Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

COMPARISON OF BREAST CANCER CLASSIFICATION USING DECISION TREE ID3 AND K-NEAREST NEIGHBORS ALGORITHM TO PREDICT THE BEST PERFORMANCE OF ALGORITHM

Zyhan Faradilla Daldiri [1*)] , Desti Fitriati [2]

Informatics Engineering [1,2]
Universitas Pancasila [1,2]
Jakarta, Indonesia
zyhandilla1308@gmail.com [1*)] , desti.fitriati@univpancasila.ac.id [2]

(*) Corresponding Author

Abstract

One of the leading causes of death is cancer. The most common cancer in women is breast cancer. Breast cancer (Carcinoma mammae) is a malignant neoplasm originating from the parenchyma. Breast cancer ranks first in terms of the highest number of cancers in Indonesia and is among the first contributors to cancer deaths. Globocan data in 2020 shows that the number of new breast cancer cases reached 68,858 (16.6%) of the total 396,914 new cancer cases in Indonesia. Meanwhile, deaths reached more than 22 thousand cases (Romkom, 2022). This death rate is increasing due to insufficient information about breast cancer's early symptoms and dangers. Of this lack of information, a system is needed that can provide information about breast cancer, such as early diagnosis. Several parameters and classification data mining techniques can predict which patients will develop breast cancer and which do not. In this study, a comparison of the classification of breast cancer using the Decision Tree ID3 algorithm and the K-Nearest Neighbors algorithm will be carried out. Attribute data consists of Menopause, Tumor-Size, Node-Caps, Deg-Malig, Breast-Squad, and Irradiant. The main objective of this study is to improve classification performance in breast cancer diagnosis by applying feature selection to several classification algorithms. The Decision Tree ID3 algorithm has an accuracy rate of 93.333%, and the K-Nearest Neighbors algorithm has an accuracy rate of 76.6667%.

Keywords: Breast cancer; Woman; Classification; Comparison; Decision Tree Algorithm; K-Nearest Neighbors Algorithm

Abstrak

Salah satu penyebab kematian yang utama adalah kanker. Kanker yang paling umum pada wanita adalah kanker payudara. Kanker payudara (Carcinoma mammae) didefinisikan sebagai suatu penyakit neoplasma ganas yang berasal dari parenchyma. Kanker payudara menempati urutan pertama terkait jumlah kanker terbanyak di Indonesia serta menjadi salah satu penyumbang kematian pertama akibat kanker. Data Globocan tahun 2020, jumlah kasus baru kanker payudara mencapai 68.858 kasus (16,6%) dari total 396.914 kasus baru kanker di Indonesia. Sementara itu, untuk jumlah kematiannya mencapai lebih dari 22 ribu jiwa kasus(Romkom,2022). Angka kematian ini meningkat karena kurangnya informasi tentang gejala awal dan bahaya dari kanker payudara itu sendiri. Dari kurangnya informasi tersebut, maka dibutuhkan sebuah sistem yang dapat memberikan informasi tentang penyakit kanker payudara seperti diagnosa secara dini. Teknik data mining klasifikasi dapat digunakan untuk memprediksi pasien mana yang terkena kanker payudara dan tidak dengan beberapa parameter yang ada. Dalam penelitian ini akan dilakukan perbandingan klasifikasi penyakit kanker payudara dengan menggunakan algoritma Decision Tree ID3 dan algoritma K-Nearest Neighbors. Atribut data yang digunakan terdiri dari Menopause, Tumor-Size, Node-Caps, Deg-Malig, Breast, Breast-Squad dan Irradiant. Tujuan utama penelitian ini adalah untuk meningkatkan peforma klasifikasi pada diagnosis kanker payudara dengan menerapkan seleksi fitur pada beberapa algoritma klasifikasi. Algoritma Decision Tree ID3 memiliki tingkat akurasi yaitu 93,333% dan algoritma K-Nearest Neighbors memilki tingkat akurasi yaitu 76,6667%.

