Learning Bayesian Network for Rainfall Prediction Modeling in Urban Area using Remote Sensing Satellite Data (Case Study: Jakarta, Indonesia)

S R Putri[1], **A W 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

*Corresponding author's e-mail: 221810596@stis.ac.id, ariewahyu@stis.ac.id

Abstract. Rainfall modeling is one of the most critical factors in agricultural monitoring and statistics, transportation schedules, and urban flood prevention. Weather anomaly during the dry season in urban coastal areas of tropical countries such as Jakarta, Indonesia has become a challenging issue that causes unexpected changes in rain patterns. In this paper, we propose the Bayesian Network (BN) approach to model the probabilistic nature of rain patterns in urban areas and causal relationships among its predictor variables. Rain occurrences are predicted using temperature, relative humidity, mean-sea level (MSL) pressure, cloud cover, and precipitation variables. Data are obtained from the remote sensing sources of National Oceanic and Atmospheric Administration (NOAA) satellite in Jakarta 2020-2021. We compare both of the score-based, i.e., Hill Climbing (HC), and hybrid structure learning algorithms of Bayesian Network including the techniques of Max-Min Hill Climbing (MMHC), General 2-Phase Restricted Maximization (RSMAX2), and Hybrid-Hybrid Parents & Children (H2PC). Further, we also compare the performance of score-based model (Hill Climbing) under five different popular scorings: Bayesian Information Criterion (BIC), K2, Log-Likelihood, Bayesian Dirichlet Equivalent (BDE), and Akaike Information Criterion (AIC) methods. The main contributions of this study are as follows: (1) insights that the hybrid structure learning algorithms of Bayesian Network models are either superior in performance or at least comparable to its score-based counterparts (2) our proposed best performed Bayesian Network model that is able to predict the rain occurrences in Jakarta with a promising overall accuracy of more than 81 percent.

1. Introduction

Global warming that impacts climate change, has been considered as one of the most fundamental concerns in Sustainable Development Goals (SDGs), especially in Goal 13 which is climate action. It is undoubtedly regarded as the main factor behind the weather anomalies [1,2]. Urban coastal areas in tropical countries such as Jakarta, Indonesia are among the most affected areas by global warming due to the high intensity of mobilities and economic activity [1-3]. In the dry season of 2020, especially in Jakarta, there was a weather anomaly that caused changes in rain patterns such as frequent rains in dry season. The Meteorological, Climatological, and Geophysical Agency of Indonesia (BMKG) stated that the frequent occurrence of rain in the dry season is caused by the interaction of three factors, sea surface temperature anomalies, atmospheric waves, and unstable atmospheric conditions [1]. This change in rain patterns continues in 2021.With changes in rain patterns, it will be more difficult to

predict the occurrence of rain in a location, even though the place is equipped by a rain observation post or a complete climatology post [2]. In fact, rain prediction is very important to overcome flood disasters, especially in the Jakarta area which is often flooded.

The occurrence of rain can be influenced by several factors including temperature [4], humidity [4], mean sea level (MSL) pressure [5], and cloud cover [6]. The difficulty of predicting rain occurrences in a location due to weather anomalies causes the need for new weather models that can describe rain occurrences [7,8]. Data corresponding to these parameters can be obtained from processing remote sensing data [9,10]. Satellite imageries data has been an important data source to model various real-world applications [11,12]. Data for the entire surface of the earth can be provided by the output of NOAA (National Oceanic and Atmospheric Administration) satellites. NOAA satellites are used to obtain information about the physical state of the oceans or oceans and the atmosphere [13]. The study about finding the best parameter in remote sensing is required, [14] uses the analytical hierarchy process (AHP) to determine the level of importance for each parameter. And the feature selection is recommended to perform the classification of global and local climate zone [15-17].

In line with the era of disruption that requires updates to what exists, this study focuses on using an Artificial Intelligence (AI) approach to model rain occurrences in Jakarta. Artificial Intelligence (AI) is a part of computer science that studies how to make machines (computers) that can do jobs like and as well as humans do, maybe even better than that [18]. One of the Artificial Intelligence (AI) algorithms that have the potential to model rain predictions is the Bayesian Network, or also known as the Belief Network because each weather parameter tends to have a causal relationship with the other. Bayesian Network is a form of probabilistic graphical model (PGM) which is used to represent patterns in data. Bayesian Network models a problem using a direct acyclic graph (DAG), this model considers the causality relationship between the variables used [19]. The formation of the DAG structure can be done automatically using automatic learning with various algorithms including Hill Climbing, MMHC (Maximum-Minimum Hill Climbing), RSMAX2 (General 2-phase Restricted Maximization), and H2PC (Hybrid HPC). Bayesian Network can be constructed using discrete data, K-Means is one of the algorithms commonly used for data discretization.

Based on the existing background and potential, the purpose of this research is to build a Bayesian Network modeling of rain occurrences in Jakarta from 2020 to early 2021 through weather parameters based on remote sensing data taken from the NOAA (National Oceanic and Atmospheric Administration) satellite. The resulted model of this study is expected to provide an overview of the causal relationship among influenced variables causing the rain patterns in Jakarta and inferences that can be used as a basis for decision making. The data used for the construction of the model is events data from January 1, 2020 - April 21, 2021. The model built is then used to predict rain occurrences from April 22, 2021 - May 13, 2021. In addition, simulations were carried out to show the performance of the built model in predicting rain occurrences for each Jakarta city on 14 May 2021 (00.00-06.00 GMT+7). The benefit of this study is weather modeling which is built using the Bayesian Network approach by comparing the different network structures and network scoring parameters. The resulted model of this study is expected to provide an overview of the causal relationship among influenced variables causing the rain patterns in Jakarta and can make inferences that can be used as a basis for decision making. The main contribution of this study is the Bayesian Network approach for weather modelling which is constructed by selecting the best performed network structures and network scoring parameters.

2. Methods

The research carried out in building the prediction model of rain occurrences in Jakarta from 2020 to early 2021 including the data collection, data preprocessing, and Bayesian Network model construction with several computational algorithms. The research framework of this study is schematically illustrated in Figure 1.

