SINOMICS JOURNAL
International Journal of Social Science, Education, Communication and Economics
ISSN (e): 2829-7350 | ISSN(p): 2963-9441

Clustering Food Security Levels based on Population, Area of Harvested Land, and Disaster Risk Index in Indonesia

Ignatius Alvin Krisnugraha[1], Kgs Muhammad Benyamin Azhary[2]
School of Architecture, Planning and Policy Development, Institut Teknologi Bandung, Indonesia
Email: alvinkrisnugraha@gmail.com[1], ben.archi12@gmail.com[2]

Abstract

Problems threatening food security are reduced agricultural land and decreased production caused by climate change. In 2 years, the rice harvested area in Indonesia has decreased by around 6.33%. This research aims to know how food security is in Indonesia.The secondary data is then processed using the Microsoft Excel application and analyzed using Machine Learning in Rstudio. Several variables were examined, Population, Area of Harvested Land, Productivity of Rice Commodity, and Level of Disaster Risk. Based on the literature, a food security ratio is calculated to get the index that will explain the deficit or surplus of food availability. This research shows that the model has an accuracy rate of 93.4%. Strengthening the results given by the elbow method, the gap statistical method ensures that the optimal number of clusters recommended by machine learning is k2, so the next process is to group districts/cities into 2 (two) clusters. a model can be formed that can be used to predict the condition of food security fairly accurately (accuracy level 90%). Cities in Indonesia when clustered based on this research can optimally be divided into two clusters which show the deficit and surplus of food.

Keywords | *food security, machine learning, agriculture, population, harvested land, disaster risk*

INTRODUCTION

Indonesia is a large country with an area of around 830 million hectares. The tropical climate and fertile soil are some of Indonesia's advantages making it suitable for agricultural activities. Rice is the main agricultural commodity in Indonesia which is a supporter of food security in Indonesia, considering that rice is still the main source of nutrition and energy for more than 90% of Indonesia's population (Agricultural Research and Development Agency, Ministry of Agriculture, 2005). In 2019, the Ministry of Agriculture said that the amount of agricultural land was almost 19 million hectares consisting of irrigated rice fields, non-irrigated rice fields, and gardens (Center for Agricultural Data and Information Systems, Ministry of Agriculture, 2020). Furthermore, based on BPS data, in 2018 Indonesia had 11,377,934 hectares of rice harvest land, but in 2020 this data has reduced to around 10,657,275 hectares of rice harvest land. In 2 years, the rice harvest area in Indonesia has decreased by 720,659 hectares or around 6.33%. Whereas in Indonesia there are around 245,000 km2 or about 24.5 million hectares of land that are considered suitable for development as wet agricultural land (rice fields) (Agricultural Research and Development Agency, Ministry of Agriculture, 2005). Problems that threaten food security are reduced agricultural land and decreased production caused by climate change (Hapsari & Rudiarto, 2017). The disaster indicator means having a stake in food security and vulnerability in an area (Ministry of Agriculture Food Security Agency, 2020). Infrastructure development which is being pursued by Indonesia in the context of national economic growth will directly or indirectly have an impact on the sustainability of agricultural land.

On the other hand, the decrease in the area of harvested land is inversely proportional to the population which continues to increase from year to year. Indonesia's population was 277.2 million in 2018 and increased by 2.49% to 284.1 million in 2020. This implies that the need for food will continue to increase, so food planning must pay attention to population growth and distribution (Government of the Republic of Indonesia, 2012). In other words, population size and density play an important role in the process of realizing food security. This research was then carried out with the aim of knowing how food security is in Indonesia at the district/city level. Conditions such as harvested area, population, and disaster risk index are then used to predict the ratio of food security or vulnerability in Indonesia, particularly at the district/city level, using secondary data for three years (2018-2020). Furthermore, a clustering process will also be carried out based on the variable characteristics of population, area of rice harvest land, and disaster risk to obtain groupings of regions in Indonesia that are vulnerable to food security hazards.

IMPLEMENTATION METHOD

The data used in this study are population data, rice commodity production data, harvested land area data, and disaster indexes obtained from the BPS website for each province and BNPB. The secondary data is then processed using the Microsoft Excel application, to then be analyzed using the Rstudio application.

