Comparison of the K-Nearest Neighbor algorithm and the decision tree on moisture classification

Odi Nurdiawan*, D.A Kurnia, Dodi Solihudin, Tuti Hartati, Tati Suprapti
Informatics Management Department, STMIK IKMI Cirebon, Cirebon, Indonesia

*odinurdiawan2020@gmail.com

Abstract. Soil moisture is a parameter needed by plants in terms of plant growth. In determining the appropriate soil moisture for plants requires a control system. This study uses a comparison of KNN and decision tree algorithms with the aim of being able to determine soil calcification with yield parameters namely moist and dry, so that it has good accuracy compared to the accuracy of the Decision Tree algorithm with an accuracy of 55.77% with dry class recall of 62.69% moist 51.92% dry precision class 58.33% humid 47.37% and K-Nearest Neighbor with 72.69% accuracy dry class recall 80.60% humid 63.16% dry precision class 72.00% humid 73.47%. The results of the above model testing can be concluded that the K-Nearest Neighbor is the most accurate algorithm for classification of moist soil.

1. Introduction
In terms of plant growth, soil moisture is a parameter those plants produce[1][2]. The moisture of soil is very dynamic, caused by soil surface evaporation, transpiration and percolation[3][4]. Humidity information can also be used for management of air reserves, drought early warning, scheduling of irrigation, and weather forecasting. In terms of plant growth, soil moisture is a parameter those plants need[5]. And determine the condition in keeping with the control system plant's needs.

Classification is a data mining technique used in the classification. Based on data interaction[6][7], data for data sets[8]. In learning the closest distance to the dataset[9], the K-Nearest Neighbor algorithm classifies objects based on the soil moisture dataset[10][11][12], so that it can produce a good learning pattern. In the meanwhile, the Decision Tree Algorithm learning pattern adopts a tree featuring roof nodes, intermediate nodes, and leaf nodes[13]-[14]. Terms of being able to estimate the soil calcification with yield parameters[15][16], namely moist and dry, this performance improvement a comparison of the k nearest neighbor and decision tree algorithms, so it has better accuracy[17][18][19].

2. Methodology
The approach used in the K-Nearest Neighbor and Decision tree design analysis. The steps for the research flow chart are as well.
2.1 The first step of data collection
The data used is the observation data from the soil moisture sensor with six test sites and different soil constructions becoming analyzed. So as to develop 124 data records through data.

2.2 Second Step Processing of Data
Preprocessing is one of the mandatory steps taken to model algorithms, so that data that has duplication, errors and gaps in the record can be minimized as best as possible. There are several techniques for preprocessing data. In this study, the data preprocessing technique used was missing values. These are some of the steps necessary taken to model the algorithm is preprocessing, so that data of duplication, errors and record gaps can be minimized as quickly as possible. For preprocessing data, there's many techniques. In this analysis, there were missing values in the data preprocessing technique used.

2.3 Third Transformation Step
This research has a data identity after preprocessing the data, namely the identity of the test with the value of the AE test site as a polynomial, the identity of the soil type, namely Andisol and Vertisol as Binominal, Test Time Identity Namely Morning Afternoon and Evening as Polynomial, Identity of Soil Results, namely numerical, Identity of Moist and Dry as Binominal classification results.

2.4 The Data Mining Model's Fourth Step
What is achieved at this step should be to model the classification by implementing training testing and evaluation tests with the Decision Tree algorithm and K-Nearest Neighbor.

2.5 Knowledge of both the Fifth Step and Action
The example that demonstrates an easy-to-understand result from the pattern process at this step becomes results to be used as information for decision-making.

3. Result and Discussion

3.1 Result
The primary objective of this study is to determine the level of accuracy in soil moisture prediction of the data mining classification algorithms. The results of the algorithm analysis using the Decision Tree and k-Nearest Neighbor technique can be compared when assessing the degree of precision. Any algorithm's performance will be checked before making a comparison. Which can be seen in Figure 2, the leading to improved is to use 10-Fold Cross Validation for the training algorithm.
3.1 Model Testing for Decision Tree

The design of the datamining model explains, based on Figure 2, that the first uses the retrieve operator that functions as a dataset, the second the multiply operator functions as a duplicate of the data that can be used by multiple different operators, the third Cross Validation functions as a sub-process for training tests and tests performed.

