Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 2, 4303 – 4310

Big Data Analytics in Library to Classification Book Publishers

[1]Ronald Maraden Parlindungan Silalahi, [2]Johanes Fernandes Andry, [3]Devi Yurisca Bernanda, [4]Hendy Tannady, [5]Enirianti

[1]*English Department, Universitas Bunda Mulia, Jakarta, Indonesia 14430, bomberrose@gmail.com*
[2]*Information Systems Department, Universitas Bunda Mulia, Jakarta, Indonesia 14430, jandry@bundamulia.ac.id*
[3]*Information Systems Department, Universitas Bunda Mulia, Jakarta, Indonesia 14430, dbernanda@bundamulia.ac.id*
[4]*Management Department, Kalbis Institute, Jakarta, Indonesia, hendytannady@gmail.com*
[5]*Information Systems Department, Universitas Bunda Mulia, Jakarta, Indonesia 14430, enirianti17@gmail.com*

Abstract

The library is a place for people to read books and a place that has various kinds of books to read. In the library has a lot of data from borrower data, library membership data, book data, and book return data. Therefore, big data is needed to process these data. Big data is data that contains a lot of information in various forms. This big data also has 3 characteristics, namely volume, velocity, and variation. Data mining is analyzing data using various methods and producing useful information for companies or organizations. The method used in this analysis is a decision tree classification method. Classification is the process of analyzing data by predicting and classifying data so that it can produce useful information. The application used in this analysis is the Rapidminer Studio application. This application is an application that is used to analyze data and generate information from the data analysis. The purpose of this study was to analyze book data according to the publisher and the year the book was published. This decision tree method will be used to predict book data based on the publisher and year of publication. To test data accuracy using cross validation. The results of the test show that the data accuracy rate is 40.52% with the prediction results of Gramedia Pustaka Utama 30.00%; prediction results of Elex Media Komputindo 40.78%; the results of the library span prediction 0.00%; Grasindo prediction results 45.07%; and the prediction of Andi Publisher 38.10%.

Keywords: Big Data, Library, Data Mining, Rapid Miner, Classification.

INTRODUCTION

Libraries have a large amount of data and information, namely knowledge books, research articles, and some reports of any kind, currently, it is possible in a physical format that can be touched or also in digital form can be in pdf format, mp3, etc. [1].

This situation provides real rationale for evaluating these factors, as it is explained that the growth of data on a large scale is referred to as "Big Data" [2]. As libraries must assess their resources and services to support data-based decisions, this panel will focus on the perspective and agenda of library data analysis / assessment in the big data era [3].

Data is not new to Libraries and Big Data provides a quantitative extension of the well-known object of study [4]. Big data presents new challenges for librarians, whose roles are changing rapidly. Library services need to use big data to handle millions of online sources.

The role of the librarian moves from paper recording and document creation to the digital and online world [5]. In many cases, a librarian and a professional can assist researchers who need guidance on various aspects of working with data [6]. Library sector has a tradition of organizing, managing, retrieving, collecting, describing, and preserving information [7].

Big data is a phenomenon caused by the rapid flow of information. The issue of big data in libraries has begun to be discussed. Meanwhile, its application in the public or government sector is apparently still limited [8]. This problem aims to address not only the problems of library management and innovation exposed by innovative library applications, technologies and services, but also technical and managerial approaches, methods, and solutions that will address the challenges faced by librarians in the era of big data [9].

For academic libraries, Big Data analytics is affected by two basic challenges: first, because of the enormous volume, selection and speed of knowledge, the storage and processing requirements of the system are somewhat redundant, and secondly, the complex analytical techniques and algorithms, which make Big Data analytics become computationally intensive task [10]. This data includes e-book circulation, browsing history and reader book downloads, daily traffic logs of the library site, and other information [11].

LITERATURE REVIEW

This section will briefly explain some of the important concepts that underlie this research, including Big Data, Libraries and Classification in classifying book publishers by utilizing Big Data Analytics.

2.1 Big Data

Big Data is like small data, but bigger in size but having bigger data requires a different approach: Techniques, tools and architectures aim to solve new or old problems in a better way. Big data generates value from the storage and processing of digital information in a very large amount that cannot be analyzed by traditional computational techniques [12].

