Classification and Prediction of Smoking Behavior and Hypertension in the Healthy Family Program with R (Case Study : Bali Provincial Government Department of Health)

I Gede Abi Yodita Utama[1*], Nyoman Pramaita[2], dan Made Sudarma[3]

[1,2,3] Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
*Email: kingkonggede89@gmail.com

Abstract The Healthy Indonesia Program with the Family Approach aims to improve the quality of human life and is a government program that starts from the family environment. There are 12 indicators marking the health status of a family. The Provincial Government of Bali in implementing the 12 indicators there are two main problems of non-communicable diseases, namely patients with hypertension and smoking behavior. An analysis in the form of classification and prediction is needed to overcome these problems. Classification and prediction is one of the techniques in data mining. Through R, a decision tree can be produced which can be used to help the classification process and produce predictions related to the problem of non-communicable diseases. The decision tree results can be predicted that the dominant hypertension are in the age group classification of 20-40 years and 41-70 years with a probability of 0.14. As for smoking behavior, the highest smoking tendency was obtained in the classification of male sex whose status worked with a probability of 0.44.

Index Terms— Classification, Data mining, Decision tree, Healthy Family Program, Prediction, R.

I. INTRODUCTION

The Healthy Family Program is a government program in the Health Sector with the aim of improving the quality of human life starting from the family environment through health efforts, community empowerment and supported by financial protection and equitable health services. To support this achievement, there are 3 main pillars that must be strengthened, namely the application of a healthy paradigm through promotion and prevention, strengthening health services through improving access and quality of health services and implementing national health insurance by expanding targets and benefits. In the framework of implementing the Healthy Family Program 12 key indicators were agreed upon as markers of a family's health status [1].

The Provincial Government of Bali in implementing the indicators of healthy families there are two main problems faced in handling indicators related to non-communicable diseases, namely hypertension sufferers and smoking behavior. Both problems of non-communicable diseases can trigger the emergence of other disease problems. Hypertension often shows no symptoms, but is at risk of causing heart failure and stroke [2]. Likewise, smoking behavior can cause other lung diseases such as inflammation, bronchitis and pneumonia [3].

The results of the Bali Province Basic Health Research also showed an increase in the two non-communicable diseases. Riskesdas data in 2018 showed an increase in hypertension by 1.07% from the results of 2013 which was 22.4%. Likewise, smoking behavior increased in 2018 by 0.77% from the results of 2013 which was 8.8%. The results also illustrate that the two non-communicable diseases are generally influenced by several factors such as age and employment status.

To overcome these problems analysis is needed through a classification and prediction process. Classification and prediction is one of the techniques in data mining [4]. Based on previous research techniques of classification and prediction of data mining has been done, with classification can run prediction future probabilities[5]. Many software that can be used to implement data mining, one of which is R. R software can help the stages of the analysis and statistical process when conducting data mining techniques classification and prediction Classification will be done based on characteristics such as age, employment status, marital status and gender. The classification process on R will produce a decision tree that can be used to predict hypertension sufferers and smoking behavior based on these characteristics. In addition, it can also provide recommendations to leaders in making policies related to prevention and control of non-communicable diseases.

II. PURPOSE OF PAPER

The purpose of this research is :

1. The results of the implementation of R in the form of a decision tree that helps in the classification of patients with hypertension and smoking behavior based on characteristics
2. Classification based on these characteristics in the form of analysis that will be used to predict patients

with hypertension and smoking behavior so that they can provide recommendations to leaders in making policies.

III. LITERATURE REVIEW

A. Data Mining

Data mining is a process for analyzing and changing information in such data into useful information or knowledge [6]. Technically, data mining involves the process of finding regularities, patterns or relationships from related databases in large amount of data [7].

In data mining there are techniques that can be used to solve structured and unstructured data problems. Techniques that can be used by data mining such as association, classification, grouping, decision tree, sequence of patterns, text mining and so on.

