Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12968
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Prediction of the Human Development Index for Equitable Development in West Sumatera Province Using the C4.5 Algorithm

Weri Sirait[1)*], Nur Azizah[2)]
[1,2)]Universitas Metamedia
[1)]werisirait@metamedia.ac.id, [2)]nur.azizah@metamedia.ac.id

Submitted : Sep 5, 2023 | **Accepted** : Sep 7, 2023 | **Published** : Oct 1, 2023

Abstract: Unequal development in Indonesia can be seen from the Human Development Index. The Human Development Index is a tool used to measure the attainment of the quality of life of a region or country and as material for economic policy on quality of life. It contains components of health level, education level and welfare level. In 2022, West Sumatera Province achieved the 9th highest Human Development Index in Indonesia, namely 73.26, with this figure the West Sumatera Province Human Development Index is above the national average. However, there are still regencies/cities in West Sumatera Province that have achievements below the national average. This factor causes the development conditions in West Sumatera Province to be uneven. Uneven human development conditions will make it difficult for the government to improve Human Resources (HR). In this research, the C45 Data Mining Algorithm was implemented to predict the Regency/City Human Development Index in West Sumatera Province. As is the method of the Central Bureau of Statistics in measuring the Human Development Index, the variables used from the Human Development Index indicators are Life Expectancy, Years of School Expectation, Average Length of Schooling, and Per Capita Expenditures. The Central Statistics Agency data used in this research covers all regencies/cities in West Sumatera during the period 2018-2022. Range levels are grouped into three groups, namely, low, medium, and high. Based on testing using RapidMiner software with the Cross Validation operator, an accuracy value of 86.61% was obtained.
Keywords: Prediction; Human Development Index; Data Mining; C4.5

INTRODUCTION

The Human Development Index is a measuring tool that influences economic policy on quality of life (Putra et al., 2018). Development is an effort made by the government to realize the welfare and prosperity of society. The success of a country in national development is not only reflected in economic growth but also in improving the quality of Human Resources (HR) (Anggraeni & Arum R, 2022). Humans are the nation's true wealth and are always the main factor in every income program. One of the measuring tools used to measure development results is the Human Development Index. The Human Development Index is intended as a planning and evaluation tool for government policies, such as the allocation of funds for regions, while the Human Development Index indicators describe the success of government development targets. So it can be said, that the Human Development Index is data that can be used in making policies by the government(Pratiwi & Wijayanto, 2019).

According to the Central Statistics Agency, 3 basic components make up the Human Development Index, namely the health dimension, the knowledge dimension, and the dimension of a decent life. The Human Development Index according to the Central Statistics Agency is divided into 4 categories or

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12968 p-ISSN : 2541-044X

groups, namely low Human Development Index if <60, medium 60≤IPM<70, high 70≤IPM<80, and ≥80 very high (Central Statistics Agency:2020) (Pratiwi & Wijayanto, 2019).

In 2022, West Sumatera Province achieved the 9th highest Human Development Index ranking in Indonesia, namely 73.26, with this figure the West Sumatera Province Human Development Index is above the national average. However, there are still regencies/cities in West Sumatera Province that have achievements below the national average. Uneven human development conditions will make it difficult for the government to improve Human Resources (HR). However, what is still a problem is the issue of data validation in several government agencies that have different data, thus creating confusion between ministries in making policies. In this research, Data Mining was implemented to predict the Regency/City Human Development Index in West Sumatera Province to identify equitable development. The Central Statistics Agency data used in this research covers all districts/cities in West Sumatera during the period 2018-2022. As per the Central Statistics Agency method in measuring the Human Development Index, the variables used from the Human Development Index indicator are Life Expectancy, Expected Years of Schooling, Average Years of Schooling, and Per Capita Expenditure.