Table 1 Variable List

No.	Variable	Variable ID on Rstudio	Unit	Data Type	Source
1	Population Number	penduduk	People	Number	BPS Website each Province 2018-2020
2	Rice Commodity Production	beras	Ton	Number	BPS Website each Province 2018-2020
3	Harvested Rice Land	lahan	Hectare	Number	BPS Website each Province 2018-2020
4	Disaster Risk Index	skor_bencana	-	Number	BNPB, Indeks Risiko Bencana Indonesia tahun 2020
5	Disaster Risk Class	risiko_bencana	-	Factor	BNPB, Indeks Risiko Bencana Indonesia tahun 2020
6	Food Availability Ratio (I_{AV})	ketahanan	-	Number	Normative consumption number divided by food availability in a day
7	Target Food Availability	Target	-	Factor	I_{AV}>1, food deficit I_{AV}<1, food surplus

Several variables were analyzed, namely Population, Area of Harvested Land, Productivity of Rice Commodity, and Level of Disaster Risk. A total of 514 (five hundred

SINOMICS JOURNAL
International Journal of Social Science, Education, Communication and Economics
ISSN (e): 2829-7350 | ISSN(p): 2963-9441

and fourteen) Regency/City data were successfully collected during the 2018-2020 period sourced from the BPS website. Based on the provincial level Food Security and Vulnerability Atlas (FSVA) for 2020, a food security ratio is calculated with the $I_{AV} < 1$ category describing a food surplus area. Regions with an I_{AV} category > 1 indicate food deficit areas.

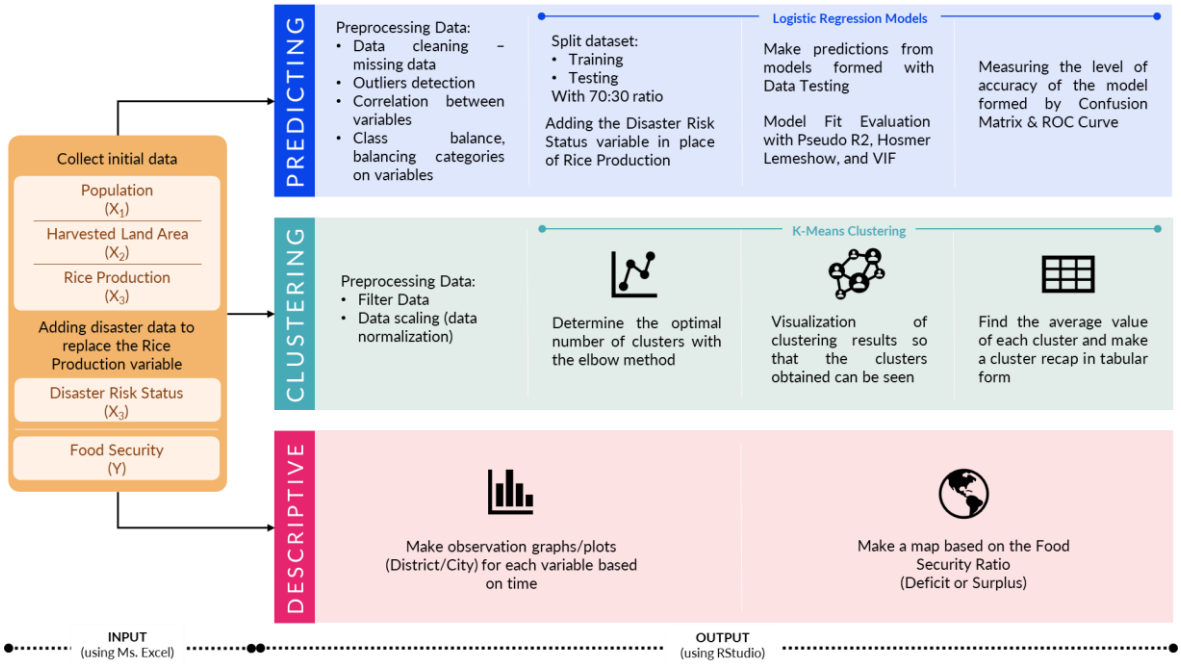

Image 1 Research Process and Methods

Two main analysis methods are used, the first is to create a Machine Learning-based predictive model about the Food Security Ratio of an area based on existing variables. The second is to create a regional cluster model based on the characteristics of the existing variables. Apart from these two main analytical methods, several visual analyzes such as line plots, scatter plots, and choropleths are also described to describe the data description.

