CLASSIFICATION OF VILLAGE DEVELOPMENT INDEX AT REGENCY/ MUNICIPALITY LEVEL USING BAYESIAN NETWORK APPROACH WITH K-MEANS DISCRETIZATION

Nasiya Alifah Utami[1], Arie Wahyu Wijayanto[1,2]

[1] Politeknik Statistika STIS, Jl. Otto Iskandardinata No.64C, Jakarta, Indonesia
[2] BPS-Statistics Indonesia, Jl. Dr. Sutomo 6-8, Jakarta, Indonesia
e-mail: [1]221810496@stis.ac.id, [2]ariewahyu@stis.ac.id

Abstract

Village development has been one of the most important targets of government policies in Indonesia in order to fully optimize its potential. Under Law 06 Year 2014 on Villages, local governments from regency/municipality level to village level are required to understand their respective village potentials in order to increase the village potentials in their regions. In this paper, we build and analyze the Bayesian network methods to classify the village development index at regency/municipality and gain a better understanding of the causal relationships between independent variables of the village potential status. Using a web scraping method of information retrieval, data are collected from the Ministry of Village, Development of Disadvantaged Regions, and Transmigration (Kemendesa) website, and Village Development Evaluation (Indeks Pembangunan Desa—IPD) of Statistics Indonesia (BPS) publication in 2018 data. Further, we combine the discretization using the K-Means clustering method to handle the continuous nature of retrieved data. An extensive comparison of different learning structures of the Bayesian Network is performed, which includes the learning structure of Naive Bayes, Maximum Spanning Tree with weighted Spearman correlation coefficient, Hill Climbing search, and Tabu Search during the construction of Bayesian networks. For fairness evaluation, all constructed models are built using 80% data as a training set and the remaining 20% as a testing set. The results show that Bayesian network approach can be applied in village development index status classification where the construction using maximum spanning tree with K-Means data discretization gain the best performance of 90.69% accuracy.

Abstrak

Pembangunan desa merupakan salah satu target penting dari kebijakan pemerintah di Indonesia dalam rangka mengoptimalkan potensinya secara maksimal. Berdasarkan Undang-Undang Nomor 6 Tahun 2014 tentang Desa, pemerintah daerah mulai dari tingkat kabupaten/kota hingga tingkat desa perlu memahami potensi desa masing-masing guna meningkatkan potensi desa di daerahnya. Dalam penelitian ini, kami membangun dan menganalisis metode Bayesian *network* untuk mengklasifikasikan indeks pembangunan desa di kabupaten/kota dan mendapatkan pemahaman yang lebih baik tentang hubungan kausal antara variabel independen dari status potensi desa. Dengan menggunakan metode *web scraping*, data dikumpulkan dari website Kementerian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi (Kemendesa), dan publikasi Evaluasi Pembangunan Desa (IPD) oleh Badan Pusat Statistik (BPS) tahun 2018. Selanjutnya, kami mengaplikasikan diskritisasi data menggunakan metode klasterisasi K-Means untuk menangani data yang bertipe kontinu. Struktur pembelajaran Bayesian *network* yang digunakan dalam penelitian ini adalah Naïve Bayes *learning structure*, *Maximum Spanning Tree* (MST) *learning structure* yang diboboti dengan koefisien korelasi Spearman, dan *learning structure* hasil optimasi dengan *Hill Climbing Search* dan Tabu *Search*. Untuk keperluan evaluasi, set data dibagi kedalam 80% data latih dan 20% data uji. Hasil penelitian menunjukkan bahwa pendekatan Bayesian *network* dapat diterapkan dalam klasifikasi status indeks pembangunan desa. Struktur pembelajaran terbaik berdasarkan hasil adalah metode *Maximum Spanning Tree* (MST) dengan diskritisasi data K-means yang memiliki akurasi mencapai 90,69%.

INTRODUCTION

The village as the lowest independent administrative area in Indonesia is one of the development subjects [1]. The development of village areas can be done by optimizing the exploration of village potential. Optimization of village potential is required to make this potential the main support for the village economic sector. Exploring village potential is one of the main focuses of the Ministry of Villages, Development of Disadvantaged Regions and Transmigration (KDPT). Exploration of village potential aims to reduce the development gap between urban and rural areas.

Under Law 06 Year 2014 on Villages, villages have to be protected and empowered in order to strengthen, advance, and be independent. According to the Law, local governments from the provincial level and regency/municipality level to village level need to know their respective village potentials in order to advance the village potentials in their regions. This matter is also discussed in Indonesia's Mid-Term Development Plan (Rencana Pembangunan Jangka Menengah – RPJMN) for 2020-2024 that increasing village potential become an effort to reduce disparities between regions [2]. Comprehensive and regular monitoring of regional village conditions by policymakers is needed. The village status data that is available from provincial to village level is important for monitoring basis. Unfortunately, it became a problem when Statistics Indonesia (BPS) was only able to publish the village development index at the district level, then the information of village potential status cannot be comprehensively obtained. Because of that, it is necessary to build a classification model for predicting village potential status at wider area such as provincial level or smaller area such as sub-district or village level.

