FEATURE SELECTION USING INFORMATION GAIN ON THE K-NEAREST NEIGHBOR (KNN) AND MODIFIED K-NEAREST NEIGHBOR (MKNN) METHODS FOR CHRONIC KIDNEY DISEASE CLASSIFICATION

TUGAS AKHIR

Disusun Sebagai Salah Satu Syarat
Untuk Memperoleh Gelar Sarjana Teknik
Pada Jurusan Teknik Informatika

Oleh

AWELDRI RAMADHAN
NIM. 11850110539

UIN SUSKA RIAU

FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SULTAN SYARIF KASIM RIAU
PEKANBARU
2023

LEMBAR PERSETUJUAN

SELEKSI FITUR INFORMATION GAIN PADA METODE K-NEAREST NEIGHBOR (KNN) DAN MODIFIED K-NEAREST NEIGHBOR (MKNN) UNTUK KLASIFIKASI PENYAKIT GINJAL KRONIS

TUGAS AKHIR

Oleh

AWELDRI RAMADHAN
NIM. 11850110539

Telah diperiksa dan disetujui sebagai Laporan Tugas Akhir
di Pekanbaru, pada tanggal 14 Desember 2023

Pembimbing I,

Elvia Budianita, S.T., M.Cs
NIP. 19860629 201503 2 007

LEMBAR PENGESAHAN

SELEKSI FITUR INFORMATION GAIN PADA METODE K-NEAREST NEIGHBOR (KNN) DAN MODIFIED K-NEAREST NEIGHBOR (MKNN) UNTUK KLASIFIKASI PENYAKIT GINJAL KRONIS

Oleh

AWELDRI RAMADHAN

NIM. 11850110539

Telah dipertahankan di depan sidang dewan penguji
sebagai salah satu syarat untuk memperoleh gelar Sarjana Teknik
pada Universitas Islam Negeri Sultan Syarif Kasim Riau

Pekanbaru, 14 Desember 2023

Mengesahkan,

Dekan,

KEMENTRIAN AGAMA
UIN SUSKA RIAU
FAKULTAS SAINS DAN TEKNOLOGI
★ UIN SULTAN SYARIF KASIM RIAU ★

Dr. Hartono, M.Pd
NIP. 19640301 199203 1 003

Ketua Jurusan,

Iwan Iskandar, M.T
NIP. 19821216 201503 1 003

DEWAN PENGUJI

Ketua	: Dr. Lestari Handayani, S.T., M.kom
Pembimbing I	: Elvia Budianita, S.T., M,Cs
Penguji I	: Fadhilah Syafria, S.T., M.Kom
Penguji II	: Siti Ramadhani, S.Pd, M.Kom

LEMBAR HAK ATAS KEKAYAAN INTELEKTUAL

Tugas Akhir yang tidak diterbitkan ini terdaftar dan tersedia di Perpustakaan Universitas Islam Negeri Sultan Syarif Kasim Riau adalah terbuka untuk umum dengan ketentuan bahwa hak cipta pada penulis. Referensi kepustakaan diperkenankan dicatat, tetapi pengutipan atau ringkasan hanya dapat dilakukan seizin penulis dan harus disertai dengan kebiasaan ilmiah untuk menyebutkan sumbernya.

Penggandaan atau penerbitan sebagian atau seluruh Tugas Akhir ini harus memperoleh izin dari Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau. Perpustakaan yang meminjamkan Tugas Akhir ini untuk anggotanya diharapkan untuk mengisi nama, tanda peminjaman dan tanggal pinjam.

LEMBAR PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan didalam daftar pustaka.

Pekanbaru, 14 Desember 2023

Yang membuat pernyataan,

AWELDRI RAMADHAN

NIM. 11850110539

LEMBAR PERSEMBAHAN

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

Alhamdulillahirabbil ‘alamin

Dengan mengucapkan syukur kepada Allah Subhanahu Wa Ta’ala, dan Shalawat serta salam kepada Nabi Muhammad Shallallahu ‘Alaihi Wasallam, alhamdulillah akhirnya Tugas Akhir ini telah Saya selesaikan.

