Contents lists available at https://citedness.com/index.php/jdsi

Data ScienceInsights

Journal Page is available tohttps://citedness.com/index.php/jdsi

Research article

Application of Decision Tree Algorithm for Classification of Rice Yields in Sumatra

Wily Chandra

Institute of Business and Technology Pelita Indonesia, Pekanbaru, Riau
e-mail: wily.candra@student.pelitaindonesia.ac.id

ARTICLE INFO

Article history:
Received June 25, 2024
Revised July 02, 2024,
Accepted July 27, 2024
Available online August 01, 2024

Keywords:
Aplication
Classification
Decision Tree
Rice Crop
Sumatera

Please cite this article in IEEE style as:
wily candra, "Application of Decision Tree Algorithm for Classification of Rice Yields in Sumatra", Data Science Insights, vol. 2, no. 2, pp. 96–104, Aug. 2024.

ABSTRACT

Rice is the main food crop in Indonesia, most of the agricultural sector in Indonesia is dominated by rice farming including on the island of Sumatra. A common problem that arises is how to find out the areas that produce the most rice each year on the island of Sumatra. This study aims to classify the areas that produce the most rice on the island of Sumatra. The dataset used in this study was taken from Kaggle with a total of 225 data and will be tested using the Decision Tree algorithm and several other algorithms. For data visualization, Tableau will be used to see which areas produce the most rice on the island of Sumatra. By using the research method using the Decision Tree algorithm, an accuracy of 97.78% was obtained with a data split of 0.8 for training data and 0.2 for testing data.

Correspondence:
Wily Chandra
Institute of Business and Technology Pelita Indonesia, Pekanbaru, Riau

1. Introduction

Food is one of the basic and fundamental needs for humans that must be met. Fulfillment of food needs is carried out as an effort so that food is always available at all times and can also be reached by all levels of society.[1]. Indonesia is an agricultural country where agriculture plays an important role because it is rich in natural resources, such as food crops and horticulture.[2][3][4][5]. As one of the industrial sectors, agriculture is part of the work and also the fulfillment of community needs such as staple food needs. Fertile land is a supporting factor for the development of plant growth and Sumatra Island has a fairly large and fertile land for agricultural areas in this case food crops.[6][7]. Rice is a food crop consumed by the majority of Indonesian people. Currently, the dependence of Indonesian people on plants is still very large, especially rice, which is the staple food of Indonesian people. Sumatra Island is one of the largest rice producing areas in Indonesia.

To get good quality rice in the future, of course, we must look at the condition of the harvest in previous years, besides that, climate conditions and temperature also greatly affect the growth and development of rice plants. There are several factors that drive the harvest of the rice plant,

namely, temperature, rainfall, humidity and also the area of land. If these factors are good, the harvest obtained will be of good quality and also the harvest will be abundant.[8][9][10].

In this study, the problem raised is which areas produce the most rice on the island of Sumatra, how much rainfall and temperature in the area so that it can produce rice in large quantities, so it is necessary to conduct research using datasets from previous years' harvests as a reference for classifying the areas that produce the most or the largest rice on the island of Sumatra.

To perform this classification, the Decision Tree algorithm method is used. Decision Tree is a tree structure, where each leaf node represents a certain class data group. The topmost node level is called the root, which has the greatest influence on a class, and can be the first rule that influences[11][12]. In this study, the accuracy results were obtained at 97.78 with a data split ratio of 0.8 (training data) and 0.2 (testing data). In addition to the Decision Tree, tests were also carried out with other algorithms such as Deep Learning.[13]and Naïve Bayes[14][15]to compare which method is best.

The purpose of this study is to increase insight into the agricultural conditions, especially for rice plants on the island of Sumatra and the largest rice-producing areas on the island of Sumatra. In addition, this paper also aims to cluster the largest rice-producing areas in Indonesia.

2. Research methodology

In this study, the method used is as shown in the following image.

