
An Approach for Early Heart Attack Prediction Systems Using K-Means Clustering and Cosine Similarity

Nanda Novita¹, Amir Saleh^{2*}, Fadhillah Azmi³

nandanovita@staff.uma.ac.id¹, amirsalehnst1990@gmail.com², azmi.fadhillah007@gmail.com³

¹Informatics Departement, Universitas Medan Area, Medan, Indonesia

²Informatics Departement, Universitas Prima Indonesia, Medan, Indonesia

³Electrical Departement, Universitas Medan Area, Medan, Indonesia

Article Information

Submitted : 1 Aug 2023

Reviewed: 5 Aug 2023

Accepted : 21 Aug 2023

Keywords

K-Means Clustering,
Cosine similarity, Heart
attacks, Prediction
system, Clinical
characteristics

Abstract

In this study, we used cosine similarity and k-means clustering to construct a system to predict heart attacks. In order to divide patient data into groups with distinct clinical profiles based on their clinical characteristics, the k-means clustering approach is used. The new patient profiles were also contrasted with predetermined risk group profiles using the cosine similarity method. Heart attack high-risk patients are those with a profile that resembles that of the high-risk category. This suggested prediction system offers numerous benefits and contributions. First, the technique helps identify individuals who are at high risk of having a heart attack, allowing for prompt intervention and treatment. Second, the technology aids in lowering the mortality and effects of a heart attack by foreseeing the possibility of one in high-risk patients. Combining the k-means clustering method and cosine similarity, this system can predict heart attacks with an accuracy and dependability of 93.71%. In order to aid medical practitioners in making wise decisions and enhancing patient care, this research offers fresh perspectives on how to understand and manage heart attacks.

A. Introduction

A heart attack is a life-threatening condition that often cannot be predicted with accuracy. In an effort to prevent heart attacks and reduce the death rate caused by this condition, early analysis and prediction are very important [1]. By identifying risk factors and classifying individuals into appropriate risk groups, we can take appropriate preventive measures and provide the necessary treatment more effectively. Early detection and prediction of heart attacks are very important in preventing and treating this disease [2][3]. So, we need a system that can handle predicting this disease as a quick treatment effort.

Machine learning advancements in recent years have created new prospects for identifying risk variables and very accurate heart attack prediction [4]. Several studies that have tested using machine learning techniques to predict heart disease have yielded good results. The use of the classification method with random forest can be applied to the prediction of heart disease and gives good results of 83% [5]. Other research tries to compare various machine learning techniques and obtains the highest accuracy of 87.28% by applying the multilayer perceptron method [6]. In addition, a simple method using k-NN can also be applied and gives accurate results of 86.89% [7]. Based on the research that has been done, machine learning techniques can be used to predict heart disease with good accuracy.

Rapid diagnosis, early treatment, and ongoing monitoring are necessary for patients with heart disease. Numerous methods have been employed in the detection and prognosis of cardiac disease in order to address these needs [8]. The aim of this research is to develop a system for analysing and predicting heart attacks. The approaches used in this study are k-means clustering and cosine similarity methods. Combining methods can be used to group them and provide more accurate prediction results than using only one method. Research combining the k-means clustering method with neural networks provides more accuracy than using conventional methods, such as kNN and Naïve Bayes [9]. In this study, the k-means clustering method will be used to group individuals based on relevant clinical features, while cosine similarity will be used to compare individual profiles with predetermined risk group profiles. Through this approach, we can identify patterns and similarities among individuals at high risk of heart attack.

The methodology in this study is divided into several stages to be able to produce a good prediction system. First, relevant clinical data such as age, sex, medical history, and other risk factors will be collected. Furthermore, the k-means clustering method will be applied to classify the results of previous patient examinations into several groups based on their clinical characteristics. Then, using cosine similarity, the results of each new patient's examination will be compared with the predetermined risk group profile obtained from the k-means clustering method. Based on the results of this comparison, these new patients can be classified into the appropriate risk groups.

The heart attack prediction and analysis system proposed in this study is expected to provide several contributions and benefits. First, by identifying high-risk groups, we can provide heart disease patients with more intensive attention and more effective care. Second, by predicting the potential for heart attack in high-risk patients, we can take the necessary precautions to reduce the likelihood

of a heart attack occurring. Third, by using k-means clustering and cosine similarity methods, we can explore patterns and relationships that may not be seen directly and can provide new insights in understanding and treatment for predicting heart disease. Thus, it is hoped that this research will make a valuable contribution to efforts to prevent and treat heart attacks more effectively.

