PRESENTATION :

Title : A Machine Learning and Rule-Based Approach for Data Quality Validation in Traffic Collision Data

Student Information

Student Name : Jayavardhan Premnath

Student Id : 20046512

Programme : MSc. in Data Analytics

Module : Applied Research Project [CA_ONE_(10%)]

Supervisor : Swati Dongre

1. Research Motivation & Objective

Core Motivation :

- **Use of Traffic Collision Data**

  Traffic collision datasets are large, complex, and safety-critical, making data quality essential for reliable analysis in traffic safety, urban planning, and policy decision-making.

- **Challenges in Real-World Data Quality**

  Real-world traffic datasets frequently contain missing values, inconsistencies, logical errors, and spatial noise, which can significantly impact analytical accuracy.

- **Limitations of Manual Data Cleaning**

  Traditional manual data cleaning approaches are time-consuming, error-prone, and not scalable for large-scale datasets.

- **Need for Automated and Scalable Solutions**

  There is a strong need for an automated, robust data validation and enhancement framework that produces analysis-ready datasets while reducing human effort and processing time.

Research Objective:

- To design and evaluate a hybrid data quality framework that combines rule-based validation techniques with machine learning methods for automated data validation and enhancement in traffic collision datasets.

2. Dataset Overview & Data Quality Challenges

- **Dataset:** NYC Motor Vehicle Collisions Dataset
- **Source:** New York City Open Data Portal
- **Dataset Size:**
  - **Rows:** 1,048,576 records
  - **Columns:** 29 attributes
- **Key Characteristics:**
  - High-volume, multi-year real-world traffic collision data
  - Mix of numerical, categorical, temporal, and geospatial features
- **Observed Data Quality Issues:**
  - Missing and inconsistent injury counts
  - Logical violations (e.g., pedestrians injured > total persons injured)
  - Invalid or noisy latitude and longitude values
  - Schema inconsistencies across files

3. Proposed Methodology – Hybrid Framework

This study adopts the Design Science Research (DSR) approach, as conceptualized by Hevner et al. (2004).
DSA approach is a problem-solving approach that focuses on the development and rigorous evaluation of a functional artefact such as a software framework or algorithm to address a specific real-world challenge.

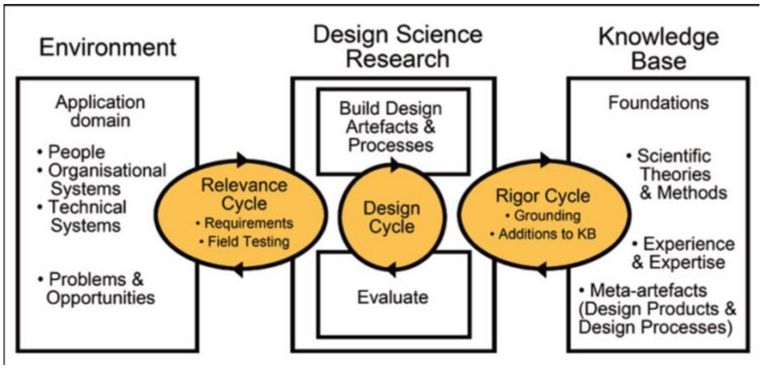


Figure1 : Research framework (Hevner et al., 2004).

