IJAIT (International Journal of Applied Information Technology)
http://journals.telkomuniversity.ac.id/ijait/

A Review: Data Quality Problem in Predictive Analytics

Heru Nugroho [a, *]

[a] *School of Applied Science, Telkom University, Bandung, Indonesia*
heru@tass.telkomuniversity.ac.id

ARTICLE INFO

Received May 7th, 2023
Revised June 7th, 2023
Accepted June 13th, 2023
Available online August 18th, 2023

Keywords
Data analytics; predictive analytics; data quality.

ABSTRACT

As the volume of data continues to expand, there has been a significant transformation in computational techniques and statistical methods used for data processing and analysis. This shift in the analytical data paradigm, moving from explicit to implicit, has opened up new possibilities for extracting insights and knowledge from data. By adopting a prospective approach, the value of new observations can be determined based on the underlying relationship between input and output variables. Data preparation plays a crucial role in predictive analytics, as high-quality data meeting specific criteria is essential for conducting accurate and reliable analyses. In the era before digital computers, data quality played a significant role in strategic decision-making and planning. However, one common challenge encountered during data analysis is the incompleteness of the raw data, often caused by missing values. The presence of missing data compromises the accuracy of predictive analyses derived from such datasets. This paper aims to address the issues associated with data quality in predictive analytics by conducting a comprehensive literature review of relevant research based on bibliometrics analysis by using R programming. Furthermore, the paper will explore the potential challenges and future directions in the field of predictive analytics concerning data quality problems.

* Corresponding author at:
School of Applied Science, Telkom University
Jl. Telekomunikasi No. 1, Terusan Buah Batu, Bandung, 40257
Indonesia.
E-mail address: heru@tass.telkomuniversity.ac.id

ORCID ID:
Author: 0000-0002-7460-7687

https://doi.org/10.25124/ijait.v7i02.5980
Paper_reg_number IJAIT000070201 2023

1. Introduction

At present, the data environment is growing in terms of complexity, size, and resolution. The ability to utilize available data and information for decision-making is needed to overcome multidisciplinary problems. A good understanding of data science is needed to successfully synthesize heterogeneous data from various sources to support holistic analysis and the extraction of new knowledge, which became known as predictive analytics [1]. The amount of data will grow very rapidly in the next few years, in line with the existence of emerging technologies and all devices. addressing the fact that 90% of the data is currently made in recent years This will be a trend that will continue in the future [2].

In the age of digital computing, the organization's primary focus is on the optimization of its technological infrastructure. It is imperative that companies could plan for unforeseen demands in international operations. Businesses are motivated to overcome the data that is produced at every turn by the competitive environment in which they operate. It is necessary to have data to acquire the knowledge required for the business expansion model. The field of predictive analytics employs several different algorithms to search through a collection of data for distinct patterns that can indicate productive behavior for commercial solutions [3].

The most difficult part of using predictive analytics is preparing the data to be analyzed. This is because the analysis process cannot be performed directly on raw data. An analyst is required to carry out the preprocessing stage to obtain data that is prepared for usage in the predictive analytics phase [3]. It is essential to have quality data if one is going to perform quality data analytics. The success of an organization is directly proportional to the quality of its data and the analytical methods it employs. It will provide highly essential insights, as well as assistance in making judgments about smart decisions [4].

Based on searches with the string “Data AND Quality AND Predictive AND Analytics” in the Scopus database with the limitation of 2012 – 2022 and the source of articles from journals or conferences obtained 1165 articles where an Annual Growth Rate 39.29% with the number of authors 3849.

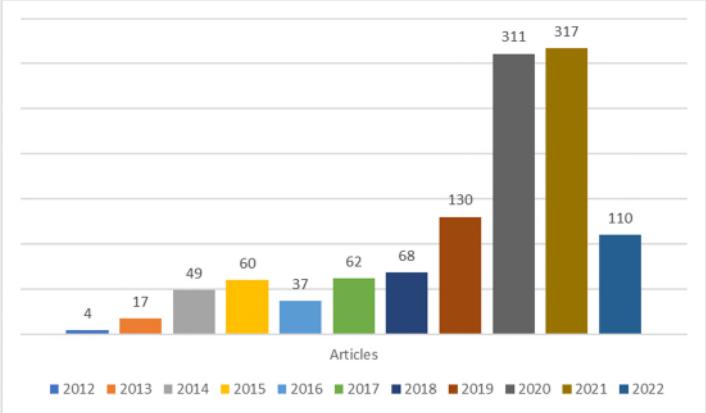

Figure 1 Annual Scientific Production

Figure 1 shows that the amount of research related to data quality and predictive analytics is growing quite rapidly between 2020 and 2021, when big data, AI, and machine learning technologies are growing too. Data quality is a key factor affecting data assessment and is still being ignored. Data quality is an integrated and diverse concept that requires knowledge of many data characteristics [5]. One problem that arises related to data quality is completeness. There are also many

cases where databases are found where half of the dataset is missing. So, it is very difficult to mine data because the available analytical methods only work with complete data [6]. Figure 2 is historiography that involves constructing a chronological overview of publications in a specific research field or topic about quality data and predictive analytics. It provides a visual representation of the temporal evolution of research and helps identify key milestones, trends, and shifts in the field over time.

