscellaneous smells. The scope of the study comprises:

- simulated the introduction of data smells into publicly accessible datasets from multiple domains;
- Neural networks and traditional machine learning algorithms for training and assessment;
- A comparison of the clean and smell-injected datasets' performance.

The suggested framework is intended for machine learning workflows that are batch-based. Unstructured data formats (such as text, audio, and images) and streaming or real-time data pipelines are specifically left out of the scope, as they are noted as areas that require further investigation.

V. METHODOLOGY

A. Overview

To assess how six different types of data smells affect the functionality of different machine learning (ML) and neural network (NN) models, this study uses a controlled experimental design. The goal is to measure the impact of each data smell category on important classification metrics, such as F1-score, recall, accuracy, and precision.

Using a constant injection volume for all categories, each data smell category is injected separately into a clean dataset. This method guarantees a consistent and equitable comparison, allowing for an accurate evaluation of the relative impact of each type of smells on model performance.

B. The dataset

The Adult Income Dataset from the UCI Machine Learning Repository is used in the experiments. There are roughly 45,222 records and 14 features in this dataset, which includes both numerical and categorical variables. The binary target variable shows whether a person makes more than \$50,000 a year. The dataset is ideal for the controlled injection and assessment of different data smell categories due to its structured format, moderate size, and variety of feature types. Reproducibility and comparability of results are further supported by its extensive use in machine learning research.

C. Smell Injection Strategy

For each data smell category, precisely 2,400 rows, or 5.3% of the entire dataset, were altered to guarantee an impartial and consistent evaluation. These rows were evenly distributed among the subtypes within each category and were chosen at random using a fixed random seed to ensure reproducibility. Only one injected smell category was present in each dataset version; all other records were the same as those in the clean baseline. The controlled isolation of each smell's unique effect on model performance was made possible by this design.

1. Believability Smells

Values that are syntactically correct but semantically implausible or logically inconsistent were inserted to create believability smells. This is how the injection was carried out:

- **Dummy Values:** For 480 records, the occupation attribute was replaced with the placeholder value "XXX" to represent missing or contextually meaningless entries.
- **Implausible Numerical Values:** For the next 480 records, the age attribute was set to an unrealistic value of 999, which represents extreme or logically incorrect entries that might pass validation checks but undermine the plausibility of the model.
- **Suspicious Precision:** To simulate artificially precise data that deviates from standard formatting or reporting standards, a high-precision floating-point value of 12345.6789 was assigned to the capital-gain field for an additional 480 records.
- **Unusual Distribution Patterns:** The hours-per-week attribute was set to a constant minimum value of 1 for 480 entries to simulate distributional anomalies. This resulted in 480 entries with unusual working hours, which could skew subsequent statistical assumptions.
- **Multiple Value Contamination:** In 480 records, the workclass field was given the combined categorical value "Private, Self-Employed," creating

ambiguity by combining several different categories into one entry.

- A total of 2,400 records (480 per subtype) were impacted by the sequential and isolated injection of each of these subtypes into the `df_believability` dataset. This methodical approach guarantees that all injected values stay contextually anomalous but type-compatible, allowing for a precise evaluation of their effect on model performance.

2. Consistency Smells

- Without changing the semantic meaning of equivalent values, consistency smells were introduced by adding variation to their representation. Using these techniques, 2,400 records in all (400 per subtype) were changed:
 - Syntax Inconsistency: To simulate inconsistent capitalization, the gender field was filled with three casing variants: "Male" (rows 0–100), "MALE" (rows 200–400), and "male" (rows 500–600).
 - Special-Character Inconsistency: For rows 100–500, "Tech#Support" was substituted for the occupation attribute, adding non-alphanumeric characters that violate uniformity.
 - Spacing Inconsistency: In rows 600–1000, values with leading and trailing spaces ("Husband") were entered into the relationship column, simulating formatting errors.
 - Separator Inconsistency: For rows 1000–1400, the workclass field was assigned "Private|Self-Employed" using a pipe character rather than the usual delimiters.
 - Separator Inconsistency: Instead of using standard delimiters, a pipe character was used to assign "Private|Self-Employed" to the workclass field for rows 1000–1400.
 - Abbreviation Inconsistency: To reflect the mixed use of abbreviations and full names, the native-country attribute was set to two synonymous forms, "US" (rows 1200–1400) and "United States" (rows 1400–1600).
 - Value- Length Inconsistency: To add variation in string length, two semantically identical but length- differing values were added to the occupation column: "Executive" (rows 1600–1800) and "Executive-Level" (rows 1800–2000).
- In the clean dataset's `df_consistency` copy, each inconsistency subtype was applied separately. This method allows a controlled investigation of how representation inconsistencies impact downstream model performance while guaranteeing the validity of type constraints.

