Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Review On Data Classification And Related Concern Technics

Le Cuong [1*]

[1*]Electric Power University, Hoang Quoc Viet Str., Bac Tu Liem District, Hanoi, Vietnam, Email: cuongle@epu.edu.vn[1]

ARTICLE INFO

Received: 15 Apr 2025
Revised: 28 May 2025
Accepted: 06 June 2025

ABSTRACT

The paper deal with the classifications of data: from unstructure to structure data. Also the data could be classified by the value it receives or the way we collect date. Another choice is once the data is available, storage, security and privacy, and updating the data at the end of its life cycle also need to be considered. Data management are also briefly described.

Keywords: Classifications data; Data preprocessing; Data management;

1.0 INTRODUCTION TO DATA CLASSIFICATIONS

Data, as a central asset in the digital era, exhibits considerable diversity in structure, format, and analytical relevance. While statistics and machine learning are the two principal disciplines dedicated to data analysis, numerous methodologies have emerged from domain-specific requirements, particularly in disciplines such as econometrics, bioinformatics, and geoinformatics. For instance, econometrics - an applied branch of economics - was developed to analyze data often collected in the form of panel datasets, characterized by both cross-sectional and temporal dimensions.

In data science, data serves as the fundamental substrate for modeling, interpretation, and informed decision-making. The exponential growth of data across domains such as finance, healthcare, media, education, and e-commerce has introduced a critical need for systematic approaches to data collection, transformation, and governance. A comprehensive understanding of the various types and structures of data is thus a prerequisite for extracting meaningful insights and unlocking latent value.

Processed data - defined as the input available prior to model construction - typically assumes a numerical representation. For example, an image with dimensions 800×600 may be converted into a vector of 480,000 pixel intensity values. Similarly, binary variables (e.g., Yes/No) are encoded numerically (e.g., 1/0), and textual data may be represented through term frequency matrices or word embeddings to enable quantitative analysis. To facilitate such transformations and promote analytical efficiency, data must be classified based on shared characteristics such as structural form, temporal or spatial attributes, or measurement type.

Data classification refers to the systematic organization of raw data into predefined categories according to common features. This process enables the conversion of unstructured data into structured formats, improving accessibility and interpretability for downstream analysis.

1.1 Types of data classification

Data can be classified along several dimensions; however, one of the most widely adopted frameworks categorizes data into **structured**, **semi-structured**, and **unstructured** types. This typology is particularly relevant within the context of **Big Data**, which is defined by five key attributes - commonly referred to as the **5Vs**: *Volume*, *Velocity*, *Variety*, *Veracity*, and *Value*. Among these, **variety** captures the heterogeneous nature of data formats and is the primary axis for structural classification.

- **Structured Data** is highly organized and adheres to predefined schemas, typically stored in relational databases. Data is arranged in rows and columns and can be accessed using structured query language (SQL). This form of data is well-established, mature, and supports robust indexing, transactional control, and relational operations. Relational database management systems (RDBMS) such as MySQL and PostgreSQL are standard tools for managing structured data.

- **Semi-Structured Data** exhibits partial organization, lacking the rigidity of relational models while still retaining identifiable structure elements such as tags or keys. It is not constrained to tabular formats but may be parsed and queried with specialized methods. Common examples include XML, JSON, and data stored in

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

NoSQL databases. Semi-structured data provides a flexible alternative in contexts requiring schema evolution or hierarchical relationships.

- **Unstructured Data** lacks a predefined data model and cannot be readily processed by conventional relational systems. It encompasses a wide range of formats such as plain text, PDF documents, multimedia files (images, audio, video), and system logs. Due to its complexity and volume, unstructured data is often stored in distributed file systems or data lakes, and processed using advanced techniques in natural language processing, image recognition, or deep learning.

