IMPLEMENTATION OF C4.5 ALGORITHM IN CLASSIFYING BREAST CANCER BASED ON MENOPAUSE AGE

Wulan Dari[1*)], Nisrina Miranda[2]

Sistem Informasi
Universitas Nusa Mandiri
www.nusamandiri.ac.id
wulan.wld@nusamandiri.ac.id [1*)], nisrinamiranda@gmail.com[2]
(*) Corresponding Author

***Abstract*—** Breast cancer is a malignant tumor that can attack breast tissue, is a disease that is most feared by women. Although based on recent findings not only women are affected by breast cancer, it turns out that men can get breast cancer, although it is still very rare. Breast cancer is one type of cancer that is often experienced by women in Indonesia. Data mining is a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to interact and identify useful information and related knowledge from large databases. Breast cancer is much feared by women as well as young and old age which can lead to death, if as early as possible for a full examination, this lump initially shrinks, but over time it enlarges, then sticks to the skin or causes changes in the skin of the breast (nipple). But the skin or nipple is pulled inward (retraction), light redness, or browning, until swelling, shrinking or sore breasts get worse the old ones will get bigger and deeper so that they can crush the one breast often smells bad and bleeds easily . C4.5 algorithm and decision tree are two inseparable models, because to build a decision tree, C4.5 algorithm is needed. Decision trees are one of the most popular classification methods because they are easy for humans to interpret. A decision tree is a predictive model using a tree structure or hierarchical structure. The concept of a decision tree is to convert data into a decision tree and decision rules.

Keywords: Breast Cancer, Data Mining, C4.5 Algorithm, Decision Tree

Intisari— Kanker payudara adalah tumor ganas yang bisa menyerang jaringan payudara, merupakan penyakit yang paling ditakuti oleh kaum wanita. Meskipun berdasarkan penemuan terakhir tak hanya kaum wanita saja yang terkena kanker payudara ternyata kaum laki-laki pun bisa terkena kanker payudara ini walaupun masih sangat jarang terjadi. Kanker payudara merupakaan salah satu jenis kanker yang sering dialami oleh perempuan di Indonesia. *Data Mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk menginteraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. Kanker Payudara banyak ditakuti oleh para wanita maupun usia muda dan usia lanjut yang dimana bisa mengakibatkan meninggal, bila sedini mungkin untuk diperiksakan keseluruhannya, benjolan ini awalnya mengecil, namun lama kelamaan membesar, kemudian menempel pada kulit atau menyebabkan perubahan pada kulit payudara (puting). Tapi kulit atau puting tertarik ke dalam (retraksi), kemerahan muda, atau kecokelatan, hingga bengkak, payudara menyusut atau sakit semakin menjadi-jadi yang lama akan menjadi lebih besar dan lebih dalam sehingga bisa menghancurkan yang satu payudara sering berbau tidak sedap dan mudah berdarah. *Algoritma C4.5* dan pohon keputusan merupakan dua model yang tak terpisahkan, kerena untuk membangun sebuah pohon keputusan, dibutuhkan algoritma C4.5. Pohon keputusan adalah salah satu metode klasifikasi yang paling populer karena mudah diinterperensi manusia. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan.

Kata Kunci: Kanker Payudara, *Data Mining*, Algoritma C4.5, Pohon Keputusan

INTRODUCTION

Breast cancer is a malignant tumor that can attack breast tissue, is a disease that is most feared by women. Although based on recent findings, not only women are affected by breast cancer, it turns out that men can get breast cancer, although it is still very rare. From observations, generally breast cancer patients can't be helped because it's too late to be treated (Arsittasari, Estiwidani, & Setiyawati, 2017)

Currently, cancer is the second leading cause of death in the world. It is estimated that 9.6 million people worldwide died from cancer in 2018.

In general, 1 in 6 cases of death is caused by cancer. Breast cancer is a type of cancer with the second highest number of occurrences after lung cancer. The incidence of breast cancer in the world in 2018 was 2.09 million cases.(Utami & Muhartati, 2020)

Breast cancer is one type of cancer that is often experienced by women in Indonesia. Breast cancer is the most dominant type of cancer in Indonesia, which beats cervical cancer. The increasing number of cancer patients is estimated to be the main cause of the increase in the economic burden because the costs to be borne are very large.