This study proposed the use of the Bayesian network method to classify the village development index status at regency/municipality level in Indonesia. The Bayesian network method is used because it can explain the causality relationship among independent variables [3]. In previous studies, many researchers have used Bayesian network models as classification tools. The research was conducted using various Bayesian network learning structures, then the classification models were compared with each other. Data discretization using the k-means method has also been carried out in previous

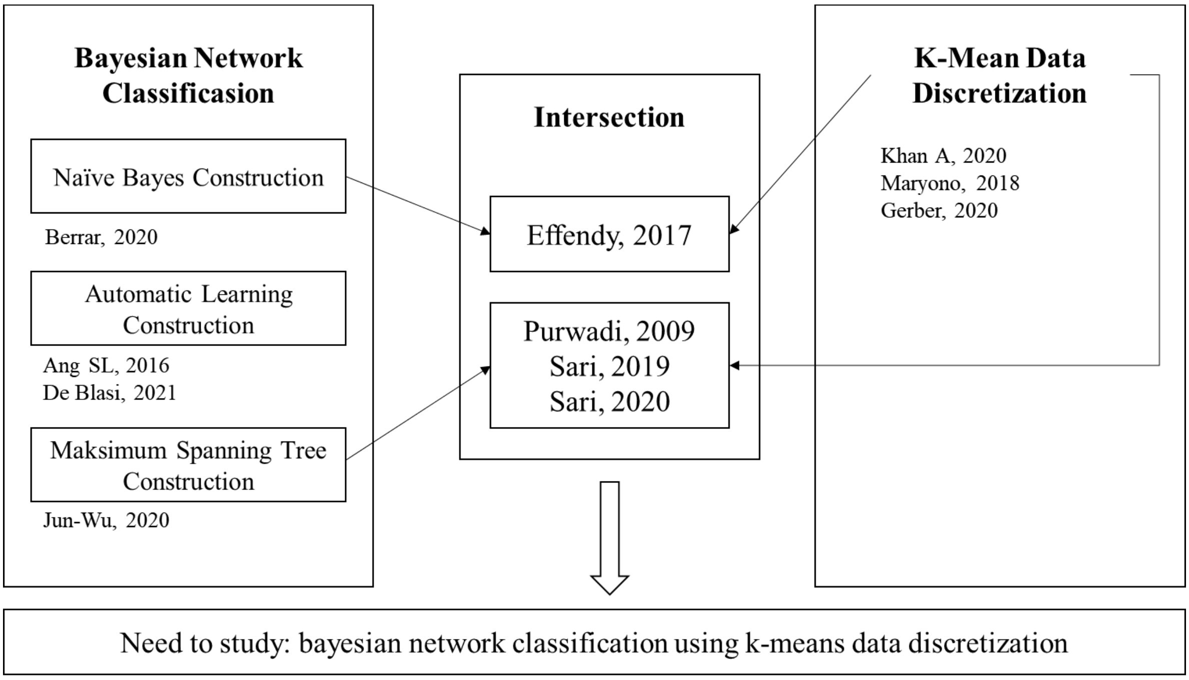

Bayesian Network Classificasion
Naïve Bayes Construction
Berrar, 2020
Automatic Learning Construction
Ang SL, 2016
De Blasi, 2021
Maksimum Spanning Tree Construction
Jun-Wu, 2020
Intersection
Effendy, 2017
Purwadi, 2009
Sari, 2019
Sari, 2020
K-Mean Data Discretization
Khan A, 2020
Maryono, 2018
Gerber, 2020
Need to study: bayesian network classification using k-means data discretization

Figure 1. Literature Map

studies, in the research of Effendy et al 2017, the classification model with data discretization using k-means has better accuracy than the classification model without k-means discretization data. Based on this, the k-means method for continuous data discretization was used in this study. The literature that forms the basis for this research is described in the literature map as follows.

This study aims to build a Bayesian network classification model to predict the village potential status at a wider or narrower area level and choose the best model from several models built. This research is expected to provide information on village potential status as a basis for monitoring. Information about village potential status can be used by local governments to focus their efforts on developing village potential in their respective regions.

THEORETICAL BACKGROUND

1. Village Development Index (VDI)

The Village Development Index (Indeks Desa Membangun - IDM) is constructed from the average of the Social Resilience Index (SocRI), Economic Resilience Index (EcoRI), and Environmental Resilience Index (EnvRI) using the following formula:

$$\boldsymbol{VDI = \frac{SocRI + EcoRI + EnvRI}{3}} \qquad (1)$$

The index calculation for each dimension, i.e., SocRI, EcoRI, and EnvRI, is constructed out using the scoring method from the aggregate of the constituent indicators which is then transformed into an index, as follows:formula:

$$\boldsymbol{I_x = \frac{\sum Indicator_y}{n_x . 5}} \qquad (2)$$

where $\boldsymbol{I_x}$is the dimensional index (SocRI, EcoRI, and EnvRI), $\boldsymbol{n_x}$is the number of indicator in corresponding dimensional index. $\boldsymbol{Indicator_y}$denotes the score of indicator in the village.

