IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 29, No. 4, August, 2024, pp. 1275-1281
Journal homepage: http://iieta.org/journals/isi

Heart Disease: Application of the K-Nearest Neighbor (KNN) Method

Diah Puspitasari[1*], Alief Juan Aprian[2], Erma Delima Sikumbang[1], Kresna Ramanda[1], Sulaeman Hadi Sukmana[1], Qudsiah Nur Azizah[1]

[1] Computer Science, Universitas Bina Sarana Informatika, Central Jakarta 10450, Indonesia
[2] Informatics Engineering, Universitas Ibn Khaldun Bogor, Bogor 16162, Indonesia

Corresponding Author Email: diah.puspitasari@bsi.ac.id

https://doi.org/10.18280/isi.290403

Received: 11 March 2024
Revised: 9 April 2024
Accepted: 2 June 2024
Available online: 21 August 2024

Keywords:
clustering, heart disease, blood sugar, K-Nearest Neighbor

ABSTRACT

The increasing prevalence of non-communicable diseases, especially heart disease, in Indonesia highlights the need to improve the effectiveness of medical procedures, especially for patients with heart disease. Data and information from the Ministry of Health of the Republic of Indonesia show that non-communicable diseases, especially heart disease, are the most prevalent diseases and account for 16% of all cases worldwide. The purpose of this study is to analyze the factors associated with the prevalence of heart disease in Indonesia, based on the frequency of gender, age, slope, blood sugar, and chest pain. This study uses the K-Nearest Neighbor algorithm method in data mining with several stages. Dataset, Preprocessing, and Clustering are stages that must be done in this research, this system is to capture patient information to be implemented. This challenge is applied by exploring some initial approaches to obtain the value of K. This is especially true for very large data. The results of this study can contribute to the understanding of data mining-based cardiac data analysis techniques using the K-Nearest Neighbor algorithm to improve the accuracy of diagnosing cardiac patients, with special attention to several types of frequencies that are key in this method. This study also obtained the results of the classification of heart disease datasets with 72.13% based on the maximum K value of KNN through the K-Nearest Neighbor clustering technique.

1. INTRODUCTION

Heart disease is a prevalent and serious health condition that affects a significant portion of the population. Various studies have been conducted to understand the epidemiology, risk factors, and implications of heart disease [1]. Heart disease can be caused by a wide range of factors, but major causes of morbidity and death, particularly in the elderly, include blood channel obstruction, inflammation, infection, and congenital abnormalities [1, 2].

Based on information from the Ministry of Health of the Republic of Indonesia's Data and Information Center, the country's projected 2016 population was 258,704,986, with 129,988,690 males and 128,716,296 women. Because there are more people in Indonesia between the ages of 0 and 14 than there are over 14, the country's total population is youthful [2, 3]. In the meantime, the percentage of people over 50 fell dramatically, as would be expected given the large middle-aged population's high death rate [4]. Non-communicable diseases account for the majority of Indonesia's mortality rate (NCDs). They harm and even kill a lot of individuals. The World Health Organization (2020) estimates that 16% of global fatalities are attributable to heart disease. The illness was the main cause of the growth since 2000, going from over 2 million fatalities to 8.9 million deaths in 2019 [5].

Thus, a mechanism must be created to ensure that medical records—especially those of patients with heart disease—are used as effectively as possible [6]. The system under consideration is a computer-based decision-making system that was created by taking relevant facts from a collection of data and reorganizing it into a different structure [7]. We refer to this system as clustering or data mining. Additional analysis may be conducted using the newly created structure produced by the data mining system [8].

High dimensionality or multi-attribute data is one of the clustering challenges. Because the data points are spread over many dimensions, the data is increasingly dispersed as the dimension grows [9]. Different data mining techniques, including naïve bayes, principal components analysis, decision trees, knearest neighbor, kernel density, bagging algorithms, and support vector machines, exhibit varying degrees of accuracy when it comes to diagnosing heart disease [10].

Three widely used data mining techniques, K-Nearest Neighbor, Naïve Bayes, and K-Means, were applied to a data set in this study. In addition, the results of integrating K-Nearest Neighbor with various initial values and number of clusters in the diagnosis of patients with heart disease are shown in the study by Singh and Rajesh [11]. One of the most popular clustering methods is K-Nearest Neighbor clustering, but the initial value problem is a crucial issue that has a great impact on the results. The use of several initial value approaches, including range, inlier, outlier, random attribute, random row, and random attribute methods, for the k-Nearest Neighbor methodology in the diagnosis of patients with heart disease is shown in this table [12]. The purpose of this study is to present a comprehensive analysis of clustering techniques using the K-Nearest Neighbor method with initial values that have the potential to improve diagnostic accuracy for cardiac patients.

2. METHOD

This study employs the frame of mind technique, which is conducted in phases to finish the research. The stages of this research method are shown in Figure 1 and are highly beneficial for researchers in their quest for the truth and a scientific knowledge of phenomena.

