ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse82932020.pdf
https://doi.org/10.30534/ijatcse/2020/82932020

Heart Disease Prediction System using Data Mining Classification Techniques: Naïve Bayes, KNN, and Decision Tree

Maria Theresa Viega[1], Eric Marvin[2], Riyanto Jayadi[3], Tuga Mauritsius[4]
[1]Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management, Bina Nusantara University, Jakarta – Indonesia 11480, maria.viega001@binus.ac.id
[2]Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management, Bina Nusantara University, Jakarta – Indonesia 11480, eric.marvin001@binus.ac.id
[3]Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management, Bina Nusantara University, Jakarta – Indonesia 11480, riyanto.jayadi@binus.edu
[4]Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management, Bina Nusantara University, Jakarta – Indonesia 11480, tmauritsus@binus.edu

ABSTRACT

Heart-related illnesses are one of the significant causes of death within the world nowadays. Most people do not realize they have heart disease until it is too late. Some parameters can be used to predict it, such as chest pain type, age, sex; fasting blood sugar; maximum heart rate. In this paper, using Naïve Bayes, Decision Tree, and K-Nearest Neighbor (KNN), a prediction of heart disease classification is presented. The results show that our proposed data mining technique using Naive Bayes can predict as high as 86% accuracy outperforming the previous works. Besides, Naïve Bayes is the best model in this study since it has the best values in terms of precision, accuracy, and specificity compared to other models.

Key words: Heart disease; Naïve Bayes, Decision Tree, K-Nearest Neighbor

1. INTRODUCTION

The heart is a healthy organ about the size of a clench hand, found directly above and simply left of the breastbone. The heart siphons blood into the cardiovascular framework, called the conduit and venous system. If it does not work appropriately, at that point, the cerebrum and different organs stop to work, and in two or three minutes, the individual passed on. Change in the way of life, worry from work, and poor dietary patterns add to the paces of a few heart illnesses.

Cardiac illness has developed as one of the foremost persuasive causes of passing around the world. The WHO expressed in 2015 that around 17 million individuals in the world die consistently because of heart illnesses. This number is approximately 31% of all global deaths. This phenomenon also happens in Indonesia. The research by Health Research Baseline in 2018 indicates that 15 out of 1.000 Indonesians have suffered heart disease. From the cause of death perspective, according to SRS (Sample Registration Survey System), in 2014, around 12,4% of death cases in Indonesia were caused by heart disease. Consequently, the plausible and exact expectation of cardiovascular sicknesses is critical.

Many research gathers data on various health-related issues. These data are dissected utilizing various procedures of artificial intelligence to get helpful bits of knowledge. These data can also be vast and noisy, too challenging to comprehend for human minds, but not for various machine learning techniques such as Naïve Bayes, KNN, and Decision Tree. Thus, in recent times, these algorithms have become very useful for reliably predicting the existence of heart-related diseases.

Data mining is a process to find models and patterns in a large dataset that involves methods for integrating machine learning, statistics and database systems, which are essential processes when implemented in extracting data patterns. Data mining can be done by way of classification, clustering, prediction, association, or time series analysis [1]. Therefore, it is possible to conclude that data mining refers to the acquisition of knowledge sourced from large amounts of data.

Many things can be done on the results of the application of data mining in various fields, one of which is the health sector. According to [2], the application of data mining can be beneficial in public health policymaking, hospital error prevention, pandemic management, as well as early detection and prevention of disease.

In this study, we present a data mining classification to determine the heart disease existence utilizing three classification models, specifically KNN, Decision Tree, and Naïve Bayes utilizing RapidMiner software. We present data preprocessing techniques and parameters on different models. With a comprehensive evaluation between the three different classification algorithms, the best model that can predict heart disease is presented.

In the next section, we present the previous works, the data mining algorithms and the methodology utilized in this study. Chapter 3 present the methodology. The results and discussion are presented in Chapter 4. Lastly, Chapter 5 presents the conclusion of this study.