3. Understandability Smells

- About 2,400 records (about 400 per subtype) were affected by Understandability Smells, which were injected to introduce comprehension and formatting challenges without changing data types. To isolate their semantic impact, all changes were made to a copy of the clean dataset (`df_understandability`) while preserving type compatibility:
 - Numerical Fields as Strings: To replicate situations in which numeric data is stored as text, the educational-num and capital-gain attributes were globally cast from numeric types to strings.
 - Special-Character Insertion: The workclass field was set to "Private!" for rows 0–400, adding unnecessary punctuation that could be confusing to users or parsers.
 - Case Inconsistency: "MARRIED" was assigned to rows 200–400 and "married" to rows 300–500 in the marital-status column, resulting in overlapping mixed-case entries that compromise uniformity.

- Unusually Verbose Strings for the Occupation Attribute Rows 600–800 ("Executive Technical Supervisor with 10+ Years Experience") and 200–400 ("Senior Executive Associate to CEO with 12+ Years of Experience including 2 awards") mimicked overly detailed entries that make them difficult to read.
- Intermingled Data Types: The capital-loss field was cast to object type and filled in with the value "200.0%" for rows 800–1000 and 200–400, combining string representation and percentage symbols with numerical content.

To guarantee that exactly one category of understandability smell appeared per dataset version, each subtype injection was done in turn. The precise measurement of the impact of such formatting and representation anomalies on the interpretability and performance of downstream models is made possible by this controlled design.

4. Distribution Smells

To skew the dataset's natural statistical distribution of features, extreme outlier values were injected to create distribution smells. In particular, an unusually high value of 1,000,000,000 was assigned to the capital-gain attribute for 2,400 records (roughly 5.3% of the dataset). By distorting feature distributions, this extreme value injection mimics uncommon but significant anomalies that can negatively impact model training stability and predictive accuracy.

5. Redundant Value Smells

Redundant Value Smells were created by incorporating features that increase the complexity and redundancy of data without offering any new predictive information. The dataset was supplemented with two primary forms of redundancy:

- Correlated Feature Duplication: Perfect correlation and redundancy were introduced by creating a new column called `education_copy`, which is an exact duplicate of the current `educational-num` feature.
- Unique Identifiers: Sequential integers ranging from 1 to the dataset size were added to a synthetic `unique_id` column. Despite being unique, this identifier may unintentionally bias the model or result in leakage rather than providing useful information for prediction.
- These additions allow for the assessment of the model's sensitivity to redundant and possibly deceptive attributes while expanding feature space without improving data quality.

6. Miscellaneous Smells

Synthetic sensitive features that serve as stand-ins for potential bias or data leakage in the dataset were added to introduce miscellaneous smells. In particular, the dataset was expanded to include a new column called `religion`, which was filled with the constant value "Unknown" in its entirety. This feature simulates situations where sensitive or proxy attributes exist in the data, potentially affecting model fairness or ethical considerations even though it lacks direct predictive information.

Only one kind of data smell was added to each modified dataset, which was then stored independently and used in conjunction with the clean version for training and assessment. The model's behaviour under each smell

condition could be directly and impartially compared thanks to this setup.

D. Model Selection

Five representative models covering linear, tree-based, ensemble, and deep learning approaches were chosen in order to thoroughly evaluate the effect of data smells on machine learning model performance:

Models Assessed for the Impact of Data Smell:

- Logistic Regression (Linear): Sensitive to noise and feature quality; interpretable; offers baseline performance.
- Tree-based Decision Tree: Interpretable; helpful for identifying irregularities in formatting and inconsistencies.
- Random Forest (Ensemble-Bagging): Resistant to redundant features and noise.
- XGBoost (Ensemble-Boosting): Extremely sensitive to outliers and feature distribution.
- Multi-Layer Perceptron (Neural Network): Affected by value distortions and data format, it captures intricate non-linear relationships.

