EasyChair Preprint

№ 15698

Data Quality Management for Real-World Graduation Prediction

Hong-Duyen Nguyen-Pham, Khoa Tan Vo, Thu Nguyen,
Tu-Anh Nguyen-Hoang, Ngoc-Thanh Dinh and Hong-Tri Nguyen

January 10, 2025

Data Quality Management for Real-World Graduation Prediction

Hong-Duyen Nguyen-Pham[1,2], Khoa Tan VO[1,2], Thu Nguyen[1,2], Tu-Anh Nguyen-Hoang[1,2], Ngoc-Thanh Dinh[3], Hong-Tri Nguyen[4]
[1]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]The Industrial University of Ho Chi Minh City, Vietnam
[4]Aalto University, Finland

***Abstract*—The rapid growth of diverse and multi-sourced data has rendered traditional data storage models inadequate to handle the sheer volume and complexity. Data Lakes, which store all raw data and all data versions in an easily accessible format, are well-suited for deep data analysis and valuable insights discovery. However, the quality of this data is not guaranteed, raising the question of how to utilize this vast repository effectively. Our research proposes a four-step data quality management process profile, implement, monitor, and improve to oversee and ensure data usability within a data lake. This process employs five commonly used evaluation criteria: accuracy, completeness, consistency, uniqueness, and timeliness. Our study focuses on higher education data, an area that has not been extensively explored in previous research, using real-world data from a university's computer science department. The application context is managing the quality of input data for a machine-learning model that predicts student graduation outcomes. Two advanced boosting machine learning models, LightGBM and CatBoost, are employed, resulting in a 5% improvement in performance. Our research aims to provide a comprehensive solution for assessing data quality in higher education, saving significant time, effort, and cost while enhancing the reliability of data utilization from data lakes.**

***Index Terms*—educational data mining, data quality management, graduation prediction, big data**

I. Introduction

The fundamental question posed is why we should care about data quality. A study conducted by a research group in Germany and published in Procedia Manufacturing in 2019 revealed that not all organizations assess data quality before making decisions [1]. Another study in 2021 affirmed that data quality still lacks adequate attention, despite its crucial role in data analysis, as the quality of analytical outcomes directly depends on the quality of underlying data [2]. High-quality data forms the basis for sound and effective decision-making [3]. Conversely, low-quality data can lead to erroneous decisions, resulting in financial losses and damage to an organization's reputation [4]. When such data is used in predictive models and alerts, it can reduce the accuracy of the results.

Our research focuses on educational data, particularly student academic and extracurricular activities within a university context. The data currently faces several challenges, including inconsistency, where formats, units, or collection methods vary. For instance, dates of birth might be stored differently across departments. Additionally, there are issues of incompleteness, where crucial information is missing, and inaccuracy, where data may be distorted or outdated. Timeliness problems also arise, as data may not reflect the current situation accurately. These issues often stem from changes in input and storage formats over time. Moreover, data duplication occurs when information is entered multiple times or copied without verification, undermining the reliability and effectiveness of data-driven decisions.

A proposed solution to address data quality issues involves developing a data quality assessment framework. A data quality assessment framework is a method to evaluate and measure the quality of data [5]. It provides a structure and process to identify data quality issues and propose measures to improve data quality. Typically, the assessment process includes analysis, evaluation, improvement, and monitoring stages. Initially, research focuses on constructing general frameworks, followed by their application to specific data types and fields. Recent trends emphasize automating data quality management, particularly the time-consuming data cleansing stage, which can consume 60-80% of a data science project's time [6]. Automating this process saves time, effort, and costs while ensuring high-quality data throughout its lifecycle. The criteria for assessing data quality form the core of such frameworks, with multiple criteria selected based on data characteristics and organizational requirements [7].

A limitation of previous studies is their tendency to focus solely on evaluating data quality without delving into its improvement, or if they do, failing to assess the effectiveness of those improvement measures. We propose to evaluate the impact of data quality improvement methods through the accuracy of predictive models, as well as provide predictions regarding students' graduation probabilities. This will assist educational managers in intervening promptly to increase on-time graduation rates and overall outcomes.

The article is structured as follows: Section II provides a summary of related research on data quality frameworks. In Section III, we analyze our higher education dataset. Section IV presents two main aspects of our approach, covering data quality management and graduation outcomes prediction. The experimental procedures and results are discussed in Section V. Finally, we wrap up in Section VI with concluding remarks and prospects.