2. LITERATURE REVIEW

The primary dataset of heart disease that is used in this research come from the UCI Machine Learning Repository[1]. Previous studies had conducted modeling of heart disease prediction on this dataset, the Naïve Bayes and Decision Tree models. Naïve Bayes can classify patients suspected of having heart disease with an accuracy rate of 85.03% [3] and in other studies with an accuracy rate of 83.49% [4]. Also, the decision tree model in previous studies can classify this dataset with an accuracy level of 84% [3] and 84.1% [5]. However, previous studies also proved that classification using other algorithms such as KNN could have better results compared to other models with an accuracy value of 79.20% [6] (Table 1).

Table 1: Results of Previous Studies Using Same Datasets

Previous Studies	**Accuracy**		
	KNN	**Naive Bayes**	**Decision Tree**
[6]		83,49%	
[7]	79%		
[5]			84,1%
[3]		85,03%	84%

Therefore, in this study, the author decides to compare the results obtained from data mining processing with the model algorithm as follows.

2.1 Naïve Bayes

Naïve Bayes depends on the Bayes Theorem, a productive strategy for classification. It infers the opportunity of indicators. For example, the attributes or characteristics should not be related to each other or should not be related to each other in any capacity. Regardless of whether there is reliance, these highlights or credits, despite everything, contribute independently to the probability, which is the reason it is called naive. Notwithstanding its straightforwardness, the Naïve Bayes classifier regularly performs shockingly well and is generally utilized because it frequently beats increasingly complex techniques for classification.

2.2 K-Nearest Neighbor

KNN is a straightforward, non-parametric, and apathetic classifier. Because of its quick combination speed and effortlessness, KNN is preferred over another classification algorithm [1].

KNN is one of the classification techniques which are mostly straightforward, however productive. It does not permit data presumptions and is regularly utilized for classification tasks where the data on data dissemination is constrained to no earlier. This algorithm involves finding and assigning the average value of the identified data points to the nearest k data points in the training set at the data point for which a target value is not accessible. The KNN calculation assumes, to put it plainly, that similar things happen in short proximity. In different terms, things indistinguishable are close to each other.

2.3 Decision Tree

Decision Tree is a decision help apparatus that utilizes a tree-like choice graph or model and its latent capacity yields, including the result of chance occasions, resource costs, and utility [8]. Decision Tree constructs a network structure with regards to classification and regression models [9]. It separates the data assortment into smaller subsets while simultaneously, step by step, making a connected decision tree. The results of the model have nodes for decision and nodes for a root.

The decision node includes two or three divisions. The node of the leaf is a ranking or judgment. The highest decision node in the tree corresponding to the root node is named the best predictor. Decision trees are capable of handling both categorical and numerical data.

2.4 CRIPS-DM

The whole process in this study is based on CRISP-DM. This framework suggests a sequence of steps that should be employed to develop a data mining model. Numerous data mining specialists utilize CRISP-DM, but it is most broadly advanced by IBM and executed in SPSS Modeler. Clients can lose their subtleties or take a more relaxed approach and utilize it as a valuable instrument.

The steps involved in CRISP-DM are [10]:

1. *Business Understanding.* It involves understanding and translating the project objectives from a business perspective into a data mining operation.
2. *Data Understanding.* Familiarization of data; the quality, amount and data are the most related to the process.
3. *Data Preparation.* Includes selection of the cleaning, transformation and details. This process is iterative.
4. *Modeling.* The method of applying data mining algorithms to the data. Parameter selection is the most vital decision in this step.

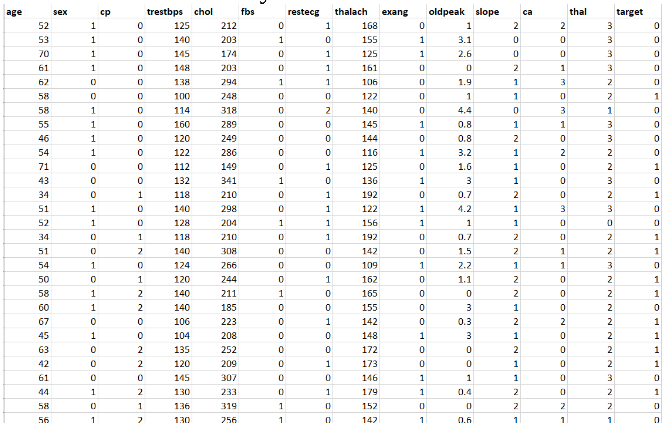
[1] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

5. *Evaluation.* It involves validation of models from standardized metrics
6. *Deployment.* It includes the incorporation of data mining models into production systems.

