International Journal of Informatics and Communication Technology (IJ-ICT)
Vol.10, No.3, December 2021, pp. 182~187
ISSN: 2252-8776, DOI: 10.11591/ijict.v10i3.pp182-187

Electronic health record to predict a heart attack used data mining with Naïve Bayes method

Johanes Fernandes Andry[1], Fabio Mangatas Silaen[2], Hendy Tannady[3], Kevin Hadi Saputra[4]
[1,2,4]Department of Information System, Universitas Bunda Mulia, Jakarta, Indonesia
[3]Department of Management, Universitas Pembangunan Jaya, Banten, Indonesia

Article Info

Article history:

Received Mar 13, 2021
Revised Sep 1, 2021
Accepted Oct 11, 2021

Keywords:

Cardiovascular diseases
Data mining
Electronic health record
Heart attack
Naïve Bayes

ABSTRACT

A heart attack is a medical emergency. A heart attack usually occurs when a blood clot blocks the flow of blood to the heart. Cardiovascular disease is a variety of diseases that attack the body's cardiovascular system including the heart and blood vessels. Cardiovascular diseases (CVD) include angina, arrhythmia, heart attack, heart failure, atherosclerosis, stroke, and so on. To resolving (CVD) is to evaluate large scores of datasets, to compare for any information that can be used to forecast, to take care of organize. The method used Naïve Bayes classification because that method can determine target which can be used to answer some questions like whether the patient has the potential for heart disease. After data analyst, authors can use data to electronic health records (EHR).

Corresponding Author:

Johanes Fernandes Andry
Department of Information System, Universitas Bunda Mulia
Jl. Ancol Barat IV, RT.12/RW.2, Ancol, Kec. Pademangan, D.K.I Jakarta 14430, Indonesia
Email: jandry@bundamulia.ac.id

1. INTRODUCTION

Cardiovascular disease (CVD) is the number one deadliest disease in the world and is on the rise in Asia. There are a number of factors that cause an increase in CVD such as sedentary lifestyle, unhealthy diet, and smoking. But you can change your life and reduce your risk of CVD and improve your quality of life [1]. Atherosclerosis is a chronic inflammatory disease; it's described the patchy intramural thickening of the subintima [2]. Cardiovascular disease (CVD) circulatory system which includes the heart and blood vessels. The circulatory system is important for keeping the body's organs functioning by transporting oxygen, nutrients, electrolytes, and hormones throughout the body. But when there is a disturbance or blockage in the heart or blood vessels, it will affect blood circulation and cause complications such as heart disease or stroke [3]. Acute myocardial infarction (AMI), or often referred to as a heart attack, is a decrease in blood flow in the coronary arteries due to occlusion, which is mostly caused by the process of atherosclerosis. Meanwhile, risk factors can be distinguished between modifiable risk factors and non-modifiable risk factors [4].

The use of big data from datasets can improve services to patients, detect the spread of disease early, generate new insights into disease mechanisms, monitor the quality of medical and health institutions and provide better treatment methods [5]. Cardiovascular disease has risk factors. Risk factors are a measure to determine the likelihood, can be seen in Table 1. Big data is a very large and quite complex way of collecting data where conventional data processing methods are not good enough. Therefore, big data will be analyzed so that patterns, or other habits, related to the organization or customers can be obtained [6]. Big data analysis refers to proper and good analysis so that it can be ensured that the decision-making process can be more accurate and the results of good performance again [7]. Characteristics of dig data is shown in Table 2.