Breast cancer is much feared by women as well as young and old age which can lead to death, if as early as possible for a full examination, this lump initially shrinks, but over time it enlarges, then sticks to the skin or causes changes in the skin of the breast (nipple). But the skin or nipple is pulled inward (retraction), light redness, or browning, until swelling, shrinking or sore breasts get worse the old ones will get bigger and deeper so that they can crush the one breast often smells bad and bleeds easily.

The C4.5 algorithm is the most popular method and also the best method and is often used by researchers to make a decision tree (Decission Tree). To build a decision tree, Algorithm C4.5 will choose an attribute as the root, create a branch for each value, divide the cases into branches, then repeat the process for each branch until all cases in the branch get the same class (Lorena, Zarman, & Hamidah, 2014). The C4.5 algorithm has advantages, namely flexible, easy to understand, and interesting, because it can be visualized in the form of images in the form of decision trees.(Gorunescu, 2011)

Several studies have been conducted by researchers in analyzing breast cancer using the C4.5 . Algorithm method.

Related research on the application of the C4.5 Algorithm for Classification of Hernia Disk and Spondylolisthesis in the Vertebral Column, namely that the process of classifying herniated disc disease and Vertebral Column Spondylolisthesis obtains an accuracy value of 89% and an average running time of 0.00912297 seconds.(Handayani, 2019)

The next research that has been carried out is on the Analysis of Data Mining Techniques "C4.5 Algorithm and K-Nearest Neighbor" To Diagnose Diabetes Mellitus In this study, comparing the C4.5 algorithm method and the K-NN algorithm used in classifying diabetes mellitus. The results of this study indicate a value of 76.105% on the C4.5 algorithm and 79,1436% on the K-NN algorithm.(Karyono, 2016)

The next study, entitled Application of the C4.5 Classification Algorithm for the diagnosis of Breast Cancer, obtained high accuracy results after an evaluation using the C4.5 algorithm, with an accuracy value for the C4.5 classification algorithm of 94.56% and for the AUC value of 0.941..(Hermawanti, 2012)

Research has also been carried out with the title Data Mining Implementation to Predict Student Study Period Using the C4.5 Algorithm. The results of this study prove that the C4.5 algorithm is more accurate than the analysis carried out by student analysts because it is able to analyze the level of timeliness of students completing their study period.(Haryati, Sudarsono, & Suryana, 2015)

Based on the background of the problems mentioned above, the advantages of the C4.5 Algorithm are that it can produce a decision tree that is very easy to interpret, is efficient in handling an attribute of discrete type, has an acceptable level of accuracy, and can handle attributes with numeric and numeric types. discrete.(Purwanto, Primajaya, & Voutama, 2020)

MATERIALS AND METHODS

Data mining is a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to interact and identify useful information and related knowledge from large databases. The term data mining has the essence as a scientific discipline whose main goal is to find, explore, or mine knowledge from the data or information that we have.(Mustafa, Ramadhan, & Thenata, 2018)

Research methodology

Research methodology is a way to find out the results of a specific problem, where the problem is also called a research problem. In Methodology, researchers also use a variety of different ways to solve existing research problems. Different sources will mention that the use of different types of methods is to solve problems.

Quantitative research is a process of finding knowledge that uses data in the form of numbers as a tool to analyze information about what you want to know (Anwar Hidayat, 2012)

Data in a research activity requires data collection and data collection methods that have an important role to provide the accuracy and quality of the data to be used in the research process. There are methods that can be used to support research, namely:

1. Literature Study

Literature study is an activity to collect information relevant to the topic or problem that is the object of research. This information can be obtained from books, scientific works, theses, dissertations, encyclopedias, internet, and other

sources. By conducting a literature study, researchers can take advantage of all the information and thoughts that are relevant to their research

