Indonesian Journal of Electrical Engineering and Computer Science
Vol. 23, No. 1, July 2021, pp. 362~369
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v23.i1.pp362-369

Formulation of city health development index using data mining

Bertalya, Prihandoko, Lilis Setyowati, Febrian Iftikhar Irawan, Syahifa Rahmita Irlianti
Faculty of Computer Science and Information Technology, Gunadarma University, Jawa Barat, Indonesia

Article Info

Article history:

Received Mar 6, 2021
Revised May 24, 2021
Accepted Jun 14, 2021

Keywords:

City health development index
Data mining
Health indicator

ABSTRACT

Every five years public health research publishes a public health development index that describes public health in indonesia. The public health development index is measured using data from the public health research and the national socio-economic survey, and the village potential survey which is obtained by surveying from sampling data. In fact, the provincial and city health offices have health profile data reports every year. For this reason, this study analyzes existing health profile data using data mining techniques to obtain indicator data that are very influential in formulating the city health development index. This city health development index was successfully formulated by adopting the model of public health development index in 2013 and using indicators from annual health profile data which obtained from the data mining process, i.e., Random Forest algorithm. The proposed model can be used as the annual report of a city to describe the health condition of that city. For the future research, the model can be adopted to measure some specific aspects of city health condition.

Corresponding Author:

Bertalya
Faculty of Computer Science and Information Technology
Gunadarma University
Jl. Margonda Raya, No.100, Pondok Cina, Depok, Jawa Barat, Indonesia
Email: bertalya@staff.gunadarma.ac.id

1. INTRODUCTION

Law Number 17 of 2007 concerning the national long term development plan (2005-2025) states that health together with education and increasing the purchasing power of the community are the three main pillars to improve the quality of human resources. Composites of these three main pillars are known as the human development index (HDI) [1]. HDI consists of three components: health, education, and economic conditions. Health indicators in the HDI are life expectancy. Life expectancy at birth is arithmetic average of ages at death but not affected by age distribution of death [2].

However, to state public health is not enough just by Life Expectancy but many other indicators that influence. For this reason, the health research and development agency, the Indonesian Ministry of Health compiles the public health development index (PHDI). PHDI is a collection of health indicators that can be easily and directly measured to describe health problems. This collection of health indicators is considered to have a direct or indirect role in increasing the long and healthy life expectancy. This PHDI can be used to see the results of the previous year's health development and can also be used to plan health development in the coming year. The PHDI measurement uses data from the public health research and the national socio-economic survey, and the village potential survey which is conducted every five years by surveying from sampling data.

For the first time, the Ministry of Health has developed a model for calculating the PHDI in 2007. This model was compiled in 2010 using data from the Public Health Research in 2007 and 2008. The 2007 PHDI model consists of 24 indicators. Then the 2007 model was developed into the 2013 PHDI Model in

2013 with 30 indicators. The determination of 30 indicators in the 2013 PHDI Model is based on the framework of the concept of health determinants which includes the health of individuals, families, communities and the health service system. The 30 indicators are divided into 7 categories consisting of; i) Toddler health, ii) Reproductive health, iii) Health services, iv) Health behavior, v) Non-communicable diseases, vi) Infectious diseases, and vii) Environmental health.

In fact, every year the provincial and city health offices in Indonesia publish health profile data for each region based on reports from the health sector in each region. The city health profile (CHPs) were developed by The WHO european healthy cities in 1995 to measure and monitor health in a city [3]. Health profile data collected is very large and varied, which is the result of health information transactions that occur dynamically every day at the community health centers and the health program section at the health office at both the city and provincial levels. The main problem faced by city health office is that they could not have any information about the position of city health condition because they don't have any formula to measure the city health index based on the CHPs that they produce every year as an annual report.

