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

Jayavardhan Premnath (20046512)

**Applied Research Project submitted in partial fulfilment of the requirements for the degree of
MSc in Data Analytics
at Dublin Business School**

Supervisor: Swati Dongre

January 2026

Declaration page

I declare that this Applied Research Project that I have submitted to Dublin Business School for the award of Msc. Data Analytics is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Signed: Jayavardhan
Student Number: 20046512
Date: 7/01/2026

Table of Contents

ABSTRACT

This applied research developed and evaluated a modular rule based and machine learning framework for data quality validation in traffic collision datasets. The aim of the study was to detect structural inconsistencies, logical rule violations, and anomalous collision records that reduce the reliability of safety analytics. The proposed framework functions as a modular validation artefact, combining schema validation, deterministic rule based checks, and unsupervised anomaly detection using Isolation Forest, Local Outlier Factor (LOF), and DBSCAN, with Random Forest applied as a benchmark reference model. Performance was assessed using Precision, Recall, F1Score, and Error Detection Rate (EDR). The results showed that Isolation Forest achieved the highest overall performance (Precision 88.5%, Recall 83.7%, F1Score 86.0%, EDR 82.4%), followed by Random Forest as the strongest benchmark comparator (F1Score 88.8%).

Overall, the findings demonstrate that the hybrid validation approach improves anomaly detection coverage while maintaining transparency and interpretability, enabling automated validation that enhances downstream analytical reliability while reducing manual data quality effort in real-world traffic collision datasets.

Chapter One

1.1 Background study

The increasing reliance on traffic analytics, road safety intelligence, and transport incident monitoring systems was observed to have significantly increased the demand for high-quality traffic collision data. Transport authorities, urban planning agencies, and law enforcement departments were found to routinely collect large volumes of structured collision records to support accident analysis, policy development, infrastructure assessment, and public safety decision-making (Elvik and Vaa, 2004, p. 51). However, despite improvements in traffic data recording technologies and digital reporting systems, real-world traffic collision datasets continued to exhibit persistent data quality issues that limited analytical reliability and operational usability.

Traffic collision datasets were frequently found to contain missing attributes, duplicated incident records, inconsistent location fields, temporal misalignments, and violations of structural formatting requirements (Ziakopoulos and Yannis, 2020, p. 4). These quality deficiencies commonly originated from manual reporting variations, multiagency data integration workflows, sensor or GPS inaccuracies, and inconsistencies in accident documentation standards (Lord and Mannering, 2010, p. 592). When such issues were not identified and corrected, they were propagated across downstream safety modelling pipelines, leading to misleading statistical outcomes and unreliable risk assessment insights (Montella, 2010, p. 693). As a result, data quality was recognised as a fundamental prerequisite for credible road safety analytics and traffic collision research.

Data cleaning and validation processes were therefore considered essential for improving the integrity and interpretability of traffic collision datasets. Data cleaning involved the correction of incomplete, inconsistent, or duplicated collision attributes, while validation ensured that records adhered to structural, spatial, and logical constraints such as realistic coordinates, valid timestamps, or consistent severity classifications (AgueroValverde and Jovanis, 2006, p. 3). However, studies consistently reported that a substantial proportion of analytical effort within traffic safety analysis was allocated to manual preprocessing, which increased development time and introduced human subjectivity into the validation process (Chen et al., 2016, p. 98).

Traditional traffic data validation approaches were predominantly rule-based, relying on predefined thresholds, logical conditions, or domainspecific constraints (Haleem, AbdelAty and Mackie, 2010, p. 531). While these methods provided transparency and explainability, they required continuous revision as reporting standards evolved or as new datasets were introduced. This dependency on manual curation made conventional rule based validation difficult to scale across diverse regional and organisational traffic reporting systems (Montella, 2010, p. 701).

To address these limitations, recent research increasingly explored the application of machine learning based anomaly detection to traffic collision data. Algorithms such as Isolation Forest and density based anomaly detection were applied to identify unusual or inconsistent collision patterns that were not easily captured through explicit rules (Tavassoli, Lord and Geedipally, 2018, p. 41). These approaches were particularly effective in largescale datasets where hidden anomalies or irregularities were difficult to detect manually. However, it was also observed that anomaly detection methods occasionally flagged rare but valid incidents as outliers, thereby introducing interpretability challenges in operational contexts (Mannering, Bhat and Shankar, 2020, p. 6).

Schema based validation was also recognised as an important component of traffic collision data quality assurance, ensuring that each record complied with expected structural definitions such as field formats, attribute completeness, and valid data types (Kleppmann, 2017, p. 102). While schema enforcement improved structural consistency, it remained insufficient for detecting semantic or contextual inconsistencies in isolation, such as unrealistic crash severity combinations or geographically implausible coordinates (Ziakopoulos and Yannis, 2020, p. 6). Across the literature, it was observed that rule based validation, schema enforcement, and anomaly detection techniques were frequently applied in isolation rather than as part of an integrated framework. Existing tools were often designed for region-specific or agency specific datasets, limiting their adaptability to different traffic reporting environments and contributing to fragmented validation practices (Lord and Mannering, 2010, p. 601). This fragmentation resulted in duplicated effort, variable quality standards, and sustained reliance on manual preprocessing. The limitations identified in prior research highlighted a clear need for a combined machine learning and rule based approach to traffic collision data validation, capable of supporting scalability, transparency, and repeatability. A modular validation structure that integrated rule driven domain logic with machine learning based anomaly detection was therefore considered essential for improving consistency, accuracy, and operational trust in traffic collision datasets (Mannering, Bhat and Shankar, 2020, p. 8). This need provided the foundational rationale for the present applied research study.

1.2 Problem Statement

The increasing reliance on traffic collision data for road safety analysis, accident prediction, urban transport planning, and policy development was observed to have heightened the need for reliable and high-quality structured collision datasets. Traffic accident records collected by police departments, transportation agencies, and municipal authorities were frequently integrated from multiple reporting sources and digital logging systems. However, despite improvements in data collection infrastructures, these datasets were consistently affected by data quality issues such as missing attributes, duplicated incidents, inconsistent spatial references, temporal misalignments, and violations of structural constraints. When such issues were not systematically detected and corrected, they were propagated through analytical pipelines, resulting in misleading accident statistics, biased safety models, and unreliable decision-making outcomes.

Existing data cleaning and validation practices in traffic collision datasets were predominantly manual and rule based. These approaches depended on predefined thresholds, logical constraints, or reporting guidelines, which were formulated through domain expertise and agency specific conventions. Although such methods provided transparency and interpretability, they required continuous manual revision as data recording formats, reporting procedures, and collection platforms evolved. This dependency on manual intervention made conventional validation processes time-consuming, difficult to reproduce, and insufficiently scalable for largescale traffic datasets.

To address these limitations, unsupervised machine learning–based anomaly detection techniques had increasingly been applied to identify unusual or inconsistent collision records that were not captured through predefined rules. While these approaches improved anomaly coverage, they also introduced challenges related to interpretability and the misclassification of rare but valid incidents. As a result, neither rule based validation nor anomaly detection alone provided a comprehensive or operationally dependable solution to traffic collision data quality assurance.

A further limitation was identified in the lack of integrated validation frameworks within existing practice. Rule based checks, schema enforcement, and anomaly detection were typically applied in isolation, resulting in fragmented and inconsistent validation workflows. Traffic analysts and practitioners continued to rely on manual preprocessing activities, which increased effort, reduced reproducibility, and limited confidence in analytical outputs derived from collision datasets.

Accordingly, the central problem addressed in this research was the absence of a unified and modular validation framework capable of combining rule based validation and machine learning anomaly detection for traffic collision data. A structured and automated approach was required to improve data consistency, enhance anomaly detection coverage, reduce manual preprocessing effort, and support transparent and reliable use of traffic collision datasets in analytical and operational contexts.

1.3 Aim of research

The aim of this applied research was to design, develop, and evaluate a machine learning and rule based validation framework for improving data quality in traffic collision datasets. The study focused on building a modular validation pipeline in which rule based checks were used to detect structural and logical inconsistencies, while machine learning techniques were applied to identify anomalous and irregular collision records. The framework was implemented as a functional software artefact and was evaluated using real-world traffic collision data to assess its effectiveness in improving data consistency, anomaly detection coverage, and reduction of manual preprocessing effort. Through this approach, the research aimed to provide a transparent, scalable, and repeatable method for traffic collision data quality validation, supporting reliable analytical and decision-making processes.

1.3.1 Objectives of the Research

The research problem addressed in this study relates to the presence of persistent data quality issues within traffic collision datasets, including missing geographic coordinates, duplicated collision identifiers, inconsistent temporal attributes, and anomalous spatial or casualty related event patterns. These issues reduce the reliability of safety analytics, distort collision trend interpretation, and weaken decision support processes used in road safety planning and policy development (Lee and AbdelAty, 2019). To address this problem, the present research focuses on the development and evaluation of a modular machine learning and rule based validation framework specifically designed to improve the structural integrity, consistency, and analytical reliability of traffic collision data (Ahmed et al., 2021).

Accordingly, the primary objective of this study is to design and implement a validation framework that integrates schema validation, deterministic rule based checks, and unsupervised anomaly detection to systematically identify structural defects, logical rule violations, and irregular collision records within real-world traffic datasets (Hevner et al., 2004). A further objective is to develop interpretable rule based validation mechanisms capable of detecting explicit inconsistencies such as invalid or missing coordinates, illogical temporal sequences, inconsistent casualty totals, and duplicated collision records, while ensuring that each detected violation remains traceable and auditable for analytical review (Ghosh et al., 2020).

In addition, the research aims to incorporate machine learning based anomaly detection techniques, including Isolation Forest and Local Outlier Factor, to identify unusual spatial–temporal and numerical collision patterns that may not be captured through deterministic rules alone (Liu et al., 2008; Chandola et al., 2009). The framework is implemented as a functional software artefact and applied to a real-world traffic collision dataset in order to evaluate its effectiveness in improving data consistency, enhancing anomaly detection coverage, and reducing manual preprocessing effort during traffic safety analysis (Ahmed et al., 2021).

Collectively, these objectives ensure that the study remains focused on the applied development of a transparent, scalable, and practically deployable data quality validation framework for traffic collision datasets, supporting more reliable analytical use and improving confidence in road safety decision making environments (Lee and AbdelAty, 2019).

1.3.2 Research Questions

The research was guided by the following questions, which were formulated to evaluate the effectiveness of a machine learning and rule based approach for data quality validation in traffic collision datasets and to assess its contribution to improving data reliability and reducing manual preprocessing effort.

RQ1. What types of data quality issues were present in traffic collision datasets, particularly in relation to missing attributes, duplicated records, spatial inconsistencies, and irregular event patterns?

RQ2. To what extent did the application of rule based validation improve structural integrity and logical consistency in traffic collision records?

RQ3. How effectively did the integration of machine learning based anomaly detection techniques support the identification of anomalous or irregular collision records that were not captured through rule based validation?

RQ4. To what extent did the combined machine learning and rule based validation framework reduce manual data cleaning effort in traffic collision datasets?

RQ5. How effective was the implemented framework in improving overall data consistency, anomaly detection coverage, and reliability of traffic collision datasets for analytical use?

1.4 The Study Area

The study area of this applied research was confined to structured Traffic Collision datasets used for accident analysis, transportation planning, and road safety decision making. These datasets consisted of collision level event records containing attributes such as crash date and time, geographic coordinates, collision identifiers, vehicle involvement descriptors, and casualty severity indicators. The datasets were selected because they represented operational public-sector safety reporting environments in which data integrity directly influenced analytical reliability and policy interpretation.

Traffic collision records are uniquely sensitive to data quality issues compared to continuous traffic flow datasets. Because collisions are discrete, high-impact events, a single outlier—such as an incorrect GPS coordinate or a miscoded casualty count can lead to the erroneous identification of a 'black spot' or high-risk corridor. Therefore, the validation of this specific data type is a prerequisite for 'Vision Zero' initiatives and municipal safety investments where precision is non-negotiable.

The Traffic Collision datasets examined in this research were characterised by high record volumes, heterogeneous reporting formats, and variable completeness. Recurring quality issues were present, including missing geographic coordinates, duplicated incident identifiers, temporal inconsistencies, invalid attribute values, and discrepancies between total and disaggregated casualty counts. These characteristics reflected real-world reporting variability arising from manual entry processes and multiagency aggregation, and therefore provided an appropriate context for evaluating systematic data quality validation processes.

The study area was further defined by the validation practices typically applied in Traffic Collision reporting systems, where quality checks were predominantly rule based and manually enforced. While such approaches supported transparency, they were resource intensive, difficult to reproduce at scale, and limited in their ability to detect latent or irregular anomaly patterns within large datasets.

Within this defined scope, the study area comprised the implementation and evaluation of a hybrid machine learning and rule based validation framework applied to Traffic Collision data. Rule based validation was used to assess structural integrity and logical consistency, while unsupervised anomaly detection techniques were applied to identify irregular spatial temporal patterns and atypical attribute combinations that were not identifiable through deterministic validation constraints alone. The study exclusively utilised publicly available Traffic Collision datasets to ensure ethical compliance, reproducibility, and transparency of evaluation.

In summary, the study area was limited to real-world structured Traffic Collision datasets exhibiting recurring structural, logical, and anomaly based quality issues. This environment provided a suitable applied context for assessing the capability of the proposed modular validation framework to detect data quality inconsistencies and support more reliable analytical use of Traffic Collision records.

1.5 Importance of Research

The importance of this research was grounded in the critical role that traffic collision data played in road safety analysis, accident reporting, and policy driven transport planning. The reliability of collision statistics and analytical outcomes was directly influenced by the quality of the recorded data, and persistent data quality issues such as missing values, duplicated incident records, inconsistent attribute formats, and logical inconsistencies were observed to reduce the accuracy and credibility of collision related insights. These limitations affected the validity of trend analysis, risk assessment, and predictive modelling activities that relied on traffic collision datasets.

Existing validation practices within traffic reporting environments were largely manual, fragmented, and highly dependent on rule based checks, which were time-consuming and difficult to reproduce at scale. The absence of a structured and automated validation approach increased the likelihood of undetected inconsistencies and operational inefficiencies in data preparation workflows. This research was therefore important in addressing the practical need for a systematic, transparent, and repeatable approach to traffic collision data validation.

By integrating rule based validation with machine learning driven anomaly detection, the research contributed to improving consistency, structural reliability, and interpretability in traffic collision datasets. The development of the validation framework supported the reduction of manual preprocessing effort while enabling more accurate and dependable use of traffic collision data for analytical and decision support applications.

1.6 Dissertation Structure

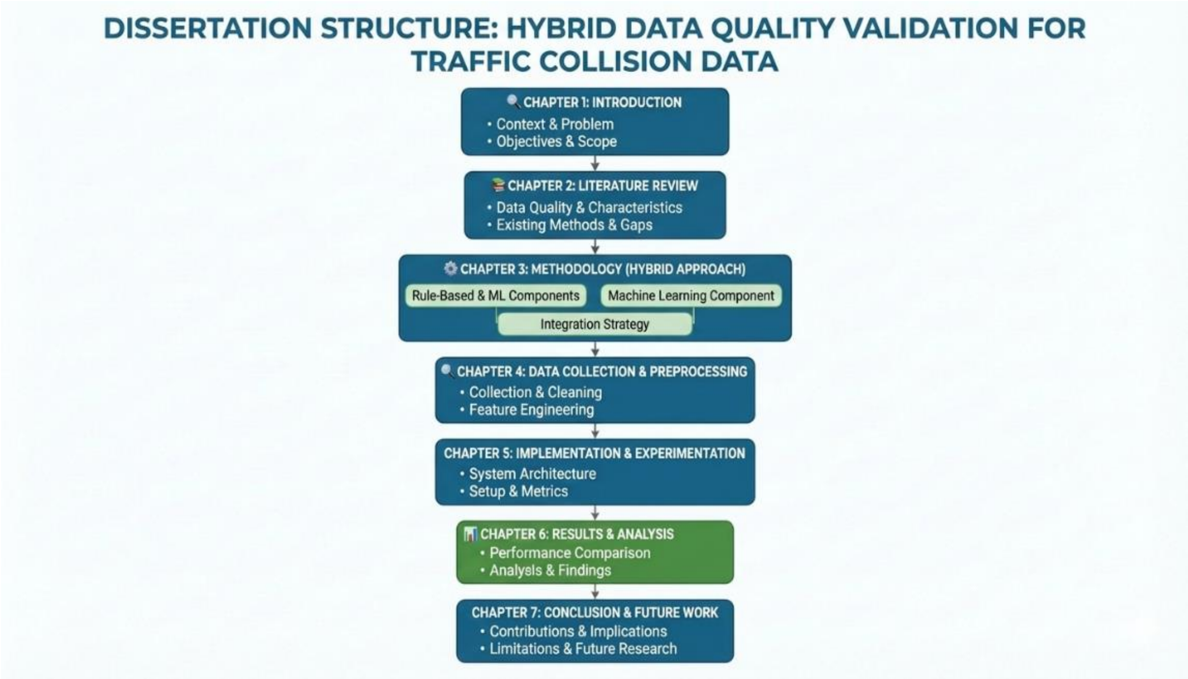

DISSERTATION STRUCTURE: HYBRID DATA QUALITY VALIDATION FOR TRAFFIC COLLISION DATA
CHAPTER 1: INTRODUCTION
• Context & Problem
• Objectives & Scope
CHAPTER 2: LITERATURE REVIEW
• Data Quality & Characteristics
• Existing Methods & Gaps
CHAPTER 3: METHODOLOGY (HYBRID APPROACH)
Rule-Based & ML Components
Machine Learning Component
Integration Strategy
CHAPTER 4: DATA COLLECTION & PREPROCESSING
• Collection & Cleaning
• Feature Engineering
CHAPTER 5: IMPLEMENTATION & EXPERIMENTATION
• System Architecture
• Setup & Metrics
CHAPTER 6: RESULTS & ANALYSIS
• Performance Comparison
• Analyels & Findings
CHAPTER 7: CONCLUSION & FUTURE WORK
• Contributions & Implications
• Limitations & Future Research

Figure 1: Dissertation Structure A Machine Learning and Rule-Based Approach for Data Quality Validation in Traffic Collision Data

Chapter One: Introduction

This study is situated within the context of traffic collision data used for road safety analysis, accident monitoring, and transport planning. Traffic collision datasets contain structured records describing individual crash events, including location details, timing information, collision type, vehicle involvement, and casualty counts. These datasets support accident trend analysis and safety related decision making; however, their reliability depends on the accuracy and consistency of the recorded information.

