information

MDPI

Article

DQMAF—Data Quality Modeling and Assessment Framework

Razan Al-Toq and Abdulaziz Almaslukh *

Information System Department, King Saud University, Riyadh 12372, Saudi Arabia; raltoq@ksu.edu.sa
* Correspondence: aalmaslukh@ksu.edu.sa

Abstract

In today's digital ecosystem, where millions of users interact with diverse online services and generate vast amounts of textual, transactional, and behavioral data, ensuring the trustworthiness of this information has become a critical challenge. Low-quality data—manifesting as incompleteness, inconsistency, duplication, or noise—not only undermines analytics and machine learning models but also exposes unsuspecting users to unreliable services, compromised authentication mechanisms, and biased decision-making processes. Traditional data quality assessment methods, largely based on manual inspection or rigid rule-based validation, cannot cope with the scale, heterogeneity, and velocity of modern data streams. To address this gap, we propose DQMAF (Data Quality Modeling and Assessment Framework), a generalized machine learning–driven approach that systematically profiles, evaluates, and classifies data quality to protect end-users and enhance the reliability of Internet services. DQMAF introduces an automated profiling mechanism that measures multiple dimensions of data quality—completeness, consistency, accuracy, and structural conformity—and aggregates them into interpretable quality scores. Records are then categorized into high, medium, and low quality, enabling downstream systems to filter or adapt their behavior accordingly. A distinctive strength of DQMAF lies in integrating profiling with supervised machine learning models, producing scalable and reusable quality assessments applicable across domains such as social media, healthcare, IoT, and e-commerce. The framework incorporates modular preprocessing, feature engineering, and classification components using Decision Trees, Random Forest, XGBoost, AdaBoost, and CatBoost to balance performance and interpretability. We validate DQMAF on a publicly available Airbnb dataset, showing its effectiveness in detecting and classifying data issues with high accuracy. The results highlight its scalability and adaptability for real-world big data pipelines, supporting user protection, document and text-based classification, and proactive data governance while improving trust in analytics and AI-driven applications.

Keywords: data quality assessment; machine learning; big data analytics

check for updates

Academic Editors: Ioannis Tsimperidis, Eleni Vrochidou and George A. Papakostas

Received: 31 August 2025
Revised: 12 October 2025
Accepted: 13 October 2025
Published: 17 October 2025

Citation: Al-Toq, R.; Almaslukh, A. DQMAF—Data Quality Modeling and Assessment Framework. *Information* **2025**, *16*, 911. https://doi.org/10.3390/info16100911

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In today's interconnected digital society, where billions of users interact with online platforms, applications, and services, data is often described as the "new oil" [1]. Just as oil fueled the industrial revolution, trustworthy data now powers secure authentication, personalized services, and user protection in the modern Internet [1]. Data has become the backbone of digital ecosystems, supporting not only analytics and decision-making but also the safety and reliability of interactions that shape social media, healthcare, e-commerce, and the Internet of Things (IoT) [2]. Generated from diverse sources—including user-generated text, financial transactions, sensor networks, and scientific exploration—data continues to grow at an unprecedented pace [3,4].

Ensuring the quality of this data is therefore a critical requirement for protecting users, since it directly influences the fairness, effectiveness, and reliability of downstream processes [2,4]. Within Data Management Systems (DMSs), quality assessment impacts not only compliance and organizational governance [5,6] but also the security of authentication workflows, the trustworthiness of user profiling, and the validity of document and text classification systems. Data quality is inherently multidimensional and context-dependent [4,7], encompassing completeness, reliability, consistency, and accuracy. These attributes capture the absence of errors, duplicates, and missing values [5], all of which are crucial for preventing biased predictions, misleading insights, or harmful user experiences.

Beyond these fundamental attributes, additional dimensions become essential in user-facing contexts. Availability ensures that data is accessible exactly when required for tasks such as real-time navigation or clinical decision support [8]. Usability reflects whether information is structured to support seamless integration into authentication or profiling pipelines, enhancing user engagement and efficiency [9]. Reliability emphasizes the trustworthiness of data, which is maintained through validation procedures and checks of provenance; for instance, users depend on accurate market feeds or verified social content [10–12]. Relevance highlights the importance of contextual accuracy—outdated or irrelevant sentiment data, for example, can severely distort analytical insights for social media or e-commerce platforms [9]. Finally, presentation quality refers to the clarity and accessibility of dashboards and recommendations presented to users, which directly contribute to their perception of trust and transparency in digital systems [13].

Despite these needs, the scale and complexity of modern data streams introduce challenges such as incompleteness, inconsistency, and noise, which compromise both organizational analytics and user safety [14]. Poor-quality data may lead to flawed recommendations, authentication errors, financial risks, or misinformation spread [2]. This impact is especially critical in IoT, healthcare, and social platforms, where inaccurate or fabricated inputs can impair patient outcomes, enable identity fraud, or expose users to misleading narratives [15–17].

Traditional methods for quality assurance—including profiling, rule-based validation, sampling, referential integrity checks, and user-driven queries [18]—struggle with the three Vs of big data: volume, velocity, and variety. They remain inadequate for real-time environments where users expect secure, responsive, and trustworthy services. As a result, Machine Learning (ML) has emerged as a scalable and adaptive alternative for automated data quality assessment [19]. ML algorithms can identify hidden patterns, detect anomalies, and classify quality levels without explicit programming [20,21], enabling systems to protect users more effectively. Prior studies have applied ML in domains such as sensor monitoring [6], oceanographic research [4], and healthcare records [20], illustrating its cross-domain applicability.

In this work, we present DQMAF (Data Quality Modeling and Assessment Framework), a generalized, machine learning–driven framework designed to safeguard users and digital services by systematically assessing and classifying data quality. DQMAF evaluates data attributes such as completeness, consistency, and duplication, then employs supervised classifiers to categorize records into high, medium, or low quality. By providing interpretable quality profiles, the framework strengthens user authentication, enhances document and text classification, and ensures more reliable profiling across big data environments. We validate the framework using a publicly available Airbnb dataset, demonstrating its ability to reveal hidden data issues with high accuracy. The approach is scalable, reusable, and adaptable across domains, including healthcare, IoT, and social media, offering a foundation for protecting Internet users while improving the trustworthiness of digital interactions.