Semua pencapaian ini saya persembahkan untuk keluarga saya yaitu ayah (Khairul) dan Ibu (Yusriati) serta Kakak, Abang, dan Adik saya (Muliati, Fathul Hidayah, dan Vazel) yang selalu memberikan motivasi, saran, dan dukungan, sehingga laporan Tugas Akhir ini dapat terselesaikan.

Dan ucapan terimakasih kepada Ibu Elvia Budianita, ST, M.Cs selaku dosen pembimbing saya, yang telah memberikan arahan dan bimbingan kepada saya. Sekali lagi terimakasih Ibu atas semua ilmu dan nasehat yang telah Ibu berikan. dan terakhir untuk teman-teman kelas TIF A 18, terimakasih atas dukungan yang telah kalian berikan.

Semoga Tugas Akhir ini bermanfaat bagi pembacanya.

SURAT PERNYATAAN

Saya yang bertandatangan di bawah ini :

Nama	: Aweldri Ramadhan
NIM	: 11850110539
Tempat/Tgl. Lahir	: Sukaddamai, 11 Desember 1999
Fakultas	: Sains dan Teknologi
Prodi	: Teknik Informatika
Judul Skripsi	: Feature Selection using Information Gain on the K-Nearest Neighbor (KNN) and Modified K-Nearest Neighbor (MKNN) Methods for Chronic Kidney Disease Classification

Menyatakan dengan sebenar-benarnya bahwa:

1. Penulisan Skripsi dengan judul sebagaimana tersebut diatas adalah hasil pemikiran penelitian saya sendiri.
2. Semua kutipan pada karya tulis saya ini sudah disebutkan sumbernnya.
3. Oleh karena itu, Skripsi saya ini, saya nyatakan bebas dari plagiat.
4. Apabila dikemudian hari terbukti terdapat plagiat dalam penulisan Skripsi saya tersebut, maka saya bersedia menerima sanksi sesuai perundang-undangan.

Demikian Surat Pernyataan ini saya buat dengan penuh kesadaraan dan tanpa paksa pihak manapun juga.

Pekanbaru, 14 Desember 2023

mbuat pernyataan,

SEPULUH RIBU RUPIAH 10000 TGL. 20 METERAI TEMPEL C4903AKX768391450

AWELDRI RAMADHAN

NIM. 11850110539

Feature Selection using Information Gain on the K-Nearest Neighbor(KNN) and Modified K-Nearest Neighbor (MKNN) Methods for Chronic Kidney Disease Classification

Aweldri Ramadhan[1*], Elvia Budianita[2], Fadhilah Syafria[3], Siti Ramadhani[4]

[1,2,3,4]Dept. of Informatics Engineering, Universitas Islam Negeri Sultan Syarif Kasim, Indonesia (8 pt)
[1]aweldriramadhan@gmail.com, [2]elvia.budianita@uin-suska.ac.id, [3]fadhilah.syafria@uin-suska.ac.id,
[4]siti.ramadhani@uin-
suska.ac.id

Abstract.
Purpose: Kidneys has an important role in the human excretory system. Unhealthy kidneys can affect kidney function. It is important to know the symptoms of chronic kidney disease. One data mining technique that can be applied is the classification technique to determine whether a person has chronic kidney disease or not based on the symptoms (attributes) obtained from medical records. The symptoms of chronic kidney disease obtained amount to 24 symptomsor attributes,
Methods/Study design/approach: In this research, the classification of chronic kidney disease is performed using the information gain feature selection method and the KNN and MKNN classification methods. The number of data used is 400 data with 2 classes, namely chronic kidney disease (CKD) and non-chronic kidney disease (non-CKD).
Result/Findings: Based on the test results, it was found that the hemo (Hemoglobin) attribute has the highest information gain value, which is 0.6255. The best accuracy for the KNN classification method is 96.61%, and for the MKNN method, it is 98%.
Novelty/Originality/Value: The purpose of information gain feature selection is to choose features or attributes that significantly influence chronic kidney disease.

Keywords: Chronic Kidney Disease, Information Gain, KNN, MKNN

Received November 2023 / **Revised** December 2023 / **Accepted** December 2023

INTRODUCTION

Kidneys are very important organs in the human excretory system and generally, every human has two kidneys in the body [1]. Based on data from the Global Burden of Disease, chronic kidney disease (CKD) was ranked 27th in the world in 1990 and 18th in 2010. According to data from the Indonesian Ministry of Health, around 2 out of every 1,000 Indonesians, or 499,800 people, had CKD in 2013. Around 6 out of every 1,000 or 1,499,400 Indonesians suffered from kidney stones [2].

