Diagnostic Analysis Of Rice Productivity Using Classification Based On Shannon And Renyi Entropy

Fajar Delli Wihartiko, Eneng Tita Tosida, Ruhul Amin

Abstract: Rice is the main commodity in Indonesia both for consumption and in terms of production. The increasing number of Indonesian population resulting in increased demand for rice is a problem that must be faced by the government to maintain national food stability. Currently Indonesian rice productivity data is available at the Central Statistics Agency and at the Ministry of Agriculture. The data is used in descriptive and diagnostic analysis. Descriptive analysis uses clustering, data visualization and entropy. The diagnostic process uses an entropy-based classification to see factors of production. The entropy function used is Shannon and Renyi Entropy. The results of using entropy in the description analysis show that production attributes have a higher level of uniformity. The results of entropy in the classification show that there are differences in the decision tree that results from Shannon and Renyi entropy. In this case Renyi Entropy has better accuracy..

Index Terms: Shannon Entropy, Renyi Entropy, Diagnostic Analysis, Rice Productivity

1 INTRODUCTION

The agricultural sector is a very important sector its role in the economy in most developing countries, especially in Indonesia. The role of the agricultural sector is to accommodate the population and provide employment opportunities for the population. Rice is a food commodity that as the main food ingredient in Indonesia. The disruption of rice production and supply will have a significant impact on other sectors such as the economy and community welfare. Indonesia was once one of the leading rice producing countries in the world. In 2014, Indonesia was the largest rice producer in the world after China and India. However, in recent years Indonesia has carried out an import policy of around 3 million tons of rice each year from Thailand and Vietnam with the aim of securing the country's rice reserves. Rice commodity is the most strategic commodity that optimization of comprehensive data management is needed. Integrated data on rice production, consumption and prices greatly affect government policies for imports and also the decision of farmers to plant rice . At present national food data is sourced from all regions managed by the Ministry of Agriculture and published by the Central Statistics Agency. Management of big data (especially rice commodities) is important not only for the government but also for farmers and stakeholders. Descriptive analysis is needed to provide a good overview of the available data, in this case the data used is sourced from the Central Statistics Agency and from the Indonesian Ministry of Agriculture.

- *FD Wihartiko, PhD student in Computer Science, IPB University, Indonesia and lecturer in the computer science department, Pakuan University, Indonesia , fajardelli@ipb.apps.ibb.ac.id; fajardelli@unpak.ac.id*
- *ET Tosida, PhD student in Computer Science, IPB University, Indonesia and lecturer in the computer science department, Pakuan University, Indonesia , enengtitatosida@apps.ipb.ac.id*
- *R Amin, PhD student in Computer Science, IPB University, Indonesia and lecturer in the informatics department, STMIK Nusaa Mandiri, ruhulamin@ipb.apps.ibb.ac.id;*

Diagnostic analysis is used to look at national rice productivity factors. Research on rice commodities has been widely carried out as by [1] who has conducted a review of machine learning using multivariate data analysis methods used in food safety. Predictions regarding production results have been carried out by [2] who use the SVM model for predictive analysis. The model is also combined with weather conditions as input variables. Research [3] discusses the diagnosis of rice disease using images. The study used the SVM model with an accuracy of 87.9%. Research [4] uses predictive analysis to estimate the origin of rice samples. The model used is EL / RF with an accuracy of 93.83%. The majority of research in rice is a predictive model. The contribution of this paper is in the development of Shannon and Reny entropy for the descriptive analysis of categorical data with different number of events per attribute. Development is also done by modifying the decision tree algorithm for the classification process which also uses Shannon and Renyi Entropy. Entropy implementation has been carried out by [5], [6] and [7].

2 MATERIAL AND METHODS

2.1 Material

Data on rice commodities has been available both at the Ministry of Agriculture (https://www.pertanian.go.id) and the Central Statistics Agency (https://www.bps.go.id). This fact shows that data and research are available to support national food policy. This paper aims to look at the descriptions and factors that influence national rice production based on Shannon and Renyi Entrophy. The data used as the target class is Indonesian rice productivity data from the Ministry of Agriculture for 2014 - 2018 which can be downloaded at the following address https://www.pertanian.go.id/home/?show=page&act=view&id 61. Data on the area of rice production and rice production was obtained from the Ministry of Agriculture. While the regional topographic data, geographical, informal workers in the agricultural sector are sourced from the Central Statistics Agency (https://www.bps.go.id/linkTableDinamis /view/id/). A description of the data before pre-processing can be seen in Table 1.

