Analysis of Big Data in Healthcare Using Decision Tree Algorithm

Evaristus Didik Madyatmadja
Information Systems Department
School of Information Systems
Bina Nusantara University
Jakarta, Indonesia 11480
emadyatmadja@binus.edu

Antonius Rianto
Information Systems Department,
Faculty Technology and Design
Universitas Bunda Mulia
Jakarta, Indonesia 14430
antoniusrianto444@gmail.com

Johanes Fernandes Andry
Information Systems Department,
Faculty Technology and Design
Universitas Bunda Mulia
Jakarta, Indonesia 14430
jandry@bundamulia.ac.id

Hendy Tannady
Department of Management,
Institut Teknologi dan Bisnis Kalbis
Jakarta, Indonesia 13210
hendy.tannady@kalbis.ac.id

Aziza Chakir
Faculty of Law Economics and Social Sciences,
Economic Department
Hassan II University
Morocco 8110
aziza1chakir@gmail.com

2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI) | 978-1-6654-4002-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCSAI53272.2021.9609734

***Abstract*—Era of technological developments, big data has been widely implemented in various any company especially healthcare. Big data has opened up new gaps in health care. Big data in healthcare has the potential to improve better healthcare. The effective use of Big data can reduce health care problems such as how to provide proper care, maximum care solutions, and improve existing systems of health care. There are 6 defining domains in Big Data, which are Vol., and etc. Big data represents a variety of opportunities to improve the quality and efficiency of healthcare. Big Data in healthcare need to expanded and explore utilize big data analytics to gain valuable knowledge. Big data analytics is used to catch value any information from all kinds of sources in healthcare that can be used to gain information for the purpose of better decision making in healthcare. Big data analytics in healthcare has the prospect to increased healthcare by discovering decision tree and understanding formats and trends in medical record data. Cardiovascular illness datasets is big data in healthcare which is one or others resources in the health sector and is used as part of facilitating the process of documenting medical records that must be analysed to offer an effective solution to solve problems in healthcare. This paper provides valuable information by using big data analytics from medical data cardiovascular disease to provide effective solutions for the problems in healthcare and also provide how important big data for healthcare is.**

Keywords— Healthcare, Analysis of Big Data, Medical Records.

I. INTRODUCTION

Right now, era of big data, information technology has been widely utilized in enterprise. Medical data are continuously explosively, there are challenges of the 21st century for managed of data, repository, and tabulation [1]. An efficient data acquisition, to process and to consumption methodology has been a theme of great attention for decades over enterprise [2]. Rich source of data has the potential to enhance understanding of disease mechanisms and better health care [3].

Big data point to wide and complex data sets that are outside the ability of classic database management systems to put away, to manage and to process [4]. Big Data can develop various challenges in data retrieval, to transferred, encryption, storage, analysis and to make visualization. Healthcare relies on medical data in the decision-making process. Big data analytics can be used to achieve valuable information from all sorts of sources that are too large, raw, or unstructured in healthcare [5]. BDA has the possibility to increase healthcare by find out associations and understanding of design and trends in the medical record [6].

In others countries, the healthcare deals with huge volumes of electronic health data such as cardiovascular disease [7]. Big Data analytics can be used to achieve valuable information from large and complicated datasets such as cardiovascular disease to improve medical treatment and healthcare [5]. Cardiovascular disease is one of others the services in the healthcare and is used as a part to facilitate the process of documenting medical records [8]. Big medical data have a big analytical capability that can be used to provide productive solutions to solve the problems in the healthcare [9].

II. BACKGROUND STUDY

A. Big Data

Big Data (BD) generally refers to the enormous volumes of data that the usual data tools and practices are not ready to handle and presents unprecedented opportunities to advance science and inform resource management through data-intensive approaches, and big data technologies are enabling new types of activism environment in the process [10].

BD has the features of wide-scale, high multi-dimensional, diversity, more-complex, unstructured, not complete, and crowded that makes it feasible to gather valuable data and information [11].

B. Big Data Healthcare

Big healthcare data to covers collected large collections of data from any various healthcare foundations followed by stored, managed, analyzed, visualized, and delivered any information for effective conclusions. Big healthcare from primary sources such as clinical DSS, electronic health records etc.) And secondary sources such as from laboratories, insurance firm sources, pharmacies, etc [12].

