VOL. 8. NO. 1 AUGUST 2022
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v8i1.3328

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

Identify Cholesterol Disease Risk Levels Using Multiple Linear Regression Algorithms

Deny Haryadi[1*]; Dewi Marini Umi Atmaja[2]

Teknologi Informasi[1*]
Institut Teknologi Telkom Jakarta
https://ittelkom-jkt.ac.id/
denyharyadi@ittelkom-jkt.ac.id

Bisnis Digital[2]
Universitas Medika Suherman
https://medikasuherman.ac.id/
dewi@medikasuherman.ac.id

Abstract—*Cholesterol is one of the fat compounds found in the bloodstream that are necessary for the formation of several hormones and new cell walls in the liver. Normal cholesterol levels in the human body are in the range of < 200 mg / dL. If cholesterol levels in the blood are abnormal or excessive, it can result in dangerous diseases such as heart disease or stroke. In this study, cholesterol disease prediction will be carried out using models formed from linear regression methods, so that the results of this study can be used as a reference for early prevention of cholesterol disease and become a means of decision making. Linear regression is one of the prediction methods in data mining that can be used to find out how dependent variables/criteria can be predicted through independent variables or predictor variables individually. In this study by utilizing some data of patients with cholesterol disease that has been stored in the database using several attributes, namely age, BMI, glucose, and cholesterol. So by applying a linear regression algorithm can be done a prediction in the identification of cholesterol diseases based on functional relationships on the attributes in the data. The results of this study showed an RMSE value of 0.347 with a standard deviation of /- 0.000. This shows that the model resulting from linear regression algorithms with the above cases is quite accurate.*

Keywords: *cholesterol disease, prediction, data mining, linear regression algorithms, RMSE.*

Intisari—Kolesterol adalah salah satu senyawa lemak yang ditemukan dalam aliran darah yang diperlukan untuk pembentukan beberapa hormon dan dinding sel baru di hati. Kadar kolesterol normal dalam tubuh manusia berada pada kisaran < 200 mg/dL. Jika kadar kolesterol dalam darah tidak normal atau berlebihan, dapat mengakibatkan penyakit berbahaya seperti penyakit jantung atau stroke. Pada penelitian ini akan dilakukan prediksi penyakit kolesterol dengan menggunakan model yang dibentuk dari metode regresi linier, sehingga hasil penelitian ini dapat dijadikan acuan untuk pencegahan dini penyakit kolesterol dan menjadi sarana pengambilan keputusan. Regresi linier merupakan salah satu metode prediksi dalam data mining yang dapat digunakan untuk mengetahui bagaimana variabel/kriteria dependen dapat diprediksi melalui variabel independen atau variabel prediktor secara individual. Dalam penelitian ini dengan memanfaatkan beberapa data penderita penyakit kolesterol yang telah disimpan dalam database dengan menggunakan beberapa atribut yaitu umur, indeks berat badan, glukosa, dan kolesterol. Sehingga dengan menerapkan algoritma regresi linier dapat dilakukan suatu prediksi dalam identifikasi penyakit kolesterol berdasarkan hubungan fungsional pada atribut-atribut pada data. Hasil penelitian ini menunjukkan nilai RMSE sebesar 0,347 dengan standar deviasi /- 0,000. Hal ini menunjukkan bahwa model yang dihasilkan dari algoritma regresi linier dengan kasus di atas cukup akurat.

Kata Kunci*:* Penyakit Kolesterol, Prediksi, Data Mining, Algoritma Regresi Linier, RMSE.

INTRODUCTION

Cholesterol is one of the fatty compounds found in the bloodstream that is necessary for the formation of several hormones and new cell walls in the liver [1]. Cholesterol in the blood is carried by lipoproteins, which are defined by three classes namely Low-Density Lipoprotein (LDL), High-Density Lipoprotein (HDL), and Triglycerides (TGA). LDL serves to carry cholesterol throughout the body through arterial blood vessels, if the levels are too high then LDL will accumulate in the walls of arteries

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.28/E/KPT/2019, September 26, 2019. Published by LPPM Universitas Nusa Mandiri

VOL. 8. NO. 1 AUGUST 2022
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v8i1.3328

(bad cholesterol)[2]. HDL serves to restore excess cholesterol to the liver to be removed from the body (good cholesterol) [3][4]. TGA is formed when the body changes the remaining unused calories in the body, if the body continues to get excessive calorie intake compared to its use, then triglyceride levels will rise which causes stroke or heart disease [5].