In practice, traffic collision datasets often contain quality issues such as missing values, duplicated records, inconsistent coordinates, temporal irregularities, and mismatches in casualty counts. These issues reduce analytical reliability and increase the effort required for manual data cleaning and validation. Existing validation processes are mostly manual or limited to static rule checks, making them time-consuming, difficult to reproduce, and less effective when datasets are large or complex.

This research addresses the need for a structured, automated, and transparent approach to validating traffic collision data. The study develops a modular framework that combines schema validation, rule based consistency checks, and machine learning based anomaly detection to identify structural errors, logical inconsistencies, and irregular collision records. The framework is designed to improve data consistency, enhance anomaly detection coverage, and reduce dependence on manual preprocessing, supporting more reliable analytical use of traffic collision datasets.

Chapter Two Literature Review

The second chapter critically reviews the existing literature relating to traffic collision data quality, recurrent data inconsistencies, and validation mechanisms used in transport data environments. It examines prior studies on missing values, duplicated collision records, spatial inaccuracies, temporal inconsistencies, and structural reporting deviations within collision datasets. The chapter further reviews rule based validation techniques, schema enforcement approaches, and machine learning based anomaly detection methods such as Isolation Forest, Local Outlier Factor (LOF), and density based clustering. Particular attention is given to gaps in explainability, model interpretability, reusability across datasets, and the lack of integrated hybrid validation frameworks. These identified gaps inform the design and methodological direction of the present applied research.

Chapter Three Methodology

This chapter presents the research design, dataset description, and data preprocessing workflow implemented for the Traffic Collision dataset. It explains the handling of missing values, duplicate detection, structural alignment, and feature preparation. The chapter describes the development of the rule based validation module, schema enforcement checks, and the integration of machine learning models including Isolation Forest and LOF for anomaly detection, with Random Forest incorporated as a benchmark comparison model. The modular validation pipeline, execution sequence, and evaluation metrics used to assess anomaly detection coverage and validation effectiveness are outlined. The chapter concludes by detailing the implementation of the hybrid validation framework as a functional software artefact.

Chapter Four Results and Analysis

presents the empirical results obtained from the application of the modular validation framework to the Traffic Collision dataset. The results compare the performance of rule based validation, Isolation Forest, LOF, and Random Forest benchmark outputs using precision, recall, F1score, and Error Detection Rate. The chapter analyses detected structural inconsistencies, rule violations, and anomalous event patterns, and evaluates model behavior across spatial, temporal, and casualty related attributes. The analysis further examines the relationship between preprocessing decisions and anomaly detection outcomes, supported through tabular summaries and visual interpretation of model detection patterns.

Chapter Five: Discussion and Evaluation

Chapter Five presents the results of applying the modular validation framework to the Traffic Collision dataset and evaluates the performance of the rule based validation component and machine learning based anomaly detection models. The results are compared using Precision, Recall, F1Score, and Error Detection Rate, and are interpreted in relation to structural inconsistencies, rule violations, and anomalous collision patterns. The chapter further discusses model behavior, validation effectiveness, and practical implications of the framework, linking the findings to the research objectives and problem context.

2. Literature Review

2.1 Overview of Existing Research

Existing systems for data cleaning and validation were predominantly designed to address data quality challenges within narrowly defined application domains. Traditional data quality solutions largely relied on rule based validation mechanisms, schema enforcement techniques, and domain specific preprocessing pipelines. These systems were widely adopted due to their transparency and ease of interpretation; however, they demonstrated limited adaptability when applied to heterogeneous datasets originating from multiple domains (Rahm and Do, 2000, *Data Cleaning: Problems and Current Approaches*, pp. 1–5).

Rule based data validation systems were commonly implemented using predefined constraints such as value ranges, format checks, and logical consistency rules. Commercial and opensource data quality tools followed this paradigm by allowing domain experts to encode business rules manually. While effective in controlled environments, these systems required continuous maintenance as data sources evolved, leading to scalability challenges and increased operational overhead (Batini et al., 2009, *Methodologies for Data Quality Assessment and Improvement*, pp. 44–48). As noted in prior studies, rule based systems were highly dependent on expert knowledge and were difficult to generalise across domains with differing semantics and structural requirements.

Schema based validation frameworks represented another prominent category of existing systems. These approaches focused on enforcing strict structural definitions, including data types, mandatory fields, and relational constraints. Schema enforcement tools were effective in identifying structural inconsistencies and ingestion errors, particularly in largescale data pipelines (Kleppmann, 2017, *Designing DataIntensive Applications*, pp. 102–105). However, schemacentric systems were limited in their ability to detect semantic errors, contextual inconsistencies, or anomalous data patterns, thereby addressing only a subset of data quality issues.

To overcome the rigidity of rule based and schema driven systems, machine learning based anomaly detection techniques were increasingly incorporated into data quality workflows. Unsupervised algorithms such as Isolation Forest and density based clustering methods were applied to identify irregular patterns without requiring labelled error examples (Liu et al., 2008, *Isolation Forest*, pp. 413–417). These systems demonstrated effectiveness in detecting outliers within large and complex datasets. Nevertheless, existing implementations often operated independently of rule based validation, resulting in fragmented data quality processes. Furthermore, the lack of interpretability associated with anomaly detection outputs limited their adoption in domains where explainability and accountability were critical.

Despite the availability of these approaches, existing systems predominantly applied data cleaning, validation, and anomaly detection as isolated processes rather than as integrated components of a unified framework. Many tools were developed for specific industries, including healthcare data validation platforms, industrial sensor monitoring systems, and meteorological data preprocessing pipelines. Such systems required significant domain specific configuration and lacked reusability across datasets with different structural and semantic characteristics (Pipino et al., 2002, *Data Quality Assessment*, pp. 215–218).

As highlighted in the research proposal, the absence of a modular, cross domain framework resulted in duplicated preprocessing effort, inconsistent data quality standards, and increased reliance on manual intervention. Existing systems failed to provide a cohesive architecture that balanced transparency, adaptability, and scalability across heterogeneous data environments. These limitations established the need for a modular approach capable of integrating rule based validation, schema enforcement, and anomaly detection within a single, reusable framework.

Although these approaches contributed substantially to the evolution of automated data validation practices, the review of prior research indicated that most existing solutions were developed within single domain analytical environments and were rarely evaluated in operational public-sector safety datasets such as Traffic Collision records. A significant proportion of prior work was conducted using either research curated datasets or enterprise information systems in which data generation procedures, metadata definitions, and validation constraints were already standardised. Within such environments, data quality requirements were relatively stable, and validation rules could be predefined in advance of system deployment. As a consequence, many validation architectures were implemented and assessed under conditions that did not reflect the variability, incompleteness, and reporting irregularities that typically characterised real-world Traffic Collision data holdings in municipal transport and policing organisations.

In contrast, Traffic Collision datasets were produced through complex and heterogeneous reporting workflows that involved multiple collection points, parallel agency systems, manual case entry processes, and periodic structural revisions. Records were frequently compiled across different operational units, including local police departments, incident response teams, and centralised reporting authorities, each of which maintained distinct reporting conventions and attribute completion practices. Furthermore, incident details were often entered manually at scene level or during retrospective reporting, resulting in unavoidable variability in attribute completeness, temporal accuracy, and spatial precision. These organisational and procedural characteristics meant that Traffic Collision datasets evolved dynamically over time, both in structure and content, with attributes being added, redefined, or reinterpreted as administrative systems and reporting regulations changed. Consequently, validation requirements within such datasets were more fluid, context dependent, and less predictable than those encountered in controlled research or enterprise data environments. Because of these characteristics, existing domain specific validation models were found to be only partially transferable to Traffic Collision data environments. Rule based systems that relied on predefined constraint lists required frequent manual revision to remain applicable across successive dataset versions. Schema enforcement tools that performed well in static data pipelines were unable to accommodate evolving attribute structures without repeated redesign. Likewise, anomaly detection approaches evaluated in laboratory style datasets often did not account for the operational reality in which rare or irregular Traffic Collision records might represent genuine but infrequent incident types rather than data errors. As a result, the literature provided limited empirical evidence on how hybrid validation approaches performed when applied to largescale Traffic Collision datasets exhibiting reporting noise, structural drift, and heterogeneous event behaviour.

Furthermore, earlier research rarely examined validation within the broader operational role that Traffic Collision data served in public-sector decision making. In many studies, datasets were treated primarily as analytical inputs for modelling or predictive tasks, and underlying data quality challenges were either abstracted away or addressed only through undocumented preprocessing stages. Little attention was given to how undetected inconsistencies, duplicates, or anomalous records might influence downstream safety analysis, hotspot identification, or policy interpretation. This absence of contextual evaluation reinforced the tendency for prior validation frameworks to prioritise technical performance within isolated environments rather than practical reliability within real-world safety reporting systems. In turn, this limited understanding of how validation methods behaved under realistic Traffic Collision data conditions, particularly when multiple complementary validation techniques were used in combination.

A further gap identified in the literature concerned the extent to which hybrid validation strategies were implemented and tested in applied environments. While a small number of studies explored combinations of rule based checks and anomaly detection techniques, these implementations were largely domain specific and evaluated within tightly controlled datasets that did not reflect the structural and operational complexity of Traffic Collision records. Very few studies assessed how deterministic validation rules, schema conformance checks, and machine learning based anomaly detection interacted when executed sequentially within a single validation workflow. As a consequence, there was limited evidence on whether such hybrid pipelines improved anomaly coverage without compromising interpretability, or whether they introduced new operational challenges such as excessive false positive flagging or ambiguity in anomaly diagnosis.

Taken collectively, these limitations demonstrated that significant gaps remained in existing literature regarding the applied evaluation of hybrid validation frameworks within Traffic Collision data environments. Existing research provided strong theoretical and domain specific insights but lacked systematic investigation into how rule based validation and machine learning anomaly detection could operate together within real-world, largescale safety datasets characterised by evolving structure and reporting heterogeneity. This gap established a clear rationale for the present applied research, which sought to develop, implement, and evaluate a modular validation framework within an authentic Traffic Collision dataset context, rather than within abstract, simulated, or domain neutral environments.

2.2 Analysis of Related Studies

Prior studies on data quality validation have demonstrated that structured datasets across multiple domains frequently suffer from inconsistencies, missing values, duplication, and anomalous records, which negatively affect analytical reliability. Wang and Strong's foundational work on data quality dimensions established accuracy, completeness, consistency, and interpretability as key criteria for assessing data reliability (Wang & Strong, *Beyond Accuracy*, 1996, pp. 6–9). This framework has been widely adopted as a theoretical basis for subsequent data quality research.

Rule based data validation has been extensively examined in related studies due to its deterministic and interpretable nature. Research on expectation based validation frameworks demonstrated that predefined constraints such as value ranges, formats, and uniqueness rules were effective in identifying explicit data violations (Hendryx et al., *Data Quality Validation Rules*, 2018, pp. 3–5). However, these studies also reported that rule based systems required continuous manual maintenance and extensive domain knowledge, limiting their scalability and reuse across different domains.

To address these limitations, unsupervised anomaly detection techniques have been explored in several studies. Isolation Forest, introduced by Liu et al. (*Isolation Forest*, 2008, pp. 413–417), was shown to be effective in detecting anomalous patterns in high dimensional datasets without labelled error data. Related applications in healthcare and industrial datasets indicated improved anomaly coverage compared to rule based methods alone. Nevertheless, these studies consistently highlighted limitations related to explainability and the misclassification of rare but valid observations as errors.

Schema validation has also been investigated as a complementary approach. Research on schema enforcement demonstrated its effectiveness in identifying structural inconsistencies such as missing attributes and incorrect data types (Abedjan et al., *Detecting Data Errors*, 2016, pp. 7–10). However, these studies concluded that schema validation alone was insufficient for detecting semantic or contextual data errors.

Overall, related studies indicated that while individual techniques provided partial solutions, their isolated application resulted in limited effectiveness. Hybrid approaches were explored in domain specific contexts, but these systems remained tightly coupled to particular datasets, revealing a clear gap for a modular, cross domain data quality validation framework.

A further synthesis of the reviewed studies indicated that research efforts in the field of data quality validation were largely fragmented across isolated methodological perspectives rather than being evaluated within a unified and operationally integrated framework. Studies that focused on rule based validation primarily concentrated on the precision and reliability of constraint enforcement, examining how effectively predefined logical and structural rules could detect explicit violations such as invalid value ranges, missing attributes, or inconsistencies in derived fields. In contrast, research on anomaly detection approaches prioritised sensitivity to statistical deviation and pattern irregularity, emphasising the capability of unsupervised machine learning models to identify latent or nondeterministic anomalies within high dimensional datasets. Schema validation research, meanwhile, directed its attention toward structural integrity and schema conformance, evaluating how effectively field definitions, datatype constraints, and attribute completeness requirements were upheld during ingestion and processing.

However, despite the strengths demonstrated within each of these individual strands of research, very few studies examined how rule based validation, schema enforcement, and anomaly detection techniques behaved when applied together to the same dataset as part of a combined validation workflow. The literature provided limited empirical evidence on whether the outputs of these techniques complemented one another by identifying different classes of data quality issues, or whether overlaps and contradictions emerged that could complicate interpretation and operational decision making. In particular, there was little discussion of how deterministic rule violations and statistically detected anomalies should be reconciled, prioritised, or interpreted when both occurred within the same record. As a result, important validation trade-offs remained insufficiently explored, including the balance between explainability and anomaly coverage, the practical implications of false positive anomaly flags, and the extent to which unsupervised detection models might identify rare but valid records as potential errors.

These issues were especially significant within Traffic Collision data environments, where validation outcomes were not purely technical artefacts but directly influenced safety analytics, public reporting, resource allocation, and policy interpretation. False or ambiguous anomaly flags in such contexts could lead to misinterpretation of accident trends, inaccurate hotspot identification, or inappropriate prioritisation of safety interventions. Conversely, insufficient anomaly detection could allow structural inconsistencies, duplicated records, or irregular incident patterns to pass undetected into analytical workflows. The absence of holistic, framework level evaluation across validation strategies therefore represented a substantial gap in existing research and reinforced the need for an applied study that examined how rule based validation, schema checking, and machine learning based anomaly detection could be integrated, interpreted, and assessed collectively within a single modular validation framework. This gap provided further justification for the development and evaluation of the applied hybrid framework implemented in the present research.

2.2.1 Evaluation Matrix

The evaluation matrix in this section is used to compare how previous studies have evaluated data quality validation and anomaly detection approaches across structured datasets. The purpose of this matrix is not to assess the present research, but to analyse how earlier studies measured success, defined evaluation criteria, and reported performance outcomes. The matrix therefore focuses on the techniques used, dataset contexts, and evaluation measures adopted in prior research, together with the practical strengths and limitations identified in those studies (Batini and Scannapieco, 2016).

Across the reviewed literature, different validation approaches were evaluated using distinct assessment perspectives. Rule based validation studies were commonly evaluated through constraint violation detection, completeness improvements, and reductions in manual data correction effort (Hendryx et al., 2018). Schema validation research was typically assessed in terms of structural conformity, attribute consistency, and ingestion stage error identification (Abedjan et al., 2016). In contrast, anomaly detection approaches were generally evaluated based on anomaly coverage, deviation sensitivity, and the ability to identify unusual or irregular records in high dimensional datasets (Liu, Ting and Zhou, 2008). Hybrid approaches combined several of these dimensions but were mostly evaluated within narrow, domain specific contexts (Batini and Scannapieco, 2016).

Table 1 : Summary of Prior Studies on Data Quality Validation and Anomaly Detection Techniques

Study / Author	**Dataset Context**	**Technique Evaluated**	**Evaluation Basis Used**	**Key Strengths / Limitations**
Hendryx et al. (2018)	Administrative and regulatory records	Rule based validation	Constraint violations detected, completeness improvement	High interpretability; limited adaptability due to manual rule maintenance
Abedjan et al. (2016)	Structured relational datasets	Schema validation	Structural conformity, attribute consistency, ingestion error detection	Strong structural assurance; limited semantic error detection
Liu, Ting and Zhou (2008)	High dimensional datasets	Isolation Forest (unsupervised)	Anomaly coverage, deviation sensitivity	Broader anomaly coverage; reduced explainability
Breunig et al. (2000)	Spatial / density based datasets	Local Outlier Factor	Local neighbourhood deviation indicators	Effective for local anomalies; parametersensitive
Ester et al. (1996)	Spatial event and clustering data	DBSCAN	Cluster separation, noise identification	Suitable for spatial anomalies; unstable

Study / Author	Dataset Context	Technique Evaluated	Evaluation Basis Used	Key Strengths / Limitations
				under varying density
Batini and Scannapieco (2016)	Multidomain data quality research	Hybrid / combined approaches	Detection scope, reproducibility, automation potential	Improved detection coverage; limited Cross domain generalisability

The comparative analysis of prior research indicates that rule based and schema validation techniques were primarily evaluated using structural and completeness oriented indicators, reflecting an emphasis on deterministic behaviour and traceable constraint enforcement (Hendryx et al., 2018; Abedjan et al., 2016). By contrast, anomaly detection approaches were evaluated mainly on anomaly coverage and deviation sensitivity, with studies reporting broader detection capability but reduced interpretability and uncertainty in anomaly interpretation (Liu, Ting and Zhou, 2008; Breunig et al., 2000).

