Information System of Agricultural Commodities Mapping Based on Machine Learning

D Nur¹, C Riyanti², and M Olivya³

¹,²,³Department of Electrical Engineering, Politeknik Negeri Ujung Pandang, Jalan Perintis Kemerdekan Km, 10 Talamanrea Makassar 90245, South Sulawesi Indonesia.

Abstract. Climate change that occurs from year to year affects the agricultural sector. People who work in agriculture need to know the compatibility of plants with climate conditions in the region. Back Propagation Neural Network (BPNN) is a multilayer Artificial Neural Network (ANN) that is used to train neural networks with input in the form of precipitation, relative humidity, and temperature data. The training produced a model that was able to classify climate data based on plant growth requirements. The model for Soybean got the highest average accuracy of 96.53% and 90.87% for Rice. Variables that influence training include the number of neurons in the hidden layer, the value of learning rate, the number of folds, and the number of epochs. Prediction results generated from the model are used as a reference to display markers on maps that can be accessed by users.

1 Introduction
In the national economic development of Indonesia, the agricultural sector has an important role. This is because Indonesia is an agrarian country with a large proportion of the population working or making a living in farming [1]. In Indonesia, the area of land that can be cultivated as agricultural land is 100.7 million hectares. Divided into wetlands (rice fields); dry land for seasonal plants; annual crop dry land [2]. The three types of land must be planted with suitable crops so that optimal agricultural yields. Suitability of plants and agricultural land is seen in the conditions for growing plants, for example is climate [3]. Climate change is an unavoidable thing. This happens due to an increase in the average temperature of the earth's surface from year to year. Climate change that occurs will have an impact on various aspects of life, including the agricultural sector. With the influence of climate on plant growth, people who work in agriculture need to know the suitability of plants in climatic conditions in the region [4].
To find out whether plants can live in accordance with a particular climate, we need growth conditions and weather information from several decades with monthly average values and year-round distribution patterns [5]. Climate elements have an influence on plant growth including air temperature (°C), humidity (%), and rainfall (mm). The data that has been collected can be analysed by machine learning to be able to find patterns and similarities in a set of data with certain techniques or methods [4].
Some research study conducted in agriculture by utilizing climate data, which is an element of plant growth. Heksaputra et. Al., [4] conducted research using the Bayes theorem method that can classify data based on the value of probability so that it can determine the good and bad growth in certain climatic conditions. The research of Kaunang et al., [6] used the Data Mining and Machine Learning technique to create a food crop prediction model in North Sulawesi province based on climate/weather using the Decision Tree J48 algorithm. However, the two studies are only in the form of calculation and prediction modelling, which data obtained is not easy to understand specifically for farmers. For that the
development of web-based information systems is required so that the information generated can be more easily understood by those who need it. This study applies the Neural Network Backpropagation algorithm to process precipitation, relative humidity, and temperature data in South Sulawesi Province, especially in Maros and Pangkajene Kepulauan (Pangkep) Regency. The results of this study are used for create an information system of agricultural commodities mapping to facilitate customers in determining suitable plants in certain areas.

2 Literature Review

2.1 Data Processing

Data pre-processing aims to: (1) make it easier to understand the data in order to facilitate the selection of appropriate data mining techniques and methods, (2) improve data quality for better data mining result, and (3) improve the efficiency and ease of the data mining process [7].

Data pre-processing methods include:

a) Data Cleaning
Before conducting data mining, dull data must be cleaning by filling in blank values and/or delete noisy data and/or removing outliers and/or recovering inconsistencies of the data.

b) Data Reduction
Data can be reduced to a much smaller level with the integrity of the original data maintained. Thus, analysis and mining of the reduced data will be more efficient, and the results will be the same (almost the same) as the results of the analysis conducted on the First Data.

c) Adding Data
Adding new dimensions sometimes desires to be done to simplify (not complicate) the data mining process.