3. METHODOLOGY

3.1 Data Understanding and Preprocessing

There are 3026 entries of data obtained from various heart disease tests. However, there are a lot of missing values in the dataset. The final data used in this study is 1026 entries (Figure 1). Table 2 shows a description of each of the attributes used in this study.



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
60	1	2	140	185	0	0	155	0	3	1	0	2	0
67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1
63	0	2	135	252	0	0	172	0	0	2	0	2	1
42	0	2	120	209	0	1	173	0	0	1	0	2	1
61	0	0	145	307	0	0	146	1	1	1	0	3	0
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
58	0	1	136	319	1	0	152	0	0	2	2	2	0
56	1	2	130	256	1	0	142	1	0.6	1	1	1	0

Figure 1: Dataset Contents

3.2 Modelling

The algorithm models used in this study are KNN, Decision Tree, and Naïve Bayes. The stages of design and testing of the model are performed using RapidMiner v9.6. Overall, the research stages depicted in is divided into the following processes:

1. *Validation*

Validation is done to validate the performance of the algorithm used in a model. The validation process in this study was carried out with the cross-validation technique. The datasets were divided into two parts: training and testing. This research used the 10-fold cross-validation technique so that 90% of entries from the dataset are to be used as the training dataset, and the remaining 10% is being used as the testing dataset. The process is repeated until ten times until all data entries have been part of testing data. Ten-fold cross-validation was used because it is proven to produce better algorithm performance.

2. *Algorithm Accuracy Measurement*

Algorithm accuracy measurement is done to prove the performance of the algorithm applied to the used dataset. At this stage, the confusion matrix is used as the performance measurement tool for the classification algorithm accuracy. The calculations are made based on the comparison of the dataset with the classification results following the actual data with the total amount of data. Confusion Matrix is a review of the classification of data mining that is represented in tabular form.

The confusion matrix consists of comparison information between the classification label and the actual label. The result of this confusion matrix is to be a level of accuracy. This accuracy level is used as the reference to measure the performance of the classification algorithms tested in this study.

Table 2: Attributes Used in This Research

Attribute	Description	Role
Age	Age in year	integer
Sex	1: Man	binominal
	0: Woman	
CP	Type of chest pains	polynomial
	0: Typical angina	
	1: Atypical angina	
	2: Non-anginal pain	
	3: Asymptomatic pain	
oldpeak	ST depression induced by exercise relative to rest	real
trestbps	Resting blood pressure (mmHg)	integer
fbs	Fasting blood sugar > 120 mg/dl	binominal
	1: True	
	0: False	
restecg	Resting electrocardiographic results	polynomial
	0: Normal	
	1: Having ST-T wave abnormality	
	2: Ventricular hypertrophy	
exang	Exercise-induced angina	binominal
	1: True	
	0: False	
thalach	Maximum heart rate achieved	integer
fbs	Fasting blood sugar > 120 mg/dl	binominal
	1: True	
	0: False	
slope	The slope of the peak exercise ST segment	polynomial
	1: Upsloping	
	2: Flat	
	3: Downsloping	
ca	Number of major vessels (0-3) colored by flourosopy	polynomial

chol	Serum cholestoral (mg/dl)	integer
thal	0: Normal	polynomial
	1: Fixed defect	
	2: Reversible defect	
target	0: Have heart disease	binominal
	1: Have a healthy heart	

In the confusion matrix, TP stands for true positive, interpreted as what we predicted is positive, and in reality, it is true. TN stands for true negative, interpreted as what we predicted is negative, and in reality, it is true. FP stands for false positive, interpreted as what we predicted is positive, and in reality, it is negative, it means the prediction is false. FN stands for false negative, interpreted as what we predicted is negative, and in reality, it is positive. From the obtained confusion matrix, then we can determine the levels of accuracy, precision, and sensitivity as follows:

1. *Precision*

The measure of closeness between a set of analytical results obtained based on several measurements of the same homogeneous sample. Precision can be measured using the equation (1).

