Naive Bayes Classifier (NBC) for Forecasting Rainfall in Banyuwangi District Using Projection Pursuit Regression (PPR) Method

Ana Ulul Azmi[1], Alfian Futuhul Hadi[2], Yuliani Setia Dewi[2], I Made Tirta[2], Firdaus Ubaidillah[2], Dian Anggraeni[2,*]

[1]*Master Student of Mathematics Department, Faculty of Mathematics and Natural Science, University of Jember*
[2]*Department of Mathematics, Faculty of Mathematics and Natural Science, University of Jember*
[*]*Corresponding author. Email: dian_a.fmipa@unej.ac.id*

ABSTRACT

Rainfall is one of the climates that has a big influence on life, such as aviation, plantations, and agriculture. Remote areas like Banyuwangi Regency are most likely to lack information on weather and climate data. Rainfall information in the future is also very decisive for the community in carrying out their daily lives, therefore prediction models or rainfall forecasting are very necessary for the community. This situation has encouraged the development of various models of approaches for forecasting rainfall. One approach for forecasting rainfall is the use of Global Circulation Model (GCM) data. GCM resolution is too low to predict local climate which is influenced by topography and land use, but it is still possible to use GCM to obtain local scale information if Statistical Downscaling (SDs) technique is used. SDs is a technique that connects GCM output as a predictor variable with local rainfall in Banyuwangi Regency as a response variable with an intermediary functional model. As for the GCM output response variable, there are three variables used in this study, namely rainfall, sea level pressure, and air temperature with a domain of 3×3 to 10×10. Forecasting rainfall in Banyuwangi Regency is carried out using the Projection Pursuit Regression (PPR) method. At the modeling stage with PPR, the optimum domain and many functions will be determined, where the chosen domain is the 6×6 domain and the optimum number of functions is m=6 with RMSEP value of 89.79. Furthermore, a process is needed to represent the forecasting results in simpler way, such as classification. The classification method used in this study is the Naive Bayes Classifier (NBC). Evidently, NBC uses PPR as a model produces forecasts classification with correct values for 18 months out of a total of 24 months with 75% of accuracy.

Keywords: *General Circulation Model (GCM), Statistical Downscaling (SDs), Projection Pursuit Regression (PPR), Classification, Naive Bayes Classifier (NBC).*

1. INTRODUCTION

The Banyuwangi Meteorology, Climatology and Geophysics Agency (BMKG) estimates that the dry season in Banyuwangi area is caused by the hot beach weather and rarely rains. Rainfall is one of the climates that has a big influence on life, such as aviation, plantations, and agriculture. Rainfall information in the future is also very decisive for the community in carrying out their daily lives, therefore prediction models or rainfall forecasting are very necessary for the community. Forecasting is an analysis of time series data that uses past events to determine the development of future events. One approach for forecasting rainfall is the use of Global Circulation Model (GCM) data.

GCM is a numerical model that is deterministic with complex computer simulations that are able to describe climatic conditions with various components that change over time [1]. GCM resolution is too low to predict local climate influenced by topography and land use, but it is still possible to use GCM to obtain local scale information if Statistical Downscaling (SDs) is used. The purpose of SDs is to find the relationship between global scale climate parameters and local scale climate parameters so that the GCM simulation projection values can be obtained on past, present, or future climates on a local scale. The SDs technique has a multivariate method, one of which is the regression method. Projection Pursuit Regression (PPR) is a non-parametric and non-linear regression method for processing large-dimensional data that can describe information in small dimensions through a projection process so that it can overcome the problems of local averages, polynomial functions, and recursive partitioning.

To convey the results of rainfall forecasting to the general public, it is not enough just to determine the amount of rainfall in the future, but more than that, namely how the weather conditions will be in the future. So we need a process that can represent the results of forecasting in the form of numbers into something that is more understandable to the public. Classification is a process to find a model or function that describes and characterizes a concept or class of data, for a particular purpose. Research on classification has been carried out by [1] and [2]. [1] analyzed the characteristics and classification of rainfall in the Maros Karst Region and [3] determined the climate in Ponorogo with the classification of Schmidt Ferguson 4 and Oldeman. In this study the classification method used is the Naive Bayes Classifier (NBC). This is because the Naive Bayes method in the process can be used for relatively small sample data. Meanwhile, other classification methods require large sample data to produce an optimal classification.