Each indicator that compose the dimensional index (SocRI, EcoRI, and EnvRI) has a score value in the interval of 0 - 5. To determine the score, the Focus Group Discussion (FGD) Analytical Hierarchy Process (AHP) is utilized. For instance, the Environmental Resilience Index (EnvRI) consists of 3 indicators, namely environmental quality indicator, disaster-prone indicator, and disaster response indicator. Sukamakmur Village has an environmental quality score of 4, a disaster-prone score of 5, and a disaster response score of 3. Hence, the EnvRI of Sukamakmur village is as follows:

$$\boldsymbol{EnvRI_{sukamakmur} = \frac{5+3+2}{3x5} = \frac{10}{15} = 0.67} \qquad (3)$$

Likewise, the EcoRI and SocRI values of Sukamakmur village can be computed in a similar manner. Classification of Village Status is determined with the following thresholds:

Table 1. Classification of Village Status based on VDI Value

Village Status	VDI Value
1. Very Underdeveloped Village	0.0000 - 0.4907
2. Underdeveloped Village	0.4907 - 0.5989
3. Developing Village	0.5989 - 0.7072
4. Developed Village	0.7072 - 0.8155
5. Independent Village	0.8155 – 1.0000

The classification of village status is constructed to assess the status of development and provide recommendations for required policy interventions. Hence, the government policies, approaches, and interventions applied to the each different village status will be different.

2. Village Development Evaluation (VDE)

Village Development Evaluation (VDE) — Indeks Pembangunan Desa (IPD) composed of several fundamental indicators, includes the Basic Services, Infrastructure Conditions, Accessibility/Transportation, Public Services, and Governance indicators.

METHODS

The research method for building Bayesian network classification model prediction of village potential status in Indonesia has five main process. The process are data collection, data preprocessing, learning structure construction, Bayesian network classification model formation, and model evaluation. The research process of this study is systematically illustrated in Figure 2.

Bayesian network classification model approach in this study is explored by comparing several learning structure method with and without k-means discretization data. The learning structure methods used are human expert learning structure which based on correlation analysis and researcher’s subjectivity, Naive Bayes learning structure, maximum spanning tree learning structure with spearman coefficient correlation weighed, and automatic learning structure which includes Hill Climbing search and Tabu search with BIC score based. The model performance evaluation carried out by using accuracy, precision, recall, and F1-Score. The best classification model is the model that has highest evaluation measurement score.

1. Data Collection Method

The data used in this study is secondary data that collected from BPS-Statistics Indonesia publication of village development index in 2018 and Kemendesa website. The data from Kemendesa website is village potential status data at district level in Indonesia. This data is collected by web scraping method using scrapy tools in Python. Web Scraping is a technique for extracting data from the World Wide Web (WWW) and saving it to a file system or database for retrieval or analysis [3].

2. Data Preprocessing

Before the classification model is applied to the data, it is necessary to clean the data from web scraping results. Then join data method is used to merge BPS data and Kemendesa data. This process is done

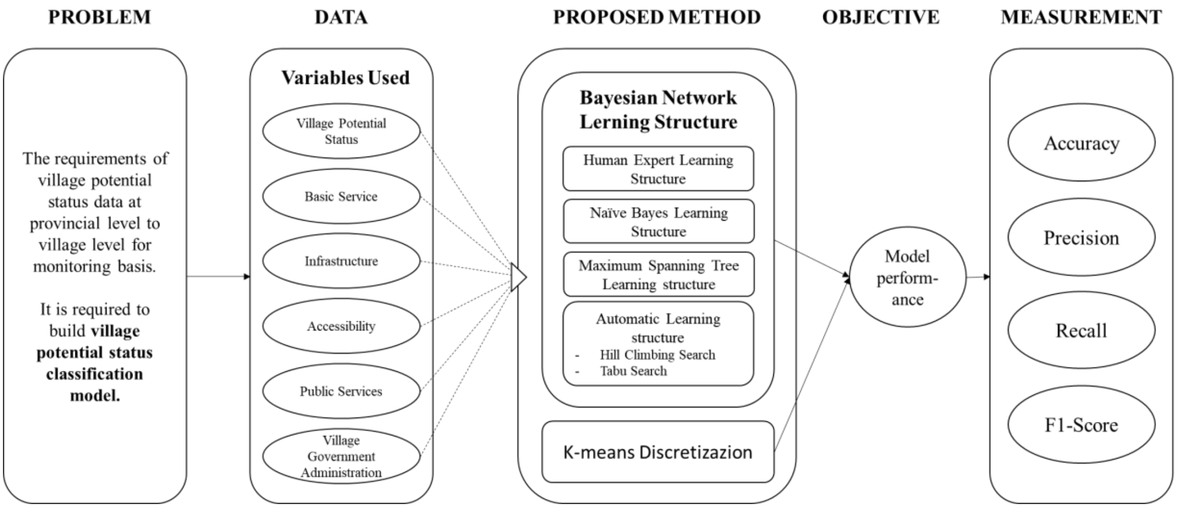

PROBLEM
DATA
PROPOSED METHOD
OBJECTIVE
MEASUREMENT
The requirements of village potential status data at provincial level to village level for monitoring basis.
It is required to build village potential status classification model.
Variables Used
Village Potential Status
Basic Service
Infrastructure
Accessibility
Public Services
Village Government Administration
Bayesian Network Lerning Structure
Human Expert Learning Structure
Naïve Bayes Learning Structure
Maximum Spanning Tree Learning structure
Automatic Learning structure
- Hill Climbing Search
- Tabu Search
K-means Discretizazion
Model perform-ance
Accuracy
Precision
Recall
F1-Score