II. Related work

Table I compares common data quality assessment frameworks, each providing definitions for a set of criteria used to evaluate data quality. Additionally, research highlights two main methods for measuring data quality: subjective assessment and objective assessment. Subjective assessment primarily uses survey questionnaires about the experiences of individuals directly interacting with the data to identify issues related to current data quality. Objective assessment relies on qualitative and quantitative measures specific to the dataset's characteristics. The objective assessment method, based on criteria, is preferred due to its structured approach, while subjective assessment is less commonly used because it incurs higher costs and requires the cooperation of multiple stakeholders. Frameworks such as TDQM [9], TIQM [10], HIQM [11], CDQ [12], COLDQ [13], DQAF [14], and TBDQ [15] were proposed in the initial stages to define the basic concepts and components of a data quality framework. Consequently, the number of criteria defined is quite diverse, and each criterion has multiple different definitions.

The following frameworks focus on the quantitative and qualitative implementation of selected criteria for datasets in specific domains. The methods used are quite varied and depend on the unique characteristics of the dataset under study. DQF4CT [16] defined specific data quality problems that impact classification models. VIoTF [17] was a framework called Valid.IoT for enhancing IoT data quality. PPF [18] cared about data quality in the pre-processing stage. IDQ-MDM [20] focused on maintaining the quality of master data and suggesting a data quality management process. As a result, the number of commonly selected criteria has become more clearly defined, including accuracy, completeness, consistency, uniqueness, validity, and timeliness. The most commonly assessed type of data is structured data. However, there have been no studies on data quality assessment in higher education or evaluations of data quality on university datasets.

Predicting graduation outcomes is a crucial research area in higher education. This issue has attracted the attention of policymakers, educators, and researchers in recent years. The time it takes for a student to complete a university program is influenced by various factors such as their prior educational background, academic performance at university, and involvement in social activities [23] [24] [25] [26]. These studies underscore the importance of data-driven methods in predicting graduation outcomes and offer insights for enhancing education quality.

Our research aims to deploy comprehensive data quality management and improve data quality in the context of educational data with a specific focus on predicting course completion times and graduation outcomes of students at a university.

III. Data Analysis

The education dataset is collected from various sources, from departments within a university. The data includes the following fields and is detailed in Table II.

- Student information: Full name, student ID, date of birth, gender, hometown, field of study, class, and other personal details.
- Lecturer information: Lecturer ID, date of birth, gender, hometown, position, academic title, academic degree, and department affiliation.
- Teaching activities: Assignment of classes to teach in each semester, list of grades for each student in each course.
- Research activities: Information on scientific research topics of lecturers.
- Extracurricular activities: Extracurricular performance scores of students summarized by semester.
- Financial activities: Information about tuition fees, fee waivers, fee extensions.

This study proposes the implementation of a data quality management process within a big data architecture. In addition to structured data tables within the system, the dataset includes discrete report files, text segments evaluating teaching quality, and weather data collected from sensors. All versions of the data are stored within the Data Lake, as depicted in Figure 1, which illustrates the diverse types of data that can be stored in it. With the increasing number of students and the need to diversify data sources and types, constructing a comprehensive profile of students will help the school better understand their learning and behavioral patterns. Consequently, timely interventions can be proposed to support students effectively, thereby enhancing the quality of education.

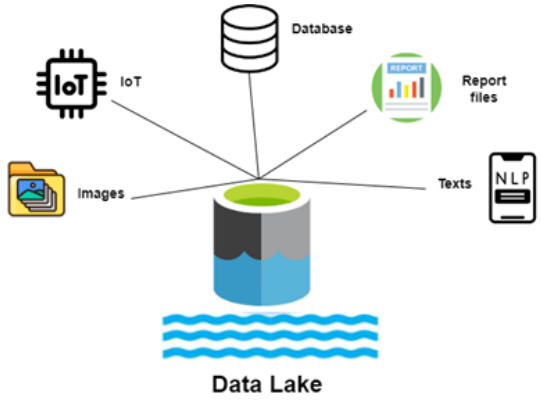

Fig. 1. The image illustrates various data types stored in their original form in the data lake

For the task of predicting graduation outcomes, we obtained a sub-dataset comprising detailed information about 6637 graduates across 14 training cohorts spanning from 2006 to 2019 within the field of information technology. These graduates underwent training programs ranging from 3.5 to 5 years. Outputs from this dataset encompass the classification of graduates into categories such as Excellent, Good, Fair, and Average. Table III shows the distribution of these labels, with the number of Excellent students being significantly lower than the others, resulting in an inherent imbalance. To preserve the real-world nature of the dataset, this imbalance was main-

TABLE I
OVERVIEW OF DATA QUALITY FRAMEWORKS, THEIR COMPONENTS, DIMENSIONS, AND TYPES OF DQ ASSESSMENT METHODS USED.

