Jurnal Teknik Informatika Vol. 17 No. 2, Oktober 2024 (190-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i2.40866

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

Using K-NN Algorithm for Evaluating Feature Selection on High Dimensional Datasets

Fina Indri Silfana[1], Mula Agung Barata[2]

[1,2]Department of Informatics Engineering, Faculty of Science and Technology, University of Nahdlatul Ulama Sunan Giri
[1,2] Jl.Ahmad Yani No. 10, Bojonegoro 62115, Indonesia

Article:
Accepted: September 12, 2024
Revised: July 27, 2024
Issued: October 29, 2024

© Silfana & Barata (2024).

This is an open-access article under the CC BY-SA license

***Correspondence Address:**
finaindri1234@gmail.com

ABSTRACT

Data mining is the process of using statistics, mathematics, artificial intelligence and machine learning to identify problems that exist in data so as to produce useful information. Based on its function, data mining is grouped into description, estimation, classification, clustering, and association. K-NN is one of the best data mining methods and is widely used in research. K-NN algorithm was introduced by Fix and Hodges in 1951. K-NN algorithm is a simple algorithm and is often used to cluster supervised data. Feature selection attribute selection is a data mining technique used in the pre-processing stage. This technique works by reducing complex attributes that will be managed at the processing and analysis stage. In this study, the most effective feature selection to improve the accuracy of the K-NN algorithm by increasing accuracy by 95.12% on the breast cancer dataset and 88.75% on the prostate cancer dataset.

Keywords : *data mining; classification; feature selection;*

Jurnal Teknik Informatika Vol. 17 No. 2, Oktober 2024 (190-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i2.40866

1. INTRODUCTION

In recent decades, large amounts of data have been stored by companies and organizations. The data comes from several formats, ranging from text, images, voice, email, sensor readings, and so on. In terms of usage, the big data will be useless if it is not useless if not processed into usable information. The method that used to process data into information is data mining [1]. Data mining is the process of using statistics, mathematics, artificial intelligence and machine learning to identify problems that exist in data so as to produce useful information. Based on its function, data mining is grouped into, description, estimation, classification, clustering, and association [2].

Classification algorithm is a process of finding a collection of patterns and functions to describe and separate data classes from one another, which is useful for grouping the data in predetermined categories. Classification is a form of data analysis that extracts data class models. Classification is included in supervised learning because it uses data that has been analyzed to be tested and classified. The classification process itself consists of learning and classification. In the learning stage there is training data and test data to ensure the rule of accuracy. Classification techniques are divided into five namely statistical-based, distance-based, decision tree-based, neural network-based and rule-based. The problem in data mining, especially classification, is high-dimensional data [3]. High-dimensional data causes the dataset size to be larger, large number of attributes and large number of sample data. Data with many attributes causes the performance of classification algorithms to be low [4].

To solve problems related to high-dimensional data is to use feature selection [5]. By using feature selection, the risk of overfitting can be avoided, accuracy becomes better because it eliminates redundant data and features that are not significant or relevant, so that it can save the time needed in the classification process because of the reduced data dimension[6]. Feature selection is also used in some studies to handle problems in high-dimensional data.

Joko Suntoro et al [7], in this study comparing the GA-SVM and SVM algorithms and combined with forward selection to handle high-dimensional data with the results showing that the GA-SVM method provides the best evaluation results compared to the SVM method with an average accuracy value in the GA-SVM WFS method of 0.902, while the SVM method is 0.874 when used on high-dimensional datasets, namely datasets with many features.

Widya Astuti et al [8], on improving the accuracy of the Naive Bayes algorithm by combining it with forward selection using breast cancer datasets. The results of the study showed an increase in accuracy value after using the forward selection method used to select features/reduce dimensions. The difference in accuracy increased by 2.92%, precision increased by 3.66% and recall by 0.93% after using the forward selection method.

Tri Ernayanti et al [9], using the Multinominal Naive Bayes algorithm and combined with Chi-Square with Tokopedia customer review datasets. With the results of feature selection using Chi-Square has an effect on reducing the number of features obtained. At a significance level of 0.05, the number of features obtained is 160 from the initial number of features of 769. The classification performance results using Multinomial Naïve Bayes without Chi-Square feature selection obtained accuracy and kappa statistics of 88% and 75.95% while using Chi-Square feature selection obtained accuracy and kappa statistics of 95% and 89.99%.

Therefore, this research will use 3 feature selections, namely forward selection, backward elimination, and chi-square method. In order to overcome the problems that arise in high-dimensional data.

K-NN is one of the best data mining methods and is widely used in research.The K-NN algorithm was introduced by Fix and Hodges in 1951.The K-NN algorithm is a simple algorithm and is often used to classify supervised data. K-NN classifies objects based on training data based on the closest distance to the object. The distance between the object and the training data can be calculated using eulidiance [10]. However, K-NN has a drawback, namely in calculating the K-NN algorithm, it must calculate the distance on each query instance together so that the results in the calculation of the K-NN algorithm using a high-dimensional dataset without the addition of other methods are low in accuracy. In some studies using the KNN

Jurnal Teknik Informatika Vol. 17 No. 2, Oktober 2024 (190-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i2.40866

algorithm by adding feature selection with increased accuracy results.