Table 1. Risk factors can be modifiable and nonmodifiable

Modifiable	Nonmodifiable
• She or he to smoke or vape environmental, • Obesity are is an excessive accumulation of fat due to an imbalance between energy intake and energy used • lifestyle of Sedentary (rarely physical activity), • Diabetes disease, a chronic disease characterized by high blood sugar (glucose) levels • High cholesterol is condition when cholesterol levels in the blood exceed normal limits. • Hypertension is condition when blood pressure is at 130/80 mmHg or more	• Gender, man is more at risk of coronary heart disease than women. Risk factors in women will increase after experiencing menopause • Age, a person's risk increases with age. Usually at the age of 40 years, a person is advised to start checking his heart health • Gen factor, heredity from a family who has had a heart attack

Table 2. Big data characteristics

No.	Big data feature	Explanation	Illustration
1	Volume	Capacity of data	Amount of data collected and stored. Capacity in MB, GB, and TB [8].
2	Velocity	Speed of data	The transfer rate of data among resource and objective [9].
3	Variety	Kind of data	Different type of data like image, movie, and sound [10].
4	Value	Importance of data	It's indicated the business value derived from big data [11].
5	Variability	Data differentiation	It's indicated to changes in data during processing and lifecycle [12].
6	Veracity	Quality of data	It's indicated 2 aspects: consistency of data and trustworthiness of data [13].

After these patterns are found, they can be used to make certain decisions for further business development [14]. Some of the steps involved in it are:

- To explore of data: The data is cleaned in the sense that nothing is lost and transformed into a different form so that other important variables which then type the data based on the problem have been determined.
- Pattern identification: Form pattern identification. Identify and choose the pattern which make the best prediction.
- Deployment: Patterns are deployed for the desired outcome.

Data mining is the process of analyzing data from different angles and summarizing results into useful information [15]. Data mining is an automated data analysis techniques to uncover previously undetected relationships among data items. Data mining also often involves the analysis of data stored in a data warehouse [16]. Data mining techniques can be applied in various aspects because data obtained from different sources can be different and out of sync. Specific technique is applied for specific types of problems to resolve efficiently [17]. Technique's data mining [18]:

- Classification, this technique usually uses machine learning or machine learning techniques. This technique classifies items or variables in a data set into predetermined groups or classes. It uses linear programming, statistics, decision trees, and artificial neural networks, among other techniques.
- Clustering, in clustering the data labeling process is not determined at the beginning, in contrast to the data group labeling classification that has been determined previously. Examples of clustering methods are K-means, C-means).
- Regression is a technique used for determining that there is a relationship between the variable that's wanting to predict (the dependent variable) and other variables (the independent variable).

Big data, in medical research, is used for electronic health record (EHR) considered "relevant" to the understanding of health and disease, including clinical, imaging, omics, data from internet use and wearable devices, and others [19]. In health care institutions, data mining tools answer the question rapidly, that are traditionally time-consuming and too complex to resolve [20]. Electronic health record (EHR) is facilitated services in terms of patient medical records. The EMR system or electronic medical record is a systematic collection of electronic-based health information that is connected and integrated with the information system in the hospital network [21]. Medical records are written or recorded information regarding identity, history taking, physical determination, laboratory, diagnosis of all medical services and actions provided to patients and treatment, both inpatients, outpatients and those receiving emergency services.

2. RESEARCH METHOD

The working of the method is described in a step by step [22]: (a) Data Selection: obtain the data resources from various sources. (b) Data preprocessing: is refer to manipulation or dropping of data before it is used in order to ensure or enhance performance and is an important step in the data mining process from the dataset. (c) Data analyst: one of data mining techniques are applied to get results. (d) Implementation: the

results from applied data mining techniques in RapidMiner application. The step of research method as shown in Figure 1.

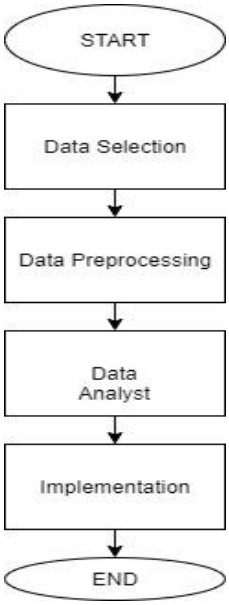

Figure 1. Research methodology step

3. RESULTS AND DISCUSSION

3.1. Data selection

Big data in healthcare refers to the vast quantities of data-created by the mass adoption of the Internet and digitization of all sorts of information, including health records- too large or complex for traditional technology to make sense of. This clinical activity produces a large number of prints including patient record any information, diagnoses, treatment schemes, notes from doctors, and sensor data [23].