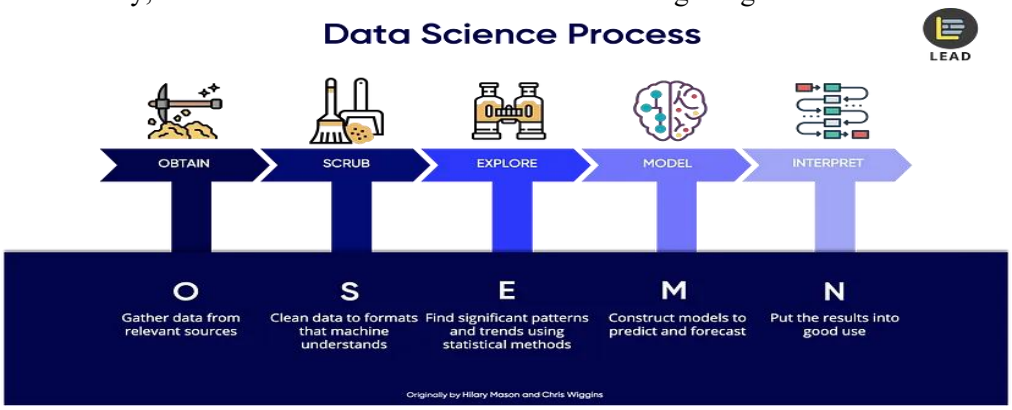


Figure 1. Data Science Process for Conducting Research

2.1 Data Obtain

Data Obtain is a collection of data on Rice Plants on Sumatra Island which is taken from a trusted source, namely Kaggel as a source of obtaining relevant data to conduct this research because data accuracy is very important so that the data obtained has high quality so that it will produce good results too.

2.2 Scrub Data

Data Scrub is the cleaning of Rice Plant data on Sumatra Island that has been taken earlier so that there is no lost data or missing data so that each research step can run smoothly, data cleaning here uses the Excel application to scrub the data.

2.3 Exploration Data

After the data is cleaned, the next step is to explore the Rice Plant data on Sumatra Island. The purpose of this data exploration is to find out what factors support the abundance of

rice harvests in Sumatra, whether there are many or not. For data exploration, the Tableau application is used to explore rainfall, temperature, and also rice production groups on Sumatra Island.

2.4 Data Modeling

After completing the data exploration, the next step is to build a model to classify the selected data. The model development here uses rapidmainer to classify which areas produce the most rice from year to year.

2.5 Data Interpret

The final step in this study process is to present the results of the research that has been done. This research will be presented in the form of a scientific journal and will be presented to the supervising lecturer as a sign that the research has been carried out in accordance with existing rules.

3. Results of Data Science Research Steps

3.1 Data Obtain

Data Obtain is a collection of Rice Plant data on Sumatra Island taken from a trusted source, namely Kaggel as a source of obtaining relevant data to conduct this research because data accuracy is very important so that the data obtained has high quality so that it will produce good results too.

Figure 2. Dataset Used From Kaggle

Provinsi	Tahun	Produksi	Luas Panen	Curah hujan	Kelembapan	Suhu rata-rata
Aceh	1993	1329536	323589	1627	82	26.06
Aceh	1994	1299699	329041	1521	82.12	26.92
Aceh	1995	1382905	339253	1476	82.72	26.27
Aceh	1996	1419128	348223	1557	83	26.08
Aceh	1997	1368074	337561	1339	82.46	26.31
Aceh	1998	1404580	365892	1465	82.6	26.84
Aceh	1999	1478712	359817	1778	82.79	26.14
Aceh	2000	1486909	336765	1974.7	90.6	27.1
Aceh	2001	1547499	295212	1688.7	69.48	28.9
Aceh	2002	1314165	315131	1296.8	68.75	29.2
Aceh	2003	1246614	367636	1507.2	70.66	29.4
Aceh	2004	1350748	370966	1097	80.84	29.4
Aceh	2005	1411650	337893	710.5	79.5	26.8
Aceh	2006	1552078	320789	506.5	80.8	26.73
Aceh	2007	1556858	360717	1414	81.5	26.38
Aceh	2008	1402287	329109	1270.4	78.5	27
Aceh	2009	1533369	359375	1577	78.7	26.9
Aceh	2010	1788738	352281	1986	81.4	27.1
Aceh	2011	1772962	380686	1268	79.4	27.1
Aceh	2012	1582393	387803	1098	79.6	26.9
Aceh	2013	2331046	419183	1623.6	80.7	27
Aceh	2014	1820062	376137	2264.4	78.3	27.1
Aceh	2015	1956940	461060	1575	80	27.1
Aceh	2016	2180754	293067	1096	83.32	27.12
Aceh	2017	2478922	294483	1905.9	85.57	26.51
Aceh	2018	1751997	329515.78	1427.8	83.98	26.48