One of the methods used in predicting the human development index is the C4.5 data mining algorithm. The C4.5 algorithm is an algorithm that processes data by calculating entropy and information gain. After the process of calculating the information gain value, the values obtained for each attribute will be compared (Siregar et al., 2021). The C4.5 algorithm is one of the algorithms contained in data mining classification and is a model or function that explains or differentiates concepts or data classes to estimate the unknown class of an object. In general, the C4.5 algorithm for building a decision tree is to select an attribute as the root, create a branch for each value, divide the cases into branches, and then repeat the process for each branch until all cases in the branch have the same class. (Siregar et al., 2021).

Based on the background above, this research is intended to help provide consideration by the government, especially to further improve programs that can build Human Development Index values in several Regencies/Cities that have Human Development Index values below the average by showing their effect on increasing indicators economy in units to measure all districts/cities in West Sumatera. So researchers use a Data Mining technique, namely the C4.5 algorithm, to extract information.

LITERATURE REVIEW

Table 1. Literature Review

No	Researcher Name (year)	Method	Data	Results
1	(Sinaga et al., 2022)	Algorithm C4.5	The questionnaire data which initially contained 102 data became 78 data that were ready to be processed.	The data set that will be used in this study is the questionnaire data for STMIK Pelita Nusantara students. For the questionnaire data, 102 data were successfully collected.
2	(Rismayanti et al., 2018)	Algorithm C4.5	Application of the C4.5 Data Mining Algorithm in Determining the Performance Track Record of STT Harapan Medan Lecturers	1. The results of the decision tree determining the performance track record of STTH Medan lecturers resulted in the decision that high-performing lecturers are lecturers who have high publication scores. 2. Lecturers who perform "adequately" are lecturers who have a publication score and a

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12968 p-ISSN : 2541-044X

No	Researcher Name (year)	Method	Data	Results
				service score of moderate value. 3. Lecturers who perform "less" are lecturers who have publication scores and service scores "None" and "Less"
3	(Fersellia et al., 2023)	Algorithm C4.5, SMOTE technique	This data set comes from public opinion on the Shopee Food application. This dataset was taken from Twitter over one month, from May 2023 to June 2023 with a total of 1005 tweets using the keyword shop foo	In conclusion, sentiment analysis on the Shopee Food application on Google Play using the Decision Tree C4.5 algorithm and the SMOTE technique can overcome data imbalances with a prediction accuracy of 0.88. This technique is more efficient than the undersampling technique and the combination of oversampling and undersampling. These results can provide developers with valuable insights to improve app quality and user satisfaction
4	(Sintawati et al., 2021)	Algorithm C4.5	Based on the sample taken in the population that has been described, it is determined that the number of parents' responses is 100 data on the influence of early childhood development on the emergence of streaming video ads on YouTube.	Based on this research, the conclusions obtained are that research on the influence of early childhood development on video streaming ads on YouTube can be applied using the calculation of the C4.5 Algorithm method. and rapid-miner software data processing with detailed data analysis results, so that they can provide the right decisions
5	(Sumpena et al., 2019)	Algorithm C4.5, Naive Bayes	Data prepared for this study consisted of 343 patients taken from the patient registers in December 2018 and January 2019.	From the results of this study by analyzing patient data using a ventilator, Central Venous Pressure (CVP), and also a diagnosis of sepsis the data is processed and classified. ICU patient accuracy results obtained, namely C4.5 algorithm has an accuracy of 81.25% and AUC 0.623, while Naive Bayes has an accuracy of 80.66% and AUC 0.795. From these results, the C4.5 algorithm has a better edge of 0.59% than the Naive Bayes algorithm
6	(Siregar et al., 2021)	Algorithm C4.5	In this study, the data that will be used are 70 prospective new students who registered at Harapan	The results of this study indicate that in mapping the selection pattern of interview attributes into level 1 nodes, the attributes of the

