

Comparative Analysis of Naïve Bayes Algorithm, K-Nearest Neighbours, Neural Network, Decision Tree, and Random Forest to Classify Data Mining in Predicting Heart Disease

Laisa Nurin Mentari¹, Fenni Agustina²

¹Master of Information Systems Management, Business Information System, Gunadarma University, Indonesia

²Faculty of Technology and Engineering, Gunadarma University, Depok, West Java, Indonesia-16424

Email: ¹laisanurin@gmail.com, ²fenni@staff.gunadarma.ac.id

Abstract— Heart disease is caused by unhealthy lifestyle patterns, smoking, unhealthy food, lack of exercise, obesity, and many more that trigger heart attacks. According to WHO data (2011), ischemic heart disease and stroke are the first and second leading causes of death worldwide. Heart disease is a proven leading cause of death, which is why accurate and precise prediction of heart disease is so important. This study compares five classification algorithms consisting of the Naïve Bayes Algorithm, K-Nearest Neighbors, Neural Network, Decision Tree, and Random Forest using RapidMiner 5 software. The results of this study show that the algorithm that shows the highest accuracy is the Naïve Bayes Algorithm. with an accuracy rate of 82.23%, the accuracy of the Neural Network Algorithm is 78.29%, the accuracy of the Decision Tree Algorithm is 74.30%, the Random Forest Algorithm is 76.62% and the accuracy of the K-Nearest Neighbors Algorithm is 62.69%. The methods that are not recommended to be used are K-Nearest Neighbors and Decision Tree because they have an AUC (Area Under Curve) value of only 0.683% which indicates that the predictor has a Poor Classification.

Keywords— Heart Disease, Prediction, Classification, Naïve Bayes Algorithm, K-Nearest Neighbor, Neural Network, Decision Tree, and Random Forest.

I. INTRODUCTION

Advances in science and technology influence change in people's behavior and lifestyle today. The tendency to live more easily and instantly often harms health. Physical activity, such as walking, becomes reluctant to do, because of the availability of transportation that facilitates mobility. Sports also become rarely done, because of the demands of busy routines. Consumption of fast food that is high in fat and low in fiber has become a daily culture. No wonder, if there is a shift in the causes of death and infectious diseases to metabolic and degenerative diseases, including cardiovascular disease. (Andarini, Titim Andri Wihastuti and Teuku Heriansyah, 2016).

Cardiovascular disease (CVD) ranks as the first cause of death in the world more than any other disease. Cardiovascular disease is a disease that is not contagious but most often causes death. According to WHO data (2011), ischemic heart disease and stroke are the first and second leading causes of death worldwide with a percentage of 12.9% and 11.4% with a total death caused of 13.2 million people.

The cardiovascular disease tends to increase from year to year. It is estimated that by 2020, cardiovascular disease will cause more than 23 million deaths per year (WHO 2013). (Andarini, Titim Andri Wihastuti and Teuku Heriansyah, 2016).

The high mortality factor caused by heart disease is due to the lack of understanding of the Indonesian people about the symptoms of heart disease caused. Therefore, it is necessary to take a step from now on to treat and prevent heart disease. This can be avoided by utilizing published patient data and then making patterns in determining heart disease using data mining so that people can find out what factors cause heart disease.

Therefore, with the classification of data mining in predicting heart disease, it can provide decisions in solving problems that can be directly addressed and provide information, as well as data that is easier and faster. In addition, it is hoped that the classification of data mining will be more effective and efficient in the process.

Based on this background, the authors aim to explore, analyze and compare heart disease by applying data mining. Here the author performs a classification technique using the Naïve Bayes Algorithm, K-Nearest Neighbors, Neural Network, Decision Tree, and Random Forest as heart disease dataset processing using RapidMiner 5 software so that it can produce accurate and precise results in analyzing heart disease datasets which later can be seen data on who have heart disease.

II. LITERATUR REVIEW

A. Analysis

The analysis is an activity in studying and evaluating a form of problem or case that occurs. (Indyah Hartami Santi, 2020).

B. Compare

In statistics, the comparative test of both independent samples and dependent samples is categorized as comparative statistics. Etymologically, comparison comes from the word compare which means "compare", comparability means "the nature can be compared or equated, comparative means that which is related to comparison, while comparison means

comparison or comparison. Comparison in the language is to compare or comparison. Comparison in the language is to compare or comparison. Yusri (2009) added that the comparative test is to find out or compare whether there is a difference between the two research samples (variables). Sugiono (2010) explains that testing comparative hypotheses means-testing population parameters in the form of comparisons through sample sizes which are also in the form of comparisons where the comparisons are made from two or more samples. (H. Fajri Ismail. 2018).