RESULTS AND DISCUSSION

Predictive Models

The prediction model is built based on data on Population, Harvested Land Area, and Total Rice Production in 2020 alone. Before conducting model analysis, preprocessing or initial data processing is carried out first. This preprocessing stage focuses on handling missing data or missing or empty data and outliers or outliers. Missing data is a real fact in a study, and whether it is necessary to replace the missing values in the missing data or not is a deep topic (Denis, 2020). Preprocessing is important so that the data used in forming the model is free from noise.

The initial data used has 514 (five hundred and fourteen) row data at the Regency/City level. In fact, some of the data collected from the National, Provincial and District BPS websites is incomplete (missing data). It is this data limitation that needs to be followed up by preprocessing, in this case for missing data. Of the 514 (five hundred and fourteen) row

data, there are 52 (fifty two) rows of which the data is incomplete or contains missing data. To avoid building an invalid model, the 52 (fifty two) incomplete data were removed. Then a new dataset is formed with the name "dataset2" with a total of 462 (four hundred and sixty two) data, which is the initial data (514 rows of data) minus incomplete data (52 rows of data) that have just been deleted.

The remaining data then needs to be observed for outliers or outliers to avoid unreasonable data values and make predictions inaccurate. The process of detecting outliers is carried out using histograms and box plots of the output variables.

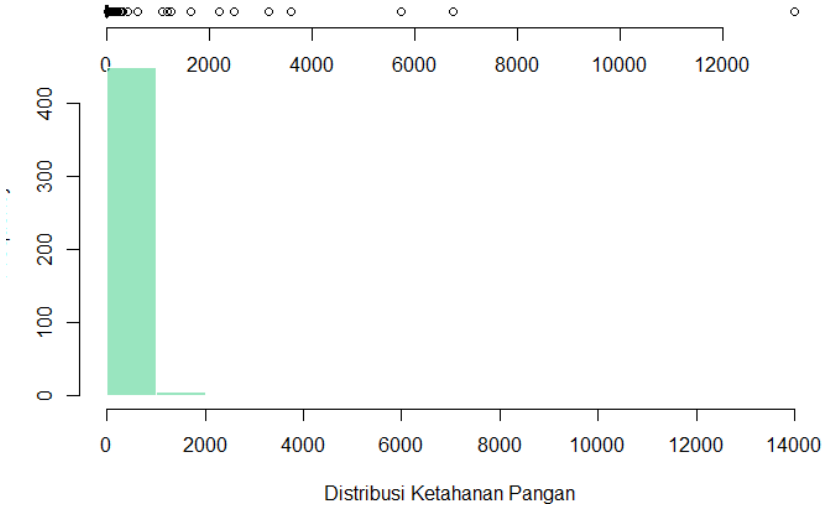

Image 2 Outliers detection using histogram and box plot of output variables

The histogram shows that there are outliers in the data. Further outliers will be removed from the dataset2 with the Quartile determination approach. If you look at the distribution of existing data, Quartile I and Quartile II are more suitable to be used as boundaries to eliminate outliers. Dataset3 is formed from the initial data which has been removed by missing data and its outliers, so that 353 rows of data are obtained. The dataset3 is then checked again for the histogram and box plot to see the distribution of the data.

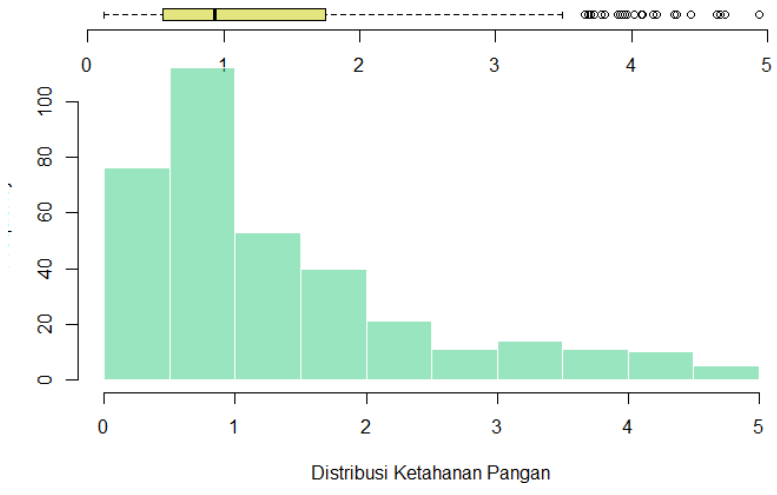

Image 3 Data distribution from the current dataset

SINOMICS JOURNAL
International Journal of Social Science, Education, Communication and Economics
ISSN (e): 2829-7350 | ISSN(p): 2963-9441