Big Data is data that contains high-volume, high-speed, and very diverse information assets that require new forms of processing so that they can improve decision making, gain new insights, and maximize process optimization [13]. While the characteristics of the big data that have been mentioned previously relate to something that is data that contains data uncertainty and has value benefits from the information generated [14].

2.2 Libraries

Most of the institutions such as libraries, of course, want to be successful in big data jobs which really have to be clear about their needs and even then, have to work efficiently with big data. librarian competency requirements in the digital era are demanded to be higher to upgrade their knowledge and skills in accordance with current changes [15].

The development of creation, acquisition, storage, and flow of Big Data has presented several challenges and opportunities for contemporary libraries or what they call modern libraries. Advances in communication and information technology have changed the modified organizational structure to keep up with environmental and social changes in society [16].

Libraries as information centers that are accessed by many people must be able to adjust and accommodate the growth of data, resources, and data provision. Libraries are very useful in data management as part of the information service process. Information services today have created a data boom in which libraries are required to improve in four main areas: (1) library organization, (2) enhancing internal data sets, (3) awareness of the power of external data sources for libraries and (4) increasing sources human resources with certain skills, especially librarians [17]. The advent of Big Data forced libraries to redesign their service patterns which they sometimes had to complete their operations. And the emerging Big Data trend is very helpful for library improvement [18].

2.3 Classification

The classification algorithm in data mining is capable of processing large amounts of data. It can be used to predict categorical class labels and classify data based on training sets and class labels and can be used to classify newly available data [19]. The classification process itself consists of two stages, namely the learning process and the testing process. The learning process is an algorithm classification process, managing the analysis of the training dataset to form classification rules, while the testing process is a process that uses a testing dataset to test the accuracy of predefined classification rules [20].

METHODOLOGY

Data mining is the process of analyzing data in various angles and summarizing the results into useful information. Data mining software is an analytical tool that allows users to analyze data from various dimensions, data categories, and summarize the relationships between data. Technically, data mining is the process of finding correlations among many fields in a very large data set [21].

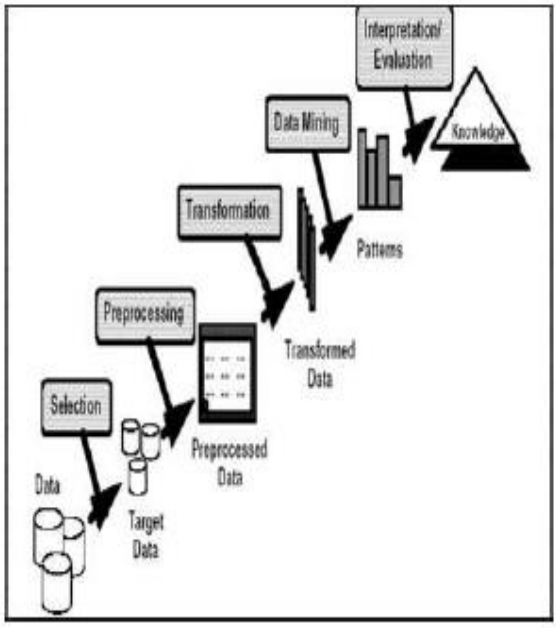

Figure 1. *Stages of Data Mining [22].*

The knowledge discovery process consists of several sequential and iterative methods as follows:

(1) Selection: Selecting data relevant to the task of a database analyst.

(2) Initial processing: Delete invalid data and inconsistent data; combine multiple data sources.

(3) Transformation: Converting data into a form suitable for data mining.

(4) Data Mining: Selecting a data mining algorithm that matches a pattern in its data properties; extract various data patterns.

(5) Interpretation / Evaluation: Interpreting various patterns into knowledge by eliminating various irrelevant patterns and the same and repetitive patterns; translating various useful patterns in terms that can be understood by ordinary people [23].

The following research methods used in this paper are shown in Figure 1. Data Mining Stages [24], [25].

ANALYSIS AND RESULTS

The magnitude of the significance that affects the assessment of book publishers in distributing published e-books according to browsing history and e-books downloaded by readers, can benefit publishers in using so much information.

