IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 3, Ser. III (May. –June. 2022), PP 69-79
www.iosrjournals.org

Performance Improvement of Naïve Bayes Algorithm Based on Information Gain and Forward Selection Features Selection for Heart Disease Classification

Widiharto[1], M. Arief Soeleman[2], Abdul Syukur[3]
[1](Magister Teknik Informatika, Universitas Dian Nuswantoro, Indonesia)
[2](PascasarjanaTeknik Informatika, Universitas Dian Nuswantoro, Indonesia)
[3](PascasarjanaTeknik Informatika, Universitas Dian Nuswantoro, Indonesia)

Abstract:
The development of technology is very fast. Technology can be used to diagnose various diseases, one of which is heart disease. To be able to predict, analyze and recognize patterns, a new science emerged, namely data mining. Data mining is a process of gathering information from big data. In data mining there are several methods, one of which is classification. Before classification is carried out on some attributes of the data it may not be relevant if it is included in the classification process. For this reason, the selection of which attributes is a process to find out which attributes are relevant to use and which are not. Attribute selection uses information gain and forward selection algorithms which are then implemented for the nave Bayes classification algorithm using the heart disease dataset. For the evaluation of the test using a confusion matrix. The results of the test show that the proposed method, namely information gain, forward selection and nave Bayes, has an accuracy value of 84.15% from the heart disease dataset, which is superior to the information gain and nave Bayes methods 82.84% and the conventional nave Bayes method with a value of 83.14%.
Key Word*: Data mining, Information Gain, Forward Selection, Naïve Bayes*

Date of Submission: 15-06-2022 Date of Acceptance: 30-06-2022

I. Introduction

In today's life the development of technology is very fast. Today's technology can be used to predict or diagnose disease. To make it easier to make decisions in analyzing, predicting and extracting data, a new branch of science emerged, namely Data Mining.

Data mining can be interpreted as a process of finding correlations and patterns from hundreds or even many fields from a very large database [1]. Collecting patient medical record data or data in the past can be used to diagnose a disease, one of which is heart disease. Heart disease is a disorder that occurs in the large blood vessel system, causing blood circulation to not function properly [2].

Classification is the process of identifying objects into a category, class or group based on predetermined procedures, definitions and characteristics. The purpose of classification is to place objects that are assigned to only one of the categories called classes [3]. According to Han (2012) [4]classification is a description of the important classes of a form of data that has been extracted.

Naïve Bayes is a simple probabilistic classification method for calculating a set of probabilities by adding up the frequency and combination of values from a given dataset. Due to its ease of construction but surprising effectiveness, Naïve Bayes continues to be one of the top 10 data mining algorithms [5]. This algorithm uses the Bayes theorem and assumes that all attributes are independent or not interdependent with a given value to a class variable[6]. It is clear that the assumption of conditional independence in naive Bayes is rarely true in reality, which would compromise its performance in applications with complex attribute dependencies. To weaken the assumption of conditional independence, many approaches are proposed. Related jobs can be broadly divided into five main categories [7]: (1) structure extension[8]–[10]; (2) attribute weighting[11]–[13]; (3) attribute selection [14], [15]; (4) instance weighting [16], [17]; (5) instance selection [17], [18].

Attribute weighting is very practical to use to reduce the main weakness of the Naives Bayes classification algorithm [11]–[13]. Some attributes may not have relevant values for data mining tasks and if irrelevant data is included it can be detrimental and disrupt the task in data mining algorithms [19]. So it is necessary to do attribute selection which is a process in identifying and eliminating attributes with irrelevant values [20]. In attribute weighting for feature selection, [21], [22]used the Forward Selection method to select the best subset of attributes to reduce training time and improve system performance.

The Naïve Bayes classification algorithm has been used by various researchers, one of which was carried out by Hamzah in 2012 [23]. The Naïve Bayes algorithm itself has several advantages, namely it is fast in carrying out calculations, the algorithm is simple and also has high accuracy. In addition to having advantages, the Naïve Bayes algorithm also has weaknesses, namely where a probability cannot measure the accuracy of a prediction, and the weakness of attribute selection so that it affects the accuracy value. Therefore, the Naïve Bayes classification algorithm needs to be optimized by means of feature selection using information gain and then classified using the Nave Bayes algorithm to work more effectively. So in this case, we propose an information gain and forward selection feature selection technique in the Naïve Bayes algorithm to diagnose heart disease which we call IG+FS+NB.

II. Material And Methods

The literature review method used is the Systematic Literature Review (SLR) which was popularized by Kitchenham and Charters in 2007. This SLR is used to identify, analyze and interpret the stages of research based on research questions [24]. The following are some of the research obtained including:

1. A decision tree-based attribute weighting filter for naive Bayes

   Research conducted by Hall in 2007 [25],to determine the weight of each attribute Hall proposed a Decision tree-based attribute weighting nave Bayes (DTAWNB) model, which calculates the dependencies between attributes through an untrimmed decision tree from training examples with random sample. In DTAWNB, the weight of an attribute is inversely proportional to the minimum depth of the decision tree, then a bagging procedure is used to stabilize the estimated weights throughout the ensemble.
2. Alleviating Naive Bayes attribute independence assumption by attribute weighting