C. Data Mining

Data mining is also a logical process to find useful information. After finding information and patterns, it can be used as a supporting tool in decision-making in developing a business. Data mining tools can provide answers to a wide range of business-related questions and are too difficult to solve. Data mining can also be used to forecast future trends that enable businesses to make effective, proactive, and dynamic decisions. The data that is processed using data mining techniques is also able to produce knowledge that is in line with expectations. For example in the health sector, there is quite a lot of data owned by hospitals, such as medical records and radiology data, but because there is no standard data collection, the data is difficult to process, so with the presence of data mining, it is expected that the data that owned by the health sector can be processed according to needs, to produce information and knowledge that can be utilized by policymakers, especially the government. (Muhammad Arhami and Muhammad Nasir. 2020).

D. Prediction

In the Big Indonesian Dictionary (KBBI) the notion of prediction is a forecast or estimate. Prediction according to the Big Indonesian Dictionary (KBBI) is an activity to predict what will happen. Some other definitions of prediction: Some other definitions of prediction, namely:

1. Prediction is defined as the use of statistical techniques in the form of a future picture based on the processing of historical figures.
2. Prediction is an internal part of management decision-making activities.
3. Forecasting is an activity that predicts what will happen in the future. The problem of decision-making is a problem that is often faced, so forecasting is also a problem that must be faced because forecasting is also closely related to making a decision. (Cahyo Prianto, Emma Ainun Novia, and Woro Isti Rahayu. 2020).

E. Classification

Classification is a process to find modals or functions that explain or distinguish concepts or data classes, to be able to estimate the class of an object whose label is unknown. Other classification methods are Bayesian, Neural Network, Genetic Algorithm, Fuzzy Case-Based Reasoning, K-Nearest Neighbors. Classification is a set of records (training set). Each record includes a set of attributes, one of which is a class. The model for class attributes is a function of the values of other attributes. A test set is used to determine the accuracy

of the model. Usually, a given data set is divided into training and test sets, where the training set is used to build the model and the test set is used to validate. (Cahyo Prianto, Emma Ainun Novia, and Woro Isti Rahayu. 2020).

F. Algorithm

The algorithm is a logical and systematic arrangement to solve a problem or to achieve a certain goal. In the world of information, algorithms play an important role in the development and construction of software. But in everyday life without you knowing the algorithm has entered into your life. (Noviana Riza, Rd. Nuraini Siti Fathonah, and Yusniar Nur Syarif Sidiq. 2020).

G. Naïve Bayes

Naïve Bayes algorithm can be used for binary and multiclass classification problems. Naïve Bayes builds and scores models very quickly and scales linearly in a number of predictions and rows. Naïve Bayes is also a classification that presents each object class based on a probabilistic conclusion or recapitulation and finds the most likely class that is appropriate for each object whose class will be determined from existing test objects based on attributes or variables whose values are known. (Muhammad Arhami and Muhammad Nasir. 2020).

H. K-Nearest Neighbours (K-NN)

K-Nearest Neighbors (K-NN) has been used in statistical estimation and pattern recognition since 1970 when K-Nearest Neighbors (K-NN) was also a non-parametric technique. K-Nearest Neighbors (K-NN) is one of the algorithms or methods for classification commonly used in data mining. K-Nearest Neighbors (K-NN) is also included in the category of regression which can also be used to predict as well as regression. (Muhammad Arhami and Muhammad Nasir. 2020).

I. Decision Tree

A decision Tree or decision tree is a model that maps observations of an item so that a conclusion is obtained about the target value of an item described in the form of a model tree. The model tree has a tree structure where the leaves represent the classification and the branches represent the relationship between the attribute values that lead to the classification. (Muhammad Arhami and Muhammad Nasir. 2020).

J. Neural Network

Neural Network or what is known as an artificial neural network (ANN) is an information processing system that has characteristics similar to biological neural networks and consists of a large number of simple processing elements called neurons, units, cells, or nodes. Each neuron is connected to other neurons using directional communication links, each with an associated weight. The weights are the information used by the network to solve problems. (Eddy Irwansyah and Muhammad Faisal. 2015).

K. Random Forest

This method can be regarded as a panacea for data science problems. Random Forest is a versatile machine learning method. This method can work well to solve regression and classification problems. In addition, it can also reduce dimensions, overcome missing values, outlier values and be able to explore other important steps. This method belongs to the category of ensemble learning where a set of weak models is combined to produce a powerful model. (Imam Tahyudin, et al. 2020).