Based on these findings, this section describes the stages of the Knowledge Discovery Database and Data Mining process which includes 5 methods, namely Data Selection, Data Preprocessing, Data Transformation, Data Mining Implementation and Data Interpretation / Evaluation.

This research will use the KDD stage using Microsoft Excel and Rapid Miner Studio. The stages of the KDD process carried out are as follows:

1. Data Selection

In this study, the data set to be used is data called Book Data which is available and stored in the library. This dataset consists of 460 rows of book data, consisting of 3 columns containing information about the book title, publisher name and the year the book was published.

The following is a list of attributes in this study and their explanations:

(1) Book Title: data containing various book titles from the library.

(2) Name of Publisher: data containing the name of the publisher, including Gramedia Pustaka Utama, Elex Media Komputindo, Bentang Pustaka, Grasindo, and Publisher Andi.

(3) Year of publication: data containing the number of years published for each book from 2007 to 2020.

The following are some samples from the book dataset that will be used in this study from a total of 460 lines of the book dataset.

2. Initial Data Processing

At the preprocessing stage, the existing book data set was checked using Microsoft Excel. This data checking process is useful for cleaning inconsistent data, correcting letter or word errors and correcting words in the data so that they are easier to read and can be processed correctly for this research.

3. Data Transformation

At this data transformation stage, selected data that has been carried out in the Preprocessing stage will be imported from Microsoft Excel into Rapid Miner to facilitate classification of Book Data.

For each data set, a data type based on the following 3 columns will be used for test data on Rapid Miner, the details are as follows:

(1) Book title, using Polynominal type data.

(2) Name of Publisher, using Polynominal data type.

(3) Year of publication, using Integer type data

4. Implementation of Data Mining

At this stage the data modeling process is carried out using the decision tree method on the data that has been collected, selected, transformed and managed from the previous KDD process stages to produce book classifications based on the decision tree results.

By using the decision tree algorithm, the data will be minimized and also changed so that the data only displays the attributes needed in the decision tree and displays all data in the form of a tree root image containing data that has been converted into existing data arranged.

Table 1: *Book Data*

Book Title	**Name of Publisher**	**Year of Issue**
Beresin Dulu Hidupmu!	Gramedia Pustaka Utama	2020
Falling Leaves Never Hate the Wind	Gramedia Pustaka Utama	2018
How to Win Friends and Influence People in the Digital Age	Gramedia Pustaka Utama	2015
Tak Mungkin Membuat Semua Orang Senang	Gramedia Pustaka Utama	2019
Segala Sesuatu Terjadi Untuk Sebuah Alasan	Elex Media Komputindo	2020
Investment Guide Series: Psychology Of Investing	Elex Media Komputindo	2019
Unleash The Real You	Elex Media Komputindo	2016
Value Investing Beat The Market In Five Minutes!	Elex Media Komputindo	2020
Average People	Bentang Pustaka	2017
Harmoni Itu Kita	Bentang Pustaka	2019
Gelandangan Di Kampung Sendiri	Bentang Pustaka	2018
Find A Way To My Heart	Bentang Pustaka	2019
Atlas Boy Adventure : Coloring Book	Grasindo	2020
Will The Real You Please Stand Up	Grasindo	2018
Teach Like Finland	Grasindo	2019
Lead or Handover to Millennial	Grasindo	2020
Coreldraw X8 For Beginners	Andi Publisher	2018
Smart Leadership: To Be The # 1 Decision Maker	Andi Publisher	2017
Kreasi Digital Dengan Photoshop Untuk Pemula	Andi Publisher	2018