   The study was conducted by Zaidi et al in 2013 [26],proposing a Naïve Bayes weighting attributes model to alleviate NB' independence assumption (WANBIA) to optimize attribute weights using gradient descent search, either by maximizing conditional log probability or minimizing mean squared error. . This method generally can achieve better classification accuracy, because the weights are determined based on performance feedback from the classifier itself.
3. A Correlation-Based Feature Weighting Filter for Naive Bayes

   Research conducted by Jiang et al in 2019 [27],proposed another improvement model called correlation-based attribute weighted naive Bayes (CAWNB). In CAWNB, the weight of each attribute is first defined as the difference between the mutual relevance (correlation between attributes and classes) the average reciprocal redundancy (redundancy between attributes). Second, a sigmoid transformation must be performed to ensure that the weights are within a realistic range, the classification accuracy of CAWNB is higher than Naïve Bayes, while maintaining the simplicity of the final model.
4. Feature Selection for Classification using Principal Component Analysis and Information Gain

   Research conducted by Omuya et al in 2021 [28],proposed a Principal Component Analysis - Information Gain (PCA-IG) hybrid model for feature selection. The proposed PCA-IG is able to reduce data dimensions, improve performance and also significantly reduce training time so that the objectives of this study are met.
5. Hybrid data mining model for the classification and prediction of medical datasets

   Raghavendra in 2016 [21],conducted research on data mining predictions for attribute selection using entropy evaluation, mean evaluation and threshold evaluation. This prediction model uses feature subset selection methods (FSM) such as the forward selection method and the backward elimination method to select the best subset of attributes to reduce training time and improve system performance.
6. On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis

   Pratiwi and Adiwijaya in 2018 [29],conducted research on a better sentiment analysis system, proposed a combination of information gain and DF thresholding feature selection (IGDFFS). IGDFFS selects features that have an IG score equal to 0.5. This means that these features are closely related to one class only. This scheme manages to reduce about 90% of unnecessary features. The proposed feature selection selects features that have high information acquisition and high occurrence. The combination of information gain and document frequency in this study proposes feature selection; IGDFFS selects subfeatures that meet the following criteria: (1) high relevance to the output class and (2) high occurrence in the data set. As a result, it builds subfeatures that achieve better performance in classification.

Of the 6 studies using attribute weighting manipulation [21], [25]–[29]it was proven to give positive results and can produce better accuracy than conventional classification methods.

Data Cleaning

Datasets that exist in the real world tend to be incomplete (missing value), noise and inconsistent. Data cleaning is carried out as an iterative two-step process consisting of difference detection and data transformation [4].Data cleaning attempts to fill in missing values, cleans up noise when identifying outliers and corrects inconsistencies in the data.

Missing values, noise and inconsistencies contribute to inaccurate data. The first step in data cleansing as a process is discrepancy detection. Differences can be caused by several factors, including poorly designed data entry, human error in data entry, intentional errors (e.g. respondents in data entry did not want to divulge their own data) and outdated data (e.g. address data that has expired). . Other sources of discrepancy include errors while recording data and errors in the system.

Discretize

Classification algorithms developed from the field of machine learning are often referred to as discrete and continuous attributes. Each type can be processed differently. Discrete attributes have an infinite or infinite set of values to be computed, which may or may not be represented as integers [4]. Discrete attributes may have numeric values, such as 0 and 1 for binary or values 0 through 110 for the age attribute.

Discretization method is used to reduce the number of values for certain continuous attributes, by dividing the attribute range into interval values, interval labels can be used to replace the actual data values and can speed up and simplify the integration of data processing from the discretization method. There are two categories of discretization, the first is unsupervised discretization and supervised discretization. The supervised discretization is for classification and regression data mining tasks while the unsupervised one is for clustering data mining [4]. An example of using discretize using the RapidMiner application using the Discretize operator is shown in Figure 2.5. By using the Discretize model, we can select the continuous attributes that we will discretize.

Information Gain

Information gain is done to determine the best attribute. Information gain uses entropy in determining the best attribute, entropy is a measure of uncertainty where the higher the entropy value, the higher the uncertainty [30]. The following is the entropy equation (1)

$$E(S) = -\sum_{j=1}^{n} f_s(j)\, log_2\, f_s\,(j) \qquad (1)$$

E(S) : entropy information of attribute S
N : the sum of the different values in the attribute S
$F_s(j)$: the frequency of the value of j in S
Log2 : binary logarithm

Information Gain from the output data or dependent variable y which is grouped based on attribute A, is denoted by gain (y, A). information gain, dain (y, A), of attribute A relative to the output data y is like equation (2)

$$gain(y, A) = Entropy(y) \sum_{c \in nilai\ (A)} \frac{y_c}{y} entropy(y_c) \qquad (2)$$

Where the value (A) is the number of possible attributes of A, and yc is a subset of y where the value of A has a value of c. The first provision in the information gain equation above is the total entropy of y and the second is the entropy after data separation is based on attribute A.

Forward Selection

Forward selection is one way to determine the most influential attribute in a dataset by negotiating the attributes one by one until the relevant attributes are obtained. Forward selection is used during data processing to select appropriate features to build models in data mining [22]. But forward selection is used to improve accuracy and prediction and reduce computational complexity.