2. MATERIAL AND METHODS

2.1. Study Region

With an area of 5,782.50 km², has a coastline of about 175.8 km that stretches along the south eastern border, and has a total of 10 islands, Banyuwangi Regency is declared to be the largest district in East Java. Banyuwangi Regency is geographically located at 7º4'–8º46' South Latitude and 113º33'-114º38' East Longitude. The Banyuwangi Meteorology, Climatology and Geophysics Agency (BMKG) suspects that the dry season in the Banyuwangi area is caused by the hot coastal weather and rarely rains. In addition, Banyuwangi is also expected to be dry area due to the influence of a little cloud growth.

The presence of potential tropical storm seeds that hit Northwest Australia and low air pressure in Northwest Australia have an impact on increasing wind speed, ocean waves and also reducing cloud growth, so that the sun's rays are not blocked by clouds. This is what causes Banyuwangi to become dry, but has quite cold air, because the heat absorbed by the earth from morning to evening will be released back into the atmosphere at night with greater intensity. This is the reason that causes Banyuwangi to become a dry area with quite cold air. The heat absorbed by the earth from morning to evening will be released back into the atmosphere at night with greater intensity.

2.2. Data Description

The variables used in this study are local response variables and global predictor variables. The response variable used in this study is rainfall in monthly time units. The data used is rainfall data for the Banyuwangi region in 2011 to 2020 taken from BMKG, Banyuwangi Meteorological Station with the link https://banyuwangikab.bps.go.id/statictable/2016/10/17/115/jumlah-curah-hujanmm-per-bulan-2011-2020.html.

The predictor variable used in this study is secondary data, namely the simulation data of the GCM Climate Model Intercomparison Project (CMIP5) obtained from the web http://climexp.knmi.nl (issued by Statistical Downscalling Modeling with KNMI Netherlands) in 2011 until 2020 with the geographical position of Banyuwangi Regency located at coordinates 7º4'–8º46' South Latitude and 113º33'-114º38' East Longitude. There are 3 types of GCM output data variables that are used as predictor variables including precipitation, air temperature (air temperature) and air pressure (air pressure at sea level).

The GCM domain used in this study is a number of square-shaped grids measuring from the smallest domain grid, which is 3 × 3 units in size, to the largest domain grid, which is 10 × 10 units. Each domain grid unit has a resolution of 2.5° × 2.5° or approximately 300 km^2. Each grid domain has different features, such as the 8 × 8 grid domain size has 64 grids for the three GCM output variables, resulting in a total of 192 features as predictor variables.

2.3. Data analysis method

2.3.1. Projection Pursuit Regresion (PPR)

Projection Pursuit Regression (PPR) is a non-parametric and non-linear regression method for processing large-dimensional data that can describe information in small dimensions through a projection process so that it can overcome the problems of local averages, polynomial functions, and recursive partitioning. The following is the PPR method [4]. Based on Friedman & Stuetzle 1981, the PPR method is:

1. Determine the initial value of the residual and the value of M (number of functions) in Equation 1.

$$r_i \leftarrow y_i, \qquad i = 1,2,\dots,t \qquad (1)$$
$$M \leftarrow 0$$

$\sum y_i = 0$. Determination of the number of functions based on the optimization of many functions $m = 1,2,3,4,5,6,7,8,$ and 9. Many functions were selected based on the best validation results.

2. Determination α and S_α in the model
For the linear combination $Z = \alpha_m X$, the smoothing function $S_\alpha(Z)$ is determined according to the Z values using the projection index $I(\alpha)$, namely in Equation 2. The coefficient vector α_{M+1} which maximizes $I(\alpha)$ or the so-called Projection Pursuit (PP) index $\alpha_{M+1} = max_a{}^{-1}(I(\alpha))$ and its smoothing function is $S_{\alpha_{M+1}}(Z)$.