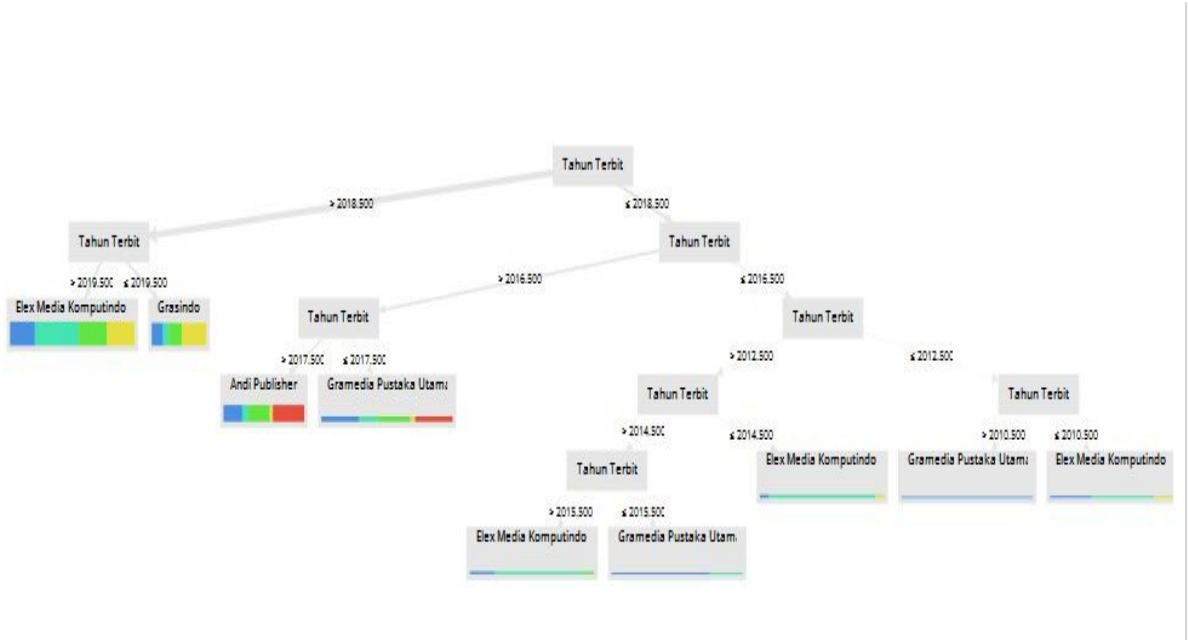

Figure 2: *Decision Tree Results*

Based on the Decision Tree Result in Figure 2 above, we are shown the number of years of publication and also the book publisher, where the book information itself will be explained through the book publisher, then the book title, and then the year the book was published.

For each book publisher, there will be re-recording of the number of books that have been submitted annually, starting from 2007 to 2020. Book data has increased every year and the number of books published has also decreased.

Meanwhile, the most stable year occurred in 2018 to 2020 where it was seen that book publishing was always above 500 books, with book publisher Gramedia Pustaka Utama publishing the highest number of books among other book publishers.

And also when we enter data in Rapid Miner, we will get several types of attributes used:

(1) Binominal: data that describes the data used can be marked as 0 or 1 or it can also be called a yes or no statement.

(2) Polynominal: data that has more than 2 types of data types used.

(3) Real: data used to describe data in decimal form.

(4) Integer: data in the form of integers, unlike real data types which are decimals.

Therefore, the following is a detailed explanation of the Text View obtained from Rapid Miner for classifying Book Data:

Tree

Year of publication> 2018,500

| Publish year> 2019.500: Elex Media Komputindo {Gramedia Pustaka Utama = 30, Elex Media Komputindo = 53, Library Span = 34, Grasindo = 33, Publisher Andi = 0}

| Publish year ≤ 2019,500: Grasindo {Gramedia Pustaka Utama = 30, Elex Media Komputindo = 14, Span Pustaka = 34, Grasindo = 64, Publisher Andi = 0}

Year of publication ≤ 2018,500

| Year of publication> 2016,500

| | Year of publication> 2017.500: Publisher Andi {Gramedia Pustaka Utama = 24, Elex Media Komputindo = 8, Bentang Pustaka = 27, Grasindo = 4, Publisher Andi = 40}

| | Publish year ≤ 2017.500: Gramedia Pustaka Utama {Gramedia Pustaka Utama = 8, Elex Media Komputindo = 4, Bentang Pustaka = 7, Grasindo = 1, Publisher Andi = 8}