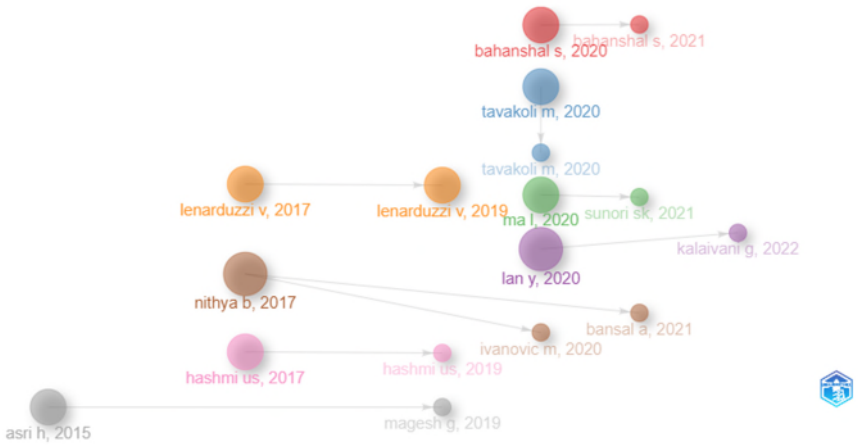

Figure 2 Historiograph

Research related to data quality and predictive analytics has been underway since 2015 but will expand by 2020. One of the issues discussed in this study is the problem of missing data. Recent research on missing data has been done enough to produce accurate methods in imputation as done by researchers in previous studies [7]–[11].

Incomplete data can affect the results of data prediction systems [12], [13]. The impact of incomplete data, for example, the presence of missing data in quantitative research, can be serious, leading to biased parameter estimates, information loss, increased standard errors, and weak generalizations of findings as results [14]. Most of the research related to data quality review in the context of predictive analytics has not utilized bibliometric analysis. The main contribution of this paper is identifying the problems related to the quality of data in predictive analytics through a literature review of related research based on bibliometrics analysis by using R programming. Additionally, it will present the challenges and directions that may occur in the domain of predictive analytics with the existence of data quality.

2. Materials and Methods

Research related to predictive analytics is associated with data quality issues and contains several studies on data analytics, predictive analytics, and data quality.

2.1. Data Analytics

Data analytics is a multidisciplinary science (theory, technology, and process) that quantitatively and qualitatively examines data to draw new conclusions or insights (exploratory or predictive) or to extract and prove (confirm or based on facts) about information for decision-making and action [15]. The objectives of analytical data are prediction and information.

The growing availability of data, whether from sensor networks or through interactions between humans and computers, gives rise to new kinds of information

systems, in which data analytics play a significant part [16]. To assist them in decision-making, corporations have utilized a wide variety of tools and strategies, each of which possesses a unique level of intelligence and sophistication. Because of the development of more sophisticated methods of analysis, businesses are now able to begin proactive decision-making with predictive and prescriptive analysis. This allows them to answer questions regarding why something occurred, what might occur in the future, and how a problem can be solved [17].

In the past forty years, there has been a shift from the analysis of data on small and simple data, along with the testing of hypotheses, to the analysis of data on vast and complicated data, intending to discover knowledge and insights without the need for hypotheses. This has led to changes in data trends from the explicit era towards the implicit era, as seen in Figure 3 [15].

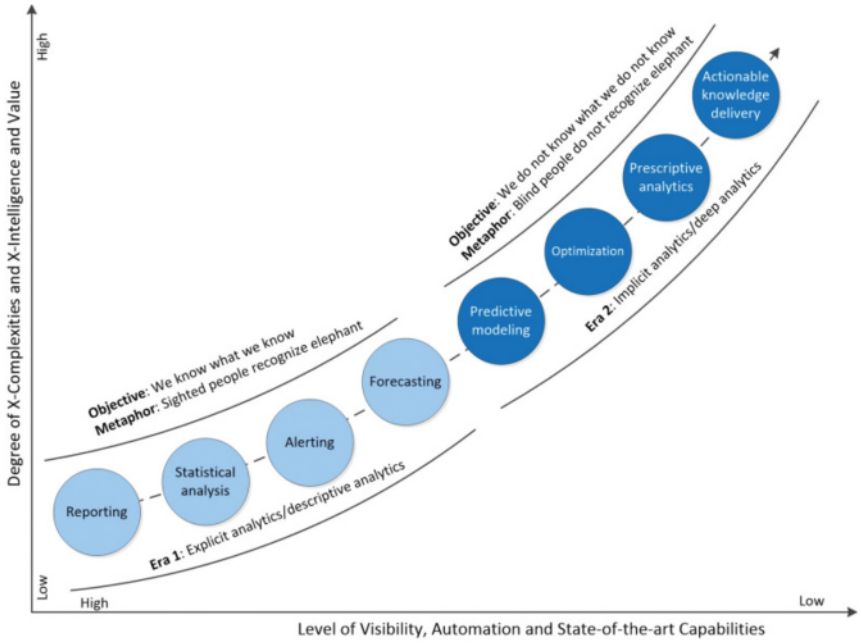

Figure 3 Explicit-to-implicit analytics spectrum and evolution [15]