The forward selection method will be applied to the prediction of heart disease using the nave Bayes algorithm. Forward selection is used to select attributes that meet the criteria, so that only selected attributes will enter the classification process. It is hoped that the selection of attributes using forward selection can overcome the problem of class imbalance and increase the accuracy of predictions.

For any data mining process based on prediction and classification models, feature selection is very important because when we build a data mining model, the data set most often contains additional information than the actual information needed to build the model. If we save the attribute columns that are not needed then

more Central Processing Unit (CPU) time and memory are required during the training process and these additional attributes can also degrade the quality of the patterns found due to reasons like the first some attributes are noise and redundant making it difficult to find meaningful patterns of data and to identify quality patterns, most data mining algorithms require much larger training datasets but very small training data in some data mining applications [21]. Forward selection processing steps as shown in Figure 1, the first is to maintain the class attribute and the first attribute, then the second is to create two variables A and B. Variable A contains a list of names based on ascending order and entropy value and or number 1. Then variable B stores the attribute names in their original order. Next, compare the list of variables A and B. Third, if they are the same, then remove the attribute from the dataset and also remove the attribute from variable 2 and evaluate. Do the second step until the last attribute in the dataset.

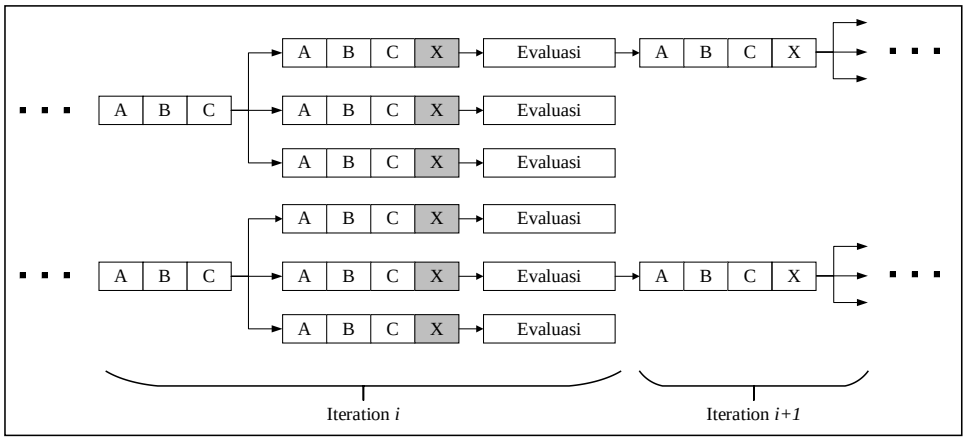

Figure 1.Forward Selection Method

Naïve Bayes

Naïve Bayes is a simple classification algorithm in which each attribute is independent and allows contributing to the final decision [31].

The basic Bayesian theorem used in programming is the Bayesian formula, which is as follows [4] :

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \quad (3)$$

Where:
X = Data whose class is unknown
H = Hypothesis data X is a specific class
P(H|X) = Probability of hypothesis H based on condition X (Posterior Probability)
P(H) = Hypothesis Probability H (Prior Probability)
P(X) = Probability X

Heart Disease

Heart Disease or Heart Attack is a very serious heart disorder when the heart muscle does not circulate blood flow, this condition interferes with the heart's function in circulating blood to all parts of the body. Cholesterol deposits lead to the formation of plaque on the walls of blood vessels. This is if people have high cholesterol, it will put a person at risk for heart disease. There are several symptoms that are commonly felt by people with heart attacks, including chest pain, shortness of breath or heavy breathing, anxiety, dizziness and cold sweats. However, there are also patients with heart attacks who do not experience symptoms immediately, and immediately experience sudden cardiac arrest [32].

Framework for Thinking

Based on Figure 2, the flowchart of the framework that will be carried out in this study is the selection of Naïve Bayes features into one framework. Starting with entering training data then whether to use attribute selection? If yes then it will calculate the information gain on each attribute, short the attribute gain value from the largest to the smallest. Next, reduce the attribute of the lowest gain value, after that we enter the forward selection method and then train using nave Bayes. Furthermore, interpretation is carried out using testing data, when testing with nave Bayes, a forward selection is also carried out after which predictive data will come out. And the last is evaluated to determine the accuracy and error rate, finished.

