Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024
DOI : https://doi.org/10.33395/sinkron.v8i3.13904
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Classification of Heart Disease Using Support Vector Machine

Tegar Haryahya Tanjung[1)], Mhd Furqan[2)]
[1,2)]Department of Computer Science, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[1)]Tegartanjung9@gmail.com, [2)]mfurqan@uinsu.ac.id

Submitted : Jul 18, 2024 | **Accepted** : Jul 23, 2024 | **Published** : Jul 27, 2024

Abstract : Heart disease is a disease that has a high mortality rate, with more than 12 million deaths occurring throughout the world. Diagnosis of heart disease is very challenging due to the complex interdependence of several attribute factors. The problem that frequently encountered is the lack of accuracy in the classification process. Thus, a system is needed to carry out early diagnosis of heart disease. The structure of this research is to take a heart disease dataset from Kaggle. Then the data will be cleaned with preprocessing. The preprocessing process carried out is changing table names, checking missing values, and normalizing. 820 data will be trained using a Support Vector Machine and 205 data will be tested to find out how well the model can perform classification. The results of training and testing from a total of 1025 data will form a classification model. The model formed using the Support Vector Machine obtained confusion matrix results of 88 is True Positive data, 93 is True Negative data, 10 is False Positive data, and 14 is False Negative data. So the results of model training produce an accuracy of 88%.

Keywords: Support Vector Machine, Heart Disease, Classification

INTRODUCTION

Heart disease is a disease that has a high mortality rate, more than 12 million deaths occur throughout the world due to heart disease (Kiruthika Devi et al., 2016). Thus, early diagnosis is very important. Diagnosis of heart disease is very challenging due to the complex interdependence of several attribute factors. The problem that is often faced is the lack of accuracy in the classification process (Jain et al., 2019).

The challenges in classifying heart disease involve various aspects, including the complexity of data related to various types of heart disease, variations in symptoms between individuals, and dynamic changes in conditions in each patient. In addition, classification problems also arise due to additional factors such as the diversity of medical data sources, including data from various sources such as test results, medical history, and physical parameters. Developing a classification model that can address this diversity is a major challenge, as it requires a deep understanding of the complexity of heart disease as well as the algorithm's ability to extract significant patterns from diverse and complex data (Furqan et al., 2020).

The main aim of this study is to evaluate the extent to which the Support Vector Machine (SVM) algorithm can be an optimal approach in the classification of heart disease based on various complex medical parameters.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13904 p-ISSN : 2541-044X

LITERATURE REVIEW

Support Vector Machine

Support Vector Machine (SVM) is an important part of machine learning theory. SVM is very efficient for many applications in science and engineering, especially for classification (pattern recognition) problems (Sahrani, 2021). The idea of SVM classification can be described as follows: suppose there are m observation samples (training set), (xi, yi), i = 1, 2, ...,m where:

$$xiT = (xi1, \dots xid) \in Rd \quad (1)$$

Where xiT is the d-dimensional feature of sample i and y ∈ {-1, +1} is its coded label class. If sample xi is assigned to the positive class, then yi is +1, and if it is assigned to the negative class, then yi is -1. This training set can be separated by the hyperplane $wTxi + b = 0$, where w is the weight vector and b is the bias. The equations of marginal hyperplanes H1 and H2 can be seen in (2.3) and (2.4).

$$H1: (wTxi + b) = 1 \quad (2)$$

$$H2: (wTxi + b) = -1 \quad (3)$$

So, correctly classified points satisfy inequality (2.5).

$$yi: (wTxi + b) \geq 1 \quad (4)$$

For xi, i = 1,2,…,m. The distance between the marginal hyperplanes is equal to 2. What training samples $\|w\|$ course that falls on hyperplanes H1 or H2, the side that defines the margin is the support vector, as shown in figure 1.