The remainder of this paper is organized as follows: Section 2 discusses limitations of existing approaches. Section 3 and 4 define the goals of building a user-centric data quality assessment framework and the research questions, respectively. Section 5 describes the DQMAF architecture and pipeline. Section 6 outlines the dataset, implementation setup, and workflow. Section 7 evaluates model performance; and Section 8 summarizes findings and proposes directions for future research.

2. Literature Review

Ensuring data quality in big data systems has emerged as a critical research focus due to the exponential growth in the volume, variety, and velocity of data generated across multiple domains. Poor-quality data, characterized by missing values, inconsistencies, redundancy, and lack of structure, can severely hinder downstream analytics and impair organizational decision-making [4,7]. Traditional rule-based or manual methods for quality assessment often fail to scale or adapt in such dynamic environments [22]. This has prompted the integration of machine learning (ML) techniques to automate, generalize, and enhance the data quality evaluation process [18,23].

Big data is commonly described using the "four Vs": Volume (massive amounts of data), Velocity (high-speed data generation), Variety (structured, semi-structured, and unstructured formats), and Veracity (uncertainty and inconsistency) [24]. For instance, social platforms such as Facebook process billions of posts, images, and videos daily, demonstrating both the challenges of scale and speed [19]. Veracity is particularly critical, as ML models depend on reliable inputs; low-quality data can propagate biases and yield unreliable or misleading predictions. Thus, a comprehensive data quality assessment prior to analytical or ML-driven applications is indispensable [25].

Data quality dimensions have been explored extensively in prior research. Beyond core dimensions such as completeness, consistency, reliability, and accuracy, other studies have introduced additional dimensions like conformance (adherence to formats), plausibility (logical validity), timeliness, and provenance, which are especially relevant in domains like healthcare and IoT [26,27]. Poor-quality data not only affects analytics but also leads to significant economic losses, incorrect business insights, and compliance-related risks [2]. This necessitates robust frameworks capable of addressing heterogeneous datasets.

2.1. Data Quality Frameworks

Multiple studies have attempted to define and formalize data quality (DQ) frameworks and standards to enable systematic evaluation. It has been emphasized that ensuring high-quality data is fundamental for extracting value and making reliable decisions [2]. A comprehensive DQ framework aligned with the big data lifecycle—encompassing data generation, acquisition, storage, and analysis—was proposed to emphasize the need for assessment at every stage [28].

Further, models for unstructured big data quality assessment have been introduced, which span from defining quality requirements to generating quality reports using profiling, sampling, and metric-based evaluations [29]. Comparative studies have analyzed a wide range of quality assessment methodologies covering profiling, cleansing, validation, enrichment, and governance, concluding that flexible and context-aware frameworks are critical for practical deployment [7]. These studies collectively suggest that static, one-size-fits-all approaches are inadequate for real-world scenarios, underscoring the need for adaptive, ML-driven solutions.

2.2. ML-Based Data Quality Assessment in Various Domains

Machine learning methods have increasingly been applied to contexts where conventional DQ approaches prove insufficient. These include dynamic and heterogeneous environments such as the following:

- IoT Systems: Supervised ML frameworks employing ensemble Decision Trees and Bayesian classifiers have been proposed to assess the quality of high-frequency and volatile oceanographic sensor data [6]. Key quality dimensions identified include accuracy, timeliness, completeness, consistency, and uniqueness [14,18].
- Healthcare: Standardized DQ terminologies and frameworks for electronic health records (EHRs) have been developed, focusing on conformance, completeness, and plausibility to ensure reliable clinical analytics [30]. Other research highlights the importance of structured and accurate health data for computing clinical quality indicators and supporting medical research [14,21].
- Social Media: Frameworks targeting the heterogeneity and velocity of user-generated content have been proposed, emphasizing lifecycle-oriented quality assessment [19,31]. Further, ML techniques leveraging sentiment analysis and natural language processing have been applied to large-scale platforms such as Airbnb reviews to derive data quality attributes and customer-related insights [32].

Recent advances emphasize that effective data quality frameworks must not only be accurate but also explainable and operationally viable in real time [33]. The field of explainable AI (XAI) [34,35] has grown rapidly, offering methods to make algorithmic decisions transparent and trustworthy for both technical and non-technical stakeholders [35,36]. In parallel, the operationalization of real-time data governance highlights the importance of continuous monitoring, anomaly detection, and quality validation at streaming scale, where latency directly affects user trust and system reliability. Existing DQ solutions often fall short on these dimensions, motivating the development of frameworks like DQ-MAF that combine interpretability with adaptability for integration into live governance pipelines [33,34].

2.3. User-Centric Perspectives on Data Quality

While existing research has predominantly emphasized organizational analytics, an equally critical dimension is the impact of data quality on protecting end-users [13]. Low-quality or manipulated data can mislead authentication systems, distort user profiling, and amplify misinformation on social platforms. For example, duplicate or inconsistent user-generated records may indicate fraudulent behavior, while incomplete or biased textual data may compromise fairness in automated decision-making [9]. High-quality data, therefore, becomes a safeguard for users, ensuring that services remain transparent, secure, and trustworthy.

Recent studies highlight the importance of integrating data quality assessment with user protection mechanisms such as anomaly detection, fake account identification, and robust textual classification [9,13]. These approaches underscore the role of data profiling not only as a tool for system optimization but also as a defense layer against risks faced by unsuspecting users. Consequently, there is growing recognition that quality assessment frameworks must evolve beyond domain-specific analytics to address user-centric concerns such as safety, fairness, and reliability in digital ecosystems.

The reviewed literature highlights a growing reliance on ML for scalable, adaptive, and automated DQ assessment. However, most existing solutions remain domain-specific and lack generalizability. This study aims to address these limitations through the proposed Data Quality Modeling and Assessment Framework (DQMAF), which integrates systematic

profiling, feature engineering, and multiple supervised ML models to achieve accurate and reusable data quality classification while reinforcing user protection across varied datasets.

3. Aim and Objectives

3.1. The Aim

This study aims to design and validate a user-centric, machine learning–driven framework (DQMAF) that not only assesses and classifies data quality in big data environments but also strengthens trust, user protection, and the reliability of Internet services. By systematically profiling datasets, evaluating critical quality dimensions, and leveraging supervised learning models, the framework seeks to safeguard end-users against unreliable services, compromised authentication, and biased decision-making processes.