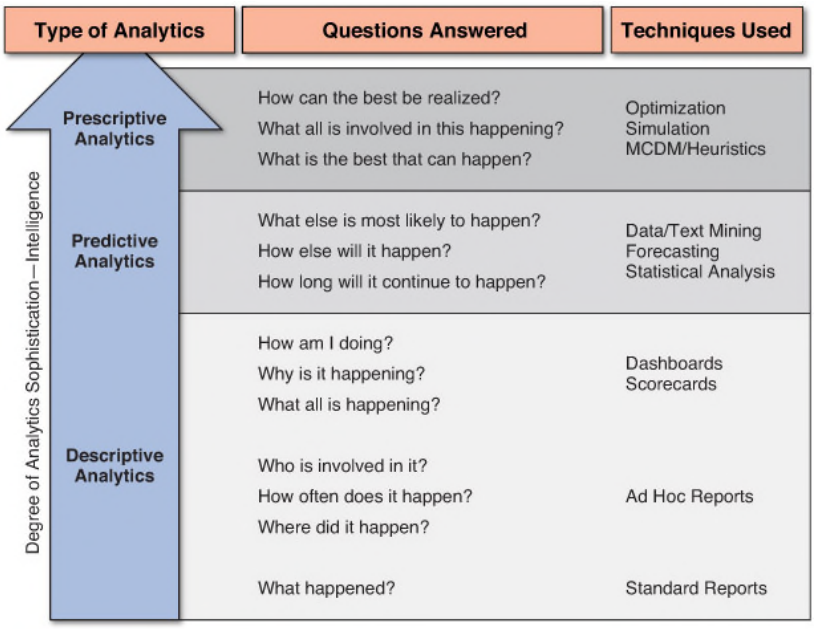

Figure 4 Three Levels of Analytics and Their Enabling Techniques [18]

Data analytics has become a very important and challenging problem in scientific disciplines such as computer science, biology, medicine, finance, and internal security. The existence of a paradigm shift from explicit to implicit becomes an inseparable part of the grouping of analytic data categories, namely descriptive analytic, predictive analytic, and prescriptive analytic [1]. A tabular illustration of the three levels of hierarchy that may be found in analytics is shown in Figure 4, along with the questions that are answered and the techniques that are employed at each level. As can be observed, the most important factor in the success of predictive analytics is the mining of data.

These three levels of analytics work in conjunction to enhance fraud prediction and prevention efforts. Descriptive analytics provides a foundation of understanding, predictive analytics identifies potential fraud cases, and prescriptive analytics offers strategies for proactive fraud management. By leveraging these enabling techniques, organizations can improve their ability to detect, prevent, and combat fraud effectively.

2.2. Predictive Analytics

Building and evaluating algorithmic models to make them empirical and predictive is what is meant by the term "predictive analytics." The purpose of prediction models is to forecast the outcomes of future observations [3]. The predictive analytic model makes use of a prospective method to estimate the value of new observations based on the structure of the relationship between input and output to extract information from the data. [4]. Predictive analytics has an important role in business because it can:

1. Detecting Fraud [19] [20] [21]
2. Optimizing the Marketing Strategy [22], [23]
3. Improved Operations [24], [25]
4. Reducing the Risk [26]

Predictive analytics helps develop and test theoretical models through different perspectives rather than explanatory statistical models and is needed in scientific research. In particular, the six roles of predictive analytics are as follows [27], [28].

1. Creating a New Theory
2. Developing Measurement Instruments
3. Comparing Competitive Theories
4. Improve the Existing Model
5. Test Relevance
6. Measuring the level of predictability (uncertainty)

The term "predictive analytics" refers to a collection of methodologies that can provide accurate forecasts of future outcomes by analyzing past data in conjunction with recent data. The goal of predictive analytics is to find patterns and identify links hidden inside data. Regression techniques (such as multinomial logit models), on the one hand, and machine learning techniques, on the other, are the two categories that can be used to classify analytic prediction strategies [14]. One further categorization can be done based on the type of result variable. When dealing with continuous variables, such as home selling prices, techniques such as linear regression are utilized, but techniques such as random forest are utilized when dealing with discrete output variables (for example, credit status) [14]. Predictive Analytics has a strong preference for the use of classification

methodologies and regression analysis. Detailed taxonomy for Predictive Analytics can be seen in Figure 5 [29].

