Systematic Literature Review: Critical Success Factor in the Application of Data Mining

Aditya Wisnuwardhana*[1], Achmad Nizar Hidayanto[1], Nur Fitriah Ayuning Budi[1], Ika Chandra Hapsari[1], Denny[1] and Ahmad Haidaroh[2]

[1] Faculty of Computer Science, Universitas Indonesia
[2] STIKOM Artha Buana Kupang

Corresponding author e-mail: bontho32@gmail.com

Abstract. The application of Data Mining has become common practice by organizations in an effort to support the achievement of the goals of the organization itself. Therefore it is important for organizations to identify critical success factors (CSF) in implementing Data Mining to prevent failures that can cause organization losses. This study uses a systematic literature review approach to the use of Data Mining with a focus on success factors. A keyword search yielded several CSF findings that had a major influence on different fields of industry. Research studies sourced from 3 online databases, including ScienceDirect, IEEE Xplore, and ProQuest. The disaggregation of the study was carried out by identifying the findings of several CSFs which became the main factors including Stakeholders, Data Availability & Data Quality, Top Management, and Communication.

1. Introduction

The development of the business environment today is very rapid. The activity also produced a large amount of data. At this time many companies began to realize that data has become an important asset for the company. The data can be reprocessed into new information that is useful for companies to help in making business decisions. This process of extracting new information is commonly known as Data Mining (DM).

Although it has been started a long time ago, DM still continues to grow until now. Various studies have shown the benefits of DM in various fields in the last five years, such as in the fields of health, environmental conservation, information technology, and banking. In the field of environmental conservation, DM is used to redefine the distribution of wolves using historical data [1]. In the health sector, DM is used to process electronic data such as insurance claims, medical records, and health surveys, in order to explore information that can be used for example to look at the relationship between diseases, improve cost efficiency, find linkages between diseases, determine patients with high health risks (high-risk), and help doctors by providing a second opinion (second opinion) in establishing the diagnosis [2]. In other research fields show that DM is useful for forming customer profiles [4] and malware detection [5].

Data and information are assets because they can create value [6]. For organizations, data can be used for more effective decision making, in addition to helping organizations operate more efficiently [6]. Furthermore, with the development of technology such as data warehouse, IoT (Internet of Things), and electronic medical records (Electronic Medical Records - EMR), the availability of data

to be processed is more abundant. However, this condition is not always followed by the ability to process and utilize data efficiently. In one study, it was mentioned that large-volume medical datasets collected from hospitals, clinics, and health care providers were mostly not well structured and could not yet be used for analytical purposes [2].

Another challenge in implementing DM appears in a survey in the medical field which states that although DM plays an important role in the development of diagnostic aid models, the reliability and credibility of this decision support system are often doubted because of the lack of collaboration between Data Miners developers and doctors (clinicians), where doctors should be involved since the initial stages of system development and not just at the time of validation of the final results [7].

Data Mining, although it has an important role with certain expectations which should ideally be applied in an organization. Can experience failure or show disappointing results that are far from expectations.

A press release from Gartner 2018 [8], found that 87% of organizations have Low BI (Business Intelligence) and Low Analytics Maturity. Data Mining, which is part of the Business Intelligence (BI) process cycle, has a large role in the success or failure of BI in the organization [9]. This is a major obstacle for any organization that has the desire to increase the value of their data assets.

Current data mining studies are detailed in their approach to methods and techniques. There are still few comprehensive studies discussing from the perspective of the scope of the organization.

Departing from this, this study will review the factors as a key to success in implementing data mining from existing research. With the aim of these success factors can be considered for organizations to increase the value of existing data assets and support in making business decisions.

2. Literature Review

2.1. Data Mining

Data mining (DM) is a decision support analysis process, which is carried out by exploring hidden, valid, and actionable information from company data and presents it so that it can be understood by business stakeholders to support decision making urgent. [9] [10] Simply stated, DM is a process of exploring new information in the form of patterns or rules contained in the data [11].

Unlike the Decision Support System (DSS) in the past which already has assumptions at the beginning (assumption-driven), DM is not directed by the initial assumption (discovery-driven). Instead DM produces a hypothesis as a result, so DM becomes a complement to previous versions of DSS, by analyzing company data automatically to find new insights for the company's business experience [9].