2. Data Type

The data collection procedure used in the research is to group it into two groups, namely primary data and secondary data (Daniel, Boyatzis, & Mckee, 2019):

a. Secondary data

Secondary data is data obtained indirectly from documentation, literature, books, journals and other information related to the problem to be studied.(Swastina, 2018)

b. Primary data

Primary data is data which in this study is the result of research. The primary data in this study is the test data using the Decision Tree C4.5 . Algorithm. (Swastina, 2018)

RESULT AND DISCUSSION

Testing the C4.5 . Algorithm Method

At this stage, experiments and testing of the method used are carried out, namely, calculating and getting the rules that exist in the proposed algorithm, namely the C4.5 algorithm.

The C4.5 algorithm is a tree structure, where there are nodes that describe attributes, on each branch that describes the results of the tested attributes, and for each leaf describes a class (Nasrullah, 2018) The C4.5 algorithm recursively visits each decision node, choosing the optimal division, until it can no longer be divided. The C4.5 algorithm uses a concept of information gain or entropy reduction to choose the optimal division (Han & Kamber, 2006)

Entropy is a probability distribution in information theory and can be adopted into the C4.5 Algorithm to measure the level of homogeneity of the class distribution of a data set. An illustration that can be drawn is that the higher the entropy level of a data set, the more homogeneous the class distribution in the data set is.(Pearce E C, 2012)

The steps taken are as follows:

1. Counting the number of right and left cases and the entropy value of all cases. From the available training data, it is known that the number of cases on the right is 19 records, and the number on the left is 31 records, the total number of cases is 50 cases, so that the total entropy is obtained.

$$Entropy\ (S) = \sum_{i=1}^{n} -\text{pi} * \log 2\text{pi} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

=(-19/50*Log2(19/50))+(-31/50*Log2(31/50))
=0.958042

2. Calculate the entropy value and gene value of each attribute. The highest gene value is the attribute that is the root of the decision tree to be made. Attribute entropy is calculated by the following formula:

$$Gain\ (S, A) = \text{Entropy (s)} - \sum_{i=1}^{n} * Entropy\ (Si) \quad \ldots\ldots (2)$$

Data Selection

In the first data mining process that is carried out is data selection from breast cancer, from the data obtained there are 10 attributes then the attributes used in the data mining process are taken 5 attributes from 10 attributes. The variables used in the data mining process use the attributes of Age, Menopause, Dig-malig, Breast and Breast-quad. Based on the case table 1, start looking for the highest entropy value from the highest gain value to determine the root for the decision tree.

Table 1. Case Table

No	Age	Meno-pouse	dig-malig	Breast	breast-quad
1	30-39	Premeno	3	left	left-bottom
2	40-49	Premeno	2	right	right-top
3	40-49	Premeno	2	left	left-bottom
4	60-69	ge40	2	right	right-top
5	40-49	Premeno	2	right	right-bottom
....					
50	40-49	Premeno	1	right	right-top

Table 2. Calculation

		Number of Cases	Right	Left	Entropy	Gain
Total		50	19	31	0,958042	
Age						0,035346
	30-39	3	1	2	0,918296	
	40-49	16	8	8	1	
	50-59	20	9	11	0,992774	
	60-69	11	2	9	0,684038	
menopouse						0,020202
	premeno	27	12	15	0,991076	
	ge40	21	8	13	0,958712	
	lt40	2	0	2	0	
dig-malig						0,018626
	3	9	2	7	0,764205	
	2	25	10	15	0,970951	
	1	16	7	9	0,988699	
breast-quad						0,192435
	left-bottom	24	4	20	0,650022	
	left-top	13	7	6	0,995727	
	right-top	8	6	2	0,811278	
	right-bottom	1	1	0	0	
	center	4	1	3	0,811278	

From the calculation in the table 2, the highest gain is breast-quad and the highest entropy is the upper left, so the case is made as follows:

Table 3. Calculation of Entropy

	Number of Cases	Right	Left	Entropy	Gain
Total	50	19	31	0,958042	

The following is in Table 3 regarding the calculation of entropy using the entropy value search formula, namely:

$$Entropy\ (S) = \sum_{i=1}^{n} -pi * log2pi \quad \text{.....(3)}$$

=((-K3/J3)*IMLOG2(K3/J3)+(-L3/J3)*IMLOG2(L3/J3))
=0,958042
Explanation:

- K3 = Number of Right Attributes
- L3 = Number of Left Attributes
- J3 = Number of Cases
- IMLOG2 = Formula for Log perhitungan calculation

Next do the calculation on the Gain

$$Gain\ (S, A) = Entropy\ (s) - \sum_{i=1}^{n} * Entropy\ (Si) \quad \text{.....(4)}$$

=(M3)-((J5/J3)*M5)-((J6/J3)*M6)-((J7/J3)*M7)-((J8/J3)*M8)
= 0,035346
Explanation:

- M3 = The sum Entropy of Total
- J5 = Number of Cases from the age attribute
- J3 = Number of Cases
- M5 = The sum of the entropy of the age attribute

Based on the calculations in table 3 will perform the same calculations so that no further calculations are needed and will make a decision tree. From these results, a decision tree can be drawn as shown in Figure 1.

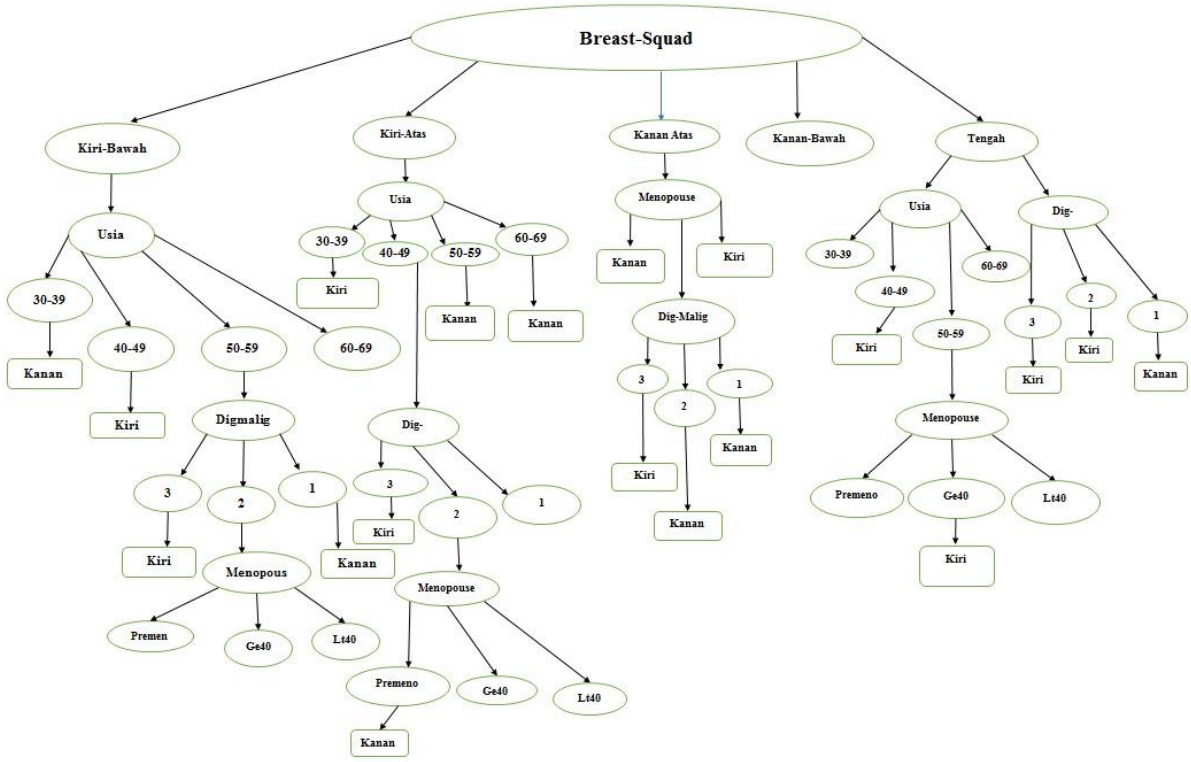

Figure 1 Decision Tree

By looking at the decision tree in Figure 1, it is known that all cases have been included in the class. Thus, the decision tree in Figure 1. is the last decision tree formed.