B. R

The R programming language was developed in the mid 1990s by Robert Gentleman Ross Lhake. R and has been widely used to carry out data mining implementations, develop statistical software and data analysis. Ease of use and extensibility R are the advantages of this programming language. In addition to data mining techniques that can be performed in R such as clustering, classification, data cleansing and prediction, R is also used to perform statistical and graphical techniques both linear and nonlinear, modeling, classical statistical processes, time series analysis and others [8].

The form of the R programming language is the command line. The user will enter the command in a command prompt and each command will be run at a time. R can combine all data manipulation processes and statistical models that are often needed by researchers to implement the results of their research. Researchers can easily build a prediction model without spending money. With R the researcher can be assisted in the process of creating designs and collecting data, showing how to do data analysis properly and illustrate the interpretation of results.

C. Classification

The classification process involves the attributes used to identify the class of certain items. Classification provides information into a category or class to predict what will happen in that class. Classification and grouping are almost defined equally. Grouping creates groups with the same content or users, while classification classifies groups from user profiles. Classification can also describe the characteristics of an object and pattern [9]. The classification process involves data for learning and classification [10]. Training data is used as reference data during the classification process [11] [12]

D. Prediction

Predictions generally carry out an analysis based on knowledge and experience. This process focuses on one aspect of data that is related to several other aspects of the data with variables called predictor variables. Prediction is used to predict some unknown results based on previous experience and history [13]. Predictions resemble the process of estimation and classification, the difference is the results of predictions indicate events that have not yet occurred [14].

E. Decision Tree

Decision tree is a way to represent rules in a hierarchical, coherent structure, where each object will produce a conclusion and make a decision. Rules refer to the logical structure presented in the form of "IF-THEN" [15].

Decision tree starts with a simple question that has two or more answers. Each answer leads to further questions that are used to classify or identify data that can be categorized, so that from that classification can be produced a prediction based on each answer.

The main components of the decision tree are nodes and branches. Nodes consist of 3 types, namely a) the root node is also called a decision node which represents a choice resulting in division into two or more subsets, b) Internal node is also called an opportunity node is one of the options that may be available in a tree structure connected to the parent node and the child node, c) Leaf nodes are also called the final node representing the final result of a combination of decisions or events. Branches represent a result of events originating from the root node and internal node. Each path from the root node through the internal node to the leaf node represents the decision rule classification process [16].

Decision trees are formed using branch hierarchies. With the branch hierarchy to make it easier for humans to see the relationship of a factor that affects a problem [4]. The most important step in building a decision tree is separation, stopping and trimming.

F. Healthy Family Program

The Healthy Indonesia Program with the Family Approach (in Indonesian abbreviated as *PISPK*) is a government program that supports the 5th nawacita agenda, namely improving the quality of human life. The aim of this program is to improve the degree of public health, the nutritional status of the community through health efforts and community empowerment supported by financial protection and equitable health services.

The program is implemented through a family approach with 12 main indicators marking the status of a family said to be healthy. The 12 indicators include 1) Families participating in the Family Planning (KB) program, 2) Mothers giving birth in a health facility, 3) Babies receive complete basic immunizations, 4) Babies receive exclusive breast milk, 5) Toddlers get monitoring growth, 6) Patients with pulmonary tuberculosis get treatment according to the standard, 7) Patients with hypertension do regular treatment, 8) People with mental disorders get treatment and are not neglected, 9) No smoking family members, 10) The family has become a member of the National Health Insurance (JKN), 11) Families have access to clean water facilities, 12) Families have access or use a toilet [1].