Comparison of Structured, Semi-Structured, and Unstructured Data

Property	Structured Data	Semi-Structured Data	Unstructured Data
Transaction Management	Mature transaction support with concurrency control	Adapted from DBMS, limited transaction maturity	No transaction or concurrency management
Version Management	Fine-grained versioning (tuples, rows, tables)	Possible versioning at tuple or graph level	Versioned as a whole object
Underlying Technology	Relational database systems (e.g., RDBMS)	XML, RDF, JSON, and NoSQL-based systems	Binary or character-based storage systems
Query Performance	Supports complex relational joins and indexing	Supports semi-structured queries over annotated nodes	Limited to keyword-based or full-text queries
Schema Flexibility	Strict schema enforcement; less flexible	Moderately flexible with partial schema	Highly flexible; absence of formal schema
Scalability	Limited horizontal scalability	Improved scalability over structured systems	High scalability, often used in Big Data systems
Robustness	Highly robust, standardized technologies	Emerging technologies; evolving robustness	Depends on implementation; often ad hoc

In conclusion, understanding the classification of data is essential for effective data management, preprocessing, and modeling in data science. The structural distinction between structured, semi-structured, and unstructured data informs not only the choice of storage and retrieval mechanisms but also influences the selection of analytical techniques. As data continues to evolve in complexity and volume, the ability to systematically classify and manage diverse data types remains a foundational competency in both academic research and real-world applications.

1.2 Classification by the value a variable can take

In data science and statistical modeling, variables are often categorized based on the nature and type of values they assume. This classification plays a crucial role in determining the appropriate analytical techniques and modeling approaches. The following sections present common types of variables according to the values they receive.

a. Continuous (Interval) Variables

Continuous variables are among the most common types in statistical analysis. These variables can take any value within a given range on the number line. Typical examples include height, weight, temperature, and revenue, where values are measurable, ordered, and theoretically infinite within certain bounds. For instance, an individual's weight can be 62.5 kg or 62.55 kg, indicating its continuous nature.
Continuous variables are typically modeled using distributions such as the normal or uniform distribution. In regression modeling - especially linear regression - it is assumed that the response variable is continuous.

Special cases of continuous variables:

- **Rate variables:** These variables are bounded within the interval [0, 1], such as conversion rates or mortality rates. Although often treated as continuous, specialized models like beta regression are more suitable when the bounded nature of the variable is critical.
- **Circular/Directional variables:** These represent angular data (e.g., 0 to 360 degrees) and are prevalent in fields such as meteorology and biology. Conventional statistical operations may be inappropriate; for example, averaging 10° and 350° yields 180°, which misrepresents their actual proximity. These variables are better analyzed using polar coordinates and directional statistics.

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

b. Binary variables

Binary variables take only two values, typically coded as 0 and 1, representing mutually exclusive outcomes such as Yes/No or True/False. In modeling, when the binary variable acts as a response, the task becomes a binary classification problem, where algorithms such as logistic regression or support vector machines (SVM) are commonly employed.

When binary variables are predictors, they act as group indicators. For example, a binary gender variable (Male/Female) divides the dataset into two subgroups. Combining multiple binary variables further partitions the data, and analysis can proceed through methods such as Analysis of Variance (ANOVA) or Analysis of Covariance (ANCOVA), which are conceptually aligned with linear regression.

c. Multinomial variables

Multinomial variables generalize binary variables to more than two unordered categories. For instance, political affiliation could be categorized as Democrat, Republican, or Independent. Although these levels may be numerically encoded (e.g., 1, 2, 3), the numbers are labels rather than ordinal values and should not be subjected to arithmetic operations.

In modeling, when multinomial variables serve as the response variable, the problem becomes a multi-class classification task. When used as predictors, they are typically one-hot encoded - represented as multiple binary variables - to avoid implying any order. A common challenge with multinomial or binary variables is class imbalance, where one class dominates. To mitigate this, techniques such as undersampling, oversampling, modifying loss functions, or employing algorithms like SVMs and ROC-AUC evaluation are applied.

d. Count variables

Count variables represent the number of occurrences of an event within a fixed period, such as the number of website visits or disease cases. Though discrete and non-negative, count variables are often treated as continuous predictors due to their ordinal nature.

When used as response variables, traditional linear models may yield negative or non-integer predictions, which are inappropriate. Instead, models such as Poisson regression or negative binomial regression are employed. These models account for skewed distributions and constraints on non-negativity. For rare event modeling - such as aviation accidents - specialized techniques are used due to the low frequency and high consequence of outcomes.

e. Ordinal variables

Ordinal variables represent categorical data with an inherent order but undefined distance between categories. A common example is the Likert scale (e.g., rating satisfaction from 1 to 10). As predictors, these variables may be treated as continuous. However, when used as response variables, their hybrid nature requires specialized treatment.