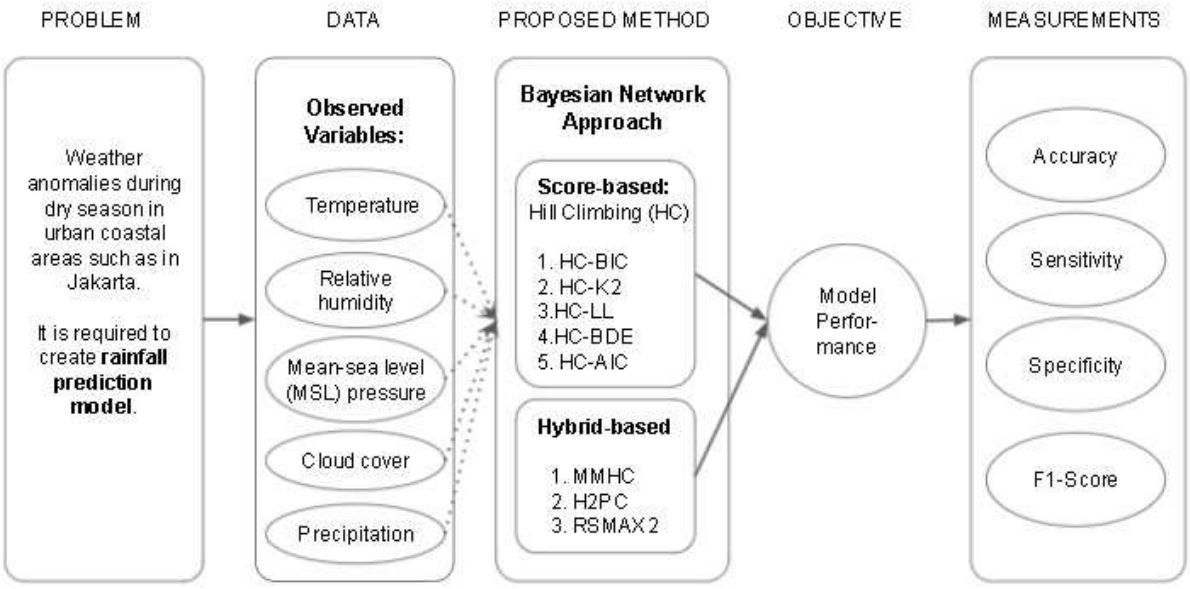


Figure 1. Research Framework

The proposed Bayesian Network learning approach is further investigated by comparing two different popular network structures: score-based algorithms, which is represented by Hill Climbing (HC) algorithm and hybrid-based learning algorithms which includes the MMHC (Maximum-Minimum Hill Climbing), RSMAX2 (2-phase Restricted Maximization), and H2PC (Hybrid HPC). The score-based Hill Climbing is compared based on its scoring criteria: Hill-Climbing Bayesian Information Criterion (HC-BIC), Hill Climbing K2 Method (HC-K2), Hill Climbing Log Likelihood (HC-LL), Hill Climbing Bayesian Dirichlet Equivalent (HC-BDE), and Hill Climbing Akaike Information Criterion (HC-AIC). The best model is the selected by evaluating the model performance in terms of the accuracy, sensitivity, specificity, and F1-score measurements.

2.1. Data Collection

The data used in this study is weather parameter data obtained from the NOAA (National Oceanic and Atmospheric Administration) satellites [22]. Data is retrieved via Google Earth Engine (GEE), a cloud-based platform designed to store and process earth data. The data taken is cumulative data for every 6 hours of observation from January 1st, 2020 to May 13th, 2021. Data from January 1st, 2020 to April 21st, 2021 is used for the construction of the model. The model built is then used to predict rain occurrences on April 22nd, 2021 to May 13th, 2021. Weather parameters that are used are available in the NOAA CFSR (Climate Forecast System Reanalysis) documentation, such as temperature, relative humidity, MSL pressure, cloud cover, and precipitation. Temperature is a measure of how hot or cold something [20]. Relative humidity is defined as the percentage ratio between partial water vapor pressure and saturated water vapor pressure, humidity itself is a level of wet air environmental conditions caused by water vapor [21]. MSL pressure is the mean pressure at sea level (MSL) in the International Standard Atmosphere (ISA) is 1013.25 hPa, or 1 atmosphere (atm), or 29.92 inches of mercury [22]. Cloud cover indicates the cloud mass that covers the area. Precipitation is the outpouring or falling of water from the atmosphere to the earth's surface in a different form namely rainfall [23].

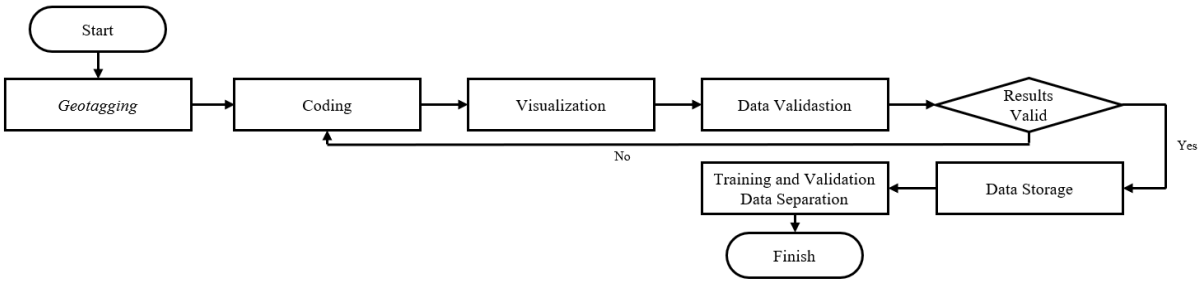


Figure 2. Data Collection Flow Chart

Figure 2 shows the flow chart of data collection. Geotagging is carried out to mark an area that becomes scope of the research, Jakarta. Geotagging is done on Google Earth Engine maps. Figure 3 shows the results of geotagging performed.

Figure 3. Geotagging Result

The coding stage is the stage of writing a program script to retrieve weather parameter data according to the corresponding band on the satellite. The code is written using the JavaScript programming language. Table 1 shows the data that is taken and the corresponding bands.

Table 1. Data Taken from NOAA Satellite

Parameter	Measure	*Band*
Temperature	K	Temperature_surface
Relative Humidity	%	Relative_humidity_entire_atmosphere_single_layer
MSL Pressure	Pa	Pressure_reduced_to_MSL_msl
Cloud Cover	kg/m^2	Cloud_water_entire_atmosphere_single_layer
Precipitation	kg/m^2	Total_precipitation_surface_3_Hour_Accumulation

Figure 4 shows an example of visualization of data obtained from temperature parameters in Jakarta from 2020-early 2021 measured in Kelvin (K).