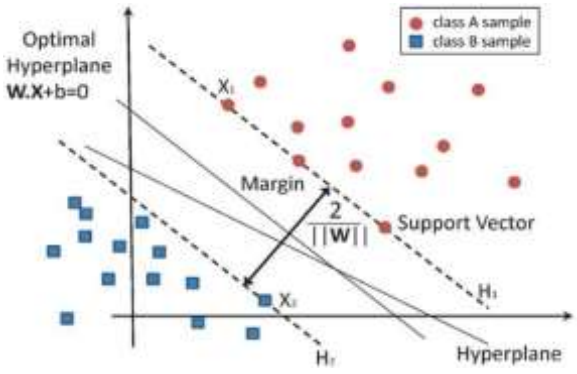

Figure 1. Support Vector Machine Classification

Kernel functions commonly used in Support Vector Machines are Linear kernel, Radial Basis Function (RBF), and Polynomial. The kernel function and parameters used in SVM analysis greatly influence the resulting accuracy (Chala Beyene, n.d.).

Support vector machine (SVM) is a classification method that was first introduced by Vapnik in 1998. Basically, this method works by defining the boundary between two classes with the maximum distance from the closest data. To get the maximum boundary between classes, the best hyperplane (dividing line) must be formed in the input space obtained by measuring the hyperplane margin and finding the maximum point. Margin is the distance between hyperplanes and the closest point of each.

Data Collection

The first stage in the research involved collecting relevant medical data from various sources, such as Kaggle datasets, which contain diverse information regarding health parameters such as blood pressure, cholesterol levels, disease history, and other medical test results. This data is the basis for

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13904 p-ISSN : 2541-044X

building a heart disease classification model using the SVM algorithm, so that good representation of health attributes is the key to effective data collection (Ghorbani & Ghousi, 2019).

Preprocessing data

This step consists of data preprocessing which includes changing table names, handling missing values, and normalizing data to equalize the value range for each attribute. Good preprocessing is needed to ensure data is clean and ready to use in the model development stage. The preprocessing steps can be seen in table 1.

Table 1. Preprocessing Step

No	Preprocessing	Function
1	Table Name Change	Increase data clarity, as well as facilitate the process of analysis and subsequent data processing.
2	Missing Values Handling	Identify and handle missing values.
3	Normalization	Equalize the value range of each attribute.

METHOD

The heart disease classification process is carried out using the support vector machine method to be able to classify data into heart disease and non-heart disease.

The research process consists of 3 stages, namely data collection, preprocessing, and forming a classification model.

Process Design

Process design is a depiction of the workflow sequence from when the process starts until the process stops. The workflow of one process with other processes can be depicted with a chart called a flowchart. This research began with the stage of collecting relevant medical data from the Kaggle dataset, which includes information on various health parameters such as blood pressure, cholesterol levels, disease history, and other test results. Then, understanding the data is carried out through exploratory analysis to understand the data structure, distribution of variables, and patterns that may exist in the dataset. After that, data preprocessing is carried out which includes handling missing values, normalizing values, and removing noise or outliers that can interfere with the model. Next, the dataset is randomly divided into training, validation, and test set subsets to train, validate, and test the model performance. The implementation of the SVM algorithm is carried out by adjusting the kernel and parameters appropriately, followed by optimizing the model using the Grid Search technique to get the best combination of parameters. Model performance evaluation was carried out using evaluation metrics such as accuracy, precision, recall, F1-score, as well as ROC and AUC curves to evaluate the model's ability to classify heart disease. Model validation is carried out with validation data sets to ensure good

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13904 p-ISSN : 2541-044X

generalization and prevent overfitting. Testing and interpretation of results were carried out on the test set data to test model performance and gain insight into the model's ability to diagnose heart disease. Finally, analysis and discussion are carried out by comparing with related research, discussing the implications of the findings, and the possibility of developing a better classification model in the future. The research process can be seen in figure 2.

