International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 04, Issue 04, October 2023
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Performance Testing of KNN and Logistic Regression Algorithms in Classifying Heart Disease Susceptibility

1st Pujo Hari Saputro, 2nd Wahyuni Fithratul Zalmi, 3rd Rendy Syahputra
Information Technology, Faculty of Engineering
Sam Ratulangi University
Manado, Indonesia
[1]Pujoharisaputro@unsrat.ac.id, [2] wahyuni.fithratul.zalmi@unsrat.ac.id, [3] rendysyahputra@unsrat.ac.id

Abstract— The annual global death toll due to cardiovascular diseases, which fall into the category of heart and blood vessel disorders, reaches 17.9 million lives. This undoubtedly requires more attention in order to anticipate the potential risk of heart attacks that can affect anyone at any time. Data analysis or data mining approaches have become a significant contribution in the field of information technology to provide valuable information regarding the risk of heart diseases. Data analysis using the K-Nearest Neighbor and Logistic Regression algorithms is expected to provide information related to the susceptibility category for heart diseases, such as age susceptibility, gender, cholesterol levels, and so on. With the information obtained from this data analysis, it is hoped that it can serve as a reference and consideration for individuals to be more vigilant in maintaining their health. The results indicate that the highest correlation with susceptibility to heart disease is based on a person's age and their body weight. The correlation coefficient between these two variables is 0.37, suggesting a relationship between a person's age and their body weight, which can make them more susceptible to heart disease. Testing with both algorithms shows a high level of accuracy, with K-Nearest Neighbor achieving an accuracy rate of 0.95, while Logistic Regression has an accuracy of 0.96.

Keywords : *Comparison, Heart attack, K-Nearest Neighbor, Logistic Regression.*

I. INTRODUCTION

In 2021, the World Health Organization (WHO) released data indicating that 1 out of every deaths worldwide is caused by heart disease. In Indonesia itself, the leading causes of heart disease are mainly attributed to improper diet, obesity, lack of physical activity, and excessive tobacco consumption [1]. Lack of information about the factors that can lead to heart disease is one of the main reasons for the delay in preventing the disease. [2], [3].

Data analysis or data mining approaches have become one of the contributions in the field of information technology to provide valuable information regarding the risk of heart disease. [4]. Data analysis approaches are expected to provide information related to the susceptibility category for heart diseases, such as age susceptibility, gender, cholesterol levels, and so on. With the information obtained from this data analysis, it is hoped that it can serve as a reference and consideration for individuals to be more vigilant in maintaining their health..

Data processing/data mining is one of the expertise areas that currently receives attention in various fields. This is because proper data processing can yield valuable information in various sectors of society. Each algorithm and method will yield different levels of accuracy, which is a consideration in determining which algorithm or method has a higher level of correctness. Performance testing by comparing the accuracy levels on the same objects is one way to find out the accuracy of using an algorithm on predetermined objects.

Based on the description above, the author believes that it is indeed necessary to conduct research to identify potential factors that may lead to heart disease. This is, of course, closely related to the high mortality rate caused by this condition. Additionally, the use of classification methods to determine the factors causing heart disease should consider accuracy levels to provide a high level of confidence.

1.1. Theoretical Foundation

In this research, data mining and classification are used as methods in problem-solving. Data Mining is a process of discovering relationships or patterns from hundreds or thousands of fields in a large relational database. Data Mining is also often referred to as a series of processes to extract added value in the form of previously unknown information. Data Mining is primarily used to search for knowledge within large databases and is often referred to as Knowledge Discovery in Databases. [5], [6].

Classification in data mining is one of the primary tasks aimed at grouping or categorizing data into specific classes or categories based on the attributes possessed by the data. The main goal of classification is to build a model that can predict the category or class of unlabeled data based on patterns found in labeled data. [7].

1. Pearson Correlation

The Pearson correlation test is a statistical method used to measure the extent of the linear relationship between two continuous variables. This method produces the Pearson correlation coefficient (r), which measures the strength and direction of the relationship between the two variables.

The Pearson correlation coefficient (r) can be calculated using the following mathematical formula. [8] [9]:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Explanation of the formula :

X_i *and* Y_i : The values of both variables

$\bar{X}$: The average of variable X

$\bar{Y}$: The average of variable Y

2. K-Nearest Neighbor

K-Nearest Neighbors (KNN) is an algorithm in machine learning used for classification and regression tasks. This algorithm operates by finding a certain number of nearest neighbors (called "K") from an unlabeled data point and then performing classification or regression based on the majority or average of the labels of those neighbors [9] .

