Rule-Based Data Mining for Diagnosis of Coronary Heart Disease

Hanung Adi Nugroho[1], Dwi Normawati[1, 2], Noor Akhmad Setiawan[1] and Widhia K.Z. Oktoeberza[1]
[1]*Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada,*
[2]*Department of Informatics Engineering, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia.*
adinugroho@ugm.ac.id

***Abstract*— Coronary heart disease is the leading cause of human death due to the presence of plaque (fat) in the blood vessels. Electrocardiograph (ECG) and treadmill tests are commonly used for coronary disease detection. However, it is costly, at risk and sometimes the diagnosis result is not accurate. This research aims to classify coronary heart disease dataset based on two rules of data mining methods, i.e. variable precision rough set (VPRS) and repeated incremental pruning to produce error reduction (RIPPER). These rules are chosen to observe the simplest pattern of rules knowledge from big data, imprecise and ambiguous data. The proposed method is evaluated on Cleveland coronary heart disease dataset taken from the UCI repository. The combination of VPRS and RIPPER obtains the best evaluation result with accuracy achieved of 92.99%. While the accuracy of VPRS and RIPPER is merely 75.22% and 88.13%, respectively. It indicates that the proposed method successfully classifies coronary heart disease dataset and has a potential to be implemented in the development of a computerised coronary heart disease diagnosis system.**

***Index Terms*— Cleveland Dataset; Coronary Heart Disease; Repeated Incremental Pruning to Produce Error Reduction (RIPPER); Variable Precision Rough Set (VPRS).**

I. INTRODUCTION

Coronary heart disease is the primary cause of death in the worldwide. The number of patients with coronary heart disease is increasing every year. According to reported data by World Health Organisation (WHO), around 17.3 million people worldwide died from cardiovascular diseases (CVDs) and 7.3 million of them is caused by coronary heart disease [1]. The main symptom of coronary heart disease is chest pain or angina. However, this symptom is ambiguous because chest pain can also often occur in conditions that may not be accompanied by coronary heart disease. This condition may lead to wrong diagnosis [2].

Along with the information technology development, computer aided diagnosis for coronary heart disease were widely developed. One of the techniques to handle large amount of data is data mining [3]. Data mining in the medical world has a great potential to find hidden pattern in medical dataset. This pattern can be used to assist the doctor to improve the quality of medical decision in revealing either presence or absence of a disease [4].

Previous researches showed that RIPPER [5, 6], artificial neural network (ANN) [7], support vector machine (SVM) [6] and decision tree were applied for coronary heart disease diagnosis. One of the disadvantages of machine learning method such as ANN is yet to show the transparency of knowledge, dependence on a big amount data, slow computation and it local optimum. SVM capable to generate the diagnosis accurately, but it is prone to overfitting and have a slow computation. Decision Tree is built with a lot of branches lead that may cause anomalies and noise in the data training. RIPPER is an optimisation algorithm that is greedy and tend to over fit, but is fast rules-based classification because RIPPER is capable to learn the multi-model dataset rules and provide a result which easy to be interpreted. Variable precision rough set (VPRS) method is a development of rough set theory [8, 9]. VPRS usage in coronary heart disease detection [10], generating less rules than the rough set method [11], but the accuracy value is not significantly different. The aim of conducting the combination of RIPPER and VPRS method is to improve the classification performance and increase the accuracy values in the diagnosis of coronary heart disease.

II. DATASET

This research employees 303 data of Cleveland heart disease dataset from UCI machine learning repository which have seven data of missing values. Missing values data is deleted in order not to affect the classification result. Cleveland heart disease dataset consist of 14 attributes with two classes of dataset which are the absence and the presence of heart disease. Table 1 describes the attributes in Cleveland heart disease dataset [12].

III. METHODOLOGY

The approach is described as follows. Three main processes consist of pre-processing, generated rules and classification process. Firstly, 296 data are selected from 303 data by deleting the missing values data by cleaning process. Two rules based of data mining methods are applied, which are VPRS and RIPPER. The research flowchart is depicted in Figure 1.