3.2. Objectives

The specific objectives of this research are as follows:

- To develop a comprehensive data profiling mechanism that extracts hidden indicators of data trustworthiness, supporting both data quality evaluation and user protection;
- To design supervised machine learning models capable of classifying data into predefined categories (high, medium, and low quality), enabling downstream applications such as authentication, document classification, and anomaly detection to operate more reliably;
- To train, optimize, and validate the models for scalable deployment across diverse data domains, including social media, IoT, healthcare, and e-commerce platforms where user safety and service trust are critical;
- To provide a reusable and extensible machine learning–based framework that integrates data quality assessment with proactive governance, thereby enhancing the fairness, transparency, and reliability of analytics and AI-driven services.

4. Research Questions

This study is guided by the following research questions:

- RQ1: How can profiling-driven validations (completeness, consistency, accuracy, structural conformity) be systematically aggregated into interpretable quality labels (high/medium/low) in a reproducible manner?
- RQ2: To what extent can supervised machine learning models (Decision Trees, Random Forest, AdaBoost, XGBoost, CatBoost) accurately classify records into quality tiers when trained on binary profiling features, without overfitting or leakage?

5. Methodology

This section outlines the methodology adopted to design, implement, and evaluate the proposed machine learning-based Data Quality Modeling and Assessment Framework (DQMAF). The methodology follows a structured pipeline consisting of data preprocessing, feature engineering, data profiling, supervised machine learning model training, and evaluation using standard performance metrics. The step-by-step approach ensures that the framework is generalizable and applicable across different big data domains.

5.1. Data Quality Modeling and Assessment Framework (DQMAF)

The DQMAF framework transforms raw heterogeneous data into quality labels (high, medium, low) through a series of well-defined stages. It addresses challenges such as missing values, inconsistent formats, and redundancy that are common in real-world big data systems. The five-stage pipeline comprises the following:

1. Data preprocessing;
2. Feature engineering;
3. Data profiling;
4. Model training;
5. Model evaluation.

This end-to-end design facilitates automated and adaptive quality classification, reducing manual effort and enhancing scalability in big data environments. Figure 1 illustrates the schematic representation of the proposed DQMAF pipeline.



Figure 1. Schematic representation of the proposed Data Quality Modeling and Assessment Framework (DQMAF).

5.2. Data Preprocessing

The initial stage involves preparing raw, heterogeneous data for machine learning analysis. Real-world datasets often suffer from issues such as missing values, inconsistent formats, duplication, and noise. The preprocessing tasks include the following:

- Identifying and handling missing values: Missingness patterns are detected and addressed using multiple imputation strategies (mean, median, and KNN-based imputation).
- Removing duplicates: Duplicate entries are identified and removed to avoid redundancy and bias.
- Data type conversions: Attributes are converted to their appropriate types (e.g., integer, float, categorical) to ensure type consistency.
- Filtering irrelevant features: Non-informative or error-prone features are removed to enhance model performance.
- Ensuring completeness and structural consistency: Logical checks are performed (e.g., validating that the number of beds is not less than the number of bedrooms).

5.3. Feature Engineering

Feature engineering extracts meaningful indicators that reflect the quality of the dataset. This stage transforms raw features into informative representations used by the classification models. Key processes include:

- Missingness Indicators: Binary flags represent the presence or absence of values for each feature.
- Cross-Field Validation: Logical dependencies between attributes (e.g., beds $\geq$ bedrooms) are verified to detect inconsistencies.
- Regular Expression Checks:Attributes such as postal codes, IDs, and emails are validated against expected formats.
- Data Type Consistency: Ensures uniformity in data formats within each column, detecting and flagging anomalies.
- Profiling Score Computation: Each validation produces binary outcomes and weights, which are aggregated into a profiling matrix.

5.4. Data Profiling

Data profiling quantitatively assesses the quality of each feature by systematically applying a set of validation checks. Each check produces a binary (pass/fail) result, which is then weighted to reflect its relative importance. The weighted outcomes are aggregated into a cumulative profiling score that determines whether a record is labeled as high,

medium, or low quality. Table 1 summarizes the full set of validations, weights, thresholds, and rationales.

Table 1. Profiling validations, weights, thresholds, and rationale.

Validation	Weight	Threshold / Rule	Rationale
Completeness	5	≥90% non-missing values per attribute	Expert-driven: Completeness is the most critical determinant of reliability and fairness in downstream analytics.
Consistency	2	Logical rules hold (e.g., beds ≥ bedrooms)	Data-driven: Cross-field errors observed in the Airbnb dataset; weight tuned to reflect their moderate prevalence.
Format Validity	1	Regex match (e.g., postal codes, IDs)	Trade-off: Structural conformity is important, but syntactic errors alone rarely invalidate analytical utility.
Data Type Consistency	1	Values match expected types (int, float, categorical)	Expert-driven: Prevents schema drift and parsing errors; low impact on semantic meaning.
Range/Domain Validity	2	Values fall within expected sets (e.g., city, bed type)	Trade-off: Balances detection of invalid values with tolerance for new/unseen categories.

To map profiling scores into *high*, *medium*, and *low* quality tiers, we employed a quantile-based thresholding approach. Specifically, the cumulative profiling score for each record was computed by aggregating the weighted outcomes of completeness, consistency, format validity, data type consistency, and range/domain checks. The distribution of these cumulative scores was then used to define class boundaries:

- Records with scores at or below the 25th percentile (Q1) were classified as Low quality.
- Records between the 25th and 50th percentiles (Q1–Q2) were classified as Medium quality.
- Records above the 50th percentile (Q2) were classified as High quality.

This data-driven stratification ensures that thresholds are not arbitrarily assigned but instead reflect the statistical distribution of profiling scores within the dataset. It guarantees balanced representation of records across the three tiers, which is particularly important for training supervised machine learning models and avoiding class imbalance issues. The supervised learning models, however, do not operate on these aggregated scores; instead, they are trained on the underlying binary profiling features, such as pass/fail indicators for completeness, regex-based format checks, and type conformity validations. By relying on these diverse and orthogonal indicators rather than the cumulative score itself, the classifiers avoid simply memorizing the label-generation rules. Furthermore, the use of ensemble methods such as Random Forest, XGBoost, and CatBoost provides additional robustness by reducing the risk of overfitting to specific rules and ensuring resilience against potential noise or redundancy in the labels.