DM is often associated with other terms such as Knowledge Discovery in Databases (KDD), Data Warehousing, and Statistics. Some consider KDD as DM [10], but there are also those who mention DM as part of Knowledge Discovery in Database (KDD) [11].

In the case of DM as part of KDD, KDD is understood as a broader business process, which of several stages, namely data selection, data cleansing, enrichment, data transformation or encoding, and data mining [9]. In addition, reporting & display of the information found can also be included as a KDD process [11].

Related to Data Warehousing (DWH), DM can obtain data from DWH. Therefore, it is not surprising that the success of DM is often associated with the quality of data at DWH. DM tools are called the most effective when using data stored in DWH or data mart, regardless of the ability of DM to process data from other data sources such as flat files [9]. In fact, DM applications are also recommended to be considered early in the DWH design stage because in very large databases (terabytes / petabytes), data warehouse construction is the initial factor that determines the success of data mining applications [11]

In general, the purpose of data mining can be divided into several classes, namely prediction, identification, classification, and optimization [11]: Prediction for example is that DM can see consumer spending behavior when getting certain promotions. Identification for example when DM recognizes intruders who enter the system. Classification for example the use of DM to group data into

several classes based on a combination of certain parameters. While optimization refers to DM's ability to optimize the use of certain resources, such as time and money.

2.2. Critical Success Factors

According to [12] [13] Critical success factors can be said as the factors that determine and measure an organization's performance success for individuals, departments, or organizations. Factors which if considered as satisfaction factors, which will ensure the success of organizational performance. This factor is the key to the success of a business. CSF appears in an industry, a company, a department, even for a manager to pursue corporate strategy.

2.3. Critical Success Factors (CSF) for Data Mining

Some of the factors of the success of data mining according to previous works include several major aspects. References [10] mentions six categories, namely: developer understanding of task domains, human factors from various levels with different training needs, datasets, appropriate tools and scalability, interpretation of results (interpretation), and the use of prior learning (using discovered knowledge) with various documentation. Other studies mention data quality, data integration, technical integration (technical integration), and expertise (expertise) to be some key factors in addition to the scope of the project, timeliness, and resources [14].

3. Research Methodology

The use of this research methodology is the system literature review (SLR) method. Researchers identify and evaluate previous studies with the formulation of problems and related topics [15]. In the absence of a systematic study that conducts a thorough study and analysis of research specifically on the application of Data Mining, the researcher intends to make SLR. The systematic review process with detailed stages will be discussed as follows.

3.1. Formulate the Problem

Finding out the critical success factor is the main objective of this literature review. what critical success factors have a big influence on the success of Data Mining implementation. To achieve these objectives, it is necessary to formulate a research problem with research questions (RQ):

- RQ: What are the critical success factors in using data mining?

3.2. Looking for Literature Study

In the search for literature in this study prioritizes the search for studies in the form of conferences and journals. Keywords are adjusted to the research question and the purpose of this study is included as a search process. In selecting a source that is looking for journals and conferences is a reputable online database, including: ScienceDirect, IEEE Xplore, ProQuest. From the search for these various sources, researchers used the keywords: "(" data mining "OR" using data mining ") AND (success OR factors) AND projects", and "(" critical success factors "OR keys) AND (" data mining "OR" using data mining ") AND" case study "". PRISMA Checklist is shown in figure 1. The studies are collected by considering keyword searches in online databases.