III. RESEARCH METHOD

Figure 1 shows the stages from research, firstly from problem formulation, next step is data collection from kaggle, and then authors will analysis from that source, after that, make a report or visualizations, next step is evaluation, when the data is ok, the stage is finish.

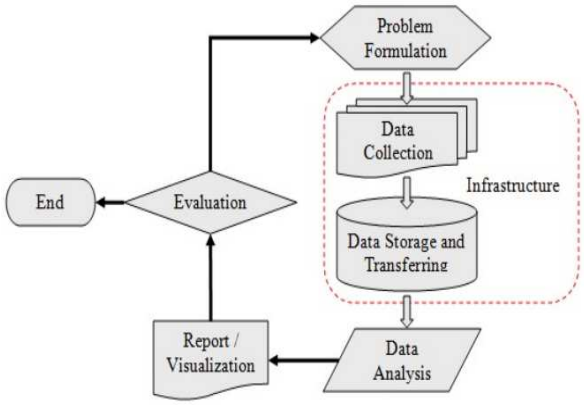

Fig. 1. Research Stages [13]

IV. RESULT AND DISCUSSION

For Massive amounts of data, driven by record keeping, regulatory compliance & requirements, and patient care are generated by healthcare. Healthcare used big data analytics to analyze data to get valuable information to improve healthcare performance.

RapidMiner is software that authors use to analyze data with algorithms to get useful information for healthcare. Authors analyzed the data using a classification method and a decision tree algorithm to classify cardiovascular disease which can be used as predictive and prescriptive analytics. Predictive analysis predicts what might happen in the future and prescriptive analysis recommends actions that can be taken to act on these results.

Authors used Cardiovascular disease datasets that were collected at the moment of medical examination in this research. There are 3 types of features in this datasets:

1. Goal: Factual any data and information
2. Calibrations: Outcome of medical
3. Subjective: Information any specified by the patient

This dataset has 70K rows records of patient's data and 12 attributes of information about patients and the results of medical examination. The following are the attributes of the cardiovascular disease datasets:

1. Age: factual information about patient's age (in days).
2. Height: factual information about patient's height (in cm).
3. Weight: factual information about patient's weight (in kg).
4. Gender: factual information about patient's gender (1 is a woman and 2 is a man).
5. Systolic blood pressure: results of medical examination from patient's systolic blood pressure (in mmHg).
6. Diastolic blood pressure: results of medical examination from patient's diastolic blood pressure (in mmHg).
7. Cholesterol: results of medical examination from patient's indicated cholesterol (1 for N = Normal; 2 for AN = Above Normal; 3 for WAN = well above normal).
8. Glucose: results of medical calibration from patient's glucose (1: normal; 2: above normal; 3: well above normal).
9. Smoking: information given by the patient about smoking (0: non-smoker; 1: smoker).
10. Alcohol intake: information given by the patient about alcohol intake (0: not an alcohol drinker; 1: alcohol drinker).
11. Physical activity: information given by the patient about physical activity (0: no physical activity; 1: have physical activity).
12. Presence or absence of cardiovascular disease: information from analytics using decision tree algorithm about the presence of cardiovascular disease (0: absence; 1: presence).

Preprocessing phase, will transform raw data into a useful and efficient format [14]. Authors explore our cardiovascular disease datasets to cleaning data. In this phase, and identify missing attributes and blank fields, cleaning or replacing missing values, duplicate or wrong data, and inconsistent data [15]. Examine the data for completeness, correctness, and consistency. Problematic data that has not been identified and analyzed can produce misleading results. This phase is important to produce good results by processing the analyzed data

After preprocessing phase, then put the data into RapidMiner to start the analytics. Before authors doing analytics to the data, set the attribute type based on data type, then have to set the attribute type correctly in order to produce correct results. Authors will analyze the data using decision tree algorithm in RapidMiner to do analytics. Decision tree algorithm is used to classify presence or absence of cardiovascular disease. First, process the data using decision tree algorithm to generate the rules and decision tree and then analyze the results. Figure 2 shows the result of decision tree.