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

2. *Accuracy*

A measurement result is only accurate if the average value of the measurement results is close to or almost equal to the correct value. The accuracy value of a model can be calculated as (2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

3. *Sensitivity*

The capability of measuring instruments in response to changes in measurement values. If the sensitivity value of a model is high, then it can be concluded that the FN of a model is low. The calculation of the sensitivity of a model is as equation (3)

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

Besides, the value of the AUC-ROC of each classification algorithm model is also be reviewed. According [11], AUC describes the capability of a model to classify between classes by explaining the overall measurement of the suitability of the model used. The closer the AUC value is to 1, the better the model is.

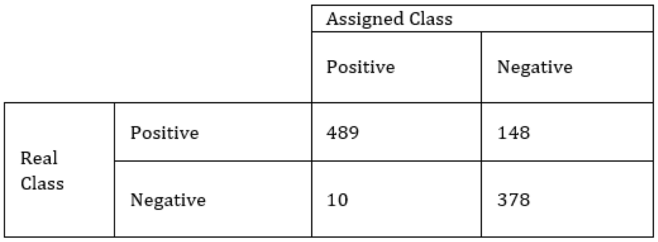
4. RESULTS AND DISCUSSION

4.1 Results

1. *K-Nearest Neighbor*

Table 3 shows the results obtained based on the data processing using the KNN classification model. Based on the confusion matrix, it may be concluded that 489 patients are correctly predicted to suffer from heart disease (TP), then 148 patients are predicted having healthy heart but suffer from heart disease (FP), ten patients are predicted to have heart disease but actually have a healthy heart (FN), and 378 patients were correctly predicted to have a healthy heart (TN). Based on these results, we get a TP rate (precision) of 0.767, an accuracy of 0.846 and a sensitivity of 0.979. The accuracy of KNN with the used k value can be seen in Table 4.

Table 3: Confusion Matrix from KNN Model



		Assigned Class	
		Positive	Negative
Real Class	Positive	489	148
	Negative	10	378

Table 4: KNN Accuracy from Various K Value

k	3	5	7	9	11	13	15
Accuracy	0,977	0,845	0,779	0,751	0,723	0,708	0,725
Precision	0,955	0,767	0,738	0,75	0,715	0,702	0,72
Sensitivity	1	0,979	0,847	0,733	0,717	0,697	0,711
AUC	0,847	0,955	0,877	0,845	0,846	0,837	0,833

The value of k = 5 is selected because although the accuracy obtained is lower than k = 3, it should be noted that the AUC value obtained when the value of k = 5 is the highest compared to when the value of k is greater than 5. AUC is a better measurement compared to accuracy [12]. At the value of k = 5, the best precision and sensitivity results are obtained compared to the other k values. The calculation is visualized into the ROC curve, as shown in Figure 2, which produces an AUC value of 0.955.

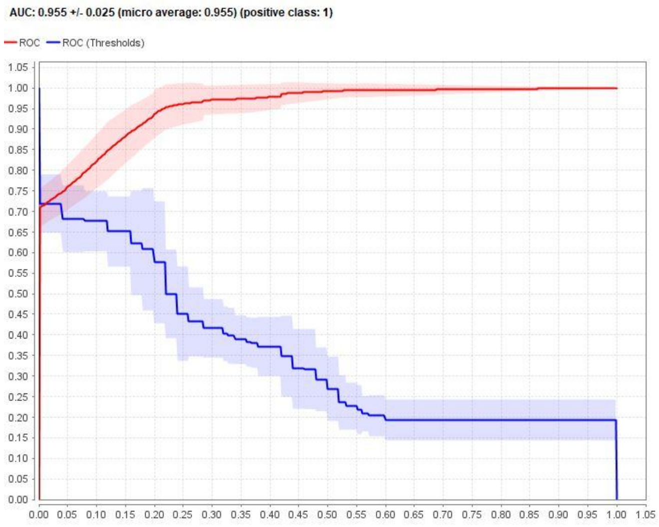


Figure 2: ROC Curve from KNN (k=5)