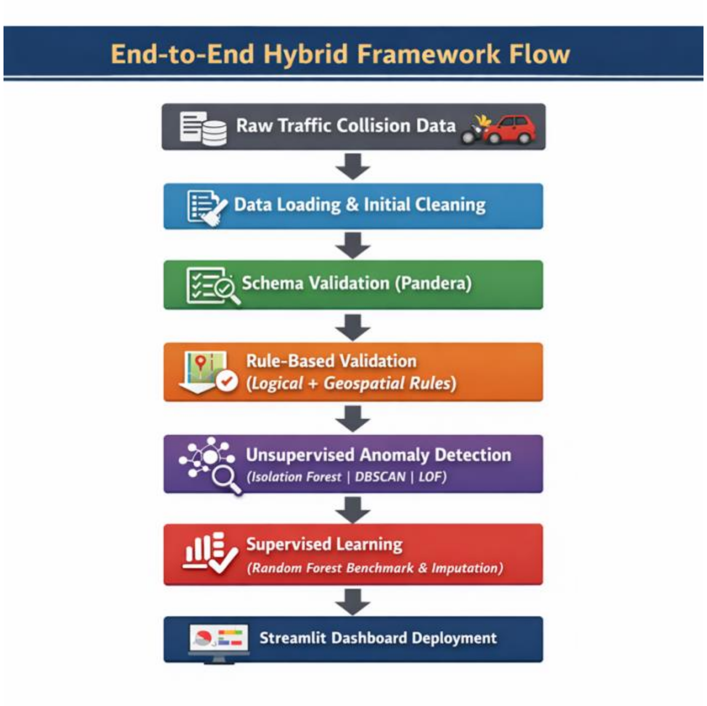

End-to-End Hybrid Framework Flow
Raw Traffic Collision Data
Data Loading & Initial Cleaning
Schema Validation (Pandera)
Rule-Based Validation
(Logical + Geospatial Rules)
Unsupervised Anomaly Detection
(Isolation Forest | DBSCAN | LOF)
Supervised Learning
(Random Forest Benchmark & Imputation)
Streamlit Dashboard Deployment

Figure 2 : Framework Flow

This flow illustrates how **rule-based validation**, **machine learning**, and **visual analytics** are integrated into a single automated data quality pipeline.