B. Research Method

The methodology for developing the heart attack analysis and prediction system using the proposed method can be seen in Figure 1 below.

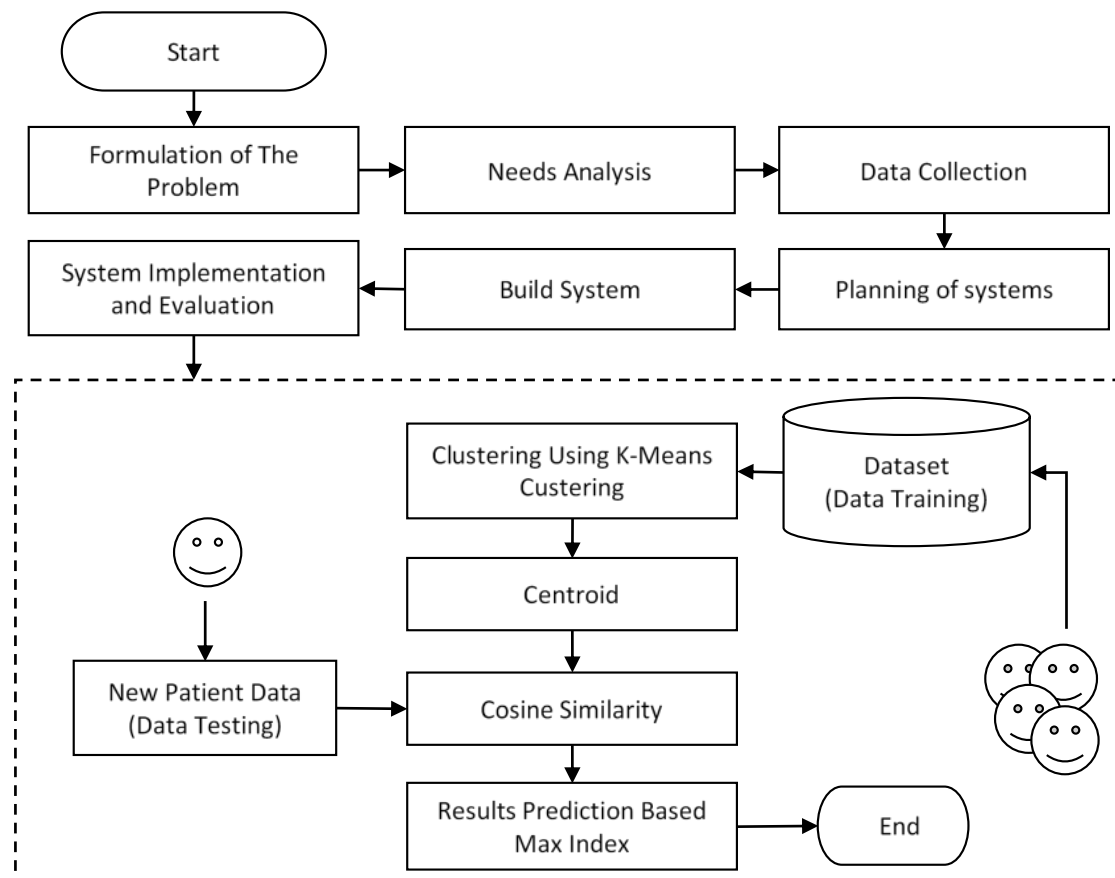


Figure 1. The proposed method to predict heart disease

The explanation of each step carried out in this study can be described as follows:

1. Formulation of the Problem

Heart disease has many risk factors and variables that are complexly interrelated. In understanding the relationship between these factors, a model is made that can be applied to the analysis and prediction of heart disease. Various methods of analysis, such as clustering or prediction algorithms, have their own advantages and disadvantages. The selection of the right method for effectively combining several parameters becomes important. The results of the analysis and

predictions can provide the right intervention, which can be a challenge in the development of the system.