IV. RESEARCH METHOD

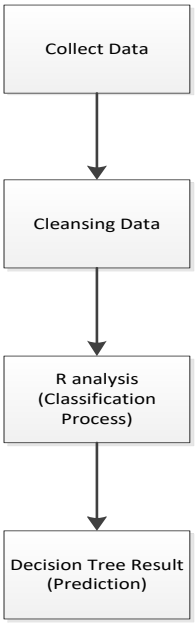

Figure 1 Research Method

1. Collect Data, Data collection is done by taking the results of healthy family applications in the form of raw data in 2018
2. Cleansing Data, Erase and correct irrelevant data from healthy family application raw data, cleansing data that will be used as input when conducting R analyzes
3. R analysis, Conducting an analysis on R, the classification is done based on the attributes / rules of characteristics in determining hypertension sufferers and smoking behavior as seen from the number of yes and no. The classification process uses training data.
4. *Decision Tree* Result, the results of the R analysis stage in the form of a decision tree used to predict patients with hypertension and smoking behavior, in addition to these results can be recommendations in policy making

V. ANALYSIS AND RESULTS

A. Collect Data

The data used in this research are primary data taken directly from healthy family applications in the form of raw data in the period 2018 taken randomly in 9 districts /cities. The sample data used were 15,601 which became training data in classifying hypertension and smoking sufferers by 85% for smoking and 94% for hypertension and the rest will be used as testing data in making predictions in implementation on R. First the data will be cleansed data before becoming training or testing data in R. The data consists of several attributes that describe a person's characteristics such as age, gender, marital status and employment. Table 1 and Table 2 are examples of training data and testing data that will be used in R implementation before data cleansing is performed

Table 1 Examples Raw Data for Healthy Family Applications (Sample Data) Smoking Behavior

No	Age	Gender	Marital Status	Proffesion	Smoking Status
1	52	Male	Married	Member of DPRD Kab./Kota	Y
2	55	Male	Married	Member of DPRD Prop.	Y
3	31	Female	Married	Pharmacist	Y
4	33	Male	Married	Pharmacist	T
5	24	Female	Married	Pharmacist	Y
6	0	Female	Single	Not yer/ Not Working	T
7	5	Male	Single	Not yer/ Not Working	T
8	9	Female	Single	Not yer/ Not Working	T
..	..	……	……	…….	…
15.601	52	Male	Married	Entrepreneur	T

Tabel 2 Examples of Raw Data for Healthy Family Applications (Sample Data) for Patients with Hypertension

No	Age	Marital Status	Proffesion	Diagnosis of Hypertension
1	52	Married	Member of DPRD Kab./Kota	Y
2	55	Married	Member of DPRD Prop.	Y
3	31	Married	Pharmacist	Y
4	33	Married	Pharmacist	Y
5	24	Married	Pharmacist	Y
6	0	Single	Not yer/ Not Working	T
7	5	Single	Not yer/ Not Working	T
8	9	Single	Not yer/ Not Working	T
..	..	……	…….	….
15.601	52	Married	Entrepreneur	Y

From the sample data above, we can make an attribute that will be used when classifying. The classification attribute refers to the results of the Basic Health Research in 2013 and 2018. Table 3 displays the attributes in carrying out the classification process.

Table 3 Attributes used for classification

Attribute	Classification of Patients with Hypertension	Smoking Behavior Classification
Age (Based on Bali Province Basic Health Research)	- Less than or equal to 10 years (<= 10 years) - 11 – 40 years - 40 – 70 years	- Less than 20 years (< 20 years) - 20 – 40 years - 41-70 years
Marital Status	- Married/Divorced/Death Divorce - Single	- Married/Divorced/Death Divorce - Single
Proffesion	- Work - Not yet/Not Working - Student	- Bekerja - Not yet/Not Working - Student
Gender	No classification based on	- Male

sex	- Female

B. Cleansing Data

Data cleansing is carried out to produce relevant and quality data by taking into account aspects of validity, completeness, uniformity and consistency. Cleansing process data is sample data from raw data of healthy family applications. Cleansing results obtained the amount of each data of 14,761 both for patients with hypertension and smoking behavior from the sample data which amounted to 15,601. The cleansing data will be used as training data in classifying 85% for smoking and 94% for hypertension. The rest will be used as data testing in making predictions. Table 4 and Table 5 show the procedures for data cleansing which are seen from the aspect of data quality.