3. End method
If $I(\alpha)$ is less than the Threshold value, then stop. If it is still large, then change the residual value and M value as follows (Equation 2):

$$r_i \leftarrow r_i - S_\alpha(Z),\ i = 1,2,\ldots,n \quad (2)$$

$$M \leftarrow M + 1$$

And back to the step of specifying α and S_α in the model. The threshold value is obtained based on the limit of the linear combination in the scatterplot between the predictor variable and the response variable. The end of the PPR method is written in Equation 3:

$$y_i = \sum_{m=1}^{M} S_{\alpha_m}(\alpha_m X) = \beta_o + \sum_{m=1}^{M} \beta_m f_m\left(\sum_{k=1}^{n} \alpha_{km} X_{ik}\right) + \varepsilon_i \quad (3)$$

$S_{\alpha_m}(\alpha_m X)$ is an unknown function, $\alpha_m = (\alpha_{1m}, \alpha_{2m}, \ldots, \alpha_{km})$ is a unit vector (direction of projection pursuit) with m basis function, $X_i = (X_{i1}, X_{i2}, \ldots, X_{ik})$ are predictor variables ke-k and observations ke-i. y_i response variable, ε_i is a random factor with $E(\varepsilon_i) = 0$ and $var(\varepsilon_i) = \sigma^2$, X_i and ε_i are independent.

2.3.2. Naive Bayes Classifier (NBC)

The Naïve Bayes Classifier (NBC) algorithm was first proposed by Revered Thomas Bayes. This algorithm is also known as Bayesian Classification, which is a statistical classification method based on the Bayes theorem which can be used to estimate the probability of class membership [5]. On the other hand, Kononenko and Langley stated that NBC is a possible class label of a data or can also be expressed as a labeled class attribute [6]. This Bayesian classification is proven to have high classification accuracy and speed when used on a big data [7]. In addition, NBC is an algorithm in data mining technique that applies Bayes theory in classification [8]. NBC can perform well in handling big data sets and can handle irrelevant data. The equation of Bayes' theorem is as follows: (where the symbol for X is the input vector containing the data and Y is the class label);

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (4)$$

where :

X = Data with unknown class

Y = Data hypothesis X is a specific class

$P(Y|X)$ = probability of hypothesis H based on condition X (posterior probability)

$P(Y)$ = Hypothesis probability H (prior probability)

It should be noted that the classification process requires several clues to determine what class is appropriate for the sample analyzed by this NBC algorithm [9]. Therefore, Equation 4 above must be adjusted into Equation 5 as follows:

$$P(Y_j|X_1, \ldots, X_n) = \frac{P(Y)P(X_1,\ldots,X_n|Y_j)}{P(X_1,\ldots,X_n)} \quad (5)$$

The variable Y_j represents the class, while the variables $X_1, \ldots, X_n$ represent the characteristics of the the data that need to be classify. The formula of Equation 5 explains that the probability of a sample with certain characteristics belonging to the Y_j (Posterior) class is the probability of the Y_j class (before there is additional information from the sample, which is called a prior) multiplied by the probability of the appearance of a sample with that particular characteristic in the Y_j class (also called likelihood), then divided by the probability of the sample characteristic that appears globally (also called evidence). Therefore, the Equation 5 above can also be written simply as follows:

$$Posterior = \frac{prior\ x\ likelihood}{evidence} \quad (6)$$

For each class in one sample, the evidence value in Equation 6 is always fixed. Furthermore, the posterior value will be compared with other posterior class values to determine the sample class to be classified. Using the following multiplication rule (Equation 7), the Bayesian formula can be further elaborated by describing $(\mathrm{Y_j}|\mathrm{X_1}, \ldots, \mathrm{X_n})$:

$$\begin{aligned} &P(Y_j|X_1, \ldots, X_n) \quad (7) \\ &= P(Y_j)P(X_1, \ldots, X_n|Y_j) \\ &= P(Y_j)P(X_1|Y_j)P(X_2, \ldots, X_n|Y_j, X_1) \\ &= P(Y_j)P(X_1|Y_j)P(X_2|Y_j, X_1)P(X_3, \ldots, X_n|Y_j, X_1, X_2) \\ &= P(Y_j)P(X_1|Y_j)P(X_2|Y_j, X_1)P(X_3|Y_j, X_1, X_2)P(X_4, \ldots, X_n|Y_j, X_1, X_2, X_3) \\ &= P(Y_j)P(X_1|Y_j)P(X_2|Y_j, X_1)P(X_3|Y_j, X_1, X_2) \ldots P(X_n|Y_j, X_1, X_2, X_3 \ldots X_{n-1}) \end{aligned}$$