d) Data Normalization
There are several techniques that can be applied to normalize data, one of them is the Min-Max Method. The equation used in the min-max method is shown in Equation 1 where $x_{\text{min}}$ is the minimum value, $x_{\text{max}}$ is a maximum value, $x_i$ is input value, $new_{x_{\text{max}}}$ is a new maximum value, $new_{x_{\text{min}}}$ is the new minimum value, dan $x'$ is normalized data.

\[ x' = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \left( new_{x_{\text{max}}} - new_{x_{\text{min}}} \right) + new_{x_{\text{min}}} \]  

(1)

2.2 Back Propagation Neural Network

One type of ANN algorithm is Back Propagation Neural Network (BPNN) which is a multilayer ANN that is used for train neural networks. Neural network training is the process of finding a set of weights and bias values so that for a set of inputs, the output produced by a neural network is very close to some known target value. Once it has weight and bias values, then it can be applied to new input values where the output value is unknown, and make predictions [7]. The BPNN algorithm consists of 3 phases [8]: forward propagation, backpropagation, and weights modification.

Figure 1. Back Propagation Neural Network with Single Hidden Layer
2.3 K-Fold Cross-Validation

K-Fold Cross Validation is one technique to estimate the error rate of testing [9]. The way of K-fold cross-validation works is by grouping a dataset consisting of separate training data and test data, then repeating the testing process for K times.

2.4 Google Maps API

Google Maps is a service by Google that offers user-friendly mapping technology. This service can be accessed through the site http://maps.google.com or http://maps.google.co.id for Google Maps Indonesia. On this site, we can see geographical information in all regions on earth [10].

Google Maps API (Application Programming Interface) is an application feature used by Google to facilitate users who want to integrate Google Maps into their websites by displaying their own data points [10].

![Google Maps Marker](developers.google.com)

**Figure 2. Google Maps Marker**

Google Maps API also has several features that can be utilized, one of which is a custom marker. This feature allows users to change the default marker to the desired image or icon so that the map display looks more interactive.

![Google Maps Custom Marker](developers.google.com)

**Figure 3. Google Maps Custom Marker**

3 Methodology

3.1 System Design

Design describes the system to be created. The first step is data collection. The dataset was taken from the National Aeronautics and Space Administration (NASA) consisting of rainfall, relative humidity, and temperature data from 2007-2018. The dataset is used as training and testing data. The ANN training model is used to predict new data labels for each sub-district in Maros and Pangkep districts and then stored in a database. Mapping is displayed in a web-based information system that is connected to the Google Maps API by retrieving data from a database. Prediction label data is a reference to display the Google Maps marker. This information system can be accessed by users.
3.2 System Flowchart

In general, the stages in this study are data processing, determination of input data and target data, ANN training process, ANN testing, and saving a model. For pre-processing data, the stages include cleaning up the data, adjusting climate data with requirements for plant growth, labeling and normalizing the data. Labels are given to the compatibility of climate data with requirements for plant growth shown in Table 1.

| Commodity | Precipitation (mm/month) | Relative Humidity (%) | Temperature (°C) | Planting Duration |
|-----------|--------------------------|-----------------------|------------------|------------------|
| Rice      | 175-500                  | 33-90                 | 19-27            | 4 months         |
| Corn      | 85-400                   | 78-83                 | 14-30            | 4 months         |
| Soybean   | 100-400                  | 70-100                | 23-30            | 3 months         |
| Shallots  | 25-208                   | 80-90                 | 25-32            | 2-3 months       |
| Chili     | <200                     | 60-80                 | 25-27            | 4 months         |

Source: [12][13][14][4][15]
The stages to conduct ANN training includes determining the number of neurons in the hidden layer, K (number of folds), learning rate value, and the number of epochs. Figure 5 shows the ANN Backpropagation flow chart used in this study. In addition, the ANN pre-processing and training functions are explained in Figures 6 and 7.