A. Pre-processing

Pre-processing stage is the initial stage of the diagnosis process. It consists of five main steps, i.e. data cleaning of missing value, data conversion from multiclass to binary class, 30 times of randomisation, discretisation using entropy/MDL algorithm and splitting dataset. Training data is conducted to seek knowledge in data in form of rules. Testing data aims to test data by matching result of knowledge with data.

B. Generated Rules

This research uses two methods to generate rules which are VPRS by using ROSE2 software and RIPPER by using WEKA software. This research uses two methods to generate rules which are VPRS by using ROSE2 software and RIPPER by using WEKA software.

Table 1
Summary of Cleveland Heart Disease Dataset

Attribute	Description	Value Description
Age	Age	Numeric
Sex	Sex	0:Female; 1:Male
Cp	Chest pain type	1 : typical angina; 2 : atypical angina; 3 : non-anginal pain; 4 : asymptomatic
Trestbps	Resting blood pressure	Numeric
Chol	Serum cholesterol	Numeric
Fbs	Fasting blood sugar >120mg/dl	0 : false; 1 : true
Restecg	Resting electrocardiographic result	0 : normal; 1 : having ST-T wave abnormality; 2 : showing probable or definite left ventricular hypertrophy by Estes' criteria
Thalac	Maximum heart rate achieved	Numeric
Exang	Exercise induced angina	0 : No; 1 : Yes
Oldpeak	Segment ST depression induced by exercise relative to test	Numeric
Slope	The slope of the peak exercise ST segment	1 : usloping; 2 : flat; 3 : downsloping
Ca	Number of major vessels coloured by fluoroscopy	0, 1, 2 and 3
Thal	Thal	3 : normal; 6 : fixed defect; 7 : reversible defect

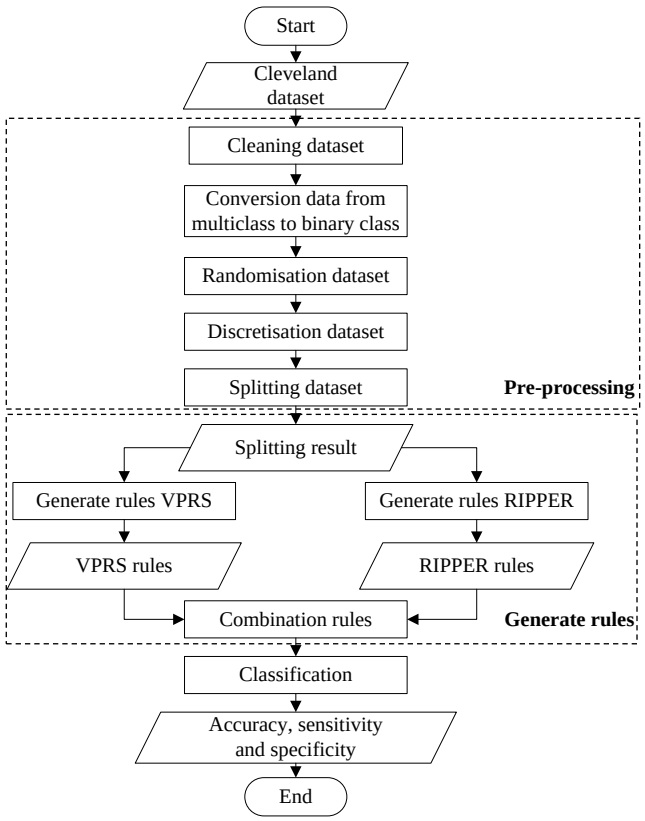

Figure 1: Research Flowchart

C. Variable Precision Rough Set (VPRS)

VPRS is the continuation of classical model of rough set. In this research, it is proposed to analyse and identify the data pattern, which is representing the functional statistic trend [9]. VPRS related to the classification of partial precision detection of parameter β. Ziarko defined the value of β as a miss-classification and ranged in value $0 \leq \beta < 0.5$. VPRS model procedure have four steps [13], namely:

Step-1: chosen a precision parameter of value (β);
Step-2: finding a full set of β-reduct;
Step-3: remove duplicate objects;
Step-4: rule extraction.