Normal cholesterol levels in the human body are in the range of < 200 mg / dL [6]. If cholesterol levels in the blood are abnormal or excessive, it can result in dangerous diseases such as heart disease or stroke[7]. Factors that cause high cholesterol levels are weight, heredity, smoking, lack of exercise, and unhealthy food. Common symptoms that occur in cholesterol disease are easily tired, drowsiness, leg pain, the nape of the head feeling sore, chest pain, yellowish wounds in a few parts of the body, impotence, and yellow patches under the eyelids [8][9].

Cholesterol disease in a person can be identified or predicted by collecting data such as age, BMI, and glucose. The data can be processed using data mining techniques so that new information or knowledge is obtained [6]. Prediction is the process of systematically estimating what is most likely to happen in the future based on past and present information owned so that the error (the difference between something that happened and the approximate result) can be reduced [10].

Linear regression is one of the algorithms in data mining that can be used to find out how dependent variables/criteria can be predicted through independent variables or predictor variables individually [11][12]. In this study, cholesterol disease prediction will be carried out using models formed from linear regression methods, so that the results of this study can be used as a reference for early prevention of cholesterol disease and become a means of decision making.

MATERIALS AND METHODS

A. *Research Stages*

To facilitate research and run systematically, a flow or stage is made in the research, as in figure 1.

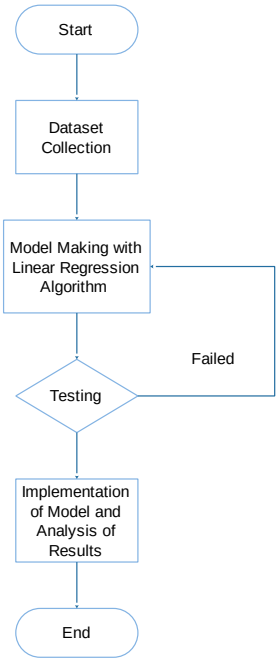

Figure 1. Research Stages

B. *Data Collection*

The data used in the study was taken from the Kaggle.com website. Then several stages of data processing will be explained as follows [13]:

a) The process of data cleaning is part of the process of cleaning, deleting data or selection of missing value data, data that is not used in a dataset, in checking the inconsistencies of a data and correcting errors in the data [14][15]. The process of cleaning data is done manually with the help of spreadsheet software or Microsoft excel.

Table 1. Data Cleaning

Age	Attributes / Variables					
	Prevalent Stroke	Prevalent Hyp	Diabets	BMI	Glucose	Cholesterol
39	0	0	0	26.97	77	0
46	0	0	0	28.73	76	0
48	0	0	0	25.34	70	0
61	0	1	0	28.58	103	1
46	0	0	0	23.1	85	0
43	0	1	0	30.3	99	0
63	0	0	0	33.11	85	1
45	0	0	0	21.68	78	0
52	0	1	0	26.36	79	0
43	0	1	0	23.61	88	0

b) The data selection process is part of data selection before entering the next stage[16]. The selection of attributes used is based on factors that influence cholesterol disease such as age, BMI, and cholesterol. Data that passes the selection will be analyzed and then determined the attributes to be used and grouped so that a dataset can be divided into training data and testing data [17][18].

c) Data Transformation is the process of converting the initial data format into a standard data format [19] which can be viewed in Table.