Framework	Year	Main components	DQ dimensions	Objective DQ	Subjective DQ
TDQM [9]	1998	Consideration of business rules and definition of data quality metrics	Accuracy, relevancy, reputation, timeliness, completeness, security	✓	✗
TIQM [10]	1999	User expectations and definition of data quality metrics	Completeness, accuracy, precision, non-duplication, accessibility, timeliness, integrity, usability	✓	✓
HIQM [11]	2006	Objective assessment through measurement algorithm suggested	Accuracy, completeness, consistency, timeliness	✓	✗
CDQ [12]	2008	User interviews and definition of data quality metrics for accuracy and currency	Structured: accuracy, completeness, currency Unstructured: currency, relevance, reliability	✓	✗
COLDQ [13]	2011	Consumer surveys and definition of various data quality metrics	Accuracy, consistency, completeness, currency, security, timeliness, relevance	✓	✓
DQAF [14]	2013	Definition of a set of data quality metrics for different types of measurement	Completeness, timeliness, validity, consistency, integrity	✓	✗
TBDQ [15]	2016	Survey questionnaire and simple ratio	Accuracy, completeness, consistency, timeliness	✓	✓
DQF4CT [16]	2018	The specific data quality issues that can impact classification tasks	Accuracy, completeness, relevance, consistency	✓	✗
VIoTF [17]	2018	A proposed framework called Valid.IoT for improving data quality in the Internet of Things (IoT)	Accuracy, timeliness, completeness, reliability	✓	✗
PPF [18]	2019	A framework for pre-processing data to improve its quality.	Completeness, validity, consistency,	✓	✗
HDQF-EF [19]	2021	A hybrid framework for data quality assessment in Environmental Footprint (EF) tools	Accuracy, completeness, uniqueness	✓	✓
IDQ-MDM [20]	2022	A proposed framework for maintaining data quality throughout the Master Data Management (MDM) implementation process	Accuracy, completeness, consistency, uniqueness, timeliness, validity	✓	✗
ISO/IEC 25012 [21]	2023	The application of the ISO/IEC 25012 framework for improving the quality of software vulnerability datasets	Accuracy, completeness, consistency, validity	✓	✗
RWDQF [22]	2024	A framework for assessing data quality in oncology research, specifically focusing on time to treatment discontinuation	Accuracy, completeness, timeliness	✓	✗
Our framework	2024	A proposed framework for assessing comprehensive higher education data quality	Accuracy, completeness, consistency, uniqueness	✓	✗

TABLE II
STATISTICAL TABLE OF INFORMATION OF DATA TABLES IN THE DATASET.

No	Table	Rows	Cols	Description
1	Student	17925	38	List of students from 2006-2022
2	Teacher	312	12	List of lecturers updated to 2022
3	GradStudent	2009	10	List of postgraduate students
4	CourseGrade	674273	15	Student learning scores for each subject
5	BehaviorGrade	111978	7	Training points for each semester
6	TeachingClass	14728	3	List of classes and instructors teaching that class
7	ExtendedTuition	10799	2	List of students whose tuition fees are extended
8	WaiverTuition	5652	7	List of students eligible for tuition exemption

tained during the experiments. The machine learning models were evaluated without adjustments to the class distribution, reflecting the true scenario where excellent students are a minority. This approach ensures that the models are tested in a realistic setting, acknowledging the challenges of working with imbalanced data. Furthermore, the dataset facilitates in-depth analysis of factors influencing students' graduation outcomes. Because of the inaccuracy of the data, we only filtered the students who studied more than 6 semesters at school.

TABLE III
THE DISTRIBUTION OF GRADUATION CLASSIFICATION LABELS FOR GRADUATE STUDENT DATA.