Kata kunci : Kanker Payudara; Wanita; Klasifikasi; Perbandingan; Algoritma Decision Tree; Algoritma K-Nearest Neighbors

INTRODUCTION

Information technology continues to progress from time to time. Information technology has provided various helpful information and data for people's lives. Technology also plays a role in advancing various institutions, one of which is health agencies. The data generated in the health sector can be in the form of data about diseases that are considered deadly such as cancer, which of course this data can be used to dig deeper into information related to cancer itself, both for the treatment or prevention of patients who have not or have experienced cancer (Ramadhan & Kurniawati, 2020).

Information from the data can be used to find new patterns of information by processing or extracting information. This new pattern can be used to classify cancer patients based on recurrence or non-recurrence of the disease. This knowledge can help the medical side handle patients to minimize the number of cancer patients who experience recurrence of the disease.

Currently, breast cancer is a type of cancer that is very frightening for women around the world, and this also applies in Indonesia. Cancer prevention can be done early by being aware of cancer at its initial appearance so that it can have a high cure rate (Hasanah, 2018). Therefore, it is necessary to carry out prevention efforts to increase public awareness in recognizing the symptoms and risks, especially breast cancer, in determining appropriate preventive measures and early detection.

Breast cancer is a malignant tumor formed from breast cells that grow and develop uncontrollably so that it can spread to organs near the breast or to other parts of the body (Buana Briliant, 2020). Breast cancer is still a disease with a high mortality rate in women. Based on data from the WHO (World Health Organization), in 2020, the number of breast cancer cases will reach 68,858 (16.6%) of 396,914 new cancer cases in Indonesia. Meanwhile, deaths reached more than 22 thousand cases (Kementrian Kesehatan Indonesia, 2022).

Judging the significant mortality rate in Indonesia makes us aware of the importance of knowing the symptoms of breast cancer to prevent increased deaths caused by breast cancer. The research data is used to test the accuracy of breast cancer diagnosis using public data from the UCI Machine Learning Repository breast cancer dataset. The dataset is tested by processing the data using data mining.

Data mining is a series of processes to find values in data sets that are not known manually (Prakarsya & Prambayun, 2020). So that the breast cancer dataset that has extensive data can be analyzed in making decisions using data mining. One of the data mining methods in the data exploration process is the data classification technique.

With this data classification technique, it is possible to classify the breast cancer dataset using the ID3 Decision Tree algorithm and the K-Nearest Neighbors algorithm with the results of the accuracy of the classification of each of these algorithms. This study will be compared for the best performance evaluation in breast cancer detection. Thus, the results of this comparison can be used as a reference for diagnosing breast cancer at an early stage.

Research objectives and methods used

1. To find out which algorithm performance produces higher accuracy.
2. To improve classification performance in breast cancer diagnosis.

Benefits of research

1. Assist in developing further research related to algorithm comparison problems.
2. It can be used as a reference in research on breast cancer.

Research contribution

This study can provide knowledge about the results of research that has been carried out in obtaining information about comparing the K-Neirest Neighbors Algorithm and the ID3 Decision Tree Algorithm in classifying breast cancer.

RESEARCH METHODS

Types of research

The data obtained in this study were sourced from the UCI Machine Learning Repository website, entitled breast cancer dataset. The data taken is data randomly seen from the classification results obtained.

Research Time and Place

The study was conducted for two weeks, the first week at the end of May 2022 and the second week at the beginning of June 2022. The study was carried out at home by analyzing and calculating the data obtained.

Data Mining

Data mining is looking for information models that can be useful in database storage. Data mining is one of a series of knowledge search processes in a database known as Knowledge Discovery In Database (KDD). KDD relates to integration techniques and scientific discoveries for interpretation and visualization so that they are easy to understand (Zulfa et al., 2020). Some techniques frequently mentioned in the Data Mining

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v5i2.406
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

literature include clustering, classification, association rule mining, and neural networks.

Classification

Classification is a data processing technique that determines data into predetermined groups or classes (Bahri & Lubis, 2020). This method helps produce functions or models that explain classes in data used to predict classes of objects that have not been labeled (Arie Wijaya et al., 2021). The classification method refers to forming data groups by applying algorithms in the classification process.