Table 2. Data Transformation

Age	Attributes / Variables		
	BMI	Glucose	Cholesterol
39	26.97	77	0
46	28.73	76	0
48	25.34	70	0
61	28.58	103	1
46	23.1	85	0
43	30.3	99	0
63	33.11	85	1
45	21.68	78	0
52	26.36	79	0
43	23.61	88	0

After data processing is completed, the next stage is data modeling [20][21]. In this study, the algorithm used was linear regression. The general form of a simple linear regression equation is written in the following formula:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (1)$$

$$a = \bar{y} - b\bar{x} \quad (2)$$

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - b\frac{\sum_{i=1}^{n} x_i}{n} \quad (3)$$

With
n: number of data pairs
Yi: the i-th non-freebie value Y
Xi: the i-th free changer value X

Calculating linear regression equations:

$$Y = a + b X \quad (4)$$

With
Y: non-free changer
X: free changer
a: constant
b: slope

RESULTS AND DISCUSSION

A. *Model Making With Linear Regression Algorithm*

The study used the Linear Regression algorithm, to identify cholesterol diseases and will obtain Root Mean Square Error (RMSE) results as well as predictions that can be used in decision-making when patients are identified with cholesterol disease. The source of the data as an object in this study is historical data taken from the Kaggle.com site. The data used in this study consisted of attributes or variables such as age, BMI, glucose, and cholesterol.

B. *Split Validation*

Validation techniques divide data into two parts at random, some as training data and others as data testing. By using Split Validation will be conducted training experiments based on a predetermined split ratio, then the rest of the split ratio data training will be considered as data testing.

Table 3. Cholesterol Dataset

Age	Attributes / Variables		
	BMI	Glucose	Cholesterol
39	26.97	77	0
46	28.73	76	0
48	25.34	70	0
61	28.58	103	1
46	23.1	85	0
43	30.3	99	0
63	33.11	85	1
45	21.68	78	0
52	26.36	79	0
43	23.61	88	0
50	22.91	76	0
43	27.64	61	0
46	26.31	64	0
41	31.31	84	0
38	21.35	70	1
48	22.37	72	0
46	23.38	89	1
38	23.24	78	0
41	26.88	65	0
42	21.59	85	0
52	34.17	113	0

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.28/E/KPT/2019, September 26, 2019. Published by LPPM Universitas Nusa Mandiri

VOL. 8. NO. 1 AUGUST 2022
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v8i1.3328

Age	Attributes / Variables		
	BMI	Glucose	Cholesterol
52	25.11	75	0
44	21.96	83	0
47	24.18	66	1
35	26.09	83	0
61	32.8	65	1
60	30.36	74	0
36	28.15	63	0
43	27.57	75	0
59	20.77	88	1
61	18.59	75	1
54	24.71	87	0
37	38.53	83	0
56	28.09	75	0
52	40.11	225	0
42	28.93	90	0
36	27.78	80	0
43	26.87	78	0
41	23.28	74	0
54	26.21	100	0
53	21.51	215	1
49	20.68	98	0
65	30.47	87	0
46	23.59	74	0
63	22.15	75	1
36	24.33	62	0
63	27.1	79	1
51	23.47	95	0
47	19.66	75	0
62	28.27	75	0
46	20.35	94	0
54	17.61	55	0
49	25.65	80	0
44	22.29	82	0
40	25.45	87	1
56	23.58	93	0
67	24.25	74	0
53	19.64	73	0
57	24.88	72	0
57	26.84	64	1
63	28.6	45	0
62	29.64	202	0
38	23.01	78	0
47	20.13	83	0
52	23.51	87	0
42	28.61	95	0
37	33.29	87	0
41	33.8	75	0
44	22.16	83	1
59	34.55	103	1
44	24.04	68	0
44	21.16	89	0
45	45.8	63	0
41	30.58	65	0
60	26.52	83	0
39	32.51	70	1
53	22.49	87	0
52	26.03	82	0
61	29.35	83	0
36	22.73	65	0
62	23.89	77	0
61	38.46	78	0
41	28.56	70	0
41	25.42	76	0
53	18.23	75	0
39	24.8	97	0
51	27.38	77	0
66	28.55	104	0
60	28.57	65	1
65	29.33	96	0
63	26.64	126	0
56	23.72	120	0
56	22.36	66	0
47	27.98	75	0
60	29.66	105	0
45	20.68	71	0
48	24.23	64	0
42	23.25	99	0
63	26.76	56	0
42	26.93	79	0

C. Linear Regression Calculation

Table 4. Cholesterol Disease Prediction

Age	Attributes / Variables		
	BMI	Glucose	Cholesterol
49	29.62	60	?
67	25.75	87	?