| Publication year ≤ 2016,500

| | Publication year> 2012,500

| | | Year of publication> 2014,500

| | | | Publish year> 2015.500: Elex Media Komputindo {Gramedia Pustaka Utama = 2, Elex Media Komputindo = 7, Bentang Pustaka = 1, Grasindo = 0, Publisher Andi = 0}

| | | | Publish year ≤ 2015.500: Gramedia Pustaka Utama {Gramedia Pustaka Utama = 3, Elex Media Komputindo = 1, Bentang Pustaka = 0, Grasindo = 0, Publisher Andi = 0}

| | | Publish year ≤ 2014.500: Elex Media Komputindo {Gramedia Pustaka Utama = 1, Elex Media Komputindo = 12, Bentang Pustaka = 0, Grasindo = 1, Publisher Andi = 0}

| | Publication year ≤ 2012,500

| | | Publish year> 2010,500: Gramedia Pustaka Utama {Gramedia Pustaka Utama = 2, Elex Media Komputindo = 0, Rentang Pustaka = 0, Grasindo = 0, Publisher Andi = 0}

| | | Published year ≤ 2010.500: Elex Media Komputindo {Gramedia Pustaka Utama = 2, Elex Media Komputindo = 3, Bentang Pustaka = 0, Grasindo = 1, Publisher Andi = 0}

Based on the explanation from Text View above, the classification of the name of the book publisher uses the year the book was published from 2007 to 2020.

Furthermore, testing is carried out using Cross Validation to test the accuracy of the data, as shown in Figure 3 below.

In the cross validation process, there is a training sub-process using a decision tree and a testing sub-process using an applied model operator and operator performance.

Select retrieval operator then enter book data to retrieval operator in fast miner design process area, then enter decision tree operator into field, connect both operators, then enter applied model and performance operator to design field and then re-enter that performance data into testing data then connect with the applicable operator model, as shown in Figure 3 Cross Validation and Figure 4 Sub-Process Cross Validation.