Latent variable models are commonly applied, where an unobservable continuous variable is assumed to underlie the ordinal responses. For instance, satisfaction could be modeled as a latent variable ranging from 0 to 10, where observed Likert scores correspond to intervals on this continuum. The model estimates cutoff points that map latent values to ordinal categories.

In summary, categorizing variables by the values they receive allows for the appropriate selection of models and techniques in data science. The next section will address how data can also be classified based on collection methods and observational structure.

1.3 Classification according to the nature of the collection process

In addition to the values that variables can assume, data can also be classified based on how it is collected. This classification influences assumptions about independence, temporal ordering, spatial correlation, and measurement completeness, all of which are critical to appropriate model selection and interpretation.

a. Cross-sectional data

Cross-sectional data refers to data collected at a single point in time, or data assumed to remain constant over the temporal or spatial domain of the study. This type of data is commonly used when the primary interest lies in the relationships between variables rather than their evolution over time. For example, examining the association between income and education level across individuals, without regard to how either variable changes over time, would utilize cross-sectional data.

A defining feature of cross-sectional data is the **independence and exchangeability** of observations. The order of data collection does not influence the results, and the identifiers (e.g., person 1 or person 1000) are not analytically distinguishable. Cross-sectional datasets typically do not include temporal markers and are

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

often modeled using standard regression and correlation techniques under the assumption of independently and identically distributed (i.i.d.) samples.

b. Time series and Panel data

In contrast, **time series data** consist of observations collected at regular time intervals. These data reflect temporal dynamics, where the value of a variable at one point is influenced by its past values. For instance, annual GDP data represent the aggregate economic output over a full calendar year, not merely the instantaneous value at year-end.

Time series data are inherently ordered and exhibit **autocorrelation**, making them unsuitable for models that assume independence. Models must account for temporal structure through techniques such as autoregressive models, seasonal decomposition, and stationarity adjustments. Applications include forecasting stock prices, GDP, or sales volume.

Panel data (also known as longitudinal data) integrate both cross-sectional and time series dimensions. Here, multiple entities are observed across multiple time periods. Panel data analysis accounts for both **time-invariant (fixed effects)** and **time-variant (random effects)** factors. This structure is prevalent in econometrics and allows for the estimation of both within-entity and between-entity variation.

c. Functional data

Functional data refer to observations recorded over a continuous domain, typically time, where each data point is a function (e.g., a curve or surface) rather than a scalar. Unlike time series, where values are recorded at discrete intervals, functional data involve **dense or continuous sampling**, yielding smooth trajectories. For example, air pollution levels measured every few seconds form a continuous curve rather than a set of isolated measurements.

An important feature of functional data is that the **measurement times may be irregular or stochastic**, requiring methods that accommodate variable timing. Functional Data Analysis (FDA) often employs basis functions (e.g., splines, Fourier series) to estimate underlying curves, making it particularly relevant in applications such as biomechanics, environmental monitoring, and handwriting recognition.

d. Spatial and Spatio - temporal Data

Spatial data are observations associated with geographic or physical locations. These data exhibit spatial correlation - observations closer in space tend to be more similar. Applications include environmental science (e.g., pollutant concentration), epidemiology, and real estate economics. Spatial data may be georeferenced (using coordinates) or defined through topological relationships (e.g., adjacency matrices).

Spatio-temporal data combine spatial and temporal dimensions. Examples include satellite images over time, climate data, and disease outbreak monitoring. These datasets often require models that simultaneously account for spatial autocorrelation and temporal dependencies. Advanced modeling approaches include **spatial-temporal autoregressive models** and **stochastic point process models**, particularly useful for predicting the occurrence of rare spatial-temporal events or changepoints.

e. Censored and missing data

Censored data arise when the full value of a measurement is not observed, but partial information is available. A common example is time-to-event data in survival analysis. For instance, if a clinical study tracks patient survival over 20 weeks and a patient survives the entire period, the exact time of death is unknown - it is only known to exceed 20 weeks. Censoring also occurs with detection limits in measurements, such as pollutant concentrations reported as " < 0.001" or " > 100." Models for censored data include **Tobit regression** and **survival models**, which must be used in lieu of omitting such observations to avoid biased estimates.