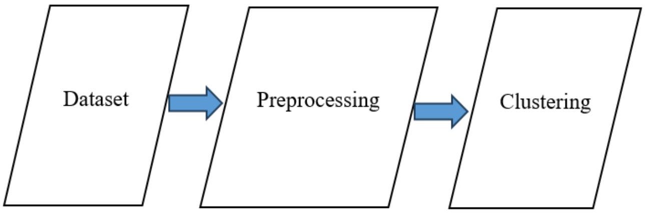


Figure 1. Research methods

2.1 Dataset

A dataset is basically a collection of data. Based on the definition from IBM, a dataset refers to a file that contains one or more records/data. In the data mining process, data alone is not enough. Structured data in the form of datasets is needed so that the results of data analysis will be faster, more accurate, efficient, and of high quality. Therefore, we need data on heart disease (Heart Disease) obtained through Kaggle data. The data will be processed and become an important source of information in this research shown in Figure 2.

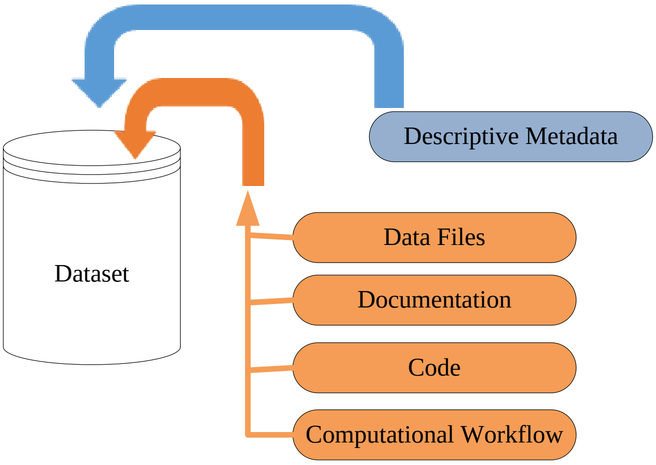


Figure 2. Visualization of datasets
(Source: https://guides.dataverse.org/en/4.6/user/dataset-management.html)

2.2 Preprocessing

Data mining is a sequence of procedures that may be broken down into several phases. Preparing and cleaning the data that is the subject of knowledge discovery in databases must first be cleaned before beginning the data mining process (KDD) [13]. Eliminating redundant data, examining inconsistent data, and fixing data errors such as printing mistakes are all part of the cleaning process [14]. Additionally, a procedure known as enrichment is carried out, which involves adding additional pertinent data or information such as necessary external data or information to the already-existing data in order to improve it for Knowledge Discovery in Databases (KDD). While the concepts in these phases differ, they are still connected to one another. Data mining is one of the stages that makes up the KDD process as a whole.

a) Selection of data before the Knowledge Discovery in Database (KDD) stage of information mining can start, a collection of operational data must be selected.

b) A file kept apart from the working database contains the chosen data that will be utilized for the data mining procedure.

The procedures involved in preprocessing are depicted in Figure 3 and include data cleansing, integration, transformation, reduction, and discretization.

- Data Cleaning: Data cleaning is the process of locating and fixing or eliminating inaccurate data from a dataset. It is also referred to as data cleansing, data scrubbing, or data cleansing. The cleansed data may contain defects that might impede the next steps in the data analysis process, such as being inconsistent, erroneous, duplicate, or mis formatted.

- Data Integration: The practice of merging data from several sources to provide a single, cohesive perspective is known as data integration.

- Data Transformation: The process of transforming data or information from one format to another is another definition of data transformation. For data processing, data analysis, and application development, data transformation is crucial.

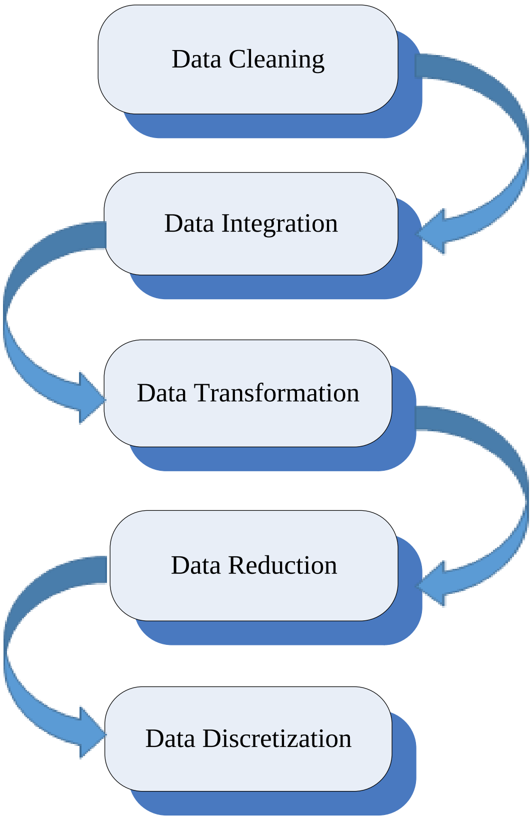


Figure 3. Data preprocessing steps

- Data Reduction: The goal of data reduction is to convert unprocessed data into a more focused, usable, and topic-appropriate format. Statistical tools or categories that facilitate data analysis for researchers might help in data reduction.
- Data Discretization: The process of breaking down a large number of data values into smaller ones to make managing and evaluating the data easier is known as data discretization. Put differently, data discretization refers to a technique that minimizes data loss while transforming the attribute values of continuous data into a limited collection of intervals.