Figure 3. Data From Inside the Dataset Downloaded From Kaggle

In Figure 2. This is the dataset used to research the cases in this journal which was sourced from Kaggle before being downloaded with the name Sumatra Rice Plant Dataset, Indonesia.

In Figure 3. It is the contents of the Kaggle dataset that has been downloaded and in the dataset there is no lost data or missing data. There are several attributes contained in the dataset (1) Province (is the place/area of rice production in Sumatra), (2) Year (is the time recorded when harvesting), (3) Production (is the amount obtained/produced from year to year), (4) Harvested Area (the total area of land harvested in one year), (5) Rainfall (the amount/volume of rain that falls each year), (6) Humidity (is the humidity condition of the surrounding air), (7) Average Temperature (the temperature of the surrounding air)

3.2 Data Scrub

Data Scrub is the cleaning of Rice Plant data on Sumatra Island that has been taken earlier so that there is no lost data or missing data so that each research step can run smoothly, data cleaning here uses the Excel application to scrub the data.

Provinsi	Skala Produksi	Produksi	Luas Panen	Curah hujan	Kelembapan	Suhu rata-rata
Aceh	Menengah	1329536	323589	1627	82	26.06
Aceh	Menengah	1299699	329041	1521	82.12	26.92
Aceh	Menengah	1382905	339253	1476	82.72	26.27
Aceh	Menengah	1419128	348223	1557	83	26.08
Aceh	Menengah	1368074	337561	1339	82.46	26.31
Aceh	Menengah	1404580	365892	1465	82.6	26.84
Aceh	Menengah	1478712	359817	1778	82.79	26.14
Aceh	Menengah	1486909	336765	1974.7	90.6	27.1
Aceh	Menengah	1547499	295212	1688.7	69.48	28.9
Aceh	Menengah	1314165	315131	1296.8	68.75	29.2
Aceh	Menengah	1246614	367636	1507.2	70.66	29.4
Aceh	Menengah	1350748	370966	1097	80.84	29.4
Aceh	Menengah	1411650	337893	710.5	79.5	26.8
Aceh	Menengah	1552078	320789	506.5	80.8	26.73
Aceh	Menengah	1556858	360717	1414	81.5	26.38
Aceh	Menengah	1402287	329109	1270.4	78.5	27
Aceh	Menengah	1533369	359375	1577	78.7	26.9
Aceh	Tinggi	1788738	352281	1986	81.4	27.1
Aceh	Tinggi	1772962	380686	1268	79.4	27.1
Aceh	Menengah	1582393	387803	1098	79.6	26.9
Aceh	Sangat Tinggi	2331046	419183	1623.6	80.7	27
Aceh	Tinggi	1820062	376137	2264.4	78.3	27.1
Aceh	Tinggi	1956940	461060	1575	80	27.1
Aceh	Tinggi	2180754	293067	1096	83.32	27.12
Aceh	Sangat Tinggi	2478922	294483	1905.9	85.57	26.51
Aceh	Sangat Tinggi	1751996.9	329515.78	1427.8	83.98	26.48

Figure 4. Scrubbed Data

In this data scrubbing process, a Production Scale table is added in the data scrubbing process so that the data can be used to run in the algorithm that the author has chosen, and the author removes or deletes the year attribute table because the table does not have much influence on the objectives to be achieved in this study.