Missing data, a broader category, can occur due to various mechanisms:

- **MCAR (Missing Completely at Random):** Missingness is unrelated to any observed or unobserved data.
- **MAR (Missing at Random):** Missingness is related to observed data but not to the missing values themselves.
- **MNAR (Missing Not at Random):** Missingness depends on both observed and unobserved data.

Each mechanism necessitates different handling strategies, such as imputation or sensitivity analysis. Ignoring non-random missingness can lead to biased models and incorrect inferences.

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

f. Complex sample designs and Meta-data

Data derived from **complex sampling designs**, such as stratified, clustered, or multi-stage sampling, require analytical adjustments to account for unequal probabilities of selection. These designs are common in large-scale surveys (e.g., national censuses). Weighting, post-stratification, and survey-specific variance estimation techniques must be applied to ensure valid population-level inferences.

Repeated measurements introduce another layer of complexity, particularly in biomedical or experimental studies where variables (e.g., blood pressure or genetic markers) are measured multiple times. Simple averaging can obscure intra-individual variability and lead to biased parameter estimates. Specialized models, such as **measurement error models** or **mixed-effects models**, are employed to account for within-subject variation.

Meta-data refer to datasets aggregated from multiple studies, surveys, or experiments. Meta-analyses aim to synthesize findings across heterogeneous sources, often requiring the use of **fixed-effects** or **random-effects models** to account for study-level variability. Proper meta-analysis must consider variations in sample design, data collection methods, and population characteristics across the included studies.

Together, these classifications based on data collection context provide a foundation for selecting appropriate modeling frameworks and interpreting results accurately in applied data science. It is also worth noting that, in the context of information security, data is commonly classified into categories such as public, restricted, sensitive, and confidential.

2.0 DATA PREPROCESSING

In the realm of data science, the **quality and structural integrity of input data** are critical determinants of the **performance, robustness**, and **validity** of both analytical and machine learning models. Real-world datasets are seldom pristine; they often contain **missing values**, **noise**, **inconsistencies**, and may exhibit **high dimensionality** or **scalability issues** that exceed the processing capabilities of traditional algorithms. Accordingly, **data preprocessing** constitutes a foundational phase within the data pipeline, aiming to transform raw data into a structured, consistent, and analytically tractable form. This section outlines key preprocessing techniques commonly adopted in modern data science workflows.

2.1 Data cleaning

Data cleaning is the initial and one of the most critical steps in the preprocessing workflow. It is essential for mitigating errors, addressing inconsistencies, and preserving the validity of subsequent analysis.

- **Handling missing values** involves strategies such as the removal of incomplete records or the imputation of missing entries using statistical methods (e.g., mean, median) or more advanced model-based techniques that leverage relationships among variables.
- **Noise removal and outlier detection** are necessary to maintain the integrity of data distributions. Techniques such as the **Interquartile Range (IQR)**, **Z-score analysis**, and **anomaly detection algorithms** are frequently employed.
- **Standardization of units and formats**, including the normalization of date/time values and measurement units, ensures dataset homogeneity and facilitates reliable analysis.

2.2 Data transformation

Data transformation enhances the dataset's suitability for modeling by converting raw data into forms more compatible with algorithmic requirements.

- **Categorical encoding** converts qualitative variables into numeric representations. Techniques such as **label encoding** (assigning ordinal integers) and **one-hot encoding** (binary vectors representing category membership) are foundational for algorithms requiring numerical inputs.
- **Normalization and scaling methods**, including **Min-Max normalization** and **Z-score standardization**, align feature scales to improve model convergence and stability—particularly important for algorithms sensitive to feature magnitude (e.g., **Support Vector Machines (SVM)**, **K-Nearest Neighbors (KNN)**).
- **Feature extraction** involves deriving informative variables from raw data. For instance, in **natural language processing (NLP)**, **term frequency–inverse document frequency (TF-IDF)** is used to represent textual data numerically. In **image processing**, features may include edges, contours, or color histograms.

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2.3 Data integration

Data integration addresses the challenge of **heterogeneous data sources**, such as relational databases, web APIs, structured files, and IoT sensor data, by consolidating them into a unified dataset.