3.3 Testing

The test aims to determine the accuracy of the model from ANN Backpropagation training to classify plants based on climate data. The test scenarios carried out are as follows:

a. Testing the effect of a combination of learning rate and the number of neurons in a hidden layer on the accuracy and MSE values. The value of learning rate tested are 0.001, 0.01, 0.1, 0.2, 0.3, 0.5, 0.8 and 1. While the number of neurons in the hidden layer is 1, 2, 3, 4, and 5.

b. Testing the effect of the number of K in k-fold cross-validation on the accuracy value and MSE using a combination of learning rate values and the number of neurons in hidden layer that have the highest accuracy in the first scenario. The number of K tested is 2, 3, 4, 5, 6, 7, 8, 9, and 10.

c. Testing the effect of the number of epochs on the accuracy value and MSE using the number of K on k-fold which has the highest accuracy results in the second scenario. Testing is done by adding the number of epochs from 100 to 500 and reducing the number of epochs from 100 to 10.

d. Accuracy testing used each model of rice, corn, soybean, shallot, and chili on the sub-districts of Maros and Pangkep.

4 Result and Discussion

4.1 Testing the Combination of Learning Rate and Neurons in Hidden Layer

The data used for the rice model training is Lau sub-district data in 2007-2018. The initial conditions when testing the first scenario by using random_seed = 5; n_input = 3; n_output = 2; n_epoch = 100; n_folds = 2.

The random seed is used to initialize the weight of the network. N_input has 3 neurons that represent perceptron, humidity, and temperature. While n_output has 2 neurons that represent the label to classifying data into 2 classes, 0 and 1. Epoch is the amount carried out by the algorithm to conduct training on the dataset. N_folds are used to repeat the training with the training data and testing data that vary in each fold.

![MSE Value of Combination Between Learning Rate (LR) and Number of Hidden Layer Neurons](image)

Figure 8. MSE Value of Combination Between Learning Rate (LR) and Number of Hidden Layer Neurons

Figure 8 represents the results of the first scenario test. From the graph, it can be seen that the MSE value fluctuates at the learning 0.001. As for the learning rate 0.01, the MSE value has increased although it is not significant. Learning rates 0.1, 0.2, 0.3, 0.5, 0.8 and 1 have MSE values that are not
much different and decrease as the number of neurons in the hidden layer increases. The MSE value of neuron 1 decreases in neuron 2 and increases in neuron 3.

Table 2. Accuracy Value of Combination Between Learning Rate (LR) and Number of Hidden Layer Neurons

| n_hidden | LR 0.001 | LR 0.01 | LR 0.1 | LR 0.2 | LR 0.3 | LR 0.5 | LR 0.8 | LR 1 |
|----------|----------|---------|--------|--------|--------|--------|--------|------|
| 1        | 68.75%   | 59.72%  | 21.53% | 15.28% | 13.19% | 9.03%  | 9.03%  | 8.33%|
| 2        | 31.25%   | 61.11%  | 90.28% | 86.81% | 86.81% | 85.42% | 59.72% | 55.56%|
| 3        | 68.75%   | 38.19%  | 6.94%  | 5.56%  | 6.25%  | 20.14% | 26.39% | 50.00%|
| 4        | 68.75%   | 65.28%  | 48.61% | 51.39% | 54.86% | 61.11% | 74.31% | 65.28%|
| 5        | 69.44%   | 62.50%  | 88.19% | 78.47% | 72.92% | 70.83% | 76.39% | 81.25%|

Figure 9. Graph of Accuracy Value of Combination Between Learning Rate (LR) and Number of Hidden Layer Neurons

Figure 9 represents the test results in Table 2. The accuracy increases in neuron 2 except for the learning rate of 0.001. The accuracy of the hidden layer with 3 neurons tends to decrease and then increase again in 4 and 5 neurons. The highest accuracy is obtained by a combination of a learning rate of 0.1 with 2 neurons. The accuracy obtained is 90.28% and MSE 0.0546. This combination will be used for testing the second scenario.