Figure 2. Research Framework

by using tools *dplyr* in RStudio. To discretize the data, K-Means clustering method is used. K-Means is a non-hierarchical clustering method that divides the available data into one or more clusters [4]. This method divides the data into k distinct classes with the same characteristic in each cluster. In general, K-Means method is described by following algorithm:

- Determine the optimum number of clusters
- Allocate the data randomly into each cluster
- Calculate the centroid value of each cluster
- Allocate the data each to the nearest centroid
- Perform the step 3 and 4 until no data moves to other cluster
- Perform profiling on the clusterization data results.

In preprocessing stage of the data, the data are divided into 80% as training set and 20% as testing test. The number of data rows is 434 data as the number of districts and cities at rural areas in Indonesia.

3. Learning Structure of Bayesian Network Construction

In this study, several methods were used to build a Bayesian network learning structure. The first method is a computational algorithm method using naive Bayes construction, maximum spanning tree with the weighted Spearman correlation, Hill Climbing search and Tabu Search. The description for each algorithm is given in the following.

Naïve Bayes Construction.

Naive Bayes is a simple probabilistic classifier that calculates a set of probabilities by adding up the frequency and combination of values from a given dataset [4]. The construction of naive Bayes learning structure forms a structure centered on the target variable with no dependencies between variables that are classification features. The following is an illustration of the construction of a naive Bayes learning structure.

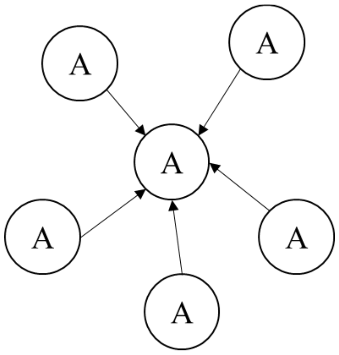


Figure 3. Illustration Of Naïve Bayes Learning Structure Construction

Maximum Spanning Tree Construction

Maximum spanning tree is a subset of a graph G that has no cycles but has the same nodes as G [6]. In this study, the weight used in the formation of the maximum spanning tree is the Spearman correlation coefficient which describes relationship between the two variables. The maximum spanning tree algorithm used to find the maximum weight value of the Spearman correlation coefficient is Prim algorithm. The following is an illustration of learning structure using maximum spanning tree.

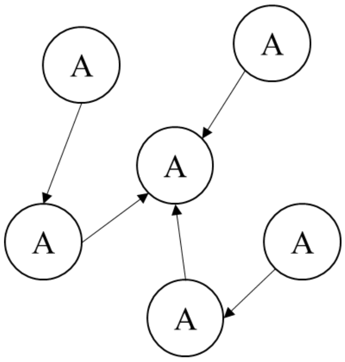


Figure 4. Illustration Of Maximum Spanning Tree Learning Structure Construction

Hill Climbing Search Algorithm

Hill climbing algorithm is a search algorithm that determines the next search step by placing the point that will appear as close as possible to the target [7]. In the construction of learning structure bayesian networks, this method is used to find the right structure by adding edges, deleting edges, and reversing edges. Each edge is weighted by using the Bayesian

Information Criterion (BIC) value. This learning structure algorithm is known as the automatic learning structure method. The `hc` function in `bnlearn` RStudio package was used to build the hill-climbing learning structure construction.

Tabu Search Algorithm

Tabu search algorithm is a sub-heuristic algorithm that simulates human memory functions [6]. For the construction process of the Bayesian network structure, the algorithm uses three operations to generate an environment without generating network loops: 1) adding edges, subtracting edges, and inverting edges, and then finding the local optimal solution in the environment and inserting it into the tabu table. Similar to the hill-climbing method, each edge is weighted by using the Bayesian Information Criterion (BIC) value. The formation of the structure with Tabu search is done by using the `tabu` function in the `bnlearn` RStudio package.

4. Bayesian Network Model

Bayesian network is a method that can describe the causal relationship between variables in a system [8]. Bayesian network can be used as a classification model. The bayesian network classification models are built by using training set which is 80% of the observation data. The classification process uses the bayes theorem as classification basis. The formulation of bayes theorem is written as follows.

$$p(c_j|d) = \frac{p(d|c_j)p(c_j)}{p(d)} \qquad (4)$$

Where$p(c_j|d)$ is the conditional probability of occurrence of c_j when d, and vice versa $p(d \vee c_j), p(c_j)$ is the probability of occurrence of event c_j and $p(d)$ is the probability of occurrence of event d.