The dataset that will be past in this research is the "Heart disease UCI" dataset. This dataset is obtained from hospital in Indonesia. This dataset contains 14 attributes, the explanation of each attribute can be seen Table 3. However, this data must be pre-processed. Data pre-processing is one of the tasks in data mining, including the preparation and conversion of data into a form suitable for mining procedures [24].

3.2. Data preprocessing

In data preprocessing, the software used in this methodology is RapidMiner. By utilizing RapidMiner, the data processing process, to determine the variables that will be used in the process of grouping data, to clean up unwanted data.

3.3. Data analyst

Data analysis looks at the existing data and implements statistical and visualization methods to test hypotheses about the data and find exceptions. Data mining looks for and finds trends in the data, which can be used for further analysis in the future. Classification algorithms learn the labels of the samples and their nominal and/or numeric values as attributes and they create a model. After that, they make predictions about these generated models [25]. Naïve Bayes classification is a probabilistic model based on Naïve Bayes theorem. Naïve Bayes defined as a statistical classification. Naïve Bayes used for supervised learning [26]. The data mining extension (DMX) query language is used to create models, model training, model predictions, and model content access. All parameters are set to default settings except for the parameters "Minimum dependency probability = 0.05" for Naïve Bayes [27]. In this paper, we use the Naïve Bayes classification algorithm. The Naïve Bayes is a simple probabilistic classifier that is easy to apply and it performs can well on data sets with a high number of instances [28]. The rules of Naïve Bayes [29].

$$P(C|E)\frac{P(E|C))P(C)}{P(E)}. \quad (1)$$

Table 3. Attributes datasets of heart disease

Age	The length of time that a person has lived or a thing has existed, age: 28 until age: 78
Sex	Refers to a set of biological attributes in humans. It is primarily associated with physical and physiological features including chromosomes, gene expression, hormone levels and function, and reproductive/sexual anatomy. [“o” = female and “1” = male]
CP	Can be divided into heart-related chest pain (cardiac chest pain) and chest pain that is not from a heart condition (non-cardiac chest pain).
Trestbps	The pressure of circulating blood against the walls of blood vessels. Most of this pressure results from the heart pumping blood through the circulatory system. [min. Bp = 92, max. Bp = 202]
Chol	Is biosynthesized by all animal cells and is an essential structural component of animal cell membranes. It is a yellowish crystalline solid. Min. Chol = 124, max. Chol = 566]
FBS	A blood sample will be taken after an overnight fast. A fasting blood sugar level less than 100 mg/dl than 120mg/dl sugar = one, if not = zero.
Resting Electrocardiographic	The heart is a muscular organ which pumps blood through rhythmic contractions induced by electric impulses generated by the sinus node, the heart's natural pacemaker. 0 = normal, 1 = has ST-T wave abnormality (T wave inversion), 2 = shows left ventricular hypertrophy.
Thalac	Is the maximum heart rate of patient?
Exang	Exercise induced angina(exang), ST depression induced by. Exercise relative to rest (old peak), the slope of the peak. Exercise ST segment(slope), number of major vessels. If Y, the value will be “1”, and “N” for not.
Old Peak	At entry indicates severe coronary lesions and large benefits of an early invasive treatment strategy in unstable coronary artery disease between 0 and 6.2.
Slope	In a cardiac stress test, an ST depression of at least 1 mm after adenosine administration indicates a reversible ischemia, while an exercise stress test requires an ST depression of at least 2 mm to significantly indicate reversible ischemia.
CA	Fluoroscopy is a type of medical imaging that shows a continuous X-ray image on a monitor, much like an X-ray movie.
Thal	A thallium stress test is a nuclear medicine study that shows your physician how well blood flows through your heart muscle while you're exercising or at rest.

