ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse124912020.pdf
https://doi.org/10.30534/ijatcse/2020/124912020

C4.5 Algorithm with Average Gain to Predict Human Development Index in Indonesia

Jajam Haerul Jaman[1], Alya Farhah Bahira[2], Tesa Nur Padilah[3], Munir[4]
[1] Faculty Computer Science, Universitas Singaperbangsa Karawang, Indonesia, jajam.haeruljaman@gmail.com
[2] Faculty Computer Science, Universitas Singaperbangsa Karawang, Indonesia, 1441177004328@student.unsika.ac.id
[3] Faculty Computer Science, Universitas Singaperbangsa Karawang, Indonesia, tesa.nurpadilah@staff.unsika.ac.id
[4] Universitas Pendidikan Indonesia, Indonesia, munir@upi.edu

ABSTRACT

Poorness makes the government's program to create national development goal rising — Benchmark for the success of human development called Human Development Index. An area that has a high Human Development Index should have good quality human, or if the Human Development Index high, then grade of the poorness is low. C4-5 algorithm method is one of the methods of creating a decision tree from the training data available. C4-5 algorithm is the expansion from ID3. The goal of this research is to predict the Human Development Index in Indonesia and make the decision tree from the data. Attributes that use in this research are Life Expectancy, Mean Years of Schooling, expect years of schooling, and per capita income. There are two an in this research are with mean and median from each attribute. The data for this research method 4473 data. The result shows that Scenario with the has classification is 1st Scenario with mean from each attribute, that to accurate 85,893%.

Key words: Average gain, C4-5 algorithm, Human development index

1. INTRODUCTION

A major problem in development is poverty or the number of people below the poverty line. Poverty is a condition in which a person or family is not able to meet the primary needs. Poverty makes government efforts to realize the ideals of national development are increasingly committed.

It is therefore formulated a new concept in measuring the development of a human-oriented country. The benchmark success of human development has been developed by the United Nations Development Programme (UNDP) known as the Human Development Index (HDI) or known as the Human Development Index (IPM) [1].

Efforts to realize that productive human development is required for ongoing monitoring. In the year 2011, the value of IPM Indonesia was ranked at 124 in the world, and in 2012 Indonesia rose three positions to rank 121. It is supported by improved indicators of the establishment, especially in life expectancy and the average length of the school, which in 2012 reaches 69.8 years and 12.9 years[2].

Human Development Index (HDI) is a breakthrough in assessing human development. The IPM includes three components that are considered fundamental to humans and are operational easily calculated to produce a measure that reflects human development efforts. These three aspects are related to the opportunity of life (longevity), knowledge, and decent living. Life opportunities are calculated based on the life expectancy of birth. Knowledge is measured based on the average length of the school, the population literacy rate of 15 years and above, and life deserves to be measured with per capita spending based on Purchasing Power Parity (purchasing power parity in rupiah) [1].

To know and map the quality of human development or the level of community welfare in Indonesia, one of them by using the Human Development Index (HDI). The IPM Data is required not only to determine the extent to which the results of Community welfare development have been conducted but also as input materials to formulate policies and intervention programs in the years to come more effectively and efficiently [1].

In this study, the algorithm of C 4.5 with the average gain will be applied by authors to predict and analyze data on the Human Development Index (HDI) obtained from the Central Statistics Agency [1].

Some research has been done in prediction development and prediction using C 4.5, including the application of Average Gain method, Threshold Pruning and Cost Complexity Pruning for Split attribute on algorithm C4.5 [3]. Implementation of Data Mining classification pattern of the customer using C4.5 algorithm of Bank BRI Batang [4]. Implementation of C 4.5 algorithm for student majors, [5]. C 4.5 algorithm to determine the qualification level of used motors to be sold, son [6].