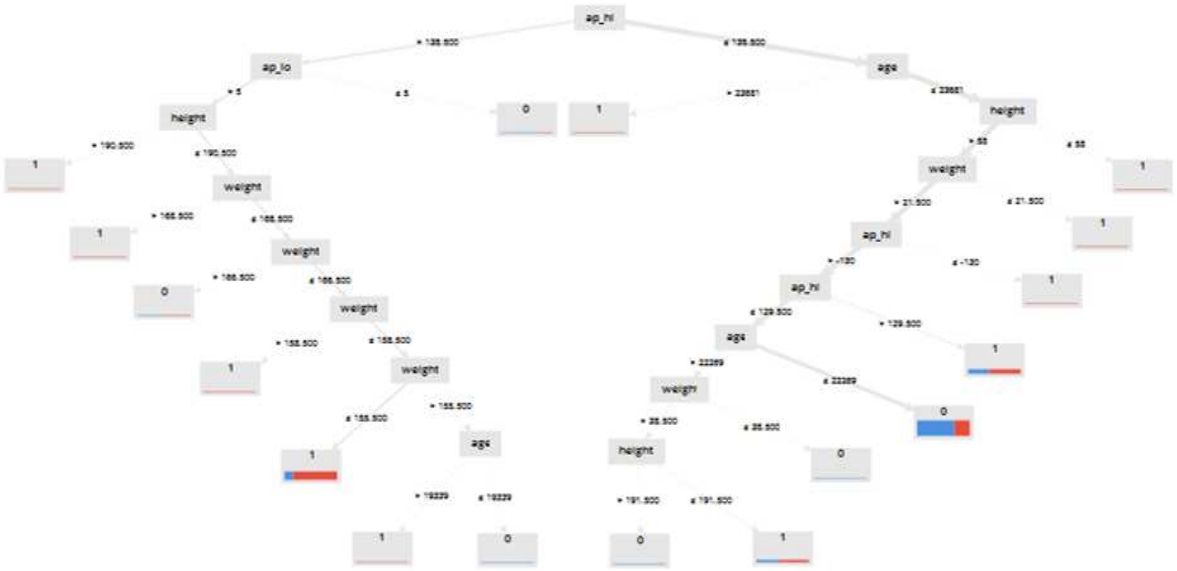

Fig. 2. Decision Tree Results

Figure 2 shows the result of decision tree and the hidden pattern that generated from cardiovascular disease datasets using decision tree algorithm in RapidMiner. Decision tree algorithm provide decision tree to find the classification rules from the data. In that decision tree, the root node or predictor is ap_hi (systolic blood pressure), internal nodes are the other attributes that contains in the data and the leaf node is the cardio. The result of decision tree from cardiovascular disease is used to explain or understand the result from classification based from the other attributes as the root node and internal nodes that determine.

The advantage of decision tree is that can be explained as a rules or description from decision tree. Rules or description from decision tree is an if-else statements. Following are the rules from the decision tree:

ap_hi > 138.500
| ap_lo > 5
| | height > 190.500: 1 {0= 0, 1= 20}
| | height ≤ 190.500
| | | weight > 168.500: 1 {0= 0, 1= 9}
| | | weight ≤ 168.500
| | | | weight > 166.500: 0 {0= 1, 1= 1}
| | | | weight ≤ 166.500
| | | | | weight > 158.500: 1 {0= 0, 1= 9}
| | | | | weight ≤ 158.500
| | | | | | weight > 155.500
| | | | | | | age > 19339: 1 {0= 0, 1= 2}
| | | | | | | age ≤ 19339: 0 {0= 3, 1= 0}
| | | | | | weight ≤ 155.500: 1 {0= 3130, 1= 16225}
| ap_lo ≤ 5: 0 {0= 8, 1= 3}
ap_hi ≤ 138.500
| age > 23681: 1 {0= 0, 1= 5}
| age ≤ 23681
| | height > 58
| | | weight > 21.500
| | | | ap_hi > -130
| | | | | ap_hi > 129.500: 1 {0= 3719, 1= 5541}
| | | | | ap_hi ≤ 129.500
| | | | | | age > 22269
| | | | | | | weight > 38.500
| | | | | | | | height > 191.500: 0 {0= 3, 1= 0}
| | | | | | | | height ≤ 191.500: 1 {0= 1853, 1= 2382}
| | | | | | | weight ≤ 38.500: 0 {0= 4, 1= 0}
| | | | | | age ≤ 22269: 0 {0= 26300, 1= 10776}
| | | | ap_hi ≤ -130: 1 {0= 0, 1= 2}
| | | weight ≤ 21.500: 1 {0= 0, 1= 2}
| | height ≤ 58: 1 {0= 0, 1= 2}