*name of corresponding author

No	Researcher Name (year)	Method	Data	Results
			University. The attributes used are: the average value of report cards, basic academic ability tests, basic computer knowledge tests, and interviews	basic computer knowledge test become the level 1 branch, the attributes of the basic academic ability test become the level 2 branch and the attribute average value of report cards becomes the level 3 branch.
7	(Sudalyo & Mukti, 2022)	Algorithm C4.5, Fuzzy MADM	The data used is data from 140 students	The initial selection used the C-45 method with the variables of GPA, parents' income, achievements, parental dependents, and cases. The results of the C4.5 calculation show that the priority is parental dependents with a Gain value of 0.007822696, followed by a GPA with a Gain value of -0.130011482, the third priority is Parents' Income with a Gain value of -0.702657067 and the last priority is an achievement. The results of the calculation are continued with Fuzzy MADM resulting in 5 rules used to determine student priorities (can) or not. The results achieved from 140 students who applied were accepted by 135 students who passed the initial stage, and out of 135 rankings, 70 students were determined to receive scholarships from the Government with the highest calculation score of 21 and the lowest of 14.4.
8	(Attamami et al., 2023)	Algorithm C4.5, Naive Bayes	The data source used is data on requests for health insurance financing assistance for the poor According to the Depok City Health Office for 2020-2022 as many as 1043 data records that have been cleaned and selected. From this dataset, 730 datasets were used as training data, and 313 datasets were used as testing data.	the C4.5 classification algorithm gets the highest accuracy value, which is 99.04%. A total of 310 data records were predicted correctly with an error rate or error of 0.96% or as many as 3 data records from 313 data tested. While the Naive Bayes classification algorithm gets an accuracy value of 92.97%. A total of 291 data records were predicted to be correct with an error rate of 7.03% or as many as 22 data records were predicted to be incorrect from the 313 data tested.
9	(Azizah et al., 2022)	Algorithm C4.5	The data processed in this study were 50 respondents	the results of the data calculation process using the C4.5 Algorithm

*name of corresponding author

No	Researcher Name (year)	Method	Data	Results
			from the results of questionnaire data distributed.	method according to the predetermined steps produce 6 rules with 3 satisfied decisions and 3 dissatisfied decisions. The resulting rule can be used to determine customer satisfaction to correct deficiencies in services provided so far.
10	(Andesti et al., 2022)	Algorithm C4.5	The data processed in this study were 90 respondents from the results of questionnaire data distributed.	The results of the discussion on the Weka 3.8.6 software, the data accuracy rate is 90 % or 81 data and the error rate is around 10 % or 9 Data. From the data of 90 respondents, the factor that influence the potential for increasing proficiency in English is Practice (C2).
11	(Yunus & Haba, 2023)	Algorithm C4.5	This research data consists of the year or period 2017 to 2021 regarding cervical cancer obtained at AlieSaboe Regional Hospital	The results achieved by this research were to obtain prediction results using the C.45 algorithm and had an accuracy of 70.37% as measured using the Confusion Matrix on the Rapid Miner tool..
12	(Nas, 2021)	Algorithm C4.5	In this research, the data processed is the results of a survey conducted on prospective new students in the D3-Informatics Management Study Program. There are 25 sample data as training datasets with 5 assessment data.	The results of testing processing using Rapid Miner Software, obtained an accuracy rate of 100% with 120 testing datasets. The result of using Data Mining with the C4.5 algorithm has been able to predict the interest of prospective new students based on assessment factors in choosing a college.
13	(Ginting et al., 2020)	Algorithm C4.5	The data used is a dataset from research conducted by Muqorobin, 2019 and implemented in programming form using the Python programming language to produce information on prediction results.	Based on the results obtained from the description of the results and prediction results with Algorithm C4.5, the values for accuracy, precision, and recall are 73%, 71%, and 71%.

There are several stages in making a decision tree in the C4.5 algorithm, namely:

1. Prepare training data. Training data is usually taken from historical data that has happened before and has been grouped into certain classes.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12968 p-ISSN : 2541-044X

2. Counting tree roots. The root will be taken from the attribute to be selected, by calculating the gain value of each attribute, the highest gain value will be the first root. Before calculating the gain value from the attribute, first calculate the entropy value.