This CHPs produced by a city or province contains data and information that illustrates the degree of health, health efforts, and health resources and the achievement of health development indicators in certain areas that can be used as a tool to evaluate the progress of health development from year to year. Health degrees include figures related to death and illness while health efforts include numbers related to health services, access to and quality of health services, community life behavior, and environmental conditions. Health resources include figures relating to facilities and health workers and health financing. These figures become the evaluation material for local governments, especially the health department to find out priority health problems, the handling that must be done, and health development planning.

Therefore, to accommodate the need for development of city health every year, CHPs generated annually is used to measure the city health development index (CHDI) in each region. The indicators and formula for measuring CHDI was adopted from the 2013 PHDI Model and processed by using data mining techniques to obtain precise and highly influential indicators. Among 2013 PHDI Model indicators, not all of them are used in the CHPs. Data mining is used by this study to determine some significant attributes that can be captured to build a new formula of CHDI, which one of the important stages in the knowledge discovery in database (KDD). Data mining is the process of finding and extracting patterns and knowledge from a very large set of data [4], [5].

The stages in KDD consist of 7 stages as the following [5]: i). Data cleaning is the process of cleaning data from noise and inconsistent data; ii). Data integration is the process of combining data from different sources; iii). Data selection is the process of selecting data from a database needed for analysis; iv). Data transformation is the process of transformation to forms suitable for data mining operations; v). Data mining is the process of extracting data patterns with intelligent methods; vi). Pattern evaluation is the process of evaluating data patterns that represent knowledge based on certain measurements; vii). Knowledge presentation is the process of presenting knowledge according to the results of mining

Data mining is a process for extracting data patterns with various techniques including classification, regression, clustering and other analysis techniques. In the health sector, data mining techniques are used to analyze and obtain useful information from patient data so that appropriate treatment can be given [6]. The commonly used model is the predictive model with the classification method [4]. This method is used to predict various types of diseases as well as assist doctors in making the right decisions medically. Techniques that are normally used include: Decision tree, k-nearest neighbor, rule-based, neural network, support vector machine, and naive bayes [4]-[10].

The technique often used in the classification stage is a decision tree with algorithms C4.5, CART, and C5.0 [5], [11]. Then the decision tree method evolved into a random forest method by combining several decision trees based on random selection of data and variables [12], [13]. In order to find significant features from the data obtained, this study had applied random forest (RF) algorithm [12] for feature selection. The principle of Random Forest is to combine a lot of binary decision trees which are constructed using several bootstrap samples derived from learning sample L and randomly selecting at each node part of the explanatory variables, X [14]. Random forest performs very well where the number of attributes is much greater than the samples [15]. Utilizing this property from random forest, it can be easily used for microarray datasets. This can be part of the method for class prediction and feature selection with microarray data [16].

Random forest has been used widely in biomedical domain. Diaz-Uriarte *et al.* [16] have worked on varied data sets and demonstrated that random forest performance is good compared to other classification methods, such as diagonal linear discriminant analysis (DLDA), support vector machines (SVM), and k-nearest neighbor (KNN). Yao *et al.* [17] used random forest algorithm to rank features and the results were evaluated by using SVM classifiers. Yang *et al.* [18], has proposed a method based on random forest classification and the SVM classifier. Random forest important variable interest score is to evaluate features that are selected and features that are eliminated. The new features obtained are then evaluated using SVM

classifier. In many researches, random forest is used as a classifier to evaluate the features. Some of them directly use variable importance level scores to support features [17], [19].

2. RESEARCH METHOD

The CHDI formulation process applied data mining techniques [4], [5] by adopting and modifying the 2013 PHDI Model as seen in Figure 1. Data mining techniques are used to obtain attributes that are very influential in the process of classifying health profile data. The influential attribute as an indicator of health will determine the weight that will be used in calculating the CHDI score.