Table 4 Procedures of Cleansing Data Hypertension

Aspect	Procedures	Cleansing Data (Deleted)	Cleansing Data (Modified)
Completeness	- The age attribute if month and year is found 0 then the data is deleted	7 5	
	- Attribute age 0-70 years If the hypertension status is found N	14	
Uniformity	then the data is deleted	751	
	The age attribute if found more than 70 years of age then the data is deleted		
validity	The age attribute if babies, toddlers and children from 0-10 years old are found to hypertension Y will be changed to T		2381
Consistency	Data on age for people with hypertension are changed by number 1 if the age is less than or equal to 10 years (<= 10 years), modified by number 2 if age 11 - 40 years and modified by number 3 if age 40 - 70 years including employment status data if the employment status of doctors, nurses, entrepreneurs, etc. then is changed to work, if the work status takes care of the household then it is changed to not yet /does not work		12380
TOTAL		**840**	**14761**

Table 5 Procedures for Cleansing Data Smoking Behavior

Aspect	Procedures	Cleansing Data (Deleted)	Cleansing Data (Modified)
Completeness	- The age attribute if month and year is found 0 then the data is deleted	7 5	
	- Attribute age 0-70 years If smoking status is found N then the data is	14	
Uniformity	deleted	751	
	The age attribute if found more than 70 years of age then the data is deleted		
validity	The age attribute if babies, toddlers and children from 0-10 years old is found smoking status Y will be changed to T		2381
Consistency	The age data for smoking behavior is changed by number 1 if the age is less than 20 years (<20 years), modified by number 2 if age 20 - 40 years and modified by number 3 if age 41 - 70 years including employment status data if the employment status of doctors, nurses, entrepreneurs, etc then is changed to work, if the work status takes care of the household then it is changed to not yet /does not work		12380
TOTAL		**840**	**14761**

C. R Analysis

Cleansing data obtained 14,761 sample data which will be input data in R. For the classification of hypertension sufferers and smoking behavior, each training data used was 85% and 94%. In R the data cannot be used 100% for classification because the rest of the use of the data will be used as testing data to make predictions in the decision tree.

Figure 2 in the decision tree result illustrates that from the training data of 85% (12,547), hypertension are classified in the age group of 20-40 years and 41-70 years with a total number of 8,657 patients (69%), whereas for the age classification below 20 years less likely to suffer from hypertension. From the age group of 20-40 years and 41-70 years, it can be classified again that hypertension are generally classified in the status of work employed and not / not working with the number of sufferers of 5,454 (63%).

Figure 3 decision tree results illustrate that from the training data used as much as 94% (13,875), smoking behavior is classified on the status of work work with the number of smokers as much as 6,938 (50%), while for classification based on the sex of men or women whose status is working the number of male smokers working

was 2081 (30%) and women working as many as 1,388 (20%). In addition, smokers are also classified based on the sex of working women, marital / divorce status (living or dead) and unmarried, age group 11-40 years. Whereas smokers for male sex whose working status can be classified by age group 41-70 years, less than 20 years and ages between 20-40 years.

D. Decision Tree Result

The results of the R are in the form of a decision tree which in addition to being used to help in classification can also be used to produce predictions for both hypertension sufferers and smoking behavior in Bali. In making predictions on the decision tree testing data used by 15% for patients with hypertension and 6% for smoking behavior. Figure 2 and figure 3 show the results of R in the form of a decision tree

From figure 2 it can be seen that hypertension in Bali are in the age group of 20-40 years and ages 41-70 years with a probability of hypertension 0.14. When seen from the status of work with the age group 20-40 years and age 41-70 years, hypertension are generally most of the status of working and not yet/ not working with a probability of 0.07 while the status of students is not diagnosed with hypertension but there is a possibility with a probability of 0.04.