To calculate the entropy value use the formula:

$$Entropy\ (H) = \sum_{i=1}^{n} - pi * log_2\ pi$$

H : Case Collection
n : Number of Partitions H
pi : Proportion of Hi to H

Then after the entropy value for each attribute has been obtained, calculate the gain value using the formula

$$Gain\ (H, A) = Entropy(H) \sum_{i=1}^{n} \frac{|Hi|}{|H|} * Entropy(Hi)$$

H : Case Collection
A : Attribute,
n : Number of attribute partitions A
|Hi|: Number of cases on partition i
|H| : Number of cases on partition H

METHOD

The stages in this research include research steps. The framework in this research is described as follows:

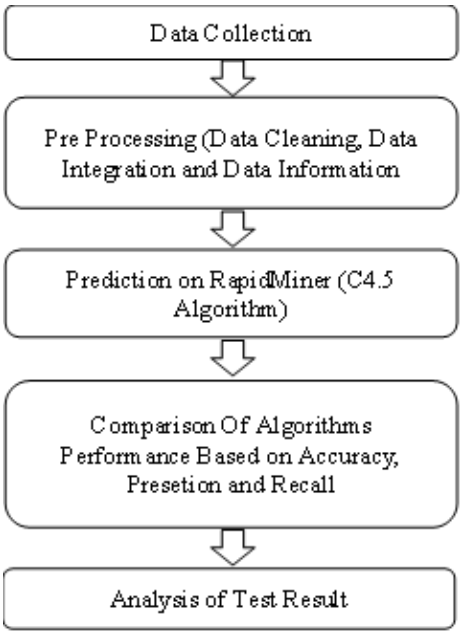


Figure 1. Research Stages

Data Collection

This research was conducted using secondary data sourced from the Central Bureau of Statistics (BPS) of West Sumatera Province which consists of 19 Regencies/Cities.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12968
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Pre-Processing

At this stage, after the data is collected, the preprocessing process is carried out which will be used for testing with the RapidMiner Application.

Data Cleaning

Data cleaning is carried out to clean data so that when all attributes are selected they do not experience a missing value, missing value, or redundancy value because the results of the data mining process can be clean from errors in data processing.

Data Integration

After cleaning the data, the cleaned data is integrated into one file as a dataset that will be used in the
data mining analysis process. In this study, data integration from several tables involved was carried out using Structured Query Language (SQL), and the results were stored in a single file in CSV format.

Data Transformation

The data obtained is based on the respondents' answers and then analyzed so that it can produce information. The information is converted into a data transformation. Data transformation is done to convert data into values with a certain format. Data that has been in a certain format is divided into small groups. Data category IPM is divided into 3 groups, namely "very high", "high" and "medium".

RESULT

Testing using the C4.5 Algorithm using RapidMiner can be seen in Figure 2 and Figure 3 below :