2. *Naïve Bayes*

Table 5 shows the result obtained based on the data processing using the Naive Bayes classification model. Based on the confusion matrix, it can be concluded 412 patients are correctly predicted to suffer from heart disease (TP), then 67 patients are predicted to have a healthy heart but suffer from heart disease (FP), 87 patients are predicted to have heart disease but actually have a healthy heart (FN), and 459 patients were correctly predicted to have a healthy heart (TN). Based on these results, we get a TP rate (precision) of 0.860, an accuracy of 0.8497 and a sensitivity of 0.82. The calculation is visualized into the ROC curve in Figure 3, which results in an AUC value of 0.919.

Table 5: Confusion Matrix from Naïve Bayes Model

		Assigned Class	
		Positive	Negative
Real Class	Positive	412	67
	Negative	87	459

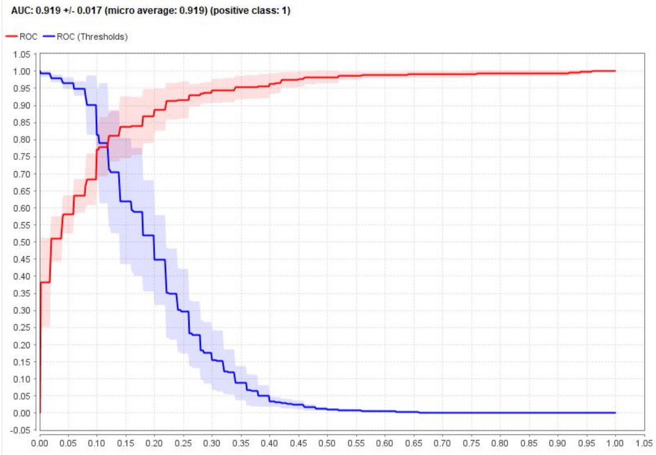


Figure 3: ROC Curve from Naïve Bayes

3. *Decision Tree*

Table 6 shows the result obtained based on the data processing using the Decision Tree classification model. The confusion matrix shows that 417 patients are correctly predicted to suffer from heart disease (TP), then 74 patients are predicted to have a healthy heart but suffer from heart disease (FP), 82 patients are predicted to have heart disease but actually have a healthy heart (FN), and 452 patients were correctly predicted to have a healthy heart (TN). Based on these results, we get a TP rate (precision) of 0.849, an accuracy of 0.8478 and a sensitivity of 0.835. Meanwhile, the calculation is visualized into the ROC curve, which results in an AUC value of 0.912, as shown in Figure 4.

Table 6: Confusion Matrix from the Decision Tree Model

		Assigned Class	
		Positive	Negative
Real Class	Positive	417	74
	Negative	82	452

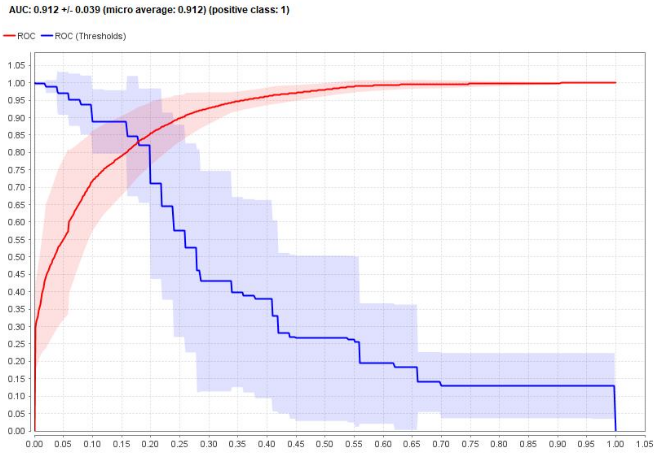


Figure 4: ROC Curve from Decision Tree

4.2 Weight of Each Attribute

The information gain is calculated performed to determine the weights of each attribute (Figure 3). The higher the weight value of an attribute, the more relevant the variable is considered. Table 3 shows the