TABLE 1
Data Description Before Pre-processing

No	Attribute	Data Type	Range Value	Remarks
1	Year	Ratio	2014-2018	
2	Province	Nominal	34 Province	
3	Production	Ratio	0 – 1000000	In Tons
4	Land Area	Ratio	0 – 1000000	In Hectares
5	Topography	Ratio	0-10000	Number of villages
6	Geographical	Ratio	0-10000	Number of villages
7	Informal Labor in the agricultural sector	Ratio	0 - 100	percentage
8	Productivity	Ratio	0-100	quintal / hectare

2.2 Method

The stages of the research used were modified from KDD [8] which were integrated with the descriptive and diagnostic stages of [9] as follows:

1. Data collection, in this stage the data is collected from various trusted sources. this study uses sourced data from the Central Statistics Agency and from the Indonesian Ministry of Agriculture.
2. Preprocessing data, in this stage the process of data discretization is done by changing the numeric data into categorical data. The discretization process uses the k-means cluster algorithm [10]. The process of determining the number of clusters is based on BPS reports [11] and histogram analysis.
3. Descriptive analytics, is the stage where will be seen how the data display after data processing. In this stage the uniformity of data will also be seen using entropy-based analysis, in this case using Shannon and Renyi entrophy [12] as follows:
   Let H(P) contained in a series of opportunities p_i ... p_N should fulfil the requirements:
   - H(P) continue function at p_i;
   - if the opportunities of p_i are equal (p_i=1/N) then H(P) must be a Monotonous function rom N;
   - H(P) can be an additive function.

   Shannon proved that the following function H(P) fulfills the three conditions above, where K is a positive constant. This function became known as Shannon Entropy.

$$H(P) = -K \sum_{i=1}^{N} p_i \ln(p_i) \quad (1)$$

   He expansion of Shannon's original work has produced many alternative steps of information or entropy. For example, Renyi entropy H_q (P) is obtained by removing the third requirements from Shanon. Renyi entropy is formulated as :

$$H_q(P) = -\frac{1}{1-q} \ln \sum_{i=1}^{N} p_i^{\,q} \quad (2)$$

   In comparing the size of entropy with different attributes, for the same number of events (i), the more uniform the data, will impact to the higher the value of entropy. The number of scales in the attribute also affects the value of the entropy. In the case of uniform distribution, the greater the number of events (i), will impact to the higher the entropy value. To compare the entropy value of different attributes of events, for example, defined Hs is the entropy value for uniform data of an event (i) of number s. ΔH is defined as the difference between Hs and H

$$\Delta H = H_s - H \quad (3)$$

   where the smaller the value of ΔH, it means the closer to the form of events with a uniform opportunity.
4. Diagnostic analysis, historical data can be measured against other data to answer the question why something happened. It is possible to explore, determine dependencies, and identify patterns. In this stage the classification process is carried out by modifying the entropy value calculation on the Decision Tree algorithm [20] where in the process initially using shannon entrophy (1) then modifying it using Renyi Entropy (2). In this paper we compare the results of the tree using Shannon and Renyi Entropy.
5. Evaluation, by using error analysis [13] based on the results of the tree obtained in step 4.
6. Knowledge, the use of KDD results to be considered further

In summary, the modified KDD process carried out can be seen in the following figure:

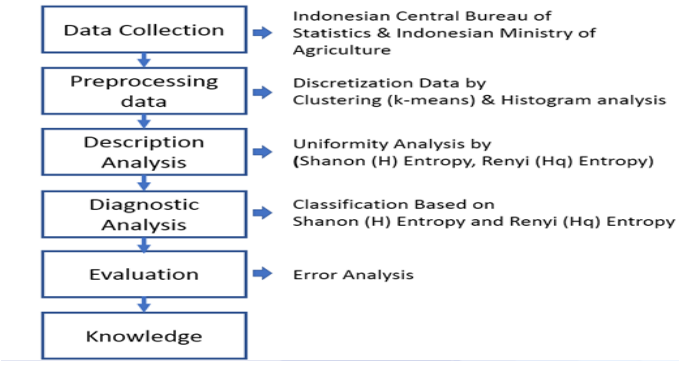

Fig. 1*. Research Method*

3 RESULTS AND ANALYSIS

3.1 Praproses

The discretization process is carried out on the available data using the k means cluster algorithm and histogram analysis. Data trends are obtained based on the results of the linear regression model for each province where the positive trend shows that production growth tends to increase every year and the negative trend shows a trend of decreasing production each year. The results of data processing can be seen in Figure 2.