3.3 Exploration Data

After the data is cleaned, the next step is to explore the Rice Plant data on Sumatra Island. The purpose of this data exploration is to find out what factors support the abundance of rice harvests in Sumatra, whether there are many or not. For data exploration, the Tableau application is used to explore rainfall, temperature, and also rice production groups on Sumatra Island.

Figure 5. Average temperature of each province as a whole

From Figure 5. Looking at the results of this tableau visualization, it can be concluded that the highest average temperature is in the province of North Sumatra with an average temperature of 27,313°C and the lowest average temperature is in West Sumatra with an average temperature of 25,651°C

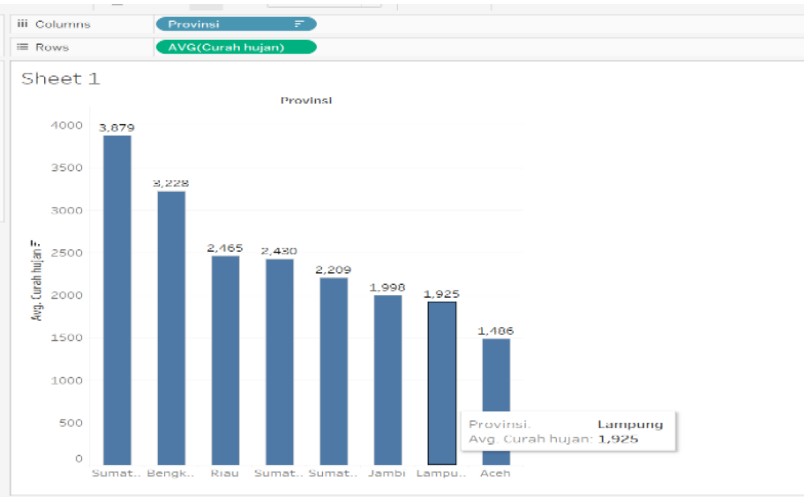


Figure 6. Average rainfall for each province as a whole

In Figure 6. Looking at the results of this tableau visualization, it can be concluded that the highest average rainfall is in West Sumatra province and the lowest average rainfall is in Aceh province.