A clean baseline dataset and six different datasets, each of which contained injected scents that matched one of the specified categories, were used to train and assess each model. A controlled comparison of model sensitivities to various data quality issues is made possible by this experimental setup.

E. Evaluation Metrics

Standard classification metrics are used to assess each model's performance:

- Precision
- The F1-Score

The relative decline in model performance brought on by each category of smell is measured by comparing the results from datasets with injected data smells to the clean baseline. Future work includes creating a Data Smell Severity Index (DSSI), even though this study does not formally suggest a single severity metric. The goal of this index is to incorporate both quantitative and qualitative elements, such as the degree of performance degradation, the intricacy of smell detection, and the influence of behavior on AI models.

F. Environment and Tools

Every experiment was carried out on a local computer using Python in a Jupyter Notebook environment. The following libraries and tools were used to implement the data smell injection, preprocessing, model training, and performance evaluation:

- scikit-learn: This tool is used to create and assess machine learning models, such as Multi-Layer Perceptron (MLP), Decision Tree, Random Forest, and Logistic Regression. Metric computation and data preprocessing were also done with it.
- XGBoost: Used to train and assess a gradient boosting classifier, chosen for its robust performance on structured data and sensitivity to data irregularities.
- Seaborn, Matplotlib, and Pandas: Used for statistical plotting, data manipulation, and visualization.
- Ataccama ONE Desktop IDE: Used both before and after data smell injection for data profiling.

Every experiment was conducted using a standard Python environment within Jupyter Notebook on a local system with a minimum of 8 GB of RAM.

VI. ANALYSIS AND FINDINGS

This section provides a thorough analysis of the effects of each data smell category on the accuracy and F1 score of machine learning models. Five models are compared: Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), XGBoost, and Multi-Layer Perceptron (MLP), which was trained on six datasets, each injected with a distinct smell category, as well as a clean dataset.

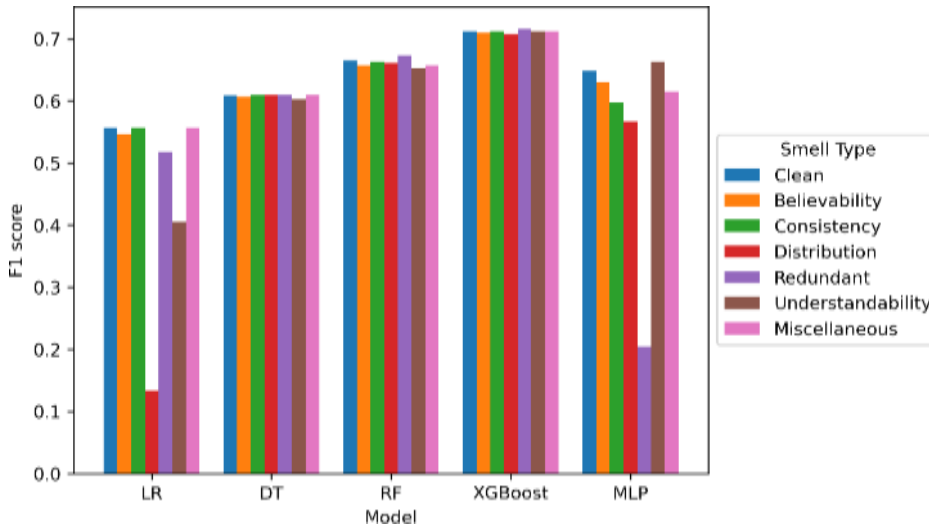
A. Model Behavior on Clean Data

The clean dataset was used to establish baseline performance. XGBoost produced the highest F1- score (0.713) and accuracy (0.872) out of the five models that were evaluated, closely followed by Random Forest and Multi-Layer Perceptron (MLP). Simpler classifiers, like Decision Tree and Logistic Regression, on the other hand, showed relatively lower F1 scores and accuracy, which is consistent with their limited ability to capture intricate relationships in the data.

B. Data Smell Categories' Effect on F1-Score

The main metric for evaluating the degree of smell-induced performance decline was the F1 Score, which is the harmonic mean of precision and recall. The analysis showed:

Fig. 1. F1-score for various smells



- Distribution Smells caused the worst degradation, especially in MLP (−12.6%) and Logistic Regression (−75.8%), demonstrating how sensitive they are to extreme outliers.
- Redundant smells, which involve redundant or irrelevant features, resulted in moderate drops in Logistic Regression and significant degradation in MLP (−68.6%).