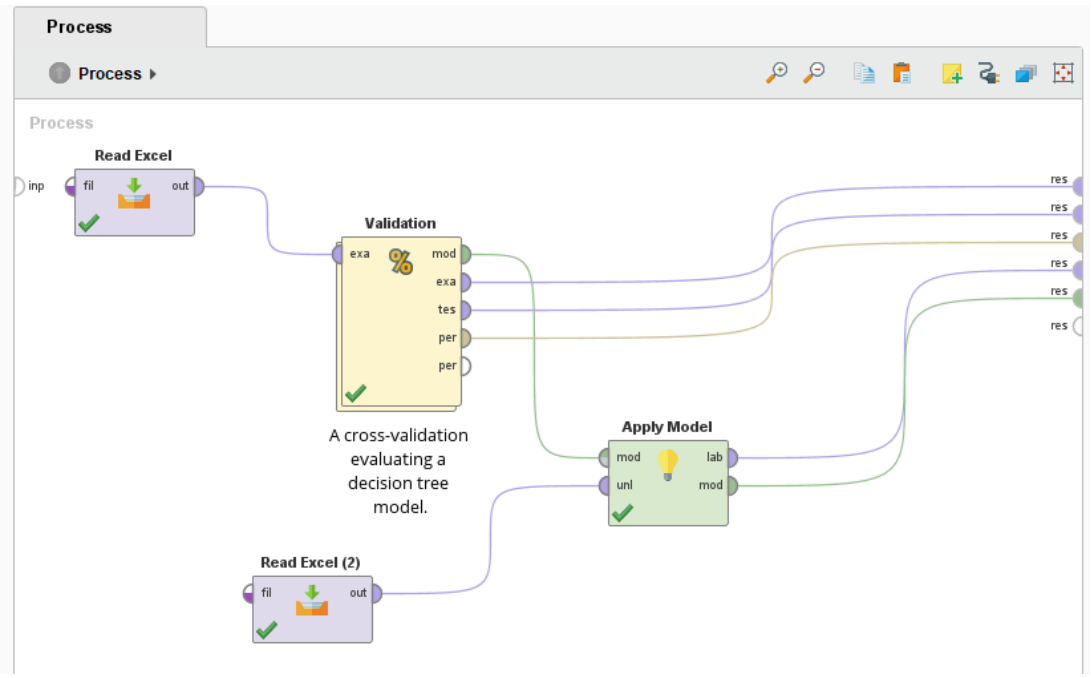


Figure 2. C4.5 Algorithm Model Design

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12968
e-ISSN : 2541-2019
p-ISSN : 2541-044X

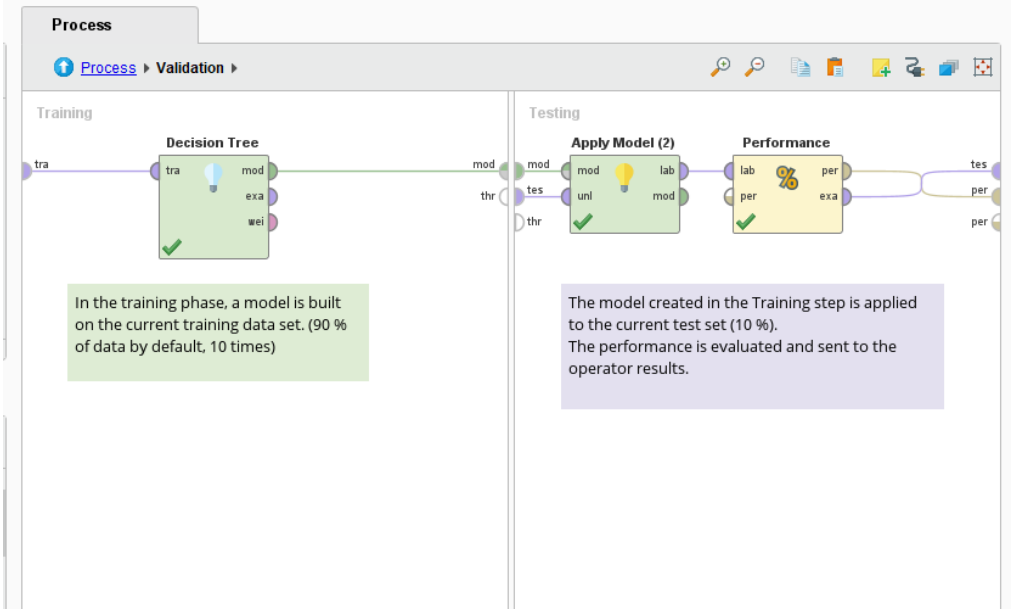


Figure 3. Design of the C4.5 Algorithm Model in Split Validation Components

From Figure 2 and Figure 3, the modeling of the C4.5 algorithm in the RapidMiner application can be explained, namely:

a. ReadExcel is an operator used to import datasets training with Excel file types (.xls, .xlsx).
b. ReadExcel (2) is an operator used to import datasets testing with Excel file types (.xls, .xlsx).
c. Validation is an operator that divides the total data from the dataset into training data and testing data. The validation method used is Split Validation which divides training and testing data based on the specified split ratio value.
d. Decision Tree is the classification method used in this research..
e. Apply Model is the operator that runs the C4.5 algorithm used in this research
f. Performance is used to measure the accuracy performance of the model