Rangga Sanjaya and Fitriyani [11]. This research combines the K-NN algorithm with Forward Selection to optimize the accuracy of the K-NN algorithm. The results of the research conducted that the model using K-NN without feature selection produces the best accuracy value of 83.40%. While the model that uses K-NN and Forward Selection (K-NN+FS) produces the best accuracy value of 85.74%.

Maxsi Ary and Dyah Rismiati[12] also conducted research with the K-NN algorithm and added Backward Elimination to increase the accuracy of the K-NN algorithm. The result of this research is the level of accuracy produced by the K-Nearest Neighbor algorithm in the classification of mesothelioma disease of 93.85%. Meanwhile, the classification results of the K-Nearest Neighbor algorithm after feature selection using Backward Elimination showed an increase in accuracy of 4.61%, so that the resulting accuracy rate was 98.46%.

Hamsir Saleh[13] also implemented the K-NN algorithm and added Chi-square feature selection. With the accuracy results with K-NN, the accuracy value is 85.78%. The Chi Square method as an attribute selection can help improve the accuracy of K-NN classification results. K-Nearest Neighbor based on Chi Square attribute selection is more accurate and effective in classifying scholarship recipients from the data used, the accuracy result is 88.53%.

In addition, the use of feature selection in other algorithms was carried out by Mula Agung Barata, et al [14] who used the C4.5 algorithm with the addition of feature selection chi-squre. In this study, feature selection was also able to increase accuracy in the C4.5 algorithm, from 93.51% to 94.27%.

From previous research, K-NN is one of the algorithms that has low accuracy and feature selection is a method that can improve the accuracy of the algorithm. So to improvise the K-NN algorithm which has low accuracy, especially in high-dimensional datasets, feature selection is used to improve its accuracy.

2. METHODS

The method used by researchers in this study is used to classify high-dimensional datasets. To find the right algorithm that is suitable for high-dimensional datasets.

2.1. KNN-Algorithm

KNN is a supervised learning algorithm, where the results of new query instances are classified based on the majority of categories in kNN. The class that appears the most will be the class of classification results [15]. KNN is one of the methods of classifying data based on similarity with data labels.K-nearest neighbors or K-NN is one of the lazy learning algorithms. K-NN is used to find the k closest group objects or those that are similar to the testing data in the new data inputted with the euclidiance distance [16]-[17]. The calculation of the euclidiance distance is with the following equation:

$$\mathbf{Euc}=\sqrt{((\boldsymbol{a\_1}-\boldsymbol{b\_1})\text{\^{}}\mathbf{2}+\cdots+(\boldsymbol{a\_n}-\boldsymbol{b\_n})\text{\^{}}\mathbf{2}} \quad (1)$$

For a=a1, a2, ..., an and b=b1, b2, ..., bn are permutations for n attribute values from two records for attributes with category values.

K-NN has a fast data processing time and is relatively good for low-dimensional datasets but tends to be weak against high-dimensional datasets. However, in this case, feature selection can be used to improve the accuracy of the K-NN algorithm by selecting features according to the feature selection criteria.

The following is the processing flowchart of the KNN algorithm :

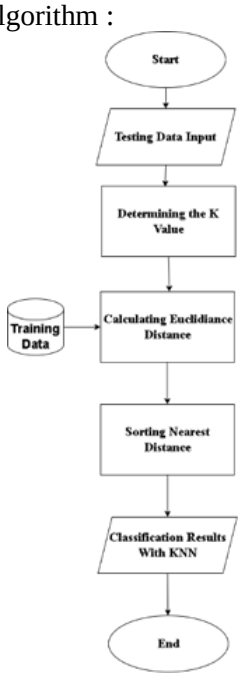

Figure 1. *KNN algorithm flowchart*

Jurnal Teknik Informatika Vol. 17 No. 2, Oktober 2024 (190-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i2.40866

With the following explanation:

1. Input the dataset that will be classified with KNN.
2. Determining the value of parameter K
3. Calculate the euclidiance or distance between training data and testing data with the following formula:

$$Euc=\sqrt{(\sum_{i=1}^{n}(pi-qi)^2} \quad (2)$$

Explanation :

pi	=	training data
qi	=	testing data
n	=	data dimension
i	=	variable data

4. sorting the distances formed
5. sorting the closest distance

2.2. Feature Selection

Feature selection attribute selection is a data mining technique used in the pre-processing stage. This technique works by reducing complex attributes that will be managed at the processing and analysis stage[18]. The attribute selection process can be seen in the following figure :

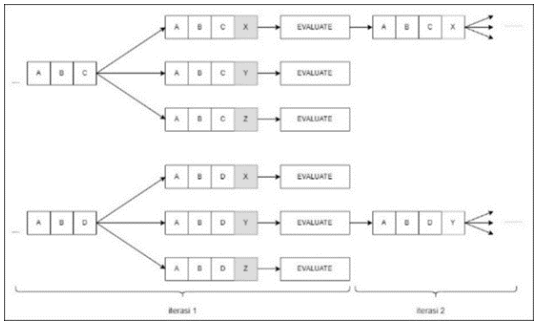

Figure 2. *Feature selection process*

In the feature selection approach there are two kinds of approaches [19], namely:

a. Filter method

Feature selection using this approach tends to be computationally cheaper as it does not involve algorithm induction in the process. There are four categories of filters, namely, Mutual Information, Fisher Score, Analysis of Variance (ANOVA), and Pearson Correlation.

b. Wrapper method

The wrapper approach uses an evaluation function based on the algorithm that will be used in algorithm processing. The first component is done after the formation of the feature space, then the search procedure is carried out which produces a subset of features to be evaluated. Then the second component performs an evaluation function as a measure of subset selection. Feature selection using the wrapper method consists of Forward Selection, Backward Elimination and Recrisive Feature Elimination (RFE).

c. Embedded Method

Embedded is a combination of Filter and Wrapper.Feature selection with this method uses a learning algorithm as a Wrapper approach and against features selected with a Filter approach. Feature selection with Embedded Category are lEast Shrinkage and Selection Operator (LASSO) and Elastic Net.