Graduation labels	Number of students
Excellent	13
Good	910
Fair	3650
Average	1720

IV. Our Approach

A. Data Quality Management

Data Quality Management is a process that includes steps and tools to maintain data quality stability over time. Data management is an important part of the data management process. Depending on the characteristics of the data and the requirements of the organization operating the data, there are many proposed processes and steps for managing data quality. Figure 2 below describes a data quality management process including four stages proposed in this study based on the characteristics of the higher education data set to predict student graduation results. It comprises four main stages: Profile, Implement, Monitor, and Improve.

In the Profile stage, data quality requirements are identified based on the task of predicting student graduation outcomes. This involves defining the necessary data quality rules for verification and selecting appropriate evaluation dimensions. The Implement stage involves deploying the data quality evaluation framework to assess data quality at both individual table levels and overall. During the Monitor stage, data quality evaluation results are continuously tracked over time, with alerts issued when data quality declines. The Improve stage proposes suitable improvement methods for each data type, comparing the accuracy of the graduation prediction model before and after improvement to assess the effectiveness of the applied enhancement methods.

Our study chooses five popular dimensions Accuracy (Acc), Completeness (Comp), Consistency (Con), Uniqueness (Uni), and Timeliness (Time) that are measured by the list of checks shown in Table IV. Accuracy ensures data accurately reflects real-world values by addressing issues like incorrect or irrelevant information, such as misspelled student names. The framework validates these details against official records. Completeness ensures all necessary data is present, resolving issues like missing grades or incomplete student records by prompting for corrections. Consistency focuses on maintaining uniform data formats and standardizing entries such as dates and grades to prevent inconsistencies. Uniqueness eliminates duplicate records by identifying and removing redundant entries, ensuring clarity. Timeliness ensures the data remains current by regularly updating outdated information and maintaining its relevance for analysis.

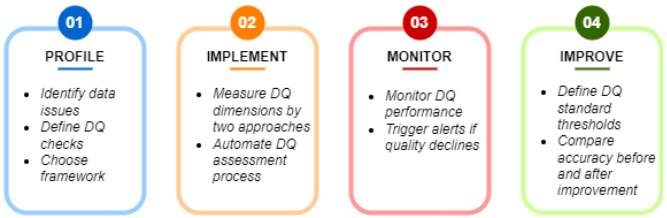

Fig. 2. A proposed four-phase data quality management process

Figure 3 illustrates the implementation of data quality management (DQM) in a cloud storage environment within the Microsoft Azure ecosystem. Data from the SQL database is integrated into the Data Lake through a pipeline designed in Data Factory. Within the Data Lake, data is divided into three containers: Bronze (raw data), Silver (cleaned data), and Gold (selected data for specific organizational tasks). Data quality in the Data Lake is managed through the proposed process outlined in Figure 2. These tasks are performed in the Databricks environment. For end users, data that has passed quality checks that are shown in Table IV and improvements to meet standards is used to visualize insights and data trends through Synapse Analytics. Additionally, the study's experimental task of predicting student graduation outcomes is conducted to evaluate the impact of data quality on the accuracy of predictive machine learning models. The level of automation is defined based on the ability to generalize DQM functions. However, since data is continuously changing and updated to meet real-world conditions and human needs, this process cannot be fully automated without the involvement of data managers for monitoring, supervision, and execution.

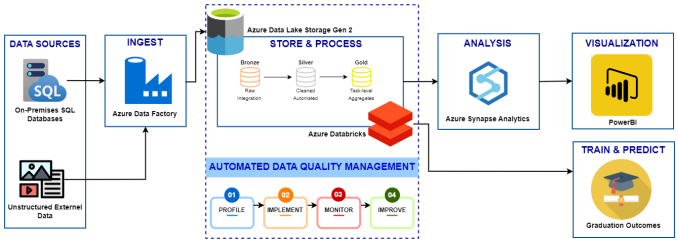

Fig. 3. Diagram integrating data quality management process for data stored in data lake on big data architecture in Microsoft Azure ecosystem.