L. RapidMiner

RapidMiner is a data science software platform developed by the same renowned company that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial, as well as for research, education, training, rapid prototyping, and application development, and supports all steps in the machine learning process including data preparation, visualization results, model validation, and optimization. RapidMiner is developed on an open core model, with RapidMiner studio Free Edition, which is limited to 1 logic processor and 10,000 rows of data, available under the APGL license. Commercial pricing starts at \$2,500 and is available from the developer. (Abdul Khamid, Agyztia Premana, and Nur Ariesanto Ramdhan. 2020).

III. RESEARCH METHOD

In this study, several stages were carried out, including data collection (datasets), data processing (data preprocessing), the classification process of the Naive Bayes algorithm, K-Nearest Neighbors, Neural Networks, Decision Trees and Random Forests, testing methods using k-fold cross-validation, evaluation using a confusion matrix and ROC curve and comparison, as shown in Figure 1 Research Stages.

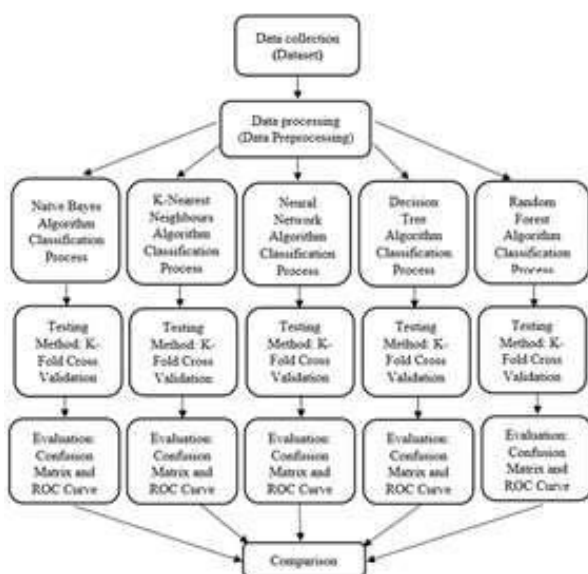


Fig. 1. Research Stage

A. Data Collection (Dataset)

The first stage is data collection. The data collection used by the researcher to collect the data came from the following website: <https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>. This heart disease data involved 303 samples with 14 attributes.

B. Data Processing (Data Preprocessing)

The second stage is data processing (data preprocessing). Data preprocessing is focused mainly on two problems, namely the data must be organized in the right form for the data mining algorithm and the data sets used to lead to the performance and quality of the model obtained from the data-mining operation. The data that will be used as a dataset in this research are training data and testing data. By separating the training data and testing data, it is intended that the model obtained can have a good ability to classify data.

C. Algorithm Classification Process

The third stage is the algorithm classification process. In the algorithm classification process, data mining is carried out by classifying the Naïve Bayes algorithm, K-Nearest Neighbors, Neural Network, Decision Tree, and Random Forest to predict heart disease. By comparing the five algorithms, it is expected to be able to handle a dataset with accurate and precise results.

D. Testing Method

The fourth stage is method testing. In testing the method, it is done by knowing the results of the calculations that have been analyzed and measuring the methods and algorithms used whether they can function properly or not. In the testing process using RapidMiner 5 software and the k-fold cross-validation method which can be expected to obtain accurate results in predicting heart disease.

E. Evaluation

The fifth stage is the evaluation of the test results. After the dataset is processed using RapidMiner 5 software, then the accuracy results are obtained from each Nave Bayes algorithm, K-Nearest Neighbors, Neural Network, Decision Tree, and Random Forest. After that, generate confusion matrix and ROC (Receiver Operating Characteristics).

F. Comparison

The sixth stage is the comparison of algorithms, at this stage, the author will compare which of the five algorithms is the best in predicting heart disease.

IV. RESULT AND DISCUSSION

Based on the results and discussion of the heart disease dataset, it can be seen, as follows:

A. Data Collection (Dataset)

The data used in this study is secondary data obtained through the following website: <https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>, as shown in Figure 2.



Fig. 2. Kaggle Website View

If the dataset from the Kaggle website has been downloaded, the heart disease dataset involves 303 samples with 14 attributes. The following is a display of the results of the download of the heart disease dataset from the Kaggle website, as shown in Figure 3.

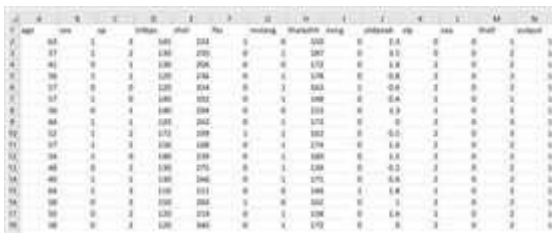


Fig. 3. Display of Downloaded Dataset Results

Before the heart disease dataset is processed using RapidMiner 5 software, the data is processed first, as shown in Figure 4.