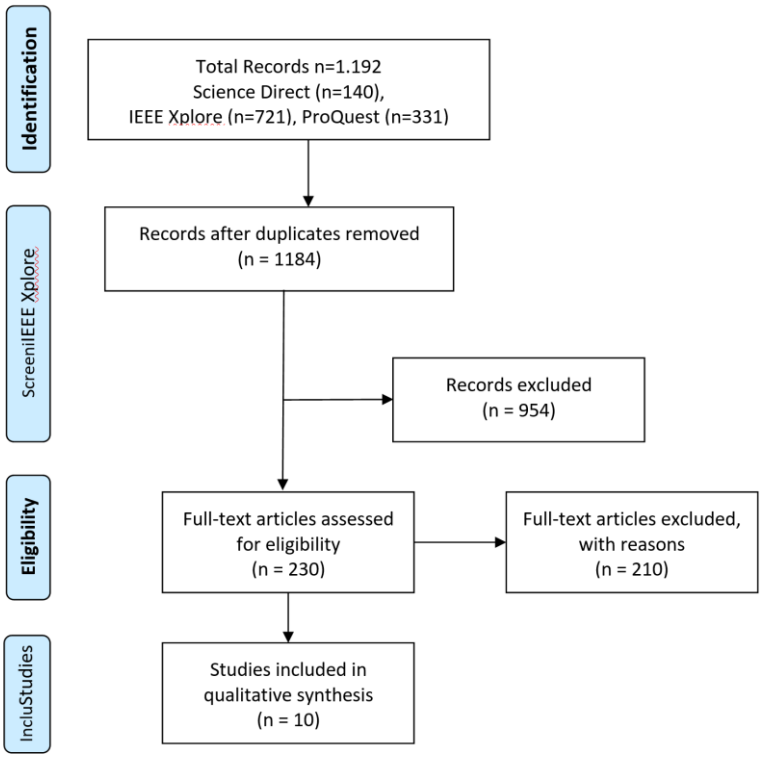

Figure 1. PRISMA Flow Diagram

Table 1. Classification Accuracy Result

Keywords	ScienceDirect	IEEE Xplore	ProQuest
("data mining" OR "using data mining") AND (success OR factors) AND project	85	482	254
(“critical success factors” OR key) AND (“data mining” OR “using data mining”) AND "case study"	55	239	77
Total	140	721	331

3.3. Inclusion and Exclusion Criteria

Determination of inclusion and exclusion criteria that researchers use is to take a study (included in the inclusion criteria) and eliminate the study (exclusion criteria). The following inclusion and exclusion criteria that researchers determine:

1) Inclusion Criteria:
- Studies published between 2014 and 2019
- Studies in the form of journals and conferences written in English
- In accordance with the problem formulation of this study

2) Exclusion Criteria:
- Study is not related to the topic and problem formulation of the research
- Study in the form of thesis and book
- Study writing does not use English
- Studies have duplication

3.4. Study Sorting

After the process of searching for literature with predetermined keywords. Next is to arrange studies that will be candidates based on the inclusion and exclusion criteria that have been determined.

Table 2. Results of Study Search by Inclusion and Exclusion

Keywords	ScienceDirect	IEEE Xplore	ProQuest
("data mining" OR "using data mining") AND (success OR factors) AND project	44	58	55
(“critical success factors” OR key) AND (“data mining” OR “using data mining”) AND "case study"	29	21	23
Total	73	79	78

Next is sorting the study with full text review which includes background, discussion, results, and conclusions, to obtain the chosen study in accordance with this research.

Table 3. Sorting full-text review study

Database	IEEE Xplore	Selected
ScienceDirect	73	6
IEEE Explore	79	1
ProQuest	78	3
Total	230	10

3.5. Conduct Analysis

At this stage the researchers conducted an analysis for each selected study. The analysis is carried out in the form of a summary and mapping of findings that have a relationship with the critical success factors of using data mining in accordance with the previous problem formulation. Next is making the grouping of literature into several categories of industry.

4. Result and Discussion

Based on the results of searches that have been carried out at the methodology stage, the researchers obtained 230 studies sourced from 3 databases, in this case researchers did not get the same type of study from different databases.

Then sorted into 10 studies on the use of CSF in data mining in various fields which become the reference for this research. From these studies identified according to their relevance to the predetermined problem formulation. After that the grouping of studies is carried out as shown in table 4.

Table 4. Grouping CSF studies

Group	Study	Quantity
Critical success factors in data mining	Bole et.al(2015) [28]; Pinzon & Souza(2016) [29]; Aghimien et.al(2018) [30]; Wong et.al(2016) [31]; Son et.al(2015) [32]: Son & Kim(2014) [33]; Larson(2018) [34]; Karaman & Kurt(2015) [35]; Fotopoulou et.al(2015) [36]; Akin et.al(2017) [37]	10

RQ. What are the critical success factors in using data mining?