These rules are generated from the decision tree starting from the root node or predictor until the leaf node. These rules give a clear analytical view of the result from the decision tree. Data can understand all the process from the decision tree using these rules.

After the decision tree is generated from the cardiovascular disease datasets using RapidMiner, and test for the performance. Next are performing performance testing to determine whether the analyzed data is accurate or not. Table 1 shows the result from performance testing.

TABLE I. RESULT OF PERFORMANCE TESTING TOPIC

Accuracy: 72.17%			
	true 0	**true 1**	**class precision**
pred. 0	26319	10780	70.94%
pred. 1	8702	24199	73.55%
Class recall	75.15%	69.18%	

Figure 2 is the result from the performance testing of classification of the cardiovascular disease datasets in RapidMiner. The performance testing shows the level of accuracy, class precision and class recall. Authors obtained 72.17% for the level of accuracy in classification. The class precision for prediction 0 or Absence is 70.94% and the class precision for prediction 1 or Presence is 73.55%. For the class recall, obtained 75.15% for true 0 or Absence and 69.18% for true 1 or Presence. The measured level of accuracy, precision and recall achieved a high performance. This means that a classification improved the efficiency and effectiveness for cardiovascular disease.

Big data analytics with classification method using decision tree algorithm for cardiovascular disease can improve the efficiency and effectiveness. The decision tree provides us explanation for the hidden pattern from the data can understand the information that gets from the data using decision tree algorithm. Next authors can analyze the results from classification using decision tree as predictive or prescriptive analytics. Healthcare can predict a patient who has cardiovascular disease and provide preventive care to patient. Big data in healthcare is important; here is the opportunity for healthcare that implemented big data:

1. Improved Preventive Care

Big data analytics using medical data in healthcare can improve prevention for their patient. With big data analytics, healthcare can capture, analyze and compare patient symptoms. Healthcare with improved preventive care can treat the patient well and prevent or delay the illness and disease. As in the research, then classify the cardiovascular disease so the healthcare can know their right treatment for their patient more effective and efficient.

2. Improved Diagnostic Symptoms

By doing big data analytics, healthcare improved their diagnostic symptoms for their patient. A diagnostic symptom is a process to determine patient with disease. Improved diagnostic symptoms using the knowledge collected from the hidden pattern of patient's data. With improved diagnostic symptoms, healthcare can diagnose their patient more effective and efficient. As in the research, can classify the cardiovascular disease that has various attributes that determine the disease so the healthcare can diagnose their patient more accurately from their symptoms.

3. Reducing Healthcare Cost

Big data can help reduce the cost of providing medical treatment. BDA for healthcare can carry out valuable about any information to improve their system by the medical record data. With analyzed data, their improved medical treatment can analyze and diagnose their patient more accurately, effective and efficient. With more accurately, effective and efficient healthcare, patient will pay less and get correct treatment than the ordinary medical treatment.

V. CONCLUSION

In this research, authors used classification methods with decision tree algorithm for cardiovascular disease datasets. The results are the decision tree and rules that classify availability of inclined cardiovascular illness. Test for the performance to determine whether the analyzed data is accurate or not. Get obtained 72.17% for the level of accuracy in classification. The class precision for prediction 0 or Absence is 70.94% and the class precision for prediction 1 or Presence is 73.55%. Class recall, obtained 75.15% for true 0 or Absence and 69.18% for true 1 or Presence. The analyzed data can be used as predictive and prescriptive analytics.