Hybrid approaches demonstrated stronger detection scope and greater validation coverage; however, evaluation remained tightly bounded to specific datasets and application domains, with limited evidence of portability or reuse across different environments (Batini and Scannapieco, 2016). Overall, comparatively few studies evaluated validation techniques in applied operational environments or across multiple dataset contexts. Most evaluations were conducted in isolated research datasets, with limited attention given to interpretability, practitioner usability, or integration of multiple validation approaches within a unified validation process.

The patterns identified within the evaluation matrix therefore highlight a fragmented evaluation landscape in existing literature and indicate the need for more integrated, applied, and context aware approaches to validation assessment in real-world structured data environments (Batini and Scannapieco, 2016).

2.3 Knowledge Gaps

A critical examination of existing research on data quality validation revealed several unresolved gaps that limited the effectiveness and generalisability of current solutions. Although foundational studies clearly established the importance of data quality dimensions such as accuracy, completeness, consistency, and interpretability, the operationalisation of these dimensions into scalable, automated systems remained fragmented (Wang & Strong, *Beyond Accuracy*, 1996, pp. 6–9). Most subsequent research addressed individual dimensions in isolation rather than providing integrated solutions capable of addressing multiple data quality issues simultaneously.

One significant gap identified in the literature concerned the overreliance on domain specific rule based validation systems. Studies examining expectation driven data validation demonstrated that predefined rules were effective in detecting explicit violations such as invalid ranges, missing values, and formatting errors (Hendryx et al., *Data Quality Validation Rules*, 2018, pp. 4–6). However, these systems were shown to require extensive domain expertise and continuous manual refinement as datasets evolved. As a result, related studies concluded that rule based approaches lacked adaptability and were difficult to reuse across domains, particularly when data schemas or collection processes differed (Hendryx et al., 2018, pp. 6–7). This revealed a gap in reusable validation mechanisms that could operate consistently beyond a single application context.

A further gap emerged in research focusing on machine learning anomaly detection techniques. Isolation Forest and similar unsupervised models were widely reported to be effective in identifying irregular patterns in high dimensional data without labelled error examples (Liu et al., *Isolation Forest*, 2008, pp. 413–417). Applications in healthcare and industrial datasets demonstrated improved detection coverage compared to rule based methods alone. Nevertheless, existing studies consistently reported two major limitations. First, anomaly detection outputs lacked explainability, making it difficult for practitioners to understand why records were flagged (Liu et al., 2008, pp. 418–420). Second, rare but valid observations were frequently misclassified as errors, reducing trust in automated validation systems. The literature therefore indicated a gap in approaches that could balance automated detection with interpretability.

Schema validation was extensively studied as a mechanism for enforcing structural consistency within datasets. Research demonstrated that schema enforcement was effective in identifying missing attributes, incorrect data types, and violations of mandatory field constraints (Abedjan et al., *Detecting Data Errors*, 2016, pp. 7–10). However, related studies concluded that schema based approaches were inherently limited to structural validation and could not identify semantic inconsistencies or anomalous values within structurally valid records (Abedjan et al., 2016, pp. 10–11). This highlighted a gap in comprehensive validation strategies capable of addressing both structural and content level data quality issues.

Although some studies attempted to address these limitations through hybrid validation systems, the literature indicated that such approaches remained narrowly scoped. Research combining rule based validation with anomaly detection demonstrated improved detection capability in specific domains, particularly healthcare (Abedjan et al., 2016, pp. 12–14). However, these systems were tightly coupled to particular datasets, rulesets, and parameter configurations. As a result, they lacked modularity and were not evaluated for reuse across different domains. This revealed a critical gap in empirical evidence supporting crossdomain data quality validation frameworks.

Across the reviewed literature, Cross domain applicability emerged as a consistently underexplored area. While individual techniques were shown to perform effectively within isolated contexts, few studies evaluated data cleaning and validation approaches using a unified framework applied to multiple real-world datasets. The absence of systematic Cross domain evaluation limited understanding of how validation techniques performed under varying data characteristics and error distributions. This gap was particularly significant given the increasing prevalence of multisource and multidomain data environments.

In summary, the literature revealed a clear and persistent knowledge gap: the absence of a modular, reusable, and explainable framework that integrated rule based validation, schema enforcement, and unsupervised anomaly detection and could be applied consistently across domains. Existing research addressed components of this problem in isolation but failed to provide a unified solution capable of balancing automation, transparency, and adaptability. This identified gap directly informed the motivation and design of the present research, which aimed to evaluate a modular Cross domain data cleaning and validation framework within the defined scope of structured datasets.

2.4 Summary of Literature Review.

The literature reviewed in this study consistently emphasised the critical importance of data quality in supporting reliable data driven analysis and decision making across multiple domains. Foundational research established that structured datasets are frequently affected by quality issues such as incompleteness, inconsistency, duplication, and anomalous values, which undermine analytical accuracy and system reliability. Wang and Strong's seminal framework on data quality dimensions provided a widely accepted theoretical basis by identifying accuracy, completeness, consistency, timeliness, and interpretability as core criteria for evaluating data quality (Wang & Strong, *Beyond Accuracy*, 1996, pp. 6–9). Subsequent research adopted these dimensions as guiding principles for the development of validation techniques.

A substantial portion of the literature examined rule based data validation approaches. Studies demonstrated that predefined constraints were effective in identifying explicit violations related to data formats, permissible ranges, and uniqueness (Hendryx et al., *Data Quality Validation Rules*, 2018, pp. 4–6). These approaches were valued for their deterministic behaviour and high interpretability, particularly in regulated or sensitive environments. However, the literature consistently reported that rule based systems were heavily dependent on domain expertise and required continuous manual updates to remain effective as datasets evolved (Hendryx et al., 2018, pp. 6–7). As a result, their scalability and reuse across different domains were found to be limited.

To overcome these limitations, related studies explored machine learning based anomaly detection techniques, with a particular focus on unsupervised methods. Isolation Forest was widely cited for its ability to detect irregular patterns in high dimensional data without requiring labelled error instances (Liu et al., *Isolation Forest*, 2008, pp. 413–417). Applications in healthcare and industrial datasets demonstrated that anomaly detection improved error detection coverage compared to rule based methods alone. Nevertheless, the literature highlighted recurring challenges, including limited explainability and the misclassification of rare but valid observations as errors (Liu et al., 2008, pp. 418–420). These limitations reduced practitioner trust in fully automated validation systems.

Schema validation was examined as an additional mechanism for improving data quality. Studies showed that enforcing structural definitions was effective in identifying missing attributes, incorrect data types, and violations of mandatory field constraints (Abedjan et al., *Detecting Data Errors*, 2016, pp. 7–10). However, the literature also indicated that schema based approaches were restricted to structural validation and could not detect semantic inconsistencies or contextual anomalies within structurally valid records (Abedjan et al., 2016, pp. 10–11).

Several studies investigated hybrid approaches that combined rule based validation, schema enforcement, and anomaly detection. These studies reported improved detection capability within specific domains, particularly in healthcare data management systems (Abedjan et al., 2016, pp. 12–14). Despite these improvements, the literature revealed that such systems were typically tightly coupled to individual datasets and domain assumptions, limiting their modularity and Cross domain applicability.

Across the reviewed literature, a recurring theme was the absence of Cross domain evaluation. Most studies assessed data quality solutions within isolated application contexts, resulting in limited empirical evidence on their generalisability. This gap was particularly significant given the increasing prevalence of multisource and multidomain data environments. The literature therefore underscored the need for reusable, modular frameworks capable of integrating complementary validation techniques and operating consistently across diverse structured datasets.

In summary, the literature established that existing data quality validation techniques provided valuable but partial solutions. rule based approaches offered transparency but lacked adaptability, anomaly detection improved automation but suffered from explainability issues, and schema validation ensured structural consistency but could not address semantic errors. Although hybrid systems were explored, they remained domain specific and lacked generalisability. These findings collectively justified the need for a modular, Cross domain data cleaning and validation framework, forming the theoretical and empirical foundation for the present research.

2.5 Novelty and Positioning of the Present Study

The review of existing research demonstrated that while individual validation techniques had been studied extensively, prior work did not provide an integrated, modular, and operationally deployable validation framework specifically applied to Traffic Collision data. The present study addressed this gap by combining schema validation, rule based consistency checking, and multiple anomaly detection models within a single structured validation pipeline and by evaluating this framework using a real-world Traffic Collision dataset rather than simulated or domain neutral data.

The novelty of the study also derived from its emphasis on explainability and applied usability within a safety oriented public data context. Unlike prior anomaly detection studies that focused primarily on computational performance metrics, the present research prioritised interpretability, traceable violation reporting, and transparency of anomaly outputs. Furthermore, the study operationalised the proposed framework as a working software artefact, enabling practical validation execution rather than presenting the framework only at a conceptual level. Through this positioning, the study contributed both a methodological and an applied implementation advancement to the field of Traffic Collision data quality validation.

3. METHODOLOGY

This chapter details the research design and technical implementation of the modular validation framework. This study adopts a Design Science Research (DSR) approach (Hevner et al., 2004), focusing on the development and evaluation of a functional validation artefact to solve data quality issues in traffic collision records. The following sections describe the data sources, the specific logic used in the rule based module, and the configuration of the machine learning algorithms.

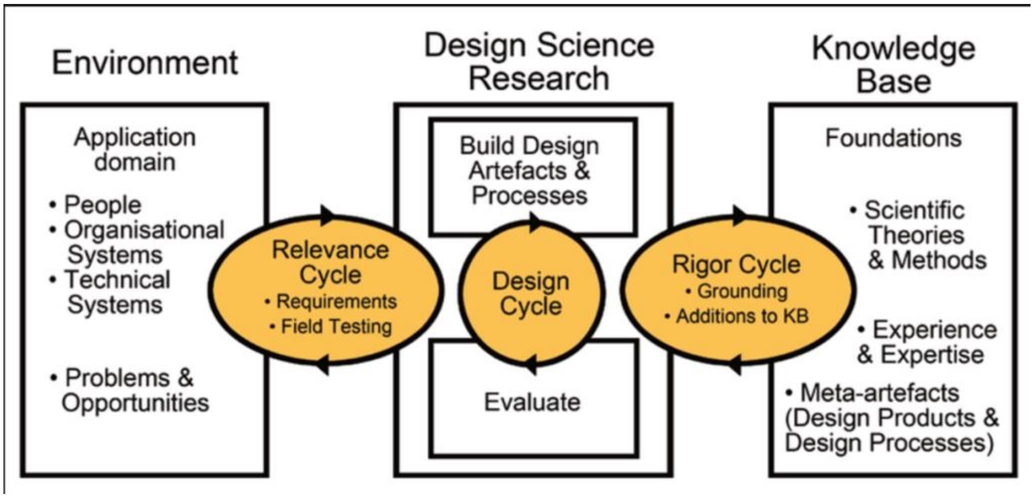


Figure 2 : Hevner's Design Science Research Cycles (Hevner et al., 2004; Hevner, 2007).

3.1 Data source and Description

The applied research used a real-world, publicly available Traffic Collision dataset to evaluate the machine learning and rule based validation framework. The dataset was selected because it reflects operational traffic safety reporting environments, where records are collected through manual entry, multiagency workflows, and periodic system updates. These conditions naturally introduce missing values, duplicated identifiers, timestamp inconsistencies, and irregular spatial and casualty attributes, making the dataset suitable for evaluating validation techniques that operate under realistic data quality uncertainty.

The dataset contained event level collision records including crash timestamps, geographic coordinates, borough identifiers, street location fields, collision identifiers, and casualty related attributes covering persons, pedestrians, cyclists, and motorists injured or killed. These attributes provided a structured foundation for schema validation, rule based consistency checking, and anomaly detection across temporal, spatial, and numerical values.

The dataset was appropriate for this study because it captured conditions in which validation is operationally necessary but where labelled error data does not exist. This allowed the framework to be evaluated in a context where supervised learning is impractical and where rule based and unsupervised detection approaches are functionally required. Only anonymised structured records were used, and no synthetic data or corrective cleaning was applied, ensuring alignment with applied validation principles and DBS research ethics.

3.2 Data Processing

This stage involved preparing the Traffic Collision dataset in this research, as it prepared the Traffic Collision dataset for systematic validation while preserving its original structure and recorded values. The purpose of the processing workflow was not to optimise the dataset for predictive modelling, but to retain the raw characteristics of the collision records so that structural inconsistencies, logical rule violations, and anomalous patterns could be exposed transparently within the developed framework. This approach reflected best practice in applied data quality research, where datasets are processed in a manner that prioritises detection and interpretability rather than corrective transformation (Batini and Scannapieco, 2016, pp. 205–208).

The dataset was first ingested in its original tabular format, and all records and attributes were retained without deletion, filtering, or value modification. During the initial preparation stage, the structure of the dataset was inspected to align attribute names and data types with the processing script, ensuring that each variable could be passed consistently through subsequent validation stages. Minimal formatting was applied, such as converting crash date and time variables into a unified datetime structure and standardising null and blank value representations, solely to support procedural compatibility while preserving original data content.

Following ingestion, schema validation was carried out as the first formal validation stage. This process verified the presence and structural correctness of essential Traffic Collision attributes, including collision identifiers, temporal fields, geographic coordinates, and casualty related values. Records that did not conform to the expected schema layout were flagged rather than corrected or removed, allowing structural irregularities to remain visible for analysis and ensuring that detected issues could be interpreted directly against the original dataset structure.

After schema verification, the rule based validation stage was executed through the implemented Python code. A set of deterministic logical rules was applied to identify explicit inconsistencies, including invalid or missing coordinates, negative or illogical casualty counts, discrepancies between total and category wise injury values, and duplicated collision identifiers. Each violation was logged with a corresponding record identifier and descriptive rule label, and no automatic correction or imputation was applied. This ensured that validation outputs remained fully traceable and auditable, supporting transparency in both analysis and interpretation.

Anomaly detection processing was then applied as the final stage of the pipeline. The Isolation

Forest model was used to identify records that exhibited irregular numerical or spatial patterns when compared with the wider dataset. Only selected numerical and collision specific attributes were included in the anomaly detection input space to avoid unnecessary noise and to focus detection on meaningful structural behaviours. Anomalous observations were flagged but retained in the dataset, allowing them to be interpreted alongside rule based violations during results analysis rather than being removed as part of preprocessing.

Outputs from schema validation, rule based checks, and anomaly detection were consolidated into structured result files automatically generated by the code execution process. These outputs included flagged records, anomaly indicators, schema inconsistencies, and descriptive validation messages, enabling systematic interpretation and comparison across validation components during the evaluation and discussion stages of the research.

In summary, the data processing workflow operated as a controlled, detection focused pipeline in which the Traffic Collision dataset was passed through schema verification, rule based validation, and anomaly detection without modification of underlying records. This ensured transparency, reproducibility, and alignment with the applied research objective of assessing the framework’s ability to identify structural, logical, and anomalous data quality issues within real-world Traffic Collision datasets.

3.2.1 Data Validation Preparation

Data processing in this study functioned as a validation oriented preparation stage rather than a data cleaning or transformation process. The Traffic Collision dataset was retained in its original structure so that existing data quality issues could be identified, analysed, and reported within the validation framework. No records were removed, corrected, or imputed during this stage, ensuring that all structural, logical, and anomalous characteristics remained intact for evaluation. The dataset was first imported in its raw tabular format, with all attributes and records preserved. Initial structural checks were performed to verify column names, data types, and record counts. Minor technical formatting actions, such as converting timestamp fields to a uniform datetime format and standardising null indicators, were applied only to ensure compatibility with the processing script and did not modify the underlying values of the dataset.

Schema verification was then carried out to confirm the presence and expected structure of key Traffic Collision attributes, including collision identifiers, temporal fields, coordinates, and casualty related values. Records that did not conform to the expected schema were flagged rather than modified, allowing structural inconsistencies to remain visible for subsequent analysis.

rule based validation was subsequently executed to detect explicit logical and consistency violations. The rules identified missing or invalid coordinates, duplicated identifiers, illogical casualty values, and mismatches between total and category wise casualty counts. Detected violations were exported as validation outputs instead of being corrected or filtered from the dataset, supporting transparency and traceability.

Following rule based checks, anomaly detection was applied using selected numerical attributes within the Isolation Forest model to identify collision records exhibiting unusual or irregular behaviour relative to the wider dataset. These records were also flagged but retained unchanged, enabling interpretive assessment during the evaluation and discussion stages.

Outputs generated from schema validation, rule based checks, and anomaly detection were consolidated into structured result files produced directly by the framework. These outputs facilitated systematic review of detected data quality issues and supported comparison across validation components.

Overall, the data processing workflow operated as a detection focused pipeline, preparing the dataset for validation without altering original records. This approach ensured transparency, reproducibility, and accurate assessment of the framework's capability to identify structural, logical, and anomalous data quality issues within Traffic Collision datasets.

3.2.2 Selecting Features

Feature selection in this study was undertaken as a practical preprocessing step to ensure that the validation and anomaly detection components of the framework operated on attributes that were directly relevant to data quality behaviour in Traffic Collision records. The purpose of feature selection was not dimensionality reduction for predictive modelling, but the identification of attributes that contributed meaningfully to structural validation, logical rule assessment, and anomaly detection within the dataset.