48	23.62	68	?

$$\Psi = \alpha + \beta_1\Xi_1 + \beta_2\Xi_2 + \beta_3\Xi_3 \quad (26)$$

$$y = -0.334 + (0{,}009 \times 49) + (-0{,}001\ x\ 29{,}62) + (0{,}001\ x\ 60)$$

$$= -0{,}334 + (0{,}441) + (-0{,}03) + (0{,}06)$$

$$= 0{,}137 \quad (27)$$

$$y = -0.334 + (0{,}009 \times 67) + (-0{,}001\ x\ 25{,}75) + (0{,}001\ x\ 87)$$

$$= -0{,}334 + (0{,}603) + (-0{,}026) + (0{,}087)$$

$$= 0{,}33 \quad (28)$$

$$y = -0.334 + (0{,}009 \times 48) + (-0{,}001\ x\ 23{,}62) + (0{,}001\ x\ 68)$$

$$= -0{,}334 + (0{,}432) + (-0{,}024) + (0{,}068)$$

$$= 0{,}142 \quad (29)$$

D. Implementation of Model

a) After the data collection process is completed, the next stage carried out is to make data modeling on the rapid miner. The model used in rapid miners uses linear regression algorithms. The first step is to enter cholesterol data into the rapidminer to be processed.

b) In this study, in figure 2. the dataset was divided into 2 parts, namely 90% of training data and 10% data testing using split validation. Split Validation is a validation technique that divides data into two parts randomly, partly as data training and some as data testing.

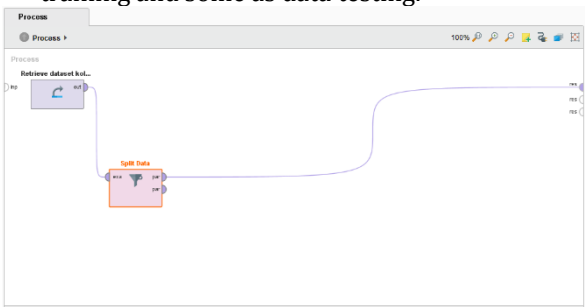

Figure 2. The process of sharing data with split validation

c) The next process in figure 3. is to set split validation parameters by dividing training data and testing data in the rapidminer.

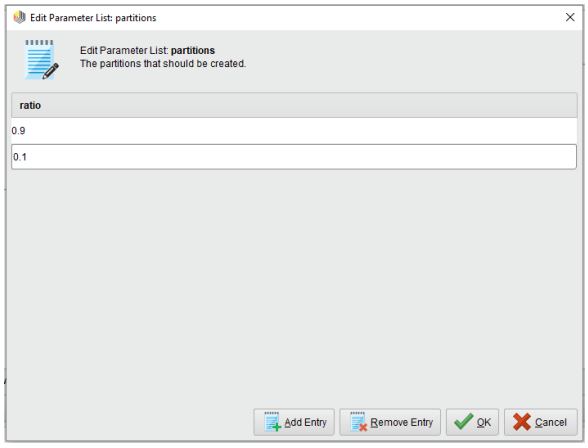

Figure 3. The process of sharing training data and data testing

d) The next process in figure 4. is to enter a linear regression algorithm to see the prediction results in the rapidminer.

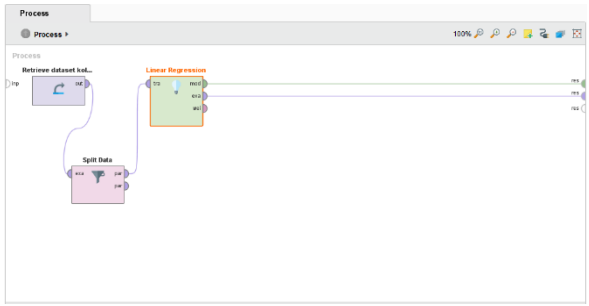

Figure 4. The process of inserting linear regression algorithms in rapid miners

e) After that, in figure 5. select attributes are carried out to find out the prediction results of the rapidminer, the results of manual calculations and test results in the rapidminer.