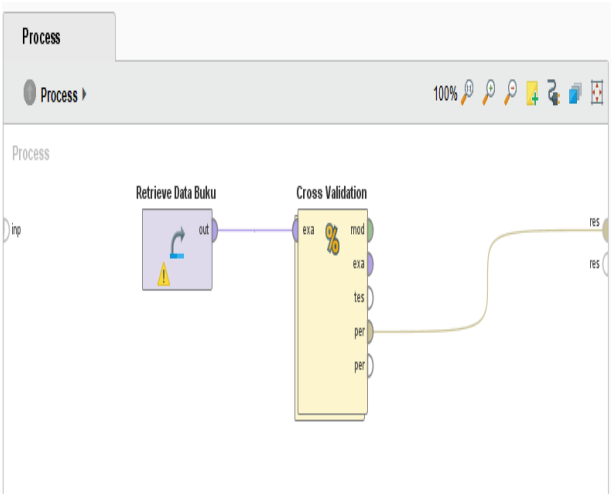

Figure 3: *Cross Validation*

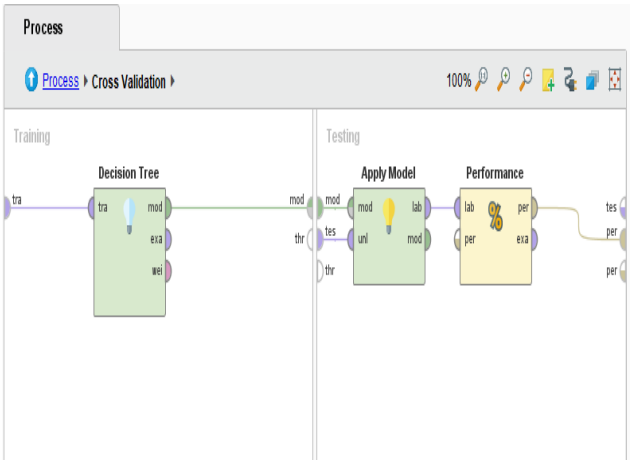

Figure 4: Sub-Process Cross Validation

5. Data Interpretation / Evaluation

At this stage the data accuracy is tested so that the data obtained is accurate. The results of the cross-validation test using operator performance showed the accuracy rate of book publishers was 40.52% with a micro average of 40.52%. With Gramedia Pustaka Utama's prediction results for class precision of 30.00%, Elex Media Komputindo's prediction results for class precision of 40.78%, Bentang Pustaka prediction results for class precision 0.00%, Grasindo prediction results for class precision 45.07%, Andi Publisher prediction results of 38.10% grade precision. As for the class recall of Gramedia Pustaka Utama's prediction results for class recall of 8.82%. The prediction result of Elex Media Komputindo for class recall is 71.57%. The results of Bentang Pustaka prediction for the 0.00% class recall. Grasindo prediction results for the class recall of 61.54%.

Andi Publisher's prediction results for 83.33% class recall. More details about the results of Accuracy Performance can be seen in Figure 5 Performance Accuracy.

accuracy: 40.52% +/- 3.08% (micro average: 40.52%)

	true Gramedia P...	true Elex Media K...	true Bentang Pus...	true Grasindo	true Andi Publisher	class precision
pred. Gramedia ...	9	7	6	1	7	30.00%
pred. Elex Media ...	36	73	35	35	0	40.78%
pred. Bentang Pu...	2	0	0	0	1	0.00%
pred. Grasindo	30	14	34	64	0	45.07%
pred. Andi Publis...	25	8	28	4	40	38.10%
class recall	8.82%	71.57%	0.00%	61.54%	83.33%	

Figure 5: *Performance Accuracy*

CONCLUSION

The research results obtained from the book data using classification methods with operator performance. We can determine the accuracy level of book publishers by 40.52% with a micro average of 40.52%.

With Gramedia Pustaka Utama's prediction results for 30.00% class precision, Elex Media Komputindo's prediction results for 40.78% class precision, Bentang Pustaka prediction results for 0.00% class precision, Grasindo prediction results for 45.07% class precision, The prediction results of Andi Publisher for class precision 38.10%.

As for the class recall of Gramedia Pustaka Utama's prediction results for class recall of 8.82%, Elex Media Komputindo's prediction for 71.57% class recall, Bentang Pustaka prediction results for 0.00% class recall, Grasindo prediction results for 61 class recall, 54%, and the prediction of Andi Publisher for class recall of 83.33%.

With this prediction, it can be seen how many books are in the library of each book publisher based on the year published. Through the results of this research is expected to be useful for librarians in compiling books in the library.

Reference

[1] C. Wang, S. Xu, L. Chen, and X. Chen, "Exposing library data with big data technology: A review," 2016 IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. ICIS 2016 - Proc., 2016, doi: 10.1109/ICIS.2016.7550937.

[2] K. Ahmad, Z. JianMing, and M. Rafi, "An analysis of academic librarians competencies and skills for implementation of Big Data analytics in libraries: A correlational study," Data Technol. Appl., vol. 53, no. 2, pp. 201–216, 2019, doi: 10.1108/DTA-09-2018-0085.

[3] H. L. Chen, P. Doty, C. Mollman, X. Niu, J. C. Yu, and T. Zhang, "Library assessment and data analytics in the big data era: Practice and policies," Proc. Assoc. Inf. Sci. Technol., vol. 52, no. 1, pp. 1–4, 2015, doi: 10.1002/pra2.2015.14505201002.

[4] K. Golub and J. Hansson, "(Big) data in library and information science: A brief overview of some important problem areas," J. Univers. Comput. Sci., vol. 23, no. 11, pp. 1098–1108, 2017.

[5] C. C. Chang, "Hakka genealogical migration analysis enhancement using big data on library services," Libr. Hi Tech, vol. 36, no. 3, pp. 426–442, 2018, doi: 10.1108/LHT-08-2017-0172.

[6] L. Federer, "Research data management in the age of big data: Roles and opportunities for librarians," Inf. Serv. Use, vol. 36, no. 1–2, pp. 35–43, 2016, doi: 10.3233/ISU-160797.

[7] C. R. Sugimoto, Y. Ding, and M. Thelwall, "Library and information science in the Big Data era: Funding, projects, and future [a panel proposal]," Proc. ASIST Annu. Meet., vol. 49, no. 1, 2012, doi: 10.1002/meet.14504901187.