- This process includes **schema alignment**, **data harmonization**, and **type normalization**, ensuring consistency across formats and structures.
- **Deduplication and synchronization** are used to detect and resolve redundant or conflicting records.

Advanced techniques such as **fuzzy matching** are employed when unique identifiers are ambiguous or absent.

2.4 Dimensionality reduction

High-dimensional datasets can hinder model interpretability and increase the risk of **overfitting** and **computational inefficiency**. Dimensionality reduction techniques aim to retain essential information while reducing the number of variables.

- Methods such as **Principal Component Analysis (PCA)**, **t-Distributed Stochastic Neighbor Embedding (t-SNE)**, and **Autoencoders** are used to project data into lower-dimensional spaces that capture the primary variance or structure.
- These techniques are not only valuable for improving model performance but also facilitate effective **data visualization**, especially in exploratory phases.

2.5 Exploratory data analysis

Exploratory Data Analysis (EDA) is a diagnostic and hypothesis-generating stage that provides insights into data structure prior to formal modeling.

- EDA employs **descriptive statistics** (e.g., mean, standard deviation, frequency distributions) and **visual tools** (e.g., histograms, boxplots, scatter plots) to identify patterns, trends, and potential anomalies.
- Through EDA, practitioners can detect **hidden structures**, such as **clusters**, **nonlinear relationships**, or **outliers**, which inform **feature engineering**, **model selection**, and **preprocessing strategy**.

Data preprocessing is an indispensable component of the **data science lifecycle**, requiring the integration of **statistical reasoning**, **algorithmic rigor**, and **domain-specific expertise**. A thoroughly preprocessed dataset enhances **model accuracy**, **generalizability**, and **interpretability -** all critical for valid and actionable insights. Therefore, **rigorous investment in preprocessing** is not merely preparatory but essential to the success and scientific soundness of any data-driven initiative.

3.0 DATA MANAGEMENT IN THE ERA OF DIGITAL TRANSFORMATION

In the context of the contemporary data explosion, **data management** has evolved beyond its traditional role as a purely technical function to become a **strategic capability** integral to the operations of organizations, enterprises, and professionals in data science, information technology, and business analytics. While data is inherently valuable, its true utility is realized only through **proper storage**, **secure handling**, **efficient processing**, and **timely utilization**. This section provides a systematic overview of the core components of modern data management: data storage architectures, data security and privacy, lifecycle management, and enabling technologies.

3.1 Data storage

Data storage forms the **foundational layer** of the data infrastructure, directly affecting information retrieval efficiency, system scalability, and data accessibility. The appropriate choice of storage solution is largely determined by the **nature and structure of the data** involved:

- **Relational databases** (e.g., *MySQL*, *PostgreSQL*) employ a tabular schema with clearly defined relationships and constraints. These systems support efficient data querying via **Structured Query Language (SQL)** and are widely deployed in transaction-oriented environments where **data integrity** and **consistency** are paramount.
- **Non-relational (NoSQL) databases** (e.g., *MongoDB*, *Cassandra*) are optimized for semi-structured or unstructured data formats such as JSON documents, system logs, and free text. Their **schema-less design** and **horizontal scalability** make them ideal for distributed, high-throughput, and real-time applications.
- **Data warehouses** (e.g., *Amazon Redshift*, *Google BigQuery*) are designed for **analytical processing**. These platforms organize data in dimensional schemas optimized for **complex query execution** over large volumes of historical data.

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- **Data lakes** (e.g., *Hadoop*, *Azure Data Lake*) serve as **flexible repositories** for raw data in its native format. Supporting both structured and unstructured data, data lakes enable schema-on-read access patterns and are widely used in **machine learning pipelines** and **exploratory analytics**.

3.2 Data security and privacy

As data becomes a central and monetizable asset in digital ecosystems, **ensuring its confidentiality, integrity, and compliance** with legal frameworks is critical.

- **Encryption and access control** mechanisms are fundamental to protecting sensitive information. Encryption ensures that data remains unreadable without proper decryption keys, while access controls define user roles and permissions for data access, modification, and deletion.
- **Regulatory compliance** is mandatory for any organization that handles personal or sensitive data. Legal frameworks such as the **General Data Protection Regulation (GDPR)** in the European Union and the **Health Insurance Portability and Accountability Act (HIPAA)** in the United States prescribe standards for data transparency, subject rights, breach notification, and accountability.