5. Evaluation Model Measure

Model Evaluation uses the testing data which is 20% of observation data. The evaluation measures used are accuracy value, precision value, recall value, and F1-Score. Accuracy is the ratio of correct predictions (positive and negative) to the overall data. The formula of accuracy measure is written as follows.

$$Accuracy = \frac{Truepositive + TrueNegative}{AllData} \qquad (5)$$

Precision is the ratio of positive correct predictions to the overall positive predicted outcome. The precision measure formula can be written as follows.

$$Precission = \frac{TruePositive}{TruePositive + FalsePositive} \qquad (6)$$

The next measure is recall value. Recall is the ratio of true positive predictions compared to the total number of true positive data. Recall value can be calculated with the following formula.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \qquad (7)$$

The last evaluation measure that used is F1-Score. The F1-score is a weighted comparison of the average precision and recall. F1-Score calculated by the formula below.

$$F1 = 2\frac{(Recall \times Precission)}{Recall + Precission} \qquad (8)$$

The best classification model is the model that has the highest accuracy, precision, recall, and F1-score value.

RESULTS

1. K-Means Discretization Results

The variables used in this study consisted of one categorical variable and five continuous variables.

Discretization using K-means is performed on continuous variables. Silhouette method used for selecting optimum number of k-means clusters. Based on silhouette method, the optimum number of clusters on five variables shows the same result which is 4 clusters. The results of clustering and profiling can be seen in Table 2.

Table 2. List of Variables

No	Variable	Type
1	Village potential status	categorical
2	Basic services	continuous
3	Infrastructure	continuous
4	Accessibility	continuous
5	Public services	continuous
6	Village government administration	continuous

2. Bayesian Network Learning Structure and Model

The first Learning Structure model is the result of human expert construction by considering the correlation coefficient value, and logical analysis based on the subjectivity of the researcher. The next learning structure model is using Naïve Bayes model.

Figure 5 describes the Naïve Bayes learning structure construction which shows that there is no causal relationships among independent variable. All independent variables directly influence the village development index status variable.

The next structure of the Bayesian network is built using the maximum spanning tree method. Spearman correlation coefficient must be calculated in order to build maximum spanning tree learning structure. The complete graph that weighted using Spearman correlation

Table 3. Discretization Result

Variable	Class	Variable	Class
Village potential status	Independent	Accessibility	Very Good
	Developed		Good
	Developing		Bad
	Least developed		Very Bad
	Undeveloped	Public services	Very Good
Basic services	Very Good		Good
	Good		Bad
	Bad		Very Bad
	Very Bad	Village government administration	Very Good
Infrastructure	Very Good		Good
	Good		Bad
	Bad		Very Bad
	Very Bad		