4. Data Preprocessing & Feature Engineering

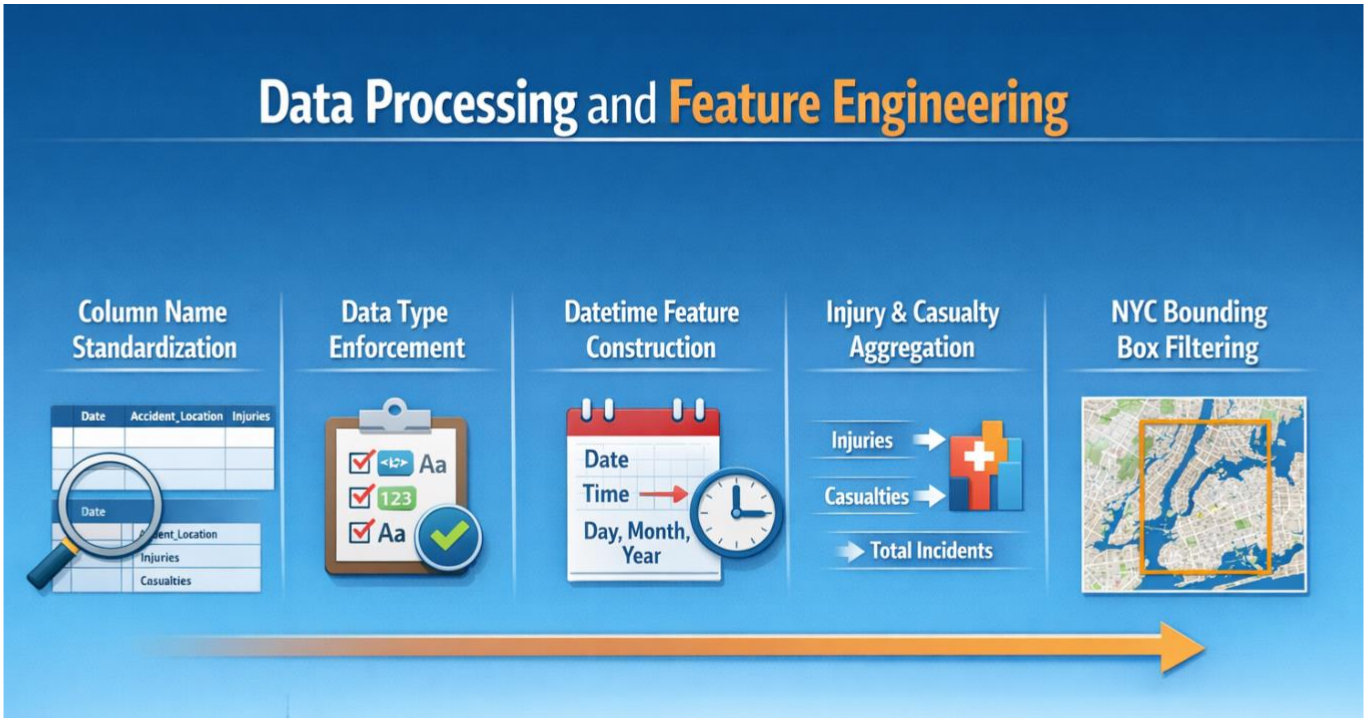


Figure 3 : Data processing and Feature Engineering

5. Rule-Based Validation Layer

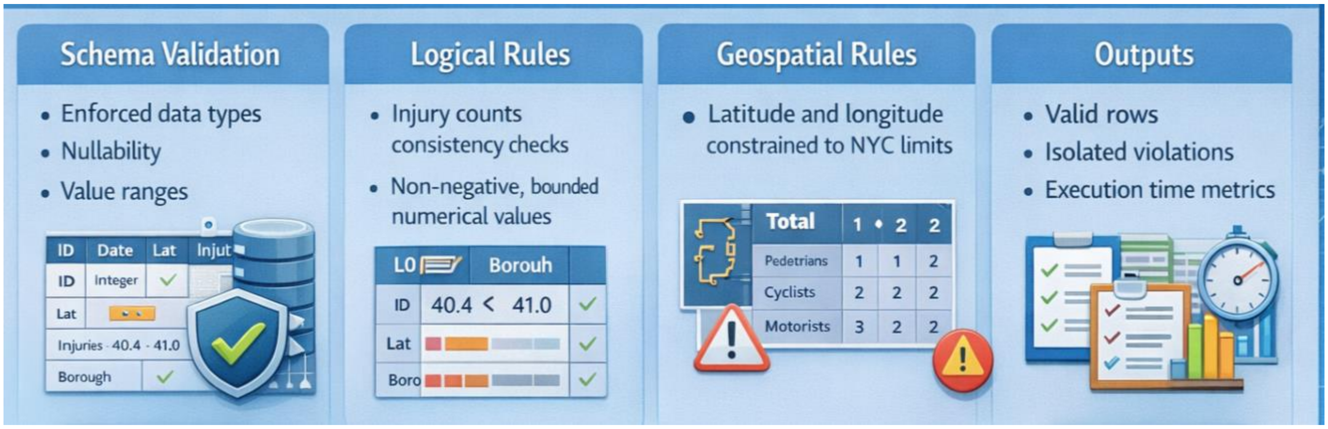


Figure 4 : Rule Based Validation Layer

6. Machine Learning Models Used

Model	Role in the System	Key Advantages	Limitations	Best Fit for This Project
Isolation Forest	Primary anomaly detection model used to identify global outliers in large-scale collision data	Fast, scalable, works well on high-dimensional data	Assumes anomalies are few and randomly distributed	**⭐ Yes – main unsupervised anomaly detector**
DBSCAN (Sampled)	Detects spatial clusters and noise in collision locations	Identifies dense regions and spatial anomalies effectively	Sensitive to parameter tuning and requires sampling	**⚠ Partially – exploratory spatial analysis**
Local Outlier Factor (LOF)	Detects local density-based anomalies in geospatial data	Effective in areas with varying spatial density	Computationally expensive on large datasets	**⚠ Supplementary – spatial refinement**
Random Forest (Rule-Based Labels)	Acts as a supervised benchmark to evaluate rule-based anomaly detection	High accuracy, robust to noise and feature interactions	Performance depends on quality of rule labels	**⭐ Yes – strong supervised benchmark**
Random Forest (Borough Imputation)	Predicts missing borough values using spatial coordinates	Improves data completeness and usability	Requires sufficient labelled training data	**⭐ Yes – data enhancement task**

Table 1 : Machine Learning Models Comparison

Key Observation :

Isolation Forest is the primary anomaly detector, Random Forest acts as a supervised benchmark and data enhancer, while DBSCAN, and LOF support spatial and density-based validation.