The histogram shows the distribution of data in dataset3, with the box plot showing the concentration of the data, where there are still some values that are outside the upper limit in the box plot. Data whose values exceed the upper limit of the box plot are then still used with the consideration that these data are quite large and may contain important information. Then check again whether there is still missing data or not. Dataset3 is clean from missing data and outliers, with 353 rows of data. The magnitude of the correlation or relationship between variables, both the dependent and independent variables and the independent and independent variables is described as one of the conditions in the formation of a predictive model.

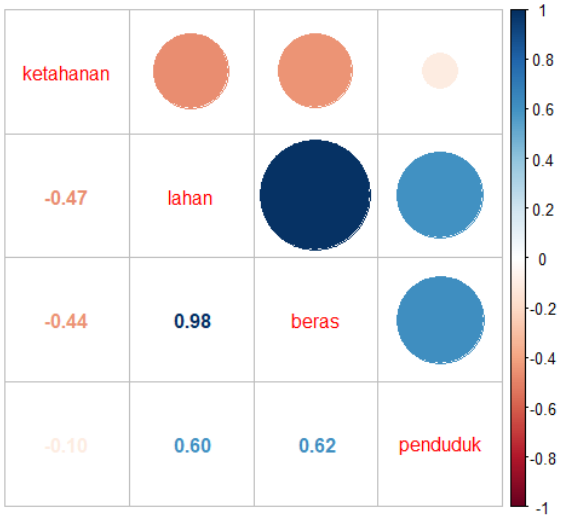

Image 4 Correlation matrix

The correlation matrix shows that there is a very strong relationship between land area and the amount of rice commodity production, with a correlation value of = 0.98. This very strong relationship can be considered reasonable, considering that the amount of rice production is the output produced from the area of agricultural land. In other words, the wider the area of paddy harvested land, the higher the amount of rice production. In forming the model, it is feared that two independent variables that have a very high correlation will interfere with the fit of the model.

The output variable consists of 2 (two) values, namely "DEFICIT" and "SURPLUS", which can then be symbolized as 1 for deficit and 2 for surplus. With these assumptions, the output variable in the "target" column needs to be converted into a factor form and then a logistic regression model is created. From the final data set, 165 regions experienced a deficit and 188 regions experienced a surplus. The comparison between the two conditions is quite balanced, so it can be a good asset in establishing a logistic regression model. The process of forming a logistic regression model begins by dividing the data into two classes, namely training data and testing data with a ratio of 70:30. Data Training is used to build models, while Data Testing is used to test or evaluate models.

Then proceed with the formation of a Prediction Model, where in the correlation analysis it is found that there is a strong correlation between the area of harvested land and

the amount of rice commodity production. This means that the total area of harvested land is directly proportional to the amount of rice commodity production. In the regression model, such multicollinearity conditions are not permitted. To test the multicollinearity that occurs, the Variance Inflation Factor (VIF) parameter is used in the model with a maximum threshold of 5 (five).

In accordance with the results of the correlation analysis, it was found that the variable area of harvested land and the amount of rice commodity production had a very high VIF (Variance Inflation Factor) value (more than 32). This shows that there is indeed multicollinearity in the model, so that one of the 2 (two) variables must be eliminated. Looking at the VIF results, the variable amount of rice commodity production has the highest value, so it will be omitted from the model.

Based on the Provincial Food Security and Insecurity Map Guidelines 2020, one of the indicators determining food security and vulnerability is the aspect of vulnerability to transient food vulnerability, which includes indicators of natural disasters (Ministry of Agriculture's Food Security Agency, 2020). Taking this into account, this study attempts to replace the variable amount of rice production and incorporate new variables related to disasters, namely the 2020 Regency/City Disaster Risk Index and Disaster Risk Class which represent the level of disaster risk in each Regency/City. The higher the value of the Disaster Risk Index, the higher the risk of a disaster occurring in that area. The included disaster variable aims to enrich the model, as well as to determine its effect on food security in an area. After adding the new variables, the process of establishing the logistic model will continue. Previously, there would be redistribution for Training data and Testing data using the new data, still with a composition of 70:30. Then the models were formed using the variables of population, area of harvested land, and class of disaster risk.