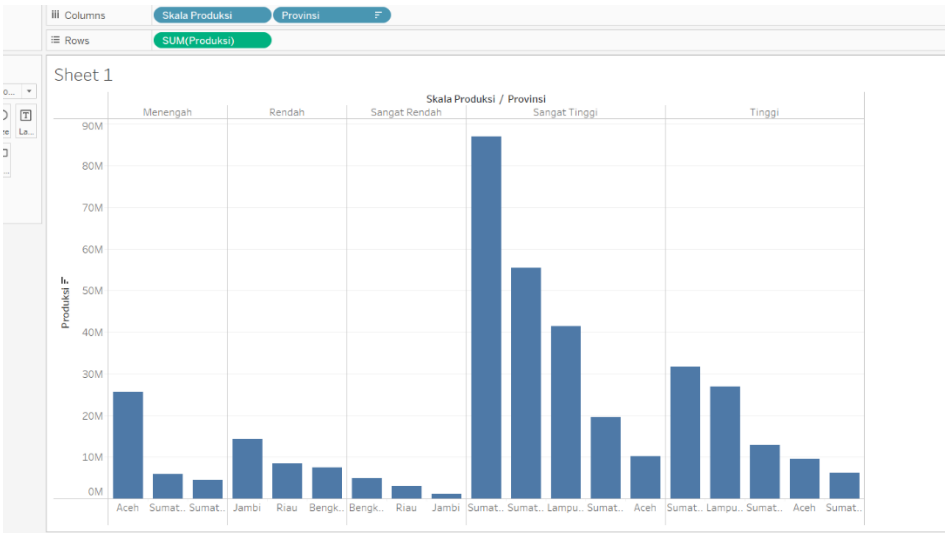


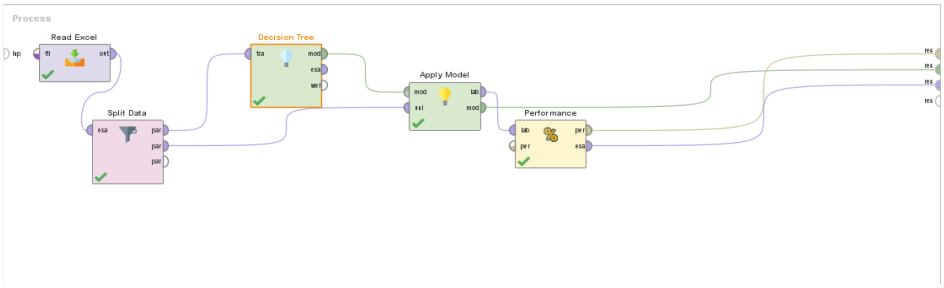
Figure 7. Harvest results from each province according to production scale

In Figure 7. Looking at the results of this tableau visualization, it can be concluded that the highest harvest is in North Sumatra province and the lowest harvest is in Jambi province. Based on rainfall and temperature, the harvest results are as in Figure 7. The conclusion is that in order for the harvest to be abundant, the air temperature should not be too cool or too hot and the rainfall should not be too high because rice can grow well in tropical weather.

3.4 Modeling

After completing the data exploration, the next step is to build a model to classify the selected data. The model development here uses rapidmainer to classify which areas produce the most rice from year to year.

1.) Decision Tree



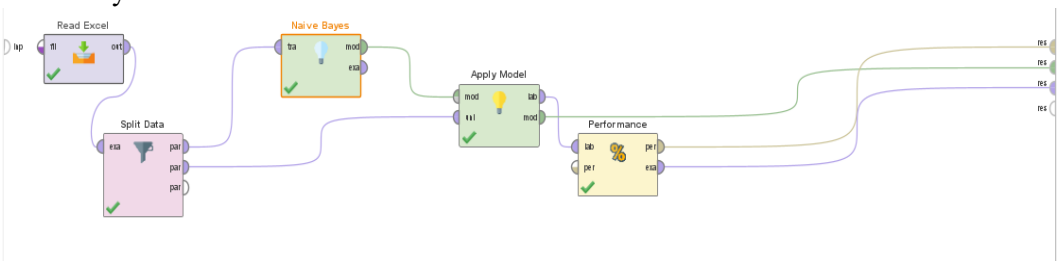
accuracy: 97.78%

	true Menengah	true Tinggi	true Sangat Tinggi	true Rendah	true Sangat Rendah	class precision
pred. Menengah	5	0	0	0	0	100.00%
pred. Tinggi	0	8	0	0	0	100.00%
pred. Sangat Tinggi	0	1	14	0	0	93.33%
pred. Rendah	0	0	0	11	0	100.00%
pred. Sangat Rendah	0	0	0	0	6	100.00%
class recall	100.00%	88.89%	100.00%	100.00%	100.00%	

Figure 8. Decision Tree algorithm and results

From the results shown in Figure 8. Using the decision tree method, the accuracy results were 97.78% with a data set division of 0.8 for training data and 0.2 for testing data. The components used in this algorithm are read excel to read the scrubbed dataset, then split data to divide the data into training data and testing data, then there is an algorithm that will be trained, namely the decision tree, apply mode to run the model, and performance to display the percentage of the method used.