2. Need Analysis

In developing a heart disease prediction system using the k-means clustering and cosine similarity methods, there are several system requirements that need to be considered, namely:

1. Complete clinical data: Access to complete and structured clinical data about the patient is required, including medical history, diagnostic test results, risk factors, symptoms, and other relevant information. This data should include groups of patients who had a heart attack and patients without a heart attack.
2. Selection of relevant features: Appropriate selection of features is very important in predicting heart disease. Features such as age, sex, blood pressure, cholesterol level, and other risk factors must be carefully selected to include those that have the most influence on the prediction of a heart attack.
3. Implementation of the k-means clustering algorithm: A good understanding and implementation of the k-means clustering algorithm are required. Steps such as initial centroid initialization, selecting the optimal number of clusters, and iterative convergence must be considered to obtain good clustering results.
4. Formation of risk group clusters: It is necessary to establish risk group clusters that reflect the characteristics of patients with a high risk of heart attack. This process involves an in-depth analysis of the clinical data of patients with heart attacks and the identification of significant risk factors that can form the basis of risk group profiles.
5. Development of the cosine similarity method: The cosine similarity method is used to compare the similarity between new patient profiles and risk group profiles to obtain a decision as a result of detection.
6. Model Evaluation: Evaluate the model used to obtain performance results such as the accuracy of heart disease prediction.

The proposed method can be used to construct a heart disease prediction system by satisfying these requirements. This approach can help with heart disease prevention, management, and better treatment decision-making.

3. Data Collection

The data used in this study were obtained from the Kaggle dataset. Age, gender, blood pressure, cholesterol levels, smoking history, family history of diabetes, family history of heart disease, and other risk factors could all be included in this data. The study and forecasting of heart attacks will be based on this data, with a total of 302 patients. Table 1 below provides a description of the dataset that was used.

Table 1. The Description Dataset

Name of Parameters	Description	Value
--------------------	-------------	-------

Age	Age of the patient	29 – 77
Sex	Sex of the patient	1: Male; 0: Female
exang	Exercise induced angina	1: Yes; 0: No
ca	Number of major vessels	0-3
cp	Chest Pain type	1: typical angina; 2: atypical angina; 3: non-anginal pain; 4: asymptomatic
trtbps	Resting blood pressure (in mm Hg)	94 – 200
chol	Cholesterol in mg/dl fetched via BMI sensor	126 – 564
fbs	Fasting blood sugar > 120 mg/dl	1: True; 0: False
rest_ecg	Resting electrocardiographic results	0: normal; 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	Maximum heart rate achieved	71 – 202
target	Chance of heart attack	0: less chance of heart attack; 1: more chance of heart attack

4. System Planning

a. Use Case Diagram

In this study, use case diagrams are utilised to show how the system is used by the administrator. Figure 2 below shows a few of the administrator's general activities.

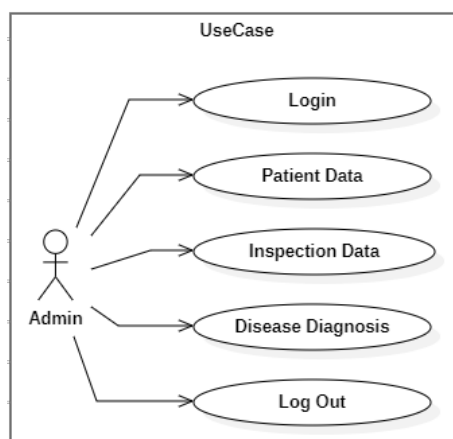


Figure 2. Use Case Diagram to Predict Heart Disease

b. Activity Diagram

An activity diagram is one type of diagram that is used to describe the flow or sequence of activities in a process. In a heart disease prediction system, activity

diagrams can be used to describe the steps or activities involved in the heart disease prediction process. The activity diagram in this study can be described in Figure 3 below.