3.3 Data lifecycle management

Data Lifecycle Management (DLM) encompasses the governance and control of data assets throughout their entire lifespan, from creation to deletion.

- The **standard data lifecycle** includes stages such as **data collection**, **storage**, **processing**, **analysis**, and **disposal**. Each phase requires well-defined procedures aligned with organizational goals and compliance requirements.

- **Metadata management** supports transparency by recording data origin, structural attributes, and version history. **Data lineage** allows traceability of data transformations, which is essential for auditing, reproducibility, and model interpretability.

Effective DLM contributes to **cost reduction**, enhances **data accuracy and availability**, and strengthens institutional readiness for **data-driven decision-making**.

3.4 Enabling Technologies and Tools

The execution of robust data management strategies relies heavily on **specialized tools and platforms** that automate and optimize critical processes:

- **ETL (Extract, Transform, Load) tools** such as *Talend* and *Apache NiFi* facilitate the migration of data from heterogeneous sources to centralized systems. These tools support both **batch** and **real-time processing**, ensuring data consistency across the pipeline.
- **Data governance platforms** (e.g., *Collibra*, *Informatica*) provide centralized mechanisms for **policy enforcement**, **data quality assurance**, **regulatory compliance**, and **metadata management**. They also enable **role-based access control**, **anomaly detection**, and **data stewardship protocols**.

Data management has become a **multidisciplinary and strategic function** that underpins the effective use of data across modern organizations. From **architecting scalable storage solutions**, **enforcing data security**, and **governing lifecycle operations**, to **leveraging advanced tools**, each aspect plays a vital role in ensuring data **utility**, **trustworthiness**, and **value extraction**. As data environments become increasingly complex and dynamic, an organization's capability to manage data effectively will serve as a **key differentiator** and a source of **sustainable competitive advantage** in the digital economy.

4.0 THE DATA ANALYSIS PROCESS IN DATA SCIENCE

In the field of data science, the data analysis process is systematically organized into a series of **interdependent stages**. This structured approach is essential for ensuring **accuracy, consistency, interpretability**, and the **practical utility** of analytical outcomes and predictive models. Each stage plays a distinct role in transforming raw data into actionable insights and is supported by techniques drawn from **statistics, computer science**, and **domain-specific knowledge**. The following sections provide a comprehensive overview of the principal phases in a modern data analysis pipeline.

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.1 Data Collection

The data analysis workflow begins with **data acquisition**, which involves gathering relevant data from diverse sources. In contemporary data-driven environments, common sources include:

- **Web scraping**, which involves the automated extraction of content from websites using tools such as *BeautifulSoup* or *Scrapy*.
- **Application Programming Interfaces (APIs)**, which facilitate structured and real-time data exchange between systems—for example, APIs for financial markets, social media platforms, or weather services.
- **Internal databases**, such as Customer Relationship Management (CRM) systems, Enterprise Resource Planning (ERP) platforms, or transactional datasets stored in relational or NoSQL databases.

Effective data collection demands careful consideration of **data relevance, structure, access permissions**, and **data privacy compliance**. Ethical and legal regulations - such as the **General Data Protection Regulation (GDPR) -** must be observed when collecting personal or sensitive information.

4.2 Data preprocessing

Raw data is often incomplete, noisy, and inconsistent. **Data preprocessing** prepares the dataset for reliable analysis and consists of several key sub-processes:

- **Data cleaning**, which includes handling missing values (through imputation or deletion), detecting and addressing outliers using statistical techniques (e.g., IQR, Z-score), and correcting formatting errors.
- **Data transformation**, which converts raw inputs into more analytically suitable forms. Common tasks include normalization (e.g., Min-Max scaling), standardization (e.g., Z-score transformation), encoding categorical variables (e.g., label or one-hot encoding), and constructing derived features through feature engineering.
- **Data integration**, which consolidates datasets from multiple sources. This process involves **schema alignment**, **deduplication**, and resolving structural conflicts to ensure consistency.