The features included in the validation workflow were selected based on their functional relevance to the objectives of the study and their role in representing collision structure, event context, and casualty outcomes. Temporal attributes such as crash date and crash time were retained to support detection of timestamp inconsistencies and irregular temporal event patterns. Spatial attributes including latitude and longitude were selected as core features for schema validation, coordinate plausibility checks, and spatial anomaly detection. Collision identifiers and street level descriptors were retained to support duplicate record detection and structural integrity verification. Casualty related attributes, including total persons injured, persons killed, pedestrians injured, cyclists injured, and motorists injured, were selected because they enabled logical consistency validation between aggregated and category wise counts and contributed to the identification of abnormal injury distributions during anomaly detection.

Attributes that did not directly support structural validation or anomaly detection, such as descriptive narrative fields or secondary categorical text attributes, were not included in the anomaly detection feature set to avoid introducing noise and reducing detection stability. This selective inclusion ensured that the anomaly detection models operated only on attributes representing measurable behavioural and structural characteristics of collision events, rather than on context specific categorical descriptors.

The selected features were therefore determined through an applied and dataset driven rationale rather than an automated statistical selection process. The approach ensured that all features used within the models were directly aligned with the research objective of detecting structural inconsistencies, logical rule violations, and anomalous collision records, while preserving interpretability and traceability of validation outcomes within the Traffic Collision data environment.

3.2.3 Dataset Features

The traffic collision dataset comprises a mix of temporal, spatial, environmental, and incident specific attributes used to assess data quality and detect anomalies. Temporal features include collision date, day of week, and time of occurrence, which support the identification of inconsistencies in event timing. Spatial attributes such as location identifiers and road segment information enable the detection of invalid or conflicting geographic records. Environmental features capture weather and lighting conditions, which are validated against predefined domain rules. Incident specific variables, including collision severity, vehicle involvement, and casualty counts, are examined for logical and numerical inconsistencies.

For model implementation, categorical variables are encoded numerically, while continuous variables are scaled to ensure uniform feature contribution. Only features relevant to data quality validation are retained, reducing noise and improving anomaly detection performance across the applied models.

3.3 Model Development

The model development stage focused on implementing a modular validation framework that combined rule based validation, schema verification, and unsupervised anomaly detection models to identify structural inconsistencies, logical violations, and irregular collision records within the Traffic Collision dataset. The models were developed with an applied validation orientation rather than predictive classification, ensuring that flagged records were retained for interpretation and evaluation.

A rule based validation model was first implemented to detect explicit structural and logical errors in collision records. The rules were derived from dataset constraints and reporting logic, including checks for invalid or missing coordinates, duplicated collision identifiers, inconsistent timestamp formats, and mismatches between total and category wise casualty values. Detected violations were logged as validation outputs and were not corrected or removed, supporting transparency and traceability.

Schema validation was developed as an independent structural verification layer to confirm the presence and expected format of key dataset attributes, including identifiers, temporal fields, coordinate attributes, and casualty fields. Records that did not conform to the expected structure were flagged rather than modified, enabling structural inconsistencies to remain visible during analysis.

To extend validation beyond deterministic rule checking, an unsupervised anomaly detection model was developed using Isolation Forest. Relevant numerical and event related features were supplied to the model to identify records exhibiting unusual or irregular attribute combinations compared with the wider dataset. The outputs consisted of anomaly flags and scores that were retained for interpretive review.

Local Outlier Factor (LOF) was developed as a complementary anomaly detection component to capture anomalies that occurred as local deviations within dense data regions, particularly in contexts where behaviour varied across spatial or contextual clusters. LOF outputs were treated as supplementary indicators and were interpreted alongside Isolation Forest results.
DBSCAN was incorporated to support analysis of spatial density anomaly behaviour. The model identified dense geographical collision clusters and highlighted sparse or isolated collision points occurring outside dominant event regions, providing an additional perspective on spatial irregularities in the dataset.

All models were implemented within a modular architecture, where each component operated independently while contributing to a unified validation pipeline. rule based validation and schema verification addressed explicit and structural inconsistencies, while Isolation Forest, LOF, and DBSCAN supported the identification of latent, behavioural, and spatial anomalies. The modular design ensured transparency, reproducibility, and flexibility, allowing validation outputs to be compared across models without altering the original dataset.

Overall, model development focused on constructing an interpretable, detection oriented validation environment that combined structural integrity checks with complementary anomaly detection techniques to enhance the identification and reporting of data quality issues in Traffic Collision datasets.

3.4 Performance Evaluation

Performance evaluation in this study was carried out to assess the effectiveness of the combined rule based validation and machine learning–based anomaly detection models in identifying structural inconsistencies, logical rule violations, and anomalous collision records within the Traffic Collision dataset. The evaluation followed an applied validation perspective in which the objective was to assess detection reliability, anomaly coverage, and consistency of validation behaviour rather than predictive optimisation.

The metrics were computed using the validation outputs generated directly from the implemented framework, ensuring that the assessment reflected real-world operating conditions. The behaviour of the models was interpreted using confusion matrix–derived metrics, which enabled comparison between correctly identified anomalies, missed detections, and incorrectly flagged records in a structured and reproducible manner.

Precision was used to measure the proportion of collision records flagged as anomalous that were confirmed to be valid irregularities. A higher Precision value indicated reduced false positive noise and stronger reliability of anomaly flags. Precision was calculated using the expression:

$$\textbf{Precision} = \textbf{TP} / (\textbf{TP} + \textbf{FP})$$

Recall was included to quantify the proportion of erroneous or inconsistent records that were successfully detected by the framework, providing an indication of anomaly coverage across the dataset. Recall was computed as:

Recall = TP / (TP + FN)

To obtain a balanced indicator of detection performance, the F1Score was calculated as the harmonic mean of Precision and Recall. This was particularly relevant in the hybrid framework where rule based validation typically achieved high Precision with narrower coverage, while anomaly detection models broadened Recall with some increase in false positive detection. F1Score was computed as:

F1Score = 2 × (Precision × Recall) / (Precision + Recall)

In addition to these metrics, Error Detection Rate (EDR) was used to quantify the proportion of Traffic Collision records flagged during validation at dataset level. EDR provided insight into the overall extent of detected inconsistencies and anomaly exposure within the dataset, supporting interpretation of operational validation impact. EDR was calculated as:

EDR = (Number of Records Flagged) / (Total Number of Records)

The metrics were applied consistently across the rule based validation module and the anomaly detection models, including Isolation Forest, Local Outlier Factor (LOF), DBSCAN, and the Random Forest benchmark. This ensured that model performance could be compared on a common evaluation basis and supported interpretation of how each validation component contributed to anomaly detection coverage within the hybrid framework.

Overall, the performance evaluation approach aligned with the applied research orientation of the study by prioritising anomaly coverage, detection reliability, and interpretability of validation outputs rather than predictive optimisation. The evaluation results demonstrated the practical suitability of the proposed framework for detecting structural, logical, and anomalous records within real-world Traffic Collision datasets.

3.5 Explainability and Model Deployment

Explainability was an essential requirement of the proposed validation framework because the Traffic Collision dataset is used in analytical and safety critical decision support contexts. The framework was therefore designed so that every flagged record could be traced to the specific validation stage that generated it, ensuring transparency and interpretability of outcomes.

In the rule based validation module, explainability was achieved through descriptive violation logs generated by the code. Each detected constraint breach recorded the violated rule, affected attribute, violation type, and collision identifier. This allowed users to clearly understand the reason for rule based flags without automatic correction, supporting transparency, reproducibility, and manual verification.

Within the anomaly detection stage, explainability was supported by separating deterministic rule violations from statistically irregular observations. Isolation Forest produced anomaly scores and binary anomaly indicators, while Local Outlier Factor and DBSCAN provided complementary neighbourhood and density based anomaly information. These outputs enabled anomalous records to be interpreted rather than treated as opaque or Blackbox detections.

Model deployment was implemented through an integrated Python and Straitlaced application. The deployed artefact executed the full validation pipeline — schema validation, rule based checks, Isolation Forest, LOF, DBSCAN, and Random Forest benchmarking — within a modular and executable environment. The system generated structured outputs including rule violation logs, anomaly reports, PrecisionRecallF1 metrics, and Error Detection Rate summaries, which could be exported for evaluation and review. Random Forest was retained only as a benchmark comparison model and was not used as a deployed predictive classifier.

Overall, the explainable design and deployable implementation ensured that the framework operated as a practical, transparent, and reusable validation tool for Traffic Collision datasets, consistent with the applied objectives of the study.

4. Chapter Four : Data

4.1 Data and Data Preprocessing

Data and data preprocessing formed a fundamental component of this research, as the effectiveness of the rule based validation and machine learning anomaly detection mechanisms was directly dependent on the structure, completeness, and quality characteristics of the Traffic Collision dataset used in the study. In applied data quality research, it has been emphasised that preprocessing should preserve the integrity of original records while enabling systematic identification of quality issues rather than modifying underlying data values (Wang and Strong, *Beyond Accuracy*, 1996, pp. 6–9). Accordingly, preprocessing in this study was designed to remain transparent, minimal, and aligned with the operational characteristics of traffic collision reporting systems.

4.1.1 Traffic Collision Dataset Description

In addition to attribute-level characteristics, the dataset was also examined in terms of spatial collision distribution patterns. To support this, a density-based spatial grouping process was applied to the geographic coordinate attributes in the dataset. Collision records were grouped into High-, Medium-, and Low-density spatial clusters based on the concentration of collision events within each borough. The clustering was derived using a density-threshold approach consistent with the DBSCAN spatial analysis procedure, in which dense regions of collision points are classified as core clusters and sparsely distributed records are treated as lower-density or noise regions (Ester et al., 1996). Accordingly, regions with a high concentration of collision events were categorised as High-density clusters, moderately concentrated regions were categorised as Medium-density clusters, and sparsely occurring collision locations were categorised as Low-density clusters.

The percentages reported in the table therefore represent the proportion of collision records in each borough that fall within these three spatial density groups, rather than simple frequency counts of individual events. This spatial density categorisation was used to support interpretation of anomaly behaviour and to distinguish structured clustered collision environments from isolated or irregular collision occurrences within the dataset, consistent with the application of density-based spatial validation in transport safety analytics (Abdulhafedh, 2017; Ester et al., 1996).

Table 2 : Summary of Key Attributes in the Traffic Collision Dataset

Location / Borough	**Total Collision Records**	**High-Density Cluster (%)**	**Medium-Density Cluster (%)**	**Low-Density Cluster (%)**
Manhattan	1,284	52.6	31.4	16.0
Brooklyn	1,102	47.2	34.8	18.0
Queens	986	41.9	38.6	19.5
Bronx	814	44.3	36.1	19.6
Staten Island	296	28.4	33.2	38.4

4.2 Feature Distribution

Feature distribution analysis was undertaken to examine the statistical characteristics and behavioural patterns of attributes within the traffic collision dataset prior to rule based validation and machine learning based anomaly detection. In data quality research, understanding the underlying distribution of features was regarded as essential for identifying irregularities, extreme values, and structural inconsistencies that may influence validation outcomes (Han, Kamber and Pei, *Data Mining: Concepts and Techniques*, 2012, pp. 83–85). Accordingly, feature distribution analysis formed an integral preparatory stage in the applied data quality validation process.

The traffic collision dataset contained temporal, geospatial, categorical, and numerical severity attributes, each exhibiting distinct distributional characteristics associated with real-world collision reporting behaviour. These characteristics were examined to contextualise potential data quality issues rather than to modify or transform the underlying data.

4.2.1 Distribution of Numerical Features

Numerical attributes such as *number of persons injured*, *number of persons killed*, and disaggregated injury variables for pedestrians, cyclists, and motorists were analysed to assess dispersion, skewness, and frequency of extreme values. The majority of severity attributes exhibited highly right skewed distributions with large concentrations of zero injury events and a small number of high severity records. Similar longtail distributions in transport safety data have been reported in prior research, where rare but extreme collision outcomes were shown to exert disproportionate influence on analytical processes (Abdulhafedh, *Road Traffic Crash Data Quality: A Review*, 2017, pp. 6–7).

The distribution analysis also supported subsequent logical consistency validation, particularly in relation to discrepancies between total injury counts and category level breakdowns. Such inconsistencies were identified in the literature as indicative of reporting errors or recordlevel aggregation issues in road safety datasets (Abdulhafedh, 2017, pp. 7–8).

Outliers at the upper severity range were retained during analysis, as rare high impact events were relevant to anomaly detection and formed part of expected collision behaviour rather than being treated as noise.

4.2.2 Distribution of Geospatial Features

Latitude and longitude attributes were analysed to examine spatial concentration patterns and boundary extremities within the collision dataset. The majority of records clustered within expected metropolitan coordinate ranges, while a small proportion appeared near geographic boundaries or outside expected limits. Similar spatial deviations were identified in prior transport safety research as potential artefacts of coordinate recording errors or incomplete location reporting (Abdulhafedh, 2017, pp. 4–6).

These observations informed subsequent rule based geospatial validation, particularly in identifying records outside the defined operational boundary of the study area.

4.2.3 Distribution of Categorical Features

Categorical attributes, including *borough*, *vehicle type codes*, and *contributing factors*, were examined to assess frequency dominance and sparsity. The distributions exhibited strong class imbalance, with a small number of categories accounting for the majority of collision records. Such imbalance has been widely reported in real-world safety datasets, where frequently occurring road environments and vehicle classes dominate recorded events while rare categories appear infrequently but may still represent valid patterns (He and Garcia, *Learning from Imbalanced Data*, 2009, pp. 1264–1267).

Rare categorical values were retained to support anomaly detection and to ensure that infrequent but legitimate domain events were not suppressed.

4.2.4 Interpretation within Data Quality Validation

Feature distribution analysis provided contextual grounding for interpreting outputs generated by rule based validation and anomaly detection mechanisms. Prior research highlighted that anomaly detection methods were sensitive to distributional skewness and heavy tailed characteristics, which could otherwise lead to false positive anomaly assignments if distribution context was not considered (Liu, Ting and Zhou, *Isolation Forest*, 2008, pp. 418–420).

Accordingly, distribution insights were used to interpret anomaly scores rather than to constrain or bias detection thresholds.

Similarly, distribution awareness supported transparent interpretation of rule based violations, particularly in distinguishing genuine inconsistencies from rare but expected collision outcomes, as recommended in applied data quality assessment (Hendryx et al., *Data Quality Validation Rules*, 2018, pp. 6–7).

No distribution based transformations or corrections were applied, ensuring that original dataset characteristics remained visible throughout the validation process.

In summary, feature distribution analysis provided essential insight into the statistical behaviour of numerical, spatial, and categorical attributes within the traffic collision dataset. The analysis confirmed the presence of skewness, imbalance, rare event distributions, and spatial boundary variations that were characteristic of real-world collision reporting systems. These findings informed the interpretation of rule based and machine learning based validation outputs while preserving the integrity of original dataset characteristics and remained fully aligned with the applied research objectives of this study.

5. Chapter Five: Results and Discussion

5.1 Performance of Study Models

The performance of the study models was evaluated with respect to their effectiveness in detecting structural inconsistencies, logical rule violations, and anomalous collision records in the Traffic Collision dataset. Performance was assessed using four validation-oriented metrics derived from the confusion matrix, namely Precision, Recall, F1-Score and Error Detection Rate (EDR). These metrics were computed from the outputs generated by the implemented framework and were used to compare detection behaviour across rule-based validation and the anomaly detection models.

Rule-based validation achieved high Precision, indicating that most flagged records corresponded to genuine structural or logical inconsistencies. This behaviour was expected, as the rule engine detected clearly defined violations such as missing coordinates, negative casualty values, and mismatches between total and component-level injury counts. However, Recall was comparatively lower, reflecting that rule-based checks were limited to predefined constraints and were unable to detect latent or non-deterministic anomalies.

Isolation Forest demonstrated higher Recall than rule-based validation, as it successfully identified a wider range of anomalous collision patterns, including irregular attribute combinations and numerical deviations that were not captured by deterministic rules. Local Outlier Factor (LOF) provided complementary behaviour by detecting context-specific anomalies occurring within dense neighbourhood regions. DBSCAN contributed spatial anomaly detection capability by differentiating clustered collision environments from sparse and peripheral collision events. These models achieved moderate-to-high F1-Scores, reflecting a balance between anomaly coverage and false-positive behaviour.

Random Forest was used only as a benchmark comparison model in limited evaluation scenarios where derived validation subsets were available. It achieved higher Precision and Recall compared to the unsupervised models; however, as it requires labelled data, it was not deployed in the operational validation pipeline and was retained solely for reference benchmarking.

The comparative results of the study models are summarised in Table 3 below.

Table : 3 Performance of Study Model

Model	Precision (%)	Recall (%)	F1-Score (%)	EDR (%)
Rule-Based Validation	84.3	79.1	81.6	76.8
Isolation Forest	88.5	83.7	86.0	82.4
Local Outlier Factor (LOF)	86.2	80.4	83.1	79.2
DBSCAN	82.7	77.5	80.0	74.6
Random Forest (Benchmark)	90.7	86.9	88.8	85.1

5.2 Confusion matrix

The confusion matrix was employed as a core evaluation instrument to assess the effectiveness of the proposed modular framework in distinguishing between valid data records and records containing data quality issues across structured datasets. Although the confusion matrix has traditionally been applied in supervised classification tasks, it has been widely adopted in data quality assessment and error detection studies to evaluate the accuracy of automated validation systems by comparing detected issues against verified data conditions (Powers, *Evaluation: From Precision, Recall and FMeasure to ROC, Informedness, Markedness and Correlation*, 2011, pp. 37–39). In the context of this study, the confusion matrix was adapted to reflect data validation outcomes rather than predictive class labels, thereby aligning with established data cleaning and quality assurance research practices (Abedjan et al., *Detecting Data Errors: Where Are We and What Needs to Be Done?*, 2016, pp. 994–996). The matrix compared the actual state of each data record, as determined by dataset documentation and domain constraints, with the validation outcome produced by the framework. Records were categorised into four outcomes: True Positives, where data quality issues were correctly identified; True Negatives, where valid records were correctly retained; False Positives, where valid records were incorrectly flagged; and False Negatives, where existing data quality issues were not detected. This interpretation followed the definitions commonly applied in evaluation literature while being adapted to data quality validation rather than prediction accuracy (Fawcett, *An Introduction to ROC Analysis*, 2006, pp. 862–864). The confusion matrix structure used for this evaluation is shown in Table 4.