3.4. Implementation

The classification technique is used to create a model that can be used to predict whether a patient with a certain attribute has stroke or not. To do this, we reduced the attributes of the dataset according to the stroke risk factors mentioned above. The attributes are ‘age’, ‘gender’, ‘hypertension’, ‘avg_glucose_level’ to indicate whether someone has diabetes, ‘heart_disease’, ‘Body mass index (BMI)’ to indicate whether someone is obese, and ‘smoking statuses. The ‘stroke’ attribute is also included as the label/class. Figure 2 shows how the operators in RapidMiner are configured to build the decision tree model. Before the optimize parameter, operator is the configuration for cleaning and reducing the dataset. This operator is a wrapper operator used to tune the parameters of the operator inside it. After being plugged into the optimize parameter operator, the dataset is split into training and testing data with 7:3 ratio respectively. Then, the training data is plugged into the decision tree operator to construct the decision tree model, with the criterion parameter set to ‘information_gain’. Other parameters such as maximal_depth, minimal_leaf_size, confidence, and ‘minimal_size_for_split will be tuned by the wrapper. The model then passed to the apply model operator together with the testing data. Which then passed to the performance operator to evaluate the accuracy of the decision tree model. The association technique is used to create association rules to find associations of the attributes in the dataset that are related to stroke. The FP-Growth operator we used accepts attributes with nominal or categorical values. Therefore, we chose the attributes ‘gender’, ‘heart_disease’, ‘hypertension’, ‘smoking_statuses, and ‘stroke’.

Figure 3 shows the configuration of the operators used to create the association rules. The first 5 operators were the same as the one used on classification technique, with exception of the select attributes operator that now only selects the attributes mentioned above. The reduced dataset is connected to the FP-growth operator with the parameter min_support is set to 0.3 and other parameters left default. Then the frequent itemset from the operator is passed to the create association rules operator to create the association rules. The parameter nonconfidence is set to 0.5, and other parameters are also left default.

Figure 4 shows the clusters made by the clustering operator. There, we can see the 2 clusters with the average values of the attributes selected above. The first cluster (cluster_0) has 708 items and the second (cluster_1) has 4234 items. As we can see, the first cluster has the highest relative proportion of patients that have stroke at 12%. This cluster is consisted of patients with an average BMI of 31, age of 58, and average glucose level of 201. From this result, we can create an assumption that elder patients that are considered obese and have diabetes are more likely to have stroke. In the second cluster, only 3,7% of the 4234 patients have stroke. Which is consisted of patients within the age of 40, BMI of 27, and average glucose level of 90.

