JESICA: Journal of Enhanced Studies in Informatics and Computer Applications
Journal homepage: jesica.itsk-soepraoen.ac.id
e-ISSN: 3046-6938 p-ISSN: 3046-6997
Vol. 1, No. 2, July 2024, Pp. 35-39

Logistic Regression's Effectiveness in Feature Selection with Information Gain in Predicting Heart Failure Patients

Mochammad Anshori*[1], M. Syauqi Haris[2], Arif Wahyudi[3]
1,2,3 Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW, Indonesia

***Corresponding author**
E-mail address:
moanshori@itsk-soepraoen.ac.id

Keywords:
Heart failure, information gain, feature selection, prediction, logistic regression

Abstract
Heart failure is a chronic illness that obstructs blood flow, which is necessary for the body to circulate oxygen. Patients with heart failure have a poor chance of survival, as evidenced by the high death rate. The hospital's infrastructure and medical facilities determine the degree of patient safety, and the patients' medical records play a significant role in ensuring that they receive the right care. As a result, a system that uses specific data to forecast the safety of heart failure patients is required. Machine learning, a computer-based approach, is one way to get around this. The logistic regression algorithm has been used to generate predictions in earlier studies. The approach for feature selection from the dataset that is suggested in this study is information gain. You can filter features that are significant to the dataset in this way. In addition, selection can enhance machine learning efficacy by decreasing the dimensions of the data. Five features—time, serum creatinine, ejection fraction, age, and serum sodium—are the outcome of information gain. After that, predictions were made using logistic regression, and a data sharing ratio of 70% training data and 30% test data resulted in an accuracy of 0.8556. This demonstrates how feature selection with Information Gain can improve the accuracy of the logistic regression model and is a very effective method.

1. Introduction
Heart failure one of the symptoms associated with cardiovascular disease (CVD), namely gangguan in the heart and darah pulsation. [1]. Heart failure is a chronic illness caused by several factors that cause the heart's aliran to contract and lead to an inadequate supply of oxygen to the entire body to meet metabolic needs [2]–[4]. Heart failure is brought on by several disorders including weak heart muscle, irregular pulse, congenital heart abnormalities, and clogged coronary arteries. However, ventricular dysfunction is typically the cause of heart failure [5].

Cardiovascular disease is a major global cause of death. Internationally, some 26 million persons have a history of heart failure, according to the European Society of Cardiology (ESC) [5]. Meanwhile, in the United States, heart failure affects more than 5.8 million citizens and around 1 million residents are hospitalized in each year. [2]. Heart failure is a disease that is difficult to cure, as seen by the high fatality rate, which also suggests a low degree of patient safety.

The infrastructure and facilities that the hospital owns determine the degree of patient safety. In addition, the percentage of patient safety is also influenced by patient medical record data if the data is supplied accurately. Thus, we require a system that can offer data to forecast patient safety. Medical science currently depends on computer-based automated technologies to provide accurate and timely diagnosis [6]. Utilizing a technique from computer science called machine learning, one can prevent and stop patient deaths caused by computers.

There are several previous studies that have discussed heart failure using machine learning. Such as *Logistic Regression* [1], Naïve Bayes [2], Deep Learning [3], Improved Random Survival Forest [4], Neural Network [5] and Support Vector Machine [7]. The Logistic Regression approach will be employed in this study. When predicting the survival rate of patients with heart failure, logistic regression, which has been used by other researchers, can yield an accuracy of 83.8% [1]. Three features that were employed were produced by feature selection in this study.

Feature selection is commonly used with the aim of speeding up computing time, improving the performance of machine learning algorithms [8]–[10] and reducing the high dimensionality of data without reducing the information contained in the data [11]. Previous research in predicting the survival rate of heart failure patients used feature selection with the chi squared technique, random forest feature selection and stratified logistic regression. The results of the research were that the best accuracy was 83.8% with 3 features obtained, namely ejection fraction, serum creatinine, and follow-up time [1]. In this research the author will use the Information Gain (IG) technique as a feature selection method. Information Gain is used to select the best features that do not represent the data and leave important features [12]. Research on GI as a feature selection method is also being conducted to evaluate the suitability

of land for clove plantations [13], another is for feature selection in Indonesian text abstract documents [12], heart disease classification [14], and airline sentiment analysis [15].

The main purpose of this research is to design a logistic regression model with feature selection using information gain to predict the survival rate of patients with heart failure. The aim is to compare accuracy results based on the features used and compare them with the results of feature selection using information gain.