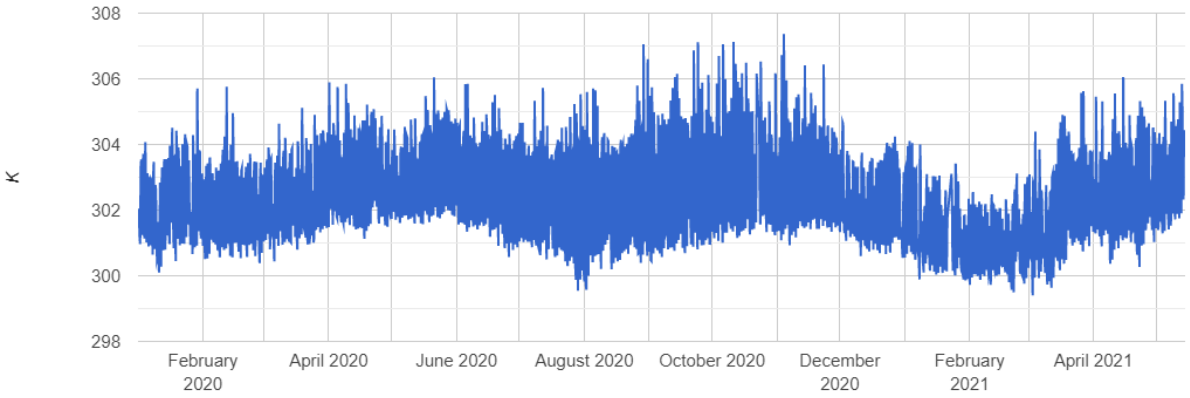


Figure 4. Temperature Data Visualization of Jakarta in 2020-early 2021

Data that has been taken is then validated to ensure that it is matched with expectations. If the data taken is not as good as what we expected, then return to the coding stage. The data is then saved in the comma separated value (csv) format. This study obtained a dataset with 1985 lines. The model was built using the first 1901 data lines (January 1, 2020 – April 22, 2021) while the remaining 84 (April 23, 2021 – May 14, 2021) will be used for validation purposes.

2.2. Data Preprocessing

Preprocessing is a very important stage in the construction of the Bayesian Network model because the model built will depend on the input data. At this stage, unit conversion, coding, and data integration are carried out. Unit conversion is carried out on temperature parameter by changing the unit to Celsius. Coding is done to change the previously continuous scale data to be discrete. The coding of precipitation variable is done by encoding the value 0 as no rain and the > 0 value as rain. The coding of the variables of temperature, relative humidity, MSL pressure, and cloud cover was performed using K-Means clustering algorithm. K-Means is an algorithm that assigns an object into a cluster that has the closest centroid [24]. The following are the steps for forming a cluster in the K-Means algorithm:

a. Divide the data into k groups, the selection of k is determined by the Elbow Method which provides information regarding the goodness of the number of clusters based on the sum of square error (SSE) value.
b. Calculate the mean in each group as a centroid.
c. Grouping each data into the nearest centroid, the distance calculation is carried out with the Euclidean distance function as follows

$$d(s,t) = \sqrt{(s_x - t_x)(s_y - t_y)} \quad (1)$$

d. Recalculate the average of each group formed into a new centroid.
e. Steps b to d are repeated so that the new group formed is the same as the previous one or is stable.
f. Conduct profiling to interpret the formed groupings.

The last process, data integration, is carried out to collect all data into one dataset.

2.3. Bayesian Network Model Construction

Bayesian Network is a form of probabilistic graphical model (PGM) which represents a causal probabilistic relationship between a set of random variables, conditional dependencies, and provides a complete representation of the joint probability distribution. Bayesian Network consists of two main parts, namely a directed acyclic graph (DAG) and a set of conditional probability distributions. If there is a probabilistic causal dependency between two random variables in the graph, the corresponding nodes will be connected by a directed edge. Since a directed arc represents a static, causal probabilistic

dependency relationship, cyclicality is not allowed in the graph. The conditional probability distribution is defined for each node in the graph [21].

The advantage of the Bayesian Network algorithm is that it can be used in the construction of predictive and descriptive models. As a predictive tool, this model provides an efficient tool for solving various inferential problems including posterior probability, abductive or diagnostic reasoning, relevance analysis, and classification. In the case of description, this model can describe the dependency relationship between random variables and construct the modeled problem domain [19]. Within a wise decision framework developed, Bayesian analysis offers reasonable and coherent way of mixing prior and posterior data information [25-27]. Bayesian Network is built based of Bayesian Statistics. Given data x and parameter θ, a simple Bayesian analysis starts with a prior probability $p(\theta)$ and likelihood $p(x|\theta)$ to compute a posterior probability $p(\theta|x) \propto p(x|\theta)p(\theta)$. Therefore, we perform the Bayesian Network which represents a causal probabilistic relationship between a set of random variables, conditional dependencies, and provides a complete representation of the joint probability distribution.

Bayesian Network model development can be done in two ways, manual construction and automatic learning. Manual construction is done by identifying the relevant nodes (variables) and the structural dependencies between them, this construction requires knowledge of the expert. Automatic learning is done by building a Bayesian Network structure with an algorithm that is applied to a dataset [21]. In this study, the construction of the Bayesian Network structure was carried out by automatic learning applying several kinds of algorithms, Hill Climbing, MMHC (Maximum-Minimum Hill Climbing), RSMAX2 (General 2-phase Restricted Maximization), and H2PC (Hybrid HPC). Hill Climbing is a local search algorithm that explores the search space by starting from the initial solution and performing a series of steps until a solution is found to maximize the value of f. In Bayesian Network construction, the steps taken include adding arcs, removing arcs, and reversing arcs on a directed acyclic graph (DAG) [28]. Several scoring or assessments that can be implemented in the Hill Climbing (HC) algorithm include BIC, K2, Log Likelihood, BDE, and AIC. Maximum-Minimum Hill Climbing (MMHC) was first proposed by Tsamardinos et al. in 2006. This algorithm is a hybrid method that combines ideas from local learning, constraint-based, and search-and-score techniques. MMHC initially studied the Bayesian Network framework with the Max-Min Parents and Children (MMPC) algorithm and then adjusted its skeletal construction by implementing the Hill Climbing algorithm [29]. General 2-phase Restricted Maximization (RSMAX2) is a general implementation of Maximum-Minimum Hill Climbing that uses a combination of constraint-based and score-based algorithms [30]. H2PC (Hybrid HPC) is a hybrid algorithm that is intended for Bayesian Network construction. H2PC constructs a Bayesian Network frame and then performs a Bayesian-scoring greedy hill climbing search to search for edge adjustments [30]. The Bayesian Network construction process is presented in Figure 5.