3.2 Model Testing for Decision Tree

The implementation of the training experimental setup or test can be seen in the following figure based on the outcome of the datamining model design.

| No | Prediction Results | Class Recall | Class Precision |
|----|--------------------|--------------|-----------------|
| 1  | Dry                | 62.69%       | 58.33%          |
| 2  | Moist              | 47.37%       | 51.92%          |

In the following table, the analysis effects of the experimental data test using the decision tree algorithm model can be seen. Meanwhile, the results of the graph decision tree can be seen in the following image.
Based on Figure 5, the result of the decision tree graph explains as follows:

Type of Test Soil = Andisols
| Testing = Test Place A: Dry {Dry = 6, Moist = 2} |
| Testing = Test Place B: Dry {Dry = 5, Moist = 3} |
| Testing = Test Place C: Dry {Dry = 6, Moist = 2} |
| Testing = Test Place D: Dry {Dry = 7, Moist = 6} |
| Testing = Test Place E: Dry {Dry = 4, Moist = 2} |
| Testing = Test Place F |
| Test time = Morning: Dry {Dry = 2, Damp = 1} |
| Test Time = Afternoon: Dry {Dry = 2, Damp = 1} |
| Test Time = Afternoon: Moist {Dry = 1, Humid = 3} |
Type of Test Soil = Vertisols: Moist {Dry = 34, Moist = 37}

3.3 The K-Nearest Neighbor Model Checking

The implementation of the learning experimental setup and test can be seen in the following figure, based upon the results of the K-Nearest Neighbor design method.

Based on Figure 6, the K-Nearest Neighbor Algorithm Model shows that the operators used are the First Operator Decision Tree, Second Apply model and the third operator Performance. Then the level of accuracy can be generated as follows. Based on Figure 7, the accuracy value of the K-Nearest Neighbor Algorithm shows that the accuracy rate is 72.69% with details of dry predictions and it turns out that 54 data is true dry, dry prediction and turns out to be true moist as much as 21 data, moist predictions and turns out to be true dry as many as 13 data, then the prediction is humid and it turns out to be true moist for 36 data. while the percentage of class recall and class precision can be seen in the following table.

| No | Prediction Results | Class Recall | Class Precision |
|----|--------------------|-------------|-----------------|
| 1  | Dry                | 80.60%      | 72.00%          |
| 2  | Moist              | 63.16%      | 73.47%          |
4. Discussion

Based on the results of the performance test, the comparison data can be seen as follows.

| No | Algorithm          | Accuracy | Prediction Results | Class Recall | Class Precision |
|----|--------------------|----------|--------------------|---------------|-----------------|
| 1  | Decision Tree      | 55.77%   | Dry                | 62.69%        | 58.33%          |
|    |                    |          | Moist              | 51.92%        | 47.37%          |
| 2  | K-Nearest Neighbor | 72.69%   | Dry                | 80.60%        | 72.00%          |
|    |                    |          | Moist              | 63.16%        | 73.47%          |

Based on table 5, it explains that the comparison results of the Decision Tree algorithm accuracy with an accuracy of 55.77% with dry class recall of 62.69% moist 51.92% dry precision class 58.33% humid 47.37% and K-Nearest Neighbor with accuracy 72.69% dry class recall of 80.60%, humid at 63.16%, dry precision class, 72.00%, moist 73.47%. Thus, from the results of testing the model above, it can be concluded that the K-Nearest Neighbor is the most accurate algorithm for classification of moist soil.

5. Conclusion

The conclusion from this research is the comparison of the accuracy of the Decision Tree algorithm with an accuracy of 55.77% with dry class recall of 62.69%, moist 51.92%, dry precision class, 58.33%, moist 47.37% and K-Nearest Neighbor with an accuracy of 72, 69% dry class recall of 80.60%, humid at 63.16%, dry precision class, 72.00%, moist 73.47%. The results of the above model testing can be concluded that the K-Nearest Neighbor is the most accurate algorithm for classification of moist soil.

Acknowledgements

STMIK IKMI Cirebon

References

[1] P. Taneja, H. K. Vasava, P. Daggupati, and A. Biswas, “Multi-algorithm comparison to predict soil organic matter and soil moisture content from cell phone images,” Geoderma, vol. 385, no. August 2020, p. 114863, 2021, doi: 10.1016/j.geoderma.2020.114863.

[2] I. Esfandiarpour-Boroujeni, M. S. Shamsabadi, H. Shirani, Z. Mosleh, M. Bagheri Bodaghabadi, and M. H. Salehi, “Comparison of error and uncertainty of decision tree and learning vector quantization models for predicting soil classes in areas with low altitude variations,” Catena, vol. 191, no. March, p. 104581, 2020, doi: 10.1016/j.catena.2020.104581.