2.2.1. Forward Selection Method

Forward selection is a feature used to select features that are not needed during the iteration process [20].

Feature selection using the Forward selection method is done before processing data in the K-NN algorithm. That is by eliminating attributes according to the following conditions which then pass the selection will be processed using the K-NN algorithm.

Forward selection is one of the modeling to find the best variable [21]. Forward selection starts with the empty feature set and then adds the features used in the first iteration. All features are selected respectively, to reduce the number of evaluations, only the best subset of features is saved. Training data is carried out in stages, starting from 1 variable to the variable with the best accuracy and small error. For example, testing with 2 data will produce a smaller error compared to 3 variables which will certainly produce a larger error. The process will be terminated when all independent variables have been tested. The forward selection algorithm will be tested on each data to produce the best accuracy as shown in Figure 2.

2.2.2. Backward Elimination Method

Backward elimination is feature selection that eliminates the least important attributes and leaves the variables that are important in a model[22]. Backward elinimanation feature selection is performed before processing the data using the K-NN algorithm. The dataset will be processed using feature selection then the selected attributes will be processed using the K-NN algorithm.

Jurnal Teknik Informatika Vol. 17 No. 2, Oktober 2024 (190-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i2.40866

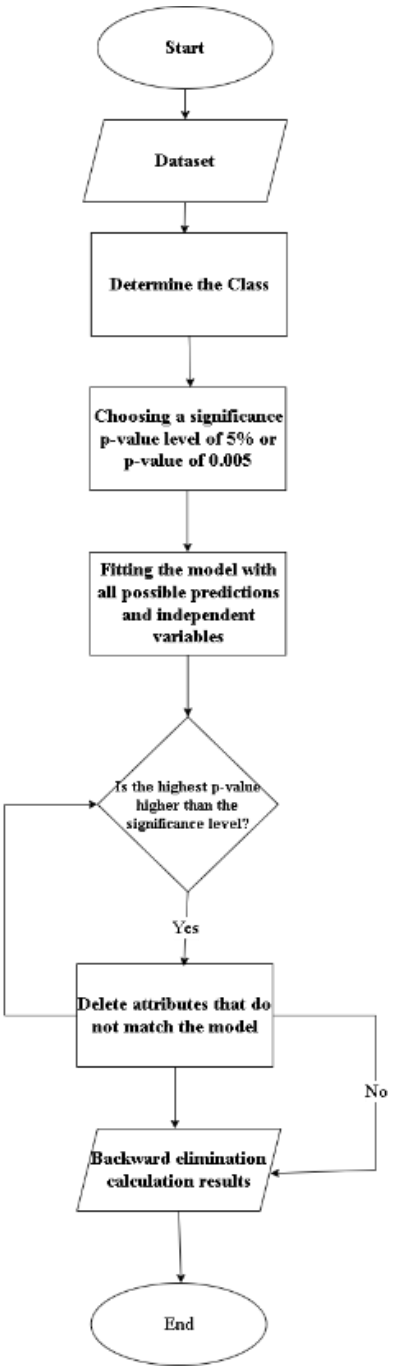

Figure 3. *Backward elimination flowchart*

Some of the steps in the backward elimination calculation include:

1. In backward elimnation all attributes will be tested in a regression model with a significance level of 0.05.
2. When the p-value of a feature is greater than its significance level (P-value > 0.05), it is omitted.
3. This step is repeated until all features become significant (P-value <0.05). Finally, the model is fitted with a new set of features.

2.2.3. Chi-square Method

One of the feature selection that has good performance is Chi-square. This method has working steps like the flowchart below from research by Mula agung barata [14]:

Testing algorithms using Chi-square there are several conditions that must be considered and must be met among others:

1. Actual Count or real frequency must have a value of 0 (zero).
2. If there is 1 cell that has an expected count or expected frequency (Fh) less than 5 or (<5) on the 2x2 contingency table then it is allowed.
3. If the number of cells with Fh<5 cannot be 20% in the 2x2 contingency table

The chi-square feature selection stage is carried out before processing the data using the K-NN algorithm. The dataset will be processed using feature selection then the selected attributes will be processed using the K-NN algorithm.

2.3. Model Evaluation

In general, to evaluate the work of the algorithm, the confusion matrix can be used. Evaluation with the confusion matrix is used to estimate the correct or incorrect object of the predicted class and compared with the actual class.The following is a confusion matrix table:

Table 1. *Confussion matrix*

conffusion matrix		**Prediction**	
		positif	**negatif**
actual	positif	TP	FP
	negatif	FN	TN

Table description:

1. TP (True Positive) is the amount of data for which both the actual and predicted classes are positive.
2. TN (True Negative) is the amount of data whose actual and predicted classes are negative.
3. FP (False Positive) is the amount of data where the actual class is negative and the prediction is positive.
4. FN (False Negative) is the amount of data where the actual class is positive and the prediction is negative.