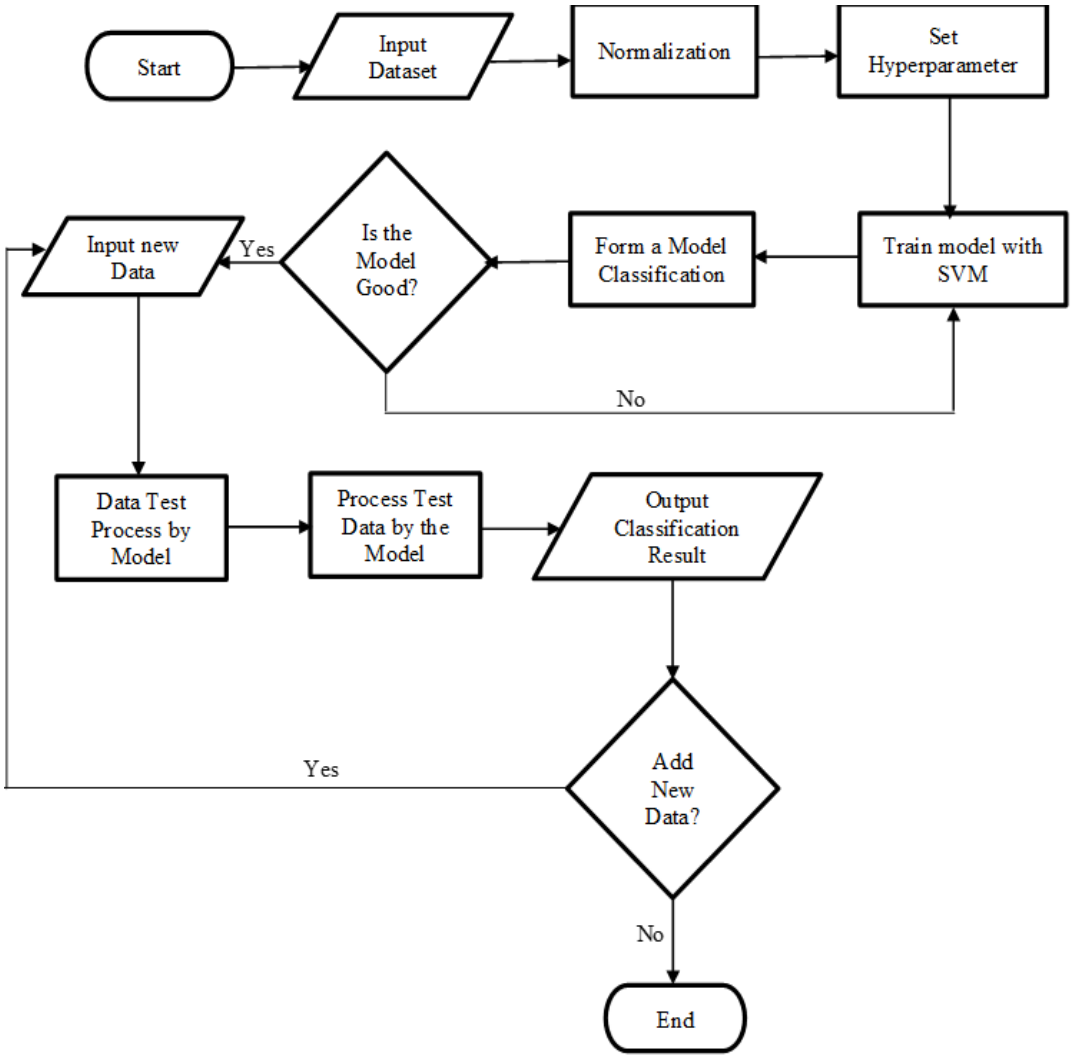

Figure 2. Research Process

Establishment of a Classification Model

After preprocessing, the dataset is randomly divided into training, validation, and test set subsets. This is done to train, validate and test the performance of the model that will be developed with predetermined proportions to ensure that the model being built is able to generalize well on new data (Annisa, 2019).

The next step involves implementing the Support Vector Machine (SVM) algorithm by selecting the appropriate kernel (such as linear, polynomial, or RBF), as well as adjusting parameters such as C (penalty parameter) and gamma (kernel coefficient). The application of SVM is the core of developing heart disease classification models. Therefore, optimization of the SVM model will be carried out using the Grid Search technique to find the best combination of parameters that produces a model with optimal

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13904 p-ISSN : 2541-044X

accuracy and generalization. The aim of this step is to improve the performance of the developed model (Farhan et al., 2022).