VPRS is an approach to analyse data that relies on two basic concepts, namely β-lower and β-upper approximations which can be expressed in Equation (1) and Equation (2).

$$\underline{C}_{\beta}(D) = \bigcup_{1-P_r(Z|x_i)\leq\beta} \left\{x_i \in E(P)\right\} \quad (1)$$

$$\overline{C}_{\beta}(D) = \bigcup_{1-P_r(Z|x_i)<1-\beta} \left\{x_i \in E(P)\right\} \quad (2)$$

Here, $\underline{C_{\beta}}(D)$ and $\overline{C_{\beta}}(D)$ are the lower and upper approximation of D with precision level β, respectively. For $E(P)$ indicates a set of equivalent classes, and class conditions based on subsets of attributes P, while $Z \subset E(D)$. P is mathematically formulated in Equation (3).

$$P_r(Z \mid x_i) = \frac{Card(Z \cap x_i)}{Card(x_{i)}} \quad (3)$$

According to [8], the classification quality measurement for VPRS model can be defined by Equation (4) as follows :

$$\gamma(P, D, \beta) = \frac{Card(\cup_{1-P_r(Z|x_i)\leq\beta} \left\{x_i \in E(P)\right\})}{Card(U)} \quad (4)$$

with $Z \subset E(D)$ and $P \subseteq C$, for certain β value. The value of $\gamma(P, D, B)$ measures the proportion of objects on set universe (U) for classification based on decision attribute D, and allowing for certain β value.

The procedure to produce a decision rule of an information system is conducted by two major steps, namely:

Step 1: select the smallest set of attributes (e.g. β-reduct value selection);
Step 2: simplify the information system by dropping the specific values of the unnecessary attributes.

Ziarko [9] indicated that every smallest set of attributes is consider as an alternative to group the attributes which used as substitute all attributes available in case based decision making.

D. Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

RIPPER is a machine learning algorithm rule based on refinement of incremental reduce error pruning (IREP) [6].

IREP does the rule reduction. After that, it is checked and it is considered not good. RIPPER adds terms to growing and pruning, as well as the generation of variation for optimisation rule. The RIPPER algorithm is briefly explained as follows [14]:

1. Building stage
   Split E into Growing and Pruning sets in the ratio 2:1.
   Repeat until:
   (a) there are no more uncovered examples of C; or
   (b) the description length (DL) of Ruleset and examples is 64 bits greater than the smallest DL found so far, or
   (c) the error rate exceeds 50%.
   1.1 Grow phase: Grow a rule by greedily adding conditions until the rule is 100% accurate by testing every possible value of each attribute and selecting the condition with greatest information gain G.
   1.2 Prune phase: Prune conditions in last-to-first order. Continue as long as the worth W of the rule increases
2. Optimisation stage
   2.1 Generate variants: for each rule R for class C, Split E afresh into Growing and Pruning sets.
   Remove all instances from the Pruning set that are covered by other Rules for C.
   Use GROW and PRUNE to generate and prune two competing Rules from the newly split data:
   R1 is a new rule, rebuilt from scratch;
   R2 is generated by greedily adding antecedents to R.
   Prune using the metric A (instead of W) on this reduced data.
   2.2 Select representative: replace R by whichever of R, R1 and R2 has the smallest DL.
3. Move up
   If there are residual uncovered instances of class C, return to the Build stage to generate more Rules based on these instances.
4. Clean up
   Calculate DL for the whole Ruleset and for the Ruleset with each rule in turn omitted; delete any rule that increases the DL.
   Remove instances covered by the Rules just generated.

E. Classification

Classification is an evaluation process by analysing the confusion matrix [15] consisting of accuracy, sensitivity and specificity values. Confusion matrix show in Table 2.

Table 2
Confusion Matrix

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

The accuracy is the success rate of classification which measure by counting the number of correct classification divided by the total classification. Sensitivity is the probability the patient said to suffer from coronary heart disease was diagnosed positive illness (sick), while specificity was diagnosed negative illness (healthy).

IV. RESULT

In the research work used 296 data taken from Cleveland heart disease dataset. The data discretisation is performed by ROSETTA software. Randomisation and rules generating process for VPRS method is conducted using ROSE2 software. Rules process for RIPPER method is determined using WEKA 3.6 software. Classification process is calculated manually by using Microsoft Excel.