Table 3: Attributes Weight

Atribut	Weight
Age	0,061
Sex	0,058
CP	0,208
trestbps	0,016
chol	0,02
fbs	0,001
restecg	0,026
thalach	0,128
exang	0,145
oldpeak	0,131
slope	0,113
ca	0,193
thal	0,208

Based on the results of the weights of each attribute written in Table 3, it can be concluded that the thal and CP attributes with a weight of 0.208 are the most influential attributes on the prediction results of patients suspected of having heart disease. Then, the fbs attribute is the attribute that has the smallest weight value, which is 0.001. It can be concluded that the fbs attribute at least affects the classification of patients suspected of having heart disease. Furthermore, the weight of variables and their effect on the decision tree model is shown in Figure 5.

4.3 Discussion

1. *Precision/TP Rate*

Given the outcomes, the most noteworthy precision esteem is acquired from the Naïve Bayes model (0.860). It very well may be presumed that contrasted with different models, for example, KNN or Decision Tree, the Naïve Bayes model is the best for classifying patients with heart illness or not.

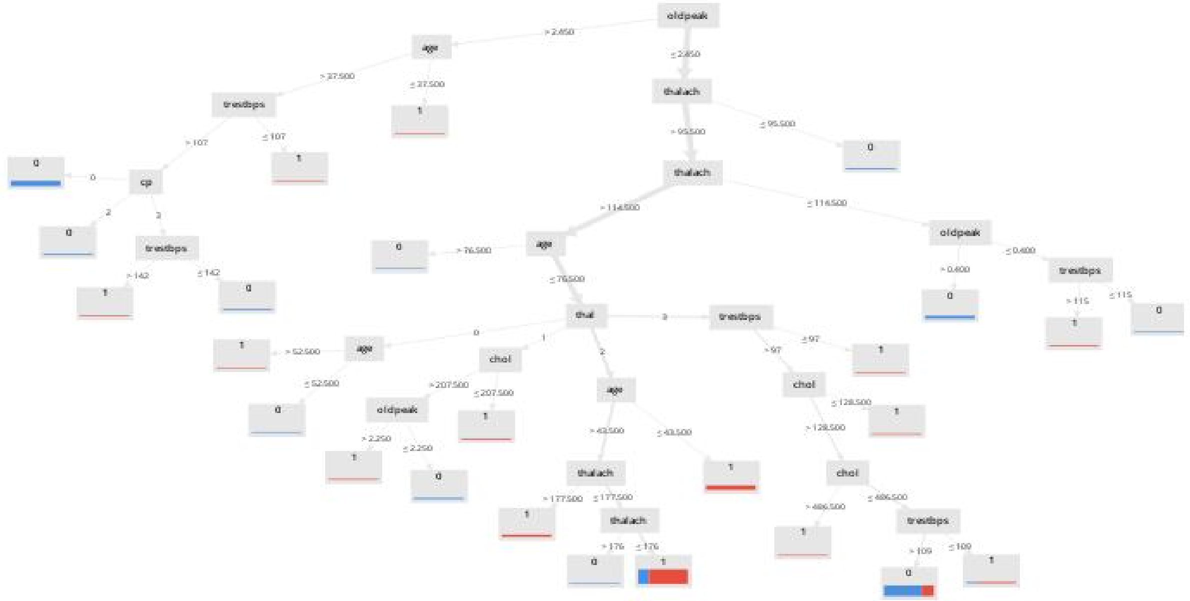


Figure 5: Decision Tree Model Result

2. *FP Rate*

Based on the results, the lowest FP rate is obtained from the Naïve Bayes model, and the highest FP rate is obtained from the KNN model. It can be concluded that KNN has the most misclassified number of patients suffering from heart disease among the other algorithm models. In contrast, Naïve Bayes has the lowest FP rate, which means that the Naïve Bayes model has the fewest misclassified number of patients suffering from heart disease.

3. *Sensitivity (Recall)*

Based on the results, the highest sensitivity value is obtained from the KNN model (0.979). It can be concluded that the KNN model is well able to show which patients have heart disease from the entire population of who suffers from heart disease.