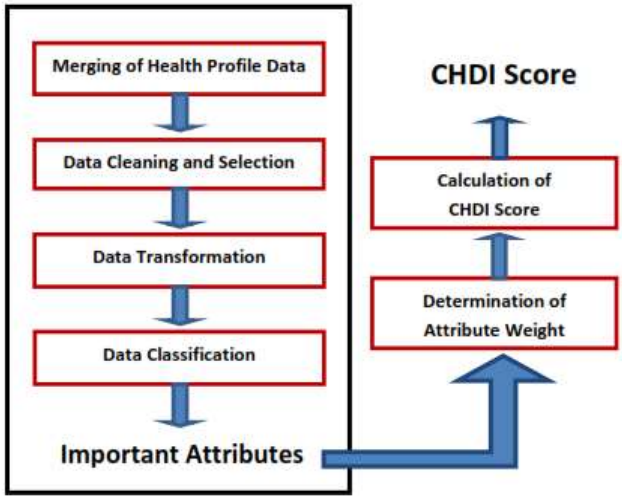

Figure 1. Process of CHDI formulation

2.1. Merging of health profile data

Health profile data used in this study obtained from provincial health offices of Bali, DKI Jakarta, DI Yogyakarta and Sulawesi Selatan in Excel format. The city health profile data from these four provinces, especially in 2017 are then combined, as seen in Table 1. In addition, PHDI score data based on the 2013 PHDI Model for each city or district of the four provinces is also used, shown in Table 2.

This city health profile data in Table 1 consists of approximately 710 attributes that are incorporated into 9 groups of indicators, namely general description, mortality, morbidity, health services, access and quality of health services, community life behavior, environmental conditions, health facilities, and health workers. For each city or district, additional attributes are given which defines the index of Health Development Index category based on the 2013 PHDI score in Table 2 with the following conditions:

- 0,7000 - 1,000 : High
- 0,5000 – 0,6999 : Medium
- 0,0000 - 0, 4999 : Low

A total of 710 attributes will be analyzed to obtain attributes that are very influential and very important in determining the city health development index by using data mining.

2.2. Data cleaning and selection

The data collected at the data collection stage often contains empty values, incorrect formatting, typos, not in accordance with the domain, or repeated [20] data cleaning. These data can affect the classification process in data mining [preprocessing]. For this reason, before data analysis, and data classification, data cleaning is carried out first [9], [21].

Data table from the combination of the four provinces are cleaned from inconsistent data and also from data that has no value. There are several attributes in one provinces table that have data values, but in other provinces tables they are null. There are also attributes which all have zero values. Therefore, this health profile data table is cleared by removing attributes that are empty or inconsistent so that only 57 attributes (denoted as A1…A57) are left out of the 8 indicators. The attributes included in the general description are not used but only the number of residents per city that is used to normalize the value of each attribute.

Table 1. Health profile data table of the Provinces of Bali, DKI Jakarta, DI Yogyakarta, and Sulawesi Tengah in 2017