Big data analytics with classification method using decision tree algorithm for cardiovascular disease can improve the efficiency and effectiveness for healthcare. Healthcare can predict a patient who has cardiovascular disease and provide preventive care to patient. With the help of big data, healthcare has the opportunity to improve better healthcare, such as improved preventive care, improved diagnostic symptoms, and reducing healthcare cost.

In further research, classification methods using decision tree algorithm can be used for another datasets and can be developed by combining or comparing using other classification algorithms to get better results.

REFERENCES

[1] Y. Zhang, M. Qiu, C. W. Tsai, M. M. Hassan, and A. Alamri, "Health-CPS: Healthcare cyber-physical system assisted by cloud and big data," IEEE Syst. J., vol. 11, no. 1, pp. 88–95, 2017, doi: 10.1109/JSYST.2015.2460747.

[2] T. Schultz, "Turning healthcare challenges into big data opportunities: A use-case review across the pharmaceutical development lifecycle," Bull. Am. Soc. Inf. Sci. Technol., vol. 39, no. 5, pp. 34–40, 2013, doi: 10.1002/bult.2013.1720390508.

[3] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: A patient-centered framework," J. Gen. Intern. Med., vol. 28, no. SUPPL.3, pp. 660–665, 2013, doi: 10.1007/s11606-013-2455-8.

[4] R. Nambiar, R. Bhardwaj, A. Sethi, and R. Vargheese, "A look at challenges and opportunities of Big Data analytics in healthcare," Proc. - 2013 IEEE Int. Conf. Big Data, Big Data 2013, pp. 17–22, 2013, doi: 10.1109/BigData.2013.6691753.

[5] L. Wang and C. A. Alexander, "Big Data Analytics in Healthcare Systems," Int. J. Math. Eng. Manag. Sci., vol. 4, no. 1, pp. 269–276, 2018, doi: 10.1109/ICoAC44903.2018.8939061.

[6] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," Heal. Inf. Sci. Syst., vol. 2, no. 1, pp. 1–10, 2014, doi: 10.1186/2047-2501-2-3.

[7] U. Srinivasan and B. Arunasalam, "Leveraging big data analytics to reduce healthcare costs," IT Prof., vol. 15, no. 6, pp. 21–28, 2013, doi: 10.1109/MITP.2013.55.

[8] A. Prasad and S. Prasad, "Imaginative geography, neoliberal globalization, and colonial distinctions: Docile and dangerous bodies in medical transcription 'outsourcing,'" Cult. Geogr., vol. 19, no. 3, pp. 349–364, 2012, doi: 10.1177/1474474012445734.

[9] M. Bouhriz and H. Chaoui, "Big data privacy in healthcare moroccan context," Procedia Comput. Sci., vol. 63, pp. 575–580, 2015, doi: 10.1016/j.procs.2015.08.387.

[10] S. S. Hasan, Y. Zhang, X. Chu, and Y. Teng, "The role of big data in China's sustainable forest management," For. Econ. Rev., vol. 1, no. 1, pp. 96–105, 2019, doi: 10.1108/fer-04-2019-0013.

[11] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," J. Big Data, vol. 2, no. 1, pp. 1–32, 2015, doi: 10.1186/s40537-015-0030-3.

[12] S. SA, “Big Data in Healthcare Management: A Review of Literature,” Am. J. Theor. Appl. Bus., vol. 4, no. 2, p. 57, 2018, doi: 10.11648/j.ajtab.20180402.14.

[13] D. P. Acharjya and K. A. P, “A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 2, pp. 511–518, 2016, doi: 10.26438/ijcse/v6i6.12381244.

[14] Madyatmadja, Evaristus Didik, and Mediana Aryuni. 2014. “Comparative Study of Data Mining Model for Credit Card Application Scoring in Bank.” Journal of Theoretical and Applied Information Technology 59 (2): 269–74.

[15] Aryuni, Mediana, and Evaristus Didik Madyatmadja. 2015. “Feature Selection in Credit Scoring Model for Credit Card Applicants in XYZ Bank: A Comparative Study.” International Journal of Multimedia and Ubiquitous Engineering 10 (5): 17–24. https://doi.org/10.14257/ijmue.2015.10.5.03.