3. CLUSTERING

According to Darmi and Setiawan, clustering [5]. Clustering, often known as classification, is a technique used to organize data into a few groups according to previously determined similarities. Clusters are collections of similar or different object data inside similar and dissimilar clusters relative to object-data that differ from one other. A subset of objects will be grouped into one or more clusters, so that objects inside one cluster will have a close relationship with one another [15]. The objects are grouped according to the principle of maximizing object similarity in a similar cluster and minimizing object dissimilarity in a different cluster. The object's similarity is usually derived from the attributes that describe the object, hence the object and object data are often represented as a set of variables in a multidimensional space [16]. By using this clustering technique, we may identify regions with low population densities, determine spatial distribution patterns, and identify significant correlations between data attributes. Within data mining, the focus is on finding effective and efficient clustering methods based on large-scale data. A few requirements for clustering in data mining are scalability, the ability to handle different attribute types that can handle high dimensionality, handling noisy data, and ease of handling [17]. The goal of data clustering, however, is to minimize the objective function that is selected throughout the clustering process, which is often concerned with minimizing the variance within a cluster [18]. And minimize inter-cluster variation. Broadly speaking, there are several methods for classifying data. The choice of clustering method depends on the type of data and the clustering goal itself. The clustering method used in this study is called K-Nearest Neighbor [19]. The final results of the previous steps in this research will yield an accuracy value using the K-Nearest Neighbor method.

4. RESULT

4.1 Dataset

The public dataset from Kaggle was used for this study; it contains text, numbers, photos, and even combinations of other kinds of data. In a specific context, this data may be utilized to find patterns, test hypotheses, and provide answers to problems. The acquired dataset contains information about heart disease. Table 1 has 303 CSV-formatted data about heart illness.

The data on heart disease, which includes age, gender, blood sugar, slope, and chest discomfort, is shown in the table below that was collected from Kaggle in CSV format.

Table 1. Heart disease data

No	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg
1	63	1	3	145	233	1	0
2	37	1	2	130	250	0	1
3	41	0	1	130	204	0	0
4	56	1	1	120	236	0	1
5	57	0	0	120	354	0	1
6	57	1	0	140	192	0	1
7	56	0	1	140	294	0	0
8	44	1	1	120	263	0	1
9	52	1	2	172	199	1	1
10	57	1	2	150	168	0	1
11	54	1	0	140	239	0	1
12	48	0	2	130	275	0	1
13	49	1	1	130	266	0	1
14	64	1	3	110	211	0	0
15	58	0	3	150	283	1	0
16	50	0	2	120	219	0	1
16	58	0	2	120	340	0	1
18	66	0	3	150	226	0	1
19	43	1	0	150	247	0	1
20	69	0	3	140	239	0	1
21	59	1	0	135	234	0	1
22	44	1	2	130	233	0	1
23	42	1	0	140	226	0	1
24	61	1	2	150	243	1	1
25	40	1	3	140	199	0	1
26	71	0	1	160	302	0	1
27	59	1	2	150	212	1	1
28	51	1	2	110	175	0	1
29	65	0	2	140	417	1	0
30	53	1	2	130	197	1	0
31	41	0	1	105	198	0	1
32	65	1	0	120	177	0	1
33	44	1	1	130	219	0	0
34	54	1	2	125	273	0	0
35	51	1	3	125	213	0	0
36	46	0	2	142	177	0	0
37	54	0	2	135	304	1	1
38	54	1	2	150	232	0	0
39	65	0	2	155	269	0	1
40	65	0	2	160	360	0	0
41	51	0	2	140	308	0	0
42	48	1	1	130	245	0	0
43	45	1	0	104	208	0	0
44	53	0	0	130	264	0	0
45	39	1	2	140	321	0	0
46	52	1	1	120	325	0	1
47	44	1	2	140	235	0	0
48	47	1	2	138	257	0	0
49	53	0	2	128	216	0	0
50	53	0	0	138	234	0	0
……							
303	57	0	1	130	236	0	0

4.2 Preprocessing

Text mining techniques and applications heavily depend on preprocessing procedures. It is the process's initial phase in text mining. The selection data and the selection data findings that will be processed for clustering are the two components of the selection data used in the study by Crone et al. [20].

Table 2 displays the first five rows of the Data Frame data to give an initial view of the dataset structure and counts the number of occurrences of each unique value in the 'target' column and displays the results. This data uses a public dataset on heart disease presented as a CSV file [21].

To normalize the data is done using Min-Max scaling; with the formula as given below, this will scale all numerical values in the range of 0 to 1.

Figure 4 provides an overview of the class distribution in the target column. Created a bar plot using Seaborn to display the class distribution in the 'target' column. On the x-axis, it has 'target'. This figure also presents the amount of data according to gender, where 0 is data on female gender which is obtained as much as 138 and 1 presents data with male gender as much as 165. In this data, it can be concluded that there are more males than females.

$$X_{normalized} = \frac{(X - X_{min})}{(X_{max} - X_{min})}$$

4.3 Clustering

Clustering or classification is a method used to divide a data set into groups based on predefined similarities [21]. This research uses the K-Nearest Neighbor Algorithm to classify the risk level of heart disease based on age and gender. Data sources as objects in this study is data taken from the Kaggle.com website. The data used in This study consists of attributes or variables such as gender, age, slope, blood sugar, and chest pain.

Figure 5 visualizes the distribution of count data based on the 'sex' column. This count plot provides information about the number of observations for each 'sex' category (0 for female, 1 for male). The plot does not display a legend and uses the 'mako_r' color palette. The x-axis shows the 'sex' category, and the y-axis shows the number of observations for each category. The program helps provide an initial understanding of the distribution of the 'sex' variable in the dataset.

Figure 6 provides useful information about the gender distribution, summary statistics related to the target, and visualization of the frequency of heart disease by age in the dataset. The figure displays a graph where the largest value of heart disease is in women based on age 58.

Table 2. Selection data

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Target
0	63	1	3	145	233	1	0	1
1	37	1	2	130	250	0	1	1
2	41	0	1	130	204	0	0	1
3	56	1	1	120	236	0	1	1
4	57	0	0	120	354	0	1	1

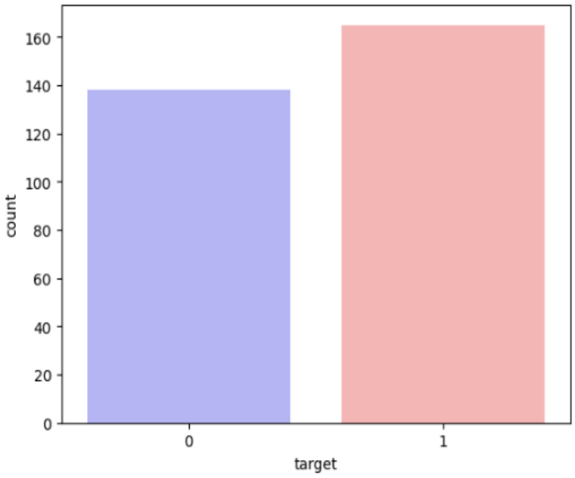


Figure 4. Result of selection data based on target

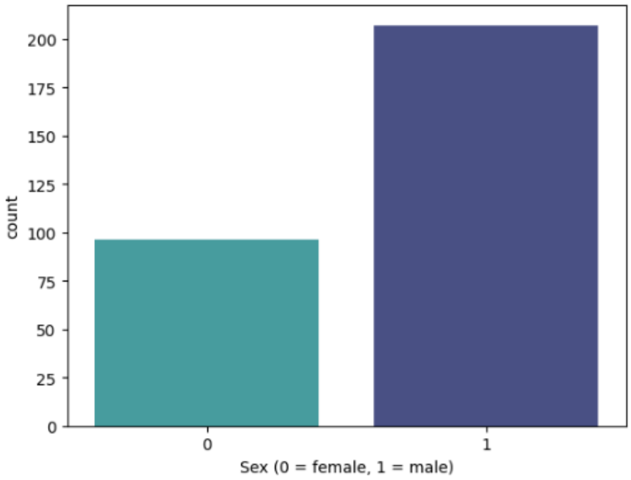


Figure 5. Count sex

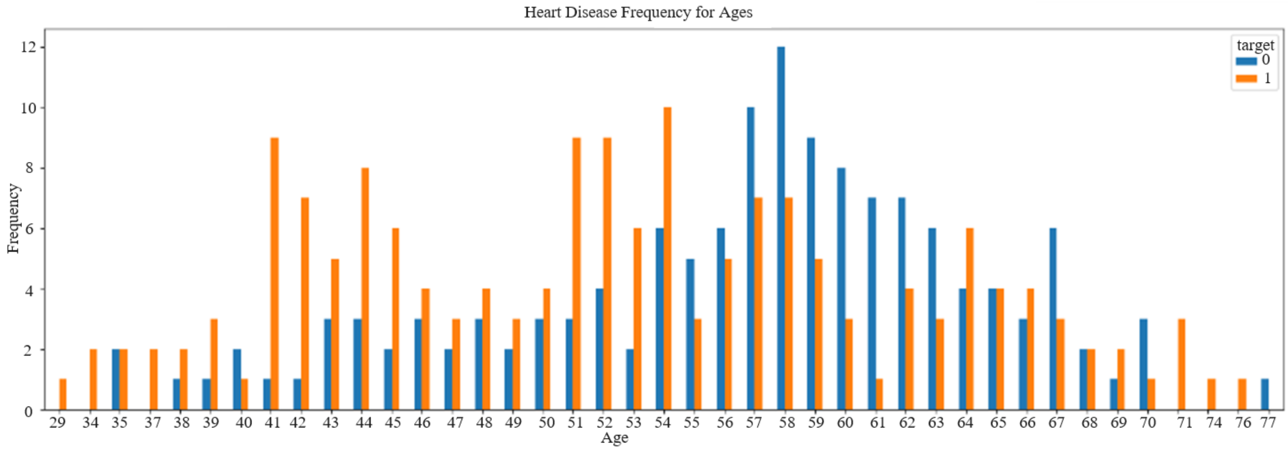


Figure 6. Frequency for age

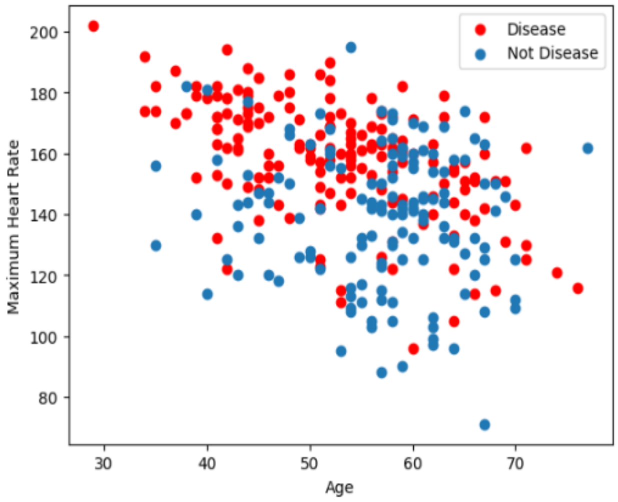


Figure 7. Maximum heart rate maximum heart rate for age

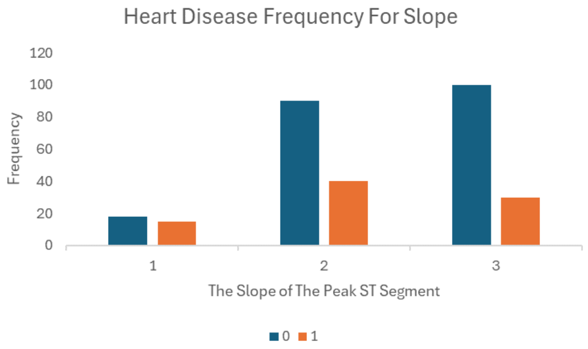


Figure 8. Heart disease frequency for slope

In Figure 7, this scatter plot helps visualize the distribution of age and maximum heart rate data for both patient groups. With the red and blue colors and the added legend, this plot makes it easier to interpret the data and understand possible patterns in the relationship between age and maximum heart rate based on heart disease status. The blue color (Not Disease) shows more of the plot.

The bar chart in Figure 8 shows the frequency of occurrence of combinations of values from the 'slope' and 'target' columns, with different colors for each target value. The relationship between the variable 'slope' (slope of the peak exercise ST segment) and the target variable ('target' which may indicate the presence or absence of heart disease). Through the bar chart, it is possible to see how often combinations of slope and target values occur, as well as a comparison of the frequency between different target values for each slope value. This helps in understanding the pattern of relationship between the two in understanding the pattern of relationship between the two variables.

The bar chart in Figure 9 visualizes the frequency of heart disease by Fasting Blood Sugar (FBS). It also clearly shows the distribution of the presence or absence of heart disease in groups with high and low FBS values. The differentiated colors make it easy to identify patterns, with the first color indicating those without heart disease and the second color indicating those with heart disease.

Figure 10 visualizes the frequency of presence or absence of heart disease based on chest pain type. Through this diagram, we can see the distribution of the number of heart disease cases for each type of chest pain. Different colors are used to distinguish between those with heart disease (first color) and those without heart disease (second color).

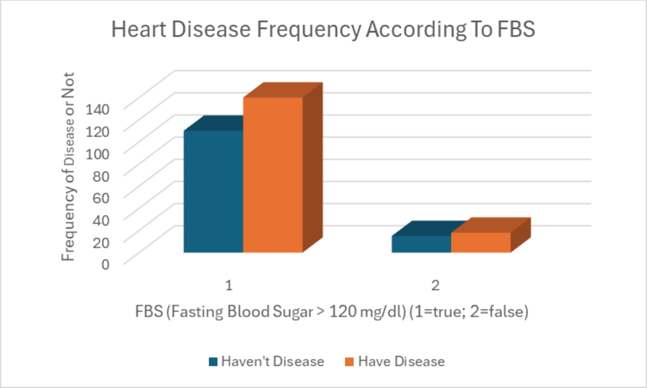


Figure 9. Heart disease frequency according to FBS

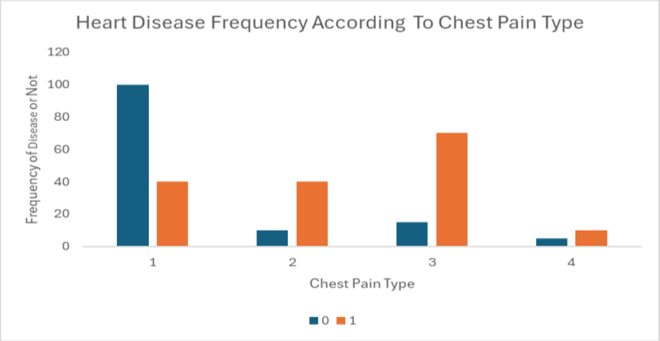
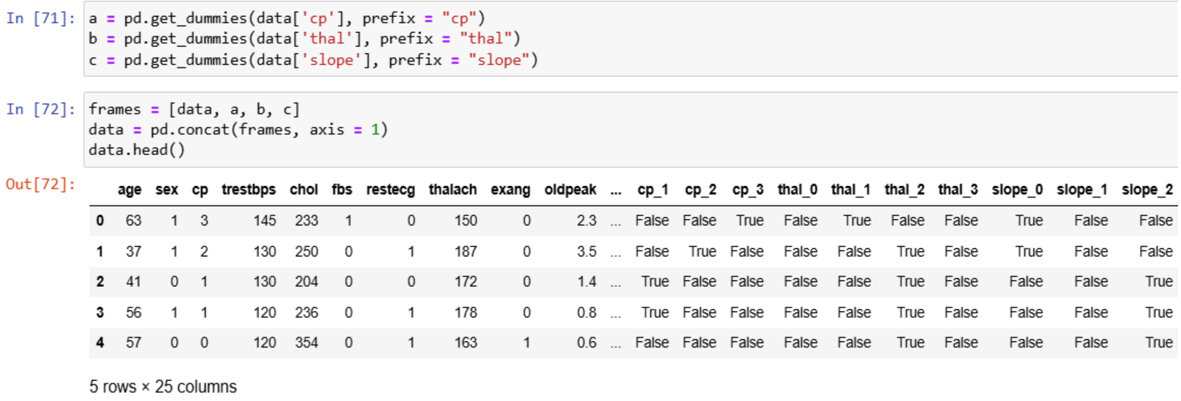


Figure 10. Heart disease frequency according to chest pain type



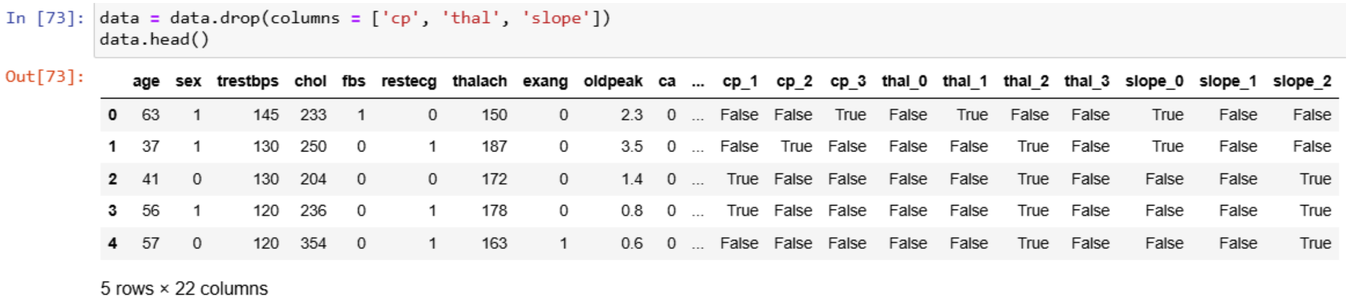
```
In [71]: a = pd.get_dummies(data['cp'], prefix = "cp")
         b = pd.get_dummies(data['thal'], prefix = "thal")
         c = pd.get_dummies(data['slope'], prefix = "slope")

In [72]: frames = [data, a, b, c]
         data = pd.concat(frames, axis = 1)
         data.head()

Out[72]:
   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  ...  cp_1   cp_2   cp_3   thal_0  thal_1  thal_2  thal_3  slope_0  slope_1  slope_2
0   63    1   3       145   233    1        0      150      0      2.3  ...  False  False  True   False   True    False   False   True     False    False
1   37    1   2       130   250    0        1      187      0      3.5  ...  False  True   False  False   False   True    False   True     False    False
2   41    0   1       130   204    0        0      172      0      1.4  ...  True   False  False  False   False   True    False   False    False    True
3   56    1   1       120   236    0        1      178      0      0.8  ...  True   False  False  False   False   True    False   False    False    True
4   57    0   0       120   354    0        1      163      1      0.6  ...  False  False  False  False   False   True    False   False    False    True

5 rows × 25 columns
```