Decision Tree

Decision Tree ID3 is a tree modeling technique that can implement a series of decisions. The ID3 Decision Tree has tree nodes that represent the attributes that have been tested, and each branch is a division of the test results, and leaf nodes represent certain class groups (Setio et al., 2020). The process in the ID3 Decision Tree is to change the form of a data table into a tree model (Rionaldi, 2022). In the calculation phase using the Decision Tree ID3 algorithm, there are several stages, namely:

1. Specifies attribute data with data values in the dataset.
2. Looking for the same data.
3. Searching for data is false.
4. Searching for data is accurate.
5. Calculate entropy with the following formula (1):

$$\text{Entropy (S)} = -\frac{Nmj}{N}\sum i\, P_{mj}^{(i)}\, log_2\, P_{mj}^{(i)} \text{ ....} (1)$$

6. Add up the entropy results.
7. Make a decision tree according to the results obtained.

K-Nearest Neighbors

The K-Nearest Neighbor (KNN) algorithm is a case search approach that calculates the closeness between new and old cases based on matching the weights of several existing features (Setiawan et al., 2020). K in K-NN is the number of neighbors that will be taken to determine decisions (Mustafa & Wayan Simpen, 2019). The K-NN algorithm in this study has several steps, which are as follows:

1. Converting qualitative data into quantitative data.
2. Calculate using the Euclidean formula on the data. The Euclidean formula is as follows:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(qi-pi)^2} \text{ ....} (2)$$

3. Determine the closest Distance from the results of the Euclidean formula calculation (2).
4. Determine the K to be taken.
5. Determine the existing prediction results by the taken K.

Procedure

In this study, a comparison was made between the Decision Tree ID3 algorithm and the K-Nearest Neighbors algorithm to get the best classification results. The stages used in this research are:

1. Dataset processing.
2. Calculations using the Decision Tree ID3 algorithm.
3. Calculations using the K-NN algorithm.
4. Perform data comparisons using the two algorithms.
5. Calculate the accuracy (3) of the two algorithms with the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}x100\% \text{ ....} (3)$$

Data, Instruments, and Data Collection Techniques

This study uses primary data and as many as 30 test data. The technique of collecting criteria and alternative data to be used in system testing was obtained by taking a dataset from the UCI Machine Learning Repository site, namely https://archive.ics.uci.edu/ml/datasets/breast+cancer. For information on breast cancer data, this study used six attributes, namely Menopause, Tumor-Size, Node-Caps, Deg-Malig, Breast, Breast-Squad, and Irradiant.

Data analysis technique

The data in this study consisted of quantitative and qualitative data. For qualitative data, researchers convert them into numbers (quantitative data) to make it easier to calculate both manual and system calculations.

RESULTS AND DISCUSSION

Based on the research stages, the first step is processing breast cancer datasets. The attributes of the dataset can be seen in Table 1.

Table 1. Attributes of the Breast Cancer Dataset

Attribute
Menopause

Attribute
Tumor Size
Node caps
Deg Malig
breast
Breast Squad

From these attributes, 30 data samples were taken randomly to be tested. The sample data can be seen in Table 2.

Table 2. Breast Cancer Sample Data

Menopause	Tumor size	inv-nodes	node-caps	deg-malig	breast	breast squad	irradiat
ge40	15-19	0-2	no	2	right	left_up	no
ge40	25-29	0-2	no	3	left	right_up	no
ge40	30-34	0-2	no	1	right	right_up	no
ge40	30-34	0-2	no	2	left	left_low	yes
ge40	25-29	0-2	no	3	left	right_low	yes
ge40	30-34	0-2	no	3	right	left_up	yes
ge40	15-19	0-2	no	3	right	left_up	yes
ge40	30-34	0-2	no	1	right	Central	no
ge40	15-19	0-2	no	1	left	right_low	no
ge40	20-24	0-2	no	1	left	left_low	no
ge40	15-19	0-2	no	3	right	left_up	yes
ge40	15-19	0-2	no	2	left	left_up	yes
ge40	30-34	0-2	yes	2	right	right_up	yes
lt40	30-34	0-2	no	1	left	left_low	no
lt40	15-19	0-2	no	2	left	left_low	no
lt40	30-34	0-2	no	3	right	left_up	no
lt40	15-19	0-2	no	3	right	left_up	no
premeno	30-34	0-2	no	2	right	right_up	yes
premeno	20-24	0-2	no	3	right	left_low	yes
premeno	35-39	0-2	yes	3	right	left_up	yes
premeno	35-39	0-2	yes	3	right	left_low	yes
premeno	25-29	0-2	no	1	right	left_low	yes
premeno	25-29	0-2	no	2	left	left_up	yes
premeno	20-24	0-2	no	2	left	right_low	no
premeno	30-34	0-2	no	3	right	left_up	yes
premeno	20-24	0-2	no	1	left	right_low	no
premeno	15-19	0-2	no	1	left	left_low	no
premeno	20-24	0-2	no	2	left	Central	no
premeno	35-39	0-2	no	2	right	right_up	no
premeno	20-24	0-2	no	3	left	left_up	yes