5.5. Model Training

The profiled dataset is split into 70% for training and 30% for testing. Five supervised machine learning algorithms are used for classification:

- Decision Trees: provide hierarchical decision-making through interpretable if-else rules. They use criteria such as Information Gain (based on entropy) or Gini Index for splitting:

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (1)$$

where p_i is the proportion of samples belonging to class i. Alternatively, the Gini Index is computed as:

$$Gini(S) = 1 - \sum_{i=1}^{c} p_i^2 \tag{2}$$

Lower Gini or entropy indicates a better split.

- Random Forest: An ensemble method that builds multiple Decision Trees on random subsets of data and features, and averages their predictions. This reduces variance and mitigates overfitting.
- AdaBoost: A boosting algorithm that assigns weights to misclassified samples in each iteration. The weight update rule is written as follows:

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right) \tag{3}$$

where ϵ_t is the classification error at iteration t. Sample weights are updated as follows:

$$w_i^{(t+1)} = w_i^{(t)} \times e^{-\alpha_t y_i h_t(x_i)} \tag{4}$$

where y_i is the true label and $h_t(x_i)$ is the weak classifier prediction.

- XGBoost: An optimized gradient boosting framework. It minimizes an objective function comprising the loss and a regularization term:

$$Obj = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{5}$$

where l is a differentiable loss function (e.g., logistic loss) and $\Omega(f_k)$ is a regularization term to control model complexity.

- CatBoost: Specifically designed for categorical features, CatBoost uses ordered boosting and target statistics to prevent overfitting and prediction shift. It efficiently handles categorical variables without explicit one-hot encoding.

Each model learns to classify data entries into predefined quality categories based on the binary profile features. Hyperparameters were tuned using cross-validation to achieve optimal performance.

5.6. Evaluation Metrics

Model performance was evaluated using widely accepted classification metrics to ensure a holistic assessment:

- Accuracy: proportion of correctly classified samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

- Precision: fraction of true positives among all predicted positives.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

- Recall: fraction of true positives among all actual positives.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

- F1-score: harmonic mean of precision and recall.

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

- Support: number of occurrences of each class label in the dataset.
- Confusion Matrix: tabular representation of actual vs. predicted classes, providing insights into classification errors for each category.

Where TP, TN, FP, and FNdenote True Positives, True Negatives, False Positives, and False Negatives, respectively. These metrics provide a comprehensive view of the model's predictive capabilities and its effectiveness for automated data quality assessment.

6. Experimental Analysis

This section presents the experimental setup and application of the proposed Data Quality Modeling and Assessment Framework (DQMAF) on the Airbnb Price Dataset. It covers the implementation environment, preprocessing, feature engineering, data profiling, and model training stages.

6.1. Implementation Environment

The implementation environment utilized Windows 11 as the base operating system, while Google Colab Proc [37] provided scalable GPU/TPU resources for accelerating the training process. The framework was implemented in Python 3.12.2, leveraging essential libraries such as `pandas` [38] for preprocessing, `scikit-learn` [39] for model development, and `impyute` [40] for missing value imputation, as well as `XGBoost` and `CatBoost` for ensemble-based classification tasks.

The hardware setup comprised an HP Spectre laptop equipped with 8 GB RAM and a 2.4 GHz Quad-Core Intel i5 processor. The combination of cloud-based GPU resources and local hardware ensured both computational efficiency and reproducibility of results. To further enhance reproducibility, all experiments were executed with fixed random seeds and version-controlled Python environments. Dependency management was handled using `pip` and `virtualenv` to ensure consistent package versions across runs.

6.2. Dataset

The dataset used for this study was the publicly available *Airbnb Price Dataset* [41] from Kaggle. It consists of 29 columns and 74,111 entries, including a mixture of categorical, numerical, and textual attributes related to Airbnb listings (e.g., property type, room type, price, location).

The dataset was selected due to its heterogeneous and unprocessed nature, making it highly suitable for evaluating the robustness of DQMAF. Its large size and diverse features mirror real-world big data challenges, including missingness, inconsistency, and redundancy. Moreover, as the dataset is publicly available, it ensures transparency and reproducibility for future research. The dataset also presents real-world anomalies, such as duplicate entries and outliers in numerical fields, making it ideal for testing the robustness of imputation and profiling strategies.

6.3. Data Preprocessing and Imputation

The raw Airbnb dataset contained missing and inconsistent values that required extensive preprocessing before applying the DQMAF framework. Several imputation and cleaning strategies were employed to ensure data integrity:

- Data Inspection: Descriptive statistics and data type summaries were generated to understand the dataset structure (e.g., `int64`, `float64`, `bool`, `object`). Missingness patterns and anomalies were identified.
- Redundant Column Removal: Irrelevant, error-prone, or non-informative columns were eliminated to reduce noise and enhance model focus. In this context, we operationalize the relevance dimension as the suitability of data on two levels: (i) the extent to which data is actually used and accessed by users, and (ii) the degree to which the data produced aligns with user needs. Attributes that are rarely used, provide little utility to end-users, or fail to match their informational needs are considered less relevant, whereas attributes frequently accessed and directly supporting decision-making are deemed highly relevant.
- Shuffling: The dataset was randomized to mitigate ordering biases and improve model generalization.
- Categorical vs Numerical Classification: Columns were categorized into numerical and categorical features to enable tailored preprocessing strategies.

To address missingness, the following imputation strategies were explored:

- Mean Imputation: For numerical attributes, missing values were replaced by the arithmetic mean, calculated as follows:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{10}$$

where $\bar{x}$ denotes the mean and x_i represents the i-th observation.
- Median Imputation: Missing numerical values were replaced with the median of the respective attribute.
- KNN Imputation: Missing entries were estimated based on the k-nearest neighbors using the Euclidean distance, calculated as follows:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{11}$$

where p and q denote two data points in an n-dimensional space.