Jurnal Teknik Informatika Vol. 17 No. 2, Oktober 2024 (190-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i2.40866

3. RESULTS AND DISCUSSION

3.1 Dataset

The object of this research is the Breast Cancer dataset and the apple qulality dataset obtained from the Kaggle.com website. This research object was taken because in research by Erlina Marfianti[20] stated that breast cancer or breast cancer is the main cancer causing death in women, while the prostate cancer dataset was chosen because in research by Adi Muzakir[23] stated that prostate cancer or prostate cancer is ranked fifth out of 5 cancers in the world in the number of sufferers.

The breast cancer dataset has 31 attributes with 569 records while the prostate cancer dataset has 8 attributes and 100 records.

Table 2. *Breast cancer dataset*

no	diagnosis	radius_mean	…	dimension _worst
1	M	17.99	…	0.1189
2	M	20.57	…	0.08902
…	……..	……..	…	……..
568	M	20.6	…	0.124
569	B	7.76	…	0.07039

Table 3. *Prostate cancer dataset*

id	diagnosis	radius	texture	…	dimension
1	M	23	12	…	0.079
2	B	9	13	…	0.057
3	M	21	27	…	0.06
4	M	14	16	…	0.097
5	M	9	19	…	0.059

3.2 Implementation of K-NN Algorithm

3.2.1. Calculating Euclidiance Distance

In this study using two different datasets, namely the breast cancer dataset and the prostate cancer dataset using the K-NN Algorithm so that to calculate the K value is as follows is the calculation of the K-NN Algorithm on the breast cancer dataset.

$$d = \sqrt{\begin{array}{c}(17.99-20.6)^2-(10.9-29.33)^2-(122.8-140.1)^2-(10.9-29.33)^2 \\ -(122.8-140.1)^2-(10.95-0.726)^2-(0.9053-1.595)^2 \\ -(8.589-5.772)^2-(153.4-86.22)^2-(0.006399-0.006522)^2 \\ -(0.04904-0.06158)^2-(0.01587-0.01664)^2 \\ -(0.03003-0.02324)^2-(0.006193-0.006185)^2 \\ -(0.9053-1.595)^2-(25.38-25.74)^2-(17.33-39.42)^2 \\ -(184.6-184.6)^2-(2119-182`)^2 \end{array}}$$

$$d=337.9305108 \quad (3)$$

The following is a table of Euclidiance distance calculation results and also the rank of each record in the breast cancer dataset:

Table 4. *K-NN calculation on breast cancer dataset*

Diagnosis	Radius_Mean	…	*euclidiance*	rank
M	17.99	…	337.9305108	43
M	20.57	…	168.5351771	12
…	…	…	…	…
B	7.76	…	811.2906793	1

Table 5. *K-NN calculation on prostate cancer dataset*

radius	…	diagnosis	euclidiance distance	Rank
23	…	M	316.6133795	67
9	…	B	684.291607	96
21	…	M	561.1782277	91
14	…	M	257.740249	61
9	…	M	655.3701262	93

3.2.2. Confusion Matrix

And here is the performance of the K-NN algorithm with breast cancer datasets measured using accuracy, precission, recall and f-1 score:

Table 6. *K-NN confusion matrix on breast cancer dataset*

conffusion matrix		Prediction	
		positif	negatif
actual	positif	154	12
	negatif	16	263

Based on the table above, the performance of the K-NN algorithm with breast cancer datasets is measured using the following accuracy, precission, recall and f-1 score :

$$Recall = \frac{TP}{TP+FN} = \frac{154}{154+16} = 0.905 \quad (4)$$

$$Precission = \frac{TP}{TP+FP} = \frac{154}{154+12} = 0.927 \quad (5)$$

$$Accuracy = \frac{TP+TN}{(TP+FP)+(FN+TN)} = \frac{154+263}{(154+12)+(16+263)} = 0.937$$

(6)

$$F1\text{-}Score = 2.\frac{Recall.Precision}{Recall+Preacission} = \frac{(0.905).(0.927)}{0.905+0.927} = 0.457$$

(7)

And here is the performance of the K-NN algorithm with the prostate cancer dataset measured using accuracy, precission, recall and f-1 score:

Table 7. *K-NN confusion matrix on breast cancer dataset*

conffusion matrix		Prediction	
		positif	**negatif**
actual	positif	44	5
	negatif	6	25

$$Recall = \frac{TP}{TP+FN} = \frac{44}{44+6} = 0.88 \quad (8)$$

$$Precission = \frac{TP}{TP+FP} = \frac{44}{44+5} = 0.897959 \quad (9)$$

$$Accuracy = \frac{TP+TN}{(TP+FP)+(FN+TN)} = \frac{44+25}{(44+5)+(6+25)} = 0.8625$$

(10)

$$F1\text{-}Score = 2.\frac{Recall.Precision}{Recall+Preacission} = \frac{(0.88).(0.897959)}{0.88+0.897959} = 0.888889$$

(11)

3.3 Implementation of backward elimination and K-NN

3.3.1 Attribute selection using backward elimination

Attribute selection in the breast cancer datasets using the backward elimination method produces 29 remaining attributes by eliminating 1 attribute.