After taking the test data, the next step is calculating the data classification using the Decision Tree ID3 algorithm. The first stage in Decision Tree ID3, which determines data attributes with values in the test data, can be seen in Table 3.

Table 3. Value Data Based on Attributes

Attribute (m)	Value Data (j)
Menopause	Ge40; Lt40; Premeno
Tumor Size	15-19; 20-24; 25-29 30-34; 35-39
Node Caps	Yes; No
Deg Malig	1; 2; 3
breast	Left; Right
Breast Squad	Left Low; Left Up; Central Right Low; Right Up

After determining the value of the test data according to the attributes, calculations can be made to determine the same data, determine false data, determine accurate data, calculate entropy, and add up the results of the entropy values in each of the data attributes with N as the amount of data being tested. The results can be seen in Table 4.

Table 4. Entropy Calculation Results

m	j	Nmj/N	FALSE	TRUE	entropy	
Menopause	**ge40**	13/30	6/13	7/13	0,4314818959	0,848017758

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v5i2.406
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

m	j	Nmj/N	FALSE	TRUE	entropy	
	lt40	4/30	4/4	0/4	0	
	premeno	13/30	5/13	8/13	0,416535862	
Tumor Size	**15-19**	8/30	5/8	3/8	0,2545157341	0,9354977189
	20-24	6/30	4/6	2/6	0,1836591668	
	25-29	4/30	1/4	3/4	0,1081704166	
	30-34	9/30	4/9	5/9	0,297322818	
	35-39	3/30	1/3	2/3	0,09182958341	
Node-caps	**yes**	3/30	0/3	3/3	0	0,8919684539
	no	27/30	15/27	12/27	0,8919684539	
Deg-malig	**1**	8/30	7/8	1/8	0,1449505182	0,8027951013
	2	10/30	5/10	5/10	0,3333333333	
	3	12/30	3/12	9/12	0,3245112498	
Breast	**left**	14/30	9/14	5/14	0,438800114	0,9478315823
	right	16/30	6/16	10/16	0,5090314682	
Breast-squad	**left low**	8/30	4/8	4/8	0,2666666667	0,8466241924
	left up	11/30	3/11	8/11	0,3099620101	
	Central	2/30	2/2	0/2	0	
	right low	4/30	3/4	1/4	0,1081704166	
	right up	5/30	3/5	2/5	0,1618250991	

From the results of the entropy calculation, tree data from deg-malig can be taken because the results obtained have the smallest value. Then proceed with calculating the entropy again based on the data from the deg-malig with value = 1, which can be seen in Table 5.

Table 5. Data Based on Deg Malig Attribute With Value=1

Menopause	Tumor size	inv-nodes	node-caps	deg-malig	breast	breast squad	irradiat
ge40	30-34	0-2	no	1	right	right_up	no
ge40	30-34	0-2	no	1	right	Central	no
ge40	15-19	0-2	no	1	left	right_low	no
ge40	20-24	0-2	no	1	left	left_low	no
lt40	30-34	0-2	no	1	left	left_low	no
premeno	25-29	0-2	no	1	right	left_low	yes
premeno	20-24	0-2	no	1	left	right_low	no
premeno	15-19	0-2	no	1	left	left_low	no

From this data, entropy calculations can be made based on the values in the deg-malig attribute data. The results of the calculation of the deg-malig attribute data with value = 1 can be seen in Table 6.