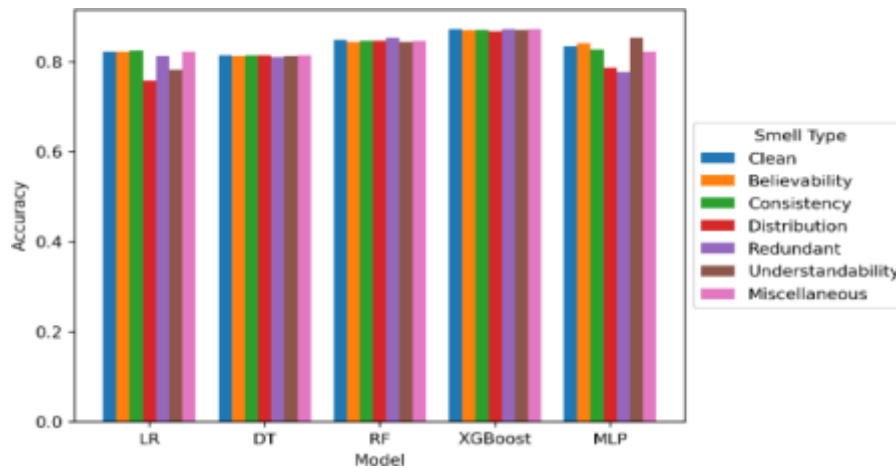
- Understandability Smells had a minor impact on Decision Tree and Logistic Regression (-27.1%), but MLP's performance slightly improved.
- Moderate F1 reductions in MLP (-2.9%) and Logistic Regression (-1.6%) were caused by Believability Smells.
- For the majority of models, Consistency and Miscellaneous Smells had little effect on F1 scores, with Decision Tree and XGBoost exhibiting the least amount of variation ($<1\%$).

These patterns suggest that models that are more susceptible to data quality problems caused by particular smell types are those with simpler architectures or those that are sensitive to numerical or structural patterns (e.g., MLP, Logistic Regression).

C. Accuracy-Based Observations

The overall patterns supported the same conclusions even though Accuracy stayed comparatively constant when compared to F1 Score:

Fig. 2. Accuracy for various smells



- Distribution Smells (-7.8% for Logistic Regression) and Redundant Smells (-6.85% for MLP) had the biggest accuracy declines.
- With variations usually less than 1% , XGBoost and Random Forest maintained high accuracy under all smell conditions.
- MLP and Logistic Regression, on the other hand, demonstrated notable performance volatility, suggesting that they are more sensitive to erratic or redundant data patterns.

D. Model Robustness Analysis

The outcomes unequivocally show how ensemble models like Random Forest and XGBoost are more resilient when different data smells are present. These models demonstrated resilience to both semantic and structural anomalies, continuously maintaining high performance. However, due to their reliance on well-conditioned input data, MLP and Logistic Regression showed significant performance degradation, particularly when statistical outliers and redundant information were present.

E. Visual Perspectives

Clear visual evidence of the relative impact of the data smells was provided by

grouped bar graphs that compared baseline performance against each smell-injected dataset. The bar graphs showed the biggest performance declines, especially under Distribution and Redundant smells, by plotting Accuracy and F1 Score side by side for each model and smell category. These plots supported the quantitative results and allowed for easy model-by-model comparisons. They demonstrated that ensemble models like Random Forest and XGBoost maintained more stable performance, while simpler models like Logistic Regression and MLP were particularly susceptible to these smells.

As a result, the visualizations enhance the numerical analysis by instantly highlighting the relative severity of each data smell.

F. Unified Framework: Comparison and Mitigation of Data Smell

A unified framework that methodically describes the six main data smell categories looked at in this study is presented in this section. We provide an overview of each category's prevalence, underlying causes, possible outcomes, detection challenges, and workable mitigation techniques. In order to improve the overall integrity and dependability of machine learning systems, practitioners and researchers can swiftly identify latent data quality issues and apply targeted remediation techniques thanks to this framework's structured organization of these smells.

1. Believability Smells

- Moderate prevalence
- Causes include mixed or multiple values, implausible numeric entries, and dummy placeholders.
- Repercussions include decreased data trust, poor generalization, and the introduction of semantic ambiguity.
- Difficulty of Detection: Medium
- Mitigation techniques include applying anomaly detection, enforcing domain constraints, and implementing range checks.