CONCLUSION

Cancer can be treated as early as possible to be examined so that lumps in the breast can be removed or removed so that they do not spread to the other breast and take regular treatment. It can be seen that the most common cancers are based on age and where the cancer is located. From the results of research that has been carried out on data sets obtained from the Uci Repository about breast cancer, it can be concluded that this method is quite accurate in determining the classification for breast cancer based on the right or left that often occurs.

REFERENCE

Anwar Hidayat. (2012, October). Pengertian dan Penjelasan Penelitian Kuantitatif – Lengkap. *Statistikian*, p. Pengertian dan Penjelasan Penelitian Kuantitatif –.

Arsittasari, T., Estiwidani, D., & Setiyawati, N. (2017). Faktor-Faktor Yang Berhubungan Dengan Kejadian Kanker Payudara Di Rsud Kota Yogyakarta Tahun 2016. *Jurnal Kebidanan*, 1–90.

Daniel, G., Boyatzis, R., & Mckee, A. (2019). Metode Penelitian. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Gorunescu, F. (2011). Data Mining: Concepts, Models, and Techniques. *Verlag Berlin Heidelberg : Springer*.

Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. *Morgan Kaufmann*.

Handayani, I. (2019). *Penyakit Disk Hernia Dan Spondylolisthesis Dalam Kolumna Vertebralis*. *1*(2), 83–88. https://doi.org/10.12928/JASIEK.v13i2.xxxx

Haryati, S., Sudarsono, A., & Suryana, E. (2015). Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu). *Jurnal Media Infotama*, *11*(2), 130–138.

Hermawanti, L. (2012). Penerapan algoritma klasifikasi c4.5 untuk diagnosis penyakit kanker payudara. *Jurnal Teknik Unisfat*, *7*(1), 57–64.

Karyono, G. (2016). Analisis Tektnik Data Mining "Algoritma C4.5 dan K-Nereset Neighbor" untuk Mendiagnosa Penyakit Diabetes Mellitus. *Seminar Nasional Teknologi Informasi*, 77–82. Retrieved from http://news.palcomtech.com/wp-content/uploads/downloads/2016/06/IT13_Giat-Karyono.pdf

Lorena, S., Zarman, W., & Hamidah, I. (2014). Analisis Dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik. *Pros. Semin. Nas. Apl. Sains Dan Teknol*, 263–272.

Mustafa, M. S., Ramadhan, M. R., & Thenata, A. P. (2018). Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Creative Information Technology Journal*, *4*(2), 151. https://doi.org/10.24076/citec.2017v4i2.106

Nasrullah, A. H. (2018). Penerapan Metode C4.5 untuk Klasifikasi Mahasiswa Berpotensi Drop Out. *ILKOM Jurnal Ilmiah*, *10*(2), 244–250. https://doi.org/10.33096/ilkom.v10i2.300.244-250

Pearce E C. (2012). *Anatomi dan Fisiologi untuk Paramedis*. Jakarta: PT. Gramedia Pustaka Utama.

Purwanto, A., Primajaya, A., & Voutama, A. (2020). Penerapan Algoritma C4 . 5 dalam Prediksi Potensi Tingkat Kasus Pneumonia di Kabupaten Karawang. *Jurnal Sistem Dan Teknologi Informasi*, *08*(4), 390–396. https://doi.org/10.26418/justin.v8i4.41959

Swastina, L. (2018). Penerapan Algoritma C4 . 5 Untuk Penentuan Jurusan Mahasiswa. *Gema Aktualita*, *2*(1), 93–98.

Utami, F. S., & Muhartati, M. (2020). Kader sadar kanker payudara. *Jurnal Inovasi Abdimas Kebidanan (Jiak)*, *1*(1), 19–22. https://doi.org/10.32536/jpma.v1i1.66