While in Indonesia, nearly 70,000 people suffer from chronic kidney disease, and this number will continue to increase every year. Based on a survey by the World Health Organization (WHO), the number of chronic kidney disease patient in Indonesia will increase by 46% from 1955 to 2025 [3]. Therefore, it is important for everyone to understand the symptoms of kidney disease, one of which is by using classification techniques in data mining.

Classification techniques in data mining allow us to understand the causes of the symptoms of a particular condition. Data mining has long been known as a method that can extract knowledge or insights from a large amount of data. In the healthcare industry, data mining is often used to provide predictions about the diagnosis of various diseases [3]. Classification is a technique for finding a model that serves to group and reveal class differences in data. Classification aims to predict the class or label so that the data can be analyzed based on the known class groups [4]. Classification is a technique for finding a model that servesto group and reveal class differences in data. Classification aims to predict the class or label so that the datacan be analyzed based on the known class groups [5]. The results of the study explain that the value of K and the number of training data have a significant effect on accuracy. This study produced the highest accuracy value of 97.61% with K=1 and 90% training data.

In this study, the method applied for classification is Modified K-Nearest Neighbor (MKNN), which is a development of K-Nearest Neighbor (KNN) by incorporating additional steps, including validity calculation and threshold of weight value [5]. In the process, the MKNN classification method uses all

DOI: 10.24014/coreit.v9i2.26834

available features, so it can sometimes cause problems because of the presence of some irrelevant features,and this affects the accuracy of the classification results. To improve the accuracy results and reduce detection errors, the feature selection process is carried out first before the classification process [4].

Feature selection is a special stage in data mining that aims to identify important attributes. Its main purpose is to remove irrelevant and redundant attributes, so that algorithms can operate more efficiently and improve accuracy[6]. Information Gain is a feature selection method that is useful for ranking attributes based on their relevance. The higher the value of an attribute on Information Gain, the more important that attribute is in data analysis [7]. Previous research that focused on Information Gain attribute selection is “Model Prediction of Diabetes Disease Using Bayesian Classification and Information Gain for Feature Selection and Adaptive Boosting for Data Weighting” [8]. From the results of the study, it can be seen that the accuracy of Naive Bayes increased from 74.01% to 79.10%, accompanied by an AUC value of 0.8722.

In this study, the application of Information Gain in the classification of chronic kidney disease using the Modified K-Nearest Neighbor (MKNN) and KNN methods on the dataset in the UCI machine learning repository will be studied, and also to compare the performance between classification using information gain and without using information gain.

METHODS

The research steps taken in this study are shown in Figure 1

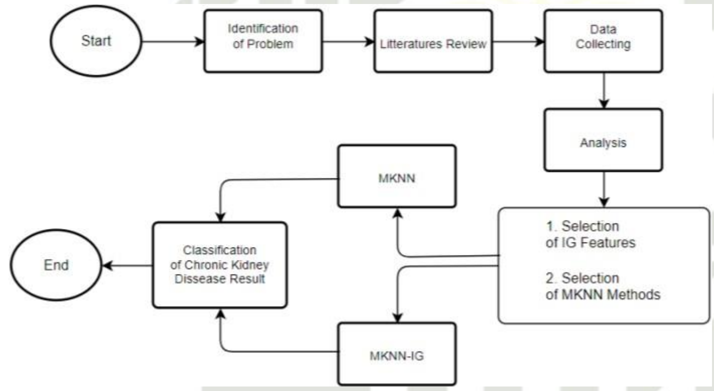

Start
Identification of Problem
Litteratures Review
Data Collecting
Analysis
MKNN
1. Selection of IG Features
2. Selection of MKNN Methods
Classification of Chronic Kidney Dissease Result
End
MKNN-IG

Figure 1. Research Steps and Methods.