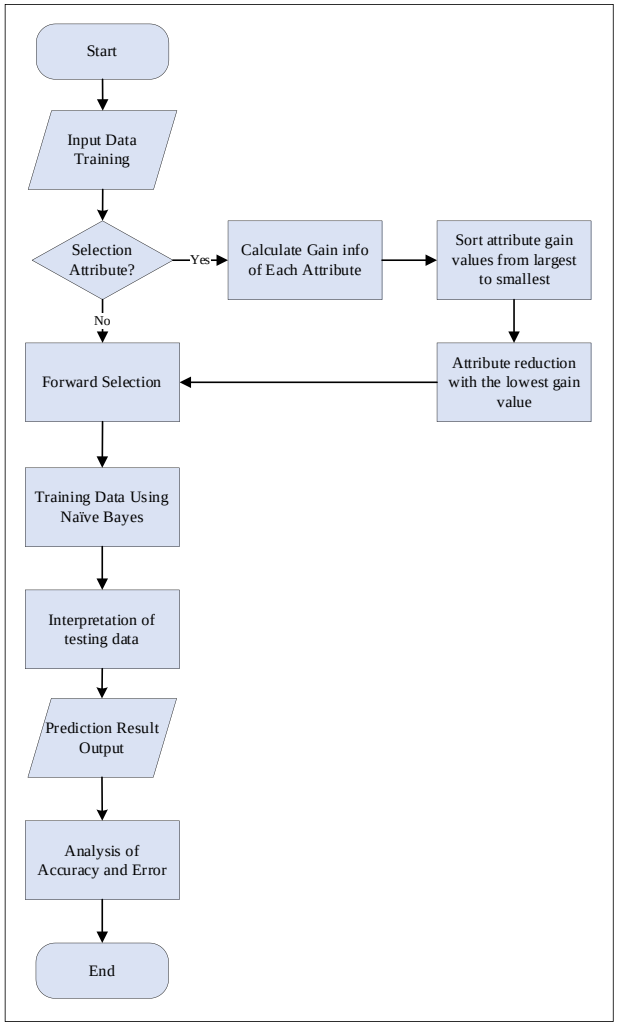

Figure 2.Flowchart of the IG+FS+NB proposed method framework

III. Research Methodology

This chapter will explain the stages in experimental research consisting of 5 stages of research carried out, namely: the first is data collection consisting of literature collection and dataset selection, the second is preprocessing of the data, the third is modeling using the proposed method, the third is modeling using the proposed method. the four test experiments used the RapidMiner Studio Free 9.9 application, and the fifth was the evaluation of the experimental results using the Confusion Matrix.

Dataset Selection

At this stage the authors use datasets for research obtained from the University California Irvinne (UCI) Repository machine learning dataset. The author uses 4 datasets as a test, namely Heart Disease Dataset, Breast Cancer Dataset, Hepatitis Dataset, Mushrooms Dataset. Each dataset has its own character, and certain datasets must be handled with special care, such as in preprocessing, namely handling missing value and noise data.

Proposed Method

At this stage the proposed method is feature selection in the nave Bayes algorithm. As in Figure 3, the first is from the dataset then pre-processing is carried out, at this stage Data Cleaning and Discretize are carried out. The second is feature selection using information gain to determine the best attribute. Then the forward selection method is carried out, after that the classification is carried out using the nave Bayes algorithm. At the time of classification, 10 fold-cross validation was carried out to select the best model. In 10 fold-cross validation, the test will be repeated 10 times and the result of the measurement is the average value of 10 tests. For each of the 10 subsets, 9 folds were used for training and 1 fold for testing. After that, an evaluation is carried out with a confusion matrix to determine Accuracy, Precision, and Recall.

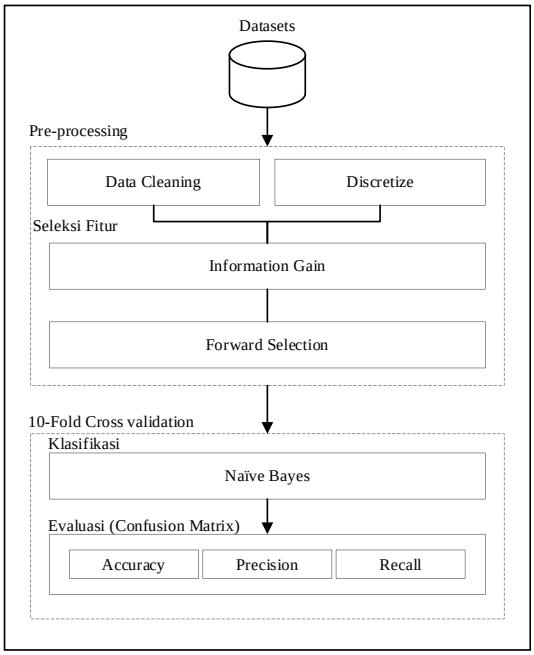

Figure 3.Proposed Method

Testing Experiment

The experimental stages and method testing were carried out using a computer with a specification of an Intel Core i5 1.60 GHz processor, 16 GB RAM and a 512 GB SSD hard drive. Using the Windows 11 64bit operating system. The application is in experimental testing using RapidMiner Studio Free 9.9. Table 3.5 displays a list of computer specifications and applications used for classification with the Nave Bayes algorithm.

Evaluation of Experimental Results

Experiments from experimental results were evaluated using measurements of accuracy, precision, and recall. Confusion Matrix is a summary of predictions on classification problems [31]. Table 1. displays a chart of the confusion matrix table.