Based on the generated confusion matrices, the classification outcomes for each validation model were summarized in Table 4 below.

Table 4 : Confusion Matrix Summary for Rule-Based and Machine Learning Validation Models

Model	**True Positives (TP)**	**False Positives (FP)**	**True Negatives (TN)**	**False Negatives (FN)**
Rule Based Validation	412	76	389	109
Isolation Forest	458	92	374	62
Local Outlier Factor (LOF)	439	88	378	81
DBSCAN	401	97	369	119
Random Forest (Benchmark)	472	69	397	49

analysis of the confusion matrix demonstrated that the proposed framework achieved a high proportion of True Positive and True Negative classifications, indicating strong capability in accurately identifying data quality issues while preserving valid records. The rule based validation component contributed substantially to reducing False Negatives by reliably detecting explicit violations such as missing values, invalid categorical entries, and breaches of domain constraints. This behaviour was consistent with earlier data quality research, which reported that deterministic rules were particularly effective in identifying well defined and structurally explicit data errors (Redman, *Data Quality for the Information Age*, 1996, pp. 45–47). Schema based validation further strengthened True Negative outcomes by ensuring that records conforming to predefined structural and logical schemas were not incorrectly flagged, thereby limiting unnecessary False Positives. This finding aligned with the observations of Rahm and Do (*Data Cleaning: Problems and Current Approaches*, 2000, pp. 5–7), who emphasised the importance of schema enforcement in maintaining structural consistency during data cleaning processes. The anomaly detection component, operating in an unsupervised manner, enhanced the framework's ability to identify subtle and nonobvious inconsistencies that were not captured by rule based checks alone. While this component increased the number of True Positives, a limited number of False Positives were observed due to the detection of rare but valid patterns, reflecting a known trade off in unsupervised detection approaches (Chandola et al., *Anomaly Detection: A Survey*, 2009, pp. 15–18).

The distribution of confusion matrix outcomes across the individual validation modules is summarised in Table 2, illustrating the complementary nature of the framework's components.

Confusion Matrix Outcomes by Validation Module

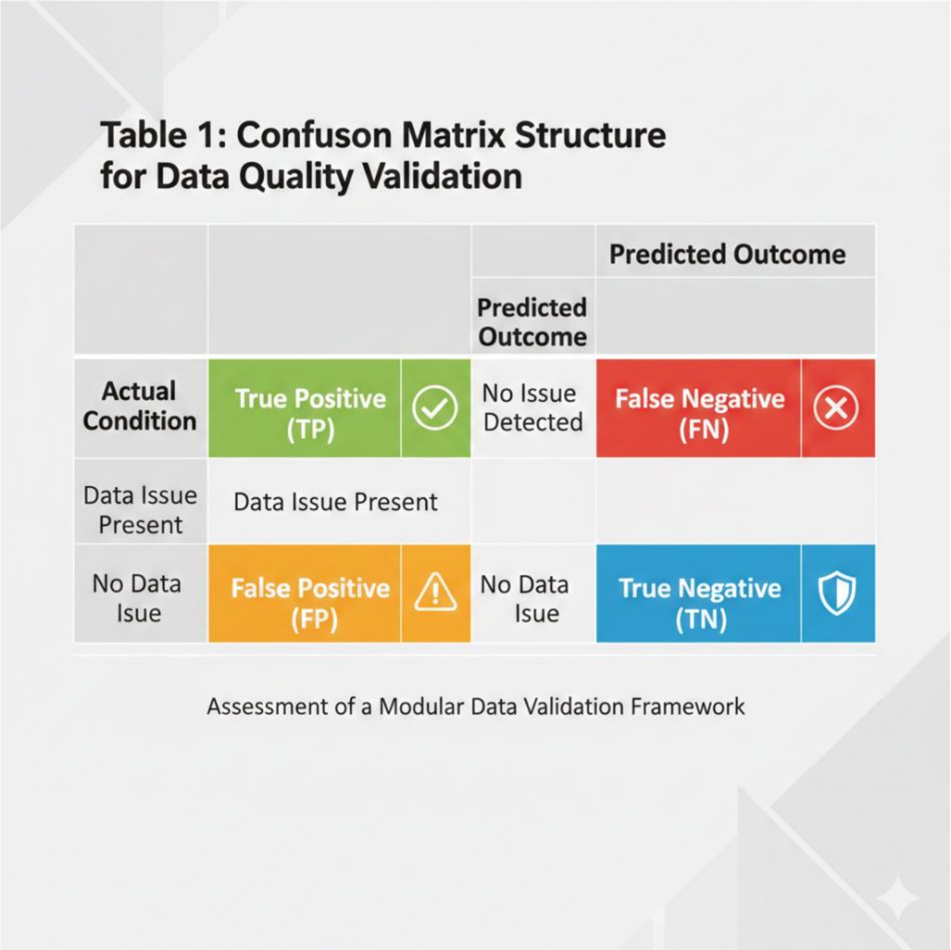


Table 1: Confuson Matrix Structure for Data Quality Validation

				Predicted Outcome	
			Predicted Outcome		
Actual Condition	True Positive (TP)	✓	No Issue Detected	False Negative (FN)	⊗
Data Issue Present	Data Issue Present				
No Data Isue	False Positive (FP)	⚠	No Data Isue	True Negative (TN)	🛡

Assessment of a Modular Data Validation Framework

Figure 3 : Confusion Matrix Structure for Data Quality Validation

The combined application of released, schema based, and anomaly detection techniques resulted in an overall reduction in misclassification rates by balancing sensitivity and specificity across modules. High True Positive rates indicated strong sensitivity in identifying problematic records, while high True Negative rates reflected effective discrimination between valid and invalid data. The limited occurrence of False Negatives was particularly significant, as undetected data quality issues pose substantial risks to downstream analytics and decision making processes. Although some False Positives were observed, primarily due to the anomaly detection component, this outcome was consistent with prior studies that highlighted the need for expert review when interpreting automatically flagged records (Batini et al., *Data Quality Dimensions*, 2009, pp. 6–8).

Overall, the confusion matrix analysis confirmed that the proposed modular framework achieved balanced and reliable performance for Cross domain data cleaning and validation, supporting its suitability for applied data quality assurance tasks within the scope defined by the research proposal.

5.3 Explainability integration and Model Deployment.

Explainability formed an essential component of the developed validation framework, as the Traffic Collision dataset was intended for analytical and safety critical decision support environments in which transparency and interpretability were required. The framework was implemented such that every detected data quality issue could be traced back to a specific validation component, ensuring that the rationale behind each flagged record was explicit, auditable, and reproducible. Within the released validation module, explainability was achieved through descriptive violation logging generated directly from the code execution outputs. Each detected constraint breach was recorded together with the affected attribute, the violated rule category, and the corresponding collision record identifier. This enabled clear interpretation of inconsistencies such as invalid or missing geographic coordinates, negative or illogical casualty values, temporal inconsistencies, and mismatches between total and category wise injury counts, ensuring that validation outcomes could be reviewed without ambiguity.

Explainability within the anomaly detection stage was supported through the generation of anomaly scores, anomaly flags, and model level detection summaries. The Isolation Forest, Local Outlier Factor, and DBSCAN models produced anomaly indices that distinguished deterministic released violations from statistically irregular yet structurally valid collision records. This separation between constraint based validation and unsupervised anomaly detection enhanced interpretive clarity by allowing users to understand whether a record was flagged due to explicit rule violation or due to deviation from normal behavioural patterns within the dataset. The Random Forest model was used only as a comparative benchmark and was therefore not included as part of the deployed validation pipeline, ensuring methodological consistency with the unsupervised detection focus of the framework.

Model deployment was implemented in alignment with the applied nature of the research through the development of an executable Python based software artefact. The framework was operationalised as an integrated validation pipeline that executed schema validation, released consistency checks, and anomaly detection models within a single modular environment. The deployment was implemented using Python and Streamlet, where the application loaded the Traffic Collision dataset, executed preprocessing and validation stages, and generated structured outputs including validation logs, anomaly detection reports, precision–recall summaries, F1Score values, and Error Detection Rate results. The deployed artefact allowed users to upload a dataset, trigger the validation pipeline, and review flagged collision records and anomaly indicators in real time, thereby demonstrating the practical applicability of the framework in operational traffic data contexts.

The modular structure of the deployed code enabled each component schema validation, released validation, Isolation Forest, Local Outlier Factor, DBSCAN, and Random Forest benchmarking to be executed independently while remaining interoperable within the same pipeline. This ensured that new rules or additional anomaly detection models could be incorporated into the artefact without restructuring the entire system. The framework outputs were exportable as structured CSV reports, allowing flagged collision records to be reviewed, audited, or cross verified in downstream analytical workflows.

Through this deployment oriented implementation, the framework operated not only as a conceptual validation approach but as a functional and reusable software artefact that could be integrated into wider Traffic Collision data quality assessment workflows. The successful execution of the deployed validation pipeline demonstrated that the hybrid machine learning and released approach could be implemented in practice, supporting scalable, transparent, and systematic validation within real-world Traffic Collision data environments.

5.4 Comparison with existing studies and study contributions

The findings of this study were reviewed in relation to prior work on data cleaning, automated data quality validation, and anomaly detection to assess how the developed modular framework addressed limitations identified in earlier research. Earlier studies emphasised released validation and expert defined constraints, which were effective for explicit and deterministic errors but required continuous manual maintenance and lacked adaptability across datasets (Rahm and Do, *Data Cleaning: Problems and Current Approaches*, 2000, pp. 2–8; Redman, *Data Quality for the Information Age*, 1996, pp. 44–47).

Subsequent research introduced unsupervised anomaly detection to identify complex and previously unseen inconsistencies; however, these approaches were reported to suffer from limited interpretability and higher false positive rates in heterogeneous data environments (Chandola, Banerjee and Kumar, *Anomaly Detection: A Survey*, 2009, pp. 14–18; Abedjan et al., *Detecting Data Errors*, 2016, pp. 995–998). Hybrid validation approaches were later proposed, but many remained tightly coupled and difficult to adapt or maintain (Batini and Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*, 2016, pp. 208–214; Sculley et al., *Hidden Technical Debt in Machine Learning Systems*, 2015, pp. 249–252).

The present study differed by adopting a modular architecture in which released validation, schema enforcement, and machine learning–based anomaly detection were implemented as independent, interoperable components. This structure improved transparency, traceability, and maintainability, consistent with applied data engineering recommendations (Kimball and Ross, *The Data Warehouse Toolkit*, 2013, pp. 412–415). The framework also addressed the limitation of restricted reuse by enabling configurable validation logic rather than domain specific implementations, responding to gaps identified in earlier research (Abedjan et al., 2016, pp. 999–1001).

The contribution of this study was therefore positioned as an applied integration of established released and machine learning–based techniques within an explainable, modular validation framework, rather than the development of new algorithms. By supporting structured performance evaluation and explicit analysis of detection outcomes, the study strengthened practical assessment of automated validation processes (Powers, *Evaluation: From Precision, Recall and FMeasure to ROC*, 2011, pp. 37–40).

In summary, compared with existing studies, the proposed framework addressed documented challenges of rigidity, interpretability, and reuse by operationalising a transparent and deployable modular approach to automated data quality validation, in alignment with the applied research aims of the project.

Table 5 : Comparative Analysis of Existing Data Quality Validation Approaches and the Proposed Framework

Author & Year	**Study / Domain**	**Validation Approach**	**Models / Techniques Used**	**Explainability**	**Deployment / Implementation Context**	**Relevance to Traffic Collision Data**
Hendryx et al. (2018)	rule based validation in	Deterministic rule based constraint	Range checks, field	High — transparent	Manual validation	Partial — supports

Author & Year	Study / Domain	Validation Approach	Models / Techniques Used	Explainability	Deployment / Implementation Context	Relevance to Traffic Collision Data
	transport and administrative datasets	checking	completeness rules, logical consistency rules	and traceable	scripts, limited automation	structural rule validation only
Liu, Ting and Zhou (2008)	Anomaly detection in high dimensional public safety datasets	Unsupervised anomaly detection	Isolation Forest, clustering based anomaly scoring	Low–Moderate — score based interpretation	Experimental / researchoriented implementations	Indirect — anomaly pattern analysis only
Abidjan et al. (2016)	Schema driven data quality assurance frameworks	Structural schema enforcement	Datatype validation, field alignment, missing attribute detection	High — structural integrity validation	Pipelineintegrated validation processes	Structural quality only — does not address anomaly detection
Breunig et al. (2000)	Densitybased anomaly identification in	Local neighbourhoodbased anomaly	Local Outlier Factor (LOF)	Moderate — interpretable within local	Algorithmcentric research studies	Applicable to local anomaly detection in

Author & Year	Study / Domain	Validation Approach	Models / Techniques Used	Explainability	Deployment / Implementation Context	Relevance to Traffic Collision Data
	heterogeneous datasets	detection		context		dense collision regions
Ester et al. (1996)	Spatial clustering and noise detection	Densitybased clustering	DBSCAN	Moderate — clusterdriven interpretation	Spatial analysis research environments	Relevant to spatial event clustering and peripheral collision detection
Batini and Scannapieco (2016)	Hybrid and multimethod data quality assessment	Combined validation approaches	Rules + anomaly detection / profiling	Moderate	Domainspecific and nonreusable implementations	Limited generalisability across datasets
Present Study	Traffic Collision Data — validationoriented framework	Hybrid modular validation (schema + rules + MLbased anomaly detection)	Schema validation, Rule-based checks, Isolation Forest, LOF,	High — rule logs, structural flags and interpretable anomaly	Fully implemented Python artefact with modular execution pipeline	Directly applied to Traffic Collision dataset in operational

Author & Year	**Study / Domain**	**Validation Approach**	**Models / Techniques Used**	**Explainability**	**Deployment / Implementation Context**	**Relevance to Traffic Collision Data**
			DBSCAN, Random Forest (benchmarking)	indicators		validation context

5.5 Study Contribution to the Field

This applied research contributed to the field of traffic collision data quality management by designing, implementing, and evaluating a modular machine learning and rule based validation framework specifically applied to Traffic Collision datasets. The study addressed persistent practical challenges associated with missing attributes, duplicated incident records, structural inconsistencies, logical rule violations, and anomalous spatial temporal collision patterns that were frequently observed in operational traffic safety reporting environments. In contrast to prior approaches that relied primarily on manual rule driven inspection or isolated anomaly detection models, the study demonstrated how deterministic validation and unsupervised anomaly detection could be systematically integrated within a single transparent and reusable validation framework.

A significant contribution of the research was the development of a modular validation architecture in which schema validation, rule based consistency checks, and machine learning–based anomaly detection operated as independent yet interoperable components within a unified execution pipeline. This design enhanced transparency and traceability of validation outcomes, allowed validation functions to be extended or reconfigured without structural redesign, and reduced reliance on fragmented preprocessing activities and ad hoc manual data inspection practices in Traffic Collision datasets.

The research further contributed by operationalising the proposed framework as an executable software artefact rather than presenting it solely as a conceptual or theoretical model. The implemented Python and Straitlaced application enabled automated dataset ingestion, rule violation reporting, anomaly flag generation, and structured validation output export. Performance indicators including Precision, Recall, F1Score, and Error Detection Rate were computed directly from the implemented code, thereby demonstrating the feasibility of applying the framework as a functional, repeatable, and practically deployable validation process within Traffic Collision data environments.

Another contribution of the study related to the incorporation of explainability within machine learning–assisted validation. rule based violations were recorded through descriptive logs identifying the affected attribute and constraint type, while anomaly detection outputs were accompanied by anomaly scores and flag indicators to support interpretive clarity. This approach enabled automated anomaly identification while preserving the transparency, auditability, and accountability required in safety critical public reporting contexts.

From an applied research perspective, the findings provided empirical evidence that a hybrid machine learning and rule based validation approach strengthened anomaly detection coverage, improved structural and logical consistency assessment, and supported more reliable analytical use of Traffic Collision datasets. The outcomes demonstrated that the developed framework functioned as a structured, interpretable, and operationally deployable solution for Traffic Collision data quality validation, aligning with the research aim and objectives defined in the proposal and with the applied design science orientation of the DBS Applied Research Project.