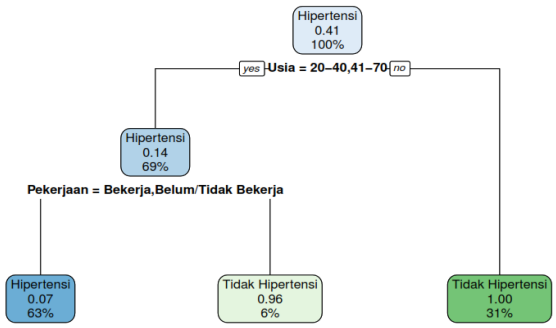

Figure 2 *Decision Tree for Patients with Hypertension*

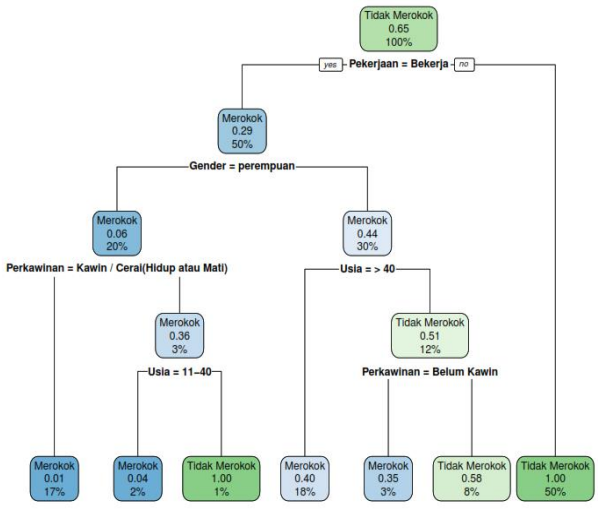

Figure 3 Decision Tree Smoking Behavior

Figure 3 can be seen that smoking behavior in Bali is generally in the status of work that works with a smoking probability of 0.29. When viewed from the sex of men or women who work generally most of the smoking with a probability of female sex working status of 0.06 and male sex working status of 0.44. In the sex of women with working status and marital / divorce status (living or dead) there is a tendency to smoke by 0.01 and the sex of women whose status is working, unmarried and ages between 11-40 years have a smoking tendency of 0.04. As for the sex of men whose status of work tends to smoke with a probability of 0.44. If viewed based on male sex working status and age group between 41-70 years there is a tendency to smoke by 0.40 and age groups 20-40 years whose status is not yet married as well as smoking tendencies by 0.35.

VI. CONCLUSION

Based on the implementation of data mining with R generated a decision tree that can be used to carry out the classification process and produce predictions of patients with hypertension and smoking behavior in a healthy family program. From the decision tree it is seen that patient with hypertension are classified in the age group of 20-40 years and 41-70 years and their working status. From these results it can also be predicted that hypertension are predominantly in the age group of 20-40 years and 41-70 years with a probability of 0.14. Whereas smoking behavior is classified based on the sex of men or women whose status is working, married or not married and the age group is 11-40 years or more or equal to 40 years. Prediction results for smoking behavior obtained the highest smoking tendency in the sex of men whose status works with a probability of 0.44.

VII. SUGESTION

To compare classification results and get more accurate prediction results, R can be used in analyzing the results in the form of the application of classification and other prediction methods such as K-Nearest Neighbors (KNN), Naive Bayes and Support Vector Machine (SVM).

REFERENCES

[1] Ministry Of Health Republic Of Indonesia. (2017,June 17). Program Indonesia Sehat dengan Pendekatan Keluarga - PISPK [Online]. Available at: http://pispk.kemkes.go.id/id/program-pispk/latar-belakang/ [Accessed:17-May-2020].

[2] Ministry Of Health Republic Of Indonesia. (2012, May 6). Masalah Hipertensi di Indonesia [Online]. Available at: https://www.kemkes.go.id/article/view/1909/masalah-hipertensi-di-indonesia.html [Accessed:17-May-2020].

[3] Ministry Of Health Republic Of Indonesia. (2015, Nov 25). Inilah 4 Bahaya Merokok Bagi Kesehatan Tubuh [Online]. Available at: https://www.kemkes.go.id/development/site/dinas-kesehatan/index.php?cid=1-15112500015&id=inilah-4-bahaya merokok-bagi-kesehatan-tubuh.html [Accessed:17-May-2020].