TABLE IV
TABLE OF PROPOSED CHECKS PERFORMED FOR EACH DATA QUALITY CRITERION AT ROW LEVEL AND TABLE LEVEL

	Data Quality Issues	Acc	Comp	Cons	Uni	Time
Row Level	Missing data		x			
	Incorrect data	x				
	Inconsistent data format			x		
	Outdated data					x
	Irrelevant data	x				
	Misspelling	x				
	Duplicated records				x	
	Incomplete records		x			
Table Level	Uniqueness constraint				x	
	Incorrect data definition	x				
	Wrong data type	x				
	Inconsistent data types			x		
	Outdated table					x
	Missing mandatory fields		x			

B. Student Outcomes Prediction

In this study, predicting students' graduation outcomes serves as a means to assess the effectiveness of proposed data quality improvement methods by training independent models on two datasets—before and after improvement. Graduation outcomes are categorized into four levels (Excellent, Good, Fair, and Average) based on cumulative GPA and average training scores. Additionally, students must fulfill specific requirements such as language proficiency certifications, national defense certificates, and payment of tuition fees. Input data comprises multiple attributes representing students' academic and training experiences during their university studies. These input attributes were selected based on Principal Component Analysis (PCA) analysis to identify influential factors affecting the prediction model's accuracy.

Two machine learning algorithms, LightGBM and CatBoost, are employed to predict student graduation outcomes due to their robust handling of classification tasks and ability to manage complex data structures. LightGBM is an open-source gradient boosting model developed by Microsoft, utilizing a "histogram-based" algorithm to find split points during tree-based learning. This algorithm offers a label distribution-based ensemble learning method that efficiently handles large datasets, making it suitable for online educational predictions [28]. CatBoost developed by the Russian company Yandex is a gradient-boosting algorithm designed specifically for classification tasks. It excels in managing categorical features and mitigating overfitting, making it ideal for predicting and classifying student academic performance [29]. However, CatBoost does not support sparse matrices and requires more training time than LightGBM. These models are particularly effective in handling imbalanced data, which is crucial for accurately predicting outcomes in educational datasets.

V. Experiments

We conducted a data quality assessment using two approaches. First, we evaluated the quality of each table within the dataset. The results in Table V represent the average of each metric across the columns of each table by implementing all the checks for five dimensions in Table IV. For each column, different calculation formulas were employed based on the evaluator's objectives, as detailed in the data profiling section. Subsequently, for columns with low scores, we investigated the causes and implemented improvement measures. The reassessed data quality results are presented in the "Improved Data" column, directly adjacent to the "Raw Data" column.

Overall, the initial assessment for the CourseGrade and Student tables revealed low scores, below 50%, for the accuracy and consistency metrics. This was attributed to significant changes in student information and individual course grades over the organization's 15-year history. Inconsistencies arose due to different data formats stored across separate departments and outdated data not being updated with new information. The uniqueness metric for all tables was nearly perfect, indicating minimal duplicate entries. However, the completeness scores for the CourseGrade and WaiverTuition tables were notably low, at 22.47% and 7.59%, respectively. This was due to inconsistencies in mandatory grade columns for courses and changes in tuition waiver policies over the years, leading to low average evaluation results.

Secondly, in the right side of Table V, we evaluated the overall data quality of the dataset containing selected attributes used to train a model for predicting student graduation outcomes. The dataset includes basic student information from the Student table, semester GPA based on individual course grades from the CourseGrade table, and conduct scores from the BehaviorGrade table. We trained and predicted using these attributes in two scenarios: before and after data quality improvement. However, since the data for graduated students is relatively complete and accurate, having been verified by the institution before graduation recognition, there was no significant difference in prediction results between the two data groups. This is evident in Table V, where the accuracy increase ranged around 5%.

VI. Conclusion and future work

Our research focuses on identifying existing issues in higher education data and proposing a data quality management process within a big data architecture to assess and maintain good data quality for the institution's overall data and graduation prediction models. The study provides a thorough examination of concepts and research trends in data quality assessment, from general overviews to detailed insights. It includes fundamental definitions of data quality frameworks and evaluation criteria and explains specific processes for a university's dataset. The research converts theoretical concepts into quantifiable visual results, allowing for the assessment of data quality levels for structured datasets within the Data Lake. We also demonstrated how the quality of data, before and after improvement, impacts the accuracy of graduation prediction models.

Predictive models may introduce bias from training data, leading to unfair treatment of overrepresented student groups. Student data contains personal information, requiring compliance with privacy regulations, encryption of sensitive data, and clear access controls. Although data quality evaluation requires assessing the unencrypted raw data to accurately evaluate criteria like accuracy, consistency, and completeness, participants must commit to strict data confidentiality throughout the research process.