Testing Using RapidMiner Tools

The results of data processing using a decision tree model according to the RapidMiner software can be seen in Figure 4 as follows :

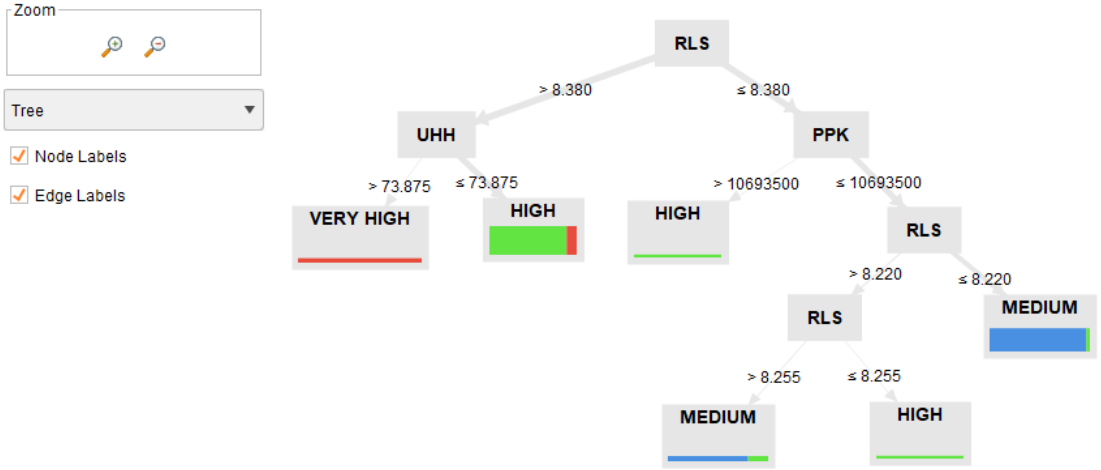


Figure 4. Decision Tree on RapidMiner

Based on Figure 4, this is a decision tree generated in Rapidminer with the rules that can be seen in the text view in Figure 5 as follows:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12968
e-ISSN : 2541-2019
p-ISSN : 2541-044X

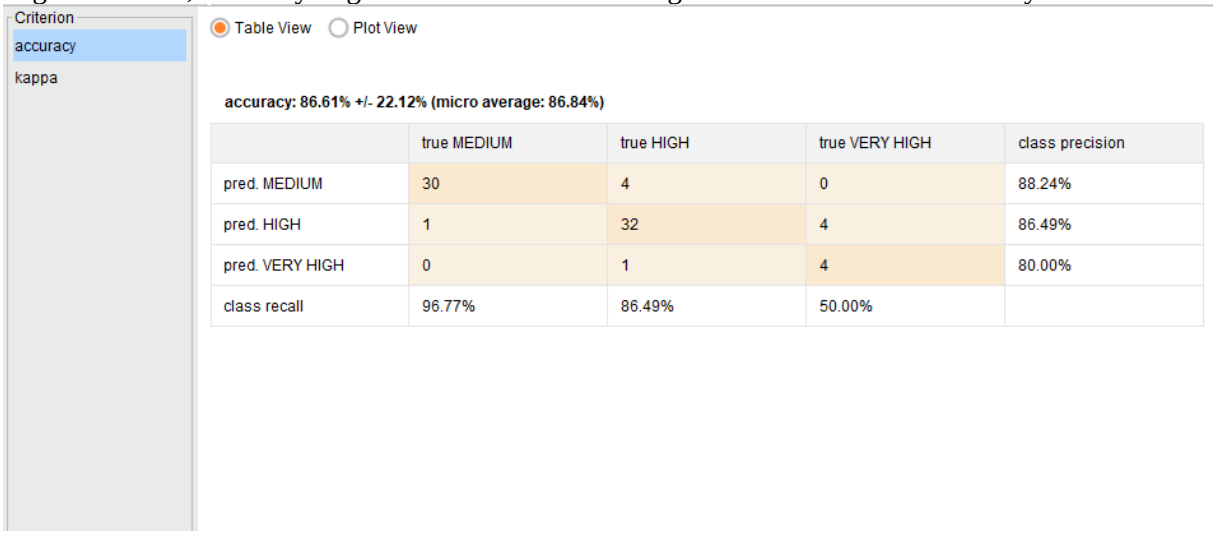
PerformanceVector

```
PerformanceVector:
accuracy: 86.61% +/- 22.12% (micro average: 86.84%)
ConfusionMatrix:
True:    MEDIUM  HIGH    VERY HIGH
MEDIUM: 30       4       0
HIGH:    1       32      4
VERY HIGH:       0       1       4
kappa: 0.771 +/- 0.375 (micro average: 0.771)
ConfusionMatrix:
True:    MEDIUM  HIGH    VERY HIGH
MEDIUM: 30       4       0
HIGH:    1       32      4
VERY HIGH:       0       1       4
```