Litterature Review

In this step, a literature review is used to gather information and references related to the research topic. These literature sources can include previous thesis reports, scientific articles, notes, and relevant literature from various experts. The purpose is for the research to be based on relevant theories and conducted systematically

Data Collecting

In this study, data from patients with chronic kidney disease were used. The data were obtained from the UCI machine learning repository. The dataset used consists of 400 data samples, consisting of 24 attributes and 2 classes, namely the Chronic Kidney Diseases (CKD) class or affected by chronic kidney disease and the not Chronic Kidney Disease (non-CKD) class or not affected by chronic kidney disease.

Data Mining

Data mining has long been known as a method that can extract knowledge or insights from large amounts of data. The main goal in Data Mining is to extract valuable knowledge from a wide dataset, by transforming it into a more understandable and useful format for future use [9]. In the healthcare industry, data mining is often used to provide predictions about the diagnosis of various diseases[3].

Features Selection

Feature selection is a specialized step in data analysis that aims to identify attributes that are relevant. Thegoal is to eliminate attributes that are not significant or redundant, so that the algorithm can run more efficiently and improve the accuracy of its results [6].

Information Gain

Information gain is a popular feature selection method that has the advantage of ranking each attribute. The larger the value of an attribute, the more relevant it is to use. This is necessary because there are some features that are not needed and can make the algorithm's performance inefficient [10].

Here are the steps involved in performing feature selection with information gain:

a. Calculating of the entropy value of the initial dataset or the entropy value of the class E(D). The initial E(D) value, or also known as the expected information value, is calculated based on the number of data and the number of classes used. In this study, the dataset consists of 400 data and is divided into two different classes, so the calculation is as follows:

N = 400
P1 = Class of CKD, total data in Class of CKD as many as 250 data
P1 = 250/400 = 0,625, log_2 = -0,6781
P2 = Class of non-CKD, total data in Class of non-CKD as many as 150 data
P2 = 150/400 = 0,375, log_2 = -1,4150

Next, the entropy value is calculated using the equation formula (1) below, so the result is,

E(D) = (-0,625*-0,6781) + (-0,375*-1,4150)
E(D) = 0,4242 + 0,5306
E(D) = 0,9544

b. Calculating the entropy value for each attribute in the dataset, with the formula of:

$$Entropy\ (S) = \sum_{i}^{c} -Pi\ log2\ Pi \tag{1}$$

After the initial E(D) value is obtained, the entropy value for each attribute is then calculated. Theattribute to be calculated is the ba attribute, as follows:

Table 1. BA Features

S	BA	CKD(S_1)	NON-CKD(S_2)
S_1	1	22	0
S	0	228	150

Calculate the entropy value for the first subset, namely BA 1 in the CKD and non-CKD classes.

N = 378
P1 = 228/378 = 0,603, log2 nya = -0,7298
P2 = 150/378 = 0,397, log2 nya = -1,3328
E(D2) = (-0,603*-0,7298) + (-0,397*-1,3328)
E(D2) = 0,4401+0,5291
E(D2) = 0,9692

c. After all of the Entropy values have been calculated, the next process is to calculate the Information Gain value using the formula of:

$$IG = E(D) - (\frac{D1}{D}E(D1) + \frac{D2}{D}E(D2)) \tag{2}$$

Which is:

E(D) = Initial Entropy dataset
D = Total of the entire sample from data
D1 = The number of values in subset1 of attribute1
D2 = The number of values in subset1 of attribute2
E(D1) = Entropy value of subset1 of attribute1
E(D2) = Entropy value of subset1 of attribute2

After the entropy value for each feature is obtained, the final step is to find the information gain value for each feature using the formula in equation (2)

IG = 0,9544 – (22/400(0) +378/400(0,9692)
IG = 0,9544 -(0+(0,945*0,9692 = 0,9159)
IG = 0,0385

Then do the same thing to all the attributes that exist. So the calculation results are obtained asshown in the following chapter.

Modified K-Nearest Neighbor (MKNN)

MKNN (Modified K-Nearest Neighbor) is an improvement over the KNN method that involves additionalsteps, including the evaluation of validity values and the setting of weight thresholds [5].