It can be seen that the complexity of the factors that affect the probability value is the result of this translation, which is almost impossible to analyze one by one. As a result, this calculation becomes difficult to do. This is where the assumption of very high independence (naive) is used, that each point $(X_1, X_2 \ldots, X_\mathrm{n})$ is independent each other. With this assumption, one thing in common is as follows (Equation 8):

$$P(X_i|Y_j) = \frac{P(X_i \cap X_j)}{P(X_j)} = \frac{P(X_i)P(X_j)}{P(X_j)} = P(X_i) \quad (8)$$

For $i \neq j$, so that

$$P(X_i|Y, X_j) = P(X_i|Y_j) \quad (9)$$

From Equation 9 above, it can be concluded that the existence of a naive independence assumption makes the probability conditions simpler, so that the calculation becomes possible. Furthermore, the translation of $P(Y_j|X_1,\dots,X_n)$ can be simplified to:

$$P(Y_j|X_1,\dots,X_n)$$
$$= P(Y_j)P(X_1|Y_j)P(X_2|Y_j)P(X_3|Y_j)\dots P(X_n|Y_j)$$
$$= (P(Y_j)\prod_{i=1}^{n} P(X_i|Y_j)) \qquad (10)$$

Information:

$P(Y_j|X_1,\dots,X_n)$ = Posterior Probability

$P(X_i|Y_j)$ = Likelihood

$P(Y_j)$ = Prior Probability

Y_{MAP} = Class with Maximum A Posterior Probability

The classification process then will use a model of Naïve Bayes theorem from the equation above (Equation 10). In classification of quantitative or continuous data, the Gauss Density formula (Equation 11) is used:

$$Y_{MAP} = \arg max_{Y_j \in Y}(P(Y_j)\prod_{i=1}^{n} P(X_i|Y_j)) \qquad (11)$$

Information:

Y_{MAP} = Class with Maximum A Posterior Probability

$P(X_i|Y_j)$= Likelihood

$P(Y_j)$ = Prior Probability

Equation (11) is a model of the Naive Bayes theorem which will then be used for the classification process. For classification with quantitative or continuous data, the Gaussian Density formula is used (Equation 12):

$$f(X_i = x_i|Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}} \qquad (12)$$

where:

f = Opportunity

X_i = Attribute to i

x_i = Attribute value to i

Y = Class sought

y_j = Y sub class you are looking for

μ_{ij} = The sample mean of the training data that belongs to y_j

2.3.3. Confusion Matrix

The classification results are recorded in the Confusion Matrix table. Measurement of classification performance is generally done by using a confusion matrix [10]. Table 1 is the result of the confusion matrix (recording the results of classification work) [11].

Table 1. Confusion matrix

		Prediction Class	
Actual Class		C1	C2
	C1	TP	FN
	C2	FP	TN

Information :

TN = The number of correct predictions is negative (True Negative)

FN = Number of false predictions is positive (False Negative)

FP = Number of wrong predictions is negative (False Positive)

TP = The number of correct predictions is positive (True Positive)

Actual is the classification of rain status in Banyuwagi which has been previously classified based on data sources. Meanwhile, prediction is the result of classification based on certain selected variables, which are generated by a program/software. The classification performance value can be calculated through the formation of a configuration matrix [12]. These values are as follows:

a. Accuracy is the proportion of correct classification prediction. The accuracy formula is written in Equation 13:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (13)$$

b. Error Rate is the proportion of the classification making prediction errors, with the calculation of the following equation (Equation 14):

$$ErrorRate = \frac{FP+FN}{TP+TN+FP+FN} \qquad (14)$$

3. RESULT AND DISCUSSION

3.1. Dataset

Most of the rainfall data in Banyuwangi Regency in the period January 2011 to December 2020 are below 300 mm.