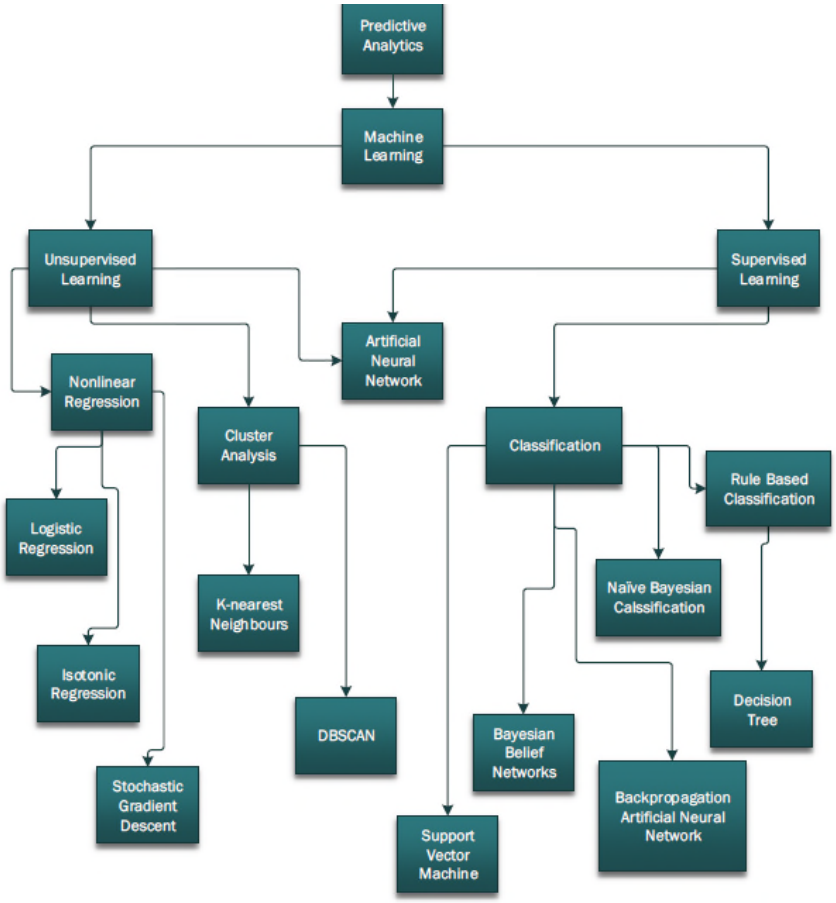

Figure 5 Predictive Analytics Taxonomy [29]

This taxonomy categorizes predictive analytics techniques used in data mining with big data. Although the specific details of the taxonomy are not provided, it likely includes various methods and approaches for predictive analytics, such as regression, classification, clustering, association rule mining, and ensemble methods. Taxonomy helps organize and classify these techniques, providing a framework for understanding and applying predictive analytics in the context of big data.

2.3. Data Quality Problem in Predictive Analytics

The quantity and quality of data are like two sides of a coin; both must be considered critically significant for data management [30]. The relatively recent development of big data has resulted in the addition of new aspects that can contribute to problems with the quality of the data. In addition, when taking into consideration the many uses of big data, the potential consequences of poor data quality have the potential to become more broad and widespread disasters [31].

As a result of the advent of the era of big data, the growth of data in a variety of businesses and fields is experiencing an explosion. Both the business world and the academic world are struggling with the same fundamental problem: how to ensure that the data they collect is of high quality, as well as how to investigate, discover, and use the information and knowledge that is concealed inside the data. If the quality of the data is poor, the efficiency with which it is used will suffer, and it may even lead to mistakes in decision-making. The availability of the data, its

usability, its dependability, its relevance, and its presentation quality are the five components that comprise data quality standards [32].

The engineering environment stores a variety of data kinds in its various databases. Most of this information is distinct, and the administration of it does not necessarily adhere to the most effective practices. This results in issues like inconsistency in maintenance and reuse, as well as difficulties in meaningfully discovering new information. The use of data-based modeling as a predictive analytics tool is particularly useful for technical data. However, the accuracy of the model is highly reliant on the precision of the data it employs [33].

The level of success that an organization achieves is directly related to how well it exploits its data in the service of eliciting relevant information or knowledge to inform decision making. However, various issues can affect the data quality that is related to the extraction of data, and these issues can affect both the truth and the usefulness of the data. Because of this, having knowledge of and an understanding of the issues surrounding data quality is quite crucial [34]. The open data paradigm is responsible for much of the success that has been achieved by previously developed technologies; there remains a degree of ambiguity regarding the quality of the data that is derived from the data set. The potential value that can be derived from the data is put at risk whenever there is uncertainty over the data's quality[35].

Research on data quality can be broken down into numerous distinct categories. Data restrictions, data integration and warehousing difficulties, data cleansing, data quality metrics, data origin, and data quality dimensions are some of the topics that are covered in this chapter [36]. A subsidiary part is devoted to every one of these facets. For instance, the sub-sections of data cleaning include things like the detection of errors, the resolution of inconsistencies, the conclusion of incomplete data, the detection of duplicates, etc. [31]. These days, quality management is made more difficult by the appearance of several different elements. In the first place, some repercussions are associated with the amount of data that is being produced by the company now. Second, during the previous several years, there has been a rise in the variety of data. This variety of data includes structured, unstructured, semi-structured, and multi-media content, which includes things like movies, maps, and photos, among other things.[36].

The problem of data quality in predictive analytics is related to the data preparation stage. Good data quality will affect the predictive analytic model produced. Some reviews of several scientific papers that discuss the relevance of data quality with the analysis process from data can be seen in Table 1.

Table 1 Review of Data Quality with Predictive Analytics

Title	Findings / Results
A Data Quality in Use Model for Big Data [37]	Data quality assessment in the context of Big Data is the focus here. The goal of the model is to solve problems that only come up with extremely huge and complicated data sets. The authors ensure the trustworthiness, precision, and use of Big Data by evaluating multiple dimensions of data quality and providing insights and instructions for doing so. The findings of this study benefit data management researchers and professionals dealing with Big Data.
Importance of Data Quality for Analytics [4]	The analytical significance of data quality is discussed. It emphasizes the need to use high-quality data in analytical procedures that yield trustworthy results. The author stresses the need for enterprises to develop efficient data quality management methods by highlighting the significance of high-quality data. This study sheds light on how important data quality is for modern analytics projects to succeed.