Here is an explanation of the classification process using the MKNN method, namely:

a. Determining the value of K, K must be odd so that ambiguity does not occur in the classification process.
b. Calculating the distance between training data using the Euclidean Distance formula. Then the calculation results are sorted from smallest to largest by selecting the nearest neighbors accordingto the predetermined value of k. Euclidean Distance formula:

$$d(xi, yi) = \sqrt{\sum}_{i=0}^{n} (xi - yi)^2 \qquad (3)$$

Which is:
d= Distance between points in
training data,x= rows in training data,
y= columns in training data

c. Calculating the validity value between training data using the formula:

$$Validitas_{(i)} = \frac{1}{k} \sum_{i=1}^{k} S\ (lbl_x, lbl\ Ni_x) \qquad (4)$$

Description:

K	: Number of nearest neighbors
Lbl (x)	: Class X
Lbl Ni (x)	: Class label of the nearest point x

d. Finding the distance between test data and training data using the Euclidean Distance formula in Equation 3.
e. Calculate the value of Weight Voting, with the formula in the following equation:

$$W_{(i)} = Validitas_{(x)}\ x\ \frac{1}{d_e + a} \qquad (5)$$

Description:

W(i)	: weight voting calculation
Validitas (x)	: Validity of training data
de	: Euclidean Distance
alfa(α)	: smoothing regulator value, the value used is 0.5

f. After all the weight voting (wv) values are obtained, then sort the values from largest to smallest. Take the wv values according to the K value.
g. The values of each weight voting class will be accumulated. The class with the largest accumulated total will be the classification result.

RESULT AND DISCUSSION

The results of classifying chronic kidney disease using information gain feature selection with the MKNNand KNN methods will be explained as follows.

Data Cleaning

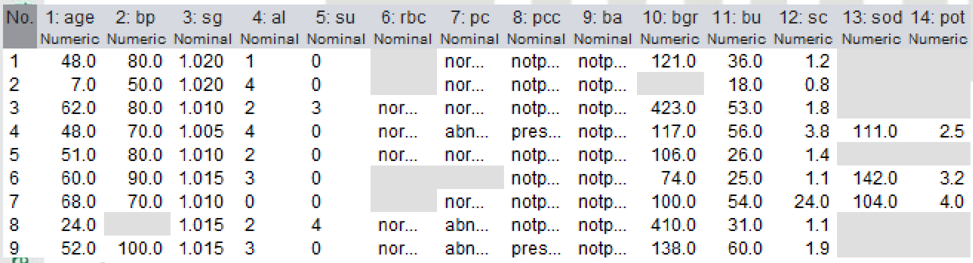
The data source for this study is a dataset obtained from the UCI Machine Learning

Repository. The data used consists of 400 data with 24 attributes and 2 classes, namely the class of patients with chronic kidneydisease/Chronic Kidney Disease (CKD) and the class of patients without chronic kidney disease/not Chronic Kidney Disease (Non-CKD). The data parameters used are presented in Table 2.

Table 2. Data Attributes

No	Abbreviation	Data Input
1	Age	*Age*
2	Bp	*Blood Pressure*
3	Sg	*Specific Gravity*
4	Al	*Albumin*
5	Su	*Sugar*
6	Rbc	*Red Blood Cells*
7	Pc	*Pus Cell*
8	Pcc	*Pus Cell Clumps*
9	Ba	*Bacteria*
10	Bgr	*Blood Glucose Random*
11	Bu	*Blood Urea*
12	Sc	*Serum Creatinine*
13	Sod	*Sodium*
14	Pot	*Potassium*
15	Hemo	*Hemoglobin*
16	Pcv	*Packed Cell Volume*
17	Wbcc	*White Blood Cell Count*
18	Rbcc	*Red Blood Cell Count*
19	Htn	*Hypertension*
20	Dm	*Diabetes Mellitus*
21	Cad	*Coronary Artery Disease*
22	Appet	*Appetite*
23	Pe	*Pedal Edema*
24	Ane	*Anemia*

The data for patients with chronic kidney disease has many missing values, so the data needs to be filled infirst so that the data to be used becomes of high quality and can obtain the best accuracy results. To fill in the missing values in this data, use the Weka application. The following is the initial data for patients with chronic kidney disease before data cleaning is performed. On the data, it can be seen that there are still many missing values in some attributes.



