Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Research, Technology, and Higher Education, Republic Indonesia with Second Grade (Peringkat 2, Sinta 2)
since year 2017 to 2021 according to the decree No. 10/E/KPT/2019.

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

RESTI JOURNAL
(System Engineering and Information Technology)
Vol. 5 No. 2 (2021) 369 - 378 ISSN Electronic Media: 2580-0760

Random Forest Algorithm to Investigate the Case of Acute Coronary Syndrome

Eka Pandu Cynthia[1], M. Afif Rizky A.[2], Alwis Nazir[3], Fadhilah Syafria[4]
[1,2,3,4]Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau
[1]eka.pandu.cynthia@uin-suska.ac.id, [2]afifrizky933@gmail.com, [3]alwis.nazir@uin-suska.ac.id, [4]fadhilah.syafria@uin-suska.ac.id

Abstract

This paper explains the use of the Random Forest Algorithm to investigate the Case of Acute Coronary Syndrome (ACS). The objectives of this study are to review the evaluation of the use of data science techniques and machine learning algorithms in creating a model that can classify whether or not cases of acute coronary syndrome occur. The research method used in this study refers to the IBM Foundational Methodology for Data Science, include: i) inventorying dataset about ACS, ii) preprocessing for the data into four sub-processes, i.e. requirements, collection, understanding, and preparation, iii) determination of RFA, i.e. the "n" of the tree which will form a forest and forming trees from the random forest that has been created, and iv) determination of the model evaluation and result in analysis based on Python programming language. Based on the experiments that the learning have been conducted using a random forest machine-learning algorithm with an n-estimator value of 100 and each tree's depth (max depth) with a value of 4, learning scenarios of 70:30, 80:20, and 90:10 on 444 cases of acute coronary syndrome data. The results show that the 70:30 scenario model has the best results, with an accuracy value of 83.45%, a precision value of 85%, and a recall value of 92.4%. Conclusions obtained from the experiment results were evaluated with various statistical metrics (accuracy, precision, and recall) in each learning scenario on 444 cases of acute coronary syndrome data with a cross-validation value of 10 fold.

Keywords: artificial intelligence, data processing, machine learning, random forest algorithm, supervised learning.

1. Introduction

Previous studies that used the random forest algorithm in solving health cases, such as the detection of congestive heart failure using ECG waves using the random forest algorithm where the study was an experiment that produced various statistical measures against the use of the random forest algorithm for the case of detection of congestive heart failure. From these studies, it can be concluded that the random forest algorithm produces a performance that is considered significant in detecting the disease being studied and provides an insight into expressing cases of congestive heart failure using ECG with 100% accuracy [1]. The random forest algorithm is also used in breast cancer detection, where this study uses a digital track record of 699 instances of 10 attributes. From these studies, it can be concluded that the random forest algorithm provides a significant performance compared to the conventional random binary classifier method with a correlation accuracy of 97% and a sensitivity of 96% [2]. Byeon conducted a study proving whether the RFA is suitable for predicting Parkinson's disease with Mild Cognitive Disorders from Parkinson's Disease with normal cognition [3]. Antoniadi et al. present prediction of caregiver burden in amyotrophic lateral sclerosis: a machine learning approach using random forests applied to a cohort study [4]. Chari et al. Classified diabetes using a random forest algorithm with feature selection [5]. Dai et al. also conducted a survey using a random forest algorithm to diagnose breast cancer [6]. Iwendi et al. used a RFA to predict patients' health exposed to Covid-19 [7].

The Random Forest algorithm is an algorithm that is often used in both classification and regression cases. This algorithm generates a random decision tree. The random forest will select the best tree and will be issued as a decision tree. The random forest algorithm has been widely researched and applied in the health sector. Alam et al. presented a random-forest-based predictor for medical data classification using feature ranking [8]. Kaur et al. combined the random forest algorithm and the internet of things (IoT) in a health monitoring system

Accepted by Editor: 06-03-2021 | Final Revision: 18-04-2021 | Online Publication: 30-04-2021

[9]. Ricordeau et al. applied the random forest algorithm to a health monitoring machine [10]. Kumar performed a health care analysis using the random forest algorithm [11]. Oliver Pauly has presented in his thesis the application of the random forest algorithm in several medical applications such as Multiple Organ Detection and Localization in multi-channel Magnetic Response scans, Multiple Organ Segmentation in CT scans, and Detection of Substantia Nigra Echogenicities in 3D Transcranial Ultra-sound. Towards Computer-Aided Diagnosis of Parkinson Disease [12].

Data science, machine learning, and artificial intelligence are becoming top-trending topics in today's technological world. It is a common thing because today and in the future, the world has entered the era of big data [13]. In processing such a lot of data, data science is the most sought-after scientific field in the period of big data like now. It happens because data science is a data processing science. Data science is not only the science of data processing, but data science functions to combine various techniques, algorithms, and principles of machine learning that are useful for finding hidden patterns in data [14]. Machine Learning is an essential topic in line with the development of knowledge and businesses looking for innovative ways to support and sustain the business realm in reaching new understanding levels [15]. It is because machine learning can help agencies and organizations have the ability to predict what will happen next. A data scientist can process data to find various clustering patterns, regression, and classification to obtain data. It can be seen what patterns, information, and knowledge can help problems such as forecasting, time series, business intelligence, data analytics, and research: prediction, sentiment analysis, and various other practitioners [16].

Data science can determine the paradigm of health cases with multiple techniques and tools in machine learning, statistics, and data visualization in the health sector. In the health sector's case, best practice data science recognizes symptoms, risks and recommends planning before these health risks occur to a person [17]. With data science, a data scientist can get a pattern based on modeling health cases and provide various visualizations for others' knowledge that can be shared with others. The World Health Organization (WHO) states that acute heart disease or coronary syndrome is currently one of the causes of the highest death cases. Every year in the world, it shows that 17.9 million people (31% of global deaths) are caused by coronary heart disease [18]. Akyol et al. have analyzed the Demographic Characteristics of Making Coronary Artery Disease Susceptible using the Random Forests Classifier [19]. Yekkala et al. Combined the random forest algorithm and selection features in predicting heart disease [20]. Ani et al. used the Random Forest Ensemble Classifier to Predict Coronary Heart Disease Using Risk Factors [21].

Based on the background and various studies conducted in the health sector using the machine learning algorithms that have been mentioned, especially in research using the RFA, the author will conduct research and analysis using data science techniques to produce a model using the RFA against cases of the ACS. This study can find various information, patterns, knowledge, and performance of the random forest algorithm in classifying whether or not cases of ACS occur.

2. Research Method

This research was conducted by building a model that can classify ACS into two classes: positive ACS and negative ACS, without using data normalization with a random forest classification algorithm. The following is the modeling workflow built-in performing the ACS classification. The flowchart for research methods is shown in Figure 1.

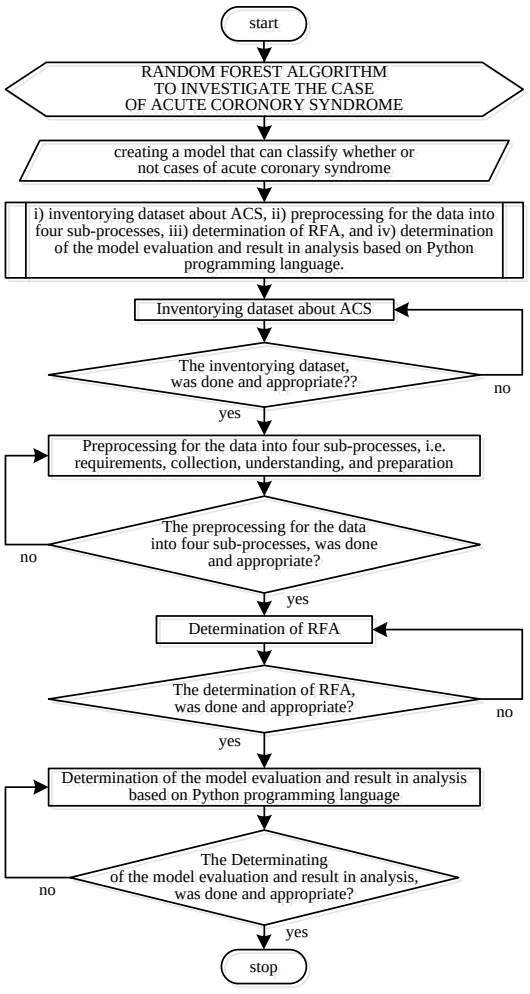


Figure 1. The flowchart for research methods

Based on Figure 1 it can be explained, that the implementation of the research methods in four stages, include i) inventorying dataset about ACS, ii) preprocessing for the data into four sub-processes, i.e.

DOI: https://doi.org/10.29207/resti.v5i2.3000

requirements, collection, understanding, and preparation, iii) determination of RFA, i.e. the "n" of the tree which will form a forest and forming trees from the random forest that has been created, and iv) determination of the model evaluation and result in analysis based on Python programming language.

2.1. Inventorying Dataset about ACS

Before data collection is carried out, it will first require data specifications suitable for research. It is done by conducting a short interview with a cardiac anesthesiologist, namely dr. Vera Muharrami, M.Ked., Sp.An and general practitioner, dr. Alfarisy Zamzami regarding the required parameters. The data used in this study were diagnostic data for patients with an acute coronary syndrome which came from two different sources, namely 303 data from Medan and 141 medical records from Arifin Achmad Hospital Pekanbaru. The data used consists of 13 input parameters that represent cases of acute coronary syndrome by cardiac anesthetists, and the data is more towards laboratory data with the following information. The description of each data parameter is shown in Table 1

Table 1. The description of each data parameter

No.	Parameter Name	Information
1	Age	Patient's Age
2	*Gender*	Patient's Gender
3	Angina's Type	Description of the type and pain in the chest experienced by the patient
4	Resting Blood Pressure	Blood pressure measurements are taken when the patient is resting / not doing activities
5	Cholesterol	Measurement of the patient's cholesterol
6	Fasting/Time Blood Sugar	Blood sugar measurements are taken when the patient is ordered to fast by a specialist
7	Electrocardiography Waveform	The shape of the wave size when the patient is placed on an electrocardiography device
8	Maximum Heart Rate	A sampling of the patient's maximum heart rate
9	Angina Activities	Patient information regarding the presence of chest pain during activities
10	ST_T ECG Line Length	Measure the length of the ST_T line from the ECG patient medical record paper
11	ST_T Slope ECG Shape	Assess the ST_T Slope shape of cardiac patients
12	Num major vessels	Enlarged size of the patient's large blood vessels
13	Thalassemia	A congenital blood disorder characterized by a lack of oxygen-carrying proteins

2.2. Preprocessing

In this study, the preprocessing stage was divided into four sub-processes, i.e. the data requirements stage, the data collection stage, the data understanding or exploratory data analysis (EDA) stage, and the data preparation stage. The data requirement stage is how to determine the correct data specifications to be used in research. In this study, the required data specifications are medical records of acute coronary syndrome patients in inpatient rooms or CVCU rooms. This data is needed because it has a history more complete than outpatient data. At the data collection stage, was collect all specified data to obtain various information such as parameter requirements and values used during the research. Based on the medical records obtained, there are different parameters available but limited to 13 essential parameters in acute coronary syndrome, as shown in Table 1.

All writing of each parameter above was done with a cardiac anesthetist specialist at RSUD Arifin Achmad Pekanbaru. After the data is collected, a table is compiled in a Comma Separated Value (CSV). The EDA stage is an essential part of data science, which means that we must understand how the data composition is seen in terms of data distribution and statistics so that we can get an idea of what is happening in the data. The EDA process for acute coronary syndrome case data includes import Python Library, read data, statistics description, and analysis. The results obtained in the statistical process describe and analyze. The showing of five variables with statistic description and analysis is shown in Figure 2.

	age	gender	cholesterol	resting_blood_pressure	maximum_heart_rate
count	444.000000	444.000000	444.000000	444.000000	444.000000
mean	53.416667	0.662162	217.085586	124.441441	133.768018
std	11.933404	0.473507	65.424827	20.944271	31.178692
min	0.000000	0.000000	71.000000	62.000000	40.000000
25%	46.000000	0.000000	170.750000	110.000000	105.750000
50%	55.000000	1.000000	216.500000	125.000000	138.500000
75%	61.000000	1.000000	259.250000	138.000000	160.250000
max	88.000000	1.000000	564.000000	200.000000	202.000000

Figure 2. The showing of five variables with statistic description and analysis

DOI: https://doi.org/10.29207/resti.v5i2.3000

The EDA process can also provide visualization related to the distribution of data and statistical values that exist in the data. Figure 3 below is an example of a visualization of cholesterol distribution. Normal cholesterol, as we know, it is <200 mg / dL. However, if seen from the results above, patients with average cholesterol had a lesser frequency of 171 cases than 273 patients with cholesterol above 200 mg/dL. The curve of a cholesterol distribution visualization is shown in Figure 3.

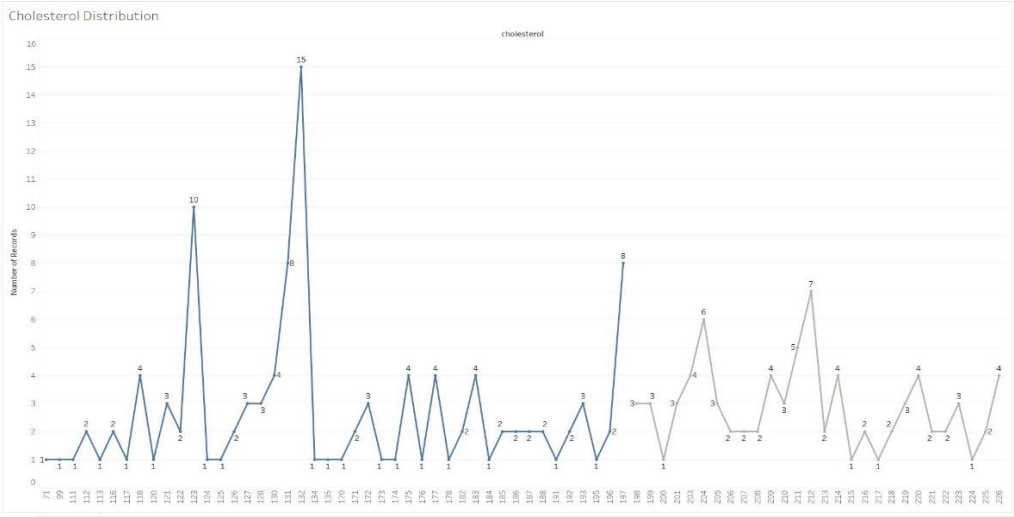


Figure 3. The curve of a cholesterol distribution visualization

Next, we visualize the correlation matrix between variables. In this visualization, we can translate more deeply about what happens to the data we have. For example, the correlation between age and target is negative -0.18, which means that the lower the age, the less likely a person is to develop the acute coronary syndrome. Meanwhile, gender also has a negative correlation, which means that women are less likely to have cases than other variables. Based on these visualization results, there are no variables that are zero or do not correlate at all with other variables, so it can be concluded that all variables are important compositions of acute coronary syndrome cases and will be used for the modeling process using the random forest algorithm. The visualization of the correlation matrix between variables is shown in Figure 4.

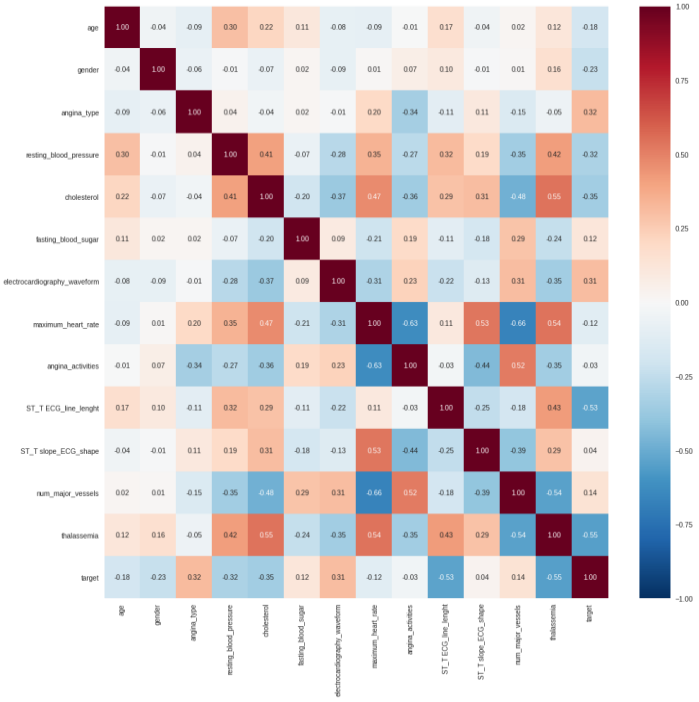


Figure 4. The visualization of correlation matrix between variables

DOI: https://doi.org/10.29207/resti.v5i2.3000

In the data preparation stage, the extract of parameters from the data used, namely the data type transformation, making the data easier to understand and train test split data. At the stage of parameter and data transformation, first, we change the data so that it is easier to understand. Second, the categorical parameter is changed from int64 to object. It is also done to make it easier for the algorithm to read the data type when it becomes the random forest algorithm's input. Third, we create a dummies variable that indicates and separates a variable with categorical values into a unique value. The transformation process can be seen in the following code listing, i.e.:

```
Value Transformation

df['Gender'][df['Gender'] == 0] = 'female'
df['Gender'][df['Gender'] == 1] = 'male'
df['angina_type'][df['angina_type'] == 0] =
'typical angina'
df['angina_type'][df['angina_type'] == 1] =
'atypical angina'
df['angina_type'][df['angina_type'] == 2] =
'non anginal pain'
df['angina_type'][df['angina_type'] == 3] =
'asymtomatics'
df["fastingtime_bloodsugar"][df['fastingtime
_bloodsugar'] == 0] = 'less than 120 mmHg'
df["fastingtime_bloodsugar"][df['fastingtime
_bloodsugar'] == 1] = 'more than 120 mmHg'
df["fastingtime_bloodsugar"][df['fastingtime
_bloodsugar'] == 2] = '120 mmHg'
df['elecrocardiography_result'][df['elecroca
rdiography_result'] == 0] = 'normal'
df['elecrocardiography_result'][df['elecroca
rdiography_result'] == 1] = 'ST_T
abnormallity'
df['elecrocardiography_result'][df['elecroca
rdiography_result'] == 2] = 'left ventrikel
hipertrophy'
df['angina_activity'][df['angina_activity']
== 0] = 'angina_activity_no'
df['angina_activity'][df['angina_activity']
== 1] = 'angina_activity_yes'
df['st_slope_ECG'][df['st_slope_ECG'] == 0]
= 'upslopping'
df['st_slope_ECG'][df['st_slope_ECG'] == 1]
= 'downslopping'
df['st_slope_ECG'][df['st_slope_ECG'] == 2]
= 'flat'
df['thalassemia'][df['thalassemia'] == 1] =
'normal'
df['thalassemia'][df['thalassemia'] == 2] =
'fixed defect'
df['thalassemia'][df['thalassemia'] == 3] =
'reversable defect'
```

```
Data Type Transformation

df['Gender'] = df['Gender'].astype('object')
df['angina_type'] =
df['angina_type'].astype('object')
df[fastingtime_bloodsugar'] =
df['fastingtime_bloodsugar'].astype('object'
)
df['electrocardiography_result'] =
df['electrocardiography_result'].astype('obj
ect')
df['angina_activity'] =
df['angina_activity'].astype('object')
df['st_slope_ECG'] =
df['st_slope_ECG'].astype('object')
df['thalassemia'] =
df['thalassemia'].astype('object')
```

```
Dummies Variable

df = pd.get_dummies(df, drop_first=True)
```

The next stage will separate the training data (train data) and test data (test data). Data split is carried out in sizes of 0.2 (80% training data, 20% testing data), 0.3 (70% training data, 30% testing data), and 0.1 (90% training data, 10% testing data). Data sharing is done using the stratified method, which means that the y_value (target class) condition is the same as the magnitude or proportion of the data separation condition. The separation of training and test data is shown in Table 2.

Table 2. The separation of training and test data

Split Model	Amount of Data	Class 0	Class 1	Split Train: Test	Y_Train Stratify	Data Testing
70%	444	138	306	310 : 133	213 : 97	92 : 41
80%	444	138	306	354 : 89	244 : 110	61 : 28
90%	444	138	306	398 : 45	274 : 124	31 : 14

2.3. Determinating of RFA

The next step is to explain the random forest algorithm's flow in processing acute coronary syndrome data. The random forest algorithm is divided into two parts. The first part is the "n" (tree) which will form a forest with random values. The second part is the algorithm for forming trees from the random forest that has been created.

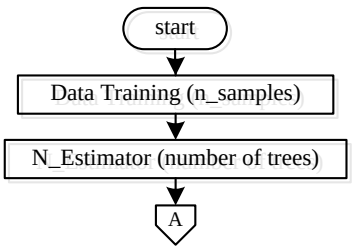
Input:

D, a dataset consisting of *d* rows

k, the number from the number of *trees*

The Random Forest method [22] first created a sample data by taking random from dataset D with replacement. Second, use the Di sample data to build the ith tree (i = 1, 2, … *k*), and third, repeat steps one and two for *k*.

The random forest algorithm begins with the initial selection of "k" samples from the data set, which is done randomly (random). Furthermore, based on the initiation, it is used to form each tree independent from the other trees. After the tree has been built to completion, each tree will give each tree a majority vote against the tree that has received the most votes from the other trees. It makes the tree contained in the random forest algorithm untraceable because it uses a random subset of the existing dataset. The flowchart of the Random Forest algorithm is shown in Figure 5.



DOI: https://doi.org/10.29207/resti.v5i2.3000

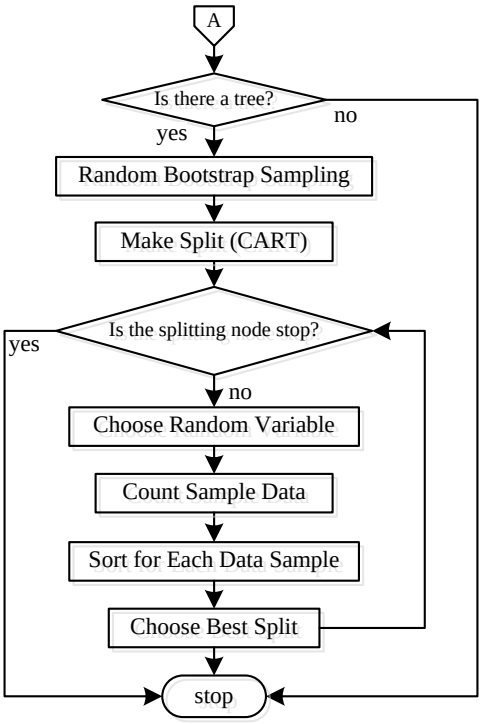


Figure 5. The flowchart of the Random Forest algorithm

2.4. Model Evaluation and Result Analysis

This study uses the Python programming language version 3.7.4 and the Integrated Development Environment (IDE) jupyter notebook/jupyterlab/googlecollab. Model evaluation Analysis of the results of modeling using the random forest algorithm in cases of the acute coronary syndrome is carried out in several stages: First, using the Confusion Matrix test [23], which is used to calculate the level of accuracy described in the table which states the number of correct test data and incorrectly classified test data. Confusion matrix testing is done based on a data set that has two classes, namely positive class, and negative class. Different classes consist of four cells, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Based on these terms, it can be described in confusion matrix. The confusion matrix is shown in Table 3.

Table 3. The confusion matrix

Classification		Result Prediction Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

The classification model performance level can be measured and evaluated by calculating the accuracy from the confusion matrix table. Accuracy is the percentage of test data that can be classified correctly by the built classification model. The equation of accuracy [23] is shown in Eq. (1).

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \times 100\% \quad (1).$$

Second, precision testing, a testing technique used to calculate performance by using a confusion matrix table of the algorithms used in a case. Precision is used to calculate the true-positive prediction ratio compared to the overall results that are predicted to be positive (true-positive predictions). The precision calculation formula is shown in Eq. (2).

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2).$$

Meanwhile, recall is a true-positive prediction compared to all true-positive (true-positive ratio). The recall calculation formula is shown in Eq. (3).

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3).$$

It is finally testing three different split scenarios. This is done to determine the best split ratio of the resulting model and assess the existing model's overall performance. The split ratios used and previously described are 70% (70:30), 80% (80:20), and 90% (90:10).

3. Results and Discussion

This study uses 444 data obtained and has been described in inventorying dataset about ACS. The parameters in the dataset of acute coronary syndrome cases are carried out in the data preprocessing stage, which can be seen in section 2.2 to make the data easier to process or use in data mining. The following is patient data on acute coronary syndrome cases used in modeling. The patient data on acute coronary syndrome cases are shown in Table 4.

Table 4. The patient data on acute coronary syndrome cases

Parameter's Name	Patient's	Medical	Record	Value
Age	63	37	41	…
Gender	1	1	0	…
Angina's Type	3	2	1	…
Resting Blood Pressure	145	130	130	…
Cholesterol	233	250	204	…
Fasting/Time Blood Sugar	1	0	0	…
Electrocardiography Waveform	0	1	0	…
Maximum Heart Rate	150	187	172	…
Angina activities	0	0	0	…
ST_T ECG Line Length	2.3	3.5	1.4	…
St slope ECG Shape	0	0	2	…
Num major vessels	0	0	0	…
Thalassemia	1	2	2	…

Based on Table 5 it can be explained, the data that has been preprocessed is divided into two parts, i.e. training and test data. The training data is used to build a model with a machine learning process against the dataset used. Furthermore, the test data is used to classify the class whose class is not yet known. In testing, the researcher used a test scenario such as the number of trees in the Random Forest (RF) with the value of n_estimator =

DOI: https://doi.org/10.29207/resti.v5i2.3000

100, the depth level = 4, and cross-validation value (K) = 10 fold. In determining the accuracy of the system from each iteration, using the most excellent accuracy. This test scenario is carried out to maximize the data used, produce a system with good performance and an accurate analysis level. These scenarios are combined in the three models used by calculating the accuracy, which will then compare the accuracy against each model to find out which model has the best performance from the Random Forest algorithm in detecting acute coronary syndrome cases. Several stages in the results and discussion, including i) classification of random forest, ii) evaluation of models, i.e. curve of cross-validation and learning, and iii) the testing and result in analysis, i.e. results from testing of the correlation matrix and precision and recall.

3.1. Random Forest Classification

The determine the many trees' parameters (n_estimator) to be built and the depth level in the random forest model initiation stage. After successful model initiation, the data that has been separated as training data will be entered into the model and will be trained by the model that has been built. In this study, the random forest model made has a depth of 4, and 100 trees were built. After the random forest model was created, training data would be entered to be trained into the model using the fit function. Then we will provide an iteration to describe the accuracy of the many n_estimators in the random forest model. The accuracy results for the loop on each data split can be seen in the table. The accuracy loop n_estimator is shown in Table 5.

Table 5. The accuracy loop n-estimator

N-Estimator	Accuracy Value's Range
70%	79%-84%
80%	75%-83%
90%	60%-77%

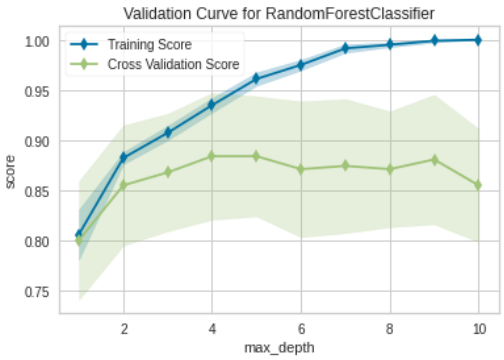
Based on Table 5 it can be explained, that the highest accuracy is owned by the model with a split ratio of 70% (70:30) with the value ranging from 79% to 84%.

3.2. Model Evaluation

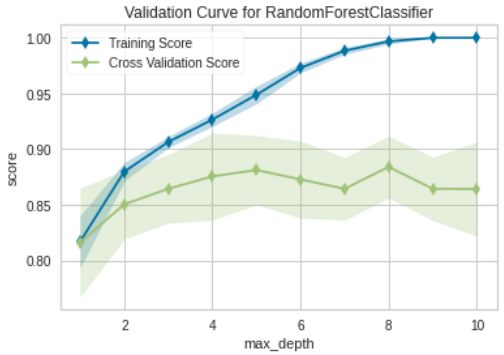
At this stage, the providing a visualization of the performance evaluation of the model that has been produced. Creating a validation curve is one way to assess the initial effectiveness. The validation curve is a plot (graph) that shows how the model's performance responds to changes in hyperparameter values. The graph shows both training data in blue stripes and validation data in green lines. Score validation allows us to infer how the model will respond to unseen data. The hyperparameters used this time are each tree's depth (max depth) and the cross-validation of 10 folds. The evaluation for the models was carried out through cross-validation curves and learning curves.

3.2.1. Cross-validation curve

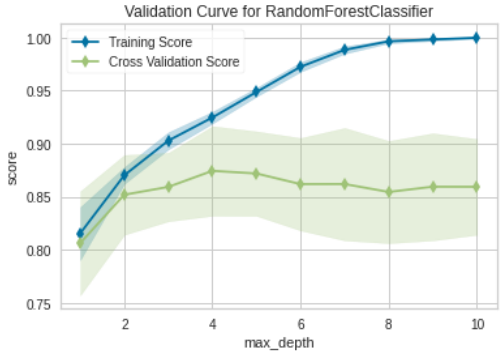
The validation for the random forest classifier is curved. The cross-validation curve is shown in Figure 6.



#a) for split data of 70%



#b) for split data of 80%



#c) for split data of 90%

Figure 6. The cross-validation curve

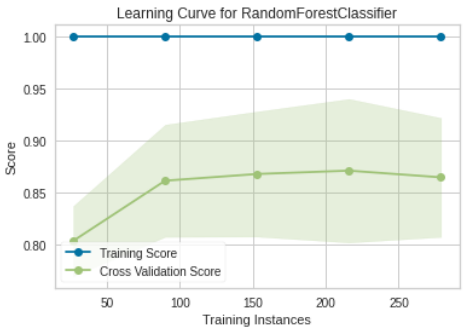
Based on Figure 6, shows that the resulting models' performance with a split ratio of (i) 70% has a cross-validation accuracy range from 80% to 89%, (ii) 80% has a cross-validation accuracy range from 83% to 88%, and (iii) 90% has a cross-validation accuracy range from 81% to 86%. As looking at the resulting curve movement patterns in each model, it can be seen that the deeper the tree level the more complex the model is, so the accuracy tends to decrease.

DOI: https://doi.org/10.29207/resti.v5i2.3000

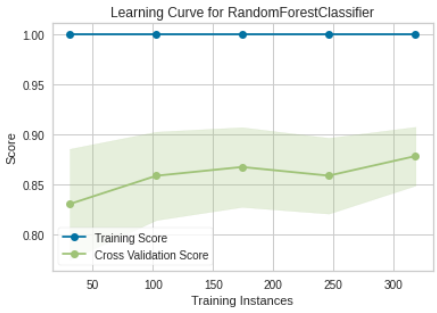
3.2.2. Learning curve

The learning curve is a plot that describes the model's learning pattern of the data that is the input value and will read the pattern from the data so that it can produce a decision. The learning curve is shown in Figure 7.

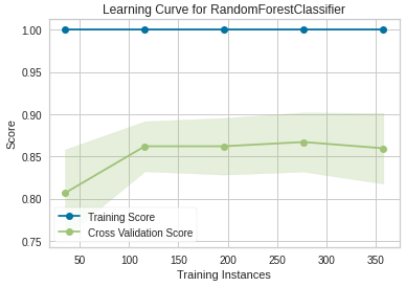
Based on Figure 7, shows that the resulting model's performance with a split ratio of (i) 70% gives a learning accuracy range from 78% to 85%, (ii) 80% has a cross-validation accuracy range from 75% to 86%, and (iii) 90% has a cross-validation accuracy range from 82% to 86%. As looking at the training and learning data scores marked with a blue line, while the cross-validation data is marked with a green line, it can be seen the cross-validation score increases, it shows that the model with the owned data has a good performance.



#a) for split data of 70%



#b) for split data of 80%



#c) for split data of 90%

Figure 7. The learning curve

3.3. Testing and Result in the Analysis

The explanation for the testing and result in the analysis was carried out through the result of correlation matrix testing and the results of precision and recall testing.

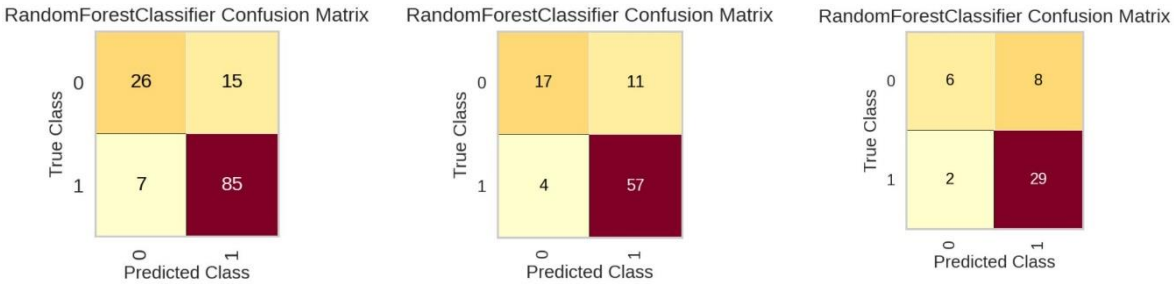
3.3.1. Correlation matrix testing result

This test is done by providing the confusion matrix visualization results to get the random forest algorithm model's overall accuracy or performance. The process of calculating confusion matrix accuracy on each split data can be seen in the table. The confusion matrix accuracy calculation is shown in Table 6.

Table 6. The confusion matrix accuracy calculation

Split Data	Identification Predicted Value	Actual Class	
		Negative	*Positive*
70%	*Negative*	26	15
	Positive	7	85
	Counting: (85+26)/(85+26+15+7)×100% = 83,45%		
		Negative	*Positive*
80%	*Negative*	17	11
	Positive	4	57
	Counting: (17+57)/(17+57+11+4)×100% = 83,14%		
		Negative	*Positive*
90%	*Negative*	6	8
	Positive	2	29
	Counting: (29+6)/(29+6+8+2)×100% = 77,78%		

Based on Table 7 it can be explained, that there is the visualizing of the correlation matrix test with the confusion matrix. The visualization of the testing correlation matrix is shown in Figure 8.



#a) for Split Data of 70% #b) for Split Data of 80% #c) for Split Data of 90%

Figure 8. The visualization of the testing correlation matrix

DOI: https://doi.org/10.29207/resti.v5i2.3000

Based on Figure 8 it can be explained, that the results of the visualization of split data of 70% model for testing using the confusion matrix has the highest accuracy value of 83.14%.

3.3.2. The Results of precision and recall testing

The process of counting the precision accuracy for each split data can be seen in the table. The precision counting is shown in Table 7.

Table 7. The precision counting

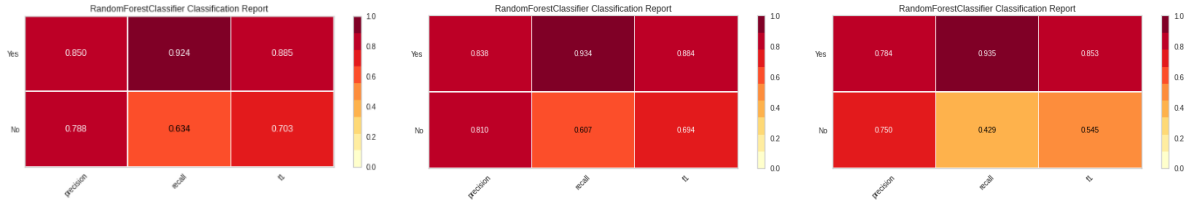
Split Model	Precision Counting = TP/(TP+FP)	Result
70%	85/(85+15)	85,00%
80%	57/(57+11)	83,82%
90%	29/(29+8)	78,38%

Based on Table 7 it can be explained, that the split data of 70% has the highest accuracy value of 85%. Meanwhile, the process of counting the accuracy of recall on each split data can be seen in the table. The recall counting is shown in Table 8.

Table 8. The recall counting

Split Model	Recall Calculation = TP/(TP+FN)	Result
70%	85/(85+7)	92,39%
80%	57/(57+4)	93,44%
90%	29/(29+2)	93,55%

Based on Table 9 it can be explained, that split data of 90% has the highest accuracy value of 93.55%. The visualization of the report precision-recall of the models with the split ratio of 70%, 80%, and 90% random forest algorithm in cases of the acute coronary syndrome can be seen in the following figures. The classification random is shown in Figure 9.



#a) for split data of 70% #b) for split data of 80% #c) for split data of 90%

Figure 9. The classification random

Based on Figure 9 it can be explained, that the results of the visualization that the precision and recall of the models built with a predetermined split ratio give results that are following the manual calculations that have been done previously.

4. Conclusion

The conclusion from modeling using the random forest algorithm for cases of the acute coronary syndrome, namely: First, in confusion matrix testing, the results of the overall accuracy or performance of the resulting model with split ratios of 80%, 70%, and 90% respectively, namely 83.14 %, 83.45%, and 77.78% with the highest accuracy owned by the model with a split ratio of 70% (70:30). Second, in a random forest, the number of trees formed (n_estimator) dramatically affects the model's performance because theoretically, the more trees, the better.

Third, based on observations of the resulting accuracy in the iteration of the testing data, each number of trees and trees' addition can provide different accuracy but tend to be more stable. Fourth, based on observations made on the validation curve, with hyperparameters, namely max_depth and ten-fold cross-validation, the deeper the tree formed will result in a more complex model so that the accuracy tends to decrease. Fifth, based on observations made on the learning curve with cross-validation of 10 fold, the model produces an accuracy of each split ratio from 80%, 70%, and 90% on an average range of 80% to 89%.

Sixth, for each model, the split ratio of 80% is 83.82%, for the model with a 70% split has a precision of 85% and for the model with a split ratio of 90% produces a precision of 78.37% with the highest precision result is owned by the model. with a split ratio of 70% with a precision of 85%. Seventh, in the recall test of the split ratio model, 80% has a recall of 93.34%, and in the split model, 70% has a recall of 92.24% then the model with a split ratio of 90% produces a recall of 93.35% with the highest recall. They are owned by a model with a split ratio of 90%. And eighth, when viewed from the classification report results, the ratio that better recognizes each class with the highest level of balance seen from the ratio of precision and recall and the accuracy of each target is a model with a split ratio of 70% (70:30) with a precision of 85 %, a recall of 92.24% and an accuracy of 83.45%.

For further research, it can be suggested to use optimization in the random forest algorithm, especially to overcome the over-fitting or under-fitting of the resulting model.

References

[1] Z. Masetic and A. Subasi, “Congestive Heart Failure Detection Using Random Forest Classifier,” *Comput. Methods Programs Biomed.*, vol. 130, pp. 54–64, 2016, doi: 10.1016/j.cmpb.2016.03.020.

DOI: https://doi.org/10.29207/resti.v5i2.3000

[2] T. I. Rohan, Awan-Ur-Rahman, A. B. Siddik, M. Islam, and M. S. U. Yusuf, "A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost," in *5th International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, IC4ME2 2019*, 2019, no. June 2020, pp. 10–14, doi: 10.1109/IC4ME247184.2019.9036697.

[3] H. Byeon, "Is the Random Forest Algorithm Suitable for Predicting Parkinson's Disease with Mild Cognitive Impairment Out of Parkinson's Disease with Normal Cognition?," *Int. J. Environ. Res. Public Health*, vol. 17, no. 7, pp. 1–14, 2020, doi: 10.3390/ijerph17072594.

[4] A. M. Antoniadi, M. Galvin, M. Heverin, O. Hardiman, and C. Mooney, "Prediction of Caregiver Burden in Amyotrophic Lateral Sclerosis: A Machine Learning Approach Using Random Forests Applied to A Cohort Study," *BMJ Open*, vol. 10, no. 2, pp. 1–8, 2020, doi: 10.1136/bmjopen-2019-033109.

[5] K. K. Chari, M. Chinna Babu, and S. Kodati, "Classification of Diabetes Using Random Forest with Feature Selection Algorithm," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 1, pp. 1295–1300, 2019, doi: 10.35940/ijitee.L3595.119119.

[6] B. Dai, R. C. Chen, S. Z. Zhu, and W. W. Zhang, "Using Random Forest Algorithm for Breast Cancer Diagnosis," in *International Symposium on Computer, Consumer and Control (IS3C)*, 2018, pp. 449–452, doi: 10.1109/IS3C.2018.00119.

[7] C. Iwendi *et al.*, "COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm," *Front. Public Heal.*, vol. 8, no. July, pp. 1–9, 2020, doi: 10.3389/fpubh.2020.00357.

[8] M. Z. Alam, M. S. Rahman, and M. S. Rahman, "A Random Forest Based Predictor for Medical Data Classification Using Feature Ranking," *Informatics Med.*, vol. 15, no. April, pp. 1–12, 2019, doi: 10.1016/j.imu.2019.100180.

[9] P. Kaur, R. Kumar, and M. Kumar, "A Healthcare Monitoring System Using Random Forest and Internet of Things (IoT)," *Multimed. Tools Appl.*, vol. 78, no. 14, pp. 19905–19916, 2019, doi: 10.1007/s11042-019-7327-8.

[10] J. Ricordeau and J. Lacaille, "Application of Random Forests To Engine," in *International Congress of the Aeronautical Sciences (ICAS)*, 2010, no. April, pp. 1–10.

[11] K. Kumar, "Health Care Analysis Using Random Forest Algorithm," *J. Chem. Pharm. Sci.*, vol. 10, no. 3, pp. 1359–1361, 2017.

[12] O. Pauly, "Random Forests for Medical Applications," Technischen Universität München, 2012.

[13] L. Pierson, *Data Science for Dummies*, 2nd ed. New Jersey: John Wiley & Sons Inc., 2017.

[14] Y. Liu, *Python Machine Learning By Example - Second Edition*, 1st ed. Birmingham: Packt Publishing, 2017.

[15] P. Mathur, *Machine Learning Applications Using Python*, 1st ed. Karnataka: Apress Media LLC, 2019.

[16] M. Kubat, *An Introduction to Machine Learning*, 2nd ed. Switzerland: Springer International Publishing AG, 2017.

[17] E. Corbett, "The Real-World Benefits of Machine Learning in Healthcare," 2017. https://www.healthcatalyst.com/clinical-applications-of-machine-learning-in-healthcare (accessed Mar. 06, 2021).

[18] W. H. Organization, "Cardiovascular diseases," 2017. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Mar. 27, 2021).

[19] K. Akyol, E. Çalik, Ş. Bayir, B. Şen, and A. Çavuşoılu, "Analysis of Demographic Characteristics Creating Coronary Artery Disease Susceptibility Using Random Forests Classifier," in *International Conference on Soft Computing and Software Engineering (SCSE)*, 2015, vol. 62, pp. 39–46, doi: 10.1016/j.procs.2015.08.407.

[20] I. Yekkala and S. Dixit, "Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection," *Int. J. Big Data Anal. Healthc.*, vol. 3, no. 1, pp. 1–12, 2018, doi: 10.4018/ijbdah.2018010101.

[21] R. Ani, A. Augustine, N. C. Akhil, and O. S. Deepa, "Random Forest Ensemble Classifier to Predict the Coronary Heart Disease Using Risk Factors," in *International Conference on Soft Computing System, Advances in Intelligent Systems and Computing*, 2016, vol. 397, pp. 701–710, doi: 10.1007/978-81-322-2671-0_66.

[22] S. Polamuri, "How The Random Forest Algorithm Works in Machine Learning," 2017. https://dataaspirant.com/random-forest-algorithm-machine-learing/ (accessed Mar. 06, 2021).

[23] J. Novakovic, A. Veljovi, S. Iiic, Z. Papic, and M. Tomovic, "Evaluation of Classification Models in Machine Learning," *Theory Appl. Math. Comput. Sci.*, vol. 7, no. 1, pp. 39–46, 2017, [Online]. Available: https://uav.ro/applications/se/journal/index.php/TAMCS/article/view/158.

DOI: https://doi.org/10.29207/resti.v5i2.3000