The developed model was evaluated using evaluation metrics such as accuracy, precision, recall, F1-score, as well as ROC and AUC curves to evaluate the model's ability to classify heart disease. This evaluation is important to determine the extent to which the model can be used effectively in the diagnosis of heart disease (Saputra, n.d.).

Final testing was carried out on the test set data to test the performance of the model that had been developed and gain insight into the model's ability to diagnose heart disease. Interpretation of model prediction results is also necessary to gain a deeper understanding of the model's capabilities.

The final stage involves analysis of the results, including comparison with related studies, discussion of the implications of the findings for heart disease diagnosis, as well as the possible development of better classification models in the future. This discussion provides an overview of the relevance of the findings to the broader scientific context and the potential for developing better models for the diagnosis of heart disease (Wibisono & Fahrurozi, 2019).

RESULT

Data Collectiom

This research uses a dataset obtained from the Kaggle website via the following link:

"https://www.Kaggle.com/datasets/johnsmith88/heart-disease-dataset". Figure 3 shows the website page where I retrieved the dataset.

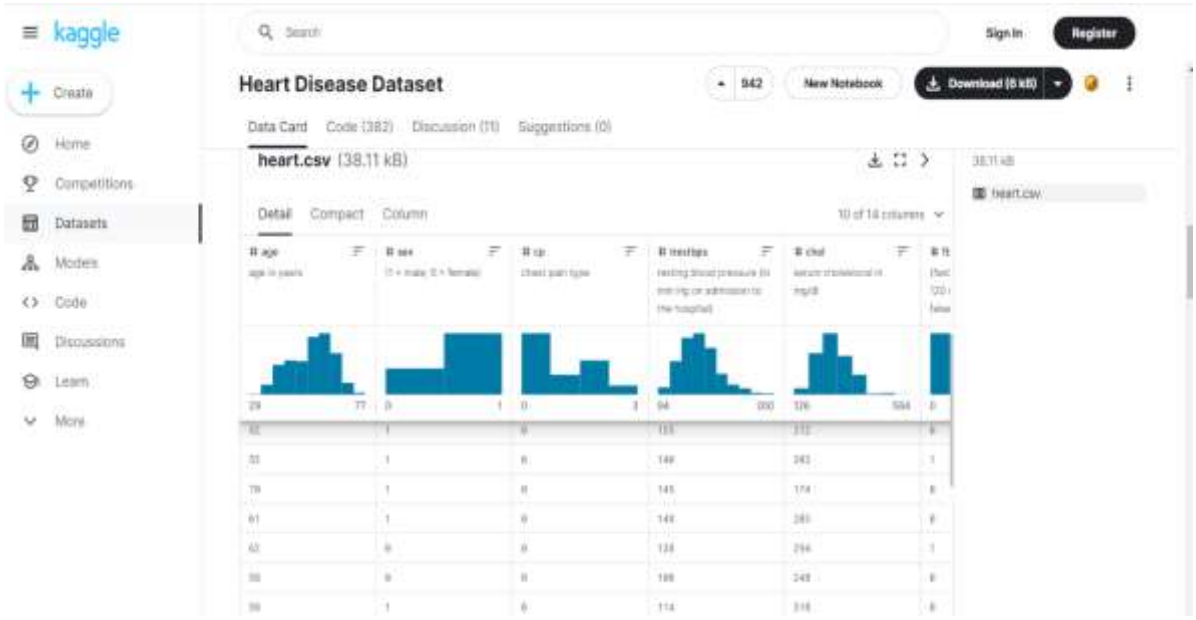

Figure 3. Website Dataset Location

This dataset provides relevant and structured data related to heart disease, which includes features such as age, gender, resting blood pressure, serum cholesterol levels and more. The dataset will be downloaded and saved into a local file with the name "heart.csv". This dataset has a total of 1025 data which will later be analyzed to form a classification model.