Title	Findings / Results
Taxonomy of data quality problems in multidimensional Data Warehouse models [38]	Data quality issues are common in the context of DWs, and the authors hope to provide a thorough methodology for recognizing and fixing them. This research sheds light on the typical difficulties encountered in DW settings by classifying data quality issues into several groups. Researchers and practitioners alike can benefit greatly from this taxonomy as they work to enhance the accuracy of data stored in multidimensional DW models.
Bias arising from missing data in predictive models [39]	The author investigates how missing data might result in false and distorted predictions, affecting the overall performance and dependability of the models. Through a discussion of different causes of bias and their consequences, the study emphasizes the significance of treating missing data effectively to enhance the accuracy and fairness of prediction models. This study focuses light on the issues faced by missing data and gives ideas for predictive modeling academics and practitioners.
Missing values in data analysis: Ignore or Impute?[40]	The authors talk about the effects of missing values and possible biases that might appear if missing values are not handled properly. They give an overview of different imputation methods that can be applied to complete datasets with missing values. The purpose of the study is to emphasize the significance of dealing with missing values in data analysis to guarantee accurate and trustworthy outcomes. It provides perceptions and direction for academics and practitioners on the selection of missing values in data analysis.
Data Mining and The Impact of Missing Data [41]	The authors look at how data mining techniques' outputs and outcomes can be impacted by missing data. They examine the various kinds and patterns of missing data and talk about the difficulties they provide when trying to draw insightful conclusions from datasets. To achieve accurate and trustworthy results from data mining, the study underlines the need for appropriate treatment of missing data. It presents suggestions for tackling this problem in data mining practice and sheds light on the effects of missing data.
Overview of data quality challenges in the context of Big Data [42]	The author draws attention to the distinctive qualities of Big Data, such as volume, velocity, and variety, which lead to problems with data quality. The dependability and usability of Big Data may be impacted by typical issues such as data inconsistency, incompleteness, and inaccuracies, which are discussed in the paper. The paper intends to improve understanding and awareness of data quality problems in the Big Data realm by shedding light on these difficulties. It stresses how crucial it is to handle data quality issues in the context of Big Data and provides useful information for researchers and practitioners working with large-scale datasets.
Improving Power Grid Monitoring Data Quality: An Efficient Machine Learning Framework for Missing Data Prediction [43]	The difficulty of missing data in power grid datasets is addressed by the authors' efficient machine learning methodology. The system forecasts missing values in the data using machine learning approaches, enhancing the quality of the data overall. The significance of trustworthy power grid data for precise monitoring and analysis is emphasized in the paper. It offers information on how to build a strong machine learning method for handling missing data in datasets related to the power grid, improving data quality in the context of power grid monitoring.
Impact of Data Quality on Predictive Accuracy of ANFIS-based Soft Sensor Models [44]	The performance of ANFIS models in forecasting sensor measurements is examined in relation to differences in data quality, such as noise or missing values. The research emphasizes the significance of high-quality data for achieving accurate and trustworthy predictions in soft sensor applications through experiments and studies. The study adds to our understanding of the connection between predicted accuracy and data quality, offering suggestions for the creation and evolution of soft sensor models based on ANFIS.
Estimating the Financial Impact of Data Quality Issues [45]	To help companies determine the costs related to data errors, inconsistencies, and inaccuracies, this article explores numerous methods and approaches for measuring the financial effect of data quality concerns. Organizations may make wise decisions and spend resources efficiently to address data quality concerns by being aware of the financial ramifications. For controlling data quality and its financial impact on enterprises, the article provides insightful analysis and recommendations.

3. Result and Discussion

As is common knowledge in this highly regulated environment, there is an ever-growing need to develop and supply protections and tools to boost the transparency and accuracy of information. These kinds of tools and methods are essential to fulfill the standards imposed by businesses and regulators. Because of this condition, data and analytics have taken on a much more important role in business in general. [4]. Reference sources that address the topics of data quality in predictive analytics on the Scopus database are many found in the ACM

International Conference Series and International Journal IEEE Access as can be seen in Figure 6.

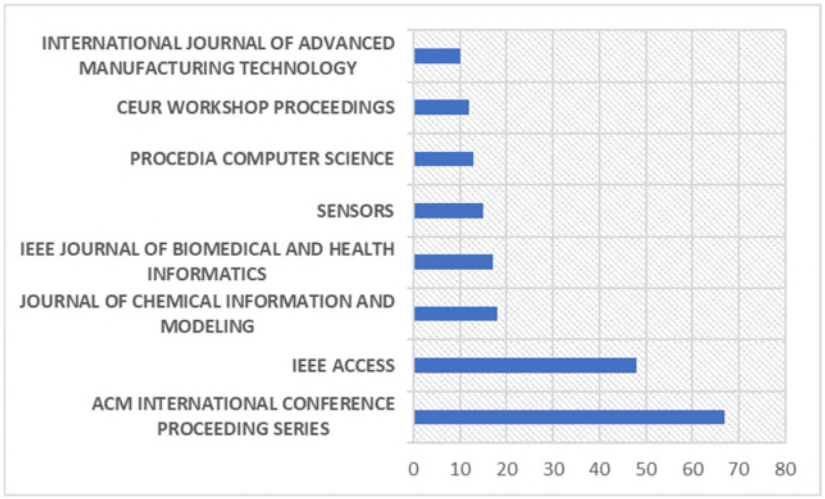

Figure 6 Most Relevant Sources

Because of its ability to depict data in a condensed, succinct format, the term co-occurrence network analysis is one of the most powerful methods that can reveal scientific tendencies and problems that have emerged over the long term. Co-occurrence analysis is used to either sketch the conceptual structure of a framework or summarize it in a bibliographic collection. This is accomplished through the utilization of a co-occurrence network to map and cluster the concepts that are derived from keywords. Based on keywords on research related to data quality and predictive analytics obtained most researchers provide keyword machine learning, predictive analytics, and big data. This is in line with the fact already shown in the previous explanation that the trend of data quality and predictive analytics research is aligned with the development of machine learning and big data technologies.