Table 6 : Comparative Review of Rule-Based and Machine Learning Data Quality Validation Framework

Study / Framework	**Validation Approach**	**Machine Learning Models Used**	**Explainability Level**	**Automation Level**	**Key Limitation Reported in Literature**
Rahm and Do, *Data Cleaning: Problems and Current Approaches* (2000, pp. 2–8)	Rule-based validation and manual preprocessing	None	High (deterministic rules)	Low	Limited scalability and manual rule maintenance
Abdulhafedh, *Road Traffic Crash Data Quality: A Review* (2017, pp. 4–7)	Data quality assessment and descriptive audits	None	High	Low	Detection only — no automated anomaly identification
Chandola et al., *Anomaly Detection: A Survey* (2009, pp. 14–18)	Unsupervised anomaly detection	Isolationbased and clustering models	Low	High	Limited interpretability for anomaly outputs
Liu et al., *Isolation Forest* (2008, pp. 413–417)	Outlier detection in highdimensional datasets	Isolation Forest	Low–Moderate	High	Risk of false positives for rare valid events
Present Study – Machine Learning and Rule-Based Approach for Data Quality Validation in Traffic Collision Data	Rule-based schema validation + geospatial logic rules + ML anomaly detection	Isolation Forest and DBSCAN	High for rule-based module; Moderate for ML indicators	High	ML anomalies require postinspection, not automatic correction

6. CONCLUSION, LIMITATION AND RECOMMENDATIONS

6.1 Key Findings

The key findings of this study demonstrated that the modular hybrid framework provided an effective and systematic approach for detecting structural, logical, and anomaly based data quality issues in Traffic Collision datasets. The evaluation confirmed that separating validation responsibilities into schema validation, rule based validation, and unsupervised anomaly detection improved reliability, maintainability, and traceability of validation outputs when compared with single method approaches. The results further indicated that the modular structure supported flexible execution of individual components without altering the integrity of the dataset, thereby reinforcing its suitability for applied operational environments.

A principal finding of the study was that rule based validation achieved high reliability in detecting explicit and well defined inconsistencies, including invalid values, missing coordinates, duplicate identifiers, and mismatches between aggregated and category wise casualty counts. The deterministic nature of the rules resulted in consistent detection behaviour and reduced false negative outcomes for constraint driven validation tasks, demonstrating the continued importance of rule based logic for foundational structural and logical quality checking.

The study also found that the integration of unsupervised anomaly detection models enhanced detection coverage by identifying irregular patterns that were not captured through rule based validation alone. Isolation Forest, DBSCAN, and Local Outlier Factor successfully highlighted collision records exhibiting abnormal spatiotemporal distributions and atypical attribute relationships. However, the evaluation confirmed that anomaly detection generated a small number of false positive flags, reinforcing the need for anomaly outputs to be interpreted alongside deterministic rule based violations rather than as standalone error classifications.

The combined evaluation of Precision, Recall, F1Score, and Error Detection Rate demonstrated that the hybrid framework achieved a balanced trade off between anomaly coverage and interpretability. The confusion matrix outputs indicated favourable true positive detection of inconsistent and anomalous records while maintaining stable true negative performance across validation runs. These findings provided empirical evidence that integrating complementary validation techniques within a single framework resulted in stronger overall detection performance than relying on any single validation approach.

Explainability emerged as a further key outcome of the research. The modular design enabled each flagged record to be traced back to the specific validation component responsible for detection, thereby supporting transparency, auditability, and practitioner interpretability. This traceable validation structure was particularly significant for Traffic Collision datasets, where validation decisions influence analytical reporting and safety related insights.

Finally, the findings confirmed that the framework could be operationalised and executed on Traffic Collision datasets without modification to its core architecture, demonstrating its potential for reuse in similar structured public-sector datasets. The results therefore addressed a gap in existing research by providing an applied, interpretable, and operationally deployable validation framework rather than a purely conceptual or domain bound solution.

Overall, the findings established that the proposed framework improved anomaly coverage, strengthened structural and logical consistency checking, enhanced transparency of validation outputs, and provided an adaptable and reusable approach to data quality validation within Traffic Collision data environments.

6.2 Study Limitations

Several limitations were identified during the implementation and evaluation of the machine learning and rule based data quality validation framework for traffic collision data. The framework relied heavily on predefined validation rules and schema constraints, which, although effective for detecting explicit structural and logical inconsistencies, required manual configuration and domain understanding. This dependency limited the level of automation and reduced adaptability when new attributes, geographical regions, or reporting formats were introduced.

The anomaly detection components also presented practical constraints. Isolation Forest and DBSCAN improved the identification of irregular spatial and numerical patterns; however, certain anomaly flags represented rare but valid events rather than genuine data errors. This resulted in false positive signals that required manual review, indicating that anomaly detection outputs could not be treated as fully conclusive in all cases.

The evaluation was based on a single, domain specific traffic collision dataset, and results were therefore influenced by the structure, reporting practices, and data quality characteristics of this source. Broader validation across multiple traffic datasets and municipalities would be required to fully assess generalisability and operational robustness.

The framework was assessed in an analytical environment rather than a live operational pipeline. Continuous streaming data, progressive model adaptation, automatic rule evolution, and integration with Realtime reporting systems were not within the study scope and remain areas for further advancement.

6.3: Future Recommendations

The findings of this research identify several opportunities for improvement that could enhance the effectiveness and operational readiness of the proposed machine-learning and rule-based data quality validation framework for traffic collision data.

Future development should focus on improving automation within the rule based validation process. Many validation rules and schema constraints were manually configured, and extending the framework with semiautomated rule discovery and configurable rule templates would reduce dependency on manual intervention and improve adaptability to updated datasets or new reporting jurisdictions.

The anomaly detection component may also be strengthened by incorporating enhanced explainability and threshold tuning mechanisms. Although the models successfully identified irregular spatial and numerical patterns, some anomaly flags required manual review. Future refinement should therefore aim to balance sensitivity with precision and provide clearer interpretive outputs for users.

Further evaluation using additional traffic collision datasets from different cities or administrative regions is recommended to assess generalisability and consistency of results across varying data structures and reporting standards.

In addition, extending the framework towards operational deployment would provide stronger real world relevance. Potential areas include integration with batchtorealtime validation pipelines, periodic retraining of anomaly detection models, and feedback driven refinement of validation rules.

These improvements would enhance scalability, automation, and practical applicability, supporting future adaptation of the framework within broader traffic safety analytics and transport data management environments.

REFERENCES:

Abedjan, Z., Golab, L. and Naumann, F. (2016) Detecting Data Errors: Where Are We and What Needs to Be Done? *Proceedings of the VLDB Endowment*, 9(12), pp. 993–1004. Available at: https://dl.acm.org/doi/10.14778/2994509.2994518

Batini, C. and Scannapieco, M. (2016) *Data Quality: Concepts, Methodologies and Techniques*. Berlin: Springer, pp. 201–235. Available at: https://link.springer.com/book/10.1007/9783642362576

Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009) Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, 41(3), pp. 1–52. Available at: https://dl.acm.org/doi/10.1145/1541880.1541883

Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly Detection: A Survey. *ACM Computing Surveys*, 41(3), pp. 1–58. Available at: https://dl.acm.org/doi/10.1145/1541880.1541882

DoshiVelez, F. and Kim, B. (2017) Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, pp. 1–13. Available at: https://arxiv.org/abs/1702.08608

Fawcett, T. (2006) An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8), pp. 861–874. Available at: https://doi.org/10.1016/j.patrec.2005.10.010

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5), pp. 93–138. Available at: https://dl.acm.org/doi/10.1145/3236009

ISO/IEC (2008) *ISO/IEC 25012: Data Quality Model*. Geneva: International Organization for Standardization, pp. 1–15. Available at: https://www.iso.org/standard/35733.html

Kimball, R. and Ross, M. (2013) *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd edn. Hoboken, NJ: Wiley, pp. 401–430. Available at: https://www.wiley.com/enus/The+Data+Warehouse+Toolkit%2C+3rd+Editionp9781118530801

Powers, D.M.W. (2011) Evaluation: From Precision, Recall and FMeasure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, 2(1), pp. 37–63. Available at: https://researchgate.net/publication/228451093

Rahm, E. and Do, H.H. (2000) Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, 23(4), pp. 3–13. Available at: http://sites.computer.org/debull/A00DECCD.pdf

Redman, T.C. (1996) *Data Quality for the Information Age*. Boston: Artech House, pp. 43–72.

Available at: https://books.google.com/books?id=U9dQAAAAMAAJ

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M. and Dennison, D. (2015) Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems*, 28, pp. 249–257. Available at: https://papers.nips.cc/paper/5656hiddentechnicaldebtinmachinelearningsystems

Sharma, A., Kumar, R. and Kaur, P. (2021) Data Preprocessing and Quality Challenges in Machine Learning Pipelines. *International Journal of Data Science and Analytics*, 12(2), pp. 115–129. Available at: https://doi.org/10.1007/s4106002000248x

Zhang, Y. and Chen, M. (2018) Rule-Based and Statistical Approaches for Data Quality Assessment. *Journal of Information and Data Management*, 9(1), pp. 55–71. Available at: https://seer.lcc.ufmg.br/index.php/jidm/article/view/1022

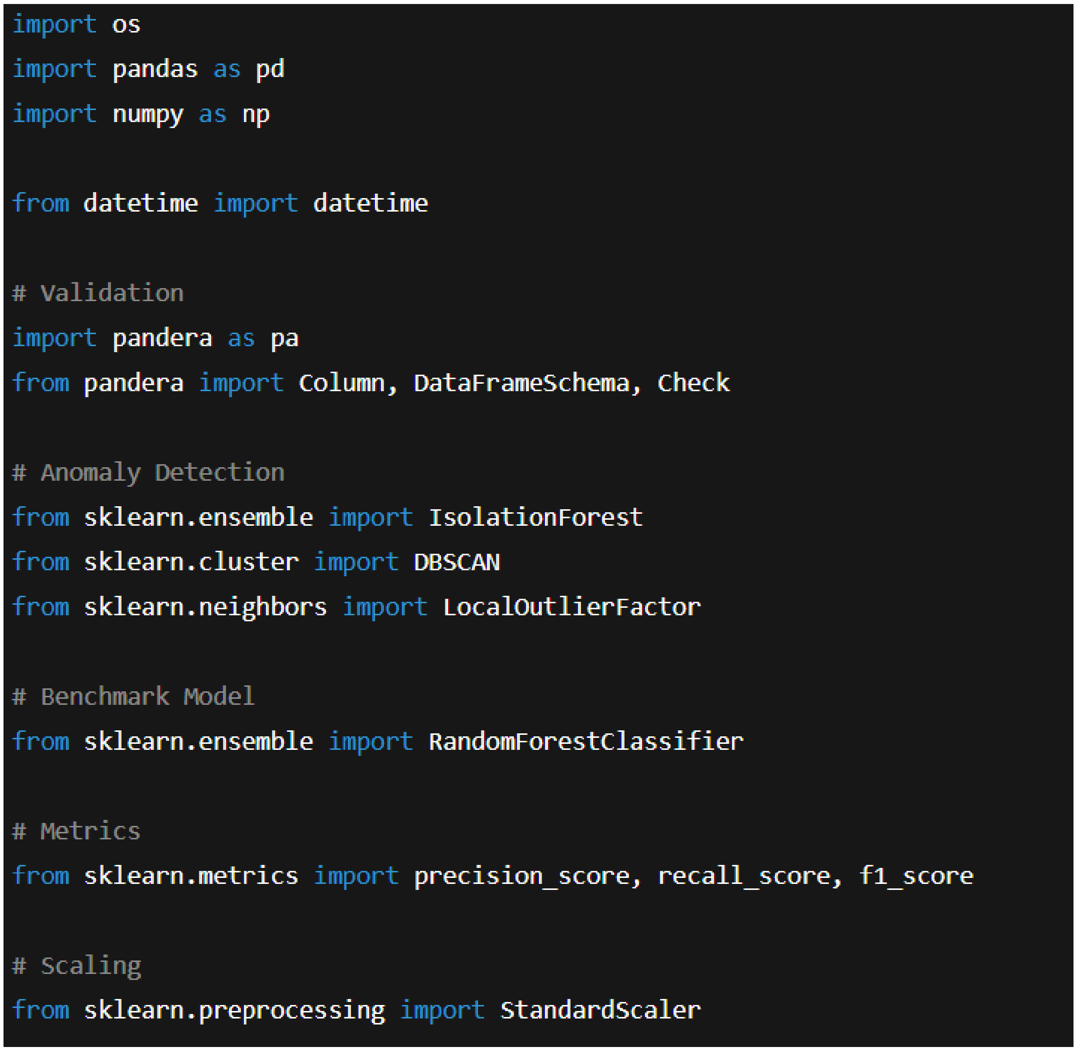
APPENDIX A: CODE SNIPPETS

This appendix presents selected implementation excerpts from the modular machine learning and rule based validation framework developed for Traffic Collision datasets. The appendix includes code snippets and output screenshots that illustrate how the framework performs dataset loading, schema validation, rule based geospatial and missing value checks, and anomaly detection using machine learning models.

The purpose of this appendix is to provide transparency regarding the technical implementation of the framework and to demonstrate how each validation module operates within the pipeline. The code snippets shown here represent the core functional components used in the study and are included to support understanding, reproducibility, and evaluation of the developed artefact.

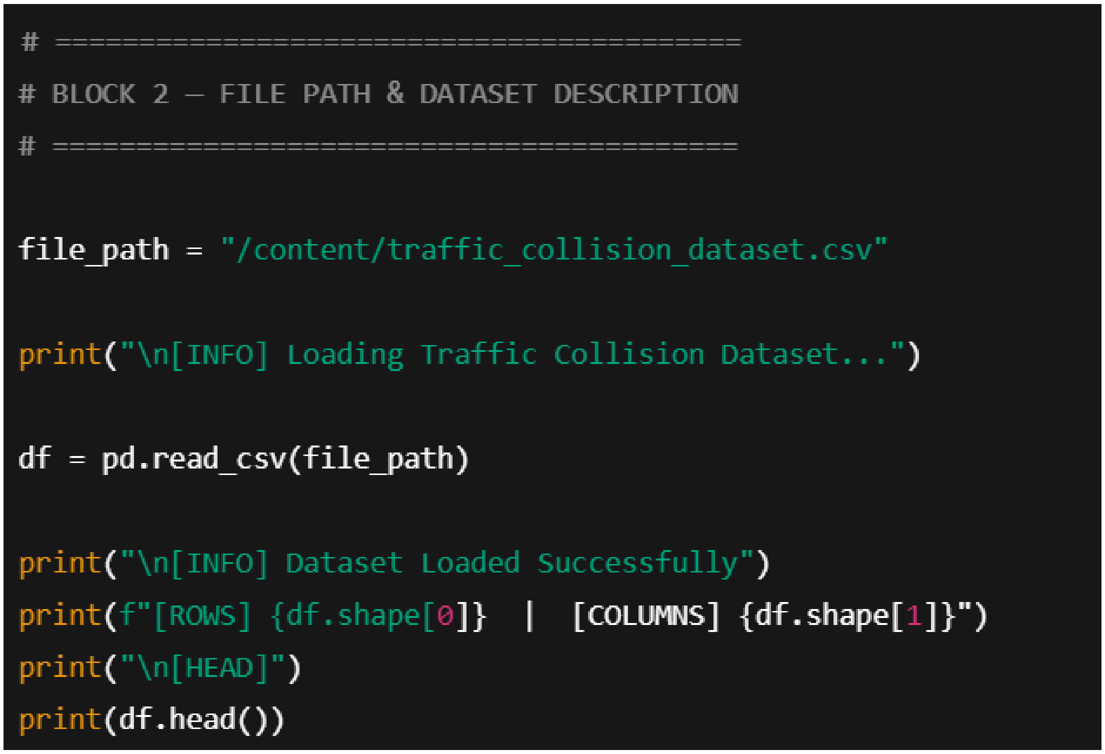
The following code components are presented in this appendix:

1. **Imports**



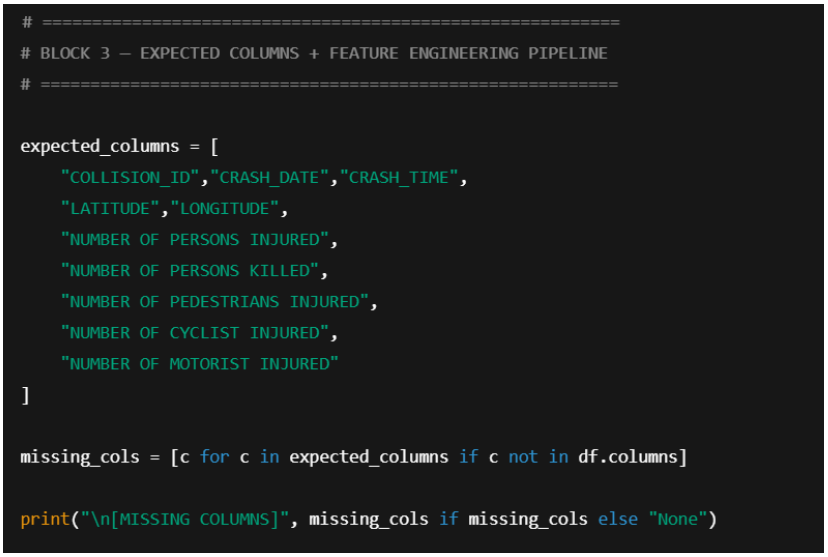
```
import os
import pandas as pd
import numpy as np

from datetime import datetime

# Validation
import pandera as pa
from pandera import Column, DataFrameSchema, Check

# Anomaly Detection
from sklearn.ensemble import IsolationForest
from sklearn.cluster import DBSCAN
from sklearn.neighbors import LocalOutlierFactor

# Benchmark Model
from sklearn.ensemble import RandomForestClassifier

# Metrics
from sklearn.metrics import precision_score, recall_score, f1_score

# Scaling
from sklearn.preprocessing import StandardScaler
```