A. Pre-processing

The first step in the pre-processing data is data cleaning process to remove the missing value in dataset Cleveland dataset, then convert from multiclass dataset into binary class dataset with an assumption that positive class is healthy (0) and negative class is sick (1).

The second step is data randomisation. Randomisation data scale is performed 30 times on all of Cleveland dataset attributes sequentially resulting in 30 new datasets with different sequences of data. These datasets are used in the next process.

The third step is data discretisation. Discretisation changes the data type of attributes from numeric type to discrete type. Some attributes with numeric type are age, trestbps, chol, thalach, oldpeak and ca, and transform into discrete type using entropy/MDL algorithm. Table 3 shows the result of data discretisation.

Table 3
The Result of Data Discretisation

	Numeric types of attributes					
	Age	Trestbps	Chol	Thalach	Oldpeak	Ca
Discrete values	[*, 71) [71, 77) [77, *)	[*, 186) [186, *)	[*, 276) [276, 277) [277, 280) [280, 295) [295, 299) [299, 301) [301, 319) [319, 320) [320, 322) [322, 324) [324, 326) [326, 338) [338, 341) [341, 342) [342, 348) [348, 354) [354, 401) [401, 413) [413, *)	[*, 148) [148, 151) [151, 162) [162, 170) [170, 172) [172, 175) [175, 176) [176, 178) [178, 183) [183, 195) [195, 199) [199, *)	[*, 2.5) [2.5, 2.7) [2.7, 3.1) [3.1, 3.5) [3.5, 3.6) [3.6, 4.3) [4.3, *)	[*,3) [3, *)

As shown in Table 3, the sign of "*" on the left indicates the "less than value (<)", but if it on the right indicates the "more than value (>)". If there is no sign, it means that "between value". Thus, the value of [*, 71) means that the age is less than 71 (>71), the value of [71, 77) means that 71 ≤ age ≤ 77, the value of [77, *) means that the age is more than 77 (>77).

Discretisation process divides the attribute into some particular intervals. The thresbps and ca attributes are discretised into two intervals. The age is discretised into three intervals while oldpeak attribute is discretised into seven intervals. For chol and thalac attribute are discretised into 19 and 12 intervals, respectively.

The last step in pre-processing data is data splitting which splits the dataset into two parts by the same amount. A total of 148 data as training data and 148 the rest is used testing data. Training dataset is used to find rules and knowledge on dataset for diagnosing coronary heart disease; while

testing dataset is used to test data with matching class prediction result of rules knowledge with class dataset.

B. Generate Rules/IF-THEN

After splitting process, the next step is generating IF-THEN rules by using VPRS and RIPPER method.

C. Variable Precision Rough Set (VPRS)

In order to get IF-THEN rules or decision rules for VPRS method, the value β = 0.15 is used by using ROSE2 software. In the research work, 30 datasets are used which resulted from data randomisation of 30 times trial. Each dataset produces different rules and numbers. Table 4 shows the sample of rules result dataset.

Table 4
Sample Rules from First Randomisation Dataset by VPRS

Rules	Attribute description
1	(sex = 0) & (chol = 0) & (exang = 0) & (thal = 3) => (class = 0)
2	(cp = 2) & (restecg = 0) & (thal = 3) => (class = 0)
…	……
43	(cp = 4) & (chol = 0) & (thalach = 2) & (exang = 0) & (slope = 1) => (class = 0)
44	(cp = 4) & (chol = 0) & (thalach = 2) & (exang = 0) & (slope = 1) => (class = 1)

D. Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

RIPPER is a rule-based machine learning algorithm. In the research work, VPRS method is merged with RIPPER method in order to increase the accuracy value. Thirty dataset resulted from data randomisation are used and generate rules after 30 times trial. The dataset for generalisation rules process is training dataset. Therefore, each datasets from 30 randomisation datasets generates one rule. Table 5 shows the sample of rules from generalisation rules process by using RIPPER method.