Figure 5. Rule Decision Tree on RapidMiner

Data Accuracy

The results of applying the C4.5 Algorithm using RapidMiner software with the X Validation operator obtained an accuracy value of 86.61%. Where for Class Precision the Middle label prediction is 88.24%, High is 86.49%, and Very High is 80.00%. The following are the results of the accuracy obtained



	true MEDIUM	true HIGH	true VERY HIGH	class precision
pred. MEDIUM	30	4	0	88.24%
pred. HIGH	1	32	4	86.49%
pred. VERY HIGH	0	1	4	80.00%
class recall	96.77%	86.49%	50.00%	

Figure 6. Algorithm Accuracy Value C4.5

DISCUSSIONS

Based on data processing conducted by research from the last 4 years, 2018 to 2021, it can be said that the HDI value is good and the category value increases every year based on the tests carried out. Processing that has been done with the rapid miner application obtained rules or rules. The results obtained from predictions with the c4.5 algorithm, obtained an accuracy value of 86.61%, for class precision in the middle label prediction of 88.24%, high label 86.49%, very high label 80.00%, and class recall of 96.77%, 86.49%, and 50.00%.

*name of corresponding author

CONCLUSION

Based on the results of the research conducted, the author draws several conclusions, including:

a. The highest factor influencing the Human Development Index is the average length of schooling.
b. The results of testing the human development indexes for the Regency/City of West Sumatera using the C4.5 Algorithm and the RapidMiner application have an accuracy value of 86.61%. It can be concluded that predictions of human development are based on the average length of schooling, life expectancy, and per capita expenditure.
c. Suggestions that can be given for further research so that it can be carried out using other prediction methods to produce a better level of accuracy, make comparisons with other data mining methods, and further research on the attributes that will later be selected.

REFERENCES

Andesti, C. L., Lonanda, F., Azizah, N., Info, A., Mining, D., & Language, E. (2022). *Potential for Improvement of Students' s English Language with*. *5*(1), 1–10.

Anggraeni, L., & Arum R, P. (2022). Analisis Cluster Menggunakan Algoritma K-Means Pada Provinsi Sumatera Barat Berdasarkan Indeks Pembangunan Manusia Tahun 2021. *Prosiding Seminar Nasional UNIMUS*, *5(1)*, 636–646.

Attamami, N., Triayudi, A., & Aldisa, R. T. (2023). *Analisis Performa Algoritma Klasifikasi Naive Bayes dan C4 . 5 untuk Prediksi Penerima Bantuan Jaminan Kesehatan*. *7*(2).