4.2 Testing the Number of Folds
The Second Condition of Rice Model Training using random_seed = 5; n_input = 3; n_output = 2; n_epoch = 100; n_hidden = 2; LR = 0.1.

Table 3. Accuracy and MSE values based on the number of folds

| n_fold | MSE    | Mean Accuracy |
|--------|--------|---------------|
| 2      | 0.0546 | 90.28%        |
| 3      | 0.0532 | 89.58%        |
| 4      | 0.0453 | 90.05%        |
| 5      | 0.0486 | 90.28%        |
| 6      | 0.0462 | 90.14%        |
| 7      | 0.0458 | 89.93%        |
| 8      | 0.0447 | 90.28%        |
| 9      | 0.0437 | 89.84%        |
| 10     | 0.0429 | 90.12%        |
Based on the test results in Table 3, it can be seen that the greater the number of folds, the smaller the MSE value. The best accuracy is 90.28% obtained with the number of folds 2 and 8 with MSE 0.0546 and 0.0447. The number of folds used in the third scenario is the one that has a smaller MSE.

### 4.3 Testing the Number of Epochs

The third condition of rice model training is testing the number of epochs by using random_seed = 5; n_input = 3; n_output = 2; n_hidden = 2; LR = 0.1 and K_fold = 8.

| n_epoch | MSE      | Mean Accuracy |
|---------|----------|---------------|
| 400     | 0.041709 | 77.08%        |
| 300     | 0.042235 | 82.54%        |
| 200     | 0.043025 | 87.80%        |
| 100     | 0.044716 | 90.28%        |
| 90      | 0.045025 | 90.38%        |
| 80      | 0.045394 | 90.38%        |
| 70      | 0.045845 | 90.68%        |
| 60      | 0.046422 | 90.87%        |
| 50      | 0.047206 | 90.48%        |
| 40      | 0.048383 | 90.28%        |

Based on the test results in Tables 4, the best accuracy is obtained by reducing the number of epochs to 60 where the accuracy is 90.87% and MSE is 0.046.

### 4.4 Plant Models

The testing with the first to the third scenario was conducted for each plant commodities in this study. Each parameter with the best accuracy will be used for plant models. This model will be used to classifying climate data for the sub-district in Maros and Pangkep. The parameter values used in the model of plants can be seen in Table 5.

| Test Parameters | Parameter Values |
|-----------------|------------------|
| Rice            | 90.87%           |
| Corn            | 88.43%           |
| Soybean         | 96.53%           |
| Shallots        | 86.81%           |
| Chili           | 89.93%           |

### 4.5 Test the Plant Models Toward the Sub-District Group

Climate data for each sub-district is grouped according to the similarity of the data. For Maros Regency, 14 sub-districts are grouped into 5 groups, whereas for Pangkep Regency, 12 sub-districts are grouped into 6 groups. The division of these groups can be seen in Table 6.
The groups were tested with a trained plant model. The accuracy results obtained can be seen in Tables 7 and 8.

Table 7. Accuracy of Plant Models in Maros Regency

| Group | Accuracy | Rice | Corn | Soybean | Shallot | Chili |
|-------|----------|------|------|---------|---------|-------|
| 1     | 91.67%   | 87.50% | 96.53% | 90.97% | 87.50% |
| 2     | 91.67%   | 84.03% | 97.92% | 87.50% | 86.11% |
| 3     | 92.36%   | 84.03% | 96.53% | 88.19% | 88.19% |
| 4     | 88.89%   | 84.72% | 93.75% | 88.19% | 84.72% |
| 5     | 81.25%   | 85.42% | 93.75% | 89.58% | 84.72% |
| Mean  | Accuracy | 89.17% | 85.14% | 95.70% | 88.89% | 86.25% |

Table 7 shows the accuracy values when testing plant models for data from sub-districts in Maros Regency. Based on the results of testing 5 plant models, the soybean model has the highest accuracy with an average accuracy of 95.70% and the lowest accuracy is the corn model with an average accuracy of 85.14%.