Attributes that pass the selection include:

1. radius_mean
2. perimeter_mean
3. area_mean
4. smoothness_mean
5. compactness_mean
6. concavity_mean
7. concave_points_mean
8. symmetry_mean
9. fractal_dimension_mean
10. radius_se
11. texture_se
12. perimeter_se
13. area_se
14. smoothness_se
15. compactness_se
16. concavity_se
17. concave_points_se
18. symmetry_se
19. fractal_dimension_se
20. radius_worst
21. texture_worst
22. perimeter_worst
23. area_worst
24. smoothness_worst
25. compactness_worst
26. concavity_worst
27. concave_points_worst
28. symmetry_worst
29. fractal_dimension_worst

And the selected attribute is :

1. texture_mean

Whereas in the prostate cancer dataset, attribute selection using the backward elimination method resulted in 7 attributes that passed the attribute selection from 8 attributes in the prostate cancer dataset.

Attributes that pass the selection include:

1. radius
2. perimeter
3. area
4. smoothness
5. compactness
6. symmetry
7. fractal_dimension

While the attributes that do not pass the selection are :

1. texture

3.3.2 Calculation using the K-NN algorithm

After selecting features, the next step is to calculate the dataset that has been selected attributes using the K-NN algorithm. The following are the results of the calculation of euclidiance distance and ranking on each reacord using breast cancer and prostate cancer datasets :

Table 8. *Backward elimination and K-NN calculation on breast cancer dataset*

Diagnosis	**Radius_Mean**	**…**	***euclidiance***	**rank**
M	17.99	…	44477.86146	34
M	20.57	…	19362.93501	19
…	…	...	…	…
B	7.76	…	1425906.921	106

Jurnal Teknik Informatika Vol. 17 No. 2, Oktober 2024 (190-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i2.40866

Table 9. *Backward elimination and K-NN calculation on prostate cancer dataset*

diagnosis	radius	…	Euclidiance distance	Rank
M	23	...	454.7549	63
B	9	...	822.1204	78
…	…	...	…	…
B	22	...	15	1

3.3.3 Confusion Matrix

Based on the table above, the performance of the K-NN algorithm with breast cancer datasets is measured using the following accuracy, precission, recall and f-1 score:

Table 10. *Confusion matrix backward elimination and K-NN on breast cancer dataset*

conffusion matrix		Prediction	
		positif	negatif
actual	positif	155	12
	negatif	15	263

$$Recall = \frac{TP}{TP+FN} = \frac{155}{155+15} = 0.9117 \quad (12)$$

$$Precission = \frac{TP}{TP+FP} = \frac{155}{155+12} = 0.8979 \quad (13)$$

$$Accuracy = \frac{TP+TN}{(TP+FP)+(FN+TN)} = \frac{44+25}{(44+5)+(6+25)} = 0.8625 \quad (14)$$

$$F1\text{-}Score = 2.\frac{Recall.Precision}{Recall+Preacission} = \frac{(0.88).(0.8979)}{0.885+0.8979} = 0,9664 \quad (15)$$

And here is the performance of the K-NN algorithm with the prostate cancer dataset measured using accuracy, precission, recall and f-1 score:

Table 11. *Confusion matrix backward elimination and K-NN on prostate cancer dataset*

conffusion matrix		Prediction	
		positif	negatif
actual	positif	39	5
	negatif	11	25

$$Recall = \frac{TP}{TP+FN} = \frac{39}{39+11} = 0.78 \quad (16)$$

$$Precission = \frac{TP}{TP+FP} = \frac{39}{39+5} = 0.886364 \quad (17)$$

$$Accuracy = \frac{TP+TN}{(TP+FP)+(FN+TN)} = \frac{39+25}{(39+5)+(11+25)} = 0.8 \quad (18)$$

$$F1\text{-}Score = 2.\frac{Recall.Precision}{Recall+Preacission} = \frac{(0.78).(0.886364)}{0.78+0.886364} = 0.829787 \quad (19)$$

3.4 R Implementation of forward selection and K-NN algorithm

3.4.1. Attribute selection using forward selection

Attribute selection in the breast cancer datasets using the forward selection method produces 3 remaining attributes by eliminating 27 attribute.

Attributes that pass the selection include:

1. radius_mean
2. texture_mean
3. perimeter_worst

And the selected attribute is :

1. perimeter_mean
2. area_mean
3. smoothness_mean
4. compactness_mean
5. concavity_mean
6. concave_points_mean
7. symmetry_mean
8. fractal_dimension_mean
9. radius_se
10. texture_se
11. perimeter_se
12. area_se
13. smoothness_se
14. compactness_se
15. concavity_se
16. concave_points_se
17. symmetry_se
18. fractal_dimension_se
19. radius_worst
20. texture_worst
21. area_worst
22. smoothness_worst
23. compactness_worst
24. concavity_worst
25. concave_points_worst
26. symmetry_worst
27. fractal_dimension_worst

Jurnal Teknik Informatika Vol. 17 No. 2, Oktober 2024 (190-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i2.40866

Whereas in the prostate cancer dataset, attribute selection using the forward selection method resulted in 2 attributes that passed the attribute selection from 8 attributes in the prostate cancer dataset.
Attributes that pass the selection include:

1. perimeter
2. compactness

While the attributes that do not pass the selection are :

1. radius
2. texture
3. area
4. smothness
5. symmetry

3.4.2. Calculation using the K-NN algorithm
3.4.3. Confussion Matrix

The performance of the K-NN algorithm with breast cancer dataset is measured using the following accuracy, precission, recall and f-1 score:

Table 12. *Confusion matrix forward selection and K-NN on breast cancer dataset*

conffusion matrix		**Prediction**	
		positif	**negatif**
actual	positif	156	8
	negatif	14	267

$$Recall = \frac{TP}{TP+FN} = \frac{156}{156+14} = 0.917647 \quad (20)$$

$$Precission = \frac{TP}{TP+FP} = \frac{155}{155+8} = 0.95122 \quad (21)$$

$$Accuracy = \frac{TP+TN}{(TP+FP)+(FN+TN)} = \frac{156+267}{(156+8)+(14+267)} = 0.950562 \quad (22)$$

$$F1\text{-}Score = 2.\frac{Recall.Precision}{Recall+Preacission} = \frac{(0.917647).(0.95122)}{0.917647+0.95122} = 0.934132$$

(23)

And here is the performance of the K-NN algorithm with the prostate cancer dataset measured using accuracy, precission, recall and f-1 score

Table 13. *Confusion matrix forward selection and K-NN on prostate cancer dataset*

conffusion matrix		**Prediction**	
		positif	**negatif**
actual	positif	43	2
	negatif	7	28

$$Recall = \frac{TP}{TP+FN} = \frac{43}{43+7} = 0.86 \quad (23)$$

$$Precission = \frac{TP}{TP+FP} = \frac{43}{43+2} = 0.95556 \quad (24)$$

$$Accuracy = \frac{TP+TN}{(TP+FP)+(FN+TN)} \quad (25)$$

$$= \frac{43+28}{(43+2)+(7+28)} = 0.8875$$

$$F1\text{-}Score = 2.\frac{Recall.Precision}{Recall+Preacission} = \frac{(0.86).(0.95556)}{0.86+0.95556} = 0.905263$$

(26)

3.5 Implementation of chi-square and K-NN algorithm

3.5.1. Attribute selection using chi-square

The chi-square implementation on the breast cancer dataset and prostate cancer dataset cannot select attributes.

3.5.2. Calculation using the K-NN algorithm

After selecting features, the next step is to calculate the dataset that has been selected attributes using the K-NN algorithm. The following are the results of the calculation of euclidiance distance and ranking on each reacord using breast cancer and prostate cancer datasets:

The following is a table of Euclidiance distance calculation results and also the rank of each record in the breast cancer dataset:

Table 14. *Chi-square and K-NN calculation on breast cancer dataset*

diagnosis	**radius_Mean**	**…**	***euclidiance***	**rank**
M	17.99	…	337.9305108	43
M	20.57	…	168.5351771	12
……..	……..	…	……..	……
B	7.76	…	811.2906793	1

Jurnal Teknik Informatika Vol. 17 No. 2, Oktober 2024 (190-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i2.40866

Table 15. *Chi-square and K-NN calculation on prostate cancer dataset*

radius	…	diagnosis	euclidiance distance	rank
23	…	M	316.6133795	67
9	…	B	684.291607	96
21	…	M	561.1782277	91
……	…	……	……	……
22	…	B	105.0952	1

3.5.3. Confussion Matrix

Then the performance of the K-NN algorithm with breast cancer datasets is measured using the following accuracy, precission, recall and f-1 score:

Table 16. *Confusion matrix chi-square and K-NN on breast cancer dataset*

conffusion matrix		prediction	
		positif	negatif
actual	positif	154	12
	negatif	16	263

$$Recall = \frac{TP}{TP+FN} = \frac{154}{154+16} = 0.905 \quad (27)$$

$$Precission = \frac{TP}{TP+FP} = \frac{154}{154+12} = 0.927 \quad (28)$$

$$Accuracy = \frac{TP+TN}{(TP+FP)+(FN+TN)} \quad (29)$$

$$= \frac{154+263}{(154+12)+(16+263)} = 0.937$$

$$F1\text{-}Score = 2.\frac{Recall.Precision}{Recall+Preacission} \quad (30)$$

$$= \frac{(0.905).(0.927)}{0.905+0.927} = 0.457$$

And here is the performance of the K-NN algorithm with the prostate cancer dataset measured using accuracy, precission, recall and f-1 score:

Table 17. *Confusion matrix chi-square and K-NN on prostate cancer dataset*

conffusion matrix		prediction	
		positif	negatif
actual	positif	44	5
	negatif	6	25

$$Recall = \frac{TP}{TP+FN} = \frac{44}{44+6} = 0.88 \quad (31)$$

$$Precission = \frac{TP}{TP+FP} = \frac{44}{44+5} = 0.897959 \quad (32)$$

$$Accuracy = \frac{TP+TN}{(TP+FP)+(FN+TN)} \quad (33)$$

$$= \frac{44+25}{(44+5)+(6+25)} = 0.8625$$

$$F1\text{-}Score = 2.\frac{Recall.Precision}{Recall+Preacission} \quad (34)$$

$$= \frac{(0.88).(0.897959)}{0.88+0.897959} = 0.888889$$

3.6 Evaluation of confusion matrix in K-NN, Backward elimination, Forward selection and Chi-Square

Confusion matrix evaluation is an evaluation of the K-NN algorithm, K-NN with backward elimination, K-NN with forward selection and K-NN with chi-square which can be compared using recall, precission, accuracy and f-1 score. The comparison can be seen in the following table