In order to answer this question a data analysis was made, then a CSF was obtained from 10 previous grouping studies.

Table 5. CSF identification

	Study	Industry	CSF
1	Bole et.al (2015) [28];	Information Management	(1)Business champion (2)External pressure (3)Stakeholder participation (4)Interdisciplinary learning (5)Focus on problem solving action (6)Data availability (7)Data quality
2	Pinzon & Souza(2016) [29]	Land Use Policy	(1)information availability (2)city manager's ability to interpret phenomena to describein theory and verified in a practical manner
3	Aghimien et.al(2018) [30]	Construction	(1)availability of skilled technical and analytical specialist (2)stakeholder awareness of problems, opportunity, and projects issue
4	Wong et.al(2016) [31];	Information System	(1)customer input (2)communication (3)senior management support
5	Son et.al(2015) [32]	Construction	Pre-project planning phase to improve cost performance
6	Son & Kim(2014) [33]	Construction	(1)Pre-project planning (2)Stakeholder to assess and identify potential success
7	Larson(2018) [34]	International Business	Communication
8	Karaman & Kurt(2015) [35]	Business and Economics	(1) Clear Requirements and Specifications (2) Realistic Schedule and Timing (3) Support from Top Management (4) User Involvement (5) Emotional Maturity (6) Optimization (7) Effective Milestone Tracking (8) Strong Business Case (9) Effective Project Management Skills/Methodologies (Project Manager) (10) Commitment of Project Team (11) Clear Objectives and Goals (12) Effective Project Planning (13) Effective Cost Estimating (14) Effective Project Measurements (15) Effective Quality Control (16) Effective Change Management (17) Strong Communication Between Stakeholders (18) Ability Of Project Team (19) Appropriate Technology (20) Continuous Monitoring And Controlling (21) High Quality Dataset
9	Fotopoulou et.al(2015) [37]	Sustainable Smart Cities	(1) Datasets Accessability (2) Linked Data Interoperability
10	Akin et.al(2017) [38]	Social Services	Stakeholder Engagement

Much research has been done in applying data mining methods with a variety of objectives and needs with diverse fields of background.

From table 5, various types of CSFs can be found in the use of data mining with different background fields. From the various types of CSF, we can find the similarity of CSF usage among different fields namely:

(1) Bole et.al (2015) [28], Pinzon & Souza (2016) [29], Karaman & Kurt (2015) [35], all three studies expressed CSF in the form of Top Management support.
(2) Bole et.al (2015) [28], Aghimien et.al (2018) [30], Son et.al (2015) [32], Son & Kim (2014) [33], Akin et.al (2017) [38], the five studies argue that the importance of the role of stakeholders.
(3) Bole et.al (2015) [28], Pinzon & Souza (2016) [29], Karaman & Kurt (2015) [35], Fotopoulou et.al (2015) [36], all four studies mention data Availability and Data Quality are part of CSF.
(4) Wong et.al (2016) [31], Larson (2018) [34], Karaman & Kurt (2015) [35], declare Communication as CSF.

The similarities in the use of CSF were previously made in the order of CSF with the aim of knowing which CSF influence has the most important role in the study of the use of Data Mining in various different fields.

Table 6. CSF Categories by Number of Studies

CSF	Studies
Top Management	3
Stakeholders	5
Data Availability & Data Quality	4
Communication	3

From table 6, we know that the support of stakeholders is the most mentioned of the 10 previous selected studies. The role of stakeholders as first order has the greatest influence on the successful use of data mining in organizations.

The composition of Stakeholders can include IT staff, Data Mining experts, and end users. Their participation is related to the success of a project implementation, especially when system requirements are unclear [39] [40]. But Stakeholders are debated that their involvement in the Information System implementation project affects the priority needs and user requirements [40]. In the case of the application of Data Mining, stakeholders need to participate, namely by their contribution to domain knowledge throughout the process. Their commitment and trust can be achieved by frequent interactions in meetings, explanations and certainty of a decision including by organizing users' expectations settings [42].