Table 6. Entropy Calculation Results Based on Deg Malig Data With Value=1

m	j	Nmj/N	FALSE	TRUE	entropy	
Menopause	**ge40**	4/8	4/4	0/4	0	0,3443609378
	lt40	1/8	1/1	0/1	0	
	premeno	3/8	2/3	1/3	0,3443609378	
Tumor Size	**15-19**	2/8	2/2	0/2	0	0
	20-24	2/8	2/2	0/2	0	
	25-29	1/8	0/1	1/1	0	
	30-34	3/8	3/3	0/3	0	
Breast	**left**	5/8	5/5	0/5	0	0,3443609378
	right	3/8	2/3	1/3	0,3443609378	
Breast-squad	**left low**	4/8	3/4	1/4	0,4056390622	0,4056390622
	central	1/8	1/1	0/1	0	
	right low	2/8	2/2	0/2	0	
	right up	1/8	1/1	0/1	0	

Repeated patterns can be used to calculate the entropy of the deg-malig attribute data with value=2 and value=3. After these calculations, the tree results will be obtained as in Figure 1.

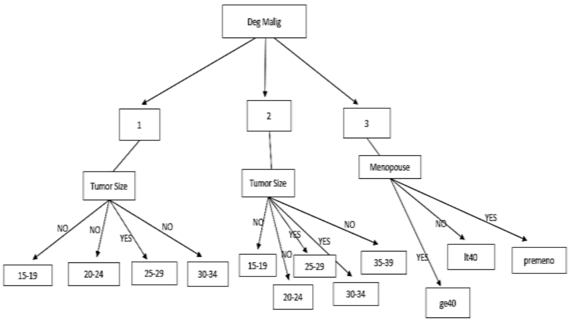

Figure 1. Classification of ID3 Decision Tree Results

Furthermore, testing the existing data with the KNN algorithm by performing the steps on the mentioned methodology by changing the qualitative data into quantitative ones. The data is as in Table 7.

Table 7. Changes in Value Data

Attribute	Value (Qualitative)	Value (Quantitative)
Menopause	Ge40	1
	Lt40	2
	Premeno	3
Tumor Size	15-19	1
	20-24	2
	25-29	3
	30-34	4
	35-39	5
Node Caps	Yes	1
	No	2
breast	Left	1
	Right	2
Breast Squad	Left Low	1
	Left Up	2
	Central	3
	Right Low	4
	Right Up	5

After changing the value of the data, the test dataset will be like in Table 8.

Table 8. Data Changes Based on Quantitative Value Data

Menopause	Tumor size	inv-nodes	node-caps	deg-malig	breast	breast squad
1	1	0-2	2	2	2	2
1	3	0-2	2	3	1	5
1	4	0-2	2	1	2	5
1	4	0-2	2	2	1	1
1	3	0-2	2	3	1	4
1	4	0-2	2	3	2	2
1	1	0-2	2	3	2	2
1	4	0-2	2	1	2	3
1	1	0-2	2	1	1	4
1	2	0-2	2	1	1	1
1	1	0-2	2	3	2	2
1	1	0-2	2	2	1	2
1	4	0-2	1	2	2	5
2	4	0-2	2	1	1	1
2	1	0-2	2	2	1	1
2	4	0-2	2	3	2	2
2	1	0-2	2	3	2	2
3	4	0-2	2	2	2	5
3	2	0-2	2	3	2	1
3	5	0-2	1	3	2	2
3	5	0-2	1	3	2	1
3	3	0-2	2	1	2	1
3	3	0-2	2	2	1	2
3	2	0-2	2	2	1	4
3	4	0-2	2	3	2	2
3	2	0-2	2	1	1	4
3	1	0-2	2	1	1	1
3	2	0-2	2	2	1	3
3	5	0-2	2	2	2	5
3	2	0-2	2	3	1	2

The change in the data value can be calculated using the Euclidean formula, determining the closest Distance in ascending order and the

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v5i2.406
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

result of the specified K. The results of these calculations are as follows.