Fig. 4. Display of Dataset Processing Results

B. Data Processing (Data Preprocessing)

The data is entered into the RapidMiner 5 software, then data preprocessing is done, this is done because the data that is available and will be used is not necessarily good due to Incompleteness, Noise, and Inconsistency. The following is data processing that has been entered into the RapidMiner 5 software, as follows:




Fig 5. Display Data Processing

To solve this problem, it can be done in RapidMiner 5 software by returning to the design menu, selecting the operator, typing replace missing value then dragging it into the process than changing the default to average, and connecting to the result on the right. After that click the Run button, as follows:



Fig. 6. Process Display to Overcome Missing Attributes

The results obtained that the missing attribute can be overcome, as follows:

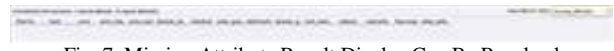


Fig. 7. Missing Attribute Result Display Can Be Resolved

C. Algorithm Classification Process

After data processing is done on the RapidMiner software, the next step is the application of an algorithm consisting of the Naïve Bayes Algorithm, K-Nearest Neighbors, Neural Network, Decision Tree, and Random Forest.

1. Naïve Bayes Algorithm

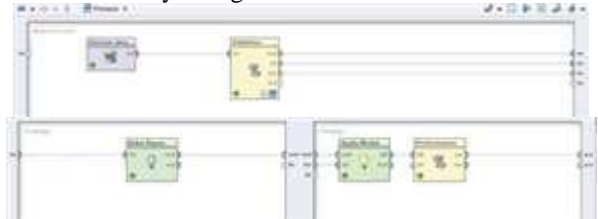


Fig. 8. Naïve Bayes Algorithm Model

2. K-Nearest Neighbors Algorithm

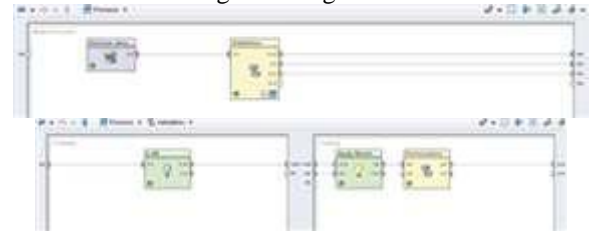


Fig 9. K-Nearest Neighbors Algorithm Model

3. Neural Network Algorithm

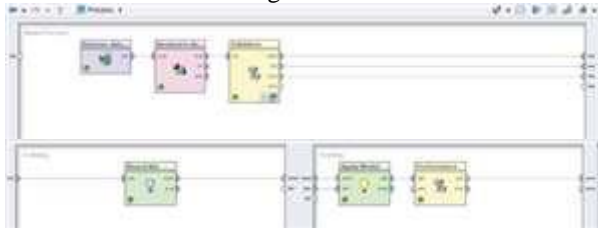


Fig. 10. Neural Network Algorithm Model

4. Decision Tree Algorithm

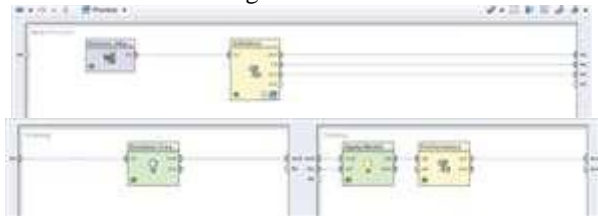


Fig. 11. Decision Tree Algorithm Model

5. Random Forest Algorithm



Fig. 12. Random Forest Algorithm Model

D. Testing Method

Method testing was carried out using k-fold cross-validation and evaluation of test results using a confusion matrix and ROC (Receiver Operating Characteristic Curve) curve and the area under the ROC curve (AUC).

1. Confussion Matrix

Confusion Matrix is used to determine the performance value of the classification algorithm modeling based on the calculation results of the accuracy value obtained from the number of correct case predictions and the number of incorrect case predictions. (Pareza Alam Jusia. 2018). The following are prediction rules, as shown in Table I.

TABLE I.

		Actual	
		True Positive (TP)	False Negative (FP)
Predicted	True	True Positive (TP)	False Negative (FP)
	False	False Negative (FN)	True Negative (TN)

(Source: Riski Annisa, 2019)

The calculation of accuracy using the confusion matrix method, as in the formula (4.1., 4.2., 4.3., and 4.4).

Accuracy describes how accurate the correctly classified models are. (Narkhede's Nest. 2018).