7. Performance Evaluation & Results

Evaluation Metrics Used - Precision - Recall - F1-score - Error Detection Rate (EDR)

Model	Precision	Recall	F1-Score	EDR
Isolation Forest	High	Moderate	Strong	High
DBSCAN (Sampled)	Moderate	Low	Moderate	Moderate
LOF	Low	Moderate	Low	Moderate
Random Forest (Supervised)	Very High	Very High	Best	Very High
Rule-Based Validation	N/A	1.00	N/A	1.00

Table 2 : Performace Summary of Models

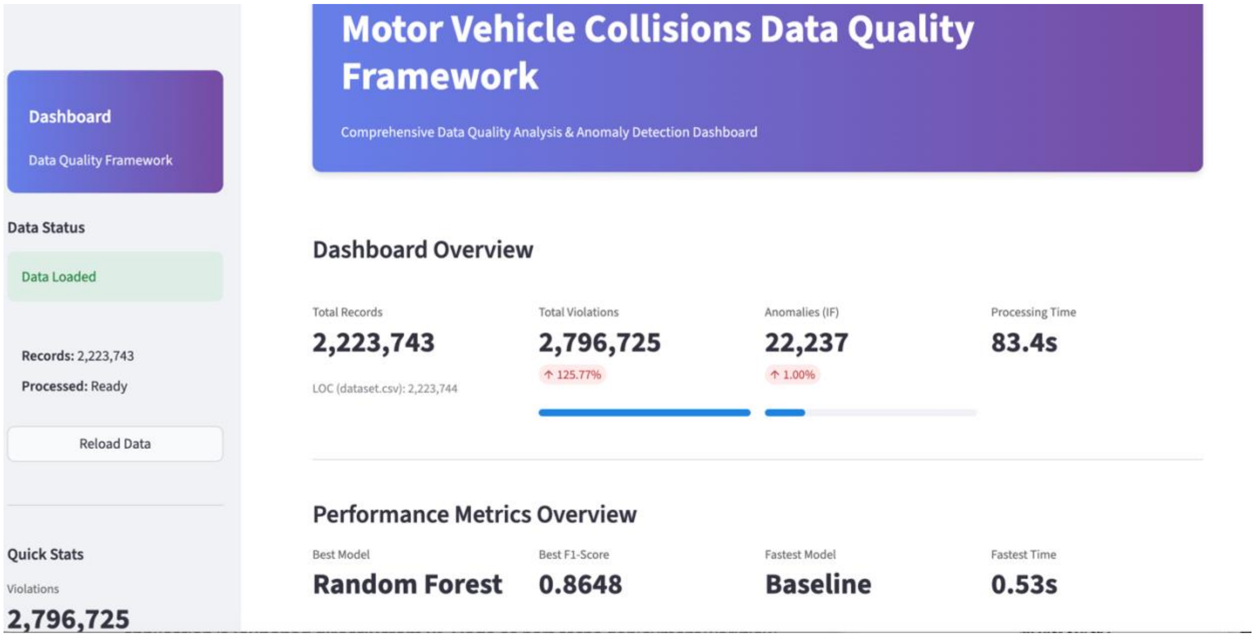
Conclusion :

The Random Forest (Supervised) model performed best overall with the highest precision, recall, F1-score, and EDR, while Isolation Forest provided strong unsupervised detection; DBSCAN showed moderate effectiveness, LOF performed weakest, and rule-based validation ensured perfect recall for known violations but could not detect complex unseen anomalies.

8. Streamlit Dashboard Demonstration

An **analytical Streamlit dashboard** was developed to provide a **high-level overview of detected anomalies and validation outcomes**, enabling users to clearly understand the nature and distribution of data quality issues.

The dashboard code is developed and maintained in **VS Code**, and the Streamlit application is **launched directly from VS Code** as part of the deployment workflow.