Year	Province	City / District	Total Population	Number Of Born Alive	Number Of Death Of Neonatal Boys & Girls	Number Of Death Of Baby Boys & Girls	Number Of Death Of Toddler Boys & Girls	Overall Number Of Mother Death
2017	DKI JAKARTA	Jakarta Pusat	921344	13705	56	69	89	13
2017	DKI JAKARTA	Jakarta Utara	1781316	36902	111	164	188	13
2017	DKI JAKARTA	Jakarta Barat	2528065	50607	139	192	235	24
2017	DKI JAKARTA	Jakarta Selatan	2226830	44237	16	25	27	11
2017	DKI JAKARTA	Jakarta Timur	2892783	60955	50	107	111	29
2017	DKI JAKARTA	Kepulauan Seribu	23897	501	4	5	7	1
2017	D.I. YOGYAKARTA	Kulon Progo	445655	5008	34	42	17	3
2017	D.I. YOGYAKARTA	Bantul	927181	12355	71	37	115	9
2017	D.I. YOGYAKARTA	Gunung Kidul	755977	7339	55	71	79	12
2017	D.I. YOGYAKARTA	Sleman	1062861	14025	49	59	61	6
2017	D.I. YOGYAKARTA	Kota Yogyakarta	412692	3621	25	33	37	4
2017	SULAWESI TENGAH	Banggai Kepulauan	116811	2513	33	35	36	3
2017	SULAWESI TENGAH	Banggai	365616	7238	23	33	38	9
2017	SULAWESI TENGAH	Morowali	117330	2614	42	44	44	4
2017	SULAWESI TENGAH	Poso	245993	4832	31	42	46	1
2017	SULAWESI TENGAH	Donggala	299174	6933	46	49	49	13
2017	SULAWESI TENGAH	Toli-Toli	230996	4952	53	70	74	6
2017	SULAWESI TENGAH	Buol	155593	3714	39	50	52	6
2017	SULAWESI TENGAH	Parigi Moutong	474339	10570	69	93	103	17
2017	SULAWESI TENGAH	Tojo Una Una	150820	3391	17	17	17	4
2017	SULAWESI TENGAH	Sigi	234588	4824	36	38	39	10
2017	SULAWESI TENGAH	Banggai Laut	72298	1756	26	27	28	3
2017	SULAWESI TENGAH	Morowali Utara	122985	2742	22	24	24	2
2017	SULAWESI TENGAH	Kota Palu	379782	7147	8	10	12	11
2017	BALI	Jembrana	274900	4605	32	48	3	5
2017	BALI	Tabanan	441000	5139	31	43	10	3
2017	BALI	Badung	643500	8693	16	26	3	5
2017	BALI	Gianyar	503900	5979	33	60	14	3
2017	BALI	Klungkung	177400	2819	6	19	5	2
2017	BALI	Bangli	225100	3274	12	23	6	4
2017	BALI	Karangasem	412800	6903	39	48	4	6
2017	BALI	Buleleng	653600	10819	30	39	4	9
2017	BALI	Kota Denpasar	914300	17333	10	11	4	8

Table 2. PHDI scores based on the 2013 PHDI Model

NO	PROVINCE	CITY/DISTRICT	PHDI SCORE
1	DKI JAKARTA	Jakarta Pusat	0,5959
2	DKI JAKARTA	Jakarta Utara	0,5994
3	DKI JAKARTA	Jakarta Barat	0,6356
4	DKI JAKARTA	Jakarta Selatan	0,6146
5	DKI JAKARTA	Jakarta Timur	0,5887
6	DKI JAKARTA	Kep. Seribu	0,5711
7	D.I. YOGYAKARTA	Kulon Progo	0,5664
8	D.I. YOGYAKARTA	Bantul	0,5772
9	D.I. YOGYAKARTA	Gunung Kidul	0,5569
10	D.I. YOGYAKARTA	Sleman	0,5805
11	D.I. YOGYAKARTA	Kota Yogyakarta	0,5578
12	SULAWESI TENGAH	Banggai Kepulauan	0,4408
13	SULAWESI TENGAH	Banggai	0,5066
14	SULAWESI TENGAH	Morowali	0,5216
15	SULAWESI TENGAH	Poso	0,5317
16	SULAWESI TENGAH	Donggala	0,4644
17	SULAWESI TENGAH	Toli-Toli	0,4255
18	SULAWESI TENGAH	Buol	0,5336
19	SULAWESI TENGAH	Parigi Moutong	0,4359
20	SULAWESI TENGAH	Tojo Una Una	0,3862
21	SULAWESI TENGAH	Sigi	0,4936
22	SULAWESI TENGAH	Banggai Laut	0
23	SULAWESI TENGAH	Morowali Utara	0
24	SULAWESI TENGAH	Kota Palu	0,6091
25	BALI	Jembrana	0,6081
26	BALI	Tabanan	0,6826
27	BALI	Badung	0,6546
28	BALI	Gianyar	0,7352
29	BALI	Klungkung	0,6203
30	BALI	Bangli	0,5776
31	BALI	Karangasem	0,5823
32	BALI	Buleleng	0,6191
33	BALI	Kota Denpasar	0,6992