4. *Accuracy*

The highest level of accuracy is obtained from the Naïve Bayes model (0.850). It can be concluded that the Naïve Bayes Model can make negative and positive predictions better than the overall data compared to the KNN and Decision Tree models.

5. *AUC*

Based on the results, the highest AUC value is obtained from the KNN model. It can be concluded that the KNN model is the most accurate in identifying the presence or absence of heart disease or differentiating between heart disease sufferers and non-heart disease sufferers.

6. *Specificity*

Based on the results, the highest specificity value is obtained from the Naïve Bayes model (0.872). It can be finalized that the Naïve Bayes model is well able to predict negatively from the overall harmful data, compared to the KNN and Decision Tree models.

Based on several indicators used to evaluate the results of the classification algorithm models in Table 8, the precision, specificity, and accuracy are the most critical indicators. The accuracy of classifying patients who are suffering from heart disease and patients with a healthy heart is critical. The model with the most occurrence of TP is also crucial because it means that the model can correctly classify patients suffering from heart disease.

The model with high FP value (represented by specificity) should be avoided since the unnecessary treatment is given to patients who are predicted having heart disease, and patients who are predicted having heart disease can be done correctly. Therefore, the best model in the case of this study is Naïve Bayes, which has better values of precision, accuracy, and specificity compared to other models. However, it should be noted that, based on the result in Table 8 and Figure 6, if misclassification of patients suffering from heart disease and vice versa is more tolerated (tolerates FP more than FN), then the recall value generated from the KNN model is more representative than the Naïve Bayes and the Decision Tree model.

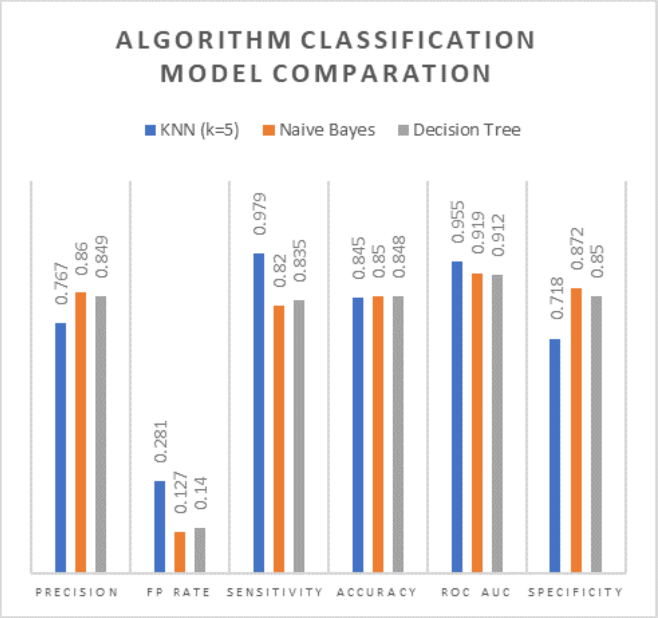


Figure 6: Algorithm Classification Model Comparisons

Table 8: Algorithm Classification Model Comparisons

	Algorithm Model		
	KNN (k=5)	**Naive Bayes**	**Decision Tree**
Precision/TP Rate	0,767	0,860	0,849
FP Rate	0,281	0,127	0,140
Sensitivity/Recall	0,979	0,82	0,835
Accuracy	0,845	0,85	0,848
AUC	0,955	0,919	0,912
Specificity	0,718	0,872	0,85

Table 9: Comparison with Previous Works

Similar Study	**Model**	**Tools**	**Accuracy**
[13]	Naïve Bayes	WEKA	0,835
[14]	Naïve Bayes	Matlab R2014a	0,833
[4]	Naïve Bayes	WEKA	0,835
Our	Naïve Bayes	Rapidminer 9.6	0,860

This study also compares the accuracy obtained as a result of the Naïve Bayes model with previous literature that used the same dataset, as shown in Table 9. The best result was obtained by this study, using the RapidMiner 9.6 tools, even though other studies also use the Naïve Bayes model with 10-fold cross-validation. Previous studies also confirmed this finding using several different tools with the Naïve Bayes model. The Naïve Bayes accuracy is best obtained by using RapidMiner tools compared to WEKA in [15] and [16].