[8] N. E. V. Anna and E. F. Mannan, "Big data adoption in academic libraries: a literature review," Libr. Hi Tech News, vol. 37, no. 4, pp. 1–5, 2020, doi: 10.1108/LHTN-11-2019-0079.

[9] S. Liu and X. L. Shen, "Library management and innovation in the Big Data Era," Libr. Hi Tech, vol. 36, no. 3, pp. 374–377, 2018, doi: 10.1108/LHT-09-2018-272.

[10] H. Al-Barashdi and R. Al-Karousi, "Big Data in academic libraries: literature review

and future research directions," J. Inf. Stud. Technol., vol. 2018, no. 2, 2019, doi: 10.5339/jist.2018.13.
[11] S. Li, F. Jiao, Y. Zhang, and X. Xu, "Problems and Changes in Digital Libraries in the Age of Big Data From the Perspective of User Services," J. Acad. Librariansh., vol. 45, no. 1, pp. 22–30, 2019, doi: 10.1016/j.acalib.2018.11.012.
[12] M. P. Kumar, S. S. Kumar, and G. Ramya, "Big Data Analytics: A Brief Survey," Int. J. Trend Sci. Res. Dev., vol. Volume-2, no. Issue-4, pp. 2264–2268, 2018, doi: 10.31142/ijtsrd15617.
[13] M. A. Beyer and D. Laney, The Importance of "Big Data": A Definition. Stamford, CT: Gartner, 2012.
[14] L. Douglas, 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note 6, 2001.
[15] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and Challenges of Big Data Research," Big Data Res., vol. 2, no. 2, pp. 59–64, 2015, doi: 10.1016/j.bdr.2015.01.006.
[16] M. Zhan and G. Widén, "Understanding big data in librarianship," J. Librariansh. Inf. Sci., vol. 51, no. 2, pp. 561–576, 2019, doi: 10.1177/0961000617742451.
[17] L. Reinhalter and R. J. Wittmann, "The Library: Big Data's Boomtown," Ser. Libr., vol. 67, no. 4, pp. 363–372, 2014, doi: 10.1080/0361526X.2014.915605.
[18] A. Affelt, The Accidental Data Scientist: Big Data Applications and Opportunities for Librarians and Information Professionals. Medford, NJ: Information Today, 2015.
[19] R. P. R. S. Suryakirani, "Comparative Study and Analysis of Classification Algorithms In Data Mining Using Diabetic Dataset," Ijsrst , vol. 4, no. 2, pp. 299–304, 2018, [Online]. Available: www.ijsrst.com.
[20] M. F. A. Saputra, T. Widiyaningtyas, and A. P. Wibawa, "Illiteracy Classification Using K Means-Naïve Bayes Algorithm," JOIV Int. J. Informatics Vis., vol. 2, no. 3, p. 153, 2018, doi: 10.30630/joiv.2.3.129.
[21] K. Sumathi, S. Kannan, and K. Nagarajan, "Data Mining: Analysis of student database using Classification Techniques," Int. J. Comput. Appl., vol. 141, no. 8, pp. 22–27, 2016, doi: 10.5120/ijca2016909703.
[22] Michael J.A. Berry and G. S. Linoff, Data Mining Techniques For Marketing, Sales, and Customer Relationship Management Second Edition, Second Edi. Wiley Publishing, Inc., 2004.
[23] A. V. D. Sano and H. Nindito, "Application of K-Means Algorithm for Cluster Analysis on Poverty of Provinces in Indonesia," ComTech Comput. Math. Eng. Appl., vol. 7, no. 2, p. 141, 2016, doi: 10.21512/comtech.v7i2.2254.
[24] E. D. Madyatmadja, Marvel, J. F. Andry, H. Tannady and A. Chakir, "Implementation of Big Data in Hospital using Cluster Analytics," 2021 International Conference on Information Management and Technology (ICIMTech), 19-20 August 2021.
[25] E. D. Madyatmadja, A. Rianto, J. F. Andry, H. Tannady and A. Chakir, "Analysis of Big Data in Healthcare Using Decision Tree Algorithm, Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI).