Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 9 No. 1 2024, pp. 1 – 7

Implementation Of Feature Selection Chi-Square To Improve The Accuracy Of The Classification Model Using The Random Forest Algorithm On Coronary Artery Disease

Ida Bagus Saya Mahendra[1], Tatik Widiharih[2]
[1,2]Universitas Diponegoro, Jalan Prof. Sudarto No.13, (024) 7460036, Semarang
E-mail : satyamahendra09@gmail.com[1], widiharih@gmail.com[2]*

Fajar Agung Nugroho*[3], Priyo Sidik Sasongko[4]
[3,4]Universitas Diponegoro, Jalan Prof. Sudarto No.13, (024) 7460036, Semarang
E-mail : fajar@live.undip.ac.id[3], priyosidiksasongko@lecturer.undip.ac.id[4]
**Corresponding author*

Abstract - A Coronary heart disease is a disease in which the occurrence of blockages in the blood vessels in the heart. Coronary heart disease is a fatal disease, it is better to get as much information about this disease as possible. Data Mining can classify whether a person has heart disease or not based on symptoms. Data mining builds a model that can predict whether a person has heart disease or not. How well a model performs classification can be determined from its accuracy value, but this accuracy value can still be improved. Increasing the accuracy value can be done by performing Feature Selection. The research object used in this research is a dataset about coronary heart disease obtained from the Kaggle website. The classification method used in this modeling is the Random Forest algorithm to classify whether a person has coronary heart disease or not. The Random Forest Algorithm is a classification algorithm consisting of Decision Trees for classifying. The Random Forest algorithm is used because it has been proven to produce good accuracy in several previous studies. The Feature Selection method used in this modeling is the Chi-Square hypothesis test to determine whether there is an effect of each independent variable on the dependent variable. This research compared the value of modeling accuracy without using Feature Selection with modeling using Feature Selection. The result of this study is that the model without Chi-Square Feature Selection produced an accuracy value of 96,05% and the model with Chi-Square Feature Selection produced an accuracy value of 97,33%.

Keywords - Chi-Square Feature Selection, Random Forest, Data Mining, Machine Learning, Coronary Artery Disease

1. INTRODUCTION

Cardiovascular diseases are a group of disorders of the heart and blood vessels. It includes coronary artery disease, where blockages are built up by fat and plaques that clog the blood flow. Cardiovascular diseases are the leading cause of death worldwide, claiming the lives of an estimated 17.9 million people each year [1]. It is best for humans to dig deeper for information about this disease to gain more knowledge for future use.

Data mining is a study to collect, clean, process, analyze, and gain important information from the data [2]. Data mining has a few techniques, including estimation, prediction, classification, clustering, and association [3]. Classification is a technique to find models in order to explain or differentiate concepts or data classes, with the aim of being able to estimate the unknown class object. Classification methods that are commonly used are support vector machines, multilayer perceptrons, nave bayes, ID3, ensemble methods, etc. [4]. The classification model’s performance can be improved by implementing other techniques, for example, feature selection.

Data is typically incomplete, dirty, and inconsistent; therefore, preprocessing the data is required to improve the accuracy and efficiency of the data mining technique that will later be used to process the data [5]. Feature selection is a technique to remove irrelevant features or transform features into a more suitable environment for analysis in order to improve the performance of the model [2]. A Statistical approach is one effective way to do a feature selection process within the data. In this research, Chi Square is chosen as a feature selection method since it has excellent performance, especially in multi-class data [6]. Earlier studies proved that implementing feature selection technique would improve the performance of the classification model. The study conducted by Hasan et al., in 2015 successfully increased the classification model’s accuracy on the KDD’99 dataset from 91,4% to 91,9% by implementing feature selection [7]. The study conducted by Prasetiyowati et al. in 2020 proved that using correlation-based feature selection improved the accuracy of the models on the urban land cover and Parkinson's datasets [8].

Random Forest is an ensemble method that contains CART decision trees to classify a class [9]. Earlier studies proved that the Random Forest algorithm for classification produced a reasonable value in terms of classification model accuracy. The study conducted by Singh et al. in 2017 built a classification model on heart disease that produced an accuracy of 85.81% [10]. Another study conducted by Pal and Parija in 2021 also built a classification model on heart disease that produced an accuracy of 86.9% [11]. Earlier studies also compared classification algorithms on heart disease datasets and proved Random Forest is superior to other classification algorithms. The study conducted by Ani et al., in 2015 compared 4 classification algorithms, with the results of random forest with the highest accuracy of 89% [12]. A Study conducted by Katarya and Meena in 2020 compared 9 classification algorithms, with the results of random forest having the highest accuracy of 95% [13]. This study compares the accuracy produced by two classification models. The first model used

Chi-Square feature selection, and the second model did not use Chi-Square feature selection. This study proposes to use Chi-Square Feature Selection and the Random Forest Algorithm on the Coronary Artery Disease dataset to improve the classification model's performance, mainly the accuracy score. K-fold cross-validation is also used in this study to get a better generalization score of the accuracy produced from the classification model. The object of this study is cardiovascular disease, specifically a coronary artery disease dataset.

2. RESEARCH METHOD

This research used a coronary artery disease dataset taken from Kaggle [14]. The dataset was processed using Google Colaboratory and Python as its programming language. To achieve an unbiased final result, the parameters of the functions used are set to default. This particular study is a comparative study. This study compares the performance score, mainly the accuracy produced by 2 classification models. The first model implements the Chi-Square feature selection technique, while the second model does not implement the Chi-Square feature selection technique.

The feature selection method used in this study is the Chi-Square hypothesis testing technique. This study uses the Chi-Square hypothesis testing technique because the majority of the features are categorical and the output variables are also categorical [15]. The dataset used has several continuous features; these features are categorized in advance so that they can be processed using the Chi-Square hypothesis testing technique.

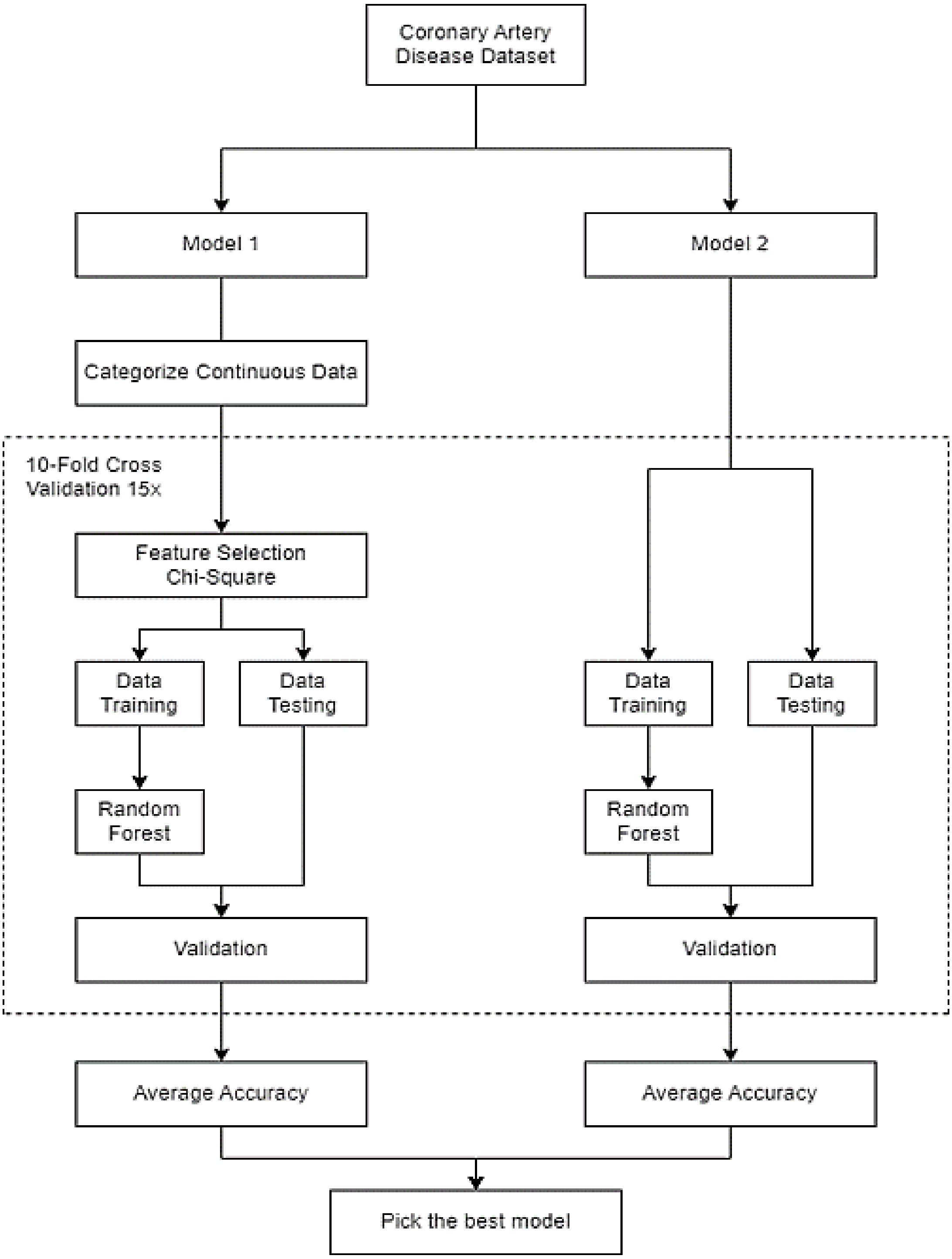

Figure 2. Research Method

The Chi-Square hypothesis test determines whether a feature is statistically relevant to its output variable. To determine whether the feature is relevant or not, first of all, determine the following hypothesis:

H_0 = There is no significant relationship between the feature and the output variable.

H_1 = There is a significant relationship between the feature and the output variable.

Then determine the significance level, or alpha value, which is 0.05. Then build a contingency table between Xi and Y features. Then calculate the Chi value based on the contingency table that was previously built. Calculating the chi-square value can be done using the following formula:

$$X^2 = \sum \frac{(O - E)^2}{E} \tag{1}$$

The chi-square value obtained is compared with the critical value based on the predetermined degrees of freedom and alpha value. If the Chi-Square value is greater than the critical value, then the feature rejects the H0 and accepts the H1, which means that the feature has a significant relationship. If the chi-square value is less than the critical value, then the feature accepts H0 and rejects H1, which means that the feature has no significant relationship. The classification algorithm used to build the classification model is Random Forest. To get more generalized results, K-Fold Cross Validation is implemented with 10 as the value of K [8]. Each test is done by changing the value of the seed to generate random data [8]. The 15 seeds are numbers that are randomly generated with a range of 0 to 1000.

3. RESULTS AND DISCUSSION

3.1 Experiment Results

The original Coronary Artery Disease dataset consists of 20 features, 1 output attribute, and 333 rows of data [14]. The 20 features in the dataset consist of 8 continuous features and 12 categorical features. The continuous data is categorized to be able to be calculated using Chi-Square. The results of the categorization of the continuous data transformed the dimension of the dataset from 20 features to 19 features. The patient's weight and height are combined into the body mass index as a new feature, resulting in a one-feature reduction.

3.2 First model (using feature selection Chi-Square)

The first model is built using only the relevant features from Chi-Square hypothesis testing. The results of the Chi-Square score for each feature can be seen in Table 1.

Table 1. Chi-Square Score of Every Features

No	Feature	Calculated Chi	DoF	Chi Table	Decision
1	Age	26,677176	3	7,815	Reject H_0
2	Sex	0,060973	1	3,841	Accept H_0
3	Smoke	0,735512	1	3,841	Accept H_0
4	Years	2,463149	4	9,488	Accept H_0
5	Ldl	1,58013	2	5,991	Accept H_0
6	Chp	2,742789	3	7,815	Accept H_0
7	bmi	0,282972	3	7,815	Accept H_0
8	Fh	1,164058	1	3,841	Accept H_0
9	Active	13,687813	1	3,841	Reject H_0
10	Lifestyle	9,993982	2	5,991	Reject H_0
11	Ihd	15,287641	1	3,841	Reject H_0

12	Hr	4,076789	2	5,991	Accept H_0
13	Dm	4,39462	1	3,841	Reject H_0
14	Bpsys	0,694821	2	5,991	Accept H_0
15	Bpdias	2,417011	2	5,991	Accept H_0
16	Htn	0,221111	1	3,841	Accept H_0
17	Ivsd	0,345502	1	3,841	Accept H_0
18	Ecgpatt	280,594753	3	7,815	Reject H_0
19	Qwave	51,384917	1	3,841	Reject H_0

There are 7 features that reject H0, which means those respected features are statistically significant. The 7 features are age, activity, lifestyle, ihd, dm, ecgpatt, and qwave. The seven features will be used to build the first model. The performance results of the first model can be seen on Table 2. The average accuracy, precision, recall, and f1 score produced by the first model, respectively, are 97,33%, 94,57%, 99,93%, and 96,88%. The first model performed really well.

Table 2. First Model's Performance Score

Seed	Accuracy	Precision	Recall	F1
143	97,575758	94,558824	100,000000	96,847291
127	97,575758	93,789593	100,000000	96,764751
928	97,272727	94,558824	100,000000	96,447291
897	97,272727	94,558824	100,000000	96,447291
722	97,575758	95,058824	99,090909	96,847291
458	97,272727	94,558824	100,000000	97,142857
448	97,575758	94,558824	100,000000	96,847291
760	97,575758	94,558824	100,000000	97,460317
954	97,272727	94,558824	100,000000	96,447291
779	96,978610	94,558824	100,000000	97,142857
61	97,272727	94,558824	100,000000	96,847291
609	97,272727	95,058824	100,000000	96,847291
214	96,978610	94,558824	100,000000	97,142857
48	97,272727	94,558824	100,000000	97,164751
879	97,272727	94,558824	100,000000	96,847291

3.3. Second Model (Without using Feature Selection Chi-Square)

The second model was built utilizing all of the features because the first model did not implement feature selection. The performance results of the second model can be seen in Table 3.

Table 3. Second Model's Performance Score

Seed	Accuracy	Precision	Recall	F1
143	96,087344	92,187029	99,166667	95,212508
127	96,087344	92,187029	99,166667	95,213957
928	95,499109	92,122926	99,166667	95,612508
897	96,087344	93,020962	99,166667	95,212508
722	95,793226	92,187029	99,166667	94,345841
458	96,675579	92,187029	99,166667	94,779175
448	96,675579	92,251131	99,166667	94,345841

760	96,087344	92,251131	99,166667	95,213957
954	96,087344	92,251131	99,166667	95,929968
779	96,087344	92,251131	99,166667	95,212508
61	95,793226	92,251131	99,166667	94,779175
609	95,793226	92,687029	98,333333	95,647291
214	96,087344	93,020362	99,166667	95,647291
48	96,087344	93,020362	99,166667	94,736318
879	95,793226	92,251131	99,166667	95,213957

The average accuracy, precision, recall, and f1 score produced by the second model, respectively, is 96,05%, 92,40%, 99,11%, and 95,14%. The second model performed excellently as well, but it’s slightly worse than the first model.

3.4. Performance Comparison

The comparison of the performance results from both models is concluded and can be seen in Table 4.

Table 4. Performance Score Comparison of Each Model

Model	Accuracy	Precision	Recall	F1
First Model	97,33%	94,57%	99,93%	96,88%
Second Model	96,05%	92,40%	99,11%	95,14%
Difference	1,28%	2,16%	0,82%	1,74%

Based on Table 4, The first model performed better than the second model, which means that the model that used feature selection can classify better than the model that didn’t use feature selection. There is an improvement of 1,28% in accuracy, 2,16% in precision, 0,82% in recall, and 1,74% in F1 score.

4. CONCLUSION

The conclusion of this study is that implementing Chi-Square feature selection to remove irrelevant features and using Random Forest algorithm on the Coronary Artery Disease dataset improved the performance score in accuracy for 1,28%, precision for 2,16%, recall for 0.82%, and f1 score for 1,74%. The classification model that implements feature selection can better determine whether someone has the Coronary Artery Disease or not.

REFERENCES

[1] World Health Organization, *Cardiovascular Disease (CVDs),* 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] C. C. Aggarwal, *Data Mining*. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.

[3] Z. R. S. Elsi *et al.*, *Utilization of Data Mining Techniques in National Food Security during the Covid-19 Pandemic in Indonesia,* Journal of Physics: Conference Series, 2020, doi: 10.1088/1742-6596/1594/1/012007.

[4] M. A. Muslim, B. Prasetiyo, E. L. H. Mawarni, A. J. Herowati, Mirqotussa’adah, S. H. Rukmana, A. Nurzahputra, *Data Mining Algoritma C4.5.*, ILKOM UNNES http://lib.unnes.ac.id/33080/

[5] S. García, J. Luengo, and F. Herrera, “Intelligent Systems Reference Library 72 Data

Preprocessing in Data Mining.” [Online]. Available: http://www.springer.com/series/8578

[6] I. Sumaiya Thaseen and C. Aswani Kumar, *Intrusion detection model using fusion of chi-squarefeature selection and multi class SVM,* Journal of King Saud University - Computer and Information Sciences, 2017, doi: 10.1016/j.jksuci.2015.12.004.

[7] Md. A. M. Hasan, M. Nasser, S. Ahmad, and K. I. Molla, *Feature Selection for Intrusion Detection Using Random Forest,* Journal of Information Security, doi: 10.4236/jis.2016.73009.

[8] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, *Feature selection to increase the random forest method performance on high dimensional data,* International Journal of Advances in Intelligent Informatics, 2020, doi: 10.26555/ijain.v6i3.471.

[9] L. Breiman, *Random Forest, Machine Learning,* 2001, https://doi.org/10.1023/A:1010933404324

[10] Y. K. Singh, N. Sinha, and S. K. Singh, *Heart disease prediction system using random forest,* Communications in Computer and Information Science,doi: 10.1007/978-981-10-5427-3_63.

[11] M. Pal and S. Parija, *Prediction of Heart Diseases using Random Forest,* Journal of Physics: Conference Series, 2021, doi: 10.1088/1742-6596/1817/1/012009.

[12] R. Ani, A. Augustine, N. C. Akhil, and O. S. Deepa, *Random forest ensemble classifier to predictthe coronary heart disease using risk factors,* Advances in Intelligent Systems and Computing,2016, doi: 10.1007/978-81-322-2671-0_66.

[13] R. Katarya and S. K. Meena, *Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis,* Health Technol (Berl), 2021, doi: 10.1007/s12553-020-00505-7.

[14] Hangaw Qadir, *Coronary Artery Disease,* Kaggle.com, 2022. https://www.kaggle.com/datasets/hangawqadir/erbil-heart-disease-dataset

[15] Jason Brownlee, *How to Choose a Feature Selection Method For Machine Learning,* machinelearningmastery.com, 2019, https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/