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100\% \tag{4.1}$$

Precision describes the accuracy between the requested data and the prediction results provided by the model.

$$Presisi = \frac{TP}{FP+TP} \cdot 100\% \tag{4.2}$$

Recall or Sensitivity describes the success of the model in retrieving information.

$$Recall = \frac{TP}{FN+TP} \cdot 100\% \tag{4.3}$$

a. Naive Bayes algorithm

The following is the Accuracy Confusion Matrix value from testing the Naive Bayes algorithm as shown in Figure 13.

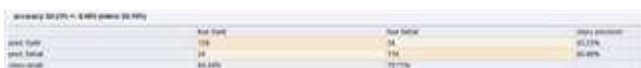


Fig. 13. Naive Bayes Algorithm Accuracy Display

Based on Figure 13, the accuracy level using the Naive Bayes Algorithm is 82.23%.

The following is the Precision Confusion Matrix value from testing the Naive Bayes Algorithm, as shown in Figure 14.



Fig. 14. Naive Bayes Algorithm Precision Display

Based on Figure 14, the precision level using the Naive Bayes algorithm is 81.12%.

The following is the Recall Confusion Matrix value from testing the Naive Bayes Algorithm, as shown in Figure 15.

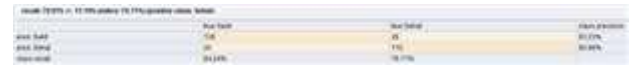


Fig. 15. Naive Bayes Algorithm Recall Display

Based on Figure 15, the recall rate using the Naive Bayes algorithm is 79.91%.

b. K-Nearest Neighbors Algorithm

The following is the Accuracy Confusion Matrix value from testing the K-Nearest Neighbors Algorithm, as shown in Figure 16.

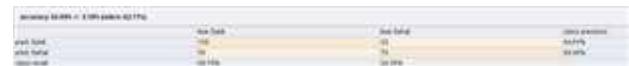


Fig. 16. K-Nearest Neighbors Algorithm Accuracy Display

Based on Figure 16, the level of accuracy using the K-Nearest Neighbors algorithm is 62.69%.

The following is the Precision Confusion Matrix value from testing the K-Nearest Neighbors Algorithm, as shown in Figure 17.



Fig. 17. Display of the Precision K-Nearest Neighbors Algorithm

Based on Figure 17, the level of precision that uses the K-Nearest Neighbors algorithm is 60.57%.

The following is the Recall Confusion Matrix value from testing the K-Nearest Neighbors Algorithm, as shown in Figure 18.

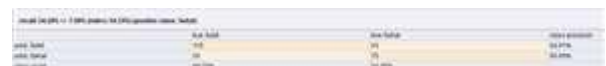


Fig. 18. K-Nearest Neighbors Algorithm Recall Display

Based on Figure 18, the recall rate using the K-Nearest Neighbors algorithm is 54.28%.

c. Neural Network Algorithm

The following is the Accuracy Confusion Matrix value from testing the Neural Network Algorithm, as shown in Figure 19.



Fig. 19. Display of Neural Network Algorithm Accuracy

Based on Figure 19, the level of accuracy using the Neural Network Algorithm is 78.29%.

The following is the Precision Confusion Matrix value from testing the Neural Network Algorithm, as shown in Figure 20.



Fig. 20. Display of Precision Neural Network Algorithm

Based on Figure 20, the precision level using the Neural Network algorithm is 78.21%.

The following is the Recall Confusion Matrix value from testing the Neural Network Algorithm, as shown in Figure 21.

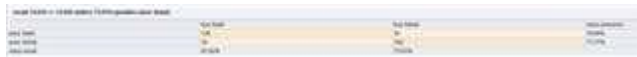


Fig. 21. Neural Network Algorithm Recall Display

Based on Figure 21, the recall rate using the Neural Network algorithm is 74.03%.

d. Decision Tree Algorithm

The following is the Accuracy Confusion Matrix value from testing the Decision Tree Algorithm, as shown in Figure 22.



Fig. 22. Display of Decision Tree Algorithm Accuracy

Based on Figure 22, the level of accuracy using the Decision Tree algorithm is 74.30%.

The following is the Precision Confusion Matrix value from testing the Decision Tree Algorithm, as shown in Figure 23.



Fig. 23. Display of Precision Algorithm Decision Tree

Based on Figure 23, the level of precision that uses the Decision Tree algorithm is 72.83%.

The following is the Recall Confusion Matrix value from testing the Decision Tree Algorithm, as shown in Figure 24.



Fig. 24. Recall Decision Tree Algorithm Display

Based on Figure 24, the recall rate using the Decision Tree algorithm is 70.40%.

e. Random Forest Algorithm

The following is the Accuracy Confusion Matrix value from testing the Random Forest Algorithm, as shown in Figure 25.



