Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre
p-ISSN 2964-1691 | e-ISSN 2964-8246

Classification of risk of death from heart disease or cigarette influence using the k-nearest neighbors (KNN) method

Muhammad Syafiq Fadhilah[1], Rini Muzayanah[2]
[1,2] Department of Informatics Engineering, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received April 20, 2024
Revised July 2, 2024
Accepted July 2, 2024

Keywords:

Heart Disease
Risk of death
Classification
K-nearest neighbors
Confusion matrix

ABSTRACT

Heart disease is one of the leading causes of death in Indonesia. In addition to coronary heart disease, smoking is the leading contributor to the death rate in Indonesia. This study aims to analyze the risk of death with the main variables of heart disease history and smoking history. This study classifies the risk of death of heart disease sufferers and smokers using the KNearest Neighbors (KNN) algorithm. The results showed that the KNN model had an accuracy of 52.38% in predicting the risk of death of smokers and heart disease patients. Confusion matrix analysis revealed that the model performed well in predicting classes 0 and 2, but had difficulty in predicting class 1. This study shows that KNN can be used to predict the risk of death of smokers and patients with heart disease with a satisfactory success rate.

1. Introduction

The heart is one of the organs that plays an important role in the human circulatory system [1]. Heart disease is one of the many diseases that can cause death in Indonesia [2]. Heart disease is one of the most common and worrying health problems today, due to its various complications, including stroke, heart attack, retinopathy, etc. [3]. Therefore, this disease requires special attention so that the treatment process due to heart disease can be handled more quickly. According to the WHO, heart disease is one of the leading killer diseases in the world. In 2021, the WHO noted that up to 17.8 million people died or one in three people was

[1] ***Corresponding Author:***
Muhammad Syafiq Fadhillah,
Department of Informatics Engineering,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia.
Email: syafiqfadhilah@students.unnes.ac.id
DOI: https://doi.org/10.52465/josre.v2i2.359

caused by this heart disease. One of the well-known heart diseases is coronary heart disease. This disease is caused by blockage of blood flow to the heart and plaque build-up in the coronary arteries that supply oxygen to the heart muscle [4], resulting in severe damage to the heart. According to the WHO, up to 45% of the 9.4 million deaths were caused by coronary heart disease. Data will continue to increase to 23.3 million in 2030 [5]. The main cause is an unhealthy diet and lifestyle [6], such as smoking, an unhealthy diet, lack of exercise, and being overweight. In Indonesia alone, coronary heart disease is the leading cause of death. According to a survey conducted by the Sample Registration System, the number of deaths caused by coronary heart disease reached 12.9% [7], [5].

In addition to coronary heart disease, the biggest contributor to the death rate in Indonesia is smoking. Indonesians are very fond of cigarettes. According to the WHO, in 2022, cigarette consumption in Indonesia reached 36.5% or one in three Indonesians is a smoker. Most Indonesians smoke inside the house, which can affect other family members [8]. Smoking activity is an activity that has a negative impact on health and the surrounding environment [9]. The data resulted in Indonesia being ranked 7th with the number of smokers spread throughout the world in 2018 [10]. This is due to the many cigarette advertisements that are attractively packaged to attract young people and women are attracted [11].

Previous studies have discussed similar topics. Research by Mesquita & Marques in 2024 has discussed building a machine learning model to detect heart disease [12]. The research presents a comparison between various classifiers and parameter tuning techniques, providing all the details needed to replicate the experiments and help future researchers working in the field. From the research that has been done, it is known that the created machine learning model can be implemented in real life to detect heart disease.

The K-Nearest-Neighbor algorithm is an algorithm that is often used for classification tasks using machine learning. K-nearest-neighbors (KNN) is a supervised machine learning algorithm used in classification and regression problems [13]. KNN classifies unlabeled data by calculating the distance between each unlabeled data point and all other points in the data set [14]. Then assign each unlabeled data point to the most identically labeled data class by finding patterns in the data set [15].

There have been many studies discussing the detection of heart disease. However, no studies have addressed the classification of mortality rates based on smoking history and heart disease history. Doll and Hill first established smoking as an independent risk factor for mortality and also became a classic cohort epidemiological study due to its high-quality experimental design [16]. Therefore, this study aims to analyze the risk of death with the main variables of history of heart disease and smoking history. This study uses dummy data sets to train machine learning classification models. This research uses a classification method

using K-Nearest Neighbor [17]. K-Nearest Neighbor is an algorithm based on the proximity of the distance of one data with other data [18]. Where the results of the query instance will be classified based on the proximity of neighboring distances [19]. This method has several advantages, namely being fast and very effective when calculating complex data [20]. This algorithm uses flexible parameters, meaning that the parameters will increase with the amount of data. This algorithm is lazy learning, which means that this algorithm does not use training data points to create a model.

2. Method

Data Collection Technique

In the process of compiling this research, we use dataset collection techniques in the form of creating dummy data or not actual data made through Python language commands through the google colab platform. The data set contains age, gender, BMI, Blood Pressure, Smoking History, Heart Disease History, and mortality risk. The data set consists of 200 rows and 7 attributes with up to 20% of the dataset in the form of missing values. In this study, the data set used is a dummy data set created using libraries in Python.

Data Preprocessing

Next, the stage carried out in this research is data preprocessing. At this stage, we will perform data labeling and missing value handling.

Data labeling is a process in which string data is converted into numeric data. The labeling process is used because there are several syntaxes in the classification process that cannot read string data types or can only read numeric data types. Missing values are information that is not available for an object (case). Missing values occur because information for something about the object is not given, is difficult to find, or simply does not exist. It needs to be addressed because it can affect the performance of the model.

Splitting Data

After performing the data preprocessing stage, the next step is the data splitting stage. The data splitting stage is a stage where the data set will be separated into two parts, namely training data and test data. With a value of 8:2 or 80% for training and 20% for test.

Modeling

The last stage is the training of the model. At this stage, the data set will be modeled using the KNN algorithm classification method. The K-Nearest Neighbor (K-NN) is a classification algorithm that uses the value of k neighbors. The formula of K-NN is represented as follows:

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (1)$$

where:

x_1 = Training data

x_2 = Testing data

i = Variabel data

d = Distance

p = Dimention

The K-Nearest-Neighbors KNN algorithm will be applied to the provided data set.

3. Results and Discussion

Dataset

The dummy data set generated in this study has a total of 7 variables, namely age, sex, BMI, blood pressure, smoking history, history of heart disease and mortality risk. The target variable in this study is the mortality variable. The number of data rows created in this study was 200 data rows. The description of the data set in this study can be found in Table 1.

Tabel 1. Results from creating a dummy dataset

Age	Gender	BMI	Blood pressure	Smoking history	Heart disease history	Mortality risk
21	Female	27.064225	110.0	No	Yes	Moderate
48	Male	26.416357	114.0	Yes	Yes	High
56	Female	38.480094	111.0	No	Yes	High
45	Male	27.064225	97.0	Yes	Yes	High
....	...	....	....	....	....	....
61	Male	39.191554	159.0	Yes	No	High

Data Preprocessing

In the data pre-processing stage, the first thing to do is to do label encoding. Label encoding was performed on 4 variables, namely, sex, smoking history, history of janyung disease, and risk of death. In the gender variable, the value "Female" was changed to the value "1", while the value "Male" was changed to the value "0". Then in the variables of smoking history and heart disease history, the value "Yes" was changed to the value "1" and the value "No" was changed to the value "0". Unlike the previous three variables, the risk of death variable was changed into 3 new values, namely value "0" to replace the low risk of death, value "1" to replace

the medium risk of death, and value "2" to replace the high risk of death. An overview of the dataset after label encoding can be seen in Table 2.

Table 2. Results of label encoding

Age	Gender	BMI	Blood pressure	Smoking history	Heart disease history	Mortality risk
21	1	27.064225	110.0	0	1	1
48	0	26.416357	114.0	1	1	2
56	1	38.480094	111.0	0	1	2
45	0	27.064225	97.0	1	1	2
....	...	....	....	....	....	....
61	0	39.191554	159.0	1	0	2

The next stage is the missing value checking stage. The results of the missing value check can be seen in Figure 1.

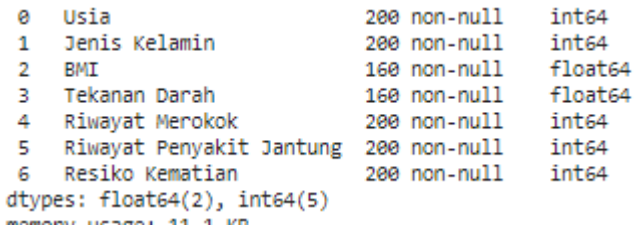

Figure 1. The result of checking for missing values in the data set.

From Figure 1, it is known in the BMI and blood pressure variables. To overcome this, the missing data are filled with the average of the values in each variable.

Model Evaluation

This section will discuss the results of data processing using the K-Nearest-Neighbor (K-NN) algorithm. When applying K-Nearest Neighbors (KNN) to classify the risk of death caused by smokers and heart disease patients, the classification results using the k value show the results in the confusion matrix which can be seen in Figure 2.

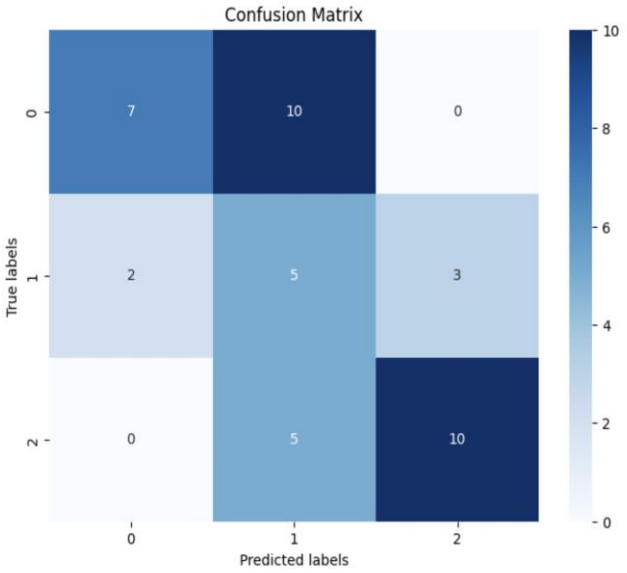

Figure 2. Classification results in the confusion matrix table

From the confusion results, an evaluation can be done using an evaluation matrix that includes precision, precision, recall, and F1 score. The evaluation results can be seen in Table 3.

Table 3. Results of performance matrix evaluation

Target	Precision	Recall	F1-Score	Accuracy
0	0.78	0.41	0.54	
1	0.25	0.50	0.33	0.52
2	0.77	0.67	0.71	

From the table above, it is known that the model can understand target 0 and target 2 quite well with precision, recall, and F1-Score values that are superior to target 1. However, the resulting accuracy results are not as good, which is around 52%. This is because the model has not been able to learn the dataset with target "1" maximally.

4. Conclusion

This research discusses how the risk of death of smokers and heart disease sufferers can be classified using the K-Nearest Neighbors (K-NN) algorithm. This study determines the optimal k-value to provide the best precision to predict the risk of death of smokers and heart disease sufferers using dummy data. Based on the results of its application, the value of k has an accuracy of 0.5238095238095238. This shows that the K-Nearest Neighbors (K-NN) algorithm can predict the risk of death of smokers and patients with heart disease with 52.38% success rate.

REFERENCES

[1] A. B. Wibisono and A. Fahrurozi, "PERBANDINGAN ALGORITMA KLASIFIKASI DALAM PENGKLASIFIKASIAN DATA PENYAKIT JANTUNG KORONER," *J. Ilm. Teknol. dan Rekayasa*, vol. 24, no. 3, pp. 161–170, Dec. 2019, doi: 10.35760/tr.2019.v24i3.2393.

[2] T. B. Anwar, "Dislipidemia Sebagai Faktor Resiko Penyakit Jantung Koroner," Universitas Sumatera Utara, 2004.

[3] A. Daza *et al.*, "Stacking ensemble based hyperparameters to diagnosing of heart disease: Future works," *Results Eng.*, vol. 21, p. 101894, Mar. 2024, doi: 10.1016/j.rineng.2024.101894.

[4] T. K. Ningsih and H. Zakaria, "Implementasi Algoritma K-Nearest Neighbor Pada Sistem Deteksi Penyakit Jantung Studi Kasus : Klinik Makmur Jaya," *Log. J. Ilmu Komput. Dan Pendidik.*, vol. 2, no. 1, 2023.

[5] L. Ghani, M. D. Susilawati, and H. Novriani, "Faktor Risiko Dominan Penyakit Jantung Koroner di Indonesia," *Bul. Penelit. Kesehat.*, vol. 44, no. 3, 2016.

[6] A. Samosir, M. S. Hasibuan, W. E. Justino, and T. Hariyono, "Komparasi Algoritma Random Forest, Naïve Bayes dan KNearest Neighbor Dalam klasifikasi Data Penyakit Jantung," *Pros. Semin. Nas. Darmajaya*, vol. 1, 2021.

[7] D. Pradana, M. Luthfi Alghifari, M. Farhan Juna, and D. Palaguna, "Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network," *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 55–60, Jul. 2022, doi: 10.56705/ijodas.v3i2.35.

[8] N. K. Noriani, I. W. G. A. E. Putra, and I. N. M. Karmaya, "Paparan Asap Rokok dalam Rumah Terhadap Risiko Peningkatan Kelahiran Bayi Prematur di Kota Denpasar," *Public Heal. Prev. Med. Arch.*, vol. 3, no. 1, pp. 55–59, Jul. 2015, doi: 10.15562/phpma.v3i1.88.

[9] K. N. Aziizah, I. Setiawan, and S. Lelyana, "Hubungan Tingkat Pengetahuan Tentang Dampak Rokok Terhadap Kesehatan Rongga Mulut dengan Tingkat Motivasi Berhenti Merokok pada Mahasiswa Universitas Kristen Maranatha," *SONDE (Sound Dent.*, vol. 3, no. 1, pp. 16–21, Jul. 2019, doi: 10.28932/sod.v3i1.1774.

[10] Ghany Vhiera Nizamie and A. Kautsar, "Analisis Faktor-Faktor Yang Mempengaruhi Konsumsi Rokok di Indonesia," *Kaji. Ekon. dan Keuang.*, vol. 5, no. 2, pp. 158–170, Nov. 2021, doi: 10.31685/kek.v5i2.1005.

[11] A. M. B. Arisani, Y. Hermawan, and N. Nurhadi, "Wanita dan Rokok (Studi Fenomenologi Dramaturgi Perilaku Merokok Mahasiswi Universitas Sebelas Maret)," *J. Pendidik. Tambusai*, vol. 7, no. 1, pp. 230–236, 2023, doi: https://doi.org/10.31004/jptam.v7i1.5284.

[12] F. Mesquita and G. Marques, "An explainable machine learning approach for automated medical decision support of heart disease," *Data Knowl. Eng.*, p. 102339, Jun. 2024, doi: 10.1016/j.datak.2024.102339.

[13] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification," 2003, pp. 986–996. doi: 10.1007/978-3-540-39964-3_62.

[14] A. Almomany, W. R. Ayyad, and A. Jarrah, "Optimized implementation of an improved KNN classification algorithm using Intel FPGA platform: Covid-19 case study," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3815–3827, Jun. 2022, doi: 10.1016/j.jksuci.2022.04.006.

[15] S. Ray, "A Quick Review of Machine Learning Algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, IEEE, Feb. 2019, pp. 35–39. doi: 10.1109/COMITCon.2019.8862451.

[16] R. Doll, R. Peto, K. Wheatley, R. Gray, and I. Sutherland, "Mortality in relation to smoking: 40 years' observations on male British doctors," *BMJ*, vol. 309, no. 6959, pp. 901–911, Oct. 1994, doi: 10.1136/bmj.309.6959.901.

[17] F. T. Admojo and Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 34–38, Jul. 2020, doi: 10.33096/ijodas.v1i2.12.

[18] Y. Pratama, A. Prayitno, D. Azrian, N. Aini, Y. Rizki, and E. Rasywir, "Klasifikasi Penyakit Gagal Jantung Menggunakan Algoritma K-Nearest Neighbor," *Bull. Comput. Sci. Res.*, vol. 3, no. 1, pp. 52–56, Dec. 2022, doi: 10.47065/bulletincsr.v3i1.203.

[19] I. P. Sari and I. H. Batubara, "Perancangan Sistem Informasi Laporan Keuangan Pada Apotek

Menggunakan Algoritma K-NN," *Semin. Nas. Teknol. Edukasi Sos. Dan Hum.*, vol. 1, no. 1, pp. 692–698, 2021, doi: https://doi.org/10.53695/sintesa.v1i1.398.

[20] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, Jul. 2020, doi: 10.33096/ijodas.v1i2.13.