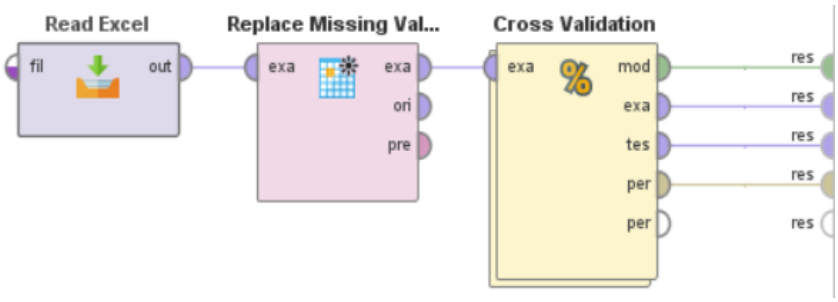

Figure 2. Operator configuration model

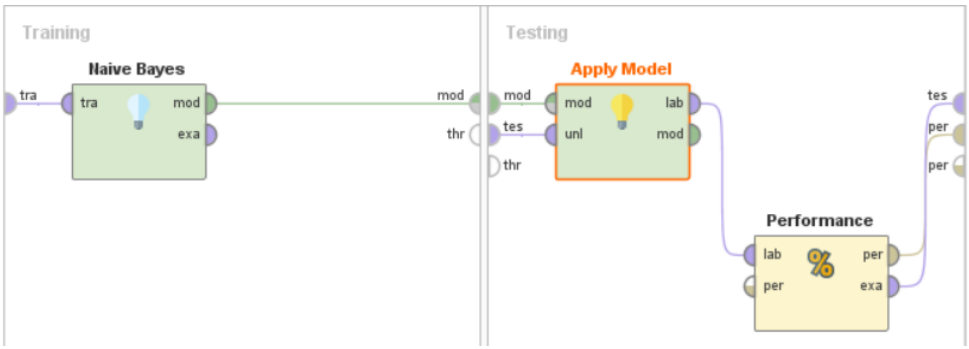

Figure 3. Operator configuration for the technique

Attribute	cluster_0	cluster_1
stroke	0.120	0.037
bmi	30.958	27.719
age	58.135	40.715
avg_glucose_level	201.393	89.528

Figure 4. Clusters made by the clustering operator

4. CONCLUSION

In this study, we can predict whether a person would potentially have heart disease or not. Therefore, the person can be treated before the disease gets worse or even prevent the disease from happening together. Classification is an appropriate data mining technique for processing heart disease datasets because the dataset used has target variables to be classified. Classification will classify data into groups of classes that already exist. There will be no formation of new groups. And the process is supervised. Different from clustering which is a process for grouping data into several clusters or groups so that the data in one cluster has similarities.

Almost every day there is an increase in the ratio of the art attacks. To reduce heart disease a system is needed to detect potential heart attacks. Big data must be analyzed first before taking a pattern that is useful for decision making. RapidMiner is used in this study to predict patients diagnosed with a heart attack using data mining techniques Naïve Bayes. We use naïve Bayes classification because in naïve Bayes classification we can determine target which can be used to answer some questions like whether the patient has the potential for heart disease. After data analyst, we can use data to electronic health record (EHR).

REFERENCES

[1] G. Vilahur, J. J. Badimon, R. Bugiardini, and L. Badimon, "Perspectives: The burden of cardiovascular risk factors and coronary heart disease in Europe and worldwide," *European Heart Journal Supplements*, vol. 16, no. suppl A, pp. A7–A11, 2014, doi: 10.1093/eurheartj/sut003.