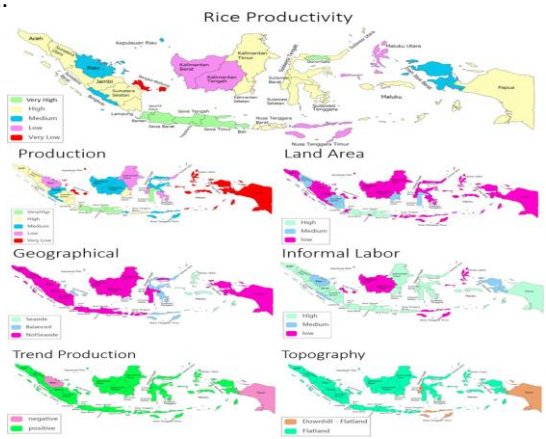

Fig2. *Cluster Results Map According to Attributes*

3.2 Descriptive Analytics

Description of data based on preprocessed data can be seen

in Table 2. From Table 2 an analysis of data uniformity will be done using Shannon and Renyi Entropy. The results of data analysis using Shannon and Renyi Entropy can be seen in Table 3From table 3 it can be seen that the production attribute is the attribute with the smallest ΔH (also ΔHq) value, which means that the attribute tends to have a chance of occurring close to uniformity. Shannon and Renyi Entropy provides the same results regarding the order of uniformity of attributes, even though the resulting values vary.

TABLE 2
Description of Cluster Results Data

Trend	:	Positive	Negative				Total
amount		30	4				34
probability		0,88	0,12				1
Topography	:	Flatland	Downhill - Flatland				
amount		31	3				34
probability		0,911764706	0,088235				1
Geographical	:	NotSeaside	Balanced	Seaside			
amount		26	6	2			34
probability		0,764705882	0,176471	0,05882353			1
Informal Labor	:	Low	medium	high			
amount		1	9	24			34
probability		0,029411765	0,264706	0,70588235			1
Area	:	Low	Medium	High			
amount		26	5	3			34
probability		0,76	0,15	0,09			1
Production	:	VeryLow	Low	sufficient	High	VeryHigh	
amount		7	7	8	8	4	34
probability		0,21	0,21	0,24	0,24	0,12	1
Productivity	:	VeryLow	Low	Medium	High	VeryHigh	
amount		2	4	4	16	8	34
probability		0,058823529	0,117647	0,11764706	0,4706	0,235294	1

TABLE 3
Descriptive Analysis using Entropy

Attribute	s	H	Hq	Hs	ΔH	ΔHq
Trend	2	0,362210557	0,232704685	0,693147181	0,331	0,460
Topography	2	0,298435813	0,175424978	0,693147181	0,395	0,518
Geographical	3	0,677908726	0,479040882	1,098612289	0,421	0,620
Informal Labor	3	0,701410139	0,563516118	1,098612289	0,397	0,535
Area	3	0,701256512	0,487456079	1,098612289	0,397	0,611
Production	5	1,583449225	1,563783323	1,609437912	0,026	0,046
Productivity	5	1,365372256	1,177790318	1,609437912	0,244	0,432

3.3 Diagnostic Analytics

The gain value in Decision Tree is the value obtained from the Entropy value of a Target class reduced by the entropy value of The higher the information gain, will impact to the more the reduction in entropy, and the better the split-point. Thus, given split points and their corresponding parts, we can score each split point and choose the one that gives the highest information gain [10]. Gain calculation results based on Shannon and Renyi Entropy for the first iteration can be seen in Table 4.

TABLE 4
Gain Calculation Results based on Shannon and Renyi Entropy

Gain	Shannon	Renyi
Trend	0,05661	0,23292
Topography	0,05498	0,21798
Geographical	0,11787	0,27921
Informal Labor	0,22257	0,34763
Area	0,25332	0,36194
Production	0,23986	0,41763

From Table 4 it can be seen that for Shannon entropy the most influential attribute on productivity is land area. In contrast to the highest gain values in the Renyi Entropy. The most influential value is Production. This causes different tree results. Figure 3 is the result of Decision Tree using Shannon Entropy, while Figure 4 is the result of Decision Tree using Renyi Entropy

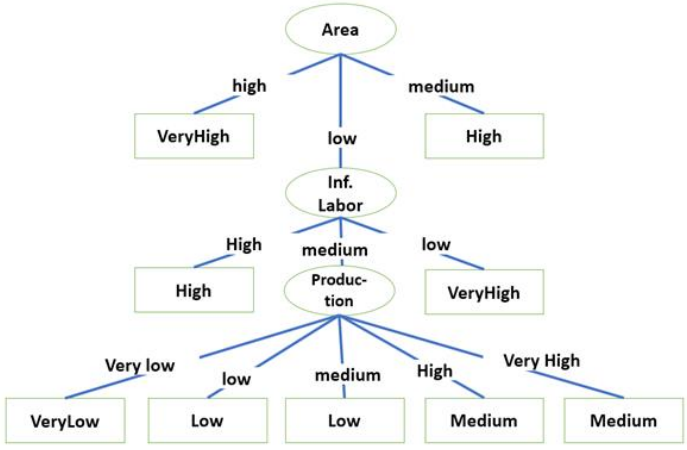

Fig 3*. Effect of Productivity in the form of Decision Tree based on Shannon Entropy*