Evaluation of Model Fit is the next stage after the training data and testing data are formed. Evaluation of this model is by looking at the suitability of the model that has been formed based on several parameters using Pseudo R2, Hosmer Lemeshow Model Fit Test, and Variance Inflation Factor (VIF). Through evaluation using Pseudo R2, the value of Cox and Snell (ML) was obtained with a value of 0.6471. This value means that the variation of the Food Security variable is 64.71%, which comes from the input variables used, namely the area of harvested land, population, and disaster risk. While using the Hosmer Lemeshow Model Fit Test, a p-value of 0.2862 was obtained. p-value is more than α (where $\alpha = 0.05$), which means that the model is fit.

```
#2 Hosmer Lemeshow model fit test
options(warn = -3)
library(generalhoslem)
Hosmer <- logitgof(train$target, fitted(model2))
Hosmer
```

```
    Hosmer and Lemeshow test (binary model)

data:  train$target, fitted(model2)
X-squared = 9.7064, df = 8, p-value = 0.2862
```

Image 5 Hosmer Lemeshow Model Fit Test

SINOMICS JOURNAL
International Journal of Social Science, Education, Communication and Economics
ISSN (e): 2829-7350 | ISSN(p): 2963-9441

Through the VIF test it can be seen that the value of each independent variable is less than 5, meaning that there is no multicollinearity problem as previously found. From three evaluation parameters used, it can be proven that the model is fit with the data. After obtaining the model and evaluating the suitability of the model, the data test will be used to make predictions and then be tested for accuracy. Confusion Matrix and Statistics shows that the model has an accuracy rate of 93.4%. Then to evaluate the accuracy of the model, for example the predictions from the model want to be equated with the actual data in the last 10 (ten) observations, the confusion matrix can be visualized. In the last 10 (ten) observations, this model succeeded in predicting exactly 9 out of 10. Then a ROC Curve and Variable Importance were made from the model to describe the variables that had the greatest influence on the model.

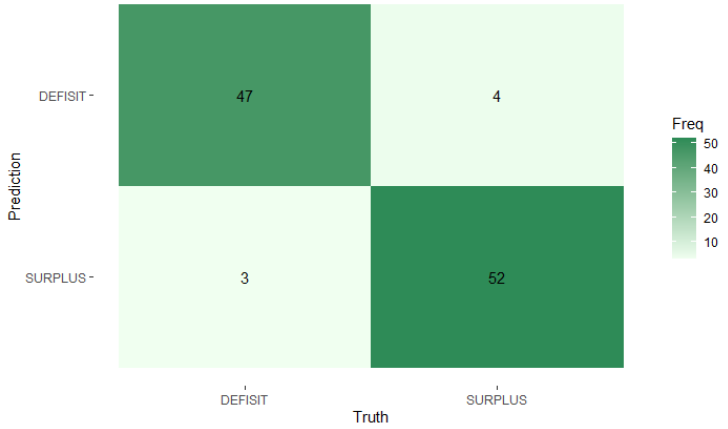

Image 6 Confusion Matrix

The graph of Overall Importance per variable shows that land area has the highest level of importance in the model, higher than the population variable which is in second position and the level of disaster risk is in third position.

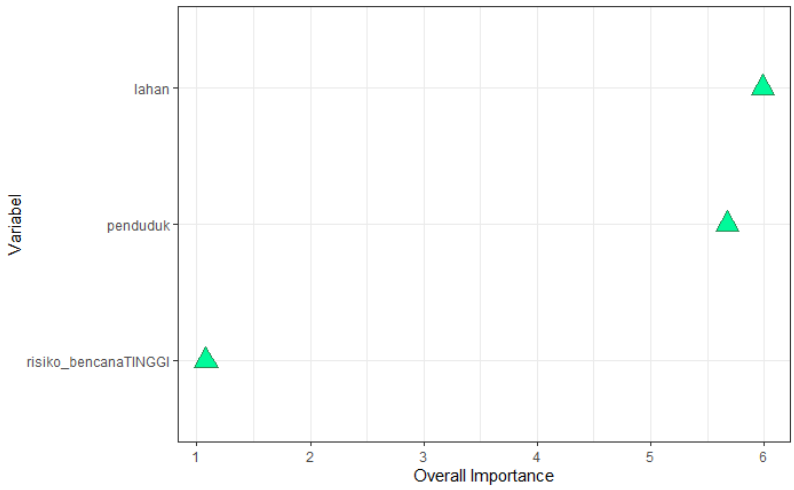

Image 7 Overall Importance Graph

After evaluating the prediction model obtained, dummy data can be used to predict the food security or vulnerability of the dummy data. Suppose you want to predict the condition