[2] A. Maria and Y. KS, "Pathogenesis of Atherosclerosis A Review," *Medical & Clinical Reviews*, vol. 2, no. 3, 2016, doi: 10.21767/2471-299x.1000031.
[3] E. Nason, "An overview of cardiovascular disease and research," 2007. [Online]. Available: https://www.rand.org/content/dam/rand/pubs/working_papers/2007/RAND_WR467.pdf.
[4] I. Al Mamoon *et al.*, "A Proposal of Body Implementable Early Heart Attack Detection System," *Malaysia Japan International Institue of Technology (MJIIT)*, no. 1–4, pp. 1–4, 2013.
[5] B. Ristevski and M. Chen, "Big Data Analytics in Medicine and Healthcare," *Journal of integrative bioinformatics*, vol. 15, no. 3, p. 20170030, May 2018, doi: 10.1515/jib-2017-0030.
[6] N. Zulkarnain and M. Anshari, "Big data: Concept, applications, & challenges," *2016 International Conference on Information Management and Technology (ICIMTech)*. IEEE, 2016, doi: 10.1109/icimtech.2016.7930350.
[7] H. J. Watson, "Tutorial: Big Data Analytics: Concepts, Technologies, and Applications," *Communications of the Association for Information Systems*, vol. 34, 2014, doi: 10.17705/1cais.03465.
[8] D. Cackett, "Information Management and Big data: A Reference Architecture," *Oracle: Redwood City, CA, USA*, 2013.
[9] B. Feldman, E. M. Martin, and T. Skotnes, "Big data in healthcare hype and hope," *Dr. Bonnie*, vol. 360, pp. 122–125, 2012.
[10] Z. Sun, K. Strang, and R. Li, "Big Data with Ten Big Characteristics," in *Proceedings of the 2nd International Conference on Big Data Research*, 2018, pp. 56–61, doi: 10.1145/3291801.3291822.
[11] A. Oguntimilehin and E. O. Ademola, "A review of big data management, benefits and challenges," *A Review of Big Data Management, Benefits and Challenges*, vol. 5, no. 6, pp. 1–7, 2014.
[12] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.
[13] S. Sami and N. Sael, "Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 3, 2016, doi: 10.14569/ijacsa.2016.070337.
[14] M. Bharati and M. Ramageri, "Data mining techniques and applications," 2010.
[15] K. Sumathi, S. Kannan, and K. Nagarajan, "Data Mining: Analysis of student database using Classification Techniques," *International Journal of Computer Applications*, vol. 141, no. 8, pp. 22–27, 2016, doi: 10.5120/ijca2016909703.
[16] H. Sahu, S. Shrma, and S. Gondhalakar, "A brief overview on data mining survey," *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, vol. 1, no. 3, pp. 114–121, 2011.
[17] H. A. Madni, Z. Anwar, and M. A. Shah, "Data mining techniques and applications — A decade review," *2017 23rd International Conference on Automation and Computing (ICAC)*. IEEE, 2017, doi: 10.23919/iconac.2017.8082090.
[18] F. Zuha and G. Achuthan, "Analysis of Data Mining Techniques and its Applications," *International Journal of Computer Applications*, vol. 140, no. 3, pp. 6–14, 2016, doi: 10.5120/ijca2016909249.
[19] H. Hemingway *et al.*, "Big data from electronic health records for early and late translational cardiovascular research: challenges and potential," *European heart journal*, vol. 39, no. 16, pp. 1481–1495, Apr. 2018, doi: 10.1093/eurheartj/ehx487.
[20] S. Thankachan and Suchithra, "Data Mining & Warehousing Algorithms and its Application in Medical Science-A Survey," 2017.
[21] P. C. McMullen *et al.*, "Electronic Medical Records and Electronic Health Records: Overview for Nurse Practitioners," *The Journal for Nurse Practitioners*, vol. 10, no. 9, pp. 660–665, 2014, doi: 10.1016/j.nurpra.2014.07.013.
[22] E. D. Madyatmadja, D. J. M. Sembiring, S. M. B. P. Angin, D. Ferdy, and J. F. Andry, "Big Data in Educational Institutions using RapidMiner to Predict Learning Effectiveness," *Journal of Computer Science*, vol. 17, no. 4, pp. 403–413, 2021, doi: 10.3844/jcssp.2021.403.413.
[23] L. Hong, M. Luo, R. Wang, P. Lu, W. Lu, and L. Lu, "Big Data in Health Care: Applications and Challenges," *Data and Information Management*, vol. 2, no. 3, pp. 175–197, 2018, doi: 10.2478/dim-2018-0014.
[24] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, 2017.
[25] B. Kolukisa *et al.*, "Diagnosis of coronary heart disease via classification algorithms and a new feature selection methodology," *International Journal of Data Mining Science*, vol. 1, no. 1, pp. 8–15, 2019.
[26] B. Afeni, T. Aruleba, and I. Oloyede, "Hypertension Prediction System Using Naive Bayes Classifier," *Journal of Advances in Mathematics and Computer Science*, vol. 24, no. 2, pp. 1–11, 2017, doi: 10.9734/jamcs/2017/35610.
[27] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *2008 IEEE/ACS International Conference on Computer Systems and Applications*. IEEE, 2008, doi: 10.1109/aiccsa.2008.4493524.
[28] S. Nikhar and A. M. Karandikar, "Prediction of heart disease using machine learning algorithms," *International Journal of Advanced Engineering, Management and Science*, vol. 2, no. 6, p. 239484, 2016.
[29] T. K and M. Wadhawa, "Analysis and Comparison Study of Data Mining Algorithms Using Rapid Miner," *International Journal of Computer Science, Engineering and Applications*, vol. 6, no. 1, pp. 9–21, 2016, doi: 10.5121/ijcsea.2016.6102.