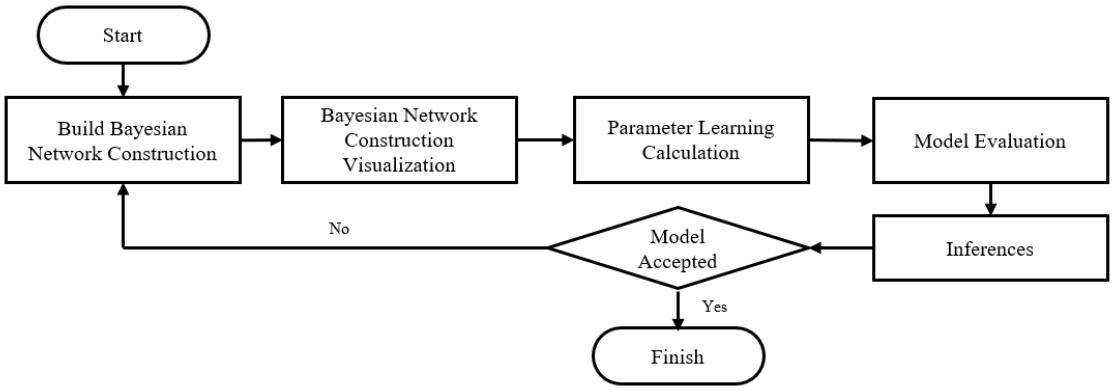


Figure 5. Bayesian Network Construction

Modeling begins by applying each Bayesian Network construction algorithm with automatic learning including Hill Climbing, MMHC (Maximum-Minimum Hill Climbing), RSMAX2 (General 2-phase Restricted Maximization), and H2PC (Hybrid HPC). After obtaining scores and dependencies for each node and the form of their connection, the results are then visualized in the form of a directed acyclic graph (DAG). The parameter is then calculated to get the conditional probability table (CPT) value, which is a value that represents the probability value of an event with the condition that other events occur. In this study, the calculation of the learning parameter value is carried out using Bayesian Estimation. The next stage is inferencing to see how the model built makes predictions. The entire construction phase of the Bayesian Network is carried out with the bnlearn library on R.

2.4. Model Validation Evaluation

Model validation checking is carried out to see how valid the built model. This evaluation is performed by predicting the test data. Then, we match the predicted results with the actual results.

2.5. Selection of The Best Model

Based on the validation tests that have been carried out previously, the best model is selected by taking into account the performance measurements of accuracy, sensitivity, specificity, and F1-score.

2.6. Analysis of Results

Based on the best selected model, an analysis was carried out to obtain a descriptive picture of the obtained model.

3. Results

3.1. Data Discretization

K-Means discretization was applied to temperature, relative humidity, MSL pressure, and cloud cover variables. Discretization is done by categorizing the data based on the closest distance to each centroid. The centroid is calculated from the mean of grouped data. The process will be repeated and stopped after the formed categories are stable. Manual discretization is applied to the precipitation variable by coding the value 0 as the occurrence of no rain and a value > 0 as the occurrence of rain. Table 2 shows the results of the discretization obtained.

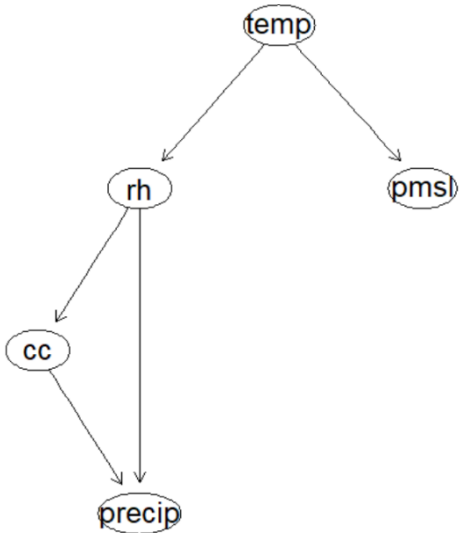
Table 2. Discretization Result

Variable	Discretization
Temperature	Low
	High
Relative Humidity	Low
	Moderate
	High
Mean Sea Level (MSL) Pressure	Low
	Moderate
	High
Cloud cover	Very Low
	Low
	High
	Very High
Precipitation	Raining
	Not Raining

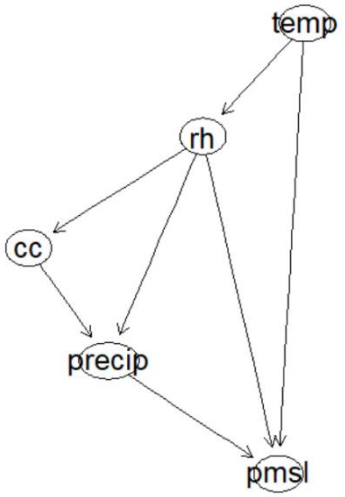
3.2. Bayesian Network Construction

In this study, the construction of the Bayesian Network structure was carried out by automatic learning applying several kinds of algorithms, such as Hill Climbing (HC), MMHC (Maximum-Minimum Hill Climbing), RSMAX2 (General 2-phase Restricted Maximization), and H2PC (Hybrid HPC). Figure 6

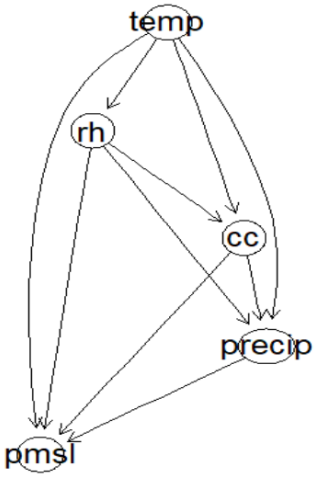
shows the directed acyclic graphs (DAGs) that illustrates the modeling results obtained by implementing these algorithms using the bnlearn library in R.