2. Consistency Smells

- High prevalence
- Reasons include unit mismatches, special character variations, spacing mistakes, and inconsistent casing.
- Repercussions include decreased parsing accuracy, more preprocessing, and duplicate or broken categories.
- Low Difficulty of Detection
- Mitigation Techniques: Standardize units and formats, employ regex-based standardization, and normalize strings.

3. Distribution Smells

- Low to Moderate Prevalence
- Causes include skewed distributions, artificial extreme values, and extreme outliers.
- Repercussions include severe model bias, poor convergence, interrupted training, and incorrect classifications.
- Medium to High Difficulty of Detection
- Mitigation Strategies: Scale or normalize data, apply IQR/z-score filtering.

4. *Redundant Smells*

- Moderate prevalence
- Reasons: Leaked identifiers, highly correlated features, and duplicate columns
- Repercussions include multicollinearity, overfitting, and more complex models.
- Difficulty of Detection: Medium
- Mitigation techniques include PCA, correlation analysis, feature pruning, and hiding unnecessary identifiers.

5. *Understandability Smells*

- High prevalence
- Reasons: Mixed data types in columns, verbose or ambiguous labels, and inconsistent encoding
- Repercussions: inconsistent input representation, misinterpretation, and parsing errors
- Difficulty of Detection: Low to Medium
- Mitigation techniques include enforcing length and format validation, cleaning labels, and coercing types.

6. *Miscellaneous Smells*

- Low prevalence
- Reasons: Sensitive or proxy characteristics (such as ID fields, race, or religion)
- Repercussions: Prejudice, moral and legal dilemmas, possible injustices • Difficulty of Detection: High Mitigation techniques include fairness audits, ethical data governance procedures, and the removal or masking of sensitive features.

G. *Synopsis of Important Findings*

The following important findings came from this empirical assessment:

- The biggest detrimental effects on model performance are caused by distribution and redundant value smells.
- Across all smell categories, ensemble models (such as Random Forest and XGBoost) show remarkable resilience.
- Under the majority of smell types, MLP and Logistic Regression show significant performance degradation due to increased sensitivity to data anomalies.
- When it comes to identifying performance, decline brought on by subtle problems with data quality, F1 score is a more sensitive metric than accuracy.
- Because of implicit regularization brought on by data distortions, some smell types may have counterintuitive effects, such as minor performance gains.

VII. LIMITATIONS AND FUTURE SCOPE

A. *Limitations*

Although this study provides a systematic and comparative analysis of the effects of six different data smell categories on machine learning performance, there are still a number of limitations:

- Synthetic Smell Injection Only: Using predetermined rules, the data smells

were manually added to a clean dataset. Although this guaranteed experimental control, it might not accurately capture the intricacy and subtlety of naturally occurring data quality problems present in real-world datasets.

- **Single Dataset Limitation:** Only the UCI Adult Income dataset was used for the experiments. Despite being structured and widely used, the findings' applicability to other domains (such as time series, unstructured text, and healthcare) has not yet been examined.
- **No Real-Time or Streaming Evaluation:** The study relies on batch training models and static tabular data. It does not evaluate the impact of data smells on models in continuous learning environments or real-time streaming pipelines.

Model-Dependent Interpretability: Each model has a different level of performance degradation. Therefore, using different models, training regimens, or hyperparameters may lead to different conclusions regarding the relative danger of each smell category.

B. Future Scope:

Based on the results of this study, a number of interesting avenues can be investigated:

- **Creation of a Data Smell Severity Index (DSSI):** Developing a single metric to rank smell categories according to their composite impact is a future goal. This DSSI would provide a weighted severity score for useful prioritization, taking into account not only performance degradation but also detection complexity and interpretability effects.
- **Extension to Diverse Datasets and Domains:** To confirm the generalizability of smell impacts across various domains, future research should expand this comparative analysis to other datasets, such as high-dimensional, imbalanced, or unstructured data (e.g., logs, sensor data, medical records).
- **Integration with Real-World Pipelines and Validation Tools:** This study's conclusions can help design monitoring systems and validation frameworks in production pipelines, particularly for machine learning systems that need to be highly reliable in the face of noisy data.
- **Automated Smell Detection and Cleaning Frameworks:** Data-centric AI development would greatly benefit from the development of smell-aware preprocessing pipelines that can identify, flag, and optionally resolve smell-induced anomalies.
- **Model-Specific Smell Tolerance Studies:** A more thorough examination of model-specific resilience may aid developers in selecting or creating architectures that are most appropriate for their data quality profile.