Besides the performance of algorithm C4.5 is outstanding in conducting calcification, average accuracy generated above 90%, in a paper titled "Implementation of C 4.5 algorithm for scholarship recipient determination", the highest precision accuracy results, and The lowest error Rate is in scenario I with seven criteria, i.e., accuracy value of 93, 58%, Precision value 95, 5%, and error value lowest Rate 6, 14% [7]. A paper titled "Predictor of Student graduation with algorithm method C 4.5", as a complete explained that for the approval seen from gender turns female gender is greater percentage compared to male gender, then Gain values gained from the age of students who work more to students with the category age 1, and the gain gained from the age of students who do not work more to students with the category Age 3 [8]. Some development is also done by the name of the paper titled "AUC 4.5: AUC-based C 4.5 decision tree algorithm for the classification of balanced data" they propose a new method of induction tree by modifying Quinlan's C 4.5 Algorithm, which we named AUC 4, 5. This method seems more suited to binary balanced data classification[9]. Adviesraad for literary optimization, the paper titled "This: Performance optimization C 4.5 decision tree Construction algorithm" proposes an optimized algorithm inspired by the RainForest. By using more sophisticated switching criteria between the two algorithms, we can benefit Kokoro even when it releases about matched stats in memory. Evaluation Products that our method can JOURS performance enhancement 2, 8 times better than traditional recursive implementations [10].

2. METHOD

The Dataset used is the Human Development Index (HDI) data from the Central Statistics Agency (BPS) as much as 4473 rows of data. This research uses quantitative research methodology with the CRISP-DM (Cross Industry Standard Process for Data Mining) method. Data obtained from BPS several empty columns are the attributes of life expectancy, literacy, old school, per capita, and HDI expenditure. The missing values are populated by calculating the median of each city/district. According to BPS [1], IPM Data is categorized as a low, medium, high, and very high, namely:
IPM < 60: Low HDI
60 ≤ IPM < 70: IPM Moderate
70 ≤ IPM < 80: High HDI
IPM ≥ 80: Very High HDI

The C 4.5 algorithm is an algorithm used to form a decision tree. Being a decision tree can mean a way to predict or clarify a powerful one. The decision tree can divide a large set of data into smaller sets of records by applying a set of decision rules. [11] stated that there are several stages in making a decision tree with C 4.5 algorithm, namely:

2.1. Reference

Related research, for data [1], methodologi and algorithm refer to [2–30], and the topic refer to [1,31–33].

2.2. Set up your training data.

2.3. Determine the roots of the tree.

The root will be extracted from the selected attribute, by calculating the gain value of each attribute, the highest gain value, which will be the first root. Before calculating the gain value of the attribute, first, calculate the value of entropy.

2.4. To calculate the value of entropy used formula:

$$Entropy\ (S) = \sum_{i=1}^{n} -pi.\log_2 pi \quad (1)$$

Description: S = Set of cases n = number of partitions S PI = Si proportion to S

2.5. Then calculate the gain value using the formula:

$$Gain\ (S,) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si) \quad (2)$$

Description:
S = Set of cases
A = Features
n = number of partition attribute A
| The | = Si proportion to S
| S | = number of cases in S

2.6. Repeat step 2 until all the records are partition.

2.7. The process of partitioning the decision tree will stop when.

(a.) All records in vertex n get the same class. (b.) There is no attribute in the repartitioning record. (c.) There is no record in the empty branch. The study used the c 4.5 algorithms with the average gain proposed by Mitchell. Average gain not only resolves weaknesses in gain information but also helps to solve problems of gain ratio. The proposed split attribute method is designed to prevent bias arising from attributes. The average gain (s, a) of attribute a, which corresponds to the s training data, can be calculated with the formula:

$$\text{AverageGain}(S, A)\ \frac{\text{Gain (S,A)}}{|A|} \quad (3)$$

Where:| A | = Number of attributes A

Gain (S, A)= information gain attribute A from the training S data

3. RESULTS AND DISCUSSION

3.1. Scenario I

The scenario I am performed by calculating the average or average of each attribute to group data to low or high.