Figure 11. Frames data categories 1



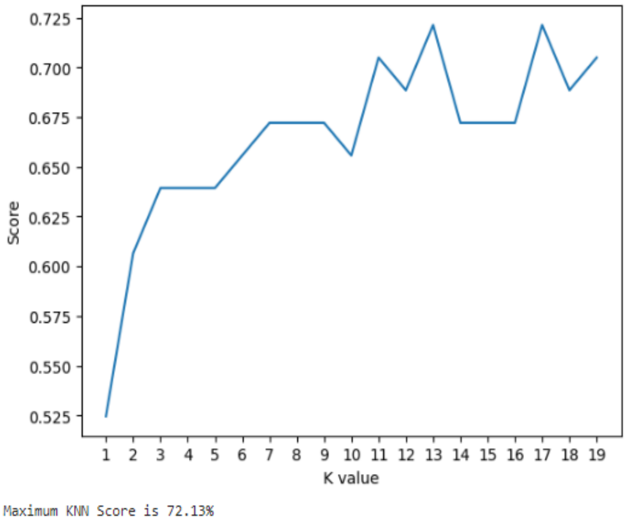
```
In [73]: data = data.drop(columns = ['cp', 'thal', 'slope'])
         data.head()
```

Out[73]:

	age	sex	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	ca	...	cp_1	cp_2	cp_3	thal_0	thal_1	thal_2	thal_3	slope_0	slope_1	slope_2
0	63	1	145	233	1	0	150	0	2.3	0	...	False	False	True	False	True	False	False	True	False	False
1	37	1	130	250	0	1	187	0	3.5	0	...	False	True	False	False	False	True	False	True	False	False
2	41	0	130	204	0	0	172	0	1.4	0	...	True	False	False	False	False	True	False	False	False	True
3	56	1	120	236	0	1	178	0	0.8	0	...	True	False	False	False	False	True	False	False	False	True
4	57	0	120	354	0	1	163	1	0.6	0	...	False	False	False	False	False	True	False	False	False	True

5 rows × 22 columns

Figure 12. Frames data categories 2



```
Maximum KNN Score is 72.13%
```

Figure 13. K-Nearest Neighbor (KNN) value

This program converts the categorical columns in the dataframe into dummy variables using pd.get_dummies (Figure 11), which converts each unique category into a column that is aligned with the current awalan. After transforming the columns "cp," "thal," and "slope" into dummies, the results are merged back into the original dataframe. This step aims to smooth the data so that it is more compatible with machine learning algorithms, which often need numerical input. Essentially, this tool preprocesses data to convert categorical columns into a format that may be used for further in-depth analysis or machine learning models.

The next step is to delete the original columns "cp," "thal," and "slope" from the dataframe data using data.drop(columns = ['cp', 'thal','slope']) (Figure 12). This is done after transforming the categorical columns "cp," "thal," and "slope" into dummy variables and merging them back with the original dataframe. This guarantees that the dataframe keeps only the created fake columns and the remaining, unaltered columns. Because the initial category columns are no longer redundant, the resulting dataframe is clearer and prepared for machine learning analysis or modeling. To confirm that the column deletion was effective and the intended data structure was achieved, the final result displays the first five rows of the dataframe that have been processed.

In Figure 13 showing the visualization results, the program looks for the K value that gives the highest accuracy on the test data. The maximum accuracy result along with the corresponding K value is then printed. Overall, this graph helps in determining the optimal parameters (number of neighbors) for the KNN model being built, with the aim of improving the performance and generalization of the model on data that has never been seen before [22]. This graph is the end result of several previous steps by displaying this plot can select the k value that gives the highest accuracy score to the KNN model, helping in decision making related to hyperparameter settings.

5. CONCLUSIONS

This study examines the clustering algorithm's performance inference using the K-Nearest Neighbor approach on a dataset of heart disease cases. We determined the performance classification results for five distinct types of frequencies in order to illustrate and evaluate the suggested classification technique. According to the experimental results, the frequency of heart disease based on gender is predominantly female; the age-based frequency is 58 years old; the slope at target 1 is higher based on frequency type; the frequency type based on blood sugar increases and leads to heart disease; and the frequency type based on chest pain is more prevalent among women. The K-Nearest Neighbor dimension reduction is used in this clustering method's performance, and it produces a final result graph to visualize the data in a higher-dimensional dataset in the clustering problem. Using the K-Nearest Neighbor clustering approach, we were able to accurately classify the heart disease dataset with 72.13% based on the K-value maximum KNN score.

REFERENCES

[1] Wartman, W.B., Hellerstein, H.K. (1948). The incidence of heart disease in 2,000 consecutive autopsies. Annals of Internal Medicine, 28(1): 41-65. https://doi.org/10.7326/0003-4819-28-1-41

[2] Purnamasari, D. (2019). The emergence of non-communicable disease in Indonesia. Acta Medica Indonesiana, 50(4): 273-274.