Ignatius Alvin Krisnugraha[1], Kgs Muhammad Benyamin Azhary[2]
DOI: https://doi.org/10.54443/sj.v2i1.109

of food security in a dummy area with a variable population of 100,000 people, a land area of 10,000 hectares, and a high disaster risk class, the result from that dummy conditions will give surplus of food availability for the result.

```
data_dummy <- data.frame(penduduk = 100, lahan = 10, risiko_bencana = "TINGGI")
predict(modelfit, data_dummy, type = "raw")
```

```
[1] SURPLUS
Levels: DEFISIT SURPLUS
```

Image 8 Predict using dummy conditions process

Cluster Models

The variables used in the formation of the prediction model are then used in forming the cluster model, but without removing the variable amount of rice production, because multicollinearity has no effect on the clustering process. The data used in the process of forming this cluster is data for 2020 as many as 353 rows of data: total population, area of harvested land, amount of rice production, and disaster risk index/score.

The next process needed is data filters and data scaling. The initial stage in clustering is selecting the variables to be used. Seeing that the data types used are all numbers with different ranges, it is necessary to carry out the normalization process. After obtaining normalized data, the next step is to find the optimal number of clusters. The method used to find the optimal number of clusters is the elbow method. Based on the curve of the elbow method, it is known that the optimal number of clusters is k=2 or k=4. Because the elbow curve does not show the optimal number of clusters clearly, the gap statistics method is used.

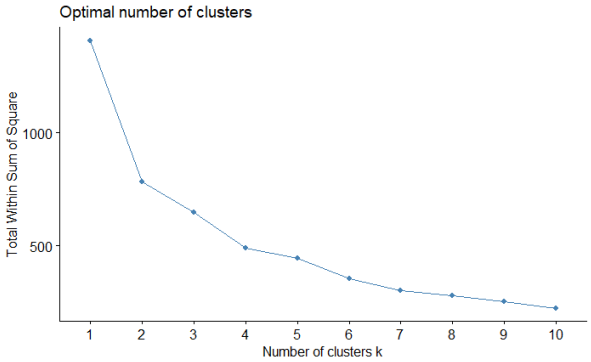

Image 9 Elbow Method Graph

Strengthening the results given by the elbow method, the gap statistical method ensures that the optimal number of clusters recommended by machine learning is k=2, so the next process is to group districts/cities into 2 (two) clusters.

SINOMICS JOURNAL
International Journal of Social Science, Education, Communication and Economics
ISSN (e): 2829-7350 | ISSN(p): 2963-9441

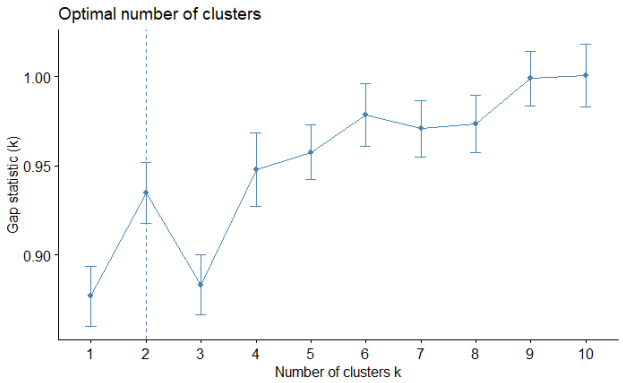

Image 10 Gap Statistic Method Graph

The clustering process is carried out with the 4 variables that have been determined earlier, using k=2. Then make a data visualization in the form of a scatter plot to see the distribution of data clusters per Regency/City. Scatter plots depict clustered areas, which can be said for areas clustered in blue means good food availability, and areas clustered in red have food vulnerability.