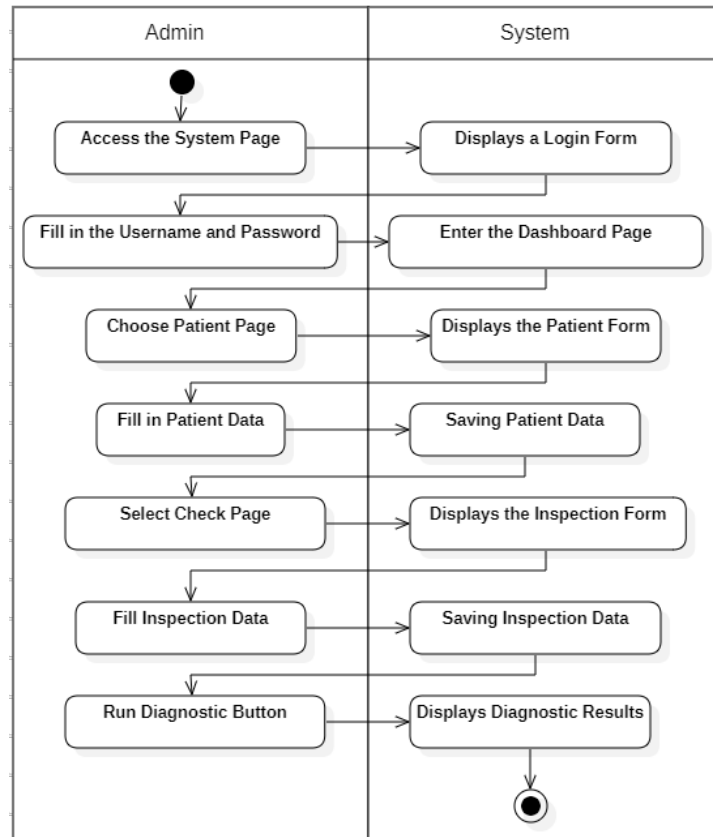


Figure 3. Activity Diagram to Predict Heart Disease

c. Entity Relationship Diagram

ERD is a type of diagram that is used to describe the relationship between entities in the database as well as important sub-tasks in the extraction of the required information [10]. In a heart disease prediction system, several main entities in ERD will be related. The database schema for a heart disease detection system can be seen in Figure 4 below.

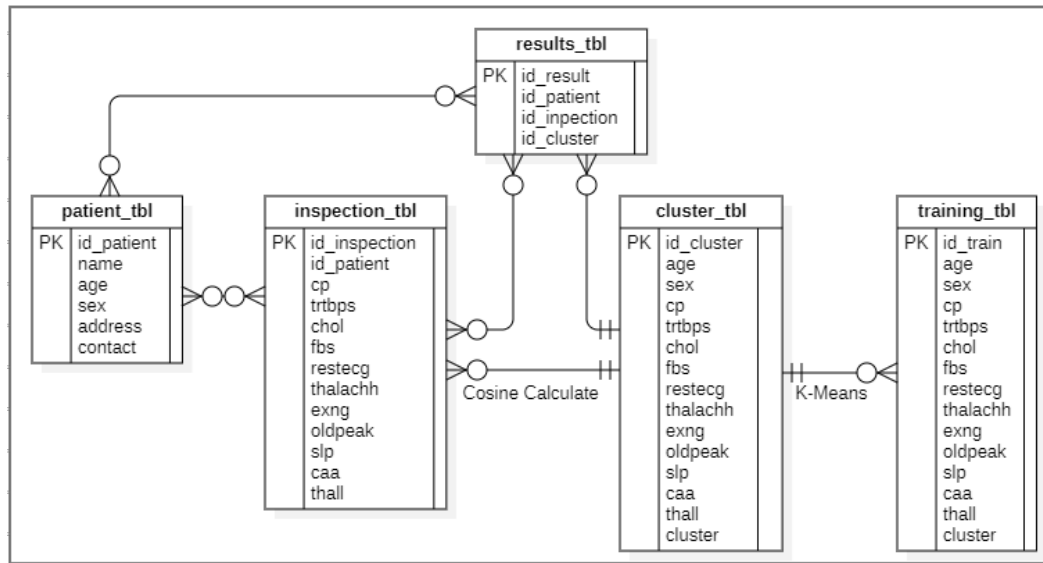


Figure 4. Entity Relationship Diagram to Predict Heart Disease

5. System Build

System development in this study uses website programming languages, such as: PHP and MySQL. The interface that is applied to the system uses a Bootstrap display so that the system attracts the user's attention and the data management process can be carried out properly. After developing the system display, the process continues with the development of machine learning methods to predict heart disease. The explanation of the method used is as follows:

a. K-Means Clustering

The k-means clustering method will be used to group patients on training data based on relevant clinical characteristics [11]. The k-means clustering process can be described in the following steps:

- Initialization: Determination of the number of data groups and the initial group center chosen randomly. In this study, the groups were divided into two groups: those who had a heart attack and those who did not.
- Clustering process (initial cluster center): Each patient will be assigned to a group based on the Euclidean distance between their clinical features and the nearest cluster center. The shortest distance is the data group that will be selected using equation 1 below. [12].

$$d = \|p(x,y) - ck\| \quad (1)$$

Where: k is the number of data clusters, d is a measure of how closely the data resemble the Euclidean equation, p(x,y) is a data feature, and ck is the centroid (cluster center) of the data.