2. Research Method

2.1 Prior Research

Research conducted by D. Chicco and G. Jurman entitled "*Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone*" [1], This study employs a variety of feature selection techniques, including stratified logistic regression, random forest feature selection, and the chi squared technique. Additionally, researchers employ a variety of machine learning methods, including naïve Bayes, support vector machines (radial and linear), logistic regression, random forests, decision trees, gradient boosting, linear regression, one rule, artificial neural networks, and K-NN. Using a variety of feature selection and classification techniques, we were able to develop a feature selection method that used logistic regression (LR) as the classification and stratified logistic regression as the feature selection.

Table 1. Comparison results of log regression using full dataset and selected features

Method	Accuracy	*F score*	MCC	ROC Area
LR (EF, SR, dan *time*)	0.838	0.719	0.616	0.822
LR (semua fitur)	0.833	0.714	0.607	0.818

Table 1 suggests that the test data should be divided into two parts: 70% of the data should be training data, and the remaining 30% should be test data. Three parameters were used in the logistic regression analysis: ejection fraction (EF), serum creatinine (SR), and follow-up time (time). The accuracy findings were 83.8%. When compared to the dataset's classification accuracy of 83.3% when all features are used, this accuracy is higher (see table above). This demonstrates how feature selection can improve the accuracy of the logistic regression model evaluation by 0.005. Logistic regression with chosen characteristics yields F-score = 71.9%, MCC = 61.6%, and ROC = 82.2%; these results are also applicable for additional evaluation measures. This means that additional feature selection methods could be able to create prediction models that are even more accurate.

2.2 Literature Study

2.2.1 Logistic Regression (LR)

A type of regression used in machine learning for categorization is called logistic regression [16]. A mathematical model based on probability estimates for every class is applied by logistic regression. In this study, binary classification with two classes was accomplished using the logistic regression model. However, logistic regression can also be used with data that has more than one class in other scenarios [17]. The logistic regression formula can be seen in the Euation (1),

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i=1}^{N} \beta_i x_i \qquad (1)$$

Where π indicates probability of a whole class, β_o as bias or intercept, N is a numb of features, β_i is a regression coefficient that associated with the group, and x_i shows feature list.

2.2.2 Information Gain (IG)

Information gain in terms is the acquisition of information. Information gain is a simple technique that can be used to filter and select features. Information gain able to noise reduction in dataset resulting from the irrelevance of features [14]. Information gain is employed in the decision tree method to generate computations for the decision tree.[13]. Equation (2) displays the formula that was utilized to determine the information gain.

$$InfoGain(S, A) = Entropy\ (S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} Entropy\ (S_v) \qquad (2)$$

Where S is the nnumber of features / attributes, A is its classes, S_v is the number of samples of each v value, v to show probabilities for class in A, Si is a list of features in i and $Values\ (A)$ is the set of possible values of class A.

$$Entropy\ (S) = \sum_i^c -p_i \log_2 p_i \qquad (3)$$

Entropy in information gain is used to quantify class uncertainty with the likelihood of a given characteristic and to identify the optimal attribute [14]. Equation (3) shows the formula to calculate the entropy. where c is the number of values in each attribute for its class and p_i is the sample total in each class of i.

3. Results and Discussion

The research conducted by Chicco and Jurman provided the dataset used in this study [1]. There are 299 data in the collection, each with 13 features. There are two classes: class 1 and 0, where 0 denotes survival and class 1 indicates death. In total, 203 rows of data belong to the survivor class, whereas 96 rows are related to the death class. If the percentage is calculated, 31.11% of the data and 67.89% of the data, respectively, are in the safe class. Table 2 below shows details about the dataset. Nothing in the data values is missing.

Table 2. Dataset details

#	Features	Data range
1	Age	40, … , 95
2	Anaemia	0, 1
3	High blood pressure	0, 1
4	Creainine phospokinase	23, …, 7861
5	Diabetes	0, 1
6	Ejection fraction	14, … , 80
7	Sex	0, 1
8	Platelets	25.01 , … , 850
9	Serum creatinine	0.5, … , 9.4
10	Serum sodium	114, … , 148
11	Smoking	0, 1
12	Time	4 , … , 285
13	(target) Death event	0, 1

Based on the primary hypothesis of this study, which is the use of knowledge gain in feature selection. One may say that data with 12 features have enormous dimensions. The goal of feature selection is to minimize features and obtain only the most significant ones. In addition, it seeks to lengthen computation times and enhance the log regression algorithm's performance.

Out of the 12 features in the dataset, 5 significant features were found once information gain was applied. Table 3 displays the outcomes of the feature selection.

Table 3. Selected feature-based information gain rank

Rank	Features	Information Gain
1	Time	0.2867
2	Serum creatinine	0.1550
3	Ejection fraction	0.1018
4	Age	0.0494
5	Serum sodium	0.0429

The next step will be classification after feature selection. LR is the algorithm used for classification. The data will be divided according to three different ratios: 70% for training data and 30% for test data, 80% for training data and 20% for test data, and 90% for training data and 10% for test data. Cross validation will be performed using k = 10. The optimal LR model is obtained by splitting the data according to a specific ratio and doing comparisons. Table 4 displays the classification's outcomes. Table 4 illustrates that the optimal outcomes are obtained when 90% of the training data and 10% of the test data of the dataset are shared. Specifically, with the highest accuracy of 0.90 in comparison to the other results. But because the accuracy offered is great and the sample size evaluated is small—just thirty data—there is a chance that this model would overfit. Next, we will contrast the findings of this study with the classification performance results of 70% training data and 30% test data. The results are displayed in Table 5.

Table 4. Classification result with log regression and information gain

	Cross validation	70% 30%	80% 20%	90% 10%
Accuraccy	0.8327	0.8556	0.883	0.90
TPR	0.688	0.647	0.727	0.8
F score	0.725	0.772	0.821	0.842
MCC	0.607	0.699	0.75	0.772
ROC	0.88	0.896	0.931	0.96

Tabel 5. Comparison results between this research result with prio research

	LR (EF, SR, time)	LR + IG
Accuracy	0.838	0.8556
F score	0.719	0.772
MCC	0.616	0.699
ROC	0.822	0.869

Table 5 above suggests that the LR classification approach and IG feature selection can enhance performance. With an accuracy of 0.8556, the accrual increased by 1.76% in comparison to the findings of earlier studies. Time, serum creatinine, ejection fraction, age, and serum sodium are the five features that are used. Additionally, information gain with log regression yields better values, according to several evaluation metrics. demonstrated by ROC = 86.9%, MCC = 69.9%, and F-Score = 77.2%. Compared to log regression utilizing the three features that have been reported in earlier studies, this value is higher.

It is evident from the results that feature selection with information gain can lower data dimensions from high to low. Additionally, it has been demonstrated that information gain can preserve key aspects of the data while enhancing the effectiveness of classification algorithms. Because it employs four attributes from the dataset, this research consumes a little bit more computation time even though it contains more features than earlier research.

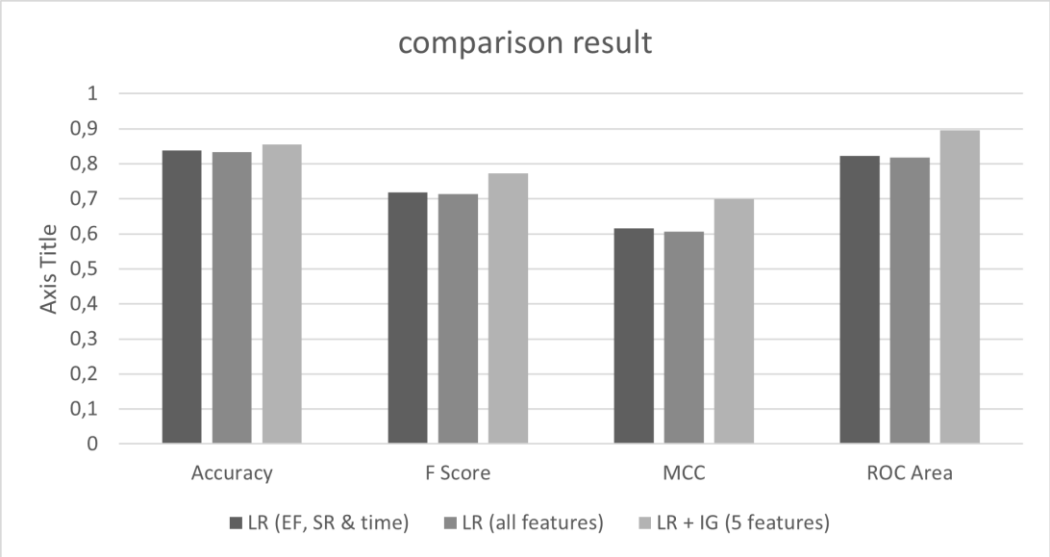

Figure 1. Overall comparison result

In Figure 1 above, the overall comparative findings are displayed. Observe the graph with the label log regression + information gain. The graph demonstrates that our suggested approach can outperform log regression with three characteristics and log regression with all features in terms of results. It has been demonstrated that the ROC area, MCC, accuracy, and F score all have higher graphic positions than alternative classification models in each measurement evaluation.