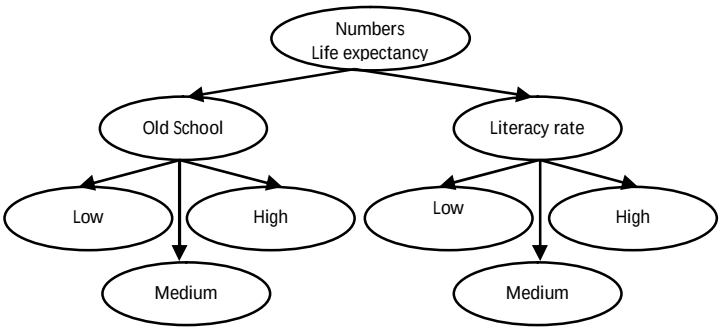

Figure 1: Mean of Decision tree

3.2. Scenario II

Scenario II is performed by categorizing using the median or middle value of each existing attribute.

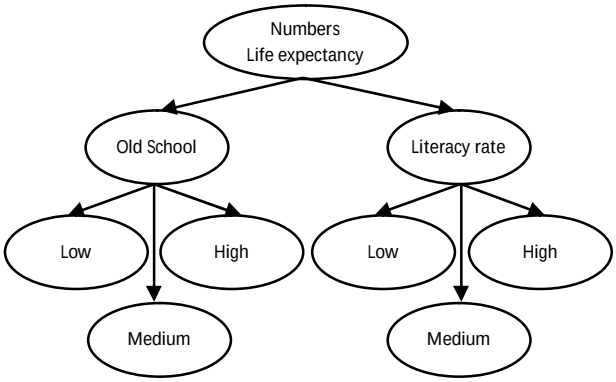

Figure 2: Decision tree median

The following is a comparison of Accuracy, Precision, and Recall results of both scenarios already done:

Table 1: Comparison of accuracy, precision and recall Results

	Scenario I	**Scenario II**
Accuracy	85.893%	83.702%
Precision	64.5686%	71.73%
Recall	56.13116%	86.2069%

The following diagram results in accuracy, Precision, and Recall of algorithm C 4.5 with Average Gain:

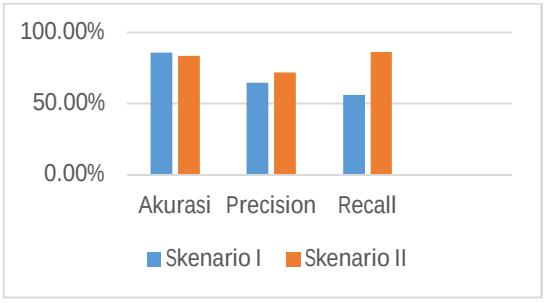

Figure 3: accuracy, precision, and recall Results

Attributes that affect classification are life expectancy, literacy, and old school numbers. While the attributes are not influential is the per capita income, because the income of Indonesian Percapita has only one category, namely low. The scenario I, i.e. by average, shows that the life expectancy, old school, and literacy rate are influential in the scenario I classification. As for scenario II, i.e. with the middle or median value of each attribute indicates that Only the literate and old school figures are influential in the scenario II classification. The research results of both scenarios have been demonstrated that determining the nominal category (e.g. low and high) of each attribute shows different results in the formation of decision trees and classification results from C45 algorithm with average gain. This research also shows that the attributes that have been most successfully classified are with an average of 85.893% accuracy

4. CONCLUSION

Based on the results of the study obtained the following conclusions:

Data processing is divided into two scenarios, namely the average and median of each attribute. Both scenarios provide different classification results.

The human development index in every city/Regency in Indonesia is increasing annually, due to the average literacy rate, life expectancy, and old school income in Indonesia, increasing from year to year.

The result of classifying algorithms is measured based on comparisons between categories obtained from the Central Statistic Agency (BPS) with categories obtained from the classification results to know the best scenarios. The classification result with the best accuracy suggests that the scenario I am with an average of 85.893% accuracy. Then from the evaluation results can be noted that the algorithm C 4.5 with the Average Gain can be used to predict the human development index in Indonesia.