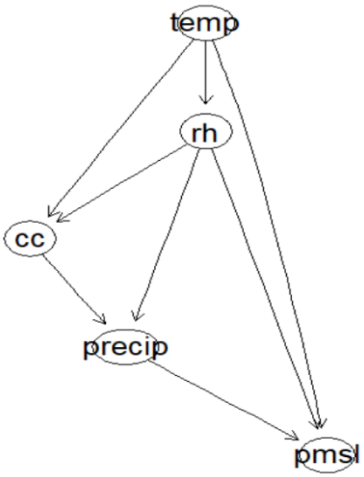
a) Using HC-BIC, HC-BDE, MMHC, RSMAX2, H2PC Algorithms



b) Using HC- K2 Algorithm



c) Using HC-LL Algorithm



d) Using HC–AIC Algorithm

Figure 6. The comparison of constructed Directed Acyclic Graph (DAG) of the Bayesian Network model for rain occurrence prediction

Figure 6a shows the directed acyclic graph (DAG) that illustrates the modeling result obtained by Hill Climbing with BIC and BDE scoring, MMHC, RSMAX2, H2PC algorithms. Based on those DAG, it can be seen that the variables that directly affect the precipitation variable *(precip)* are cloud cover *(ca)* and relative humidity *(rh)*. The cloud cover variable *(cc)* is influenced by the relative humidity variable *(rh)*. The relative humidity variable itself is influenced by the temperature variable *(temp)*. The only variable that does not affect the precipitation variable is the MSL pressure variable *(pmsl)*. The MSL pressure variable is influenced by temperature variable *(temp)*.

Figure 6b shows the directed acyclic graph (DAG) that illustrates the modeling result obtained by Hill Climbing with K2 scoring algorithm. Based on those DAG, it can be seen that variables that directly affect the precipitation variable *(precip)* are cloud cover *(cc)* and relative humidity *(rh)*. The

cloud cover variable *(cc)* is influenced by relative humidity variable *(rh)*. The relative humidity variable *(rh)* itself is influenced by temperature variable *(temp)*. The only variable that does not affect the precipitation variable is MSL pressure variable *(pmsl)*. The MSL pressure variable *(pmsl)* is directly affected by temperature *(temp)*, relative humidity *(rh)*, and precipitation *(precip)* variables.

Figure 6c shows the directed acyclic graph (DAG) that illustrates the modeling result obtained by Hill Climbing with Log Likelihood scoring algorithm. Based on those DAG, it can be seen that the variables that directly affect the precipitation variable *(precip)* are temperature *(temp)*, cloud cover *(cc)* and relative humidity *(rh)* variables. The cloud cover variable *(cc)* is influenced by relative humidity *(rh)* and temperature *(temp)* variables. The relative humidity variable itself is influenced by temperature variable *(temp)*. The only variable that does not affect the precipitation variable is the MSL pressure variable *(pmsl)*. The MSL pressure variable *(pmsl)* is directly affected by *(temp)*, relative humidity *(rh)*, and precipitation *(precip)* variables.

Figure 6d shows the directed acyclic graph (DAG) that illustrates the modeling result obtained by Hill Climbing with AIC scoring algorithm. Based on those DAG, it can be seen that the variables that directly affect the precipitation variable *(precip)* are cloud cover *(cc)* and relative humidity *(rh)* variables. The cloud cover variable *(cc)* is influenced by relative humidity *(rh)* and temperature *(temp)* variables. The relative humidity variable itself is influenced by the temperature variable *(temp)*. The only variable that does not affect the precipitation variable is the MSL pressure variable *(pmsl)*. The MSL pressure variable *(pmsl)* is directly affected by temperature *(temp)*, relative humidity *(rh)*, and precipitation *(precip)* variables.

Based on the described modeling result, similar results were obtained, the precipitation variable *(precip)* is influenced by the temperature variable *(temp)*, relative humidity variable *(rh)*, and cloud cover variable *(cc)* either directly or indirectly. This shows that to predict rain occurrences based on those algorithms, it can be done using the values of temperature, relative humidity, and cloud cover.

3.3. Bayesian Network Models Evaluation

The results of Bayesian Network construction with automatic learning for rain modeling at the previous point are then evaluated. The evaluation was carried out by looking at several measures of the goodness of the model such as accuracy, sensitivity, and specificity, and F1-score with the occurrence of rain as a positive class. Evaluation is used to predict rain on 84 data that are not used in model construction, weather parameter data on April 22nd, 2021 – May 13th, 2021. The results of the evaluation can be seen in Table 3.

Table 3. Model Evaluation Result

Model	Accuracy	Sensitivity	Specificity	F1-Score
Model 1 (HC-BIC, HC-BDE, MMHC, RSMAX2, H2PC)	**81.18%**	**78.26%**	**84.62%**	**81.81%**
Model 2 (HC-K2)	72.94%	58.97%	84.78%	77.23%
Model 3 (HC-LL)	72.94%	53.85%	89.13%	78.10%
Model 4 (HC-AIC)	72.94%	58.97%	84.78%	77.23%

Based on Table 3, it can be seen that the best accuracy value was obtained using Model 1, other measures, sensitivity, specificity, and F1-Score also received quite high results. Models 2, 3, and 4 have very low sensitivity values, so it can be concluded that these three models are not good at predicting negative class. Therefore, the best model chosen in the case of rain prediction in Jakarta is Model 1.

3.4. Bayesian Network Best Model Parameter Analysis

Based on the best-selected model, the learning parameters are calculated using Bayesian Estimation. The calculation is done using training data to get the CPT (Conditional Probability Table) value for

each node. In this discussion, CPT analysis is presented for the precipitation *(precip)* and relative humidity *(rh)* nodes.