Additionally, categorical attributes with missing values were imputed using the mode or the most frequent value strategy. Outliers beyond three standard deviations from the mean were flagged and treated separately to maintain consistency. These preprocessing techniques ensured data completeness, improved reliability, and enhanced the performance of subsequent ML models.

6.4. Application of DQMAF on Airbnb Dataset

The DQMAF pipeline was applied to the Airbnb dataset following a structured multi-stage process.

6.4.1. Feature Engineering

In this stage, meaningful features were derived to serve as indicators of data quality. These features included the following:

- Categorical Feature Encoding: Categorical columns (e.g., `property_type`, `room_type`, `bed_type`) were transformed using binary or one-hot encoding to facilitate machine learning.
- Statistical Summaries: Descriptive statistics for numerical attributes were computed to identify trends, variability, and outliers.

- Quality Labeling: A target variable, quality, was defined with three classes: high, medium, and low. This label was inferred from profiling scores and utilized for supervised classification.

6.4.2. Data Profiling

Data profiling is the core of DQMAF, where each column is evaluated based on the following predefined quality metrics:

- Completeness Check (Missingness Indicators): Each cell was examined for null values. Presence of data received a weight of 5, while absence was assigned 0.
- Consistency Checks:
  - Cross-Field Validation: Logical relationships between fields (e.g., `beds` $\geq$ `bedrooms`) were enforced; matches received a weight of 2.
  - Regular Expression Validation: Columns such as `zipcode` and IDs were checked against regex-based format rules.
  - Data Type Consistency: All column values were validated against their expected data types (e.g., integer, string); consistent entries were weighted.
  - Range Consistency: Values for categorical attributes like `city` and `bed_type` were verified against expected ranges.
- Profile Matrix Creation: Each column was transformed into a binary representation based on these checks, with cumulative scores computed.
- Quality Classification: Based on total profile weights, data quality was categorized as high, medium, or low.
- New Dataset Formation: The original dataset was transformed into a structured representation with 49 binary profile features and one quality label.

6.5. Model Training and Evaluation

The final profiled dataset was partitioned into 70% training and 30% testing subsets. Multiple supervised ML models (Decision Tree, Random Forest, AdaBoost, XGBoost, and CatBoost) were trained to map the binary profile features to quality labels (high, medium, and low).

The evaluation phase involved standard classification metrics such as Accuracy, Precision, Recall, and F1-score to compare the performance of models and assess the effectiveness of the DQMAF framework. Confusion matrices were also analyzed to examine class-wise performance and detect any systematic misclassification patterns. Additionally, hyperparameter tuning was conducted using grid search and cross-validation to optimize model performance further.

6.6. Generalizability Across Domains

While the empirical evaluation of DQMAF was conducted on the Airbnb dataset, the framework is not domain-specific. Its profiling dimensions—completeness, consistency, format validity, type conformity, and relevance—are task-agnostic and therefore applicable across heterogeneous data sources. Importantly, the use of a quantile-based thresholding strategy, rather than fixed cutoffs, ensures that the mapping of profiling scores into high, medium, and low quality tiers remains adaptive to datasets with different scales or quality distributions. This makes the approach inherently robust to domain shifts.

Furthermore, DQMAF can be customized to domain-specific requirements: for instance, stricter plausibility and conformity checks may be required in clinical or financial datasets, while higher tolerance for variability may be necessary in IoT or streaming contexts. These adaptations allow the framework to be transferred across application areas

such as healthcare, e-commerce, social media, and sensor-based systems, thereby extending its applicability beyond the dataset presented in this study.

7. Results and Discussion

This section presents and analyzes the performance of various supervised machine learning models applied within the proposed DQMAF framework. Figure 2 illustrates the accuracy scores of all evaluated classifiers, with the x-axis representing their respective accuracy levels. Among the tested models, Random Forest, XGBoost, and CatBoost demonstrated the highest performance levels, while Decision Tree and AdaBoost exhibited comparatively lower accuracy.

To further investigate model behavior, a confusion matrix was generated for the best-performing model—Random Forest—depicted in Figure 3. The confusion matrix provides insights into classification reliability by revealing true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) across the three quality classes: high, medium, and low. This analysis enables a deeper understanding of how well the model distinguishes between each category.

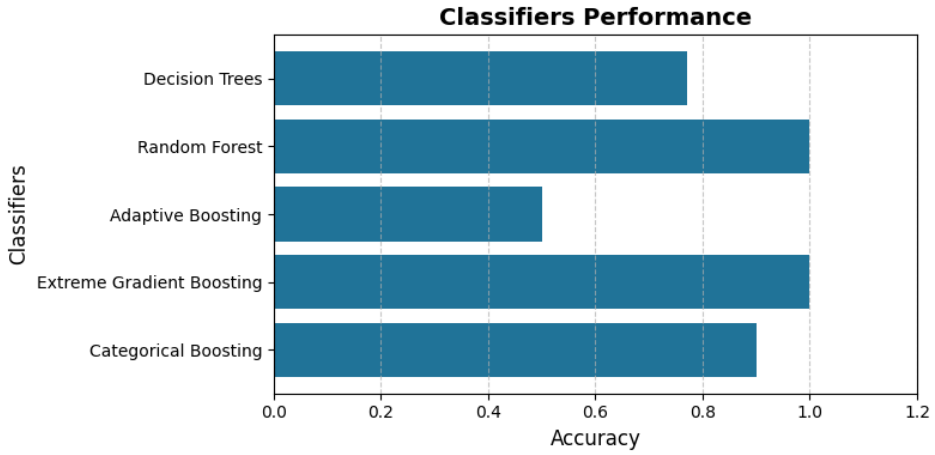


Figure 2. Classification accuracy of different algorithms on the Airbnb Price Dataset. The x-axis represents the accuracy achieved by each model.