No.	1: age Numeric	2: bp Numeric	3: sg Nominal	4: al Nominal	5: su Nominal	6: rbc Nominal	7: pc Nominal	8: pcc Nominal	9: ba Nominal	10: bgr Numeric	11: bu Numeric	12: sc Numeric	13: sod Numeric	14: pot Numeric
1	48.0	80.0	1.020	1	0		nor...	notp...	notp...	121.0	36.0	1.2		
2	7.0	50.0	1.020	4	0		nor...	notp...	notp...		18.0	0.8		
3	62.0	80.0	1.010	2	3	nor...	nor...	notp...	notp...	423.0	53.0	1.8		
4	48.0	70.0	1.005	4	0	nor...	abn...	pres...	notp...	117.0	56.0	3.8	111.0	2.5
5	51.0	80.0	1.010	2	0	nor...	nor...	notp...	notp...	106.0	26.0	1.4		
6	60.0	90.0	1.015	3	0			notp...	notp...	74.0	25.0	1.1	142.0	3.2
7	68.0	70.0	1.010	0	0		nor...	notp...	notp...	100.0	54.0	24.0	104.0	4.0
8	24.0		1.015	2	4	nor...	abn...	notp...	notp...	410.0	31.0	1.1		
9	52.0	100.0	1.015	3	0	nor...	abn...	pres...	notp...	138.0	60.0	1.9		

Figure 2. Initial data of Chronic Kidney Dissease Patient

Data Transformation

The initial data for patients with chronic kidney disease has some features that are still in text form, so it needs to be transformed first so that the data can be used. Below is the data after the transformation is done.

Table 3. Data Transformation

No	*age*	*bp*	*sg*	*al*	…	…	*appet*	*pe*	*ane*	*class*
1	48	80	1.02	1	…	…	1	0	0	1
2	7	50	1.02	4	…	…	1	0	0	1
3	62	80	1.01	2	…	…	0	0	1	1
4	48	70	1.005	4	…	…	0	1	1	1
5	51	80	1.01	2	…	…	1	0	0	1

…	…				…	…	…	…	…	…
397	55	80	1.02	0	…	…	1	0	0	0
398	42	70	1.025	0	…	…	1	0	0	0
399	12	80	1.02	0	...	...	1	0	0	0
400	17	60	1.025	0	…	…	1	0	0	0
401	58	80	1.025	0	…	…	1	0	0	0

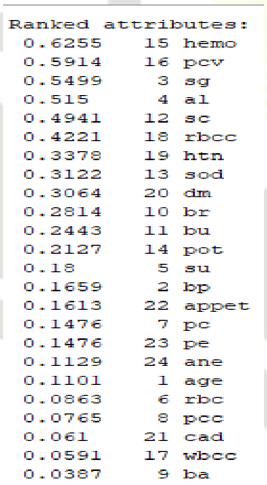
After all the data is transformed into numeric form, because the difference in values in the data is too large, normalization is performed. The normalization used is min-max normalization. The results of the normalization are as follows.

Table 4. Data after normalization

No	*age*	*bp*	*sg*	*al*	…	…	*appet*	*pe*	*ane*	*class*
1	0,523	0,231	0,75	0,2	…	…	1	0	0	1
2	0,057	0	0,75	0,8	…	…	1	0	0	1
3	0,682	0,231	0,25	0,4	…	…	0	0	1	1
4	0,523	0,154	0	0,8	…	…	0	1	1	1
5	0,557	0,231	0,25	0,4	…	…	1	0	0	1
…	…				…	…	…	…	…	…
397	0,602	0,231	0,75	0	…	…	1	0	0	0
398	0,455	0,154	1	0	…	…	1	0	0	0
399	0,114	0,231	0,75	0	...	...	1	0	0	0
400	0,170	0,077	1	0	…	…	1	0	0	0
401	0,636	0,231	1	0	…	…	1	0	0	0

Information Gain

In this study, the feature ranking process was carried out using the Weka application with the ranking resultsof each attribute as follows:



```
Ranked attributes:
 0.6255   15 hemo
 0.5914   16 pcv
 0.5499    3 sg
 0.515     4 al
 0.4941   12 sc
 0.4221   18 rbcc
 0.3378   19 htn
 0.3122   13 sod
 0.3064   20 dm
 0.2814   10 br
 0.2443   11 bu
 0.2127   14 pot
 0.18      5 su
 0.1659    2 bp
 0.1613   22 appet
 0.1476    7 pc
 0.1476   23 pe
 0.1129   24 ane
 0.1101    1 age
 0.0863    6 rbc
 0.0765    8 pcc
 0.061    21 cad
 0.0591   17 wbcc
 0.0387    9 ba
```