[3] D. C. Owen, M. T. Bensi, A. P. Davis, and A. H. Aydilek, “Measuring soil coverage using image feature descriptors and the decision tree learning algorithm,” Biosyst. Eng., vol. 196, pp. 112–126, 2020, doi: 10.1016/j.biosystemseng.2020.06.002.

[4] A. Beucher, A. B. Møller, and M. H. Greve, “Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark,” Geoderma, vol. 352, no. September, pp. 351–359, 2019, doi: 10.1016/j.geoderma.2017.11.004.

[5] R. J. Engel, J. E. Witt, and H. Esvaran, “The classification, distribution, and extent of soils with a xeric moisture regime in the United States,” Catena, vol. 28, no. 3–4, pp. 203–209, 1997, doi: 10.1016/S0341-8162(96)00038-0.

[6] F. F. Ting, Y. J. Tan, and K. S. Sim, “Convolutional neural network improvement for breast cancer classification,” Expert Syst. Appl., vol. 120, pp. 103–115, 2019, doi: 10.1016/j.eswa.2018.11.008.

[7] C. Pelletier, G. Webb, and F. Petitjean, “Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series,” Remote Sens., vol. 11, no. 5, p. 523, Mar. 2019, doi: 10.3390/rs11050523.
[8] R. Chauhan and H. Kaur, Predictive analytics and data mining: A framework for optimizing decisions with R tool, 2013.

[9] Y. Hu and Y. Wang, “Probabilistic soil classification and stratification in a vertical cross-section from limited cone penetration tests using random field and Monte Carlo simulation,” Comput. Geotech., vol. 124, no. January, p. 103634, 2020, doi: 10.1016/j.compgeo.2020.103634.

[10] G. Yao, T. Lei, and J. Zhong, “A review of Convolutional-Neural-Network-based action recognition,” Pattern Recognit. Lett., vol. 118, pp. 14–22, 2019, doi: 10.1016/j.patrec.2018.05.018.

[11] L. Chen, C. Wang, J. Chen, Z. Xiang, and X. Hu, “Voice Disorder Identification by using Hilbert-Huang Transform (HHT) and K Nearest Neighbor (KNN),” J. Voice, 2020, doi: 10.1016/j.jvoice.2020.03.009.

[12] C. Reale, K. Gavin, L. Librić, and D. Jurić-Kaćunić, “Automatic classification of fine-grained soils using CPT measurements and Artificial Neural Networks,” Adv. Eng. Informatics, vol. 36, no. April, pp. 207–215, 2018, doi: 10.1016/j.aei.2018.04.003.

[13] I. D. Mienye, Y. Sun, and Z. Wang, “Prediction performance of improved decision tree-based algorithms: A review,” Procedia Manuf., vol. 35, pp. 698–703, 2019, doi: 10.1016/j.promfg.2019.06.011.

[14] H. F. Zhou, J. W. Zhang, Y. Q. Zhou, X. J. Guo, and Y. M. Ma, “A feature selection algorithm of decision tree based on feature weight,” Expert Syst. Appl., vol. 164, no. February 2020, p. 113842, 2021, doi: 10.1016/j.eswa.2020.113842.

[15] Nurdin et al., “The Implementation of Backtracking Algorithm on Crossword Puzzle Games Based on Android,” J. Phys. Conf. Ser., vol. 1363, no. 1, 2019, doi: 10.1088/1742-6596/1363/1/012075.

[16] E. Basti, C. Kuzey, and D. Delen, “Analyzing initial public offerings’ short-term performance using decision trees and SVMs,” Decis. Support Syst., vol. 73, pp. 15–27, 2015, doi: 10.1016/j.dss.2015.02.011.

[17] W. Gumindoga, A. Murwira, D. T. Rwasoka, F. B. Jahure, and L. Chikwiramakomo, “The spatio-temporal soil moisture variation along the major tributaries of Zambezi River in the Mbire District, Zimbabwe,” J. Hydrol. Reg. Stud., vol. 32, no. October, p. 100753, 2020, doi: 10.1016/j.jhr.2020.100753.

[18] F. Hu, Z. Wei, W. Zhang, D. Dorjee, and L. Meng, “A spatial downscaling method for SMAP soil moisture through visible and shortwave-infrared remote sensing data,” J. Hydrol., vol. 590, no. March, p. 125360, 2020, doi: 10.1016/j.jhydrol.2020.125360.

[19] R. D. Dana, A. R. Dikananda, D. Sudrajat, A. Wanto, and F. Fasya, “Measurement of health service performance through machine learning using clustering techniques,” J. Phys. Conf. Ser., vol. 1360, no. 1, 2019, doi: 10.1088/1742-6596/1360/1/012017.