Table 5
The Sample of Rules from First Randomisation Dataset by RIPPER

Rules	Attribute description
1	(exang = 1) and (thal = 7) => class=1
2	(sex = 1) and (thalach <= 0) => class=1
3	(restecg = 2) and (thal = 7) => class=1
4	(oldpeak >= 2) => class=1
5	class=0

E. VPRS and Ripper Rules Combination

The combination of VPRS and RIPPER methods is conducted to observe the accuracy value result. The merger rules process is done manually by using Microsoft Excel software. Merging rules is applied into 30 randomisation datasets. Table 6 shows the merger rules of VPRS and RIPPER results.

F. Classification

The classification based on VPRS and RIPPER has been done either individually or merger condition. Rules generated by individually or combined methods are tested on the testing dataset. The examination is applied into 30 testing dataset, so the confusion matrix is obtained for each dataset. Table 7 shows the confusion matrix value of VPRS, RIPPER and merger of VPRS & RIPPER rules.

As shown in Table 7, the RIPPER yields the highest average of TP rate with around 79.9 sick data are correctly classified as sick data. But the lowest average of FP rate is obtained by combination of VPRS and RIPPER at 0.23. It indicates that the less healthy data is classified as sick data. Moreover, combination of VPRS and RIPPER also obtains the lowest average of FP rate which indicates only around 10.13 sick data are classified as healthy data. The lesser the false rate the better result the performance evaluation.

Table 6
Sample Merge Rules VPRS and RIPPER

Rules		Attribute description
VPRS	1	(sex = 0) & (chol = 0) & (exang = 0) & (thal = 3) => (class = 0)
	2	(cp = 2) & (restecg = 0) & (thal = 3) => (class = 0)
	…	……
	43	(cp = 4) & (chol = 0) & (thalach = 2) & (exang = 0) & (slope = 1) => (class = 0)
	44	(cp = 4) & (chol = 0) & (thalach = 2) & (exang = 0) & (slope = 1) => (class = 1)
RIPPER	1	(exang = 1) and (thal = 7) => class=1
	2	(sex = 1) and (thalach <= 0) => class=1
	3	(restecg = 2) and (thal = 7) => class=1
	4	(oldpeak >= 2) => class=1
	5	class=0

G. Performance Evaluation

In the medical contexts, there are only two classes “sick” or “healthy”, which “sick” is more important than “healthy”. The medical diagnosis purpose is to focus on the improvement of the accuracy of “sick” class or sensitivity and maintain the accuracy of “healthy” class or specificity. The accuracy, sensitivity and specificity values can be calculated from the confusion matrix for each method by using Equation (5) until Equation (7).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} x100\% \quad (5)$$

$$Sensitivity = \frac{TP}{TP + FN} x100\% \quad (6)$$

$$Specificity = \frac{TN}{TN + FP} x100\% \quad (7)$$

Table 8 shows that the combination of VPRS and RIPPER generates a better accuracy compared to the separated methods. The evaluation of performance method of VPRS and RIPPER shows on the calculation of sensitivity and specificity values.

V. Conclusion

This study provides a data mining scheme for coronary heart disease diagnosis by performing randomisation and classification using Cleveland dataset. From this research some points can be concluded as follows:

1. The results of diagnosis of coronary heart disease tend to be biased if performed on VPRS and RIPPER methods individually as a consideration in the final decision maker.

2. A combination method of VPRS and RIPPER is able to increase the accuracy value to 92.99% of classification performance.

The comparison result from testing process shows that the diagnosis process of coronary heart disease by using the combination of VPRS and RIPPER methods indicated by better accuracy than that of individually separated methods.

Table 7
Confusion Matrix Values of 30 Datasets Randomisation with VPRS, RIPPER and Merger of VPRS & RIPPER