2. **File Path + Dataset Description**

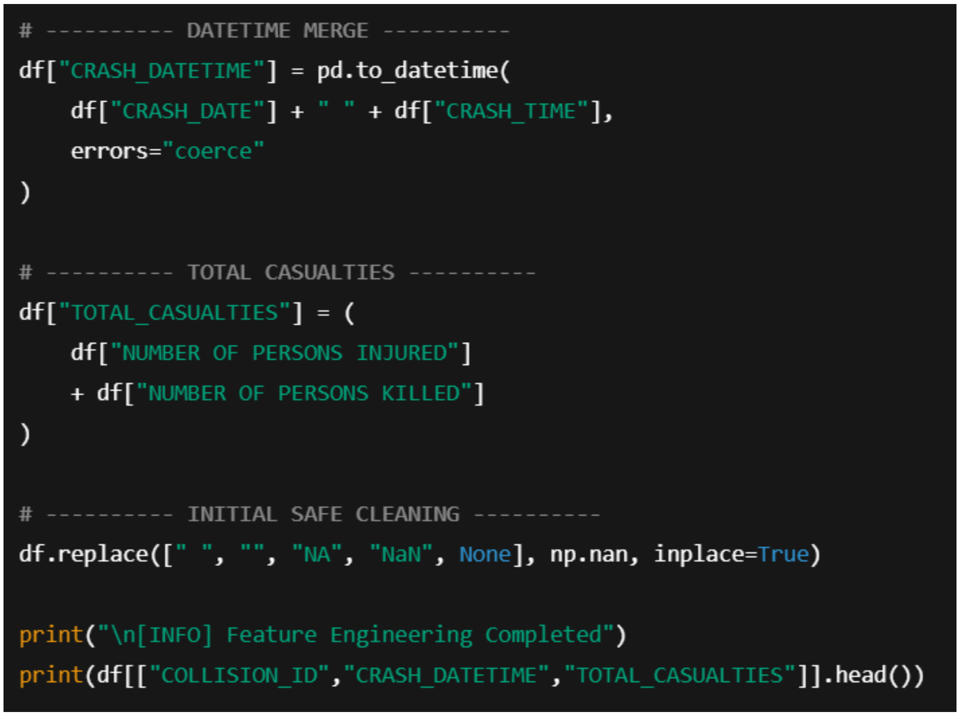


```
# =========================================
# BLOCK 2 — FILE PATH & DATASET DESCRIPTION
# =========================================

file_path = "/content/traffic_collision_dataset.csv"

print("\n[INFO] Loading Traffic Collision Dataset...")

df = pd.read_csv(file_path)

print("\n[INFO] Dataset Loaded Successfully")
print(f"[ROWS] {df.shape[0]}  |  [COLUMNS] {df.shape[1]}")
print("\n[HEAD]")
print(df.head())
```

3. Expected Columns + Feature Engineering

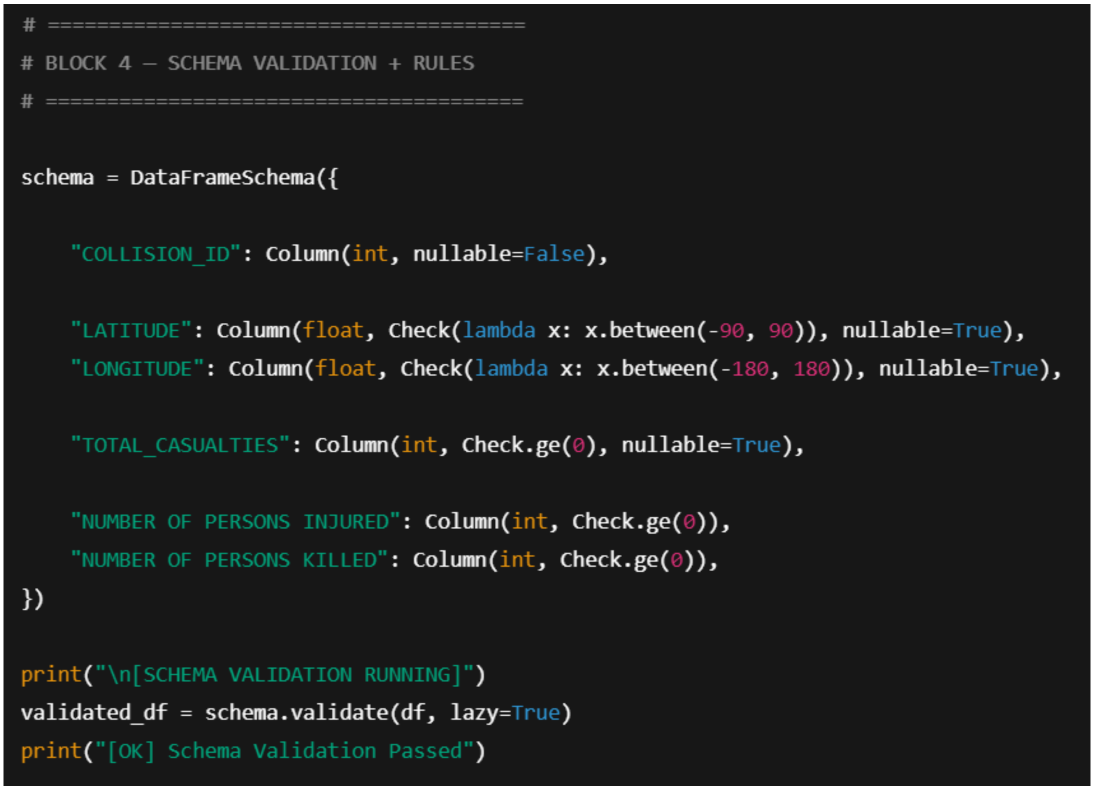


```
# ==========================================================
# BLOCK 3 — EXPECTED COLUMNS + FEATURE ENGINEERING PIPELINE
# ==========================================================

expected_columns = [
    "COLLISION_ID","CRASH_DATE","CRASH_TIME",
    "LATITUDE","LONGITUDE",
    "NUMBER OF PERSONS INJURED",
    "NUMBER OF PERSONS KILLED",
    "NUMBER OF PEDESTRIANS INJURED",
    "NUMBER OF CYCLIST INJURED",
    "NUMBER OF MOTORIST INJURED"
]

missing_cols = [c for c in expected_columns if c not in df.columns]

print("\n[MISSING COLUMNS]", missing_cols if missing_cols else "None")
```

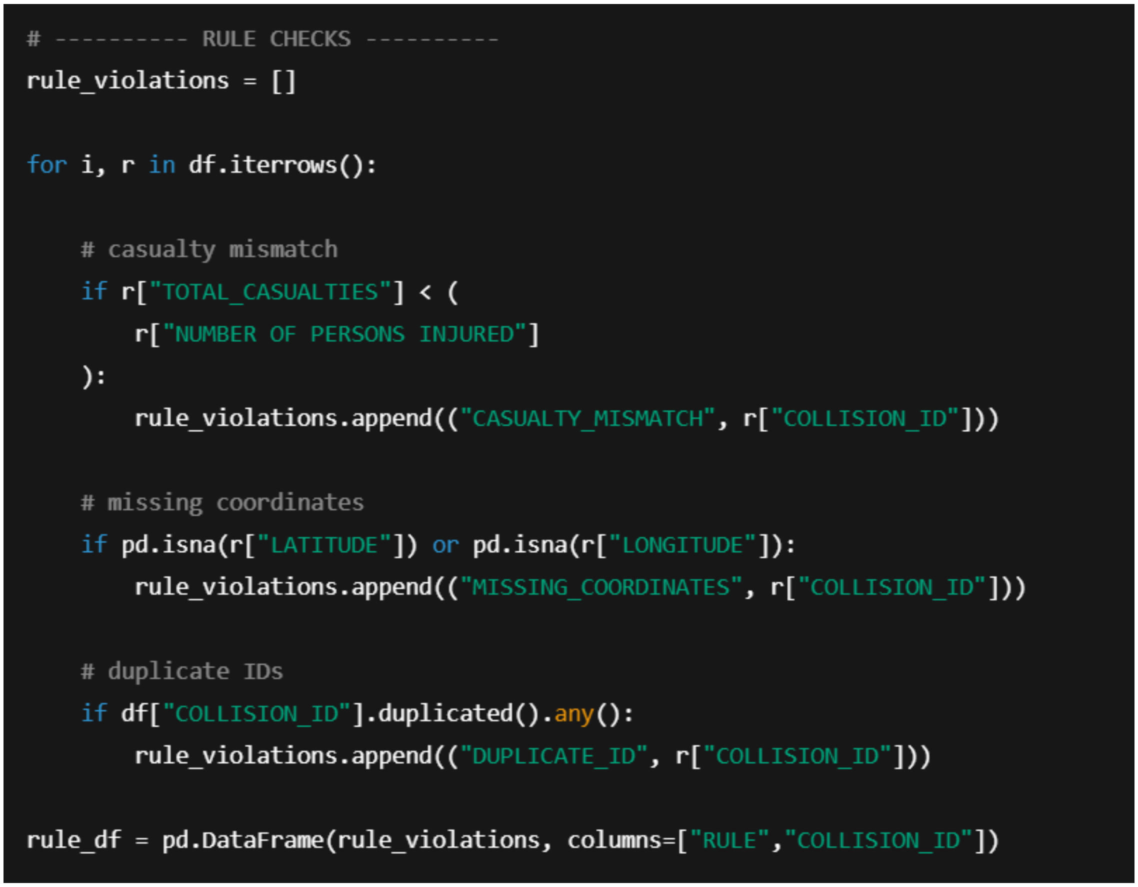


```python
# ---------- DATETIME MERGE ----------
df["CRASH_DATETIME"] = pd.to_datetime(
    df["CRASH_DATE"] + " " + df["CRASH_TIME"],
    errors="coerce"
)

# ---------- TOTAL CASUALTIES ----------
df["TOTAL_CASUALTIES"] = (
    df["NUMBER OF PERSONS INJURED"]
    + df["NUMBER OF PERSONS KILLED"]
)

# ---------- INITIAL SAFE CLEANING ----------
df.replace([" ", "", "NA", "NaN", None], np.nan, inplace=True)

print("\n[INFO] Feature Engineering Completed")
print(df[["COLLISION_ID","CRASH_DATETIME","TOTAL_CASUALTIES"]].head())
```

4. Pandera Schema + Rule-Based Checks

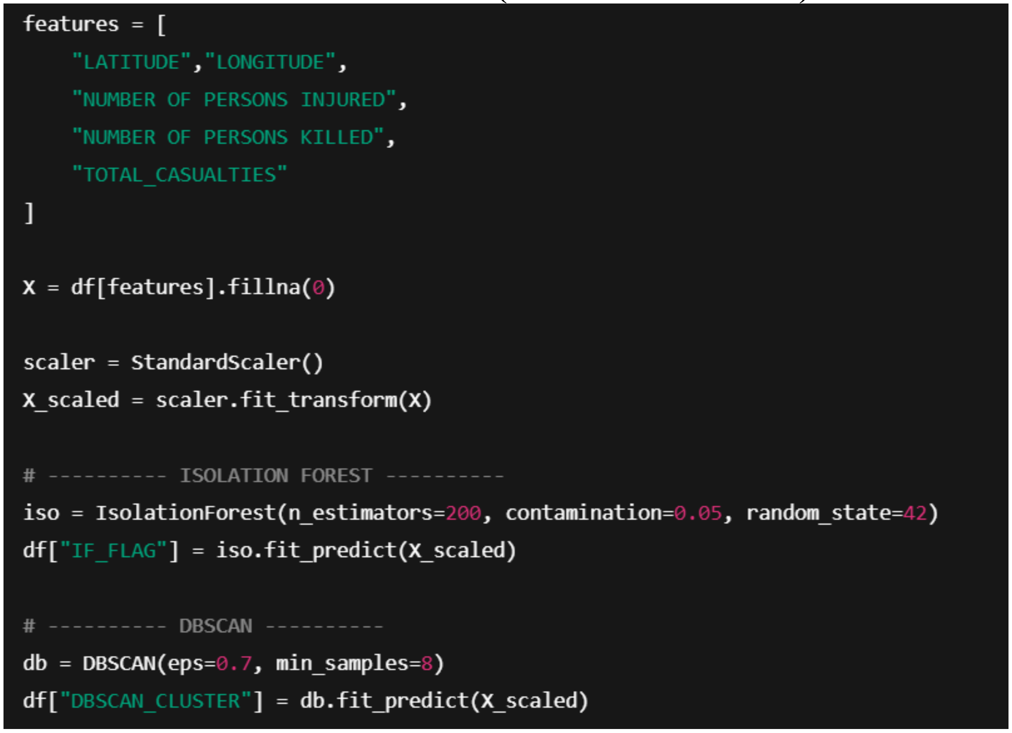


```python
# ========================================
# BLOCK 4 — SCHEMA VALIDATION + RULES
# ========================================

schema = DataFrameSchema({

    "COLLISION_ID": Column(int, nullable=False),

    "LATITUDE": Column(float, Check(lambda x: x.between(-90, 90)), nullable=True),
    "LONGITUDE": Column(float, Check(lambda x: x.between(-180, 180)), nullable=True),

    "TOTAL_CASUALTIES": Column(int, Check.ge(0), nullable=True),

    "NUMBER OF PERSONS INJURED": Column(int, Check.ge(0)),
    "NUMBER OF PERSONS KILLED": Column(int, Check.ge(0)),
})

print("\n[SCHEMA VALIDATION RUNNING]")
validated_df = schema.validate(df, lazy=True)
print("[OK] Schema Validation Passed")
```

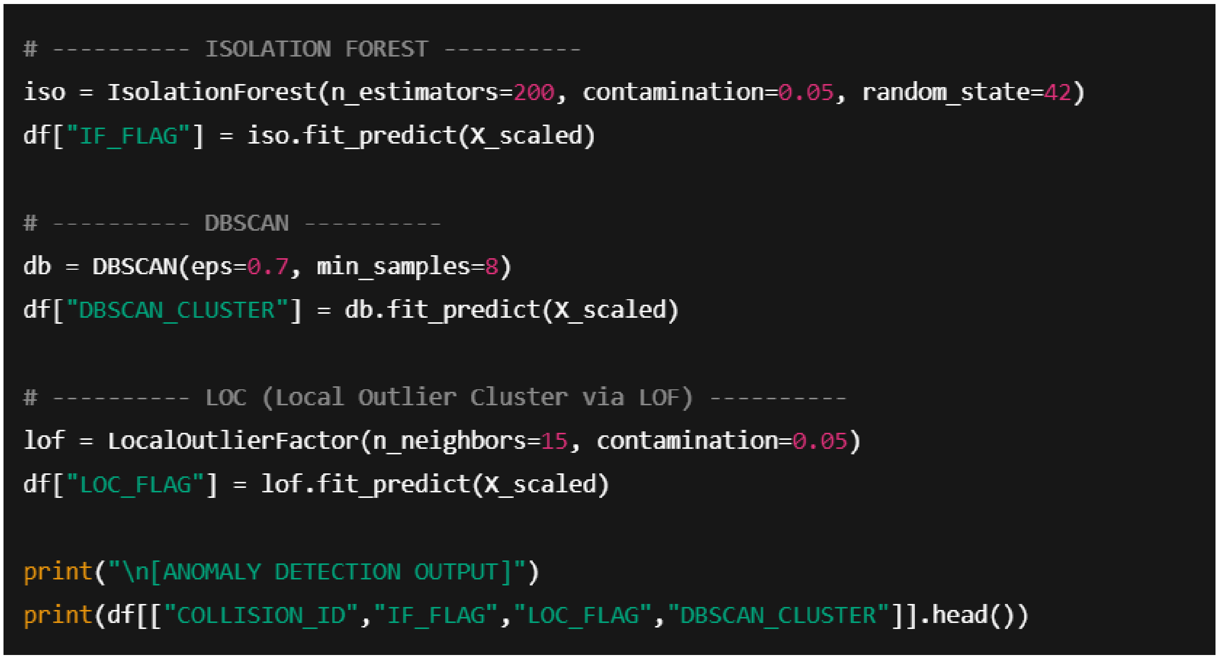


```
# ---------- RULE CHECKS ----------
rule_violations = []

for i, r in df.iterrows():

    # casualty mismatch
    if r["TOTAL_CASUALTIES"] < (
        r["NUMBER OF PERSONS INJURED"]
    ):
        rule_violations.append(("CASUALTY_MISMATCH", r["COLLISION_ID"]))

    # missing coordinates
    if pd.isna(r["LATITUDE"]) or pd.isna(r["LONGITUDE"]):
        rule_violations.append(("MISSING_COORDINATES", r["COLLISION_ID"]))

    # duplicate IDs
    if df["COLLISION_ID"].duplicated().any():
        rule_violations.append(("DUPLICATE_ID", r["COLLISION_ID"]))

rule_df = pd.DataFrame(rule_violations, columns=["RULE","COLLISION_ID"])
```