Fig. 25. Random Forest Algorithm Accuracy Display

Based on Figure 25, the accuracy level using the Random Forest algorithm is 76.62%.

The following is the Precision Confusion Matrix value from testing the Random Forest Algorithm, as shown in Figure 26.

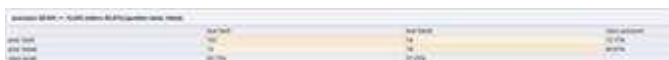


Fig. 26. Display of Random Forest Algorithm Precision

Based on Figure 26, the precision level using the Random Forest algorithm is 88.69%.

The following is the Recall Confusion Matrix value from testing the Random Forest Algorithm, as shown in Figure 27.

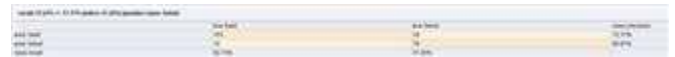


Fig. 27. Display of Random Forest Algorithm Recall Results

Based on Figure 27, the recall rate using the Random Forest algorithm is 57.67%.

E. Evaluation

The ROC (Receiver Operating Characteristic Curve) curve is used to obtain a cut-off point in a diagnostic test (cut off point) through a variable in numerical form, for example, to diagnose IDA (Iron Deficiency Anemia) in infants used by Wibisono (pediatrics), RDW (Red Blood Cell Distribution Width). (Julius H. Lolombulan, 2020).

ROC (Receiver Operating Characteristic Curve) curve analysis is feasible if the Area Under the Curve (AUC) value is > 0.7 . The Area Under the Curve (AUC) category consists of (Gorunescu, 2011), as shown in Figure 28.

No	Nilai AUC	Klasifikasi
1	0,90 - 1,00	Excellent
2	0,80 - 0,90	Good
3	0,70 - 0,80	Fair
4	0,60 - 0,70	Poor
5	0,50 - 0,60	Failure

Fig 28. AUC Value and Classification (Source: Gorunescu, 2011)

a. Naïve Bayes Algorithm

AUC (Optimistic)

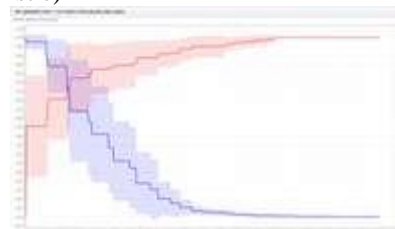


Fig. 29. Results of AUC (optimistic) Naive Bayes

AUC (Area Under the Curve)



Fig. 30. Results of AUC (Area Under the Curve) Naive Bayes

AUC (Pessimistic)

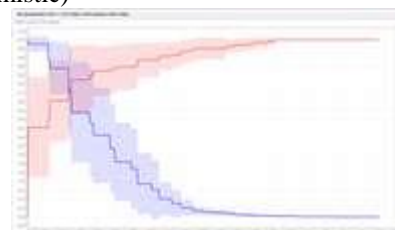


Fig. 31. Results of AUC (pessimistic) Naive Bayes

b. K-Nearest Neighbors Algorithm

AUC (Optimistic)



Fig. 32. Results of AUC (optimistic) K-Nearest Neighbors

AUC (Area Under the Curve)

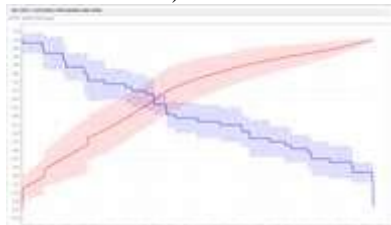


Fig. 33. Results of AUC (Area Under the Curve) K-Nearest Neighbors

AUC (Pessimistic)

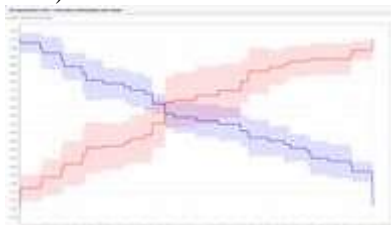


Fig. 34. Results of AUC (pessimistic) K-Nearest Neighbors

c. Neural Network Algorithm

AUC (Optimistic)



Fig. 35. Results of AUC (optimistic) Neural Network

AUC (Area Under the Curve)

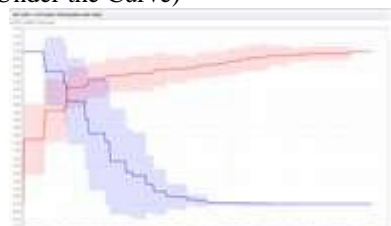


Fig. 36. Results of AUC (Area Under the Curve) Neural Network

AUC (Pessimistic)