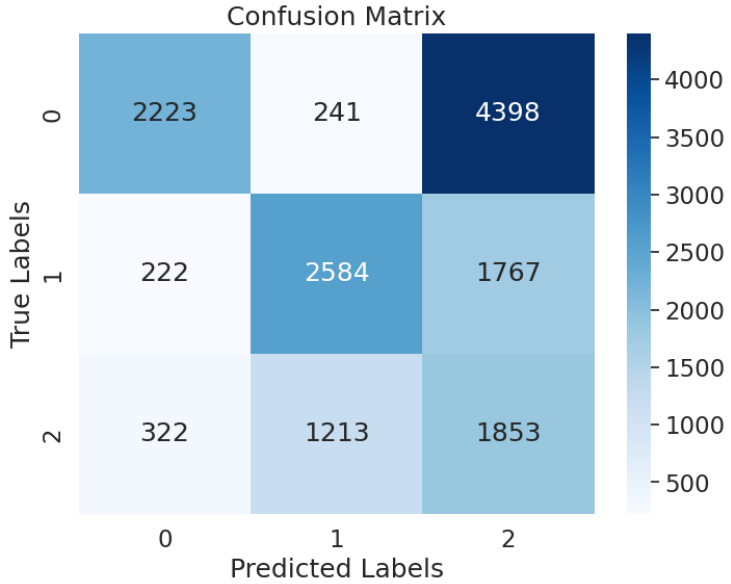


Figure 3. Confusion Matrix for Random Forest Classifier showing its performance in classifying high, medium, and low quality labels.

To quantitatively assess the models, Tables 2–5 report precision, recall, F1-score, and class support for each classifier. As shown in Table 2, the Decision Tree achieved moderate performance with a weighted average F1-score of 0.76. However, the model struggled with the median class, attaining only 0.41 F1-score, which suggests difficulty in correctly identifying mid-quality records.

Table 2. Classification report for Decision Tree model.

	Precision	Recall	F1-Score	Support
high	0.92	0.92	0.92	6862
low	0.65	0.92	0.76	4573
median	0.68	0.30	0.41	3388
macro avg	0.75	0.71	0.70	14,823
weighted avg	0.78	0.78	0.76	14,823

In contrast, the Random Forest model (Table 3) achieved perfect performance across all metrics, with precision, recall, and F1-score values of 1.00 for all classes. This highlights the model's strong ability to separate high, medium, and low quality data instances, likely due to its ensemble nature and robustness to overfitting.

Table 3. Classification report for Random Forest model.

	Precision	Recall	F1-Score	Support
high	1.00	1.00	1.00	6862
low	1.00	1.00	1.00	4573
median	1.00	1.00	1.00	3388
macro avg	1.00	1.00	1.00	14,823
weighted avg	1.00	1.00	1.00	14,823

Similarly, XGBoost also demonstrated perfect classification (Table 4), indicating its capability to effectively handle the profiled dataset. CatBoost (Table 5) also performed exceptionally well, achieving near-perfect scores with a weighted average F1-score of 0.97. This reinforces the strength of gradient boosting approaches in handling categorical variables and capturing complex feature interactions.

Table 4. Classification report for XGBoost model.

	Precision	Recall	F1-Score	Support
high	1.00	1.00	1.00	6862
low	1.00	1.00	1.00	4573
median	1.00	1.00	1.00	3388
macro avg	1.00	1.00	1.00	14,823
weighted avg	1.00	1.00	1.00	14,823

Table 5. Classification report for CatBoost model.

	Precision	Recall	F1-Score	Support
high	0.96	1.00	0.99	6862
low	0.99	0.92	0.95	4573
median	0.91	0.97	0.94	3388
macro avg	0.96	0.96	0.96	14,823
weighted avg	0.97	0.97	0.97	14,823

AdaBoost, on the other hand, underperformed compared to the ensemble boosting models (Table 6). It achieved a weighted average F1-score of only 0.47, again showing poor

classification for the median class. This can be attributed to its sensitivity to noisy data and reliance on weak learners.

Table 6. Classification report for AdaBoost model.

	Precision	Recall	F1-Score	Support
high	0.80	0.32	0.46	6862
low	0.64	0.57	0.60	4573
median	0.23	0.55	0.32	3388
macro avg	0.56	0.48	0.46	14,823
weighted avg	0.62	0.45	0.47	14,823

Overall, the results confirm that the proposed data profiling approach within DQMAF enhances the discriminative power of ML classifiers. Ensemble-based methods such as Random Forest, XGBoost, and CatBoost leverage these engineered features effectively, yielding superior classification performance. The consistently high metrics for these models also demonstrate the scalability and adaptability of DQMAF for complex, heterogeneous big data environments.

7.1. Robustness and Sensitivity Analysis

To assess the stability of the proposed DQMAF framework, additional robustness and sensitivity analyses were performed. These experiments evaluated whether small perturbations in profiling parameters or data splits significantly affected classification outcomes.

Sensitivity Analysis: Profiling weights assigned to the five validation dimensions—completeness, consistency, format validity, data type consistency, and range/domain validity—were independently varied by $\pm 20\%$ around their baseline values. The cumulative profiling scores and corresponding High/Medium/Low labels were recalculated for each variation, followed by model retraining. As shown in Table 7, the performance of the top three classifiers (Random Forest, XGBoost, and CatBoost) remained highly stable, with fluctuations of less than $\pm 1.5\%$ in overall accuracy and weighted F1-scores. This confirms that DQMAF's outcomes are not sensitive to small parameter perturbations and that the learned models generalize consistently under moderate threshold shifts.

Table 7. Sensitivity analysis under $\pm 20\%$ variation in profiling weights.

Model	Baseline Accuracy	Accuracy Range Under $\pm 20\%$ Weight Variation	F1-Score Range
Random Forest	1.00	0.99–1.00	0.99–1.00
XGBoost	1.00	0.99–1.00	0.99–1.00
CatBoost	0.97	0.96–0.98	0.96–0.97
Decision Tree	0.78	0.76–0.79	0.75–0.77
AdaBoost	0.47	0.45–0.49	0.44–0.46

Robustness Analysis: To further validate generalization robustness, each classifier was trained and tested across five different random splits (70/30 ratio). The mean and standard deviation of accuracy across these trials are presented in Table 8. The results reveal minimal variance (standard deviation < 0.01) for the top-performing ensemble models, underscoring the reliability and reproducibility of the DQMAF pipeline. Models with lower baseline accuracy, such as Decision Tree and AdaBoost, exhibited slightly higher variability, consistent with their simpler architectures and sensitivity to class boundaries.

Table 8. Robustness analysis across multiple random train-test splits.