Table 18. *Evaluation using confusion matrix on breast cancer dataset*

Evaluation	K-NN	Backward, K-NN	*Forward*, K-NN	*Chi*, K-NN
Recall	0.905882	0.911765	0.917647	0.905882
Precissiion	0.927711	0.928144	0.95122	0.927711
Accuracy	0.937079	0.939326	0.950562	0.937079
F1-Score	0.916667	0.919881	0.934132	0.916667

The visualization of the evaluation and comparison of the algorithms and each feature selection on the breast cancer dataset is as follows:

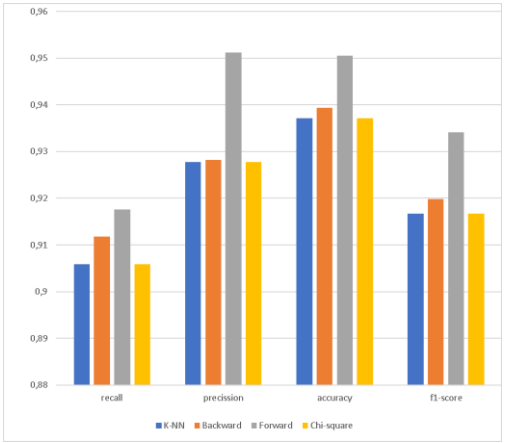

Figure 4. *Evaluation graph of algorithm and feature selection on breast cancer dataset*

While on the breast cancer dataset, the evaluation obtained using the confusion matrix is as follows:

Jurnal Teknik Informatika Vol. 17 No. 2, Oktober 2024 (190-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v17i2.40866

Table 19. *Evaluation using confusion matrix on prostate cancer dataset*

Evaluation	K-NN	Backward, K-NN	*Forward*, K-NN	*Chi*, K-NN
Recall	0.88	0.78	0.86	0.88
Precissiion	0.897959	0.886364	0.955556	0.897959
Accuracy	0.8625	0.8	0.8875	0.8625
F1-Score	0.888889	0.829787	0.905263	0.888889

The visualization of the evaluation and comparison of the algorithms and each feature selection on the breast cancer dataset is as follows:

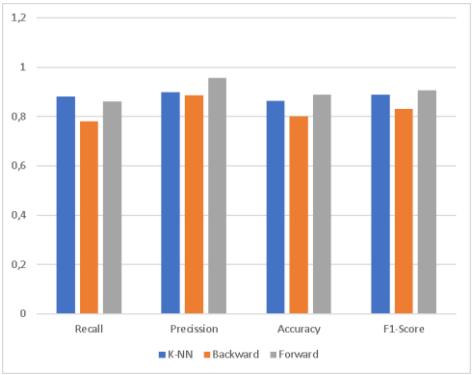

Figure 5. *Evaluation graph of algorithm and feature selection on prostate cancer dataset*

3.7 Evaluation Using T-Test

The T-test is a test step on the research hypothesis that affects the independent variables partially. The T-test is carried out to determine false and True in the hypothesis [24]. The T-test criteria are as follows:

1. If the t-test singnication value >0.05 then H0 is accepted and Ha is rejected.
2. If the t-test significance value is <0.05 then H0 is rejected and Ha is accepted.

Here are the results of the T-test calculation on the breast cancer dataset :

Table 20. *T-test results on breast cancer dataset*

T-test	K-NN	Backward, K-NN	Forward, K-NN	Chi, K-NN
	0.937 +/- 0.043	0.935 +/- 0.051	0.937+/- 0.032	0.935+/- 0.039
0.937 +/- 0.043		0.921	0.991	0.890
0.935 +/- 0.051			0.920	0.982
0.937 +/- 0.032				0.882
0.935 +/- 0.039				

The result of the test is that forward selection is more dominant than other feature selection or the K-NN algorithm without feature selection.

Meanwhile, the following are the results of the T-test on the prostate cancer dataset

Table 21. *T-test results on prostate cancer dataset*

T-test	K-NN	Backward, K-NN	*Forward*, K-NN	*Chi*, K-NN
	0.840 +/- 0.084	0.840 +/- 0.107	0.830 +/- 0.149	0.850 +/- 0.127
0.840 +/- 0.084		1.000	0.856	0.838
0.840 +/- 0.107			0.866	0.851
0.830 +/- 0.149				0.751
0.850 +/- 0.127				
0.935 +/- 0.039				

The result of the test is that forward selection is more dominant than other feature selection or the K-NN algorithm without feature selection.

CONCLUSION

From research from the initial stage to the final stage using feature selection sbackward elimination, forward selection and chi-square which is processed using the K-NN algorithm to evaluate the performance of feature selection on improving K-NN accuracy, the following conclusions can be obtained

Improving the accuracy of the K-NN algorithm is done by adding feature selection forward selection, backward elimination and chi-square then evaluating using the confusion matrix on each feature selection that has been input into the K-NN algorithm. From the results of this experiment, feature selection forward selection has the highest accuracy of 95.12% on the breast cancer dataset and 88.75% on the prostate cancer dataset.

The algorithm used in this research is only limited to the K-NN algorithm. So in future research can use other algorithms so that there are improvisations as well as more varied and diverse research models.

Feature selection used in this research is limited even though there are many feature selections either in wrap, filter or embedded types that can be applied to other classification algorithms or K-NN.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, *Techniques to Improve Classification Accuracy*. 2012.