Data Availability & Data Quality as part of the data mining process. Data Availability includes the availability of data in accordance with the requirements. Integration of Data Mining with several other systems such as search engines, database systems, data warehouse systems, including cloud computing systems is a supporting factor for the availability of data for further processing [43]. Data Quality consists of several factors including, accuracy, completeness, and consistency. An inaccurate, incomplete, inconsistent data condition is common in large databases and data warehouses. For example, some customer address data that has not been updated for a long time makes the data inaccurate. Errors in the form of errors in data transmission ie the transfer from one device to another can occur and cause an incomplete data [43]. This makes Data Quality an important part of the data mining process, which is categorized as CSF.

The role of Top Management is important in terms of commitment which includes quickly and precisely making decisions to solve problems, also allocating the necessary resources and ensuring coordination between departments [44].

According to [35], Communication is a key success factor that influences a project's success or not. The failure of a project has an impact on costs and loss of investment. Another impact of the project failure is the loss of momentum on the market, which has a direct effect on financial loss.

5. Conclusion

Many types of companies classified as large or small in various fields of academia, construction, business economics have utilized Data Mining. Information gathering from company data assets is done in an effort to support the management elite to make an important decision. In order to support the success in implementing Data Mining, it is necessary to consider special factors that are critical (critical success factors). From this research study, CSF obtained, among others, the role of stakeholders, data availability & data quality, commitment from active top management and communication from every part of the company.

From this research, it can be seen that there have not been many studies on Data Mining that discuss the relationship with the business of a company (Business-driven). Another case with the study of techniques or methods from Data Mining itself. The hope of researchers from this study can open up opportunities for other researchers to conduct other Data Mining case studies with a business perspective.

References

[1] Gallant D, Gauvin L Y, Berteaux D and Lecomte N 2016 *The importance of data mining for conservation science: a case study on the wolverine Biodiversity and Conservation* **25** 2629-2639

[2] Tekieh M H and Raahemi B 2015 *Importance of Data Mining in Healthcare: A Survey* in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15) Paris

[3] Rancati T 2016 *SP-0310: Growing importance of data-mining methods to select dosimetric/clinical variables in predictive models of toxicit* Radiotherapy and Oncology 119

[4] Hassan M M and T M 2018 *Customer Profiling and Segmentation in Retail Banks Using Data Mining Techniques* International Journal of Advanced Research in Computer Science **9** 24-29

[5] Souri A and Hosseini R 2018 *A state-of-the-art survey of malware detection approaches using data mining techniques* Human-centric Computing and Information Sciences **8** 1

[6] Bradley Beach: Technics Publications LLC DAMA International 2017 *DAMA-DMBOK : Data Management Body of Knowledge 2nd ed* 20

[7] Itani S, Lecron F and Fortemps P 2019 *Specifics of medical data mining for diagnosis aid: A survey* Expert Systems With Applications **118** 300-314

[8] Moore S 2018 *Gartner Data Shows 87 Percent of Organizations Have Low BI and Analytics Maturity* Gartner

[9] Anil P and Maheshwari K 2015 *Business and Intelligence Data Mining* New York: Business Expert Press

[10] Kleissner C 1998 *Data Mining for the Enterprise* in Proceedings of the Thirty-First Hawaii International Conference on System Sciences Kohala Coast

[11] Sim J 2014 *Consolidation of Success Factors in Data Mining Projects* GSTF Journal on Computing (JoC) **4** 66-73

[12] Elmasri R and Navathe S B 2012 *FUNDAMENTALS Of Database Systems, 6th ed* Addison-Wesley

[13] Rockart J F 1979 *Chief executives define their own needs* Harvard business review **57** 81-93

[14] Bullen C V and Rockart J F 1981 *A Primer on Critical Success Factors*

[15] Nemati H R and Barko C D 2003 *Key factors for achieving organizational data-mining success* Industrial Management & Daya Systems **103** 282-292

[16] K B and Charters S 2007 *Guidelines for Performing Systematic Literature Reviews in Software*

Engineering

[17] Singhal N G and Hedman K W 2019 *A Data-Driven Reserve Response Set Policy for Power Systems With Stochastic Resources*

[18] Pinelli F, Calabrese F and Bouillet E 2015 *A Methodology for Denoising and Generating Bus Infrastructure Data*