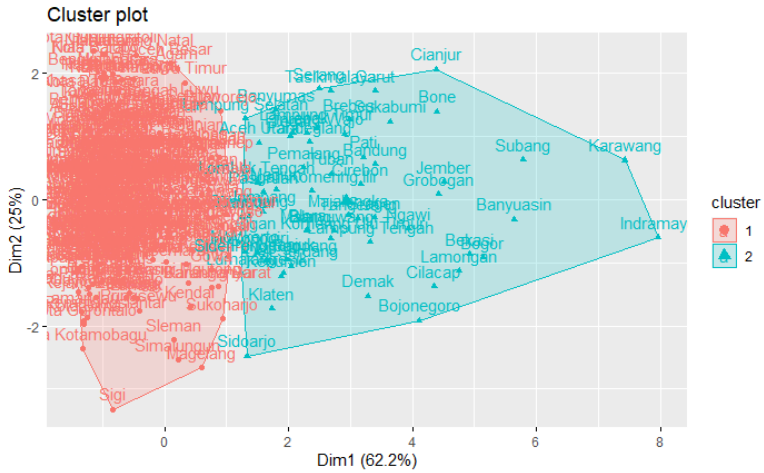

Image 11 Scatter Plot of Data Clusters per Regency/ City

Descriptive (Visualization)

Visualization of food security data will be presented in the form of choropleths. Therefore, the visualization will be sorted based on the surrounding area.

Ignatius Alvin Krisnugraha[1], Kgs Muhammad Benyamin Azhary[2]
DOI: https://doi.org/10.54443/sj.v2i1.109

Image 12 Food Security Conditions Choropleths

CONCLUSION

Through this research several conclusions were obtained, like the area of paddy harvested land, population, and disaster risk sequentially have a significant influence from the highest to the lowest on food security in Indonesia. Using secondary data in variables used for 2020, a model can be formed that can be used to predict the condition of food security fairly accurately (accuracy level > 90%). In addition, Regencies/ Cities in Indonesia when clustered based on data on the area of harvested land, population, amount of rice production, and disaster risk can optimally be divided into 2 clusters (results from the elbow method and the gap statistical method).

REFERENCES

Agricultural Research and Development Agency, Ministry of Agriculture. (2005). Prospek dan Arah Pengembangan Agribisnis Padi. 1-10.

Data Center for Agricultural and Information Systems, Ministry of Agriculture. (2020). Statistik Lahan Pertanian Tahun 2015-2019. Jakarta: Pusat Data dan Sistem Informasi Pertanian, Sekretariat Jenderal - Kementerian Pertanian.

Denis, D. J. (2020). Univariate, Bivariate, and Multivariate Statistics Using R. John Wiley & Sons, Inc.

Directorate of Food Crops, Horticulture and Urban Statistics, Ministry of Agriculture. (2022). Luas Panen dan Produksi Padi di Indonesia 2021. Jakarta: Badan Pusat Statistik.

Government of the Republic of Indonesia. (2012). Undang-Undang Republik Indonesia Nomor 18 Tahun 2012 tentang Pangan.

Hapsari, N., & Rudiarto, I. (2017). Faktor-Faktor yang Mempengaruhi Kerawanan dan Ketahanan Pangan dan Implikasi Kebijakannya di Kabupaten Rembang. Jurnal Wilayah dan Lingkungan, 125-140.

ISSN (e): 2829-7350 | ISSN(p): 2963-9441

Ministry of Agriculture's Food Security Agency. (2020). Panduan Penyusunan Peta Ketahanan dan Kerentanan Pangan (Food Security and Vulnerability ATLAS/FSVA) Provinsi 2020. Jakarta: Ministry of Agriculture's Food Security Agency.

Muryono, S., & Utami, W. (2020). Pemetaan Potensi Lahan Pertanian Pangan Berkelanjutan Guna Mendukung Ketahanan Pangan. Jurnal Agraria dan Pertanahan, 6, 201-218.

Tambunan, N. ., Aprilia, S. ., & Pangesti Rahayu, N. . (2022). Study Literature: Dampak Kenaikan Bbm Bagi Perekonomian Rakyat. Sibatik Journal: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, Dan Pendidikan, 2(1), 329–336. https://doi.org/10.54443/sibatik.v2i1.550

Zafira, A. ., Kustiawati, D. ., Fajria Putri Noor, J. ., & Farhan Sopyan, M. . (2022). Library Research: Elastisitas Penawaran Terhadap Beberapa Bahan Pangan. Sibatik Journal: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, Dan Pendidikan, 2(1), 115–120. https://doi.org/10.54443/sibatik.v2i1.506

Ignatius Alvin Krisnugraha[1], Kgs Muhammad Benyamin Azhary[2]
DOI: https://doi.org/10.54443/sj.v2i1.109