Table 8. Accuracy of Plant Models in Pangkep Regency

| Group | Accuracy | Rice | Corn | Soybean | Shallot | Chili |
|-------|----------|------|------|---------|---------|-------|
| 1     | 91.67%   | 87.50% | 96.53% | 90.97% | 87.50% |
| 2     | 75.00%   | 86.11% | 92.36% | 92.36% | 77.78% |
| 3     | 75.00%   | 81.25% | 88.89% | 73.61% | 97.78% |
| 4     | 77.78%   | 86.81% | 94.44% | 86.11% | 80.56% |
| 5     | 79.86%   | 90.97% | 91.67% | 92.36% | 77.78% |
| 6     | 86.11%   | 89.58% | 95.14% | 86.11% | 89.58% |
| Mean  | Accuracy | 80.90% | 87.04% | 93.17% | 87.15% | 84.72% |

Table 8 shows the accuracy values when testing plant models for data from sub-districts in Pangkep Regency. Based on the results of testing 5 plant models, the soybean model also has the highest accuracy with an average accuracy of 93.17% and the lowest accuracy is the corn model with an average accuracy of 80.90%.

The accuracy while training in the rice model is 90.87%. However, the rice model turned out to have the lowest average value when tested with data from the sub-district group in Pangkep Regency. In Table 8 it can be seen that Groups 2, 3, 4, and 5 show low accuracy values. This indicates that the model for rice experiences overfitting so that it is only able to predict the value if the data provided has a pattern that is quite similar to the training data.
4.6 Plan Models

The testing with the first to the third scenario was conducted for each plant commodities in this study. Each parameter with the best accuracy will be used for plant models. This model will be used to classifying climate data for the sub-district in Maros and Pangkep. The parameter values used in the model of plants can be seen in Table 9.

Table 9. Parameter Values for Model of Plants

| Test Parameters | Parameter Values |
|-----------------|------------------|
| Rice | Corn | Soybean | Shallots | Chili |
| random_seed | 5 | 1 | 5 | 1 | 1 |
| n_input | 3 | 3 | 3 | 3 | 3 |
| n_output | 2 | 2 | 3 | 2 | 2 |
| n_hidden | 2 | 2 | 5 | 5 | 4 |
| LR | 0.1 | 1 | 0.8 | 0.3 | 0.3 |
| K_fold | 2 | 4 | 2 | 2 | 3 |
| n_epoch | 60 | 100 | 80 | 80 | 90 |
| Mean Accuracy | 90.87% | 88.43% | 96.53% | 86.81% | 89.93% |
| MSE | 0.046 | 0.0995 | 0.0215 | 0.0943 | 0.0511 |

4.7 Mapping Result.

The markers displayed on the map are adjusted to the prediction results, see figure 10. The marker will only appear if the data matches the requirements, i.e. if there is a label 1 in a row during the planting time for each plant commodities in the same year. For example, rice needs four months of planting time. Then the marker will only appear if the data in the sub-district has label 1 four times in a row. If there are sub-districts that do not meet these requirements, then markers in the sub-district are not displayed.

Figure 10. Agricultural Commodities Mapping Information Systems Interface

This information system has two types of filters to make it easier for users to find the specific information that they want. Filters available are plant type and year. Each marker has an info window that will appear when the marker is clicked. The info window provides information on the amount of perceptron, average humidity and temperature of the month that meets the requirements of the selected plant growth. The info window display is shown in Figure 11.
Figure 11. Info Window Display

5. Summary
The Agricultural commodity mapping information system can map plants according to the prediction results from the rice, corn, soybean, shallots, and chili models based on labels according to the planting time needed by the plants. The results obtained indicate that the districts of Maros and Pangkep, suitable for rice and soybean cultivation, are seen with an average accuracy of 90.8% and 96.53%. While corn, chili, and onion can be alternatives that are planted during the dry season.