Data Split

Data splitting is done to divide the dataset into two parts, namely train data and test data. Train data will be determined at the beginning, namely 80% of the total dataset and test data 20% of the total dataset. The split data process is carried out using Python syntax, namely:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13904 p-ISSN : 2541-044X

A. *train_test_split* which is used to divide the dataset into two parts: training data (train) and test data (test).
B. *X_normalized* as a variable that contains dataset features that have been normalized using MinMaxScaler.
C. *y* as a variable that contains the target or label of the dataset.
D. *test_size=0.2* as a parameter that determines the proportion of data that will be allocated as test data.
E. *random_state=42* as an optional parameter that determines the seed value for the random number generator.

Hyperparameter Search

Researchers have carried out a series of tests and evaluations using various combinations of hyperparameter values such as C, gamma, and kernel type. This research uses a value of C = 100, because it wants to give the SVM model a greater ability to fit the training data well. Gamma with a 'scale' value, the model automatically adjusts the gamma to match the scale of the data to be studied. Poly or polynominal kernels are used when the relationship between features and targets is not straight. This kernel is a tool to help the model see more complex patterns in the data. With this kernel, the model can understand the patterns that weave in the data.

Manual Calculation

Determining the class of new data using the manual SVM formula involves several mathematical equations. The following are the calculation steps:

A. In the first stage the author will take 5 examples of random patient data from the dataset. Example data can be seen in Tables 2 and 3.

Tables 2. Training Data Examples

No.	umur	Jenis_kelamin	Jenis_nyeridada	tekanan_darah	kolestrol
1	58	0	1	136	319
2	48	0	2	130	275
3	44	1	0	120	169
4	58	1	0	100	234
5	56	1	1	130	221

Tables 3. Training Data Examples

No.	gula_darah	elektrokardiografi	detak_jantung	target
1	1	0	152	0
2	0	1	139	1
3	0	1	144	0
4	0	1	156	0
5	0	0	163	1

B. Calculate the weight vector (w) with the following equation:

$$w = (X^T \times X)^{-1} \times (X^T) \times y \quad (5)$$

X^T= The transpose of the matrix X is obtained by swapping its rows and columns.
$(X^T \times X)^{-1}$= Generates a matrix that is the inverse of a matrix (Inverse Matrix).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13904 p-ISSN : 2541-044X

C. C. It is known that the values of X and y are as follows:

X = [
[58, 0, 1, 136, 319, 1, 0, 152],
[48, 0, 2, 130, 275, 0, 1, 139],
[44, 1, 0, 120, 169, 0, 1, 144],
[58, 1, 0, 100, 234, 0, 1, 156],
[56, 1, 1, 130, 221, 0, 0, 163]
]
y = [0, 1, 0, 0, 1]

D. New data will be classified using the following equation:

$$F(x_{baru}) = w^T \times x_{baru} + b \quad (6)$$

x_{baru}= new data matrix

The author will provide a table of 5 examples of new data used to test the classification model. New data can be seen in table 4.

Table 4. New Data Process

No	Matriks baru	f(x_baru)	Prediksi
1	[55, 1, 0, 128, 200, 0, 0, 130]	0.82460439	Having Heart Disease
2	[62, 0, 1, 140, 250, 0, 1, 150]	1.3401105	Having Heart Disease
3	[40, 1, 0, 110, 180, 1, 0, 115]	-0.3841715 3	Not Having Heart Disease
4	[50, 0, 1, 135, 230, 1, 1, 145]	0.29661959	Having Heart Disease
5	[58, 1, 0, 125, 190, 0, 0, 120]	1.30744992	Having Heart Disease

From the results of the table above, it is known that the new data provided predicts that 4 people will have heart disease and 1 person will not have heart disease.