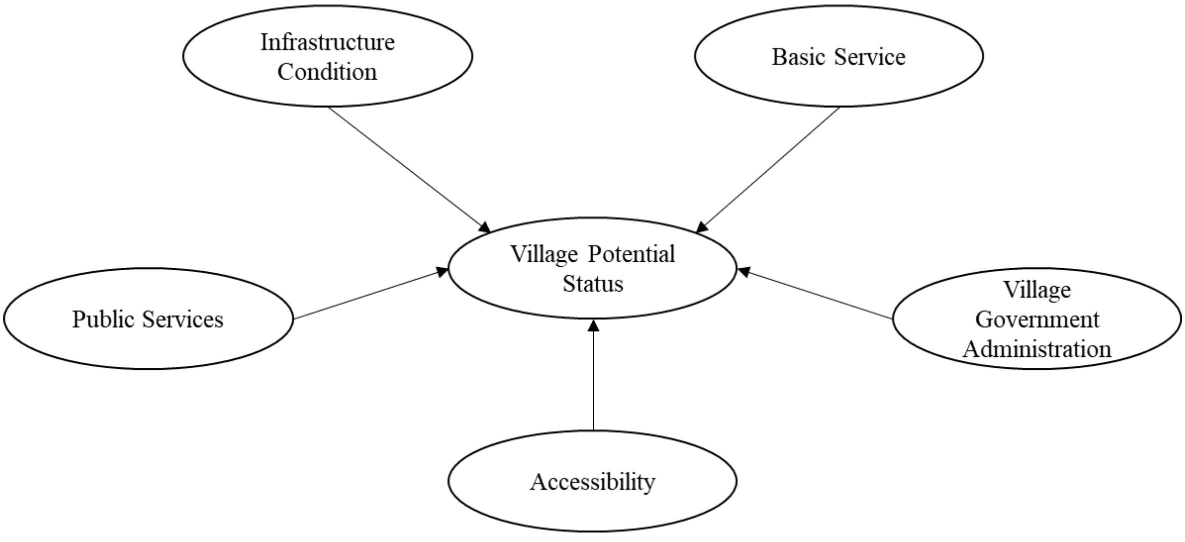
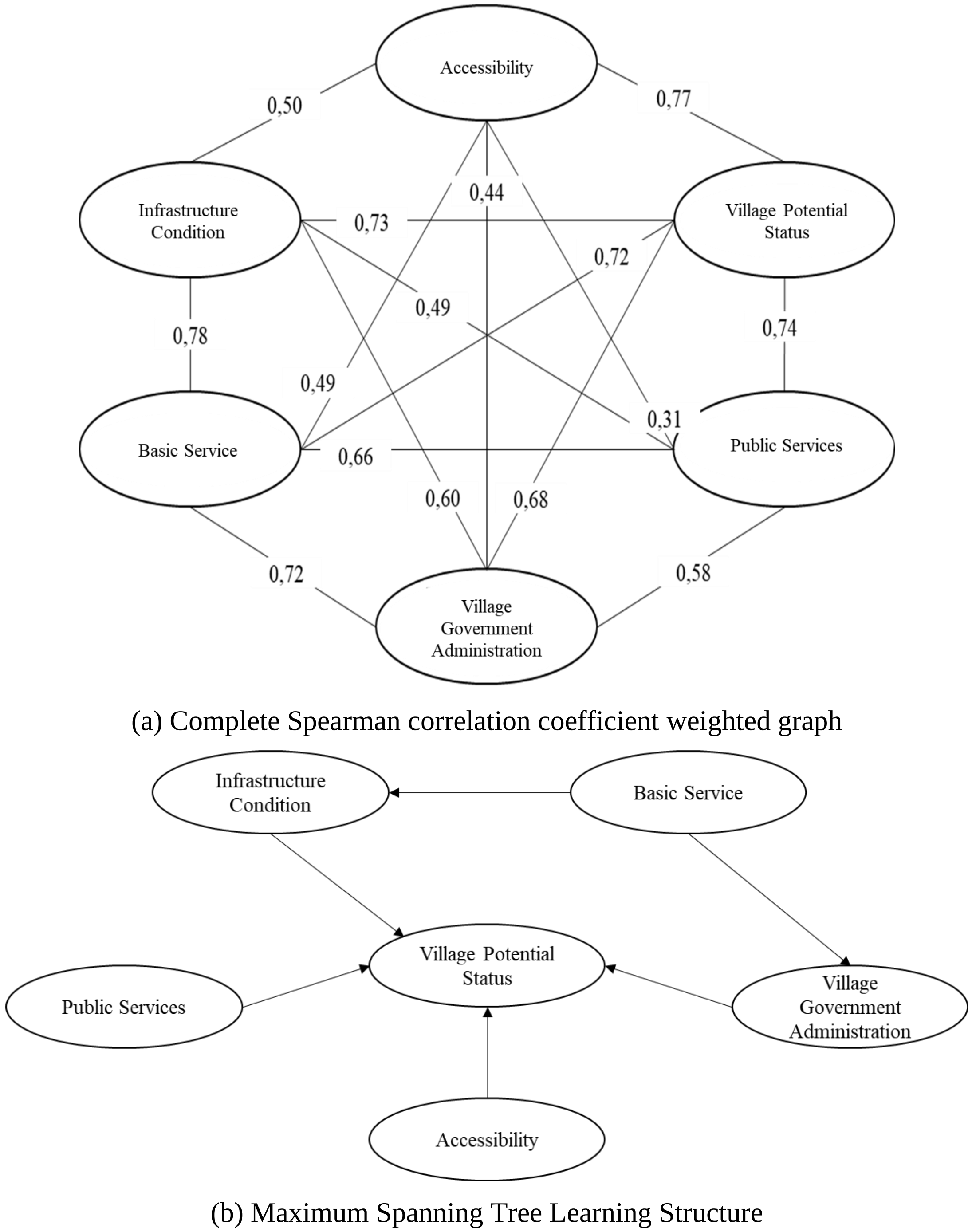


Figure 5. Learning Structure using Naïve Bayes construction

coefficient between variables depicted in Figure 6a.

From the graph in Figure 6a, maximum spanning tree is formed using Prim’s algorithm. The results of the maximum spanning tree construction can be seen in Figure 6b. Figure 6b shows that there are three variables that directly affected village potential status variables which includes public services, infrastructure condition, accessibility, and village government administration. The causal relationships among independent variables in this structure are basic services that influence infrastructure condition and village government administration variable.

The next learning structure method are Hill Climbing search and Tabu Search. These method known as automatic learning structure method. The construction of learning structure bayesian network using Hill Climbing and Tabu Search depicted in



(a) Complete Spearman correlation coefficient weighted graph

(b) Maximum Spanning Tree Learning Structure

Figure 6. Maximum Spanning Tree Learning Structure (a) Complete Spearman correlation coefficient weighted graph (b) Maximum Spanning Tree Learning Structure

Figure 7. From the graph in Figure 6a, maximum spanning tree is formed using Prim's algorithm. The results of the maximum spanning tree construction can be seen in Figure 6b. Figure 6b shows that there are three variables that directly affected village potential status variables which includes public services, infrastructure condition, accessibility, and village government administration. The causal relationships among independent variables in this structure are basic services that influence infrastructure condition and village government administration variable.

The next learning structure method are Hill Climbing search and Tabu Search. These method known as automatic learning structure method. The construction of learning structure bayesian network using Hill Climbing and Tabu Search depicted in Figure 7.