2.) Naive Bayes



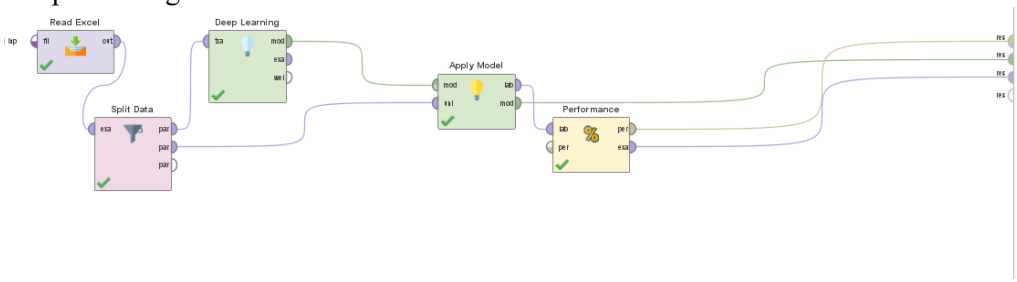
accuracy: 86.67%

	true Menengah	true Tinggi	true Sangat Tinggi	true Rendah	true Sangat Rendah	class precision
pred. Menengah	3	0	0	0	0	100.00%
pred. Tinggi	1	8	1	0	0	80.00%
pred. Sangat Tinggi	1	1	13	0	0	86.67%
pred. Rendah	0	0	0	10	1	90.91%
pred. Sangat Rendah	0	0	0	1	5	83.33%
class recall	60.00%	88.89%	92.86%	90.91%	83.33%	

Figure 9. Naïve Bayes algorithm and results

From the results shown in Figure 9. Using the decision tree method, the accuracy results were 86.63% with a data set division of 0.8 for training data and 0.2 for testing data. The components used in this algorithm are read excel to read the scrubbed dataset, then split data to divide the data into training data and testing data, then there is an algorithm that will be trained, namely the decision tree, apply mode to run the model, and performance to display the percentage of the method used.

3.) Deep Learning



accuracy: 73.33%

	true Menengah	true Tinggi	true Sangat Tinggi	true Rendah	true Sangat Rendah	class precision
pred. Menengah	3	1	1	0	0	60.00%
pred. Tinggi	2	8	2	0	0	66.67%
pred. Sangat Tinggi	0	0	11	0	0	100.00%
pred. Rendah	0	0	0	11	6	64.71%
pred. Sangat Rendah	0	0	0	0	0	0.00%
class recall	60.00%	88.89%	78.57%	100.00%	0.00%	

Figure 10. Deep Learning algorithm and results

From the results shown in Figure 10. Using the decision tree method, the accuracy results were 73.33% with a data set division of 0.8 for training data and 0.2 for testing data. The components used in this algorithm are read excel to read the scrubbed dataset, then split data to divide the data into training data and testing data, then there is an algorithm that will be trained, namely the decision tree, apply mode to run the model, and performance to display the percentage of the method used.

4. Conclusion

Rice is the main food crop in Indonesia, most of the agricultural sector in Indonesia is dominated by rice farming, including on the island of Sumatra. A common problem that arises is how to find out the areas that produce the most rice each year on the island of Sumatra. This study aims to classify the areas that produce the most rice on the island of Sumatra. The dataset that will be used is taken from Kaggle with a total of 225 data and will be tested using the Decision Tree algorithm and several other algorithms. For data visualization, Tableau will be used to see which areas produce the most rice on the island of Sumatra. By using the research method using the Decision Tree algorithm, an accuracy of 97.78% was obtained with a data split of 0.8 for training data and 0.2 for testing data. In addition, it can also be concluded that production results can be influenced by many factors such as air temperature, rainfall, humidity and others, so this rice cannot live in conditions that are too watery/wet and cannot be too dry. Farmers can do certain things to maximize the care of rice plants so that they can always produce maximum rice production so that the food and economic needs of the Indonesian people can be met properly.

References

[1] H. Tohari, S. Harini, M. A. Yaqin, I. B. Santoso, and C. Crysdian, "Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Produktivitas Padi," *J. Comput. Syst. Informatics*, vol. 5, no. 1, pp. 175–183, 2023, doi: 10.47065/josyc.v5i1.4538.