Azizah, N., Andesti, C. L., & Sirait, W. (2022). Implementasi Algoritma C4. 5 dalam Menentukan Tingkat Kepuasan Pelanggan di Kadai Kopi Lasi. *IndraTech*, *3*(2), 1–7. http://ojs.stmikindragiri.ac.id/index.php/jit/article/view/103%0Ahttps://ojs.stmikindragiri.ac.id/index.php/jit/article/download/103/84

Fersellia, F., Utami, E., & Yaqin, A. (2023). Sentiment Analysis of Shopee Food Application User Satisfaction Using the C4.5 Decision Tree Method. *Sinkron*, *8*(3), 1554–1563. https://doi.org/10.33395/sinkron.v8i3.12531

Ginting, V. S., Kusrini, K., & Taufiq, E. (2020). Implementasi Algoritma C4.5 untuk Memprediksi Keterlambatan Pembayaran Sumbangan Pembangunan Pendidikan Sekolah Menggunakan Python. *Inspiration: Jurnal Teknologi Informasi Dan Komunikasi*, *10*(1), 36–44. https://doi.org/10.35585/inspir.v10i1.2535

Nas, C. (2021). Data Mining Prediksi Minat Calon Mahasiswa Memilih Perguruan Tinggi Menggunakan Algoritma C4.5. *Jurnal Manajemen Informatika (JAMIKA)*, *11*(2), 131–145. https://doi.org/10.34010/jamika.v11i2.5506

Pratiwi, I. A. A. S., & Wijayanto, A. W. (2019). Klasifikasi Indeks Pembangunan Manusia dengan Metode K-Nearest Neighbor dan Support Vector Machine di Pulau Jawa. *Jurnal Ilmu Komputer*, *15*(1), 8–21. https://ojs.unud.ac.id/index.php/jik/article/download/68565/44248

Putra, R. M., Asril, E., & Taslim, T. (2018). Prediksi Indeks Pembangunan Manusia Menggunakan Algoritma C4.5 di Kabupaten Kampar. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, *9*(2), 204–214. https://doi.org/10.31849/digitalzone.v9i2.1584

Rismayanti, R., Damayanti, F., & Khairunnisa, K. (2018). Penerapan Data Mining Algoritma C4.5 dalam Menentukan Rekam Jejak Kinerja Dosen STT Harapan Medan. *SinkrOn*, *3*(1), 99–104. https://doi.org/10.33395/sinkron.v3i1.173

Sinaga, B., Manurung, J., Mayana, N., Tarigan, B., Feronika, S., & Sitepu, B. (2022). *Application of with C4 . 5 algorithm to measure the level of student satisfaction with student services*. *7*(3), 2134–2143.

Sintawati, I. D., Widiarina, W., & Mariskhana, K. (2021). Application of the C4.5 Algorithm on the Effect of Watching YouTube Videos On the Development of Early Childhood Creativity. *SinkrOn*, *6*(1), 120–126. https://doi.org/10.33395/sinkron.v6i1.11116

Siregar, Y. S., Sembiring, B. O., Hasdiana, H., Dewi, A. R., & Harahap, H. (2021). Algorithm C4.5 in mapping the admission patterns of new Students in Engineering Computer. *SinkrOn*, *6*(1), 80–90. https://jurnal.polgan.ac.id/index.php/sinkron/article/view/11154

*name of corresponding author

Sudalyo, R. A. T., & Mukti, B. (2022). Analysis University of Surakarta KIP Scholarship Recipients Using the Fuzzy MADM Method and C-45. *Sinkron*, *7*(1), 9–16. https://doi.org/10.33395/sinkron.v7i1.11221

Sumpena, Akbar, Y., Nirat, & Henky, M. (2019). Comparison of C4 . 5 Algorithm and Naïve Bayes for Last Information on ICU Patients. *Journal Publications & Informatics Engineering Research*, *4*(1), 88–94.

Yunus, W., & Haba, A. R. K. (2023). Implementasi Algoritma C.45 Dalam Prediksi Penyakit Kanker. *Jurnal Indonesia : Manajemen Informatika Dan Komunikasi*, *4*(1), 70–76. https://doi.org/10.35870/jimik.v4i1.114

*name of corresponding author