[4] Kastawan P. W, Wiharta Dewa M, M. Sudarma, "Implementasi Algoritma C5.0 pada Penilaian Kinerja Pegawai Negeri Sipil," Majalah Ilmiah Teknologi Elektro, [S.l.], v. 17, n. 3, p. 371-376, dec. 2018, ISSN 2503-2372, 2018.

[5] W. P Dewa Ayu, P. B Kadek Ary, M. Sudarma. "Prediction of Days in Hospital Dengue Fever Patients using K-Nearest Neighbor," International Journal of Engineering and Emerging Technology, [S.l.], v. 3, n. 1, p. 23-25, july 2018. ISSN 2579-597X.

[6] A.U. Begum. (2019, August). Data Mining Techniques For Big Data. International Journal of Advanced Research in Science, Engineering and Technology" [Online]. Vol.6..Available.:.https://www.researchgate.net/publication/336197482_Data_Mining_Techniques_For_Big_Data_Vol_6_Special_Issue [Accessed:17-May-2020].

[7] S. Putri, B. Adi, M. Sudarma, "The Optimization of Feature Selection Using Genetic Algorithm with Naïve Bayes

Classification for Home Improvement Recipients," International Journal of Engineering and Emerging Technology, [S.l.], v. 3, n. 1, p. 66-70, july 2018, ISSN 2579-597X, 2018.

[8] T. U. Sawant. (2016, May). R: Data Mining Tool And Its Applications. International Journal of Advanced Computer Technology & Management (IJACTM) [Online].. Available at :.https://www.researchgate.net/publication/338853853_R_Data_Mining_Tool_And_Its_Applications [Accessed: 17-May-2020].

[9] Andisana I Putu Gd. S, M. Sudarma, W. I Made Oka, "Pengenalan dan Klasifikasi Citra Tekstil Tradisional Berbasis Web Menggunakan Deteksi Tepi Canny, Local Color Histogram dan Co-Occurrence Matrix," Majalah Ilmiah Teknologi Elektro, [S.l.], v. 17, n. 3, p. 401-408, dec. 2018, ISSN 2503-2372, 2018.

[10] A. S Mahendra I G. N, Leo Mahadya Suta I. B, M. Sudarma, "Classification of Data Mining with Adaboost Method in Determining Credit Providing for Customers," International Journal of Engineering and Emerging Technology, [S.l.], v. 4, n. 1, p. 31--36, oct. 2019, ISSN 2579-597X, 2019.

[11] M. Sudarma, Harsemadi I Gede, "Design and Analysis System of KNN and ID3 Algorithm for Music Classification based on Mood Feature Extraction," International Journal of Electrical and Computer Engineering, Vol. 7, Iss. 1, p. 486-495, (Feb 2017), 2017

[12] Harsemadi Gede, M. Sudarma, N. Pramaita, "Implementasi Algoritma K-Nearest Neighbor pada Perangkat Lunak Pengelompokan Musik untuk Menentukan Suasana Hati," Majalah Ilmiah Teknologi Elektro, [S.l.], v. 16, n. 1, p. 14-20, july 2016, ISSN 2503-2372, 2016

[13] Madni A.H, Anwar Zahid, Shah Ali M, "Data Mining Techniques and Applications – A Decade Review," COMSATS Institute of Information Technology, Pakistan, 2017.

[14] A. I Made Dwi; A. A Made Pasek, M. Sudarma, "Data Mining, Evaluation, K-means Evaluation of Supporting Work Quality Using K-Means Algorithm," International Journal of Engineering and Emerging Technology, [S.l.], v. 3, n. 1, p. 52-55, july 2018, ISSN 2579-597X, 2018.

[15] Grabusts Pēteris, Borisovs Arkādijs, Aleksejeva Ludmila, "Decision Tree Creation Methodology Using Propositionalized Attributes," Information Technology and Management Science, Vol. 19, pp 34-38, De Gruyter Open, Riga Technical University, Latvia, 2016.

[16] Song Yan-Yan, LU Ying, "Decision tree methods: applications for classification and prediction," Shanghai Archives of Psychiatry, Vol.27, No. 2, Shanghai, 2015.