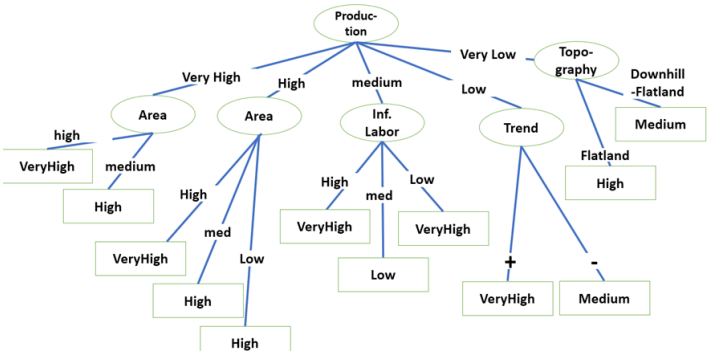

Fig 4*. Effect of Productivity in the form of Decision Tree based on Renyi Entropy.*

3.4 ANALYSIS

The evaluation process is done by looking at the accuracy of the data with the model. Evaluation of the Decision Tree model to the available data shows that the accuracy value generated from the Decision Tree model with an Shannon Entropy is 47.06%. While the accuracy of the Decision Tree model using Renyi Entropy is 58.82%. In this case it shows that the accuracy of Renyi Entropy is superior when compared to Shannon Entropy.

4 CONCLUSION

The implementation of Shannon and Renyi entropy has been carried out on Indonesian rice production data from Central Statistics Agency and the Ministry of Agriculture. The results of the application of data clustering show that the production data has higher entropy values than the broad attributes and trends which indicate that the production attributes have a higher level of uniformity. There is no difference in the results in the order of attributes based on the results of the value of Shannon and Renyi entropy. The results of the application of Shannon and Renyi entropy in the classification show different tree results and accuracy. In this case the classification model based on Renyi entropy has better accuracy. The challenge of future research is to improve the performance of the classification model.

5 REFERENCES

[1] Maione, C.; Barbosa, R.M. Recent applications of multivariate data analysis methods in the authentication of rice and the most analyzed parameters: A review. Crit. Rev. Food Sci. Nutr. 2018, 1–12.

[2] Su, Y.; Xu, H.; Yan, L. Support vector machine-based open crop model (SBOCM): Case of rice production in China. Saudi J. Biol. Sci. 2017, 24, 537–547.

[3] Chung, C.L.; Huang, K.J.; Chen, S.Y.; Lai, M.H.; Chen, Y.C.; Kuo, Y.F. Detecting Bakanae disease in rice seedlings by machine vision. Comput. Electron. Agric. 2016, 121, 404–411.

[4] Maione, C.; Batista, B.L.; Campiglia, A.D.; Barbosa, F.; Barbosa, R.M. Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. Comput. Electron. Agric. 2016, 121, 101–107.

[5] Mays DC, Boris A. Faybishenko, and Stefan Finsterle (2002). Information entropy to measure temporal and spatial complexity of unsaturated flow in heterogeneous media. Water Resources Research, VOL. 38, NO. 12, 1313, doi:10.1029/2001

[6] Prasetyo, A., Koestoer, R. H., & Waryono, T. (2016). Pola Spasial Penjalaran Perkotaan Bodetabek : Studi Aplikasi Model Shannon ' S Entropy, 16, 144–160. Jurnal Pendidikan Geografi, Volume 16, Nomor 2, Oktober 2016

[7] Purvis B, Yong M, Darren R (2019)Entropy and its Application to Urban Systems.Entropy. 21, 56; doi:10.3390/e21010056

[8] Han & Kamber.2013.Data mining 3rd Edition.Elsevier.United States of America

[9] Bekker A, 2018. 4 Types of Data Analytics to Improve Decision-Making https://www.scnsoft.com/

[10] M. J. Zaki, Jr. W. Meira and W. Meira. Data mining and analysis: fundamental concepts and algorithms. Cambridge: Cambridge University Press, 2014.

[11] BPS. 2018. Ringkasan Eksekutif Luas Panen dan Produksi Beras di Indonesia 2018. ISSN / ISBN : 978-602-438-237-7

[12] Bromiley, P. A., & Thacker, N. A. (2010). Shannon Entropy , Renyi Entropy , and Information, (2004). Statistics and Segmentation Series (2008-001)

[13] Mathews JH, Kurtis DF. Numerical Methods using Matlab. New Jersey: Pearson Education International. 2004.