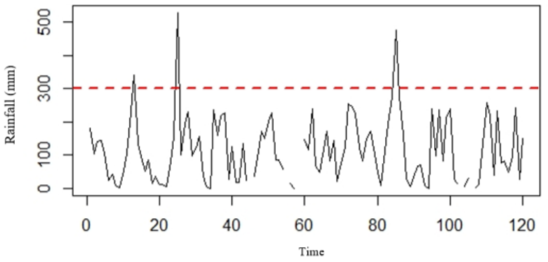

Figure 1 Graph of rainfall data for Banyuwangi Regency 2011-2020.

Figure 1 shows a broken line indicating that there is an empty value at that time. Empty data must be filled with interpolation or other techniques. For this reason, Banyuwangi Regency rainfall data is then modeled using the SDs method. For modeling using the SDs method, other variables are needed as explanatory variables (predictors) and as rainfall variables (response variables). There are 3 types of predictor variables used, namely precipitation, air temperature and air pressure at sea level.

The initial data used 300 predictor variables to be used with details of 100 predictor variables of precipitation, 100 predictor variables of air pressure, and 100 predictor variables of temperature showed in Figure 2. This study has a large data dimension where local data is only 120 data but the predictor variable is 300 data.

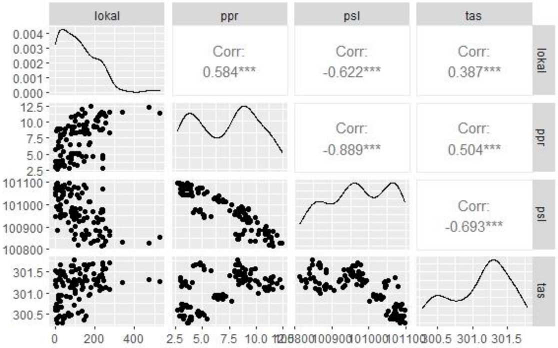

Figure 2 Pairsplots of rainfall, precipitation, air pressure, and temperature.

3.2. *Rainfall Forecasting Using the PPR Model*

In this study, the optimum PPR model was obtained in a domain measuring 6×6 units and the number of projections was 6 (m=6). Based on the beta coefficients generated in this study, the selected PPR model produced is

$$y_i = 115{,}095 + 100{,}805 f_1(\sum_{k=1}^{108} \alpha_{k1} X_{ik}) + 41{,}609 f_2(\sum_{k=1}^{108} \alpha_{k2} X_{ik}) + 38{,}255 f_3(\sum_{k=1}^{108} \alpha_{k3} X_{ik}) + 28{,}911 f_4(\sum_{k=1}^{108} \alpha_{k4} X_{ik}) + 24{,}275 f_5(\sum_{k=1}^{108} \alpha_{k5} X_{ik}) + 29{,}087 f_6(\sum_{k=1}^{108} \alpha_{k6} X_{ik}) + \varepsilon_i \quad (15)$$

Equation 15 is a functional rule that connects 108 GCM output predictor variables with monthly rainfall in Banyuwangi Regency.

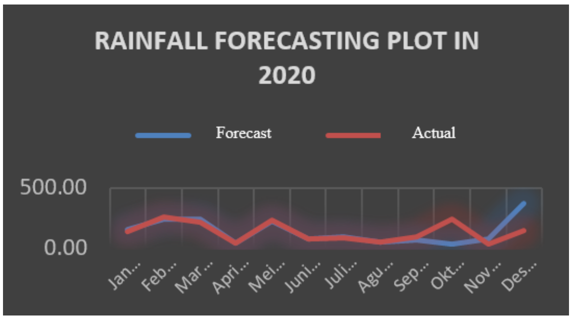

Figure 3 Monthly rainfall forecasting plot Banyuwangi Regency in 2020

Figure 3 shows the difference between the forecast result and the actual value. The red line is the actual value while the blue line is the forecast result. Finding the RMSEP (Root Mean Square of Prediction Error) value can use the difference between the forecast value and the actual value, where the complete formula is in Equation 16 as follows:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (16)$$

where :

$\hat{y}_i$ = conjecture dependent variable

y_i = true dependent variable

n = amount of data

The RMSEP value is obtained from the root of the sum of the differences in the squares of the forecast value and the actual value divided by the amount of data. So that the RMSEP value obtained is 89.79mm. Where this value proves that the PPR model is good enough for forecasting.