The results showed that the human development index figures were influenced by the life expectancy, old school, and literacy numbers based on the scenario I decision tree with the highest accuracy results.

5. AUTHORS' NOTE

The author(s) declare(s) that there is no conflict of interest regarding the publication of this article. The authors confirmed that the data and the paper are free of plagiarism.

REFERENCES

1. BPS. *Index Pembangunan Manusia*. Vol 53. Badan Pusat Statistik; 2013. doi:10.1017/CBO9781107415324.004
2. Purnamasari SB, Yasin H, Wuryandari T. Pemilihan

Cluster Optimum pada Fuzzy C Means (Studi Kasus Pengelompokan Kabupaten/Kota di Provinsi Jawa Tengah Berdasarkan Indikator Indeks Pembangunan manusia). *J Gaussian Vol 3 No 3 tahun 2014*. 2014;3:491-498. doi:10.3856/vol38-issue3-fulltext-15

3. Rahayu ES, Wahono RS, Supriyanto C. Penerapan Metode Average Gain, Threshold Pruning dan Cost. *J Intell Syst*. 2015;1(2):91-97.
4. Oktaviana AR. Penerapan data mining klasifikasi pola nasabah menggunakan algoritma c4.5 pada bank bri batang. *Fik, UDINUS*. 2016;1(1). doi:10.1021/jf901375e
5. Swastina L. Penerapan Algoritma C4.5 untuk Penentuan Jurusan Mahasiswa. *J Gema Aktual*. 2013;2(1):93-98.
6. Putra DWT. Algoritma C4.5 untuk Menentukan Tingkat Kelayakan Motor Bekas yang Akan Dijual. *J TEKNOIF*. 2016;4(1):16-22.
7. Jaman JH, Astuti NIP. Penerapan Algoritma C4.5 Untuk Penentuan Penerima Beasiswa. *J Ilmu Komput dan Teknol Inf*. 2018;3(1):25-29.
8. Jaman JH. Prediksi Kelulusan Mahasiswa Dengan Metode Algoritma C4.5. *Syntax*. 2013;2(2):1-6. doi:10.1002/jhet.1722
9. Lee JS. AUC4.5: AUC-Based C4.5 Decision Tree Algorithm for Imbalanced Data Classification. *IEEE Access*. 2019;7:106034-106042. doi:10.1109/ACCESS.2019.2931865
10. Yang Y, Chen W. Taiga: Performance optimization of the C4.5 decision tree construction algorithm. *Tsinghua Sci Technol*. 2016;21(4):415-425. doi:10.1109/TST.2016.7536719
11. Kusrini K. *Algoritma Data Mining*. Yogyakarta: Andi; 2009.
12. Kaur KA, Bhutani L. A Review on Classification Using Decision Tree. *IJCAT -International J Comput Technol*. 2015;2(02):42-46. www.IJCAT.org.
13. Jabber B, Sai Venkat P, Sri Sai Nikhil K, Lakshmi Avinash B. A novel sampling approach for balancing the data and providing health care management system by government. *Int J Adv Trends Comput Sci Eng*. 2019;8(6):2753-2761. doi:10.30534/ijatcse/2019/12862019
14. Mothukuri R. Data Mining on Prediction of Crime and Legal Judgements: A State of an Art. *Int J Adv Trends Comput Sci Eng*. 2019;8(6):3670-3679. doi:10.30534/ijatcse/2019/153862019
15. Buot NS. Multiple intelligences and reading comprehension of senior high school students: A response evaluation through educational data mining technique. *Int J Adv Trends Comput Sci Eng*. 2019;8(6):2871-2876. doi:10.30534/ijatcse/2019/30862019
16. Böhringer KF. Modeling and controlling parallel tasks in droplet-based microfluidic systems. *Des Autom Methods Tools Microfluid Biochips*. 2006;25(2):301-327. doi:10.1007/1-4020-5123-9_12