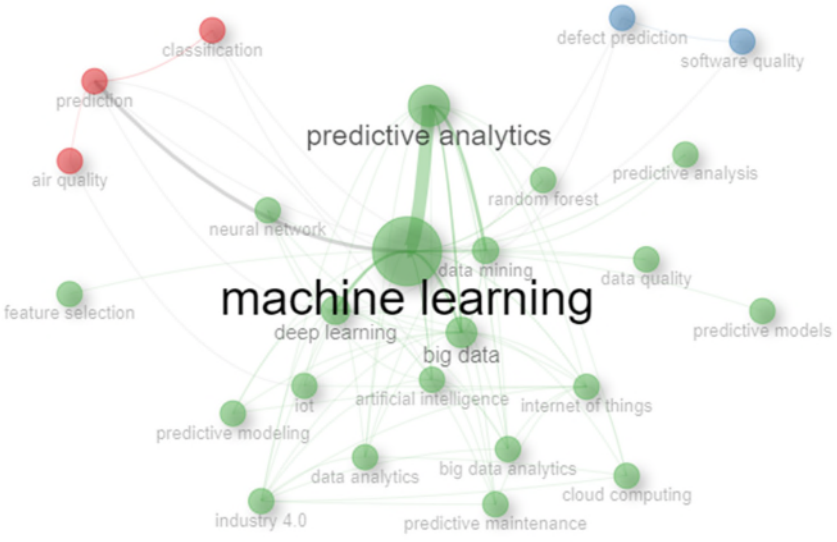

Figure 7 Co-occurrence Network

Opportunities and challenges for research in the predictive analytics domain about previous studies related to predictive analytics methods and predictive analytics models. Some studies address hybrid methods for conducting predictive analytics. Other studies compare the methods and compare the results of their accuracy. Some studies propose new methods as a modification of existing

methods for predictive analytics. Research related to the quality of data associated with data analytics needs shows that:

1. To produce good analytical data output, it needs to be supported by good data quality.
2. Predictive models can be influenced by the quality of data that is not good in this case the existence of missing data.
3. The desired benefits in the presence of analytical data will be affected by handling inappropriate quality data.

Amazing things called prediction models are becoming increasingly useful in today's era of big data, which has led to a rise in the number of possible applications for these models. Despite developments in modeling options and algorithms that promote higher robustness and accuracy, performance remains to be governed by the quality of the data that is fed into the model. It is a future challenge how to determine the method or propose a new method for predictive analytics if data with inappropriate quality is obtained but can still produce predictive results with good accuracy. One of the things that will be done for further research is how data quality problems are more specific, such as missing data which is then associated with predictive analysis.

It is possible for missing data to have a significant impact on the reliability of research findings [46]. The missing of some data in quantitative studies can result in estimations of parameter values that are biased [47]–[50]. The outcomes of the analysis are affected in a meaningful way when missing data are ignored [40], [51], Additionally, the approach taken to deal with incorrectly missing data might affect the overall performance of the model in the prediction model [48], [52]. Researchers can learn the appropriate strategies to address the challenges posed by missing data when such challenges arise. In addition, research that is associated with predictive analytics is highly susceptible to being researched in terms of the approaches that are employed for predictive analytics. Another difficulty is to connect the two opportunities where it is currently possible to execute predictive analytics with constraints such as missing data while maintaining the desired predictive outcomes with optimal computation time [53].

The review results show that one of the causes of poor-quality data that cannot be used directly for data processing is missing data. Various studies show that methods to handle missing data have been developed by many researchers and the techniques are grouped into two techniques, statistical-based imputation, and machine learning-based imputation.

4. Conclusions

The quality of data is not influenced by the nature of the data. In addition, the quality of data is influenced by analytical processes. To support data analysis, high-quality data is needed. One indicator that good quality data is the completeness of the data. Therefore, in preparation for analysis and cross-checking, data plays an important role. Various types of analytics related to data can be specified according to the objectives and needs of the business. Good data combined with the right analytical techniques is the key to organizational success. With good quality data and the right analytics, it will produce very important insights and help in making good decisions.

Once a company has access to data of sufficient quality, it can open a wide variety of options for predictive optimization, which can lead to significantly increased levels of operational efficiency. Research possibilities to find the best

approach can be found in problems linked to the quality of the data. In addition, research pertaining to predictive analytics is open enough to investigate the best practices for using predictive analytics. One more difficulty connects the two opportunities: now, it is possible to do predictive analytics despite the presence of limitations, which take the shape of poor-quality data; nonetheless, one can still achieve the desired level of accuracy in one's forecasts.

Bibliography

[1] K. Gibert, J. S. Horsburgh, I. N. Athanasiadis, and G. Holmes, "Environmental Data Science," *Environmental Modelling and Software*, vol. 106, pp. 4–12, 2018, doi: 10.1016/j.envsoft.2018.04.005.

[2] O. Y. Al-jarrah, P. D. Yoo, S. Muhaidat, and G. K. Karagiannidis, "Efficient Machine Learning for Big Data : A Review," *Big Data Research*, vol. 2, no. 3, pp. 87–93, 2015, doi: 10.1016/j.bdr.2015.04.001.