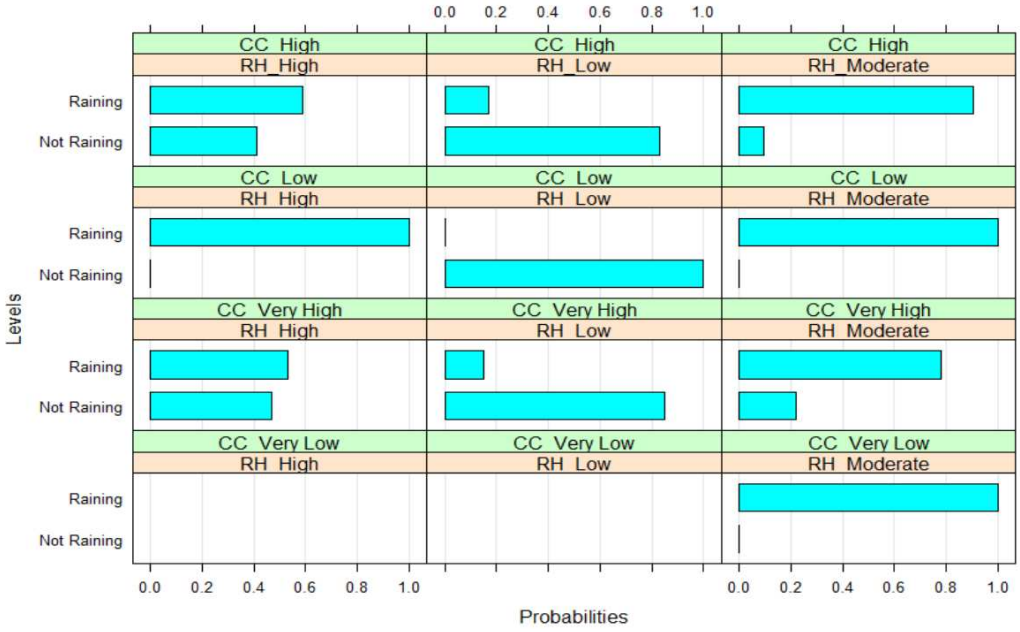


Figure 7. Conditional Probabilities of Precipitation Variable towards the Relative Humidity *(rh)* and Cloud Cover *(cc)*

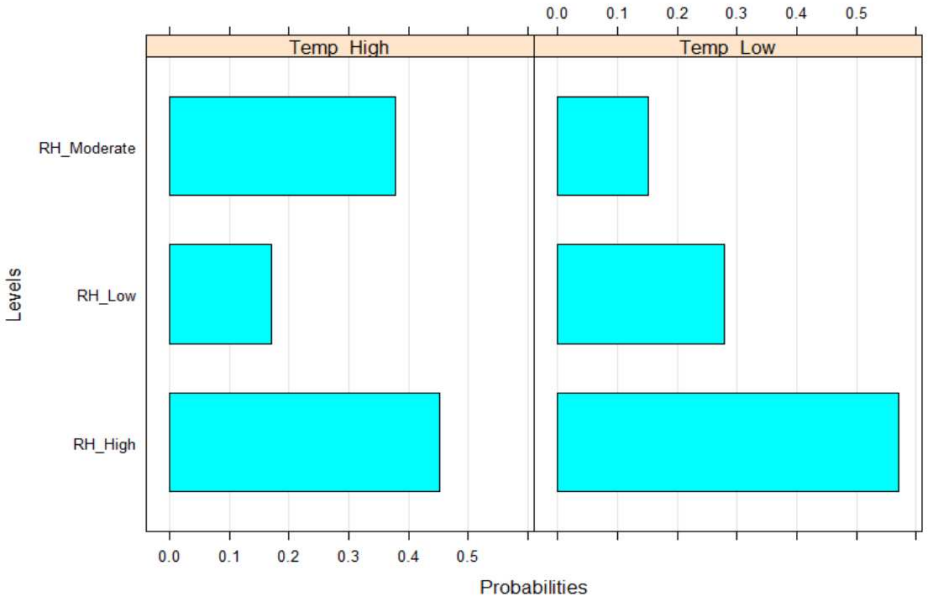


Figure 8. Conditional Probabilities of the relative humidity (rh) variable towards temperature (temp)

Figure 7 shows the conditional probabilities for the precipitation node. It can be seen that in conditions of very low cloud cover *(cc)*, but moderate relative humidity, the probability of rain occurrence is high. In conditions of low cloud cover *(cc)*, a high probability of rain occurs if the relative humidity is moderate and high, a high probability of not raining is high when the relative humidity is low. In conditions of high cloud cover *(cc)*, the probability of rain falling is high when the relative humidity *(rh)* is moderate or high, the probability of not raining is high when the relative humidity is when the relative humidity is low. In very high cloud cover *(cc)* conditions, the probability of rain occurrence is high when the relative humidity *(rh)* is moderate or high.

On the other hand, Figure 8 describes the conditional probabilities for the relative humidity *(rh)* node. It can be seen that at low temperatures, the probability of occurrence of high relative humidity is quite high compared to the probability of occurrence of moderate and low relative humidity. At high temperatures, the probability of occurrence of moderate and high relative humidity tends to be higher than the probability of occurrence of low relative humidity.

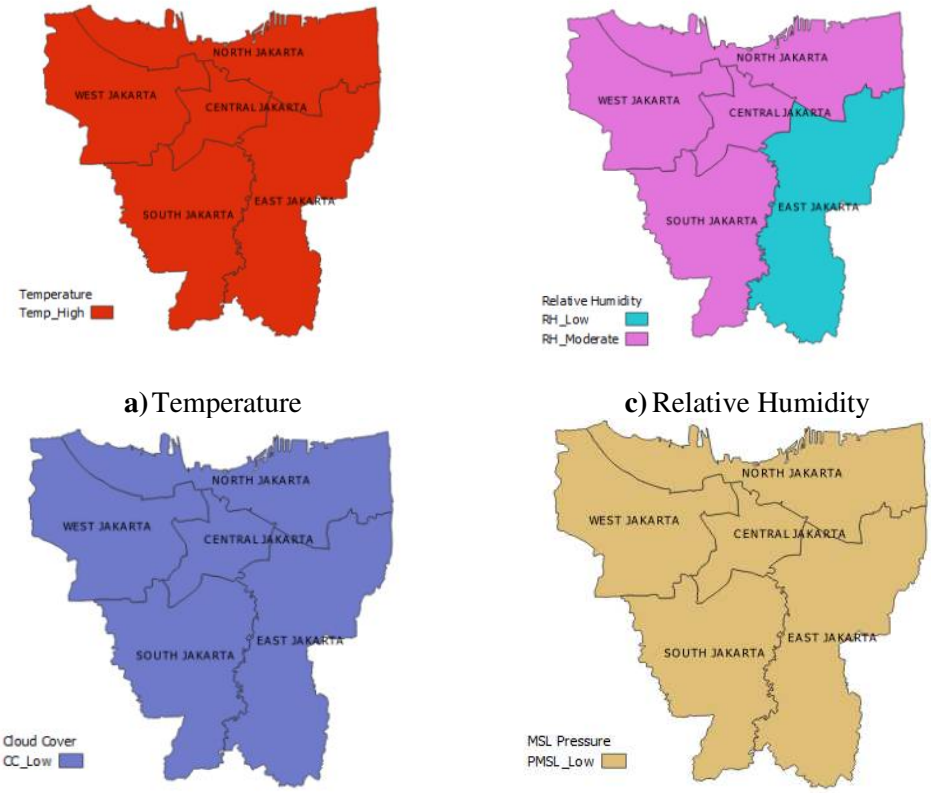
3.5. Inferences

Based on the construction of the Bayesian Network model in the form of a directed acyclic graph (DAG) and conditional probabilities, several inferences can be made with the following examples:

- Probability of rain in Jakarta when the temperature is high is 57.99%
- Probability of rain in Jakarta when the temperature is high and the relative humidity is low is 16.27%
- Probability of rain in Jakarta when the temperature is high and cloud cover is high is 81.38%
- Probability of rain in Jakarta when the temperature is low and the relative humidity is low is 14.26%
- Probability of rain in Jakarta when the temperature is low and the relative humidity is medium is 82.21%
- Probability of rain in Jakarta when the temperature is high, cloud cover is high, and relative humidity is high is 60.34%

3.6. Simulation

The simulation was carried out to see how the best model applied in predicting rain in Jakarta at a certain time. Here, we predict rain occurrences on May 14, 2021, from 00.00 to 06.00 GMT+7 for each area of Jakarta. Figure 9 shows the visualization of input data.

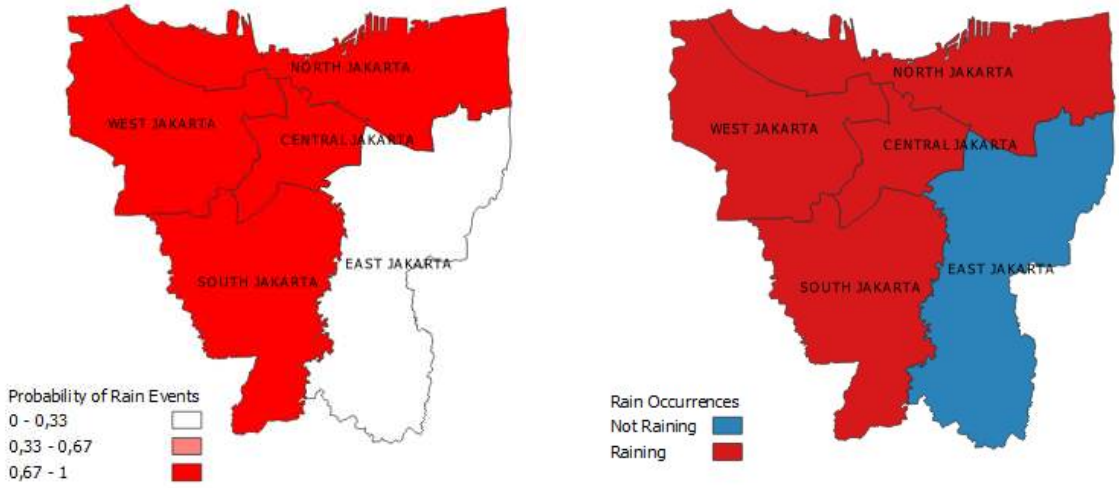


a) Temperature **c)** Relative Humidity

d) Cloud Cover **f)** MSL Pressure

Figure 9. Simulation Input

In Figure 9, we can see that the entire city of Jakarta is classified as high temperature (Temp_High), low cloud cover (CC_Low), and low MSL pressure (PMSL_Low). East Jakarta is classified as low relative humidity (RH_Low) while the rest is classified as moderate relative humidity (RH_Moderate). To get the probability value of rain events that time, inference has done based on the best model built with the previous data. Figure 10 shows the result of this simulation.



a) Probability of Rain Events **c)** Rain Occurences

Figure 10. Simulation Result

Figure 10 shows the comparison between the probability of rain occurrences and rain events that actually occurs. We can see that the probability of rain events in East Jakarta is very low. In fact, there is no rain at those area. On the other hand, the probability of rain events in North Jakarta, Central Jakarta, West Jakarta, and South Jakarta is high. In fact, rain falls at those area.

4. Conclusions

Based on the Bayesian Network modeling that has been constructed in this study, it can be concluded that the construction of the Bayesian Network model with automatic learning is influenced by not only the learning algorithm but also the selection of scoring criterion. It is found that the hybrid structure learning algorithms of Bayesian Network models are either superior in performance or at least comparable to their score-based counterparts. We predict the rain occurrence in Jakarta based on atmospheric parameter data obtained from the NOAA satellite (temperature, relative humidity, precipitation, cloud cover, and mean sea level pressure). The best performed models are the model built using the Hill Climbing with BIC and BDE scoring, MMHC, RSMAX2, and H2PC algorithms with overall accuracy on predicting future rain incidents is 81.18%. The represented directed acyclic graph (DAG) structure of those models reveals that the occurrence of rain in the Jakarta area are directly affected by the relative humidity and cloud cover variables. The temperature variable indirectly affects the occurrence of rain. The mean sea level (MSL) pressure variable does not affect the occurrence of rain either directly or indirectly. The pattern of rain occurrences according to weather parameter data from the NOAA satellite (temperature, relative humidity, cloud cover, and MSL pressure) is found so that inference can be useful as a basis for decision making.

Three main suggestions can be inferred from the results of this study. Firstly, this study still only uses weather parameter data in the Jakarta area taken from the NOAA satellite so that other research is needed using other parameters. Second, more detailed research is needed to classify rain occurrences according to their quantity (heavy, light, or otherwise) so that more accurate decisions can be made. Lastly, we suggest to use the higher resolution satellite for better estimations.

References

[1] S. Lestari, A. King, C. Vincent, D. Karoly, A. Protat, "Seasonal dependence of rainfall extremes in and around Jakarta, Indonesia", Weather and Climate Extremes, Vol. 24, 100202, 2019

[2] K. Kataoka, F. Matsumoto, T. Ichinose, M. Taniguchi, "Urban warming trends in several large Asian cities over the last 100 years", Science of The Total Environment, Volume 407, Issue 9, Pages 3112-3119, 2009

[3] S. Siswanto, G. J. van Oldenborgh, G. van der Schrier, R. Jilderda, B. van den Hurk, "Temperature, extreme precipitation, and diurnal rainfall changes in the urbanized Jakarta city during the past 130 years", International Journal of Climatology, 36: 3207-3225, 2016

[4] D. Prakoso, "Analisis Pengaruh Tekanan Udara, Kelembaban Udara, dan Suhu Udara Terhadap Tingkat Curah Hujan di Kota Semarang," 2018, [Online]. Available: http://lib.unnes.ac.id/36742/1/4112314008.pdf.