Visual Representation of SVM

A visual representation of the SVM model can be seen in Figure 4.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13904 p-ISSN : 2541-044X

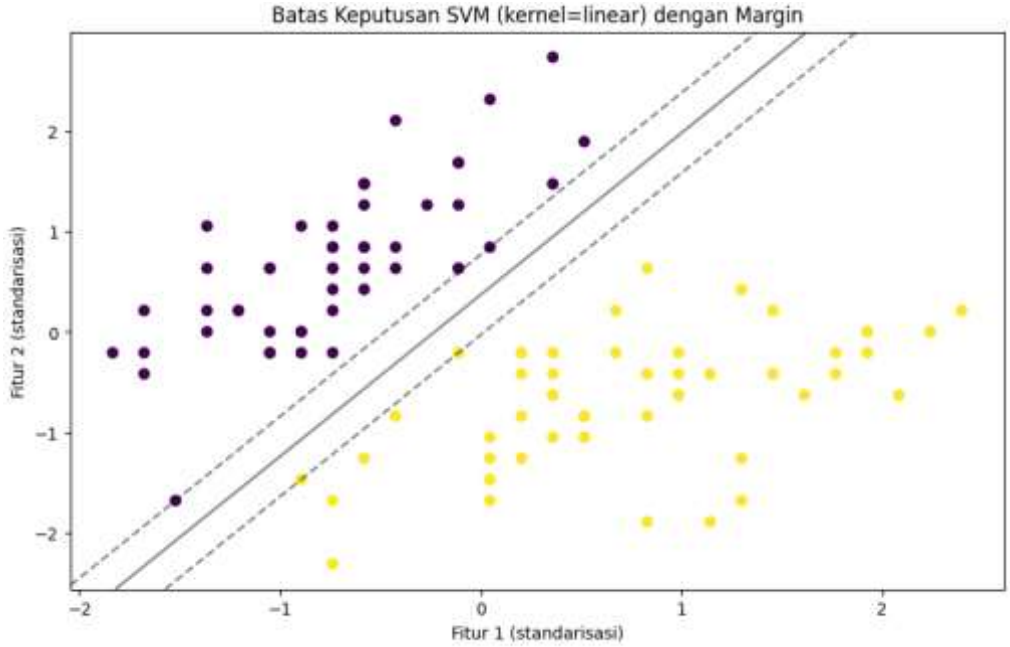

Figure 4. Visual Model SVM Representation

How to read the graph above is as follows:

Data Point: Each point in the graph represents one data sample from the Iris dataset. Each sample is displayed in two dimensions, namely feature 1 (x-axis) and feature 2 (y-axis). The color of each point indicates the class or target of the sample. Red indicates class 1, purple class 2, and green class 3.

Decision Boundary: The straight line seen in the graph is the decision boundary made by the SVM model. This is the threshold at which the model decides to separate samples into different classes.

Margin: The dotted line around the decision boundary shows the margin generated by the SVM model. Margin is the distance between the decision boundary and the closest points of each class. SVM models attempt to maximize this margin to improve generalization and prediction performance.

Support Vectors: The points located above or below the margin are support points (support vectors). They are the most influential samples in determining the decision boundaries and margins of the SVM model.

DISCUSSION

The total data used in this research was 1025 data. 820 data will be trained and 205 data will be tested using the Support Vector Machine. The quality measures that will be measured to determine the performance of the system are accuracy, precision, recall, and F1score. Then the average number of quality measures will be calculated.

The classification model that is formed after going through the testing process using SVM will provide several questions related to heart disease which will later be filled in by the user. The results of the model can be seen in Figure 5.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13904 p-ISSN : 2541-044X

```
Masukkan usia: 45
Masukkan jenis kelamin (1 untuk pria, 0 untuk wanita): 1
Masukkan jenis nyeri dada (1-4): 4
Masukkan tekanan darah istirahat (mm Hg): 150
Masukkan kadar serum kolesterol (mg/dl): 165
Masukkan kadar gula darah puasa (1 jika > 120 mg/dl, 0 jika tidak): 0
Masukkan hasil elektrokardiografi istirahat (0-2): 2
Masukkan detak jantung maksimal tercapai: 220
Berdasarkan data yang Anda masukkan, kemungkinan terdapat penyakit jantung.
```

Figure 5. Classification Model

Some related questions that must be filled in by the user are age, gender, type of chest pain that is usually reported to medical personnel, blood pressure, serum cholesterol levels, blood sugar levels, results of electrocardiography, and maximum heart rate. After the user enters appropriate data, the model will make predictions from the existing classification, namely whether the patient has the possibility of heart disease or not.