3.3. *Naive Bayes Classifier on Rainfall*

Banyuwangi rainfall data can be classified into 4 classes, namely dry months, humid months, wet months, and very wet months. Classification using NBC in this study was carried out using PPR forecasting results.

Table 2. Confusion matrix of NBC classification testing results using PPR

		Prediction Class			
		Dry	Moist	Wet	Very Wet
Actual Class	Dry	8	1	0	1
	Moist	1	3	0	0
	Wet	0	1	0	0
	Very Wet	0	0	2	7

Table 2 is a confusion matrix that shows the number of correct values in each category. Meanwhile, NBC uses PPR as a model that produces classification forecasts with correct values for 18 months out of a total of 24 months. The correct values consist of 8 wet months, 3 wet months and 7 very wet months. From equation (13) the confusion matrix in Table 1 produces an accuracy rate of 75%.

4. CONCLUSION

The use of statistical downscaling techniques with the Projection Pursuit Regression (PPR) model for forecasting monthly rainfall in Banyuwangi Regency using a $\mathbf{6} \times \mathbf{6}$ grid domain with 3 variables, namely precipitation, air temperature (air temperature) and air pressure (air pressure at sea level) provides good performance with RMSEP value of 89.79. Classification using Naive Bayes Classifier (NBC) in this study was carried out on the $\mathbf{6} \times \mathbf{6}$ domain using Projection Pursuit Regression (PPR). The research data were first classified into 4 classes, namely dry months, humid months, wet months, and very wet months. The results of the NBC classification using PPR become a model that produces a classification forecast with a correct value of 18 months out of a total of 24 months or in other terms it produces an accuracy rate of 75%.

REFERENCES

[1] A.H. Wigena, Pemodelan Statistical Downscaling dengan Regresi Projection Pursuit untuk Peramalan Curah hujan Bulanan, *Disertasi,* (in Indonesian), Bogor: Institut Pertanian Bogor, 2006.

[2] M. Arsyad, Sulistiawati, A.T. Vistarani, Analisis Karakteristik dan Klasifikasi Curah Hujan di Kawasan Karst Maros, (in Indonesian), Proceeding Internasional Seminar On mathematics, Science, and Computer Education, 2015, 57-62.

[3] R.A. Sasminto, A. Tunggul, J. B. Rahadi. Analisis Spasial Penentuan Iklim Menurut Klasifikasi Schmidt-Ferguson Oldeman di Kabupaten Ponorogo, (in Indonesian), Jurnal Sumberdaya Alam & Lingkungan, 2013, 1(1): 51-56.

[4] H.F. Jerome, W. Stuetzle. Projection Pursuit Regression, Journal of The American Statistical Association, 1981, 76(376): 817-823.

[5] D.J. Hand, K. Yu, Idiot’s Bayes: Not So Stupid after All, International Statistical Review, 2001, 69 (3): 385-398.

[6] P. Langley, S. Sage, Induction of Selective Bayesian Classifier, Proceeding of The Tenth Conference on Uncertainty in Artificial Intelligence, 29 Juli 1994, Morgan Kaufman: 399-406.

[7] D.D. Lewis, Naïve (Bayes) at forty: The independence assumption in information retrieval, Proceedings of the Tenth European Conference on Machine Learning, April 1998, Berlin: 4-15.

[8] M.H. Ridwan, Suyono, M. Sarosa, Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Metode Naive Bayes Classifier, (in Indonesian), Jurnal EECCIS, 2013, 7(1): 59-64.

[9] A. Saleh, Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga, (in Indonesian), Citec Journal, 2015, 2(3): 207–217.

[10] E. Prasetyo, Data Mining Konsep dan Aplikasi Menggunakan MATLAB, (in Indonesian), Yogyakarta: ANDI Yogyakarta, 2012.

[11] A. Novandya, I. Oktria, Penerapan Metode Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi. Jurnal Format, 2017, 6(2): 98-106.

[12] M. Faisal, D. T. Nugrahadi, Belajar Data Science Klasifikasi dengan Bahasa Pemrograman R, (in Indonesian), Banjarbaru: Scripta Cendekia, 2019.

.