6. References
[1] F. Latief, “Pengaruh Curah Hujan, Kelembapan Udara, dan Luas Panen terhadap Hasil Produksi Jagung di Sumatera Utara,” Universitas Sumatera Utara, 2018.
[2] Pujiasmanto, B., & Yunus, A. (2018). Potensi Keaneragaman Hayati dalam Mendukung Indonesia Sebagai Lumbung Pangan Dunia. Makalah yang disajikan dalam Seminar Nasional Dalam Rangka Dies Natalis ke 42 UNS. Fakultas Pertanian Universitas Sebelas Maret. Solo, 12-13 April 2018.
[3] N. K. Dewi, “Kesesuaian Iklim Terhadap Pertumbuhan Tanaman,” J. Ilmu-Ilmu Pertan., vol. 1, no. 2, pp. 1–15, 2005.
[4] D. Heksaputra, Y. Azani, Z. Naimah, and I. Iswari, “Penentuan Pengaruh Iklim Terhadap Pertumbuhan Tanaman dengan Naïve Bayes,” J. Fak. Huk. UII, pp. 34–39, 2013.
[5] As-syakur, A. R., Suarna, I. W., Rusna, I. W., & Dibia, I. N. (2011). Pemetaan Kesesuaian Iklim Tanaman Pakan Serta Kerentanannya Terhadap Perubahan Iklim Dengan Sistem Informasi Geografi (SIG) Di Provinsi Bali. Pastura, 1(1), 9–15.
[6] F. J. Kaunang, R. Rotikan, and G. S. Tulung, “Pemodelan Sistem Prediksi Tanaman Pangan Menggunakan Algoritma Decision Tree Crop Prediction System Using Decision Tree Algorithm,” Cogito Smart J., vol. 4, no. 1, pp. 213–218, 2018.
[7] W. Budiharto, MACHINE LEARNING & Computational Intelligence. Yogyakarta: Penerbit ANDI, 2016.
[8] G. Dhaneswara, V. S. Moertini, and I. R. HS, “Jaringan Saraf Tiruan Propagasi Balik Untuk Klasifikasi Data,” Integral, vol. 9, no. 3, pp. 117–131, 2004.
[9] S. Y. Riska, L. Cahyani, and M. I. Rosadi, “Klasifikasi Jenis Tanaman Mangga Gadung dan Mangga Madu Berdasarkan Tulang Daun,” J. Buana Inform., vol. 6, no. 1, pp. 41–50, 2015.
[10] N. Muslim and A. Sunyoto, “Sistem Informasi Geografis Berbasis Petaan Potensi Panas Bumi Di Indonesia Menggunakan Google Maps,” J. Dasi, vol. 13, no. 2, pp. 60–64, 2012.
[11] G. Maps, “Adding a Google Map with a Marker to Your Website,” 2019. [Online]. Available: https://developers.google.com/maps/documentation. [Accessed: 16-Jun-2019].
[12] “PADI (Oryza Sativa).” Kantor Deputi Menegristek Bidang Pendayagunaan dan Pemasyarakatan Ilmu Pengetahuan dan Teknologi, Jakarta, pp. 1–16, 2000.
[13] B. I. Cahya, Adiwijaya, and U. N. Wisesty, “Prediksi Kalender Tanam Berdasarkan Curah Hujan Dengan Menggunakan Jaringan Syaraf Tiruan dan Algoritma Genetika,” e-Proceeding Eng., vol. 2, no. 2, pp. 6174–6181, 2015.
[14] G. A. Pryanto, K. D. Hartomo, and M. A. I. Pakereng, “Perancangan Kalender Tanam Berdasarkan Data Klimatologi Menggunakan Metode Forecasting Holt-Winters,” Universitas Kristen Satya Wacana, 2013.
[15] H. Imtiyaz, B. H. Prasetio, and N. Hidayat, “Sistem Pendukung Keputusan Budidaya Tanaman Cabai Berdasarkan Prediksi Curah Hujan,” J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 1, no. 9, pp. 733–738, 2017.