[2] F. T. Admojo and Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 34–38, 2020, doi: 10.33096/ijodas.v1i2.12.

[3] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using Joint Mutual Information Maximisation," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8520–8532, 2015, doi: 10.1016/j.eswa.2015.07.007.

[4] J. Suntoro and C. N. Indah, "Average Weight Information Gain Untuk Menangani Data Berdimensi," *J. Buana Inform.*, vol. 8, pp. 131–140, 2017.

[5] R. S. Wahono, N. Suryana, and S. Ahmad, "Metaheuristic Optimization based Feature Selection for Software Defect Prediction," *J. Softw.*, vol. 9, no. 5, 2014, doi: 10.4304/jsw.9.5.1324-1333.

[6] A. Bengnga and R. Ishak, "Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Correlation Matrix with Heatmap," *Jambura J. Electr. Electron. Eng.*, vol. 4, no. 2, pp. 169–174, 2022, doi: 10.37905/jjeee.v4i2.14403.

[7] A. Rifa'i, J. Suntoro, and G. G. Setiaji, "GA-SVM Wrapper Feature Selection untuk Penanganan Data Berdimensi Tinggi," *J. Transform.*, vol. 21, no. 2, p. 64, 2024, doi: 10.26623/transformatika.v21i2.8886.

[8] L. W. Astuti, I. Saluza, F. Faradilla, and M. F. Alie, "Optimalisasi Klasifikasi Kanker Payudara Menggunakan Forward Selection pada Naive Bayes," *J. Ilm. Inform. Glob.*, vol. 11, no. 2, 2021, doi: 10.36982/jiig.v11i2.1235.

[9] T. Ernayanti, M. Mustafid, A. Rusgiyono, and A. R. Hakim, "Penggunaan Seleksi Fitur Chi-Square Dan Algoritma Multinomial Naïve Bayes Untuk Analisis Sentimen Pelangggan Tokopedia," *J. Gaussian*, vol. 11, no. 4, pp. 562–571, 2023, doi: 10.14710/j.gauss.11.4.562-571.

[10] I. A. Angreni, S. A. Adisasmita, M. I. Ramli, and S. Hamid, "Pengaruh Nilai K Pada Metode K-Nearest Neighbor (Knn) Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan," *Rekayasa Sipil*, vol. 7, no. 2, p. 63, 2019, doi: 10.22441/jrs.2018.v07.i2.01.

[11] R. Sanjaya and F. Fitriyani, "Prediksi Bedah Toraks Menggunakan Seleksi Fitur Forward Selection dan K-Nearest Neighbor," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, p. 316, 2019, doi: 10.26418/jp.v5i3.35324.

[12] M. A. D. A. F. Rismiati, "SATIN – Sains dan Teknologi Informasi Ukuran Akurasi Klasifikasi Penyakit Mesothelioma Menggunakan Algoritma K-Nearest Neighbor dan Backward Elimination Maxsi Ary," vol. 5, no. 1, 2019.

[13] H. Saleh, "K-Nearest Neighbor Berbasis Seleksi Atribut Chi Square Untuk Klasifikasi Penerima Beasiswa," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 14, no. 1, pp. 1–10, 2023, doi: 10.24176/simet.v14i1.9178.

[14] M. A. Barata, Edi Noersasongko, Purwanto, and Moch Arief Soeleman, "Improving the Accuracy of C4.5 Algorithm with Chi-Square Method on Pure Tea Classification Using Electronic Nose," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 2, pp. 226–235, 2023, doi: 10.29207/resti.v7i2.4687.

[15] T. A. Setiawan and M. A. A. Karomi, "Penerapan Metode Sample Bootstrapping untuk Meningkatkan Performa kNearest Neighbor pada Dataset Berdimensi Tinggi," *J. STMIK IC-Tech*, vol. XII, no. 1, pp. 9–14, 2017, [Online]. Available: http://jurnal.stmik-wp.ac.id

[16] E. K. Garcia, S. Feldman, M. R. Gupta, and S. Srivastava, "Completely lazy learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 9, pp. 1274–1285, 2010, doi: 10.1109/TKDE.2009.159.

[17] A. Riski, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019, [Online]. Available: https://jurnal.kaputama.ac.id/index.php/JTIK/article/view/141/156

[18] I. made B. Adnyana, "Penerapan Feature

Selection untuk Prediksi Lama Studi Mahasiswa," *J. Sist. Dan Inform.*, vol. 13, pp. 72–76, 2019.

[19] D. S. Ramadhansyah, "Perbandingan Metode Seleksi Fitur Filter, Wrapper, dan Embedded Prediksi Kandungan Vitamin C Pada Buah Mangga Meggunakan Metode Linear Regression dan Random Forest Regression [Skripsi]," vol. Yogyakarta, p. Universitas Islam Indonesia, 2022.

[20] E. Nurlia and U. Enri, "Penerapan Fitur Seleksi Forward Selection Untuk Menentukan Kematian Akibat Gagal Jantung Menggunakan Algoritma C4.5," *J. Tek. Inform. Musirawas) Elin Nurlia*, vol. 6, no. 1, p. 42, 2021.

[21] H. Harafani and H. A. Al-Kautsar, "Meningkatkan Kinerja K-Nn Untuk Klasifikasi Kanker Payudara Dengan Forward Selection," *J. Pendidik. Teknol. dan Kejuru.*, vol. 18, no. 1, p. 9 9, 2021,