Model	Mean Accuracy (5 Splits)	Std. Deviation
Random Forest	0.999	0.001
XGBoost	0.998	0.002
CatBoost	0.972	0.004
Decision Tree	0.776	0.006
AdaBoost	0.462	0.009

7.2. Future Work

Future research can extend DQMAF to include unsupervised and semi-supervised models for scenarios where labeled data is scarce. Incorporating streaming architectures (e.g., Apache Spark or Flink) would enable real-time quality monitoring, while additional quality dimensions such as timeliness and provenance could support domain-specific adaptations. Moreover, integrating explainable AI (XAI) techniques can enhance interpretability for stakeholders. To ensure the framework's generalizability, future work can involve extensive testing across diverse datasets and deployment within real-time big data pipelines.

8. Conclusions

The growing complexity and heterogeneity of big data continue to introduce significant challenges in ensuring and maintaining data quality, particularly with respect to key dimensions such as completeness, consistency, accuracy, and structural conformity. Poor-quality data has far-reaching implications, affecting not only analytical outcomes but also user trust, authentication processes, and the reliability of digital services. In response to these challenges, this study presented DQMAF, a generalized five-stage, machine learning–based framework designed to systematically transform raw, heterogeneous data into structured, quality-labeled outputs. The proposed framework integrates comprehensive data profiling, feature engineering, and supervised machine learning algorithms to automate and enhance data quality assessment while directly contributing to user protection and safer digital interactions. By evaluating the Airbnb dataset, which reflects real-world big data challenges due to its volume, diversity, and textual attributes, DQMAF demonstrated exceptional performance across multiple models. Notably, Random Forest and XGBoost achieved 100% classification accuracy, while CatBoost achieved 97%, highlighting the framework's robustness, scalability, and ability to capture nuanced indicators of trustworthiness effectively. These state-of-the-art results illustrate the strong potential of DQMAF to serve as a foundation not only for scalable data governance solutions but also for applications in user profiling, document and text classification, and anomaly detection in online platforms. Beyond its empirical results, the study contributes a reusable and adaptable methodology that can be applied across multiple domains where both data quality and user safety are critical, such as healthcare, IoT, and social media analytics. In summary, DQMAF addresses a crucial gap by providing an effective, scalable, and user-centric machine learning–driven approach for automated data quality assessment, with substantial implications for protecting end-users, strengthening authentication systems, and improving the fairness, transparency, and trustworthiness of analytics and AI-driven decision-making in modern big data ecosystems. Looking ahead, it is also important to acknowledge the ethical implications of using automated systems for data quality assessment. While DQMAF already promotes interpretability through its reliance on transparent profiling features, future extensions could further enhance transparency and stakeholder trust by incorporating explainable AI techniques and user-facing audit mechanisms. These additions would

ensure that automated decisions remain not only accurate and scalable but also accountable and aligned with responsible data governance principles.

Author Contributions: Conceptualization, A.A.; methodology, A.A.; software, R.A.-T.; validation, R.A.-T.; formal analysis, R.A.-T.; investigation, R.A.-T.; resources, A.A.; data curation, R.A.-T.; writing—original draft preparation, R.A.-T.; writing—review and editing, A.A.; visualization, R.A.-T.; supervision, A.A.; project administration, A.A.; funding acquisition, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ongoing Research Funding program, (ORF-2025-1253), King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are openly available in the Airbnb Price Dataset on Kaggle.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zeng, S.; Jin, W.; Sun, T. The Value of Data in the Digital Economy: Investigating the Economic Consequences of Personal Information Protection. 2025. Available online: https://ssrn.com/abstract=5147162 (accessed on 10 October 2025).
2. Cai, L.; Zhu, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci. J.* **2015**, *14*, 2. [CrossRef]
3. Li, G.; Zhou, X.; Cao, L. Machine Learning for Databases. In Proceedings of the AIMLSystems 2021: The First International Conference on AI-ML-Systems, Bangalore, India, 21–23 October 2021; ACM: New York, NY, USA, 2021; pp. 1–2. [CrossRef]
4. Warwick, W.; Johnson, S.; Bond, J.; Fletcher, G.; Kanellakis, P. A Framework to Assess Healthcare Data Quality. *Eur. J. Soc. Behav. Sci.* **2015**, *13*, 92–98. [CrossRef]
5. Priestley, M.; O'Donnell, F.; Simperl, E. A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *J. Data Inf. Qual.* **2023**, *15*, 11. [CrossRef]
6. Rahman, A.; Smith, D.V.; Timms, G. A Novel Machine Learning Approach Toward Quality Assessment of Sensor Data. *IEEE Sens. J.* **2014**, *14*, 1035–1047. [CrossRef]
7. Cichy, C.; Rass, S. An Overview of Data Quality Frameworks. *IEEE Access* **2019**, *7*, 24634–24648. [CrossRef]
8. Tencent Cloud Techpedia. What Is the Difference Between Data Availability and Data Quality? 2025. Available online: https://www.tencentcloud.com/techpedia/108108 (accessed on 2 September 2025).
9. Declerck, J.; Kalra, D.; Vander Stichele, R.; Coorevits, P. Frameworks, Dimensions, Definitions of Aspects, and Assessment Methods for the Appraisal of Quality of Health Data for Secondary Use: Comprehensive Overview of Reviews. *JMIR Med. Inform.* **2024**, *12*, e51560. [CrossRef] [PubMed]
10. Data Reliability in 2025: Definition, Examples & Tools. 2024. Available online: https://atlan.com/what-is-data-reliability/#:~:text=Data%20reliability%20means%20that%20data,business%20analytics%2C%20or%20public%20policy (accessed on 2 September 2025).
11. The Three Critical Pillars of Data Reliability—Acceldata. 2025. Available online: https://www.acceldata.io/guide/three-critical-pillars-of-data-reliability (accessed on 2 September 2025).
12. He, D.; Liu, X.; Shi, Q.; Zheng, Y. Visual-language reasoning segmentation (LARSE) of function-level building footprint across Yangtze River Economic Belt of China. *Sustain. Cities Soc.* **2025**, *127*, 106439. [CrossRef]
13. Randell, R.; Alvarado, N.; McVey, L.; Ruddle, R.A.; Doherty, P.; Gale, C.; Mamas, M.; Dowding, D. Requirements for a quality dashboard: Lessons from National Clinical Audits. *AMIA Annu. Symp. Proc.* **2020**, *2019*, 735–744.
14. Reimer, A.P.; Milinovich, A.; Madigan, E.A. Data quality assessment framework to assess electronic medical record data for use in research. *Int. J. Med. Informatics* **2016**, *90*, 40–47. [CrossRef]
15. Janssen, M.; van der Voort, H.; Wahyudi, A. Factors influencing big data decision-making quality. *J. Bus. Res.* **2017**, *70*, 338–345. [CrossRef]
16. Jerez, J.M.; Molina, I.; García-Laencina, P.J.; Alba, E.; Ribelles, N.; Martín, M.; Franco, L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif. Intell. Med.* **2010**, *50*, 105–115. [CrossRef]
17. Mahdavinejad, M.S.; Rezvan, M.; Barekatain, M.; Adibi, P.; Barnaghi, P.; Sheth, A.P. Machine learning for internet of things data analysis: A survey. *Digit. Commun. Netw.* **2018**, *4*, 161–175. [CrossRef]