Accuracy is used to measure classification performance based on the accuracy of the classification method used. Precision is the level of accuracy between the information requested by the user and the answer provided by the system. Recall is the level of success of the system in retrieving information. F1Score is the harmonic mean of the recall and precision values. This calculation is useful for finding out how precise and reliable the system's performance is in classifying classes. The results of the model evaluation can be seen in Figure 6.

	precision	recall	f1-score	support
0	0.90	0.86	0.88	102
1	0.87	0.90	0.89	103
accuracy			0.88	205
macro avg	0.88	0.88	0.88	205
weighted avg	0.88	0.88	0.88	205

Figure 6. Classification Model Evaluation

It can be seen in the figure that the accuracy obtained by this model is 88%. The precision obtained by this model is 90%. The recall obtained by this model is 86%, the f1-score obtained by this model is 88%. This shows that the model formed can operate well.

CONCLUSSION

This research succeeded in forming a classification model for heart disease using a dataset taken from Kaggle, trained using a Support Vector Machine (SVM). The results obtained show that the best hyperparameter configuration for this dataset is C 100, gamma scale, and kernel poly. This configuration produces an SVM model with optimal fitting capabilities to complex data, tight decision margins, and dynamically adjusted gamma scaling. With an accuracy rate of 88%, the results of this study show that SVM is an excellent algorithm for classifying heart disease. These findings indicate the potential of SVM in supporting medical diagnosis and further development in cardiac disease classification.

*name of corresponding author

REFERENCES

Annisa, R. (2019). ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK PREDIKSI PENDERITA PENYAKIT JANTUNG. *Jurnal Teknik Informatika Kaputama (JTIK)*, *3*(1).

Chala Beyene, M. (n.d.). *Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques*. http://www.ijpam.eu

Farhan, R., Pohan, R., Ratnawati, D. E., & Arwani, I. (2022). *Implementasi Algoritma Support Vector Machine dan Model Bag-of-Words dalam Analisis Sentimen mengenai PILKADA 2020 pada Pengguna Twitter* (Vol. 6, Issue 10). http://j-ptiik.ub.ac.id

Furqan, M., Kurniawan, R., & HP, K. I. (2020). Evaluasi Performa Support Vector Machine Classifier Terhadap Penyakit Mental. *JURNAL SISTEM INFORMASI BISNIS*, *10*(2), 203–210. https://doi.org/10.21456/vol10iss2pp203-210

Ghorbani, R., & Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. In *International Journal of Data and Network Science* (Vol. 3, Issue 2, pp. 47–70). Growing Science. https://doi.org/10.5267/j.ijdns.2019.1.003

Jain, A., Ahirwar, M., Pandey, R., & Pandey, R. A. (2019). Review on Intutive Prediction of Heart Disease Using Data Mining Techniques. *International Journal of Computer Sciences and Engineering*. https://doi.org/10.26438/ijcse/v7i7.109113ï

Kiruthika Devi, S., Krishnapriya, S., & Kalita, D. (2016). Prediction of heart disease using data mining techniques. *Indian Journal of Science and Technology*, *9*(39). https://doi.org/10.17485/ijst/2016/v9i39/102078

Sahrani, L. (2021). Classification of Tomato Leaf Based on Gabor Filter Extraction And Support Vector Machine Algorithm. *International Journal of Information System & Technology Akreditasi*, *4*(2), 677–681.

Saputra, K. (n.d.). Perbandingan Kinerja Fungsi Kernel Algoritma Support Vector Machine Pada Klasifikasi Penyakit Padi. *IJCCS*, *x*, *No.x*, 1–5.

Wibisono, A. B., & Fahrurozi, A. (2019). PERBANDINGAN ALGORITMA KLASIFIKASI DALAM PENGKLASIFIKASIAN DATA PENYAKIT JANTUNG KORONER. *Jurnal Ilmiah Teknologi Dan Rekayasa*, *24*(3), 161–170. https://doi.org/10.35760/tr.2019.v24i3.2393

*name of corresponding author