- Cluster Center Update: The new cluster center will be calculated based on the average of the clinical features of the patients in the same group. The calculation of the new cluster center can be seen in equation 2 below [12].

$$ck = \frac{1}{k} \sum_{y \in ck} \sum_{x \in ck} p(x, y) \quad (2)$$

Where: k is the number of data clusters, p(x,y) is the data feature, and ck is the data centroid (cluster center).

- Iteration: The clustering center clustering and updating steps will be repeated until there is no significant change in the clustering. The end result of this process is the final cluster center, which is used as the main data for the process of detecting heart disease.

b. Cosine Similarity

After the clustering is done, the final cluster center will be generated for the data matching process. Then, each new patient data examination result will be compared with the center of the cluster using the cosine similarity method. This method involves a comparison of the clinical features of patients with cluster centers to identify similarities and potential risks of heart attack. The use of this method will produce the highest similarity value of 1 and the lowest similarity of 0 [13]. The result of calculating the maximum similarity between the two cluster centers is the result of heart disease detection. Obtaining similarity using the cosine similarity method can be seen in equation 3 below [14].

$$\cos \theta = \frac{a \cdot b}{\|a\| \cdot \|b\|} = \frac{\sum_{i=1}^n a_i \cdot b_i}{\sqrt{\sum_{i=1}^n (a_i)^2} \cdot \sqrt{\sum_{i=1}^n (b_i)^2}} \quad (3)$$

where a and b are the features being compared, and Cos θ is a measure of how comparable the characteristics are (values in the range of 0-1).

6. System Implementation Testing

In an effort to improve system performance, black box testing is used to test systems. All system features are tested to check if they operate correctly and in accordance with the previously completed design. It is clear from the results of the tests conducted on all features that the system functions effectively and produces reliable results.

7. Evaluation Method

The evaluation used in this study is a test conducted to obtain performance in the form of accuracy for the proposed method for predicting heart disease. Accuracy is an important parameter in measuring the performance of methods that are widely applied as support systems. Measuring the accuracy of the method proposed in this study can be done using equation 4 below [13].

$$\text{Accuracy (\%)} = \frac{\text{The numbers of true data}}{\text{All numbers of data}} \quad (4)$$

C. Result and Discussion

1. Application Testing for System Prediction

- a. Log in to the system

Before users as a whole use the heart disease prediction system, it is crucial to test the system on the login display to make sure it works properly and is trustworthy. Prior to the system's introduction into a production environment or use by end users, this system testing aims to find and fix any potential issues or errors. The login system display can be seen in Figure 5 below.

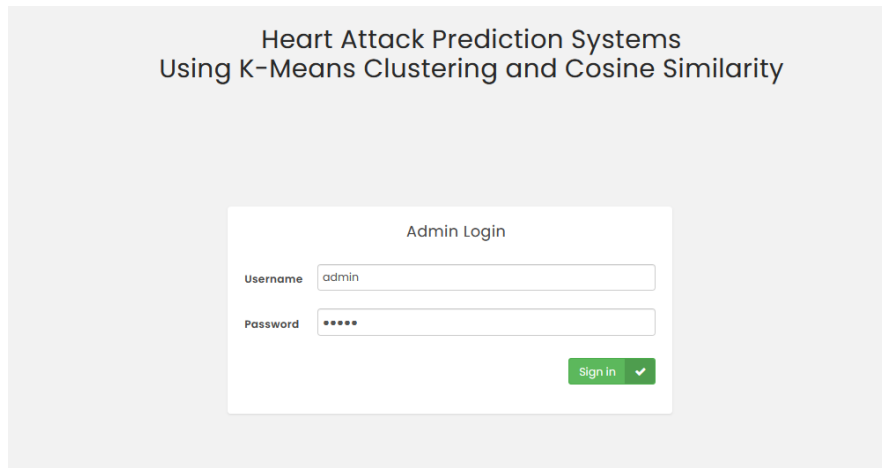


Figure 5. Login Display for the Heart Disease Prediction System

In this login view, functional testing will be carried out, which aims to ensure that all elements of the login view function properly, including the input fields for username, password, and login button. In addition, the system will provide proper verification and response when the user enters valid or invalid login information. The application created provides valid results and can respond to the user when entering the username and password provided.