Figure 7a and Figure 7b shows the same formation of learning structure construction. Figure 7a visualize hill

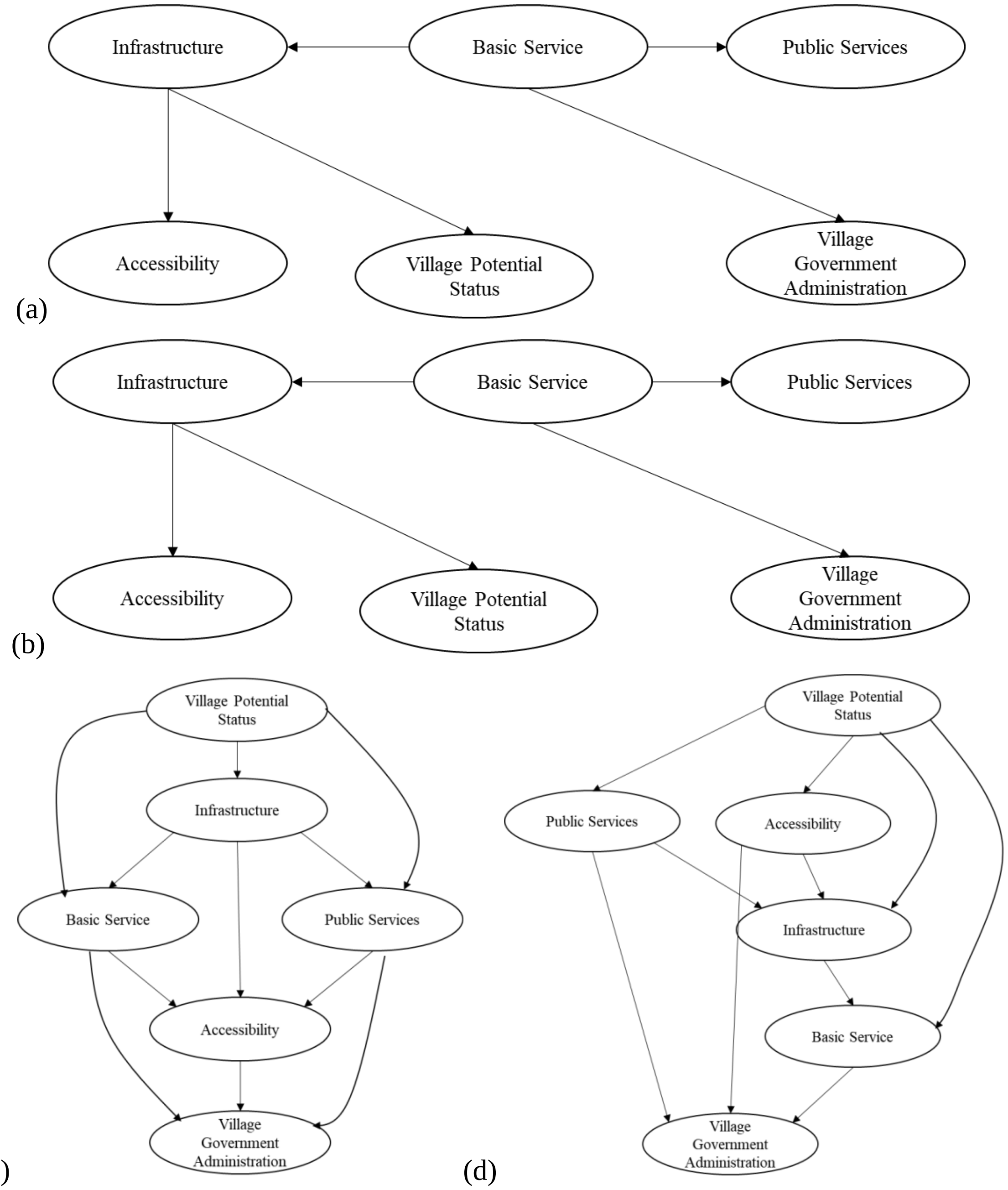


Figure 7. Automatic Learning Structure (a) Hill Climbing method with discretization, (b) Tabu Search method with discretization, (c) Hill Climbing search without discretization, (d) Tabu Search method without discretization.

climbing learning structure with k-means discretization data and Figure 7b visualize tabu learning structure with k-means discretization data. The DAG that illustrated in Figure 6a and 7b describe that village potential status variable is directly influenced by infrastructure condition. The independent variables have causal relationships with the other independent variables. Basic services variable affect the public services, village government administration and infrastructure condition variable. The infrastructure condition variable influence the accessibility variable.

Figure 7c shows directed acyclic graph (DAG) that describe hill climbing learning structure without k-means discretization method. Based on the DAG, the dependent variable which is village potential status is not affected by the independent variables. Otherwise, the village potential status variable directly affect some of independent variables. The same case is also found in Figure 7d that illustrate DAG from tabu learning structure without k-means discretization. The village potential status variable influenced several independent variable.

3. Model Evaluation

The following are the results of the model evaluation using the accuracy measure:

Based on Table 3, in general, the accuracy of classification models with k-means discretization data show better result then the classification models without k-means discretization. Maximum spanning tree model with discretization using kmeans method has the best accuracy value of 90,7%. The following is another evaluation value of the maximum spanning tree model with k-means discretization which is presented in Table 4.

The best model obtained based on the results of the model evaluation is the Bayesian network model with maximum spanning tree learning structure with Spearman correlation coefficient weighed using K-means discretization method. This model can be used for classifying village potential status of new data or records.