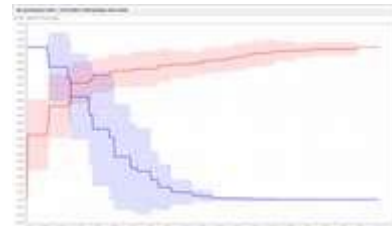


Fig. 37. Results of AUC (pessimistic) Neural Network

d. Decision Tree Algorithm

AUC (Optimistic)

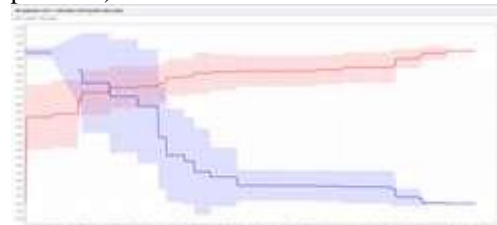


Fig. 38. Results of AUC (optimistic) Decision Tree

AUC (Area Under the Curve)

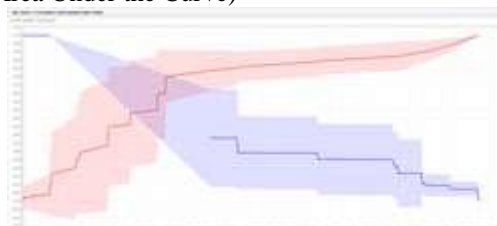


Fig. 39. Results of AUC (Area Under the Curve) Decision Tree

AUC (Pessimistic)

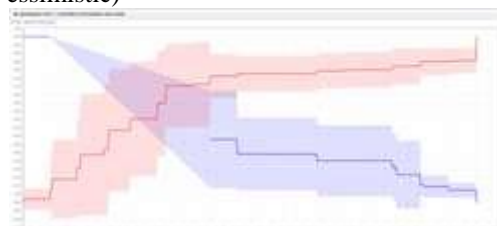


Fig. 40. Results of AUC (pessimistic) Decision Tree

e. Random Forest Algorithm

AUC (Optimistic)

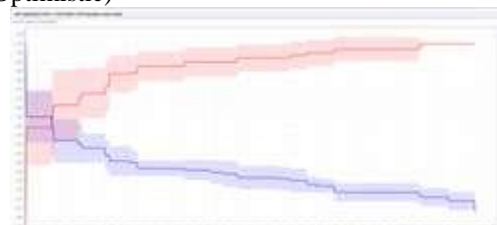


Fig. 41. Results of AUC (optimistic) Random Forest

AUC (Area Under the Curve)

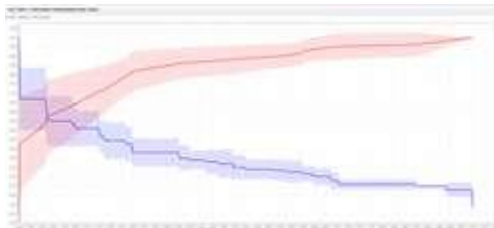


Fig. 42. Results of AUC (Area Under the Curve) Random Forest

AUC (Pessimistic)

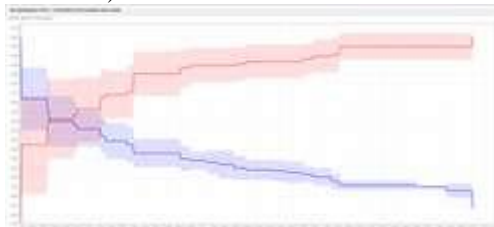


Fig. 43. Results of AUC (pessimistic) Random Forest

F. Comparison

Based on the results of tests carried out on five algorithms consisting of the Naïve Bayes algorithm, K-Nearest Neighbors, Neural Network, Decision Tree, and Random Forest, it can be concluded as in table II.

TABLE II.

	Naive Bayes	K-Nearest Neighbors	Neural Network	Decision Tree	Random Forest
Accuracy	82.23%	62.69%	78.29%	74.30%	76.62%
Precision	81.12%	60.57%	78.21%	72.83%	88.69%
Recall	79.91%	54.28%	74.03%	70.40%	57.67%
AUC (Optimistic)	0.891	0.772	0.863	0.831	0.879
AUC (Area Under the Curve)	0.891	0.683	0.863	0.683	0.846
AUC (Pessimistic)	0.891	0.594	0.863	0.633	0.813

V. CONCLUSION AND SUGGESTION

A. Conclusion

Thus the algorithm that shows the highest level of accuracy is the Naïve Bayes algorithm, which has an accuracy rate of 82.23% as the best method for predicting heart disease. Meanwhile, the AUC (area under the curve) values with good classification accuracy are Naïve Bayes, Neural Network, and Random Forest. Meanwhile, the AUC (area under the curve) value with poor classification accuracy is K-Nearest Neighbors (KNN) and the Decision Tree is 0.683.