Table 9. KNN Calculation Results With 1st Data Trial

Menopause	Tumor size	inv-nodes	node-caps	deg-malig	breast	breast squad	irradiat	Euclidean	Nearest Distance (k)	k=5	Classify
1	1	0-2	2	2	2	2	N	0	1	Y	N
1	3	0-2	2	3	1	5	N	15	24	N	
1	4	0-2	2	1	2	5	N	19	25	N	
1	4	0-2	2	2	1	1	Y	11	18	N	
1	3	0-2	2	3	1	4	Y	10	14	N	
1	4	0-2	2	3	2	2	Y	10	14	N	
1	1	0-2	2	3	2	2	Y	1	2	Y	Y
1	4	0-2	2	1	2	3	N	11	18	N	
1	1	0-2	2	1	1	4	N	6	8	N	
1	2	0-2	2	1	1	1	N	4	7	N	
1	1	0-2	2	3	2	2	Y	1	2	Y	Y
1	1	0-2	2	2	1	2	Y	1	2	Y	Y
1	4	0-2	1	2	2	5	Y	19	25	N	
2	4	0-2	2	1	1	1	N	13	22	N	
2	1	0-2	2	2	1	1	N	3	6	N	
2	4	0-2	2	3	2	2	N	11	18	N	
2	1	0-2	2	3	2	2	N	2	5	Y	N
3	4	0-2	2	2	2	5	Y	22	27	N	
3	2	0-2	2	3	2	1	Y	7	9	N	
3	5	0-2	1	3	2	2	Y	22	27	N	
3	5	0-2	1	3	2	1	Y	23	29	N	
3	3	0-2	2	1	2	1	Y	10	14	N	
3	3	0-2	2	2	1	2	Y	9	13	N	
3	2	0-2	2	2	1	4	N	10	14	N	
3	4	0-2	2	3	2	2	Y	14	23	N	
3	2	0-2	2	1	1	4	N	11	18	N	
3	1	0-2	2	1	1	1	N	7	9	N	
3	2	0-2	2	2	1	3	N	7	9	N	
3	5	0-2	2	2	2	5	N	29	30	N	
3	2	0-2	2	3	1	2	Y	7	9	N	

Judging from Table 9, the KNN calculation is based on the test data on the first data to find out the prediction results obtained in the KNN classification. In this study, 30 existing data will be tested using the KNN algorithm so that the calculation is repeated with test data on the 2nd to 30th data. After calculating the KNN classification data on the 30 data, the KNN classification can be generated in Table 10.

Table 10. KNN Calculation Classification Results

No	Menopause	Tumor size	inv-nodes	node-caps	deg-malig	breast	breast squad	irradiat	Result
1	ge40	15-19	0-2	N	2	right	left_up	N	**Y**
2	ge40	25-29	0-2	N	3	left	right_up	N	**Y**
3	ge40	30-34	0-2	N	1	right	right_up	N	**N**
4	ge40	30-34	0-2	N	2	left	left_low	Y	**N**
5	ge40	25-29	0-2	N	3	left	right_low	Y	**Y**
6	ge40	30-34	0-2	N	3	right	left_up	Y	**Y**
7	ge40	15-19	0-2	N	3	right	left_up	Y	**Y**
8	ge40	30-34	0-2	N	1	right	Central	N	**N**
9	ge40	15-19	0-2	N	1	left	right_low	N	**N**
10	ge40	20-24	0-2	N	1	left	left_low	N	**N**
11	ge40	15-19	0-2	N	3	right	left_up	Y	**Y**
12	ge40	15-19	0-2	N	2	left	left_up	Y	**Y**
13	ge40	30-34	0-2	Y	2	right	right_up	Y	**Y**
14	lt40	30-34	0-2	N	1	left	left_low	N	**Y**

No	Menopause	Tumor size	inv-nodes	node-caps	deg-malig	breast	breast squad	irradiat	Result
15	lt40	15-19	0-2	N	2	left	left_low	N	**N**
16	lt40	30-34	0-2	N	3	right	left_up	N	**Y**
17	lt40	15-19	0-2	N	3	right	left_up	N	**Y**
18	premeno	30-34	0-2	N	2	right	right_up	Y	**Y**
19	premeno	20-24	0-2	N	3	right	left_low	Y	**Y**
20	premeno	35-39	0-2	Y	3	right	left_up	Y	**Y**
21	premeno	35-39	0-2	Y	3	right	left_low	Y	**Y**
22	premeno	25-29	0-2	no	1	right	left_low	Y	**Y**
23	premeno	25-29	0-2	no	2	left	left_up	Y	**Y**
24	premeno	20-24	0-2	no	2	left	right_low	N	**N**
25	premeno	30-34	0-2	no	3	right	left_up	Y	**Y**
26	premeno	20-24	0-2	no	1	left	right_low	N	**N**
27	premeno	15-19	0-2	no	1	left	left_low	N	**Y**
28	premeno	20-24	0-2	no	2	left	Central	N	**N**
29	premeno	35-39	0-2	no	2	right	right_up	N	**N**
30	premeno	20-24	0-2	no	3	left	left_up	Y	**Y**