Preprocessing not only enhances **data quality** but also supports **model generalizability**, ensuring the robustness of downstream analytical tasks.

4.3 Data analysis and modeling

With the data preprocessed, the analytical phase can commence. This stage typically involves both **exploratory** and **predictive** components:

- **Descriptive and inferential statistical analysis** is used to understand data distributions, detect patterns, and evaluate relationships. Methods include measures of central tendency, variance analysis, hypothesis testing, and correlation matrices.
- **Machine learning modeling** aims to generate predictive insights and discover latent structures. Techniques can be broadly categorized as:
  - **Supervised learning** (e.g., linear regression, logistic regression, decision trees, random forests, support vector machines), which is used for classification and regression tasks involving labeled data.
  - **Unsupervised learning** (e.g., k-means clustering, hierarchical clustering, Principal Component Analysis), which is used to identify hidden patterns or reduce dimensionality in unlabeled data.
- **Model selection and validation** are critical for ensuring reliability. Techniques such as **cross-validation**, **grid search**, and performance metrics (e.g., **accuracy**, **precision**, **recall**, **F1-score**) are used to evaluate model effectiveness.

This analytical phase lays the foundation for **evidence-based decision-making** and informs downstream processes such as deployment, policy design, or business strategy alignment.

4.4 Summary of the process

In summary, the data analysis process in data science is a **methodical and iterative framework** that transforms raw data into valuable insights through a sequence of **well-defined and interrelated stages**. From the acquisition of relevant data, through rigorous preprocessing, to the application of sophisticated analytical and modeling techniques, each phase contributes to the overall **integrity, interpretability**, and **utility** of the analytical outcome. A deep understanding and meticulous implementation of these steps is essential for the success of any data science initiative—both in academic research and industrial practice.

Journal of Information Systems Engineering and Management
2025, 10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.0 CONCLUSION

Data is an indispensable resource in the digital age. The ability to **understand, classify, and manage data** -in conjunction with effective **storage, security, preprocessing, and modeling -** is a foundational competence in data science. Every stage of the data lifecycle, from cleaning and integration to machine learning and decision support, contributes to the value extraction process. As technological advancement continues to accelerate, **data literacy** and **data management capabilities** are becoming increasingly essential - not only within data science but across all disciplines and industries engaged in digital transformation.

FUNDING

This research is funded by Electric Power University under research 2025.

REFERENCES

[1] Han, J., Kamber, M., & Pei, J., "Data Mining: Concepts and Techniques (3rd ed.)", *Morgan Kaufmann*, 2011.
[2] Shmueli, G., Bruce, P. C., & Patel, N. R., "Data Mining for Business Analytics", *Wiley,* 2016
[3] Rahm, E., & Do, H. H., "Data Cleaning: Problems and Current Approaches", *IEEE Bulletin on Data Engineering*, 7, 182-185, 2003.
[4] Aggarwal, C. C., "Data Mining: The Textbook", *Springer*, 2015.
[5] Batini, C., Scannapieco, M., "Data Quality: Concepts, Methodologies and Techniques", *Springer,* 2015.
[6] Kimball, R., & Ross, M., "The Data Warehouse Toolkit (3rd ed.)", *Wiley*, 2013.
[7] James, G., Witten, D., Hastie, T., & Tibshirani, R., "*An Introduction to Statistical Learning*", *Springer*, 2013.
[8] Kuhn, M., & Johnson, K., "Applied Predictive Modeling", *Springer*, 2013.
[9] Kitchin, R., "The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences", SAGE, 2014.
[10] Provost, F., & Fawcett, T., "Data Science for Business", *O'Reilly Media*, 2013.
[11] Han, J., Kamber, M., & Pei, J., "Data Mining: Concepts and Techniques (3rd ed.)", *Morgan Kaufmann*, 2011.
[12] Shmueli, G., Bruce, P. C., & Patel, N. R., "Data Mining for Business Analytics", *Wiley*, 2016.
[13] Mayer-Schönberger, V., & Cukier, K., "Big Data: A Revolution That Will Transform How We Live, Work, and Think", Houghton Mifflin Harcourt. 2013.
[14] Davenport, T. H., & Patil, D. J., "Data Scientist: The Sexiest Job of the 21st Century", *Harvard Business Review*., 2012.