17. Jebamalar DRJPA. Accuracy Improvement of C4.5 using K means Clustering. *Int J Sci Res*. 2017;6(6):2755-2758. https://www.ijsr.net/archive/v6i6/ART20174834.pdf.
18. Black A, Korb K, Nicholson A. Learning Dynamic Bayesian Networks: Algorithms and Issues. *AbnmsOrg*. http://abnms.org/conferences/abnms2013/presentations/ABNMS2013 - Black - Learning Dynamic Bayesian Networks Algorithms and Issues.pdf.
19. Widiartha KK. Rancang Bangun Sistem Fuzzy Takagi Sugeno untuk Prediksi IPM dan Rekomendasi Anggaran APBD. *FMIPA UGM*. 2014.
20. Larrañaga P, Karshenas H, Bielza C, Santana R. A review on evolutionary algorithms in Bayesian network learning and inference tasks. *Inf Sci (Ny)*. 2013;233:109-125. doi:10.1016/j.ins.2012.12.051
21. Aminian M, Couvin D, Shabbeer A, et al. Predicting mycobacterium tuberculosis complex clades using knowledge-based bayesian networks. *Biomed Res Int*. 2014;2014. doi:10.1155/2014/398484
22. Zelterman D. *Bayesian Artificial Intelligence*. Vol 47.; 2005. doi:10.1198/tech.2005.s836
23. Sandri M, Berchialla P, Baldi I, Gregori D, De Blasi RA. Dynamic Bayesian Networks to predict sequences of organ failures in patients admitted to ICU. *J Biomed Inform*. 2014;48:106-113. doi:10.1016/j.jbi.2013.12.008
24. Rajeev Bedi Sunil Kumar Gupta D. Review of Decision Tree Data Mining Algorithms: Id3 and C4.5. 2015;(August):5-8.
25. Kaur S, Kaur H. Available Online at www.ijarcs.info Review of Decision Tree Data mining Algorithms : CART and C4 . 5. 2017;8(4):2015-2018.
26. Ngoc PV, Ngoc CVT, Ngoc TVT, Duy DN. A C4.5 algorithm for english emotional classification. *Evol Syst*. 2019;10(3):425-451. doi:10.1007/s12530-017-9180-1
27. Muslim MA, Rukmana SH, Sugiharti E, Prasetiyo B, Alimah S. Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis. *J Phys Conf Ser*. 2018;983(1). doi:10.1088/1742-6596/983/1/012063
28. Elistia E, Syahzuni BA. the Correlation of the Human Development Index (Hdi) Towards Economic Growth (Gdp Per Capita) in 10 Asean Member Countries. *Jhss (Journal Humanit Soc Stud*. 2018;2(2):40-46. doi:10.33751/jhss.v2i2.949
29. Chebrolu S, Sanjeevi SG. Attribute Reduction in Decision-Theoretic Rough Set Model using Particle Swarm Optimization with the Threshold Parameters Determined using LMS Training Rule. *Procedia Comput Sci*. 2015;57:527-536. doi:10.1016/j.procs.2015.07.382
30. Wang G. Applying Customer Loyalty Classification with RFM and Naïve Bayes for Better Decision Making. *2019 Int Semin Appl Technol Inf Commun*. 2019:564-568. doi:10.1109/ISEMANTIC.2019.8884262
31. Sumatera DAN, Dengan U, Regresi M, Ordinal L.

Pemodelan ipm provinsi jawa timur, jawa tengah, jawa barat dan sumatera utara dengan metode regresi logistik ordinal. 2009:1-9.

32. Sari BN, Priati P. Identifikasi Keterkaitan Variabel dan Prediksi Indeks Pembangunan Manusia (IPM) Provinsi Jawa Barat Menggunakan Dynamic Bayesian Networks. *J INFOTEL - Inform Telekomun Elektron*. 2016;8(2):150. doi:10.20895/infotel.v8i2.123

33. Suseno N. ANALISIS FAKTOR–FAKTOR YANG MEMPENGARUHI INDEKS PEMBANGUNAN MANUSIA DI PROVINSI SUMATERA UTARA. 2015.