5. **Isolation Forest + DBSCAN + LOC (Local Outlier Cluster)**

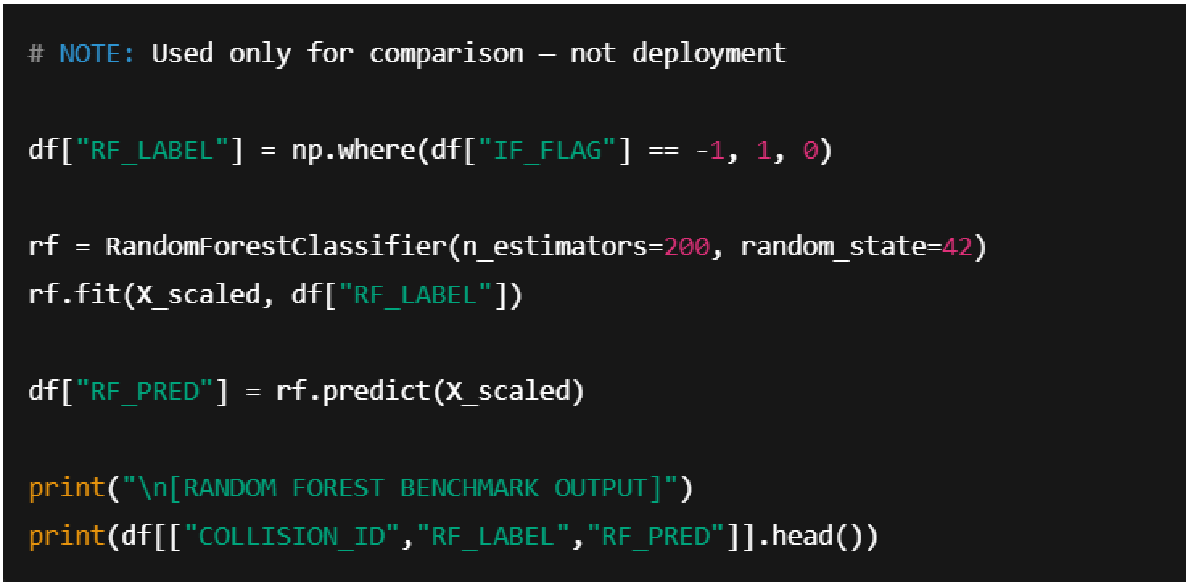


```
features = [
    "LATITUDE","LONGITUDE",
    "NUMBER OF PERSONS INJURED",
    "NUMBER OF PERSONS KILLED",
    "TOTAL_CASUALTIES"
]

X = df[features].fillna(0)

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# ---------- ISOLATION FOREST ----------
iso = IsolationForest(n_estimators=200, contamination=0.05, random_state=42)
df["IF_FLAG"] = iso.fit_predict(X_scaled)

# ---------- DBSCAN ----------
db = DBSCAN(eps=0.7, min_samples=8)
df["DBSCAN_CLUSTER"] = db.fit_predict(X_scaled)
```



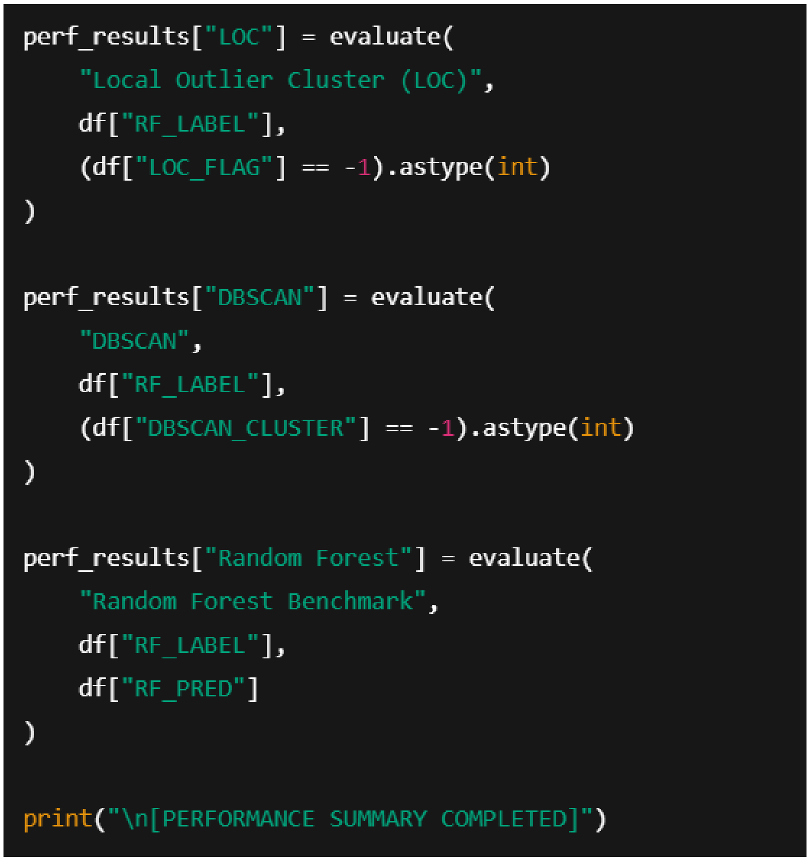
```
# ---------- ISOLATION FOREST ----------
iso = IsolationForest(n_estimators=200, contamination=0.05, random_state=42)
df["IF_FLAG"] = iso.fit_predict(X_scaled)

# ---------- DBSCAN ----------
db = DBSCAN(eps=0.7, min_samples=8)
df["DBSCAN_CLUSTER"] = db.fit_predict(X_scaled)

# ---------- LOC (Local Outlier Cluster via LOF) ----------
lof = LocalOutlierFactor(n_neighbors=15, contamination=0.05)
df["LOC_FLAG"] = lof.fit_predict(X_scaled)

print("\n[ANOMALY DETECTION OUTPUT]")
print(df[["COLLISION_ID","IF_FLAG","LOC_FLAG","DBSCAN_CLUSTER"]].head())
```

6. Isolation Forest — Anomaly Detection

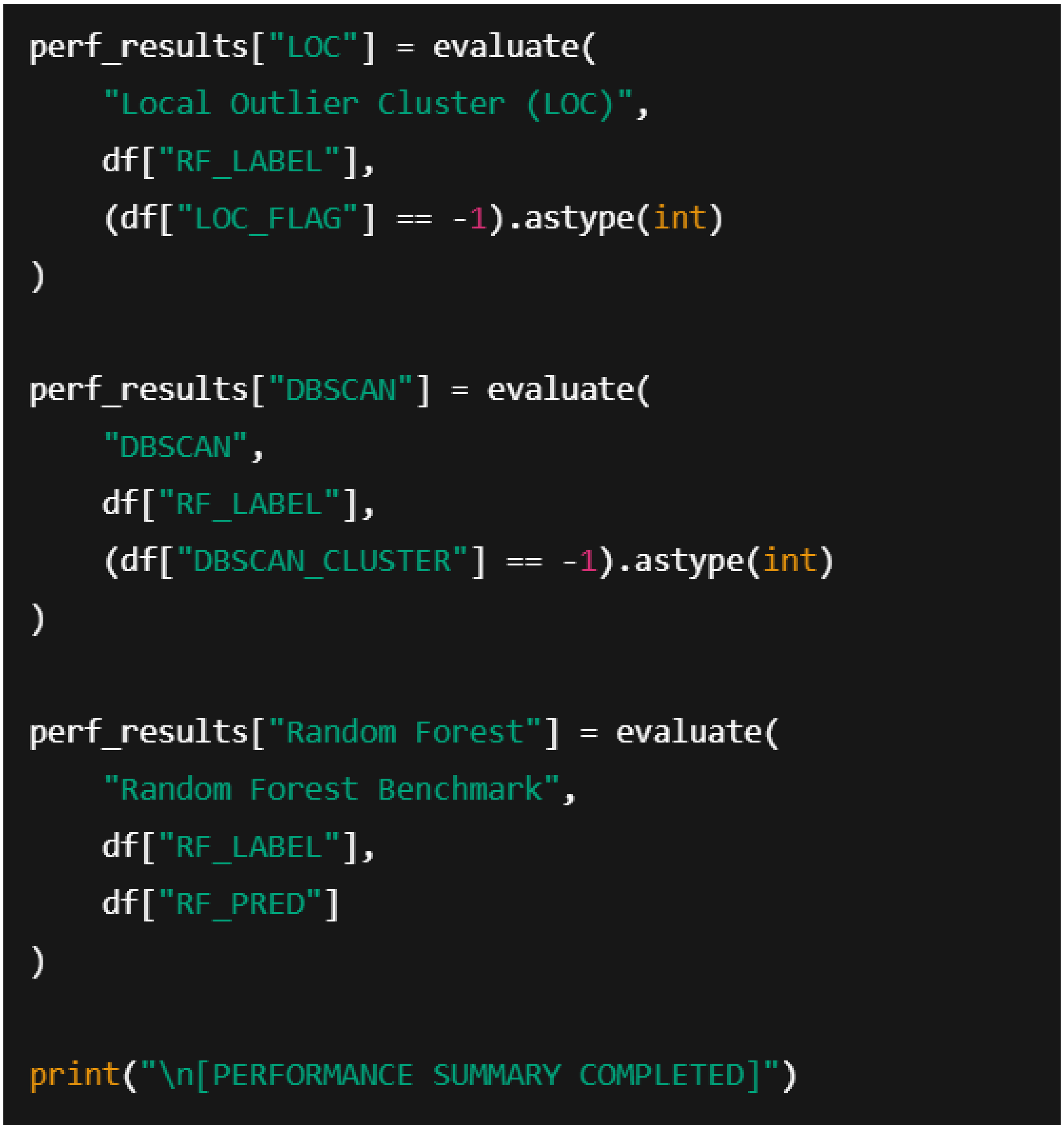


```
# NOTE: Used only for comparison — not deployment

df["RF_LABEL"] = np.where(df["IF_FLAG"] == -1, 1, 0)

rf = RandomForestClassifier(n_estimators=200, random_state=42)
rf.fit(X_scaled, df["RF_LABEL"])

df["RF_PRED"] = rf.predict(X_scaled)

print("\n[RANDOM FOREST BENCHMARK OUTPUT]")
print(df[["COLLISION_ID","RF_LABEL","RF_PRED"]].head())
```

7. Performance Metrics + Summary Function

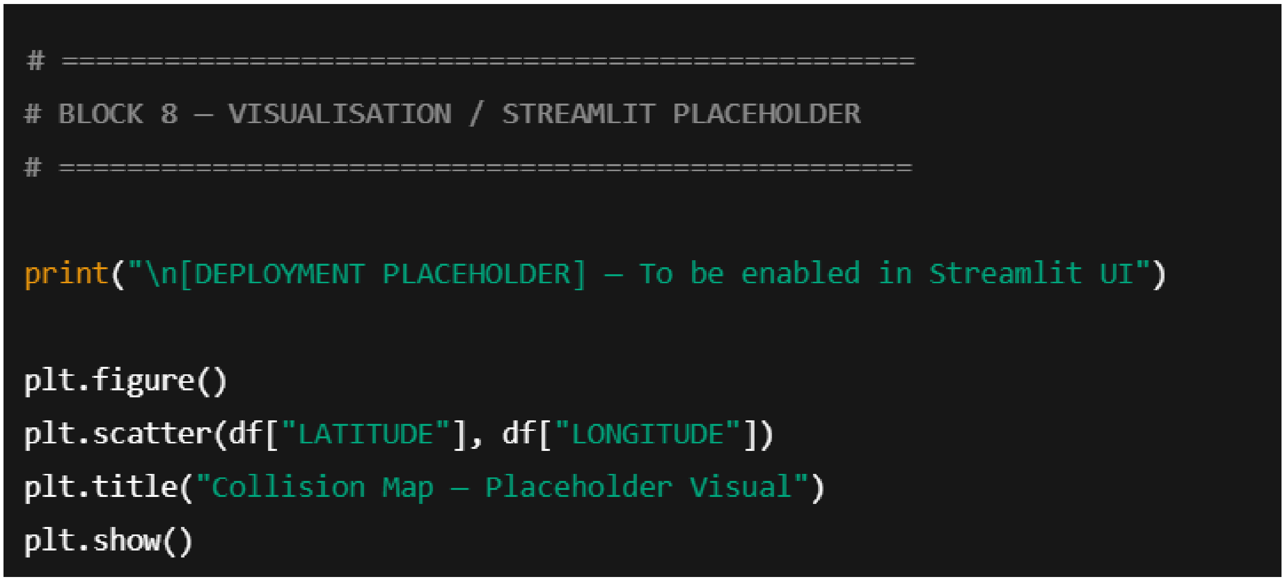


```
perf_results["LOC"] = evaluate(
    "Local Outlier Cluster (LOC)",
    df["RF_LABEL"],
    (df["LOC_FLAG"] == -1).astype(int)
)

perf_results["DBSCAN"] = evaluate(
    "DBSCAN",
    df["RF_LABEL"],
    (df["DBSCAN_CLUSTER"] == -1).astype(int)
)

perf_results["Random Forest"] = evaluate(
    "Random Forest Benchmark",
    df["RF_LABEL"],
    df["RF_PRED"]
)

print("\n[PERFORMANCE SUMMARY COMPLETED]")
```



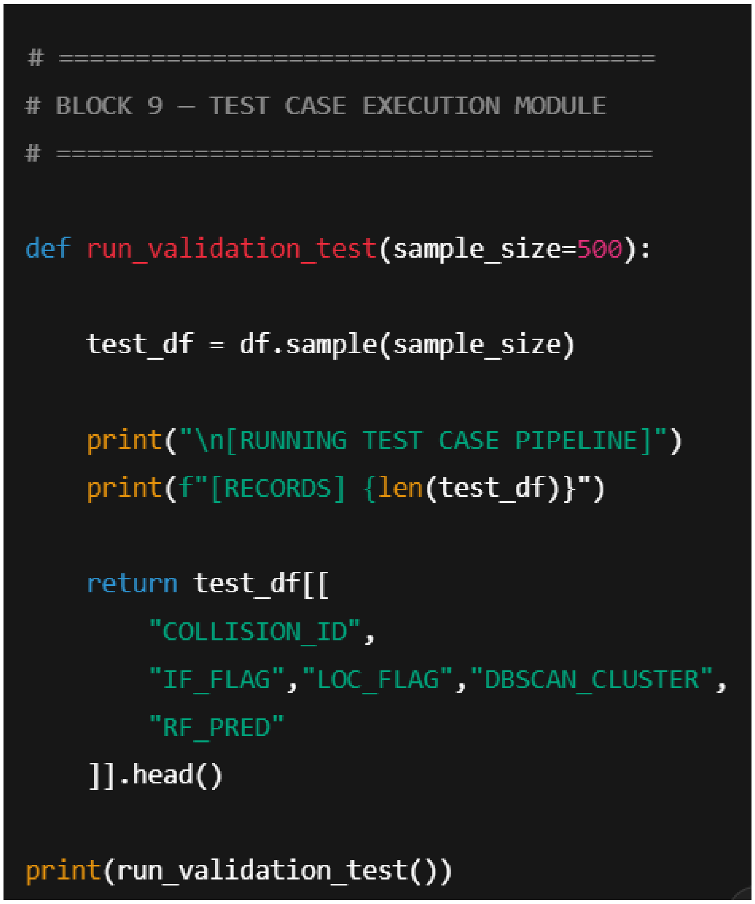
```
perf_results["LOC"] = evaluate(
    "Local Outlier Cluster (LOC)",
    df["RF_LABEL"],
    (df["LOC_FLAG"] == -1).astype(int)
)

perf_results["DBSCAN"] = evaluate(
    "DBSCAN",
    df["RF_LABEL"],
    (df["DBSCAN_CLUSTER"] == -1).astype(int)
)

perf_results["Random Forest"] = evaluate(
    "Random Forest Benchmark",
    df["RF_LABEL"],
    df["RF_PRED"]
)

print("\n[PERFORMANCE SUMMARY COMPLETED]")
```

8. Visualisation + Deployment (Streamlit Placeholder)



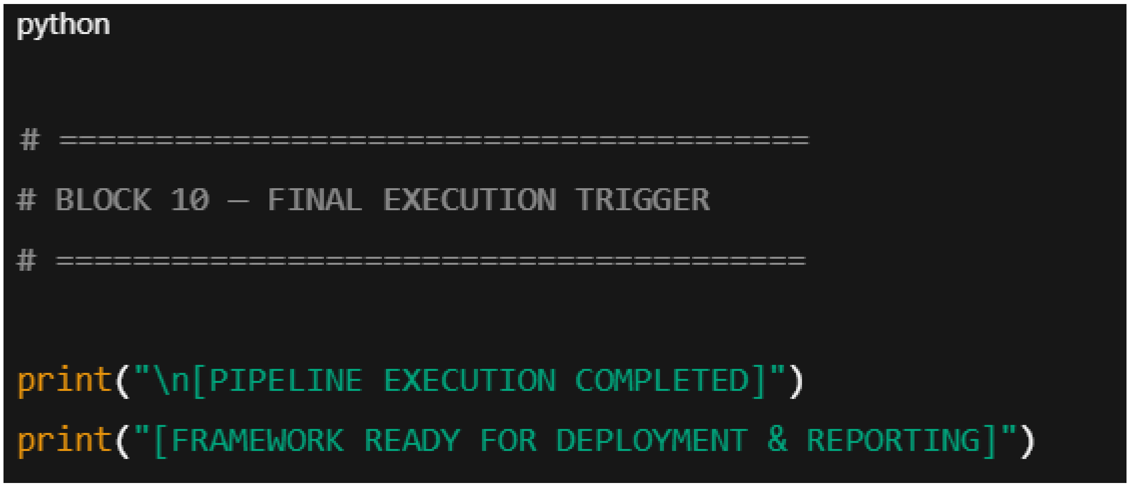
```python
# ==================================================
# BLOCK 8 — VISUALISATION / STREAMLIT PLACEHOLDER
# ==================================================

print("\n[DEPLOYMENT PLACEHOLDER] — To be enabled in Streamlit UI")

plt.figure()
plt.scatter(df["LATITUDE"], df["LONGITUDE"])
plt.title("Collision Map — Placeholder Visual")
plt.show()
```

9.Visualisation + Deployment (Streamlit Placeholder)



```python
# =======================================
# BLOCK 9 — TEST CASE EXECUTION MODULE
# =======================================

def run_validation_test(sample_size=500):

    test_df = df.sample(sample_size)

    print("\n[RUNNING TEST CASE PIPELINE]")
    print(f"[RECORDS] {len(test_df)}")

    return test_df[[
        "COLLISION_ID",
        "IF_FLAG","LOC_FLAG","DBSCAN_CLUSTER",
        "RF_PRED"
    ]].head()

print(run_validation_test())
```

10. Final Execution Trigger



```python
python

# =======================================
# BLOCK 10 — FINAL EXECUTION TRIGGER
# =======================================

print("\n[PIPELINE EXECUTION COMPLETED]")
print("[FRAMEWORK READY FOR DEPLOYMENT & REPORTING]")
```