Figure 3. Information Gain Value

From the test results above, it was found that the hemo (Hemoglobin) attribute has the largest value with avalue of 0.6255, which means that the hemo attribute has the greatest influence on the dataset. Similarly, the ba (bacteria) attribute has the smallest value with an attribute value of 0.0387

Results of Information Gain Feature Selection Classification with KNN and MKNN

The classification results are presented in the form of a table comparing the accuracy of information gain feature selection between the KNN and MKNN methods with the ratio of training and test data used, namely90:10, 80:20, and 70:30. Here are the results:

a. Information gain and MKNN method testing with a 90:10 ratio

Table 5. IG+MKNN 90:10

No	Total Attributes	K-Value	Accuracy
1	24	3	97,5%
		5	97,5%
		7	97,5%
2	20	3	97,5%
		5	100%
		7	100%
3	15	3	97,5%
		5	97,5%
		7	97,5%
4	10	3	97,5%
		5	97,5%
		7	97,5%
5	5	3	100%
		5	97,5%
		7	97,5%
	Accuracy Means		98%

The next, is information gain and KNN method testing with a 90:10 ratio.
The test starts from using the existing 24 features, then continues with the test using 20 features, 15 features, 10 features, and finally using 5 features. The variation of the k value used is 3, 5, and 7. The average accuracy obtained in this test is 96.17%.

Table 6. IG+KNN 90:10

No	Total Attributes	K-Value	Accuracy
1	24	3	97,5%
		5	97,5%
		7	92,5%
2	20	3	97,5%
		5	97,5%
		7	97,5%
3	15	3	97,5%
		5	95%
		7	95%
4	10	3	95%
		5	95%
		7	95%
5	5	3	100%
		5	95%
		7	95%
	Accuracy Means		96,17%

b. Information gain and MKNN method testing with a 80:20 ratio
In this test, the test will be carried out using a data training and test ratio of 80:20. The test starts from using the existing 24 features, then continues with the test using 20 features, 15 features, 10 features, and finally using 5 features. The variation of the k value used is 3, 5, and 7. The average accuracy obtained in this test is 97.75%.

Table 7. IG+MKNN 80:20

No	Total Attributes	K-Value	Accuracy
1	24	3	97,5%
		5	98,33%
		7	98,33%
2	20	3	97,5%
		5	98,33%
		7	98,33%
3	15	3	97,5%
		5	97,5%
		7	97,5%
4	10	3	96,66%
		5	96,66%
		7	97,5%
5	5	3	96,66%
		5	96,66%
		7	96,66%
	Accuracy Means		97,44%

The next is, information gain and KNN method testing with a 80:20 ratio.
The test starts from using the existing 24 features, then continues with the test using 20 features, 15 features, 10 features, and finally using 5 features. The variation of the k value used is 3, 5, and 7. The average accuracy obtained in this test is 96.5%

Table 8. IG+KNN 80:20

No	Total Attributes	K-Value	Accuracy
1	24	3	97,5%
		5	96,25%
		7	97,5%
2	20	3	97,5%
		5	97,5%
		7	98,75%
3	15	3	97,5%
		5	96,25%
		7	96,25%
4	10	3	96,25%
		5	96,25%
		7	96,25%
5	5	3	96,25%
		5	93,75%
		7	93,75%
	Accuracy Means		96,5%

c. Information gain and MKNN method testing with a 70:30 ratio
The test starts from using the existing 24 features, then continues with the test using 20 features, 15 features, 10 features, and finally using 5 features. The variation of the k value used is 3, 5, and 7. The average accuracy obtained in this test is 97.44%.