Table 1. Confusion Matrix

		Aktuals	
		Positif	Negatif
Prediksi	Positif	**TP**	**FP**
	Negatif	**FN**	**TN**

Where:
TP (True Positive): The actual number of sick patients and the model predicts illness.
TN (True Negative): The actual number of not being sick and the model predicting not being sick.
FP.(False Positive) : Actual number of no pain but model predicting illness.
FN.(False Negative) : The actual number of sick but the model predicts no pain.

$$Akurasi = \frac{\text{TP .+. TN}}{\text{TP.+.FP.+.FN.+.TN}} \quad (4)$$

$$Presisi = \frac{\text{TP.}}{\text{TP.+.FP}} \quad (5)$$

$$Recall = \frac{\text{TP.}}{\text{TP.+.FN}} \quad (6)$$

$$TPR = \frac{\text{TP.}}{\text{TP.+.FN}} \quad (7)$$

$$FPR = \frac{\text{FP.}}{\text{FP.+.TN}} \quad (8)$$

$$AUC = \frac{1 + \text{TPrate} - \text{FPrate}}{2} \quad (9)$$

Accuracy is the proportion of the total number of correct predictions. Precision is the predicted correct proportion of relevant pages. While recall is the proportion of relevant pages that are correctly identified. The Area Under Curve (AUC) can be calculated based on the average trapezoidal area estimate for the curve created

by Precision and Recall. (AUC) is calculated as a measure of the Receiver Operating Characteristics (ROC) curve area using equation [22]. The ROC curve is the ratio of the two characteristics True Positive Rate (TPR) and False Positive Rate (FPR).

IV. Results and Discussion

This chapter describes the results of testing and discussion of classification in predicting heart disease using the nave Bayes algorithm and information gain and forward selection as parameters in attribute weighting. In training and classification tasks using the Rapidminer application. Based on the results of training and testing using the Rapidminer application, it will be concluded whether the nave Bayes algorithm with attribute selection using information gain and forward selection can increase the accuracy value compared to the conventional nave Bayes algorithm.

Result

The results of the tests carried out to predict heart disease using the Heart Disease dataset using the Rapidminer application. Here the author also uses other datasets such as the Wisconsin Breast Cancer, Hepatitis and Mushrooms dataset. In the classification process, three classification comparisons are carried out, the first is classification using the nave Bayes algorithm, the two features are weighted using information gain and then classified using the nave Bayes algorithm, and the third features are weighted using information dain and forward selection and then classified using the nave Bayes algorithm. The following are the results of the modeling using the Rapidminer application. The results of the tests will be shown in sub-chapter 4.2.1 on Heart Disease Dataset, sub-chapter 4.2.2 on Breast Cancer Dataset, sub-chapter 4.2.3 on Hepatitis Dataset, and sub-chapter 4.2.4 on Mushrooms Dataset.

Table 2. Test results using rapid miner

Dataset	Akurasi			Presisi			Recall			AUC		
	NB	IG+NB	IG+FS+NB	NB	IG+NB	IG+FS+NB	NB	IG+NB	IG+FS+NB	NB	IG+NB	IG+FS+NB
Heart Disease	83,14%	82,84%	84,15%	83,44%	83,62%	85,24%	79,78%	79,12%	79,51%	0,901	0,902	0,878
Breast Cancer	73,45%	73,10%	74,79%	60,29%	57,44%	64,07%	49,58%	47,22%	34,58%	0,689	0,683	0,682
Hepatitis	81,83%	83,96%	86,50%	66,00%	63,33%	82,35%	53,33%	60,00%	42,50%	0,794	0,823	0,833
Mushrooms	99,56%	99,59%	99,80%	99,69%	99,60%	99,62%	99,45%	99,62%	100,00%	1,000	1,000	1,000

Heart Disease Dataset

1. Naïve Bayes (NB)
   Table 2 showing that the accuracy of the classification of the Heart disease dataset using the Nave Bayes algorithm is 83.14%, for the precision value in this classification is 83.44%, recall is 79.78% and AUC value 0.901.
2. Information Gain + Naïve Bayes (IG+NB)
   Table 2 shows that the accuracy is 82.84%, the precision is 83.62%, then the recall is 79.12% and the AUC is 0.902.
3. Information Gain + Forward Selection + Naïve Bayes (IG+FS+NB)
   Table 2 which shows the accuracy value is 84.15%, the result of precision is 85.24%, then the result of recall is 79.51% and the result of AUC is 0.878.

Breast Cancer Dataset

1. Naïve Bayes (NB)
   Table 2 with an accuracy of 73.45%, then the results of the precision obtained are 60.29%. For recall, the result is 49.58% and AUC is 0.989.
2. Information Gain + Naïve Bayes (IG+NB)
   Table 2 the result of the accuracy value is 73.10%, precision is 57.44%, recall is 47.22% while the AUC value is 0.683.
3. Information Gain + Forward Selection + Naïve Bayes (IG+FS+NB)
   Table 2 with an accuracy of 74.79%, then a precision of 64.07%, for recall it gets a value of 34.58% and for AUC it gets 0.682.

Hepatitis Dataset

1. Naïve Bayes (NB)

Table 2 where the accuracy value of the hepatitis dataset classification using the nave Bayes algorithm is 81.83%, then for precision the value is 66.00%, then for recall the value is 53.33% while the AUC value is 0.794.

2. Information Gain + Naïve Bayes (IG+NB)
   Table 2 where the accuracy value of the hepatitis dataset classification using the nave Bayes algorithm is 83.96%, then for precision the value is 63.33%, then for recall the value is 60.00% while the AUC value is 0.823.
3. Information Gain + Forward Selection + Naïve Bayes (IG+FS+NB)
   Table 2 which for accuracy obtained is 86.50%, then for precision is 82.35%, the recall value obtained is 42.50, while the AUC value itself is 0.833.