4. Conclusion

This study demonstrates that significant features in the data can be preserved while employing information gain as a feature selection method. It may also be stated that information acquisition raises the value of assessment measures and enhances log regression's classification performance. According to the study's findings, the ROC area

was 0.896, the accuracy was 0.8556, the F score was 0.772, and the mcc was 0.699. Five screening-derived parameters are used for classification: age, serum sodium, ejection fraction, time, and serum creatinine. Without omitting any crucial information, these 5 features accurately capture the entirety of the dataset. To determine which selection feature method is optimal, additional research could be conducted using a different approach.

References

[1] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–16, 2020, doi: 10.1186/s12911-020-1023-5.

[2] P. R. Degregory, J. Tapia, T. Wong, J. Villa, I. Richards, and R. M. Crooks, "Managing Heart Failure at Home with Point-of-Care Diagnostics," *IEEE J. Transl. Eng. Heal. Med.*, vol. 5, no. August, pp. 1–6, 2017, doi: 10.1109/JTEHM.2017.2740920.

[3] M. Gjoreski, A. Gradisek, B. Budna, M. Gams, and G. Poglajen, "Machine Learning and End-to-End Deep Learning for the Detection of Chronic Heart Failure from Heart Sounds," *IEEE Access*, vol. 8, pp. 20313–20324, 2020, doi: 10.1109/ACCESS.2020.2968900.

[4] F. Miao, Y. P. Cai, Y. X. Zhang, X. M. Fan, and Y. Li, "Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest," *IEEE Access*, vol. 6, pp. 7244–7253, 2018, doi: 10.1109/ACCESS.2018.2789898.

[5] B. Wang *et al.*, "A Multi-Task Neural Network Architecture for Renal Dysfunction Prediction in Heart Failure Patients with Electronic Health Records," *IEEE Access*, vol. 7, pp. 178392–178400, 2019, doi: 10.1109/ACCESS.2019.2956859.

[6] D. Derisma, "Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining," *J. Appl. Informatics Comput.*, vol. 4, no. 1, pp. 84–88, 2020, doi: 10.30871/jaic.v4i1.2152.

[7] G. G. N. Geweid and M. A. Abdallah, "A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique," *IEEE Access*, vol. 7, pp. 149595–149611, 2019, doi: 10.1109/ACCESS.2019.2945527.

[8] A. Harris and A. E. Mintaria, "Komparasi Information Gain , Gain Ratio , CFs-Bestfirst dan CFs-PSO Search Terhadap Performa Deteksi Anomali," vol. 5, pp. 332–343, 2021, doi: 10.30865/mib.v5i1.2258.

[9] I. made B. Adnyana, "Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa," *J. Sist. Dan Inform.*, vol. 13, pp. 72–76, 2019.

[10] S. J. Pasha and E. S. Mohamed, "Ensemble Gain Ratio Feature Selection (EGFS) Model with Machine Learning and Data Mining Algorithms for Disease Risk Prediction," *Proc. 5th Int. Conf. Inven. Comput. Technol. ICICT 2020*, pp. 590–596, 2020, doi: 10.1109/ICICT48043.2020.9112406.

[11] A. Ridok, N. Widodo, W. F. Mahmudy, and M. Rifa'i, "A hybrid feature selection on AIRS method for identifying breast cancer diseases," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 728–735, 2021, doi: 10.11591/ijece.v11i1.pp728-735.

[12] I. Maulida, A. Suyatno, H. Rahmania Hatta, and U. Mulawarman, "Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain," *JSM STMIK Mikroskil*, vol. 17, no. 2, pp. 249–258, 2016.

[13] D. A. Bimantoro and S. ' Uyun, "Pengaruh Penggunaan Information Gainuntuk Seleksi Fitur Citra Tanah Dalam Rangka Menilai Kesesuaian Lahan Pada Tanaman Cengkeh," *Jiska*, vol. 2, no. 1, pp. 42–52, 2017.

[14] A. A. Syafitri Hidayatul AA, Yuita Arum S, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 9, pp. 2546–2554, 2018.

[15] A. B. P. Negara, H. Muhardi, and I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes dan Seleksi Fitur Information Gain," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 3, p. 599, 2020, doi: 10.25126/jtiik.2020711947.

[16] M. Anshori, F. Mar'i, and F. A. Bachtiar, "Comparison of Machine Learning Methods for Android Malicious Software Classification based on System Call," *Proc. 2019 4th Int. Conf. Sustain. Inf. Eng. Technol. SIET 2019*, pp. 343–348, 2019, doi: 10.1109/SIET48054.2019.8985998.

[17] W. Książek, M. Gandor, and P. Pławiak, "Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma," *Comput. Biol. Med.*, vol. 134, p. 104431, 2021, doi: 10.1016/j.compbiomed.2021.104431.