This study compared test data using the Decision Tree ID3 algorithm and the KNN algorithm as the classification method. The results of the predictions using the Decision Tree ID3 algorithm and the KNN algorithm are equated with the results of the existing test data so that the data is obtained, like in Table 11.

Table 11. Data Comparison

No	Menopause	Tumor size	inv-nodes	node-caps	deg-malig	breast	breast squad	irradiat	ID3	KNN
1	ge40	15-19	0-2	no	2	right	left_up	no	no	yes
2	ge40	25-29	0-2	no	3	left	right_up	no	yes	yes
3	ge40	30-34	0-2	no	1	right	right_up	no	no	no
4	ge40	30-34	0-2	no	2	left	left_low	yes	yes	no
5	ge40	25-29	0-2	no	3	left	right_low	yes	yes	yes
6	ge40	30-34	0-2	no	3	right	left_up	yes	yes	yes
7	ge40	15-19	0-2	no	3	right	left_up	yes	yes	yes
8	ge40	30-34	0-2	no	1	right	Central	no	no	no
9	ge40	15-19	0-2	no	1	left	right_low	no	no	no
10	ge40	20-24	0-2	no	1	left	left_low	no	no	no
11	ge40	15-19	0-2	no	3	right	left_up	yes	yes	yes
12	ge40	15-19	0-2	no	2	left	left_up	yes	no	yes
13	ge40	30-34	0-2	yes	2	right	right_up	yes	yes	yes
14	lt40	30-34	0-2	no	1	left	left_low	no	no	yes
15	lt40	15-19	0-2	no	2	left	left_low	no	no	no
16	lt40	30-34	0-2	no	3	right	left_up	no	no	yes
17	lt40	15-19	0-2	no	3	right	left_up	no	no	yes
18	premeno	30-34	0-2	no	2	right	right_up	yes	yes	yes
19	premeno	20-24	0-2	no	3	right	left_low	yes	yes	yes
20	premeno	35-39	0-2	yes	3	right	left_up	yes	yes	yes
21	premeno	35-39	0-2	yes	3	right	left_low	yes	yes	yes
22	premeno	25-29	0-2	no	1	right	left_low	yes	yes	yes
23	premeno	25-29	0-2	no	2	left	left_up	yes	yes	yes
24	premeno	20-24	0-2	no	2	left	right_low	no	no	no
25	premeno	30-34	0-2	no	3	right	left_up	yes	yes	yes
26	premeno	20-24	0-2	no	1	left	right_low	no	no	no
27	premeno	15-19	0-2	no	1	left	left_low	no	no	yes
28	premeno	20-24	0-2	no	2	left	Central	no	no	no
29	premeno	35-39	0-2	no	2	right	right_up	no	no	no
30	premeno	20-24	0-2	no	3	left	left_up	yes	yes	yes

The accuracy between the algorithms is calculated from the comparison data to determine which algorithm is more accurate. The accuracy test is carried out to measure how much accuracy is obtained from the tests carried out. Accuracy can be calculated using the formula (3) above. By calculating the accuracy, the accuracy of the

Decision Tree ID3 algorithm is 93.333%, while the accuracy of the KNN algorithm is 76.6667%.

CONCLUSIONS AND SUGGESTIONS

Conclusions

From this research, it can be concluded that the classification calculation using the Decision Tree ID3 Algorithm has more accurate results than the classification calculation using the KNN Algorithm on the breast cancer dataset. These results can be seen through the accuracy value obtained by calculating the classification using the Decision Tree ID3 algorithm, which results in 93.333% having a more excellent accuracy value than the KNN algorithm, which results in 76.6667%. So it can be said that using Decision Tree ID3 classification calculations on breast cancer estimates has accurate results on the incidence of breast cancer so that it can be used as a reference in predicting and preventing breast cancer early.