[5] M. Karimini, S. P. Nugroho, S. Tikno, S. Nuryanto, B. P. Sitorus, and S. Bahri, "Aplikasi Teknologi Modifikasi Cuaca Untuk Meningkatkan Curah Hujan di Das Citarum - Jawa Barat 12 Maret S.D. 10 April 2001," J. Sains Teknol. Modif. Cuaca, vol. 2, pp. 1–9, 2009.

[6] S. Mujiasih, "Pemanfatan Data Mining Untuk Prakiraan Cuaca," J. Meteorol. dan Geofis., vol. 12, pp. 189–195, 2011.

[7] A. Sharma and M. K. Goyal, "Bayesian network model for monthly rainfall forecast," 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2015, pp. 241-246

[8] A. Sharma and M. K. Goyal, "Bayesian network for monthly rainfall forecast: a comparison of K2 and MCMC algorithm", International Journal of Computers and Applications, 38:4, 199-206, 2016

[9] D. W.Triscowati, B. Sartono, A. Kurnia, D. D. Domiri, A. W. Wijayanto, "Multitemporal remote sensing data for classification of food crops plant phase using supervised random forest", in Sixth Geoinformation Science Symposium, 11311, 1131102, 2019

[10] A. W. Wijayanto, D. W. Triscowati, A. H. Marsuhandi, "Maize Field Area Detection in East Java, Indonesia: An Integrated Multispectral Remote Sensing and Machine Learning Approach", 2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE), 2020

[11] Y. Nurmasari and A. W. Wijayanto, "Oil Palm Plantation Detection in Indonesia using Sentinel-2 and Landsat-8 Optical Satellite Imagery (Case Study: Rokan Hulu Regency, Riau Province)", International Journal of Remote Sensing and Earth Sciences (IJReSES), vol. 18, no. 1, 2021, LAPAN.

[12] T. D. T. Saadi and A. W. Wijayanto, "Machine Learning Applied to Sentinel-2 and Landsat-8 Multispectral and Medium-Resolution Satellite Imagery for the Detection of Rice Production Areas in Nganjuk, East Java, Indonesia", International Journal of Remote Sensing and Earth Sciences (IJReSES), vol. 18, no. 1, 2021, LAPAN.

[13] NOAA, "CFSR: Climate Forecast System Reanalysis," 2021. [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/NOAA_CFSR#bands.

[14] L. N. Syahid et al., "Determining Optimal Location for Mangrove Planting Using Remote Sensing and Climate Model Projection in Southeast Asia," Remote Sens., vol. 12, no. 22, pp. 1–29, 2020.

[15] J. Hu, P. Ghamisi, and X. Zhu, "Feature Extraction and Selection of Sentinel-1 Dual-Pol Data for Global-Scale Local Climate Zone Classification," ISPRS Int. J. Geo-Information, 2018, doi: 10.3390/ijgi7090379.

[16] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," IEEE Trans. Geosci. Remote Sens., vol. 48, no. 5, pp. 2297-2307., 2010, doi: 10.1109/TGRS.2009.2039484.

[17] M. Belgiu and L. Dr?gu, "Random forest in remote sensing: A review of applications and future directions," ISPRS J. Photogramm. Remote Sens., vol. 114, pp. 24–31, 2016, doi: 10.1016/j.isprsjprs.2016.01.011.

[18] M. Dahria, "Kecerdasan Buatan (Artificial Intelligence)," SAINTKOM, vol. 5, pp. 185–196,

2008, [Online].
[19] M. Horný, “Bayesian Networks,” Tec. Rep. Bost. Univ., vol. 5, p. 8, 2014.
[20] M. Adrinta, M. Ihsan, A. Syahputra, and R. I. Ghani, “Alat Ukur Suhu Udara Digital Berbasis Atmega 32,” Univ. Sumatera Utara, 2017.
[21] S. Indarati, S. M. B. Respati, and Darmanto, “Kebutuhan Daya Pada Air Conditioner Saat Terjadi Perbedaan Suhu dan Kelembaban,” Momentum, vol. 15, pp. 91–95, 2019.
[22] C. Sanders, “Atmospheric Pressure and Altimeters,” Front, vol. 1, pp. 1–6, 2002.
[23] C. Asdak, Hidrologi dan Pengeolaan Daerah Aliran Sungai. Yogyakarta: Gadjah Mada University Press, 1995.
[24] A. F. Febrianti, A. H. Cabral, and G. Anuraga, “K-Means Clustering dengan Metode Elbow Untuk Pengelompokan Kabupaten dan Kota di Jawa Timur Berdasarkan Indikator Kemiskinan,” SNHRP, 2018.
[25] H. Jamil and W. Bergsma, “Bayesian variable selection using I-priors,” 2020. doi: 10.1007/978-3-030-42553-1_5.
[26] I. G. N. M. Jaya, B. Tantular, and Y. Andriyana, “A Bayesian approach on multicollinearity problem with an Informative Prior,” in Journal of Physics: Conference Series, 2019, p. 012021, doi: 10.1088/1742-6596/1265/1/012021.
[27] R. E. Caraka, N. T. Nugroho, S.-K. Tai, R. C. Chen, T. Toharudin, and B. Pardamean, “Feature Importance of The Aortic Anatomy on Endovascular Aneurysm Repair (EVAR) using Boruta and Bayesian MCMC,” Commun. Math. Biol. Neurosci., vol. 2020, no. 1–23, 2020, doi: https://doi.org/10.28919/cmbn/4584.
[28] J. A. Gamez and J. L. Mateo, “Learning Bayesian Networks By Hill Climbing: Efficient Methods Based on Progressive Restriction of The Neighborhood,” Data Min. Knowl. Discov. Jose M. Puerta, vol. 22, no. January 2011, 2011, doi: 10.1007/s10618-010-0178-6.
[29] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, “The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm,” Mach Learn, vol. 65, pp. 31–78, 2006.
[30] M. Scutari, “Package ‘bnlearn,’” 2013.