In the future, we plan to develop data quality assessment methods for other data types in the Data Lake, such as text and images. Furthermore, we aim to collect more data from sources like social media to diversify the dataset, enhance the prediction of graduation outcomes, and gain better insights into student behavior. Additionally, we will explore methods for addressing data imbalance, such as applying weighting techniques and identifying key data clusters, to further improve the accuracy and reliability of our predictive models.

TABLE V
THE DATA QUALITY ASSESSMENT RESULTS ON INDIVIDUAL TABLE AND GRADUATION PREDICTION DATASET

Individual Table-level Data Quality Assessment									Dataset-level Data Quality Assessment			
Table	Completeness		Accuracy		Consistency		Uniqueness		Prediction Model	Metric	On Raw Data	On Improved Data
	Raw Data	Improved Data	Raw Data	Improved Data	Raw Data	Improved Data	Raw Data	Improved Data				
Student	96.88	97.67	**45.99**	**87.53**	**35.67**	**90.00**	100	100	LightGBM	Accuracy	**80.65**	**85.05**
Teacher	90.38	93.42	66.40	74.06	56.25	86.47	100	100		Precision	78.23	87.20
GradStudent	**44.30**	**58.34**	63.47	76.12	59.38	92.46	100	100		Recall	61.40	69.74
TeachingClass	100	100	75.00	75.00	75.00	75.00	99.86	100		F1-score	68.80	77.50
CourseGrade	**22.47**	**22.47**	**48.29**	**94.17**	**47.92**	**59.98**	100	100	CatBoost	Accuracy	**79.32**	**85.16**
BehaviorGrade	90.6	90.6	79.12	98.70	52.68	87.9	100	100		Precision	74.88	81.62
WaiverTuition	**7.59**	**65.00**	86.53	100	50.89	83.54	100	100		Recall	61.04	66.48
ExtendedTuition	100	100	85.05	100	75.00	100	99.91	100		F1-score	65.80	71.32

ACKNOWLEDGMENT

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS2024-26-02.

REFERENCES

[1] Günther, L.C., Colangelo, E., Wiendahl, H.H. and Bauer, C. (2019) Data Quality Assessment for Improved Decision-Making: A Methodology for Small and Medium-Sized Enterprises. Procedia Manufacturing, 29, 583-591.

[2] Tute, Erik, Nagarajan Ganapathy, and Antje Wulff. "A data driven learning approach for the assessment of data quality." BMC Medical Informatics and Decision Making 21 (2021): 1-11.

[3] Zha, Daochen, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, and Xia Hu. "Data-centric ai: Perspectives and challenges." In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), pp. 945-948. Society for Industrial and Applied Mathematics, 2023.

[4] Wang, Jingran, Yi Liu, Peigong Li, Zhenxing Lin, Stavros Sindakis, and Sakshi Aggarwal. "Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality." Journal of the Knowledge Economy (2023): 1-20.

[5] Cichy, Corinna, and Stefan Rass. "An overview of data quality frameworks." Ieee Access 7 (2019): 24634-24648.

[6] Makarov, Artem, and Dmitry Namiot. "Overview of data cleaning methods for machine learning." International Journal of Open Information Technologies 11, no. 10 (2023): 70-78.

[7] Ehrlinger, Lisa, and Wolfram Wöß. "A survey of data quality measurement and monitoring tools." Frontiers in big data 5 (2022): 850611.

[8] Cichy, Corinna, and Stefan Rass. "An overview of data quality frameworks." IEEE Access 7 (2019): 24634-24648.

[9] English, Larry P. "Total Quality data Management (TQdM) Methodology for Information Quality Improvement." In Information and database quality, pp. 85-109. Boston, MA: Springer US, 2002.

[10] Wang, Richard Y. "A product perspective on total data quality management." Communications of the ACM 41, no. 2 (1998): 58-65.

[11] Cappiello, Cinzia, Paolo Ficiaro, and Barbara Pernici. "HIQM: A methodology for information quality monitoring, measurement, and improvement." In Advances in Conceptual Modeling-Theory and Practice: ER 2006 Workshops BP-UML, CoMoGIS, COSS, ECDM, OIS, QoIS, SemWAT, Tucson, AZ, USA, November 6-9, 2006. Proceedings 25, pp. 339-351. Springer Berlin Heidelberg, 2006.