2.3. Data transformation

Data transformation is the stage of converting the previous data format to the new data format. One of the operations of data transformation is normalization [22], [23]. The normalization process is needed to make the classification process effective. Therefore, data values of health profile are normalized by dividing each attribute value by the population in the area and multiplying by 10 so that the value is neither too large nor too small. Furthermore, the data is divided into two parts, namely training data and test data with a composition of 70:30, 75:25, and 80:20.

2.4. Data classification

Algorithms of data classification are decision trees, rule-based methods, probabilistic methods, SVM methods, instance-based methods, and neural networks [24], [25]. Data classification of these health data is carried out by using decision tree C5.0 and random forest techniques. Each attribute will be classified into certain groups so that the most dominant attribute involvement can be identified to get the weight that will determine in the calculation of CHDI score. The calculation is done using R programming.

3. RESULTS AND ANALYSIS

The results of the classification using the random forest technique are displayed in graphical form as shown in Figure 2. The A43 attribute gets the highest value on the variable importance, which means that the A43 attribute is very influential in determining the classification and measurement of accuracy of the random forest method in the comparison of train data and test data 75:25. Then it's followed by attributes A41, A4 and others. Figure 2 illustrates the grouping of 57 attributes so that the weight value for each attribute can be determined with the following conditions:

- $X < 0,1$: weight 1
- $0,1 \leq X < 0,2$: weight 2
- $0,2 \leq X < 0,3$: weight 3
- $0,3 \leq X < 0,4$: weight 4
- $X \geq 0,4$: weight 5

The 57 indicators are weighted according to the predetermined categories as shown in the example of Table 2. Attribute or indicator of the number of live births having a random forest value of 0.4553, which means that it has a weight of 5. This weight will be used as one component of the CHDI formula.

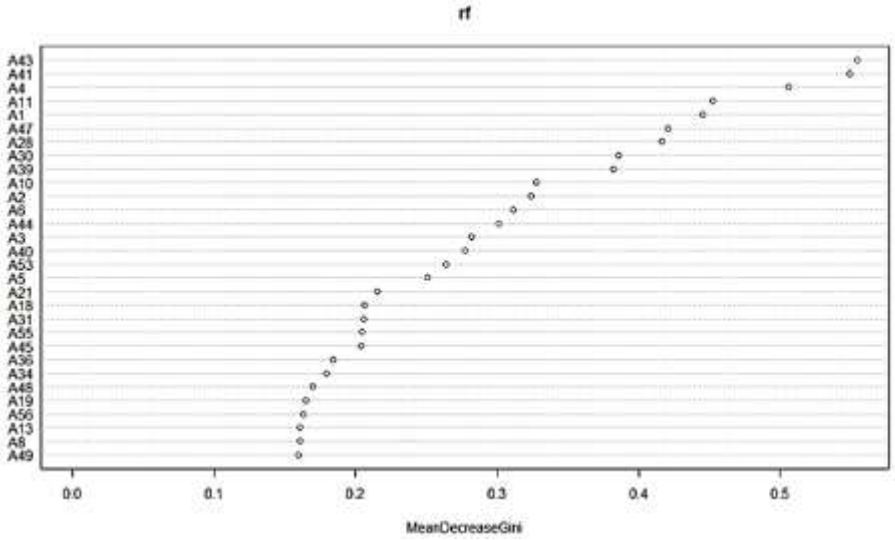

Figure 2. Graphs of classification results with random forest

The formulation of CHDI refers to the calculation of the 2013 PHDI Model as follows [1]:

1) Calculate the weight proportion of each indicator in one group :

$$Proportion\ of\ indicator\ weight = \frac{indicator\ weights}{the\ total\ weight\ of\ the\ indicator\ group} \quad (1)$$

2) Calculate the index of each group of indicators by adding up the results of the multiplication of the indicator index values by the proportion of weights in one group :

$$Indicator\ group\ index = (Indicator\ Index\ Value\ (1) * Weight\ Proportion\ (1)) + (Indicator\ Index\ Value\ (2) * Weight\ proportion\ (2)) + \ldots + (Indicator\ Index\ Value\ (n) * Weight\ proportion\ (n) \quad (2)$$

3) After obtaining six index values for the indicator group, then proceed with calculating CHDI:

$$CHDI = \frac{Indicator\ group\ index(1) + \cdots + Indicator\ group\ index\ (n)}{the\ number\ of\ indicator\ groups/n} \quad (3)$$

A simulation of the CHDI calculation with 5 indicators illustrated in Table 3. The results of the CHDI measurement for provinces of DKI Jakarta, DI. Yogyakarta, Sulawesi Tengah, and Bali are shown in Table 4. The CHDI score was successfully calculated but the comparison between CHDI and PHDI is still quite large.

Table 3. Simulation of CHDI measurement

No (a)	Variabels	Indicators (b)	Values (c)	Weight (d)	Proportion of Indicator Weight (e)	(c)*(e) (f)	Indicator Group Index (g)	CHDI Score (h)
		1. Death Rate					0,05386	0,74878
1	A1	Number of Live Births	0,209650	5	0,25000	0,052412		
2	A2	Number of Neonatal deaths	0,001674	4	0,20000	0,000335		
3	A3	Number of Dead Babies	0,002092	3	0,15000	0,000314		
4	A4	Number of Dead Toddler	0,002929	5	0,25000	0,000732		
5	A5	Number of maternal deaths	0,000418	3	0,15000	0,000063		

Table 4. Simulation for Provinces of DKI Jakarta, DI. Yogyakarta, Sulawesi Tengah, and Bali

NO	PROVINCE	CITY/DISTRICT	PHDI SCORE	CHDI SCORE
1	DKI JAKARTA	Jakarta Pusat	0,5959	0,49039
2	DKI JAKARTA	Jakarta Utara	0,5994	0,51373
3	DKI JAKARTA	Jakarta Barat	0,6356	0,17723
4	DKI JAKARTA	Jakarta Selatan	0,6146	0,40701
5	DKI JAKARTA	Jakarta Timur	0,5887	0,33689
6	DKI JAKARTA	Kep. Seribu	0,5711	0,49039
7	D.I. YOGYAKARTA	Kulon Progo	0,5664	0,49039
8	D.I. YOGYAKARTA	Bantul	0,5772	0,51373
9	D.I. YOGYAKARTA	Gunung Kidul	0,5569	0,17723
10	D.I. YOGYAKARTA	Sleman	0,5805	0,40701
11	D.I. YOGYAKARTA	Kota Yogyakarta	0,5578	0,33689
12	SULAWESI TENGAH	Banggai Kepulauan	0,4408	0,27347
13	SULAWESI TENGAH	Banggai	0,5066	0,28354
14	SULAWESI TENGAH	Morowali	0,5216	0,25877
15	SULAWESI TENGAH	Poso	0,5317	0,31878
16	SULAWESI TENGAH	Donggala	0,4644	0,25604
17	SULAWESI TENGAH	Toli-Toli	0,4255	0,34186
18	SULAWESI TENGAH	Buol	0,5336	0,22629
19	SULAWESI TENGAH	Parigi Moutong	0,4359	0,23566
20	SULAWESI TENGAH	Tojo Una Una	0,3862	0,25579
21	SULAWESI TENGAH	Sigi	0,4936	0,12273
22	SULAWESI TENGAH	Banggai Laut	0	0,19379
23	SULAWESI TENGAH	Morowali Utara	0	0,25489
24	SULAWESI TENGAH	Kota Palu	0,6091	0,47766
25	BALI	Jembrana	0,6081	0,47619
26	BALI	Tabanan	0,6826	0,40554
27	BALI	Badung	0,6546	0,19052
28	BALI	Gianyar	0,7352	0,41922
29	BALI	Klungkung	0,6203	0,49998
30	BALI	Bangli	0,5776	0,44201
31	BALI	Karang Asem	0,5823	0,38532
32	BALI	Buleleng	0,6191	0,40596
33	BALI	Kota Denpasar	0,6992	0,36725