3. Logistic Regression

Logistic regression is one of the techniques in statistics and machine learning used for regression analysis in classification problems. Despite having the word "regression" in its name, logistic regression is actually used to classify data into two or more categories or classes based on a set of attributes or features. It is a very commonly used binary classification algorithm. Logistic regression models the probability that a sample of data belongs to one of the two possible categories or classes (usually referred to as the positive class and the negative class). Mathematically, the logistic regression model models the probability of the positive class (y = 1) as a function of the log-odds of independent variables (features or attributes) and model parameters. [10].

- Logit Function (Log-Odds):

  The logistic regression model calculates the log-odds (logit) of the probability of the positive class as follows:

  $$\text{Logit} \quad \big(P(y=1)\big) = \beta_0 + \beta_{1x1} + \beta_{2x2} + \cdots + \beta_{pxp}$$

  - logit(P(y=1)) is the log-odds value that a sample of data belongs to the positive class
  - $\beta_0, \beta_1, \ldots, \beta_p$ are the coefficients or model weights that need to be estimated during training..
  - $x_1, x_2, \ldots, x_p$ are the values of attributes (features) from the data sample..

- Sigmoid Function (Logistic Function):

  The log-odds values are transformed into probabilities using the sigmoid function (logistic function), which produces probability values between 0 and 1:

  $$\big(P(y=1)\big) = \frac{1}{1 + e^{-logit(P(y=1))}}$$

  Explanation:

  P(y=1) : the probability that a sample of data belongs to the positive class,

  e : Euler's number (mathematical constant).

  logit(P(y=1)): the log-odds value calculated in the previous step.

II. RESEARCH METHODS

This research begins with the collection of the dataset to be used. Once the dataset is obtained, the data will be processed using the Python programming language. The first step is to perform the Pearson correlation test to determine the correlation or the level of relationship between variables that are susceptible to heart disease. After the Pearson correlation test is completed, testing is conducted using the K-Nearest Neighbor and Logistic Regression algorithms to determine the accuracy level that can be achieved using both algorithms.

The classification process using both algorithms starts with training on the dataset, followed by testing on the data. From these two processes, conclusions can be drawn regarding the accuracy values of both algorithms, and they can be compared. This is further illustrated in Figure 1 below.

International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 04, Issue 04, October 2023
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

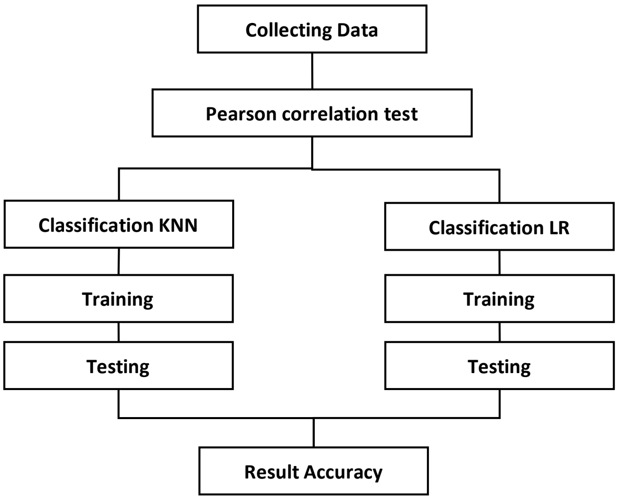

Figure 1. Research Flow

III. RESULT AND ANALYSIS

In this research, the dataset used can be shown in Figure 2, which contains attributes such as age, sex, smoking status, etc.

	gender	age	hypertension	heart_disease	ever_married	work_type
0	Male	67.0	0	1	Yes	Private
1	Male	80.0	0	1	Yes	Private
2	Female	49.0	0	0	Yes	Private
3	Female	79.0	1	0	Yes	Self-employed
4	Male	81.0	0	0	Yes	Private
...	...	...	...	...	...	...
4976	Male	41.0	0	0	No	Private
4977	Male	40.0	0	0	Yes	Private
4978	Female	45.0	1	0	Yes	Govt_job
4979	Male	40.0	0	0	Yes	Private
4980	Female	80.0	1	0	Yes	Private

4981 rows × 11 columns

Figure 2. Datasets Research part 1

work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Private	Urban	228.69	36.6	formerly smoked	1
Private	Rural	105.92	32.5	never smoked	1
Private	Urban	171.23	34.4	smokes	1
Self-employed	Rural	174.12	24.0	never smoked	1
Private	Urban	186.21	29.0	formerly smoked	1
...	...	...	...	...	...
Private	Rural	70.15	29.8	formerly smoked	0
Private	Urban	191.15	31.1	smokes	0
Govt_job	Rural	95.02	31.8	smokes	0
Private	Rural	83.94	30.0	smokes	0
Private	Urban	83.75	29.1	never smoked	0

Figure 3. Datasets Research part 2

3.1. *Pearson Correlation Test*

The Pearson correlation test starts with importing the required libraries, followed by preparing the dataset in the form of NumPy arrays. Next, the Pearson correlation and p-value are calculated, and from these calculations, the correlation coefficient is analyzed. The results of the Pearson correlation test in this research show a Pearson coefficient of 0.027 and a p-value of 0.05. The relationship between variables is illustrated in Figure 4.

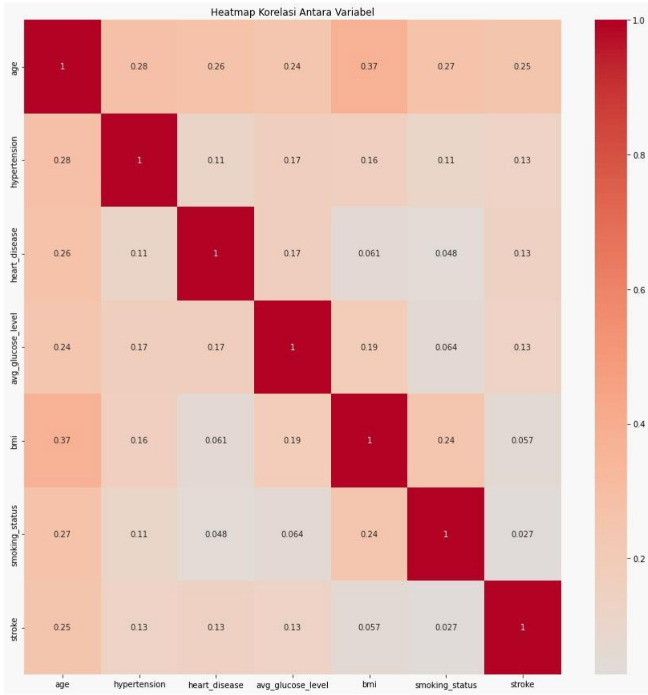

Figure 4. Correlation Between Variables.

From Figure 4, it can be observed that the highest correlation occurs between age and BMI (Body Mass Index), which is 0.37, indicating a relationship that can lead to the occurrence of heart disease. The second-highest correlation is between age and Hypertension with a correlation of 0.28, showing a connection between the two variables that is at risk of heart disease. Further details are shown in Table 1.

Table 1. Correlation Between Variables.

No	Variabel 1	Variabel 2	Correlation
1	Age	BMI	0,37
2	Age	Hypertension	0,28
3	Age	Smooking Status	0,27
4	Age	Heart Disease	0,26
5	Age	Stroke	0,25
6	Age	Avg Glucose Level	0,24

International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 04, Issue 04, October 2023
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

7	Hypertension	Heart Disease	0,11
8	Hypertension	Avg Glucose	0,17
9	Hypertension	BMI	0,16
10	Hypertension	Smooking Status	0,11
11	Hypertension	Stroke	0,13
12	Heart Disease	Avg Glucose Level	0,17
13	Heart Disease	BMI	0,061
14	Heart Disease	Smooking Status	0,048
15	Heart Disease	Stroke	0,13
16	Avg Glucose Level	BMI	0,19
17	Avg Glucose Level	Smooking Status	0,064
18	Avg Glucose Level	Stroke	0,13
19	BMI	Smooking Status	0,24
20	BMI	Stroke	0,057

The results from Table 1 indicate that the age variable has a significant influence on the risk of heart disease. From the correlation test results, it is found that there is a correlation between several variables indicating a risk of heart disease. However, the values obtained show that the correlation is at a low level. This may occur due to several reasons, one of which is the dataset used.

3.2. *K-Nearest Neighbor Test*

The application of K-Nearest Neighbor in this classification is carried out by taking 25% of the data for testing purposes. After testing, the results obtained are shown as follows.

```
              precision    recall  f1-score   support

           0       0.95      1.00      0.98      1187
           1       0.00      0.00      0.00        59

    accuracy                           0.95      1246
   macro avg       0.48      0.50      0.49      1246
weighted avg       0.91      0.95      0.93      1246
```

Figure 5. Classification Report KNN

Figure 5 is a report of the data processing results using K-nearest neighbor, which can be further simplified in Table 2.

Table 2. Detailed Results of K-Nearest Neighbor

No	Prosentase	Accuration	Precision	Recall
1	25	95%	0,91	0,95

3.3. *Logistic Regression Test*

The implementation of the Logistic Regression method in classification also uses a scenario using 25% of the data. The results of this scenario are as follows:

```
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      1187
           1       0.00      0.00      0.00        59

    accuracy                           0.96      1246
   macro avg       0.49      0.51      0.49      1246
weighted avg       0.92      0.97      0.95      1246
```

Figure 6. Classification Report Logistic Regression

Figure 6 is the report of the data processing results using Logistic Regression, which can be further simplified in Table 3.

Table 3. The detailed results of Logistic Regression.

No	Prosentase	Accuration	Precision	Recall
1	25	96%	0,92	0,97

3.4. Interpretation of the results.

Based on the conducted tests in this research, it was found that the variable with the highest percentage of susceptibility to heart disease is age combined with an unhealthy body weight. This indicates that individuals of a certain age who do not maintain an ideal body weight, based on this research, are more susceptible to heart disease.

In accordance with the scenarios conducted, which involved testing data using 25% of the dataset, an accuracy of 95% was obtained for K-Nearest Neighbor, while Logistic Regression, using the same scenario, achieved an accuracy of 96%. Therefore, it can be concluded from the conducted tests that the Logistic Regression method or algorithm has a higher level of accuracy compared to the K-Nearest Neighbor method or algorithm.

VI. CONCLUSION

Based on the research conducted, it can be concluded that age and BMI variables have the highest correlation that can affect someone's susceptibility to heart disease. Furthermore, the logistic regression method exhibits a higher level of accuracy compared to the KNN method. For a more comprehensive presentation, the author provides the following:

1. The best method based on the conducted tests is the Logistic Regression method with an accuracy of 96%. Meanwhile, the accuracy for the K-Nearest Neighbor method was found to be 0.95.

2. The highest risk factor for heart disease in this research is the correlation between age and BMI, with a correlation coefficient of 0.37. Based on this, it is known that there is a relationship between a person's age and their body weight, leading to the occurrence of heart disease.
3. It is worth noting that the correlation values obtained from the tests tend to be low with the dataset used. This suggests that future research should consider using different datasets or expanding the percentage of data used.

REFERENCES

[1] A. F. D. Zakha MaisatEka Darmawana, "Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM," *Teknologi: Jurnal Ilmiah Sistem Informasi,* pp. 8-15, 2023.

[2] A. F. P. Novanto Yudistira, "Algoritma Decision Tree Dan Smote Untuk Klasifikasi Serangan Jantung Miokarditis Yang Imbalance," *Jurnal Litbang Edusaintech,* vol. 2, nr 2, pp. 112-122, 2021.

[3] S. S. N. L. Z. M. &. C. X. N. Satish Chandra Reddy, "Classification and Feature Selection Approaches byMachine Learning Techniques: Heart Disease Prediction," *International JournalofInnovative Computing* , vol. 9, nr 1, pp. 39-46, 2019.

[4] M. P. a. S. Parija, "Prediction of Heart Diseases using Random Forest," *Journal of Physics: Conference Series,* vol. 1817, nr 1, p. 012009, 2021.

[5] N. T. S. T. a. B. T. Nguyen, "Intelligent Information and Database Systems," i *9th Asian Conference, ACIIDS 2017*, Kanazawa, Japan, 2017.

[6] B. W. E. a. D. T. Sherrill, "Analysis of Student Data for Retention Using Data Mining Techniques.," i *7th Annual National Symposium on Student Retention,* Oklahoma, 2011.

[7] Suyanto, Data Mining untuk klasifikasi dan klasterisasi data., Bandung: Informatika Bandung , 2017.

[8] J. C. Y. H. &. I. C. Jacob Benesty, Noise Reduction in Speech Processing : Pearson Correlation Coefficien, Springer, 2009.

[9] L. Wang, "Research and implementation of machine learning classifier based on KNN.," i *IOP Conference Series: Materials Science and Engineering,* IOP Publishing, 2019.

[10] R. a. K. P. Ewing, Basic quantitative research methods for urban planners, London: Routledge, 2020.