B. Suggestion

There are several suggestions that the author can give for further development, which are as follows:

1. By using different classification methods, such as using the ID3 method, Decision Stump, Random Tree, Bayesian Boosting, AdaBoost, and so on.
2. Taking heart disease datasets that can be directly collected from various sources in the hospital, for example a special hospital for heart patients.
3. Using software other than RapidMiner in order to find

more accurate and precise results in predicting heart disease.

REFERENCES

- [1] Andarini, Titim Andri Wihastuti, Teuku Heriansyah, "Patofisiologi Dasar Keperawatan Penyakit Jantung Koroner : Inflamasi Vaskular," UB Press, 2016.
- [2] Indyah Hartami Santi, "Analisa Perancangan Sistem," PT. Nasya Expanding Management, 2020.
- [3] H. Fajri Ismail, "Statistika Untuk Penelitian Pendidikan dan Ilmu-Ilmu Sosial," Kencana, 2018.
- [4] Muhammad Arhami, Muhammad Nasir, "Data Mining Algoritma dan Implementasi," CV Andi Offset, 2020.
- [5] Noviana Riza, Rd. Nuraini Siti Fathonah, Yusniar Nur Syarif Sidiq, "Metode Klasifikasi Menentukan Kenaikan Level UKM Bandung," CV. Kreatif Industri Nusantara, 2020.
- [6] Cahyo Prianto, Emma Ainun Novia, Woro Isti Rahayu, "Sistem Perbandingan Algoritma K-Means dan Naïve Bayes Untuk Memprediksi Prioritas Pembayaran Tagihan Rumah Sakit Berdasarkan Tingkat Kepentingan," Kreatif Industri Nusantara, 2020.
- [7] Eddy Irwansyah, Muhammad Faisal, "Advanced Clustering Teori dan Aplikasi," Depublish, 2015.
- [8] Abdul Khamid, Agyzitia Premana, Nur Ariesanto Ramdhan, "Modul Belajar Data Mining dan RapidMiner," Penerbit Lakeisha, 2020.
- [9] Prima Dina Atika, I Wowon Priatna, "Modul Data Mining," Fakultas Ilmu Komputer, Universitas Bhayangkara Jakarta, 2020.
- [10] Imam Tahyudin, dkk, "Pengenalan Machine Learning Menggunakan Jupyter Notebook," Zahira Media Publisher, 2020.
- [11] Julius H. Lolombulan, "Analisis Data Statistika Bagi Peneliti Kedokteran Dan Kesehatan," Andi Publisher, 2020.
- [12] Hanna Willa Danny, "Performa Algoritma K-Nearest Neighbour dalam Memprediksi Penyakit Jantung," Seminar Nasional Informatika (SENATIKA) Prosiding SENATIKA, 2021.
- [13] Donny Maulana, Rezayadi Yahya, "Implementasi Algoritma Naïve Bayes Untuk Klasifikasi Penderita Penyakit Jantung Di Indonesia Menggunakan RapidMiner," SIGMA, Jurnal Teknologi Pelita Bangsa, Volume 10 Nomor 2, 2019.
- [14] Fidia Deny, M. Sabransyah, Tisna Amijaya, Yuki Novia Nasution, "Aplikasi Metode Naïve Bayes dalam Prediksi Risiko Penyakit Jantung," Jurnal EKSPONENSIAL Volume 8, Nomor 2, 2017.
- [15] Abdul Rohman, "Komporasi Metode Klasifikasi Data Mining untuk Prediksi Penyakit Jantung," Jurnal Neo Teknika Volume 2 Nomor 2, hal. 21-28, 2016.
- [16] Pareza Alam Jusia, "Analisis Komparasi Pemodelan algoritma Decision Tree menggunakan Metode Particle Swarm Optimization Dan metode Adaboost untuk Prediksi awal penyakit Jantung," Seminar Nasional Sistem Informasi, Fakultas Teknologi Informasi-UNMER Malang, 2018.
- [17] Ahmad Habibullaah, Agus Navirgo, "Implementasi Data Mining Dengan Algoritma Berbasis Tree Untuk Klasifikasi Serangan Pada Intrusion Detection System (IDS)," Jurnal Sistem Informasi & Manajemen Basis Data (SIMADA) Volume 2 Nomor 2. Jurusan Sistem Informasi Institut Informatika dan Bisnis Darmajaya, 2018.
- [18] Riski Annisa, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," Jurnal Teknik Informatika Kaputama (JTik) Volume 3 , Nomor 1, 2019.
- [19] Sarang Narkhede, "Understanding Confusion Matrix," <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>, 2018.