[2] A. M. Siregar and A. Fauzi, "Klasifikasi Kab Kota Provinsi Jawa Barat Berdasarkan Pendapatan Dari Sektor Pertanian Dengan Algoritma Decision Tree," *Fakt. Exacta*, vol. 13, no. 1, p. 1, 2020, doi: 10.30998/faktorexacta.v13i1.5542.

[3] S. Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, A. Nurkholis, and T. Susanto, "Terakreditasi SINTA Peringkat 2 Algoritme Spatial Decision Tree untuk Evaluasi Kesesuaian Lahan Padi Sawah Irigasi," *Masa Berlaku Mulai*, vol. 1, no. 3, pp. 978–987, 2017.

[4] A. Nurkholis, M. Muhaqiqin, and T. Susanto, "Analisis Kesesuaian Lahan Padi Gogo Berbasis Sifat Tanah dan Cuaca Menggunakan ID3 Spasial (Land Suitability Analysis for Upland Rice based on Soil and Weather Characteristics using Spatial ID3)," *JUITA J. Inform.*, vol. 8, no. 2, pp. 235–244, 2020.

[5] A. Satria, R. M. Badri, and I. Safitri, "Prediksi Hasil Panen Tanaman Pangan Sumatera dengan Metode Machine Learning," *Digit. Transform. Technol.*, vol. 3, no. 2, pp. 389–398, 2023, doi: 10.47709/digitech.v3i2.2852.

[6] A. Nurkholis and I. S. Sitanggang, "Optimization for prediction model of palm oil land suitability using spatial decision tree algorithm," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 3, pp. 192–200, 2020, doi: 10.14710/jtsiskom.2020.13657.

[7] G. Engineering *et al.*, ", Ryo Anugrah," vol. 2, no. 1, pp. 36–40, 2023.

[8] R. Adolph, "済無No Title No Title No Title," vol. 5, pp. 1–23, 2016.

[9] M. Masnur and M. Ali, "Sistem Penunjang Keputusan Penentuan Kesesuaian Budidaya Tanaman Padi Pulu' Mandoti Menggunakan Metode Forward Chaining," *J. Sintaks Log.*, vol. 1, no. 3, pp. 146–152, 2021, doi: 10.31850/jsilog.v1i3.1084.

[10] R. P. Fhonna, Y. Afrillia, Zulfan, J. Aqmal, and S. Abadi, "Klasifikasi Penentuan Jenis Tanah yang Sesuai Terhadap Tanaman Pangan Sebagai Solusi Ketahanan Pangan di Kabupaten Pidie Jaya Menggunakan Metode Random Forest," *J. Inf. dan Teknol.*, vol. 5, no. 4, pp. 12–18, 2023, doi: 10.60083/jidt.v5i4.402.

[11] F. Joanda Kaunang, R. Rotikan, and G. Stella 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.

[12] N. Putri Setyadini, "Penerapan Data Mining Untuk Prediksi Hasil Produksi KaretMenggunakan Algoritma Decision Tree C4.5," *Informatika*, vol. 2, no. 7, pp. 1–11, 2022.

[13] J. T. Elekterika *et al.*, "Klasifikasi Penyakit Tanaman Jagung Melalui Citra Daun Dengan Menggunakan Metode Deep Learning," vol. x, no. x.

[14] A. Ramadhani and M. A. Sembiring, "Sistem Kendali Berbasis Machine Learning Menggunkan Model Neive Bayes Pada Pengeringan Padi Otomatis," *J. Sci. Soc. Res.*, vol. 5, no. 3, p. 690, 2022, doi: 10.54314/jssr.v5i3.1040.

[15] A. I. Widyatami and V. M. Reistiani, "Clustering Wilayah Potensi dan Strategi Pengembangan Komoditas Unggulan Tanaman Hortikultura dan Palawija Level Kecamatan di Sumatera Barat Tahun 2021," *Semin. Nas. Off. Stat.*, vol. 2023, no. 1, pp. 41–52, 2023, doi: 10.34123/semnasoffstat.v2023i1.1737.