Mushrooms Dataset

1. Naïve Bayes (NB)
   Table 2 with an accuracy value of 99.56%. Then for the results of the precision that is 99.69%, for recall the value is 99.45%, while for the AUC value is 1,000.
2. Information Gain + Naïve Bayes (IG+NB)
   Table 2 namely the accuracy value is 99.59%, then the precision value obtained is 99.60%. The recall value obtained is 99.62%, and the AUC value is 1,000.
3. Information Gain + Forward Selection + Naïve Bayes (IG+FS+NB)
   Table 2 where the accuracy value is 99.80%, then the precision value is 99.62%. As for the recall, the score is 100.00% and the AUC is 1,000.

Discussion

From the results of the tests carried out, the results are summarized to compare the results. In Table 2 is the accuracy of the test dataset, for the heart disease dataset the NB model gets an accuracy value of 83.14%, then the IG+NB model gets an accuracy value of 82.84% and the IG+FS+NB model gets an accuracy value of 84.15%. Of the three models, the proposed model IG+FS+NB obtains greater accuracy than the NB and IG+NB models. In the breast cancer dataset, the test for the NB model gets an accuracy value of 73.45%, then for testing using the IG+NB model it gets an accuracy value of 73.10% and the next test is the proposed model IG+FS+NB gets an accuracy value of 74, 79%. Therefore, from the three proposed models, the proposed model, namely IG+FS+NB, gets a greater accuracy value than the NB model and the IG+NB model. The next test of the dataset is to use the hepatitis dataset. The NB model gets an accuracy value of 81.83%, then the IG+NB model gets an accuracy value of 83.96% and for the proposed model, the IG+FS+NB model gets an accuracy value of 86.50%. Of the three models tested using the hepatitis dataset, the highest accuracy value was obtained, namely the IG+FS+NB model which was higher than the NB and IG+NB models. The last test is using the mushrooms dataset, which is obtained for the accuracy value of NB modeling which is 99.56%, from the next model using the IG+NB model, the accuracy value is 99.59% and for the proposed model, namely IG+FS+NB, it is obtained 99.80% accuracy value. From testing using the mushrooms dataset, the proposed model, namely the IG+FS+NB model, obtained better accuracy results than the NB and IG+NB modeling.

Comparison of the accuracy of the modeling of NB, IG+NB, and IG+FS+NB is presented in the form of a comparison diagram as shown in Figure 4 which can be seen in the diagram that the four datasets used for the proposed modeling, namely IG+FS+NB, outperform other models such as NB and IG+NB.

As shown in Figure 4 in the heart disease dataset, the best method is the IG+FS+NB method with an accuracy rate of 84.15%, 1.31% superior to the IG+NB accuracy and 1.01% superior to the NB method. Furthermore, the breast cancer dataset with the best accuracy rate is the IG+FS+NB method with a value of 74.79%, 1.34% superior to the NB method and 1.69% superior to IG+NB. In the hepatitis dataset, the IG+FS+Nb method is again superior with an accuracy value of 86.50%, 2.54% superior to IG+NB and 4.67% outperforming the NB method. In the mushroom dataset, the difference between IG+NB and IG+FS+NB is 0.21%, but the proposed modeling IG+FS+NB is superior.

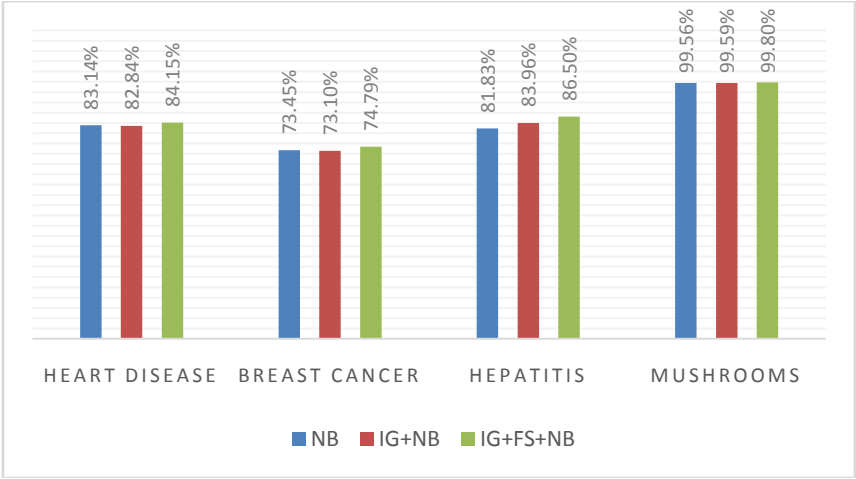

Figure 4. Accuracy comparison chart of the test model

Table 2 shows a comparison of the results of precision, namely for the heart disease dataset with the NB model 83.44%, then the IG+NB model 83.62% and the IG+FS+NB model getting a value of 85.24%. Then, the breast cancer dataset with the NB model gets a value of 60.29%, then the IG+NB model gets a value of 57.44% and the IG+FS+NB model gets a value of 64.07%. For further testing using the hepatitis dataset with the NB model getting a value of 66.00%, the IG+NB model getting a value of 63.33% and the IG+FS+NB model getting a value of 82.35%. And the last dataset is mushrooms, with the NB model getting a value of 99.69%, then the IG+NB model getting a value of 99.60% and the IG+FS+NB model getting a value of 99.62%.

In this study, the results of recall are shown in Table 2 with the heart disease dataset for the NB model getting a value of 79.78%, then the IG+NB model getting a value of 79.12% and for the IG+FS+NB model getting a value of 79.51%. For the next dataset, namely breast cancer in the NB model, it gets a recall value of 49.58%, then for the IG+NB model it gets a value of 47.22% and for the IG+FS+NB model it gets a value of 34.58%. Furthermore, for the hepatitis dataset, the NB model gets a recall value of 53.33%, then for the IG+NB model it gets a value of 60.00% and for the IG+FS+NB model it gets a value of 42.50%. The last dataset, namely moshrooms dataset on the NB model, gets a recall value of 99.45%, then for the IG+NB model it gets a value of 99.62% and for the IG+FS+NB model it gets a 100% value.

The test results from the AUC are shown in Table 2 where the heart disease dataset for the NB model gets a value of 0.901, for the IG+NB model it gets a value of 0.902 and for the IG+FS+NB model it gets a value of 0.878. Furthermore, the breast cancer dataset for the NB model gets a value of 0.689, then for the IG+NB model it gets a value of 0.683 and for the IG+FS+NB model it gets a value of 0.682. The next dataset used is the hepatitis dataset with the NB model getting a value of 0.794, for the IG+NB model it gets a value of 0.823 and for the IG+FS+NB model it gets a value of 0.833. And the last one is the mushrooms dataset, namely for the NB, IG+NB and IG+FS+NB models, they both get an AUC value of 1,000.

V. Conclusion and Sugestion

Conclusion

The conclusion in this study is the proposed method in dealing with the problem of attribute independence in the nave Bayes algorithm by weighting the attributes in nave Bayes. In this test, the authors use datasets from the UCI repository, which include the Heart Disease Dataset, Breast Cancer Dataset, Hepatitis Dataset and Mushrooms Dataset. Attribute weighting using information gain and forward selection with the nave Bayes algorithm produces a better accuracy value on the heart disease dataset for prediction of heart disease with an accuracy value of 84.15% compared to the information gain method and nave Bayes results in an accuracy of 82.84%. and the nave Bayes method only reached 83.14%. The test also uses other datasets such as the breast cancer dataset using the proposed method, namely information gain and forward selection with the nave Bayes algorithm, which also gets the best accuracy, with an accuracy value of 74.79%, superior to the information gain method and nave Bayes with an accuracy value of 73. ,10% and nave Bayes method 73.45%. In the hepatitis dataset, the information gain and forward selection methods using the nave Bayes algorithm are also superior to the nave Bayes and information gain methods and the nave Bayes method with an accuracy value of 86.50% compared to 83.96% and 81.83%, respectively. The author also uses the mushrooms dataset for this test with the proposed method of information gain and forward selection using the nave Bayes algorithm outperforming the nave Bayes and information gain and nave Bayes methods with accuracy values of 99.80% compared to 99.59% and 99.56%. With the results of this test, although using different datasets, the results of

the proposed method, namely information gain and forward selection using the classification algorithm or IG+FS+NB, outperform the Naïve Bayes or NB and information gain nave Bayes or IG+NB methods.

Sugestion

With the completion of this research in the discussion of attribute weighting using information gain and forward selection on the nave Bayes algorithm to predict heart disease, the authors suggest several things for future research updates, namely:

1. Weighing the attributes using other methods from the author such as the GainRatio, Greedy or ChiSquare methods.
2. The test uses a classification algorithm other than the nave Bayes algorithm.
3. The test uses a dataset with a larger number of records.

References

[1] A. P. Windarto, "Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method," *Int. J. Artif. Intell. Res.*, vol. 1, no. 2, p. 26, 2017, doi: 10.29099/ijair.v1i2.17.

[2] N. A. Widiastuti, S. Santosa, and C. Supriyanto, "ALGORITMA KLASIFIKASI DATA MINING NAÏVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION UNTUK DETEKSI PENYAKIT JANTUNG," *J. Pseudocode*, vol. 1, no. 1, pp. 11–14, 2014.

[3] M. Bramer, *Principles of data mining*. 2016.

[4] J. Han, M. Kamber, and J. Pei, *DATA MINING : Concepts and techniques*. 2012.

[5] X. Wu *et al.*, *Top 10 algorithms in data mining*. 2008.

[6] A. Saleh, "Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," *Creat. Inf. Technol. J.*, vol. 2, no. 3, pp. 207–217, 2015.

[7] L. Jiang, Z. Cai, and D. Wang, "Improving Naive Bayes for Classification IMPROVING NAIVE BAYES FOR CLASSIFICATION," vol. 7074, no. April, 2016, doi: 10.2316/Journal.202.2010.3.202-2747.

[8] L. Jiang, H. Zhang, and Z. Cai, "A Novel Bayes Model : Hidden Naive Bayes," vol. 21, no. 10, pp. 1361–1371, 2009.

[9] G. I. Webb, "Not So Naive Bayes : Aggregating One-Dependence Estimators," pp. 5–24, 2005.

[10] J. Wu, S. Pan, X. Zhu, P. Zhang, and C. Zhang, "SODE : Self-Adaptive One-Dependence Estimators for classi fi cation," vol. 51, pp. 358–377, 2016, doi: 10.1016/j.patcog.2015.08.023.

[11] L. Jiang, C. Li, S. Wang, and L. Zhang, "Engineering Applications of Arti fi cial Intelligence Deep feature weighting for naive Bayes and its application to text classi fi cation," *Eng. Appl. Artif. Intell.*, vol. 52, pp. 26–39, 2016, doi: 10.1016/j.engappai.2016.02.002.

[12] L. Jiang, L. Zhang, L. Yu, and D. Wang, "Class-specific attribute weighted naive Bayes," *Pattern Recognit.*, vol. 88, pp. 321–330, 2019, doi: 10.1016/j.patcog.2018.11.032.

[13] H. Zhang, L. Jiang, and L. Yu, "Class-specific attribute value weighting for Naive Bayes," *Inf. Sci. (Ny).*, vol. 508, pp. 260–274, 2020, doi: 10.1016/j.ins.2019.08.071.

[14] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," 2000.

[15] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Syst.*, p. 105361, 2019, doi: 10.1016/j.knosys.2019.105361.

[16] L. Jiang and Y. Guo, "Learning lazy Naive Bayesian classifiers for ranking," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 2005, pp. 412–416, 2005, doi: 10.1109/ICTAI.2005.80.

[17] L. Jiang, D. Wang, and Z. Cai, "Discriminatively weighted naive Bayes and its application in text classification," *Int. J. Artif. Intell. Tools*, vol. 21, no. 1, pp. 1–19, 2012, doi: 10.1142/S0218213011004770.

[18] W. Xu, L. Jiang, and L. Yu, "An attribute value frequency-based instance weighting filter for naive Bayes," *J. Exp. Theor. Artif. Intell.*, vol. 31, no. 2, pp. 225–236, 2019, doi: 10.1080/0952813X.2018.1544284.

[19] B. Azhagusundari and A. S. Thanamani, "Feature Selection based on Information Gain," *Int. J. Innov. Technol. Explor. Eng.*, vol. 2, no. 2, pp. 18–21, 2013.

[20] R. Abraham, J. B. Simha, and S. S. Iyengar, "Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical datamining," *Int. J. Comput. Intell. Res.*, vol. 5, no. 2, pp. 116–129, 2009, doi: 10.5019/j.ijcir.2009.175.

[21] S. Raghavendra and M. Indiramma, "Hybrid data mining model for the classification and prediction of medical datasets," *Int. J. Knowl. Eng. Soft Data Paradig.*, vol. 5, no. 3/4, pp. 262–284, 2016, doi: 10.1504/ijkesdp.2016.10005501.

[22] A. Saifudin, Ekawati, Yulianti, and T. Desyani, "Forward Selection Technique to Choose the Best Features in Prediction of Student Academic Performance Based on Naïve Bayes," *J. Phys. Conf. Ser.*, vol. 1477, no. 2, 2020, doi: 10.1088/1742-6596/1477/3/032007.

[23] A. Hamzah, "Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis," *Pros. Semin. Nas. Apl. Sains Teknol. Periode III*, no. 2011, pp. 269–277, 2012, doi: 1979-911X.

[24] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," *EBSE Tech. Rep. Version 2.3,EBSE-2007-01*, 2007.

[25] M. Hall, "A decision tree-based attribute weighting filter for naive Bayes," *Knowledge-Based Syst.*, vol. 20, no. 2, pp. 120–126, 2007, doi: 10.1016/j.knosys.2006.11.008.

[26] N. A. Zaidi, J. Cerquides, M. J. Carman, and G. I. Webb, "Alleviating Naive Bayes attribute independence assumption by attribute weighting," *J. Mach. Learn. Res.*, vol. 14, pp. 1947–1988, 2013, doi: 10.13039/501100000923.

[27] L. Jiang, L. Zhang, C. Li, and J. Wu, "A Correlation-Based Feature Weighting Filter for Naive Bayes," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 201–213, 2019, doi: 10.1109/TKDE.2018.2836440.

[28] E. Odhiambo Omuya, G. Onyango Okeyo, and M. Waema Kimwele, "Feature Selection for Classification using Principal Component Analysis and Information Gain," *Expert Syst. Appl.*, vol. 174, no. January, p. 114765, 2021, doi: 10.1016/j.eswa.2021.114765.

[29] A. I. Pratiwi and Adiwijaya, "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis," *Appl. Comput. Intell. Soft Comput.*, vol. 2018, 2018, doi: 10.1155/2018/1407817.

[30] M. Slocum, "Decision Making Using Id3 Algorithm," *InSight RIVIER Acad. J.*, vol. 8, no. 2, pp. 1–12, 2012.

[31] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training

Web Pages," *Int. J. Comput. Sci.*, vol. 4, no. 1, pp. 16–23, 2009, [Online]. Available: http://cogprints.org/6708/.
[32] R. Hajar, "Risk factors for coronary artery disease: Historical perspectives," *Hear. Views*, vol. 18, no. 3, pp. 109–114, 2017, doi: 10.4103/heartviews.heartviews_106_17.

Widiharto, et. al. "Performance Improvement of Naïve Bayes Algorithm Based on Information Gain and Forward Selection Features Selection for Heart Disease Classification." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 24(3), 2022, pp. 69-79.