b. Fill in the training data

Testing the system on the training data entry display for a heart disease prediction system is a critical process to ensure that the training data entered into the system is valid, of high quality, and can produce an accurate predictive model. This test focuses on how the user fills in the training data and how the system processes and validates the data before it is used to train a heart disease prediction model. The login system display can be seen in Figure 6 below.

Add Training Data

Patient Number

Age of the person

Gender of the person ☒ Male ☐ Female

Chest Pain type chest pain type ☒ Typical angina ☐ Atypical angina ☐ Non-anginal pain ☐ Asymptomatic

Resting blood pressure (in mm Hg)

cholestoral in mg/dl

Fasting blood sugar > 120 mg/dl ☒ False ☐ True

Resting electrocardiographic results ☒ Normal ☐ Having ST-T wave abnormality ☐ Showing probable

Figure 6. Fill in the training data display

In this view, testing will be carried out by ensuring the system can validate the data entered by the user correctly. In addition, the system will verify whether the user has filled in all the required data attributes in the correct format. The system will display an error message if the user fills in data in an invalid format or if an attribute is missing. The application created provides valid results and can respond to the user when entering the training data provided.

c. Fill in the new data patient and inspection result

Testing the system on the patient data entry display for the heart disease prediction system is an important step in ensuring that the data entered into the system is accurate, complete, and in accordance with the prediction model requirements. This test aims to verify that the patient data entry display functions properly and can receive relevant data to train and run a heart disease prediction model. The display of patient data entry in the system can be seen in Figure 7 below.

Add Patient Data

Name

Age of the person

Gender of the person ☒ Male ☐ Female

Adress

No. HP

Figure 7. Fill in the new patient data display

After filling in the patient's data, the next step is to fill in the data from the patient's examination results for needs and to determine whether or not you have heart disease. The display for checking data entry can be seen in Figure 8 below.

Inspection Data Patient

Patient Number

Age of the person

Gender of the person ☒ Male ☐ Female

Chest Pain type chest pain type ☒ Typical angina ☐ Atypical angina ☐ Non-anginal pain ☐ Asymptomatic

Resting blood pressure (in mm Hg)

cholesterol in mg/dl

Fasting blood sugar > 120 mg/dl ☒ False ☐ True

Resting electrocardiographic results ☒ Normal ☐ Having ST-T wave abnormality ☐ Showing probable

Figure 8. Fill in the inspection result display

The two displays in Figures 7 and 8 will be tested for validation to see if the system can validate the data entered by the user correctly. In addition, it will provide verification of whether the system recognises and addresses invalid or out-of-reach values for each attribute, such as patient age, blood pressure, cholesterol level, etc. The application created provides valid results and can respond to the user when entering patient data and the examination results provided.

d. Carry out the diagnosis process

Testing the system on the display of prediction results for a heart disease prediction system is very important to ensure that the results displayed to users are accurate, informative, and easy to understand. The diagnosis result display is an interface that serves as the output of a heart disease prediction model, and testing on this view is intended to verify that the system provides consistent, relevant, and reliable results. The display of system test results can be seen in Figure 9 below.

Clustering Results:

Show entries Search:

id_patient	Age	Sex	Hasil Diagnosis
1	63	Male	More chance of heart attack
2	37	Male	More chance of heart attack
3	41	Female	More chance of heart attack
4	56	Male	More chance of heart attack
5	57	Female	Less chance of heart attack
6	57	Male	More chance of heart attack
7	56	Female	Less chance of heart attack
8	44	Male	Less chance of heart attack
9	52	Male	More chance of heart attack
10	57	Male	More chance of heart attack
id_patient	Age	Sex	Hasil Diagnosis

Showing 1 to 10 of 302 entries Previous **1** 2 3 4 5 ... 31 Next

Figure 9. Carry out the prediction process

The outcomes of the prediction, which were acquired from the machine learning model utilised, will be tested and displayed in this view. It will also be checked to see if the prediction accurately reflects the patient's health situation, such as whether or not the patient has been given a heart disease diagnosis. The developed application offers reliable results, can reply to the user during the diagnosis process, and presents results that are reasonably correct.

2. Evaluation System Testing Accuracy

The purpose of this evaluation is to see how well the model can distinguish between patients with and without cardiac disease. Two approaches—the conventional k-means clustering approach and the suggested approach—will be tried in this study to see which yields the best accuracy results. Figure 10 below shows the heart disease prediction outcomes using the usual k-means clustering technique.

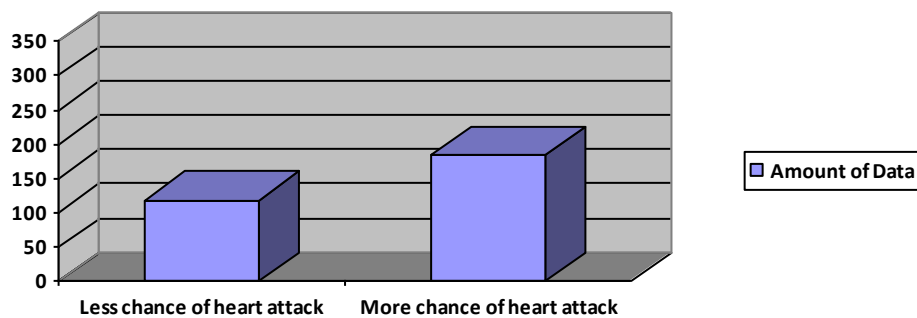


Figure 10. The results of the accuracy of the heart disease detection system with k-means clustering

The prediction results for heart disease obtained in Figure 10 are 91.06%. These results are quite good at predicting heart disease. However, some disadvantages of using the usual k-means clustering method are that it depends on the Euclidean distance. K-means clustering uses Euclidean distance to measure the closeness between the data and the cluster center. Although this metric is commonly used, the Euclidean distance does not always describe the relationship between data well in the case of complex multidimensional data or data with non-linear spatial features. So this study adds the cosine method to classifying patient data. The grouping results obtained using the proposed method can be seen in Figure 11 below.

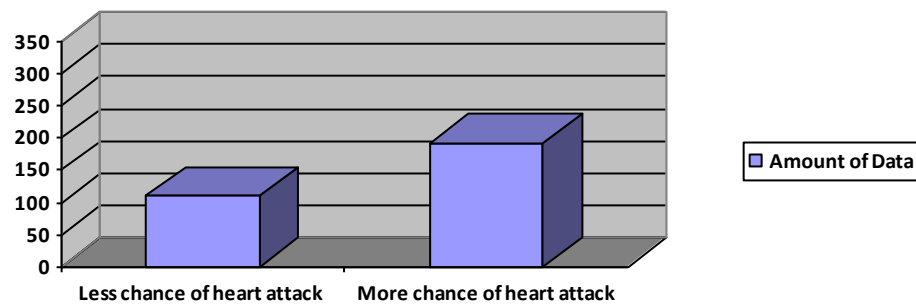


Figure 11. The results of the accuracy of the heart disease detection system with k-means clustering and cosine similarity

The prediction results for heart disease obtained in Figure 11 are 93.71%. These results obtained a fairly good increase in accuracy of 2.65%. The k-means clustering and cosine similarity methods are techniques that can be used in heart disease prediction systems to group patient data into similar groups or to measure the similarity between patient data based on relevant attributes.

In contrast to the k-means clustering approach, which divides patient data into k groups (clusters) based on attribute similarity, the cosine similarity method uses the direction and angle of the vector between two patient data sets to determine how similar they are. The effectiveness of cardiac disease prediction systems can be increased by combining the two techniques or using them independently. Using k-means clustering, for instance, we can classify patient data into groups according to particular characteristics, and then cosine similarity can be used to assess how similar the patient data are within each group. As a result, we can find similar patients with similar symptoms, which can help with further diagnosis and therapy.

The simplicity of deployment and speedy grouping of patient data are benefits of employing k-means clustering. In situations where the magnitude of the characteristic is unimportant but the direction is crucial, the advantage of employing cosine similarity is the ability to assess similarity based on the direction of the vector. The combination of these two approaches can offer a greater knowledge of the patterns and similarities among patient data, which can help with heart disease diagnosis, treatment, and overall understanding.