Dashboard

Data Quality Framework

Data Status

Data Loaded

Records: 2,223,743

Processed: Ready

Quick Stats

Violations

2,796,725

Motor Vehicle Collisions Data Quality Framework

Comprehensive Data Quality Analysis & Anomaly Detection Dashboard

Dashboard Overview

Total Records
2,223,743
LOC (dataset.csv): 2,223,744

Total Violations
2,796,725
↑ 125.77%

Anomalies (IF)
22,237
↑ 1.00%

Processing Time
83.4s

Performance Metrics Overview

Best Model
Random Forest

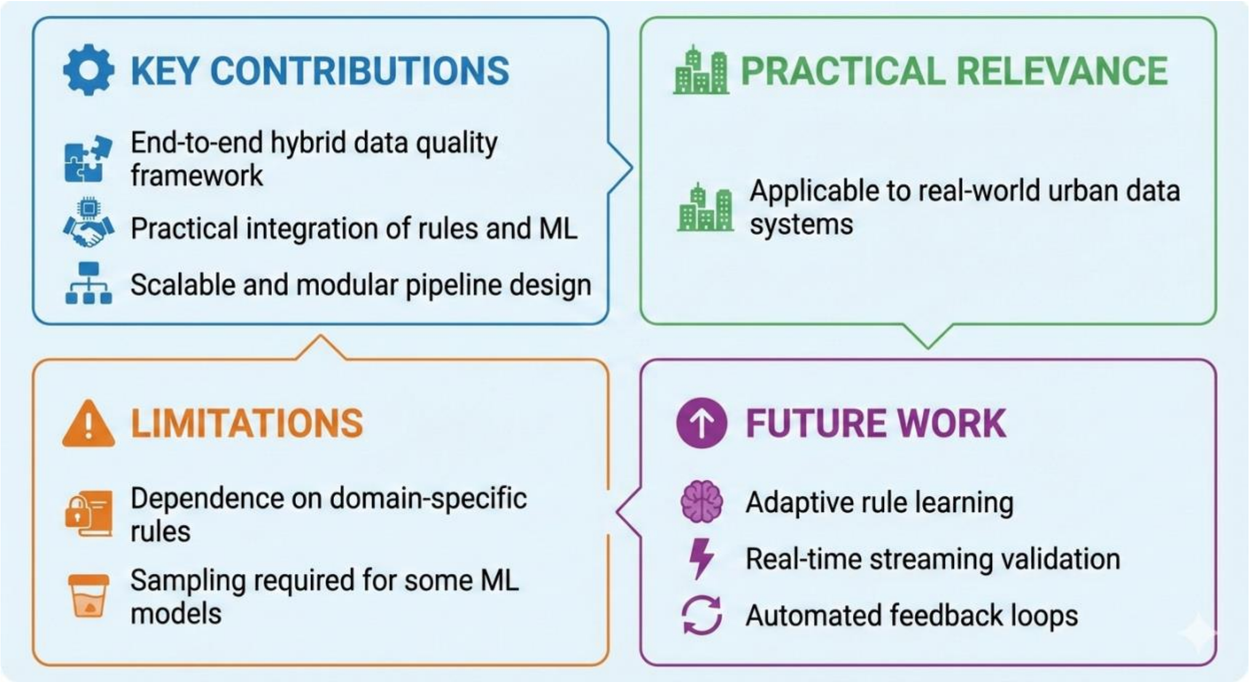
Best F1-Score
0.8648

Fastest Model
Baseline

Fastest Time
0.53s

Figure 5 : Streamlit Dashboard

9. Conclusion & Contributions



KEY CONTRIBUTIONS

- End-to-end hybrid data quality framework
- Practical integration of rules and ML
- Scalable and modular pipeline design

PRACTICAL RELEVANCE

- Applicable to real-world urban data systems

LIMITATIONS

- Dependence on domain-specific rules
- Sampling required for some ML models

FUTURE WORK

- Adaptive rule learning
- Real-time streaming validation
- Automated feedback loops

Figure 6 :Conclusions and Contributions

End of Presentation

Thank You !