According to Table 4, the cities with the biggest difference of PHDI and CHDI scores occur in the city of Badung and Jakarta Barat. This means that the condition of these cities during 2013 to 2017 had changed significantly. The government should pay more attentions to these cities. In other side, the cities with the smallest difference of PHDI and CHDI scores are Bantul, Kulon Progo, and Toli-Toli. This means that there is no significant progress in the development of city health growth in these cities.

The city with the lowest PHDI score is the city of Tojo Una Una with PHDI score of 0.3862, and its CHDI score is 0.25579. The city with the lowest CHDI score is the city of Sigi with CHDI score of 1.2273, and its PHDI score is 0.4936. These two cities are in the same province, Sulawesi Tengah. The city with the highest PHDI is the city of Gianyar with PHDI score of 0.7352, and its CHDI score is 0.41922. The city with the highest CHDI score is Jakarta Utara and Bantul with their score of 0.51373, where their PHDI scores are 0.5994 and 0.5772, respectively. These figures confirm that our proposed CHDI model, which can be produced once a year, can be used as the index of the city health condition beside PHDI, which is built once within five years. This CHDI index would be very beneficial for the city health office to be used as the baseline for their annual planning programs in developing the city health.

4. CONCLUSION

Every year, the city health office should create a city health program planning for developing the city health. At present, they are still depending upon the PHDI index that are delivered once within five years as their baseline figures. This study assists the city health office by creating a new model, city health development index (CHDI) that can produce a city health index once in a year. The CHDI formulation was successfully formulated by adopting the 2013 PHDI Model and using indicators from city health profile produced annually by each city. The CHDI indicators are obtained by identifying some significant influential indicators as the attributes used to build the index. This identification process or feature selection is carried out by using data mining technique, i.e., random forest algorithm. The CHDI formula has been examined and analysed by comparing the scores with that of PHDI scores. The result shows that the different between PHDI and CHDI is not significantly big, in terms of the scores and the cities that are measured. This means that CHDI formula can be used by city health office in creating annual program planning of their city health.

For future research, the CHDI model can be adopted to measure some specific aspects of city health condition, such as the reproductive health, city health services, public health behavior, infectious diseases, or environmental.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support from Directorate for Research and Community Services, Ministry of Research, Technology, and Higher Education, Republic of Indonesia.