[19] Cheng J C and Ma L J 2014 *A data-driven study of important climate factors on the achievement*

[20] Sousa M D M and Figueiredo R S 2014 *Credit Analysis Using Data Mining: Application In The Case Of A Credit Union*

[21] Lara J A, Lizcano D, Martínez M A and Pazos J 2014 *Data preparation for KDD through automatic reasoning based on description logic*

[22] Lu N, Jiang B and Lu J 2014 *Data mining-based flatness pattern prediction for cold rolling process with varying operating condition*

[23] Zhao X, Zang W, Lv W and Cui W 2018 *Effective Information Filtering Mining of Internet of Brain Things Based on Support Vector Machine*

[24] Chou J S and Pham A D 2014 *Hybrid computationalmodel for predicting bridge scour depth near piers and abutments*

[25] Shameer K, Rodriguez M M P, Bachar R, Li L, Johnson A, Johnson K W and Glicksberg B S 2017 *Pharmacological risk factors associated with hospital readmission rates in a psychiatric cohort identified using prescriptome data mining* in The 7th Translational Bioinformatics Conference Los Angeles

[26] Altujjar Y, Altamimi W, Al-Turaiki I and Al-Razgan M 2016 *Predicting Critical Courses Affecting Students Performance: A Case Study* in Symposium on Data Mining Applications Riyadh

[27] Jun M and Cheng J C 2017 *Selection of target LEED credits based on project information and climatic factors using data mining techniques*

[28] Koskinen J 2018 *How to Build Competencies for a Data-Driven Business: Keys for Success and Seeds for Failure*

[29] Bole U, Popovic A, Zabkar J, Papa G and Jaklic J 2015 *A case analysis of embryonic data mining success*

[30] Pinzón D F D B and Souza F T D 2016 *A data based model as a metropolitan management tool: The Bogotá-Sabana region case study in Colombia*

[31] Aghimien D, Aigbavboa C and Oke A 2018 *A Review of the Application of Data Mining For Sustainable Construction in Nigeria* in 10th International Conference on Applied Energy Hong Kong

[32] Wong T, Chan H K and Lacka E 2016 *An ANN-based approach of interpreting user-generated comments from social media*

[33] Son H, Lee S and Kim C 2015 *An Empirical Investigation of Key Pre-project Planning Practices Affecting the Cost Performance of Green Building Projects* in International Conference on Sustainable Design, Engineering and Construction

[34] Son H and Kim C 2014 *Early prediction of the performance of green building projects using pre-project planning variables: data mining approaches*

[35] Larson D 2018 *Exploring Communication Success Factors in Data Science And Analytics Projects*

[36] Karaman E and Kurt M 2015 *How PMBOK Addresses Critical Success Factors For IT Projects*

[37] Fotopoulou E, Zafeiropoulos A, Papaspyros D, Hasapis P, Tshiolis G, Bouras T, Mouzakitis A

and Zanetti N 2015 *Linked Data Analytics in Interdisciplinary Studies: The Health Impact of Air Pollution in Urban Areas*

[38] Akin B A, Goltzman J S and Camargo C C 2017 *Successes and challenges in developing trauma-informed child welfare systems: A real-world case study of exploration and initial implementation*

[39] Meensel J V, Lauwers L, Kempen I, Dessein J and Huylenbroeck G V 2012 *Effect of a participatory approach on the successful development of agricultural decision support systems: The case of Pigs2win*

[40] Wixom B H and Watson H J 2001 *An Empirical Investigation of the Factors Affecting Data Warehousing Success*

[41] Yeoh W and Koronios A 2009 *Critical Success Factors for Business Intelligence Systems*

[42] Viaene S and den Bunder A B 2011 *The Secrets to Managing Business Analytics Projects*

[43] Han J, Kamber M and Pei J 2012 *Data Mining Concepts and Techniques* Morgan Kaufmann

[44] Leyh C 2014 *Critical Success Factors for ERP Projects in Small and Medium-sized Enterprises in The Perspective of Selected German SMEs*

Acknowledgement
This study was supported by PIT9 Research Grant No NKB-0005/UN2.R3.1/HKP.05.00/2019 Universitas Indonesia.