18. Rahm, E.; Do, H.H. Data Cleaning: Problems and Current Approaches. *IEEE Data Eng. Bull.* **2000**, *23*, 3–13.
19. Immonen, A.; Paakkonen, P.; Ovaska, E. Evaluating the Quality of Social Media Data in Big Data Architecture. *IEEE Access* **2015**, *3*, 2028–2043. [CrossRef]
20. Suzuki, K. Pixel-Based Machine Learning in Medical Imaging. *Int. J. Biomed. Imaging* **2012**, *2012*, 792079. [CrossRef]
21. Dentler, K.; Cornet, R.; Teije, A.T.; Tanis, P.; Klinkenbijl, J.; Tytgat, K.; Keizer, N.D. Influence of data quality on computed Dutch hospital quality indicators: A case study in colorectal cancer surgery. *BMC Med. Inform. Decis. Mak.* **2014**, *14*, 32. [CrossRef]
22. Marupaka, D. Machine Learning-Driven Predictive Data Quality Assessment in ETL Frameworks. *Int. J. Comput. Trends Technol.* **2024**, *72*, 53–60. [CrossRef]
23. Frank, E. Machine Learning Models for Data Quality Assessment. EasyChair Preprint 13213, EasyChair, 2024. Available online: https://easychair.org/publications/preprint/cktz (accessed on 10 October 2025).
24. Nelson, G. Data Management Meets Machine Learning. In Proceedings of the SAS Global Forum, Denver, CO, USA, 8–10 April 2018. Available online: https://support.sas.com/resources/papers/proceedings18/ (accessed on 10 October 2025).
25. Zhou, Y.; Tu, F.; Sha, K.; Ding, J.; Chen, H. A Survey on Data Quality Dimensions and Tools for Machine Learning Invited Paper. In Proceedings of the 2024 IEEE International Conference on Artificial Intelligence Testing (AITest), Shanghai, China, 15–18 July 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 120–131. [CrossRef]
26. Batini, C.; Scannapieco, M. Data Quality Dimensions. In *Data and Information Quality*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–51. [CrossRef]
27. Gong, Y.; Liu, G.; Xue, Y.; Li, R.; Meng, L. A survey on dataset quality in machine learning. *Inf. Softw. Technol.* **2023**, *162*, 107268. [CrossRef]
28. Taleb, I.; Serhani, M.A.; Dssouli, R. Big Data Quality Assessment Model for Unstructured Data. In Proceedings of the 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 18–19 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 69–74. [CrossRef]
29. Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A.V. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **2017**, *237*, 350–361. [CrossRef]
30. Kahn, M.G.; Callahan, T.J.; Barnard, J.; Bauck, A.E.; Brown, J.; Davidson, B.N.; Estiri, H.; Goerg, C.; Holve, E.; Johnson, S.G.; et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMs (Generating Evid. Methods Improv. Patient Outcomes)* **2016**, *4*, 18. [CrossRef]
31. Reda, O.; Zellou, A. Assessing the quality of social media data: A systematic literature review. *Bull. Electr. Eng. Inform.* **2023**, *12*, 1115–1126. [CrossRef]
32. Amat-Lefort, N.; Barravecchia, F.; Mastrogiacomo, L. Quality 4.0: Big data analytics to explore service quality attributes and their relation to user sentiment in Airbnb reviews. *Int. J. Qual. Reliab. Manag.* **2022**, *40*, 990–1008. [CrossRef]
33. Papastergios, V.; Gounaris, A. Stream DaQ: Stream-First Data Quality Monitoring. *arXiv* **2025**, arXiv:2506.06147. [CrossRef]
34. Sarr, D. Towards Explainable Automated Data Quality Enhancement Without Domain Knowledge. *arXiv* **2024**, arXiv:2409.10139. [CrossRef]
35. Angelov, P.P.; Soares, E.A.; Jiang, R.; Arnold, N.I.; Atkinson, P.M. Explainable artificial intelligence: An analytical review. *WIREs Data Min. Knowl. Discov.* **2021**, *11*, e1424. [CrossRef]
36. Costa e Silva, E.; Oliveira, O.; Oliveira, B. Enhancing Real-Time Analytics: Streaming Data Quality Metrics for Continuous Monitoring. In Proceedings of the ICoMS 2024: 2024 7th International Conference on Mathematics and Statistics, Amarante, Portugal, 23–25 June 2024; ACM: New York, NY, USA, 2024; pp. 97–101. [CrossRef]
37. Google Colab. 2025. Available online: https://colab.research.google.com/ (accessed on 23 August 2025).
38. Pandas: Python Data Analysis Library. 2025. Available online: https://pandas.pydata.org/ (accessed on 23 August 2025).
39. Scikit-Learn: Machine Learning in Python. 2025. Available online: https://scikit-learn.org/ (accessed on 23 August 2025).
40. Impyute: Missing Data Imputation Library. 2025. Available online: https://impyute.readthedocs.io/en/master/ (accessed on 23 August 2025).
41. Airbnb Price Dataset on Kaggle. 2025. Available online: https://www.kaggle.com/datasets/rupindersinghrana/airbnb-price-dataset (accessed on 23 August 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.