[12] Batini, Carlo, Federico Cabitza, Cinzia Cappiello, and Chiara Francalanci. "A comprehensive data quality methodology for web and structured data." International Journal of Innovative Computing and Applications 1, no. 3 (2008): 205-218.

[13] Carlo, Batini, Barone Daniele, Cabitza Federico, and Grega Simone. "A data quality methodology for heterogeneous data." International Journal of Database Management Systems 3, no. 1 (2011): 60-79.

[14] L. Sebastian-Coleman, Measuring Data Quality for Ongoing Improvement. Waltham, MA, USA: Morgan Kaufmann, 2013.

[15] Vaziri, Reza, Mehran Mohsenzadeh, and Jafar Habibi. "TBDQ: A pragmatic task-based method to data quality assessment and improvement." PLoS One 11, no. 5 (2016): e0154508.

[16] Corrales, David Camilo, Agapito Ledezma, and Juan Carlos Corrales. "From theory to practice: A data quality framework for classification tasks." Symmetry 10, no. 7 (2018): 248.

[17] Kuemper, Daniel, Thorben Iggena, Ralf Toenjes, and Elke Pulvermueller. "Valid. IoT: A framework for sensor data quality analysis and interpolation." In Proceedings of the 9th ACM Multimedia Systems Conference, pp. 294-303. 2018.

[18] Juneja, Ashish, and Nripendra Narayan Das. "Big data quality framework: Pre-processing data in weather monitoring application." In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 559-563. IEEE, 2019.

[19] Salemdeeb, Ramy, Ruth Saint, William Clark, Michael Lenaghan, Kimberley Pratt, and Fraser Millar. "A pragmatic and industry-oriented framework for data quality assessment of environmental footprint tools." Resources, Environment and Sustainability 3 (2021): 100019.

[20] Benkherourou, Chafika, and Abdelhabib Bourouis. "A framework for improving data quality throughout the MDM implementation process." In 2nd International Conference on Industry 4.0 and Artificial Intelligence (ICIAI 2021), pp. 164-169. Atlantis Press, 2022.

[21] Croft, Roland, M. Ali Babar, and M. Mehdi Kholoosi. "Data quality for software vulnerability datasets. In 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)." (2023): 121-133.

[22] Ru, Boshu, Arthur Sillah, Kaushal Desai, Sheenu Chandwani, Lixia Yao, and Smita Kothari. "Real-World Data Quality Framework for Oncology Time to Treatment Discontinuation Use Case: Implementation and Evaluation Study." JMIR Medical Informatics 12 (2024): e47744.

[23] Aiken, John M., Riccardo De Bin, Morten Hjorth-Jensen, and Marcos D. Caballero. "Predicting time to graduation at a large enrollment American university." Plos one 15, no. 11 (2020): e0242334.

[24] Alyahyan, Eyman, and Dilek Düştegör. "Predicting academic success in higher education: literature review and best practices." International Journal of Educational Technology in Higher Education 17 (2020): 1-21.

[25] Iatrellis, Omiros, Ilias . Savvas, Panos Fitsilis, and Vassilis C. Gerogiannis. "A two-phase machine learning approach for predicting student outcomes." Education and Information Technologies 26 (2021): 69-88.

[26] Demeter, Elise, Mohsen Dorodchi, Erfan Al-Hossami, Aileen Benedict, Lisa Slattery Walker, and John Smail. "Predicting first-time-in-college students' degree completion outcomes." Higher Education (2022): 1-21.

[27] Ehrlinger, L., Haunschmid, V., Palazzini, D., and Lettner, C. (2019). "A DaQL to monitor the quality of machine data," in Proceedings of the International Conference on Database and Expert Systems Applications (DEXA), volume 11706 of Lecture Notes in Computer Science. (Cham: Springer), 227–237.

[28] Zhang, Long, Shu Kai, Huang Keyu, and Zhang Ruiqiu. "An approximation of label distribution-based ensemble learning method for online educational prediction." INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS CONTROL 16, no. 3 (2021).

[29] Joshi, Abhisht, Pranay Saggar, Rajat Jain, Moolchand Sharma, Deepak Gupta, and Ashish Khanna. "CatBoost—An ensemble machine learning model for prediction and classification of student academic performance." Advances in Data Science and Adaptive Analysis 13, no. 03n04 (2021): 2141002.