Randomisation Dataset	VPRS				RIPPER				VPRS & RIPPER			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
1	56	51	24	17	80	54	0	14	80	63	0	5
2	59	54	22	13	80	59	0	9	80	61	1	6
3	67	48	13	19	80	53	0	15	80	59	0	9
4	73	56	7	12	80	46	0	22	78	57	2	11
5	61	43	17	25	80	45	0	23	80	49	0	19
6	51	51	29	17	80	51	0	17	80	61	0	7
7	57	47	24	20	80	46	0	22	79	53	2	14
8	60	42	19	26	80	54	0	14	80	58	0	10
9	55	50	25	18	80	58	0	10	80	63	0	5
10	61	51	19	17	80	48	0	20	80	54	0	14
11	63	44	17	24	80	48	0	20	80	55	0	13
12	59	56	20	13	80	54	0	14	80	60	0	8
13	64	58	16	10	80	57	0	11	80	65	0	3
14	64	52	16	16	80	48	0	20	80	57	0	11
15	65	52	17	14	80	49	0	19	80	61	2	5
16	60	45	17	26	80	46	0	22	77	52	0	19
17	65	50	15	18	80	48	0	20	80	56	0	12
18	55	51	25	17	80	56	0	12	80	59	0	9
19	66	52	14	16	80	54	0	14	80	60	0	8
20	56	58	24	10	80	46	1	21	80	62	0	6
21	60	47	15	26	80	50	0	18	75	59	0	14
22	56	45	24	23	80	56	0	12	80	58	0	10
23	62	50	18	18	78	46	2	22	80	54	0	14
24	56	42	24	26	80	54	0	14	80	57	0	11
25	71	49	9	19	80	47	0	21	80	56	0	12
26	61	54	19	14	80	45	8	15	80	60	0	8
27	61	46	19	22	80	51	0	17	80	61	0	7
28	68	46	12	22	80	55	0	13	80	61	0	7
29	73	47	7	21	80	42	1	25	80	51	0	17
30	61	54	19	14	79	50	1	18	80	58	0	10
Average	**61.53**	**49.7**	**18.2**	**18.43**	**79.9**	**50.53**	**0.43**	**17.13**	**79.63**	**58**	**0.23**	**10.13**

Table 8
The Comparison Evaluation Result of VPRS, RIPPER and Merger of VPRS & RIPPER

Rules	Accuracy (%)	Sensitivity (%)	Specificity (%)
VPRS	75.22	77.11	73.55
RIPPER	88.13	82.49	99.14
VPRS+RIPPER	92.99	88.87	99.60

REFERENCES

[1] W. H. Organisation. (2017, 29 May). *About cardiovascular diseases*. Available: http://www.who.int/cardiovascular_diseases/about_cvd/en/

[2] B. L. Zaret, L. S. Cohen, and M. Moser, *Yale university school of medicine heart book*: William Morrow and Co., 1992.

[3] T. J. Peter and K. Somasundaram, "Study and development of novel feature selection framework for heart disease prediction," *International Journal of Scientific and Research Publications*, vol. 2, pp. 1-7, 2012.

[4] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, pp. 43-48, 2011.

[5] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the twelfth international conference on machine learning*, 1995, pp. 115-123.

[6] M. Kumari and S. Godara, "Comparative study of data mining classification methods in cardiovascular disease prediction 1," 2011.

[7] A. H. Chen, S.-Y. Huang, P.-S. Hong, C.-H. Cheng, and E.-J. Lin, "HDPS: Heart disease prediction system," in *Computing in Cardiology, 2011*, 2011, pp. 557-560.

[8] W. Ziarko, "Variable precision rough set model," *Journal of computer and system sciences*, vol. 46, pp. 39-59, 1993.

[9] W. Ziarko, "Probabilistic decision tables in the variable precision rough set model," *Computational Intelligence*, vol. 17, pp. 593-603, 2001.

[10] N. A. Setiawan and H. A. Nugroho, "Deteksi Penyakit Jantung Koroner Menggunakan Model Variable Precision Rough Set dan Logika Fuzzy," Universitas Gadjah Mada, 2014.

[11] B. Tripathy, D. Acharjya, and V. Cynthya, "A framework for intelligent medical diagnosis using rough set with formal concept analysis," *arXiv preprint arXiv:1301.6011*, 2013.

[12] K. Bache and M. Lichman, "UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. University of California, School of Information and Computer Science," *Irvine, CA*, 2013.

[13] C.-T. Su and J.-H. Hsu, "Precision parameter in the variable precision rough sets model: an application," *Omega*, vol. 34, pp. 149-157, 2006.

[14] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2016.

[15] J. Han, "Micheline kamber, 2006."Data Mining: Concepts and Techniques," ed: Mortgan Kaufmann Publishers, 2005.