4. Classify New Data Using the Best Model

The best model obtained in this study used to predict new data at provincial level. A case study of village potential status prediction will be conducted in Banten Province. The value of variables for the village potential status prediction of Banten

Table 3. Accuracy value of the model (%)

Learning Structure	With K-Means Discretization	Without K-Means Discretization
Naïve Bayes	82,76	74,71
Maximum Spanning Tree	90,69	85,71
Hill Climbing Search	75,86	70,11
Tabu Search	77,38	67,81

Table 4. Evaluation measure of maximum spanning tree model (%)

Class	Precision	Recall	F1-Score
Independent	88.00	69	77.35
Developed	81.00	89	84.81
Developing	100.00	100	100.00
Least developed	76.00	83	79.34
Undeveloped	79.00	67	62.75

Province obtained through Statisitics Indonesia (BPS) data, is presented in Table 5.

Table 5. Classification Results on Banten Province Data

Variable	Value	Profile
Village potential status	65,55	Good
Basic services	51,09	Bad
Infrastructure	80,08	Very Good
Accessibility	58,26	Bad
Public services	72,26	Good

From Tabel 5, the conditional probability is calculated based on the Bayesian network model. It turns out that when the basic services of an area are in a good category, infrastructure is in a bad category, accessibility is in the very good category, public services are in a bad category, and the administration of the village government is in a good category, the village potential status of the area is “Developing” with 75.6% probability value. The results show Banten Province is a province with developing village potential status.

CONCLUSION

In this paper, we introduce the utilization of Bayesian network methods to classify the village potential status using the village development index score at regency/municipality to gain a better understanding of the causal relationships between independent variables of the village potential status. Bayesian network classification models can be used for predicting village potential status in Indonesia. The Bayesian network models built from the k-means discrete data have better performance accuracy than the models built from the data without discretization. The result of this research shows that the maximum spanning tree with the weighted Spearman coefficient correlation using the k-means discretization method has the best classification model with 90.69% accuracy.

The result of this study can be potentially beneficial to support the local government in monitoring their respective village potential status and focusing their work on exploring village potential in their region so that villages in their area can develop and become independent. Our future work includes the investigation of the use of different learning structure algorithms to produce models with better accuracy.

REFERENCES

[1] Badan Pusat Statistik Republik Indonesia (BPS RI). 2018. “Indeks Pembangunan Desa 2018”. (pp.1-149)

[2] Bappenas, “Narration: The National Medium-Term Development Plan for 2020-2024”, pp.1-320, 2020.

[3] L.A. Schintler, C.L. McNeely. “Encyclopedia of Big Data”, Springer International Publishing AG. 2017. DOI 10.1007/978-3-319-32001-4_483-1

[4] Effendy, D. A., Kusrini, K., & Sudarmawan, S. (2017). Algoritma K-Means untuk Diskretisasi Numerik Kontinyu Pada Klasifikasi Intrusion Detection System Menggunakan Naive Bayes. *E-Proceedings KNS&I STIKOM Bali*, 61-66.

[5] Purwadi, I. (2009). Penerapan bayesian network dalam penetapan daerah tertinggal [skripsi]. Bogor: Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor.

[6] Zhang, Z., Zhang, J., Wei, Z., Ren, H., Song, W., Pan, J., ... & Qiu, L. (2019). Application of tabu search-based Bayesian networks in exploring related factors of liver cirrhosis complicated with hepatic encephalopathy and disease identification. *Scientific reports*, *9*(1), 1-8. [5] Gámez, J. A., Mateo, J. L., & Puerta, J. M. (2011). Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Mining and*

Knowledge Discovery, *22*(1), 106-148.

[7] Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, *29*(2), 131-163.

[8] Ang, S. L., Ong, H. C., & Low, H. C. (2016). Classification Using the General Bayesian Network. *Pertanika Journal of Science & Technology*, *24*(1).

[9] De Blasi, R. A., Campagna, G., & Finazzi, S. (2021). A dynamic Bayesian network model for predicting organ failure associations without predefining outcomes. *Plos one*, *16*(4), e0250787.

[10] Jun-wu, L. I., Guo-ning, L. I., & Ding, Z. H. A. N. G. (2020). Application of CS-PSO algorithm in Bayesian network structure learning. *Journal of Measurement Science & Instrumentation*, *11*(1).

[11] Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands, 403-412.

[12] Khan, A., & Zubair, S. (2020, July). Expansion of Regularized Kmeans Discretization Machine Learning Approach in Prognosis of Dementia Progression. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[13] Sari, D. P., Rosadi, D., & Effendie, A. R. (2019). K-means and bayesian networks to determine building damage levels. *Telkomnika*, *17*(2).

[14] Sari, D. P., Rosadi, D., Effendie, A. R., & Danardono, D. (2021). Discretization methods for Bayesian networks in the case of the earthquake. *Bulletin of Electrical Engineering and Informatics*, *10*(1), 299-307.

[15] Maryono, D., Hatta, P., & Ariyuana, R. (2018, March). Implementation of numerical attribute discretization for outlier detection on mixed attribute dataset. In *2018 International Conference on Information and Communications Technology (ICOIACT)* (pp. 715-718). IEEE.

[16] Gerber, S., Pospisil, L., Navandar, M., & Horenko, I. (2020). Low-cost scalable discretization, prediction, and feature selection for complex systems. *Science advances*, *6*(5), eaaw0961.