In heart disease prediction systems, the k-means clustering and cosine similarity approaches can be applied jointly or independently based on the previous justification. The outcomes of these two approaches can help medical practitioners uncover pertinent health trends and support them in choosing the best course of action for their patients. Additionally, putting the system to the test and validating it using these two techniques will give us a better understanding of how well the system predicts cardiac disease as a whole

D. Conclusion

The k-means clustering and cosine similarity methods can be useful tools in heart disease prediction systems to provide insight into patient health patterns and look for similarities between patients with an accuracy of 93.71%. The

integration of these two methods can improve predictive performance and provide health professionals with more complete information for better diagnosis and treatment.

E. Acknowledgment

We would like to express our gratitude to Universitas Medan Area for its assistance and contributions in developing this collaborative project.

F. References

- [1] N. Biswas *et al.*, "Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques," *Biomed Res. Int.*, vol. 2023, 2023, doi: 10.1155/2023/6864343.
- [2] J. Talukdar and T. P. Singh, "Early prediction of cardiovascular disease using artificial neural network," *Paladyn*, vol. 14, no. 1, 2023, doi: 10.1515/pjbr-2022-0107.
- [3] K. Rohit Chowdary, P. Bhargav, N. Nikhil, K. Varun, and D. Jayanthi, "Early heart disease prediction using ensemble learning techniques," *J. Phys. Conf. Ser.*, vol. 2325, no. 1, 2022, doi: 10.1088/1742-6596/2325/1/012051.
- [4] J. S. Rose, P. Malin Bruntha, S. Selvadass, M. V. Rajath, M. Bill Christ Mary, and D. Minni Jenifer, "Heart Attack Prediction using Machine Learning Techniques," *2023 9th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2023*, vol. 10, no. 11, pp. 210–213, 2023, doi: 10.1109/ICACCS57279.2023.10113045.
- [5] V. Chang, V. R. Bhavani, A. Q. Xu, and M. A. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthc. Anal.*, vol. 2, no. September 2021, p. 100016, 2022, doi: 10.1016/j.health.2022.100016.
- [6] P. Singh and I. S. Virk, "Heart Disease Prediction Using Machine Learning Techniques," *2023 Int. Conf. Artif. Intell. Smart Commun. AISC 2023*, pp. 999–1005, 2023, doi: 10.1109/AISC56616.2023.10085584.
- [7] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012046.
- [8] A. Yazdani, K. D. Varathan, Y. K. Chiam, A. W. Malik, and W. A. Wan Ahmad, "A novel approach for heart disease prediction using strength scores with significant predictors," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–16, 2021, doi: 10.1186/s12911-021-01527-5.
- [9] A. Malav, K. Kadam, and P. Kamat, "PREDICTION OF HEART DISEASE USING K-MEANS and ARTIFICIAL NEURAL NETWORK as HYBRID APPROACH to IMPROVE ACCURACY," *Int. J. Eng. Technol.*, vol. 9, no. 4, pp. 3081–3085, 2017, doi: 10.21817/ijet/2017/v9i4/170904101.
- [10] W. Zheng, W. Hou, and J. C. W. Lin, "A Deep Learning based Feature Entity Relationship Extraction Method for Telemedicine Sensing Big Data," *Mob. Networks Appl.*, no. 0123456789, 2022, doi: 10.1007/s11036-022-02024-3.
- [11] R. Singh and E. Rajesh, "Prediction of Heart Disease by Clustering and Classification Techniques," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 5, pp. 861–866, 2019, doi: 10.26438/ijcse/v7i5.861866.
- [12] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image Segmentation Using

- K-means Clustering Algorithm and Subtractive Clustering Algorithm," *Procedia Comput. Sci.*, vol. 54, pp. 764–771, 2015, doi: 10.1016/j.procs.2015.06.090.
- [13] A. Saleh, T. Tulus, and S. Efendi, "Analysis of Accurate Learning in Radial Basis Function Neural Network Using Cosine Similarity on Leaf Recognition," 2019, doi: 10.4108/eai.20-1-2018.2281924.
- [14] A. S. Nasution, A. Alvin, A. T. Siregar, and M. S. Sinaga, "KNN Algorithm for Identification of Tomato Disease Based on Image Segmentation Using Enhanced K-Means Clustering," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 4, no. 3, 2022, doi: 10.22219/kinetik.v7i3.1486.