REFERENCES

[1] S. Fukuda-Parr, "Indikator of human development and human rights-overlap, differences and what about the human development index ?," *Statistical Journal of the United Nations ECE 18*, 2001, 239-248, IOS Press, 2001.
[2] S. Ghislandi, W. C. Sanderson and S. Scherbova, "A simple Measure of Human Development: The Human Life Indicator," *Population and Development Review*, vol. 45, no. 1, pp. 219-233, 2018, doi: 10.1111%2Fpadr.12205.
[3] P. Webster and A. Lipp, "The Evolution of The WHO City Health Profiles: A Content Review," *Health Promotion International*, vol. 24, no. 1, pp. 56-63, 2009, doi: 10.1093/heapro/dap055.
[4] N. Jothi and W Husain., "Data Mining in Healthcare-A Review," *Procedia Computer Science*, vol. 72, pp. 306-313, 2015, doi: 10.1016/j.procs.2015.12.145.
[5] J. Han, Micheline Kamber and Jian Pei., "Data Mining; Concept and Techniques," The Morgan Kaufmann Series in Data Management Systems, USA, 2012.
[6] V. Manikandan, S. Latha, "Predicting the analysis of heart disease symptoms using medical data mining methods," *International Journal of Advanced Computer Theory and Engineering*, vol. 2, no. 1, pp. 46-51, 2013.
[7] Kevin Daimi and S. Banitaan., "Using Data Mining to Predict Possible Future Depression Cases," *International Journal of Public Health Science*, vol. 3, no. 4, pp. 231-240, 2014.
[8] L. Muflikhah, Nurul Hidayat, and Dimas Joko Hariyanto., "Prediction of hypertention drug therapy response using K-NN imputation and SVM algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 1, pp. 460-467, 2019
[9] S. Adnan Diwan Alalwan, "Diabetic analytics: proposed conceptual data mining approaches in type 2 diabetes dataset," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 1, pp. 88-95, 2019
[10] D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241-266, 2013, doi: 10.14257/ijbsbt.2013.5.5.25.
[11] H. Sharma and S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," *International Journal of Science and Research (IJSR)*, 2013.
[12] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
[13] K. Fawagreh, M. M. Gaber, E. Elyan., "Random forests: from early developments to recent advancements," *Systems Science and Control Engineering*: *An Open Access Journal*, vol. 2, pp. 602-609, 2014, doi: 10.1080/21642583.2014.956265.
[14] R. Genuer, V. Michel, E. Eger and B. Thirion., "Random forests-based feature selection for decoding fMRI data," *Proceedings Compstat*, vol. 267, pp. 1-8, 2010.
[15] G. Biau, and E. Scornet, "A random forest guided tour," *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research, Springer; Sociedad de Estadística e Investigación Operativa*, vol. 25, no. 2, pp. 197-227, June, 2016.
[16] R. Díaz Uriarte and S. A. Andres. "Geneselection and classification of microarray data using random forest," *BMC Bioinformatics*, 2006.
[17] D. Yao, J. Yang, X. Zhan, X. Zhan and Z. Xie, "A novel random forests-based feature selection method for microarray expression data analysis," *International Journal of Data Mining and Bioinformatics*, vol. 13, no. 1, pp. 84-101, 2015, doi: 10.1504/IJDMB.2015.070852.
[18] J. Yang, D. Yao, X. Zhan, and X. Zhan., "Predicting disease risks using feature selection based on random forest and support vector machine," *Proceedings of the 10th International Symposium on Bioinformatics Research and Applications*, pp. 1-11, 2014, doi: 10.1007/978-3-319-08171-7_1.
[19] R. Genuer, J. M. Poggi, C. Tuleau-Malot., "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225-2236, 2010, doi: 10.1016/j.patrec.2010.03.014.
[20] I. F., Ilyas and X. Chu, *Data Cleaning*, ACM Books Series, electronic, ACM, 2019
[21] A. S. Nayak, A. P. Kanive, N. Chandavekar, and R. Balasubramani, "Survey on Pre-Processing Techniques for Text Mining," *International Journal of Engineering and Computer Science*, vol. 5, no. 6, pp. 16875-16879, 2016, doi: 10.18535/ijecs/v5i6.25.
[22] K. Swati and S. Kumar, "A Comparative Study of Various Data Transformation Techniques in Data Mining," *International Journal of Scientific Engineering and Technology*, Vol.4, No. 3, pp. 146-148, 2015.
[23] L. Canchen, "Preprocessing Methods and Pipelines of Data Mining: An Overview," *arXiv preprint arXiv:1906.08510*, 2019.
[24] C. C. Aggarwal, *Data Classification : Algorithms and Application*, CRC Press, 2015
[25] T. C. Sharma and M. Jain, "WEKA Approach for Comparative Study of Classification Algorithm," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 4, pp. 1931-199, 2013.