[3] P. Wazurkar and R. S. Bhadoria, "Predictive Analytics in Data Science for Business Intelligence Solutions," in *International Conference on Communication Systems and Network Technologies*, 2017, pp. 367–370. doi: 10.1109/CSNT.2017.70.

[4] R. Jugulum, "Importance of Data Quality for Analytics," in *Quality in the 21st Century*, Springer International Publishing, Ed., 2006, pp. 23–32. doi: 10.1007/978-3-319-21332-3.

[5] H. Yu and M. Zhang, "Data pricing strategy based on data quality," *Computers & Industrial Engineering*, vol. 112, pp. 1–10, 2017, doi: 10.1016/j.cie.2017.08.008.

[6] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognition*, vol. 41, pp. 3692–3705, 2008, doi: 10.1016/j.patcog.2008.05.019.

[7] H. Nugroho, N. P. Utama, and K. Surendro, "Comparison Method for Handling Missing Data in Clinical Studies," in *9th International Conference on Software and Computer Applications (ICSCA)*, Langkawi, Malaysia, 2020, p. 6.

[8] H. Nugroho, N. P. Utama, and K. Surendro, "Performance Evaluation for Class Center-Based Missing Data Imputation Algorithm," in *Proceedings of the 2020 9th International Conference on Software and Computer Applications*, Langkawi Malaysia: ACM, Feb. 2020, pp. 36–40. doi: 10.1145/3384544.3384575.

[9] H. Nugroho, N. P. Utama, and K. Surendro, "Class center-based firefly algorithm for handling missing data," *Journal of Big Data*, vol. 8, no. 1, p. 37, Dec. 2021, doi: 10.1186/s40537-021-00424-y.

[10] H. Nugroho, N. P. Utama, and K. Surendro, "Normalization and outlier removal in class center-based firefly algorithm for missing value imputation," *J Big Data*, vol. 8, no. 1, p. 129, Dec. 2021, doi: 10.1186/s40537-021-00518-7.

[11] H. Nugroho, N. P. Utama, and K. Surendro, "Smoothing Target Encoding and Class Center-Based Firefly Algorithm for Handling Missing Values in Categorical Variable," *Journal of Big Data*, doi: 10.21203/rs.3.rs-1667499/v1.

[12] P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation," *Neurocomputing*, vol. 72, pp. 1483–1493, 2009, doi: 10.1016/j.neucom.2008.11.026.

[13] M. R. Malarvizhi and A. S. Thanamani, "K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation," *IOSR Journal of Computer Engineering (IOSRJCE)*, vol. 6, no. 5, pp. 12–15, 2012.

[14] Y. Dong and C. J. Peng, "Principled missing data methods for researchers," *SpringerPlus*, vol. 2004, pp. 1–17, 2013.

[15] L. Cao, "Data Science : A Comprehensive Overview," *ACM Computing Surveys*, vol. 50, no. 3, 2017.

[16] M. Bichler, A. Heinzl, and W. M. P. Van Der Aalst, "Business Analytics and Data Science : Once Again ?," *Business & Information Systems Engineering*, vol. 59, no. 2, pp. 77–79, 2017, doi: 10.1007/s12599-016-0461-1.

[17] A. Banerjee, T. Bandyopadhyay, and P. Acharya, "Data Analytics: Hyped Up Aspirations or True Potential?," *VIKALPA*, vol. 38, no. 4, pp. 1–11, 2013.

[18] Dursun Delen, *Real-World Data Mining: Applied Business Analytics and Decision Making*. Pearson FT Press, 2014.

[19] J. West and M. Bhattacharya, "Intelligent financial fraud detection : A comprehensive review," *computers & security*, vol. 57, pp. 47–66, 2016, doi: 10.1016/j.cose.2015.09.005.

[20] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system : A survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016, doi: 10.1016/j.jnca.2016.04.007.

[21] J. L. Perols, R. M. Bowen, C. Zimmermann, and B. Samba, "Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Prediction," *THE ACCOUNTING REVIEW*, vol. 92, no. 2, pp. 221–245, 2017, doi: 10.2308/accr-51562.

[22] P. Ducange, R. Pecori, and P. Mezzina, "A glimpse on big data analytics in the framework of marketing strategies," *Soft Computing*, vol. 22, no. 1, pp. 325–342, 2018, doi: 10.1007/s00500-017-2536-4.

[23] S. Erevelles, N. Fukawa, and L. Swayne, "Big Data consumer analytics and the transformation of marketing," *Journal of Business Research*, vol. 69, no. 2, pp. 897–904, 2016, doi: 10.1016/j.jbusres.2015.07.001.

[24] R. Blackburn, K. Lurz, B. Priese, and I. Darkow, "A predictive analytics approach for demand forecasting in the process industry," *International Transactions in Operational Research*, vol. 22, pp. 407–428, 2015, doi: 10.1111/itor.12122.

[25] T. Ashalatha and N. L. N, "Improving operational efficiencies using Big Data for Financial Services," *Journal of Computer Engineering*, vol. 18, no. 4, pp. 75–77, 2016, doi: 10.9790/0661-1804067577.

[26] H. Ghasemkhani, S. Reichman, and G. Westerman, "Using Predictive Analytics to Reduce Uncertainty in Enterprise Risk Management," in *Thirty-sixth International Conference on Information Systems*, 2015.