Suggestions

In further research, the classification algorithm with Decision Tree ID3 is expected to be used in the health sector and other fields by re-comparing other classification algorithms to find the best accuracy results.

REFERENCE

Arie Wijaya, Y., Bahtiar, A., Kaslani, K., & R, N. (2021). Analisa klasifikasi menggunakan algoritma decision tree pada data log firewall. *Jurnal Sistem Informasi Dan Manajemen*, *9*(3), 256–264. https://ejournal.stmikgici.ac.id/index.php/jursima/article/view/303/196

Bahri, S., & Lubis, A. (2020). Metode klasifikasi devision tree untuk memprediksi juara english premier league. *Jurnal Sintaksis: Pendidikan Guru Sekolah Dasar, IPA, IPS, Dan Bahasa Indonesia*, *2*(1), 63–70. https://jurnal.stkipalmaksum.ac.id/index.php/Sintaksis/article/view/47

Buana Briliant, J. (2020). *Proses asuhan gizi terstandar dan kualitas hidup pada pasien kanker payudara di RS Panti Rapih Yogyakarta*. http://eprints.poltekkesjogja.ac.id/id/eprint/2734

Hasanah, A. (2018). *Gambaran faktor yang mempengaruhi pemeriksaan diri ke pelayanan kesehatan pada penderita kanker payudara*. http://repository.unair.ac.id/id/eprint/73250

Kementrian Kesehatan Indonesia. (2022, February 4). *Kanker payudara paling banyak di Indonesia, Kemenkes targetkan pemeratan pelayanan kesehatan*. Https://Www.Kemkes.Go.Id/Article/View/22020400002/Kanker-Payudara-Paling-Banyak-Di-Indonesia-Kemenkes-Targetkan-Pemerataan-Layanan-Kesehatan.Html.

Mustafa, Syukri. M., & Wayan Simpen, I. (2019). Implementasi algoritma K-Nearest Neighbor (KNN) untuk memprediksi pasien terkena penyakit diabetes pada Puskesmas Manyampa Kabupaten Bulukumba. *Pro Siding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, *VIII*(1), 1–10. https://ejurnal.dipanegara.ac.id/index.php/sisiti/article/view/1%20-10

Prakarsya, A., & Prambayun, A. (2020). Implementasi data mining untuk prediksi pembayaran virus HIV/AIDS di Bandar Lampung dengan teknik decision tree. *Jurnal Siskomti*, *3*(2), 18–26. https://www.ejournal.lembahdempo.ac.id/index.php/SISKOMTI/article/download/120/93/

Ramadhan, I., & Kurniawati, K. (2020). Data mining untuk klasifikasi penderita kanker payudara berdasarkan data dari University Medical Center menggunakan algoritma naïve bayes. *JURIKOM (Jurnal Riset Komputer)*, *7*(1), 21–27. https://doi.org/10.30865/jurikom.v7i1.1755

Rionaldi, R. (2022, July 24). *Klasifikasi data mining menggunakan algoritma decision tree*. Https://Lab_adrk.Ub.Ac.Id/Id/Klasifikasi-Data-Mining-Menggunakan-Algoritma-Decision-Tree/.

Setiawan, A., A. I. Nahusuly, B. J., Aulia Yuliandi Putri, F., Raditya, A., & Susrama Mas Diyasa, I. G. (2020). Case based reasoning menggunakan algoritma K-Nearest Neigbors untuk penanganan penyakit ikan cupang hias. *Jurnal Teknologi Informasi Dan Komunikasi*, *XV*(2), 1–5. https://doi.org/https://doi.org/10.33005/scan.v15i2.2082

Setio, P. B. N., Saputro, D. R. S., & Winarno, B. (2020). Klasifikasi dengan pohon keputusan berbasis algoritme C4.5. *Prosiding Seminar Nasional Matematika* *3*, *3*, 64–71. https://journal.unnes.ac.id/sju/index.php/prisma/

Zulfa, I., Rayuwati, R., & Koko, K. (2020). Implementasi data mining untuk menentukan strategi penjualan buku bekas dengan pola pembelian konsumen menggunakan metode apriori. *Teknika: Jurnal Sains Dan Teknologi*, *16*(1), 69–82. https://doi.org/10.36055/tjst.v16i1.7601