Table 9. IG+MKNN 70:30

No	Total Attributes	K-Value	Accuracy
1	24	3	97,5%
		5	98,33%
		7	98,33%
2	20	3	97,5%
		5	98,33%
		7	98,33%
3	15	3	97,5%
		5	97,5%
		7	97,5%
4	10	3	96,66%
		5	96,66%
		7	97,5%
5	5	3	96,66%
		5	96,66%
		7	96,66%
	Accuracy Means		97,44%

The next is, information gain and KNN method testing with a 70:30 ratio.
The test starts from using the existing 24 features, then continues with the test using 20 features, 15 features, 10 features, and finally using 5 features. The variation of the k value used is 3, 5, and 7. The average accuracy obtained in this test is 96.61%.

Table 10. IG+KNN 70:30

No	Total Attributes	K-Value	Accuracy
1	24	3	
		5	98,33%
		7	98,33%
2	20	3	97,5%
		5	98,33%
		7	98,33%
3	15	3	96,67%
		5	96,67%
		7	96,67%
4	10	3	95,83%
		5	95,83%
		7	95%
5	5	3	95,83%
		5	94,17%
		7	94,17%
	Accuracy Means		96,61%

CONCLUSION

Based on several tests that have been conducted, it can be concluded that information gain feature selection has an impact on improving the accuracy results, both using the MKNN method and the KNN method. In the test with a data comparison of 90:10, the best accuracy in the test using information gain and the MKNNmethod is at the number of features 20 with the value k=5, with an accuracy of 100%, and at the number offeatures 5 with the value k=3, the accuracy is 100%. Then for those using the KNN method, the best accuracy is at the number of features 5 with the value k=3, with an accuracy of 100%. However, it should be noted that if you want to use information gain feature selection, you first need to set a threshold or limit, so only features with the smallest gain value are deleted. This is because based on the results of the tests that have been conducted, if too many features with high gain values are deleted, it can affect or even reducethe accuracy results.

REFERENCES

[1] Harmayani and L. Sitorus, “Diagnosa Penyakit Ginjal Kronis Menggunakan Metode Klasifikasi Naïve Bayes,” *J. MEDIA Inform. BUDIDARMA*, vol. 4, no. 3, pp. 850–854, 2020, doi: 10.30865/mib.v4i3.2292.

[2] G. A. M. Pratama *et al.*, “Diagnosis Penyakit Ginjal Kronis dengan Algoritma C4.5, K-Means dan BPSO,” *J. Elektron. Ilmu Komput. Udayana*, vol. 10, no. 4, pp. 371–381, 2022.

[3] P. Studi and M. Informatika, “Perbandingan Metode Data Mining Svm dan Nn Untuk Klasifikasi Penyakit Ginjal Kronis,” *J. PILAR Nusa Mandiri*, vol. 14, no. 1, pp. 1–6, 2018.

[4] M. R. Hasibuan and Marji, “Pemilihan Fitur dengan Information Gain untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN),” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 11, pp. 10435–10443, 2019.

[5] M. I. P. Putra, D. T. Murdiansyah, and A. Aditsania, “Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Klasifikasi Penyakit Kanker Payudara,” *E-proceeding Eng.*, vol. 6, no. 1, pp. 2431–2441, 2019.

[6] M. R. Maulana and M. A. Al Karomi, “Information Gain untuk Mengetahui Pengaruh Atribut terhadap Klasifikasi Persetujuan Kredit,” *J. Litbang Kota Pekalongan*, vol. 9, 2015.

[7] F. Y. Nabella, Y. A. Sari, and R. C. Wihandika, “Seleksi Fitur Information Gain Pada Klasifikasi Citra Makanan Menggunakan Hue Saturation Value dan Gray Level Co-Occurrence Matrix,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 2, 2019.

[8] I. S. Bakti and Ivandari, “Model Prediksi Penyakit Diabetes Menggunakan Bayesian Classification dan Information Gain untuk Seleksi Fitur dan Adaptive Boosting untuk Pembobotan Data,” *IC-Tech*, vol. XI, no. 1, pp. 28–37, 2019.

[9] S. I. Fernanda, D. E. Ratnawati, and P. P. Adikara, “Identifikasi Penyakit Diabetes Mellitus Menggunakan Metode Modified K- Nearest Neighbor (MKNN),” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 6, pp. 507–513, 2017.

[10] S. Z. Hr, A. Aziz, and W. Harianto, “Optimasi algoritma k-nearest neighbor (knn) dengan normalisasi dan seleksi fitur untuk klasifikasi penyakit liver,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 439–445, 2022.