[27] D. Delen and H. M. Zolbanin, "The analytics paradigm in business research," *Journal of Business Research*, vol. 90, no. April, pp. 186–195, 2018, doi: 10.1016/j.jbusres.2018.05.013.

[28] G. Shmueli and O. R. Koppius, "Predictive Analytics in Information Systems Research," *MIS Quarterly*, vol. 35, no. 3, pp. 553–572, 2011.

[29] D. Lam, "A Survey of Predictive Analytics in Data Mining with Big Data," Athabacha University, 2014.

[30] W. Fan, "Data Quality : From Theory to Practice," *SIGMOD Record*, vol. 44, no. 3, 2015.

[31] D. Rao, "Data Quality Issues in Big Data," in *IEEE International Conference on Big Data*, 2015, pp. 2654–2660.

[32] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal*, vol. 14, no. 2, pp. 1–10, 2015.

[33] N. Micic, D. Neagu, F. Campean, and E. H. Zadeh, "Towards a Data Quality Framework for Heterogeneous Data," in *IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2017. doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.28.

[34] P. Oliveira, "A Formal Definition of Data Quality Problems.," in *International Conference on Information Quality (MIT IQ Conference)*, 2005.

[35] S. Sadiq and M. Indulska, "Open data : Quality over quantity," *International Journal of Information Management*, vol. 37, no. 3, pp. 150–154, 2017, doi: 10.1016/j.ijinfomgt.2017.01.003.

[36] S. Sadiq, N. K. Yeganeh, and Marta Indulska, "20 Years of Data Quality Research : Themes, Trends and Synergies," in *Proceedings of the Twenty-Second Australasian Database Conference (ADC 2011)*, 2011.

[37] M. Jorge, C. Ismael, R. Bibiano, S. Manuel, and P. Mario, "A Data Quality in Use model for Big Data," *Future Generation Computer Systems*, vol. 63, pp. 123–130, 2016, doi: 10.1016/j.future.2015.11.024.

[38] W. G. De Almeida *et al.*, "Taxonomy of data quality problems in multidimensional Data Warehouse models," in *8th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, 2013.

[39] M. H. Gorelick, "Bias arising from missing data in predictive models," *Journal of Clinical Epidemiology*, vol. 59, pp. 1115–1123, 2006, doi: 10.1016/j.jclinepi.2004.11.029.

[40] N. C. Guan, M. Saiful, and B. Yusoff, “Missing values in data analysis: Ignore or Impute?,” *Education in Medicine Journal*, vol. 3, no. 1, pp. 6–11, 2011, doi: 10.5959/eimj.3.1.2011.or1.

[41] M. L. Brown, J. F. Kros, and N. Carolina, “Data mining and the impact of missing data,” *Industrial Management and Data System*, vol. 103, no. 8, pp. 611–621, 2003, doi: 10.1108/02635570310497657.

[42] S. Juddoo, “Overview of data quality challenges in the context of Big Data,” in *International Conference on Computing, Communication and Security (ICCCS)*, IEEE, 2015.

[43] W. Shi *et al.*, “Improving Power Grid Monitoring Data Quality : An Efficient Machine Learning Framework for Missing Data Prediction,” in *17 th International Conference on High Performance Computing and Communications (HPCC)*, 2015, pp. 417–422. doi: 10.1109/HPCC-CSS-ICESS.2015.16.

[44] S. Jassar, S. Member, Z. Liao, L. Zhao, and S. Member, “Impact of Data Quality on Predictive Accuracy of ANFIS based Soft Sensor Models,” in *Proceedings of the World Congress on Engineering and Computer Science (WCECS)*, 2009.

[45] R. De Jong, “Estimating the Financial Impact of Data Quality Issues Management summary,” University of Twente, 2016.

[46] C.-H. Wang, H.-Y. Cheng, and Y.-T. Deng, “Using Bayesian belief network and time-series model to conduct prescriptive and predictive analytics for computer industries,” *Computers & Industrial Engineering*, vol. 115, pp. 486–494, Jan. 2018, doi: 10.1016/j.cie.2017.12.003.

[47] Y. Dong and C.-Y. J. Peng, “Principled missing data methods for researchers,” *SpringerPlus*, vol. 2, no. 1, p. 222, Dec. 2013, doi: 10.1186/2193-1801-2-222.

[48] M. H. Gorelick, “Bias arising from missing data in predictive models,” *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1115–1123, Oct. 2006, doi: 10.1016/j.jclinepi.2004.11.029.

[49] S. Ferreira Morais, “Dealing with Missing Data: An Application In The Study Of Family History Of Hypertension,” University of Porto, Porto, 2013.

[50] N. Demirel, “The Problem of Missing Data in Regression Analysis,” Dokuz Eylul University, Turkey, 2007.

[51] M. Pampaka, G. Hutcheson, and J. Williams, “Handling missing data: analysis of a challenging data set using multiple imputation,” *International Journal of Research & Method in Education*, vol. 39, no. 1, pp. 19–37, Jan. 2016, doi: 10.1080/1743727X.2014.979146.

[52] M. L. Yadav and B. Roychoudhury, “Handling missing values: A study of popular imputation packages in R,” *Knowledge-Based Systems*, vol. 160, pp. 104–118, Nov. 2018, doi: 10.1016/j.knosys.2018.06.012.

[53] H. Nugroho and K. Surendro, “Missing Data Problem in Predictive Analytics,” in *9th International Conference on Software and Computer Applications (ICSCA 2019)*, Penang: ICSCA 2019, 2019.