VIII. CONCLUSION

The impact of six different data smell categories—Believability, Consistency, Distribution, Redundant, Understandability, and Miscellaneous—on machine learning model performance is systematically and comparatively examined in this study. We quantitatively assessed each smell type's effect on five distinct classifiers by isolating it and injecting it into a clean, structured dataset. We then measured changes in F1 Score and Accuracy.

Our results show that the most significant negative effects on model performance are caused by Distribution and Redundant smells, particularly

for algorithms like Logistic Regression and Multi- Layer Perceptrons that are highly sensitive to numerical and feature-level distortions. On the other hand, ensemble models such as Random Forest and XGBoost showed significant resilience to changes in data quality across all types of data smells.

Lastly, this work suggests a useful comparative framework that characterises data smells according to their detection complexity, performance impact, and prevalence, along with specific mitigation techniques for each category. For practitioners looking to improve data preprocessing and validation within AI pipelines, these insights provide practical advice. Overall, by clarifying the frequently disregarded but important role that data smells play in influencing predictive performance, this research advances the developing field of data-centric AI. In order to support data quality management in production settings, future research will investigate automated smell detection and correction techniques as well as a formal severity index.

REFERENCES

1. David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean- Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015).
2. H. Foidl, M. Felderer and R. Ramler, "Data Smells: Categories, Causes and Consequences, and Detection of Suspicious Data in AI-based Systems," 2022 IEEE/ACM 1st International Conference on AI Engineering – Software Engineering for AI (CAIN), Pittsburgh, PA, USA, 2022, pp. 229-239, doi: 10.1145/3522664.3528590
3. G. Recupito, R. Rapacciuolo, D. Di Nucci and F. Palomba, "Unmasking Data Secrets: An Empirical Investigation into Data Smells and Their Impact on Data Quality," 2024 IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN), Lisbon, Portugal, 2024, pp. 53-63.
4. Polyzotis, N., Zinkevich, M., Roy, S., Breck, E. and Whang, S., 2019. Data validation for machine learning. *Proceedings of machine learning and systems*, 1, pp.334-347.
5. Elouataoui, W., El Alaoui, I., El Mendili, S. and Gahi, Y., 2022. An advanced big data quality framework based on weighted metrics. *Big Data and Cognitive Computing*, 6(4), p.153.
6. V. Agrawal, "Data Validation : A Complex Challenge in Modern AI Systems," *International Scientific Journal of Engineering and Management*, vol. 04, no. 01, pp. 1–9, Jan. 2025, doi: 10.55041/isjem01311.
7. S. E. Whang, *Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI*

9. Perspective, VLDB Journal, 2023.
10. L. Fischer et al., AI System Engineering—Key Challenges and Lessons Learned, 2020.
11. H. Chen, Data Quality Evaluation and Improvement for Machine Learning, 2022.
12. T. Manju, Quality Challenges in Deep Learning Data Collection in Perspective of Artificial Intelligence, 2023.
13. V. Golendukhina, H. Foidl, M. Felderer, and R. Ramler, Preliminary findings on the occurrence and causes of data smells in a real-world business travel data processing pipeline, 2022.
14. H. Foidl and M. Felderer, Risk-based data validation in machine learning-based software systems, 2019.
15. D. Costal, C. Gómez, S. del Rey, and S. Martínez-Fernández, Using metrics for code smells of ML pipelines, 2024.
16. A. Torralba and A. A. Efros, Unbiased Look at Dataset Bias, CVPR, 2011.
17. A. Mikołajczyk and M. Grochowski, A Survey on Bias in Machine Learning Research, 2024.
18. M. Zaharia et al., Accelerating the Machine Learning Lifecycle with MLflow, IEEE Data Engineering Bulletin, 2018.
19. M. Hort et al., Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey, ACM Journal on Responsible Computing, 2023.
20. S. Siddique et al., Survey on Machine Learning Biases and Mitigation Techniques, Digital Journal, 2023.
21. T. Rekatsinas et al., HoloClean: Holistic Data Repairs with Probabilistic Inference, VLDB, 2017.