5. CONCLUSION

In this study, data mining modeling has been conducted to classify patients with heart disease and with a healthy heart. Different data classification models such as Naïve Bayes, KNN, and Decision Tree are utilized with the assistance of RapidMiner software. Based on the results in this study and the indicators such as precision, accuracy, and specificity, it can be concluded that the Naïve Bayes model is the best for the classification of patients with and without heart disease. The best model in the case of this study is Naïve Bayes, which has better balance values of precision, accuracy, and specificity compared to other models. The decision of the best model to be applied in this case is expected to help predict the possibility of cardiovascular disease in patients early on so that treatment can be done on time.

REFERENCES

1. Jabbar, A, Deekshatulu, B. L, & Chandra, P. (2013). **Heart Disease Classification Using Nearest Neighbor Classifier with Feature Subset Selection**. *Annals. Computer Science Series.*
2. Canlas, R D (2009). **Data mining in healthcare: Current applications and issues**. *School of Information*

Systems & Management, Carnegie Mellon University, Australia.

3. Venkatalakshmi, B, & Shivsankar, M V (2014). **Heart Disease Diagnosis Using Predictive Data Mining**. *International Journal of Innovative Research in Science, Engineering and Technology*.
4. Mutyala, N. K, Koushik, K. V, & Krishna, K. D. (2018). **Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools**. *International Journal of Scientific Research in Computer Science*.
5. Shouman, M, Turner, T, & Stocker, R (2011). **Using Decision Tree for Diagnosing Heart Disease Patients**. In *Proceedings of the Ninth Australasian Data Mining Conference*.
6. Koushik, K. V, Kumar, M. N, & Deepak, K. (2018). **Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools.** *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*.
7. Khateeb, N, & Usman, M (2017). **Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique**. *Proceedings of the International Conference on Big Data and Internet of Thing*. https://doi.org/10.1145/3175684.3175703
8. Nikhar, S, & Karandikar, A. M (2016). **Prediction of Heart Disease Using Machine Learning Algorithms**. *International Journal of Advanced Engineering, Management and Science*.
9. Christ, Mogi & Rahmanto, Nikolaus. (2019). **Lending Club Default Prediction using Naïve Bayes and Decision Tree.** *International Journal of Advanced Trends in Computer Science and Engineering*. 8. 2528-2534. 10.30534/ijatcse/2019/99852019.
10. Mauritsius, T, Braza, A, & Fransisca (2019). **Bank Marketing Data Mining using CRISP-DM Approach**. *International Journal Of Advanced Trends In Computer Science And Engineering*, 8(5), 2322-2329. doi: 10.30534/ijatcse/2019/71852019.
11. Fawcett, T, & Provost, F (2013). ***Data Science for Business***. O'Reilly Media, Inc.
12. Ling, C, Huang, J, & Zhang, H. (2003). **AUC: A Better Measure than Accuracy in Comparing Learning Algorithms**. *Proceedings of The 16th Canadian Society for Computational Studies of Intelligence Conference on Advances in Artificial Intelligence*. https://doi.org/10.1007/3-540-44886-1_25
13. Chaki, D, Moinul, Z, & Das, A (2015). **A Comparison of Three Discrete Methods for Classification of Heart Disease Data.** *Bangladesh Journal of Scientific and Industrial Research*. https://doi.org/10.3329/bjsir.v50i4.25839
14. Liu, X, Wang, X., Su, Q, Zhang, M., Zhu, Y, Wang, Q, & Wang, Q (2016). **A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method**. *Computational and Mathematical Methods in Cardiovascular Diseases*. https://doi.org/10.1155/2017/8272091
15. Naik, A, & Samant, L (2016). **Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime**. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2016.05.251
16. Sharma, A, Sharma, D, & Mansotra, V (2017). **Performance Analysis of Data Mining Techniques on Lifestyle Diseases**. *International Journal for Research in Applied Science & Engineering Technology*.