[3] Kushwah, S., Sharma, N., Das, S. (2022). Novel E-Focused crawler and enhanced k-mean (n-gram) clustering technique for automatic classification of attribute level customer healthcare sentiments. Journal of Algebraic Statistics, 13(2): 68-94. https://publishoa.com/index.php/journal/article/view/141/129

[4] Shouman, M., Turner, T., Stocker, R. (2013). Integrating clustering with different data mining techniques in the diagnosis of heart disease. Journal Computer Science and Engineering, 20(1): 1-10.

[5] Darmi, Y.D., Setiawan, A. (2016). Penerapan metode clustering k-means dalam pengelompokan penjualan

produk. Jurnal Media Infotama, 12(2): 148-157. https://doi.org/10.37676/jmi.v12i2.418

[6] Thangamani, M., Vijayalakshmi, R., Ganthimathi, M., Ranjitha, M., Malarkodi, P., Nallusamy, S. (2020). Efficient classification of heart disease using K-Means clustering algorithm. International Journal of Engineering Trends and Technology, 68(12): 48-53. https://doi.org/10.14445/22315381/IJETT-V68I12P209

[7] Ziasabounchi, N., Askerzade, I.N. (2014). A comparative study of heart disease prediction based on principal component analysis and clustering methods. Turkish Journal of Mathematics and Computer Science (TJMCS), 16: 18.

[8] Mirmozaffari, M., Alinezhad, A., Gilanpour, A. (2017). Heart disease prediction with data mining clustering algorithms. Int'l Journal of Computing, Communications & Instrumentation Engg, 4(1): 16-19. https://doi.org/10.15242/ijccie.dir1116009

[9] SundarV, B., Devi, T., Saravanan, N. (2012). Development of a data clustering algorithm for predicting heart. International Journal of Computer Applications, 48(7): 8-13. https://doi.org/10.5120/7358-0095

[10] Irwansyah, E., Pratama, E.S., Ohyver, M. (2020). Clustering of cardiovascular disease patients using data mining techniques with principal component analysis and K-medoids. Preprints 2020, 2020080074. https://doi.org/10.20944/preprints202008.0074.v1

[11] Singh, R., Rajesh, E. (2019). Prediction of heart disease by clustering and classification techniques prediction of heart disease by clustering and classification techniques. International Journal of Computer Sciences and Engineering, 7(2): 861-866. https://doi.org/10.26438/ijcse/v7i5.861866

[12] Ripan, R.C., Sarker, I.H., Hossain, S.M.M., Anwar, M.M., Nowrozy, R., Hoque, M.M., Furhad, M.H. (2021). A data-driven heart disease prediction model through K-means clustering-based anomaly detection. SN Computer Science, 2(2): 112. https://doi.org/10.1007/s42979-021-00518-7

[13] Roostaee, S., Ghaffary, H.R. (2016). Diagnosis of heart disease based on meta heuristic algorithms and clustering methods. Journal of Electrical and Computer Engineering Innovations (JECEI), 4(2): 105-110. https://doi.org/10.22061/jecei.2016.570

[14] Gárate-Escamila, A.K., El Hassani, A.H., Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19: 100330. https://doi.org/10.1016/j.imu.2020.100330

[15] Berger, L.A. (1998). Imaging in the diagnosis of infections in immunocompromised patients. Current Opinion in Infectious Diseases, 11(4): 431-436. https://doi.org/10.1097/00001432-199808000-00006

[16] Mirkin, B. (2005). Clustering for data mining: A data recovery approach. Chapman and Hall/CRC. https://doi.org/10.1201/9781420034912

[17] Fu, Z., Hu, W., Tan, T. (2005). Similarity based vehicle trajectory clustering and anomaly detection. In IEEE International Conference on Image Processing 2005, Genova, Italy, pp. 602-605. https://doi.org/10.1109/ICIP.2005.1530127

[18] Aichelin, U. (2022). Book selection. Libr. Journal of the Operational Research Society, 57(3): 332-335. https://doi.org/10.1057/palgrave.jors.2602163

[19] Jabbar, M.A., Deekshatulu, B.L., Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technology, 10: 85-94. https://doi.org/10.1016/j.protcy.2013.12.340

[20] Crone, S.F., Lessmann, S., Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. European Journal of Operational Research, 173(3): 781-800. https://doi.org/10.1016/j.ejor.2005.07.023

[21] Omran, M.G., Engelbrecht, A.P., Salman, A. (2007). An overview of clustering methods. Intelligent Data Analysis, 11(6): 583-605. https://doi.org/10.3233/ida-2007-11602

[22] Song, Y., Huang, J., Zhou, D., Zha, H., Giles, C.L. (2007). Iknn: Informative k-nearest neighbor pattern classification. In 11th European Conference on Principles of Data Mining and Knowledge Discovery in Databases, Warsaw, Poland, pp. 248-264. https://doi.org/10.1007/978-3-540-74976-9_25