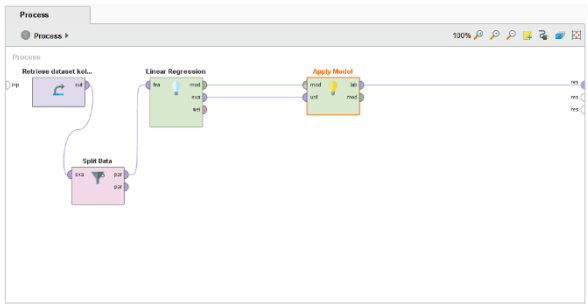

Figure 5. Model evaluation process in rapid miner

f) In figure 6. this process aims to enter training data and testing data that will be tested to produce predictions on class attributes.

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.28/E/KPT/2019, September 26, 2019. Published by LPPM Universitas Nusa Mandiri

VOL. 8. NO. 1 AUGUST 2022
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v8i1.3328

Row No.	Cholesterol	prediction(C...	Age	BMI	Glucose
1	0	0.060	39	26.970	77
2	0	0.128	46	28.730	76
3	0	0.116	48	25.340	70
4	1	0.301	61	28.580	103
5	0	0.112	46	23.100	85
6	0	0.149	43	30.300	99
7	1	0.314	63	33.110	85
8	0	0.084	45	21.680	78
9	0	0.172	52	26.360	79
10	0	0.094	43	23.610	88
11	0	0.130	50	22.910	76
12	0	0.071	43	27.640	61
13	0	0.095	46	26.310	64
14	0	0.112	41	31.310	84
15	1	0.008	38	21.350	70

Figure 6. Predicted results

g) In figure 7. when the prediction has been sought, the next step is to measure how accurate the prediction results have been made.

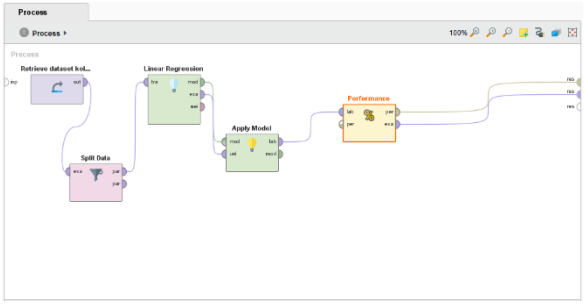

Figure 7. Root mean squared error search process

h) In figure 8. to facilitate the reading of cholesterol data, it is necessary to input performance tools to find Root Mean Squared Error. Here are the results:

root_mean_squared_error

root_mean_squared_error: 0.347 +/- 0.000

Figure 8. Root test results mean squared error

i) The second step is the implementation of linear regression algorithms using rapidminer tools. Here are the stages in the application of linear regression algorithms:
Determine the prediction of test data carried out by the rapidminer and produce a confidence value that has been predicted.
Specifies performance with output to find Root Mean Squared Error

j) In the split validation model there are two parts, namely the training section (used for classification algorithms) and the testing section (using the Apply Model feature to apply the model to the testing data and the Performance feature to display root mean squared error).

In 100 records of the training data dataset, after calculating the values for X_1Y, X_2Y, X_3Y, X_1X2X_3, X_1^2, X_2^2 and X_3^2, the overall results of each are 4963 for the total value of the X_1 variable, the total value of the X_2 variable is 2624.7, the total the value of the X_3 variable is 8414, the total value of the Y variable is 17, the total value of X_1Y is 914, the total value of X_2Y is 443.6, the total value of X_3Y is 1482, the total value of $X_1X_2X_3$ is 11203795, the total value of X_1^2 is 254035, the total value of X_2^2 is 71174.93, and the total value of X_3^2 is 778358. The total value above is used to convert the formula to find the value of coefficient a, coefficient b_1, coefficient b_2 and coefficient b_3 so that the value of a is -0.334, the value of b_1 is 0.009, the value of b_2 is -0.001, and the value of b_3 is 0.001. These three coefficient values will be used in the application of the multiple linear regression algorithm equation. From the coefficient values above, by using a simple mathematical model, we can determine the equation of the Linear Regression algorithm with the model Y = -0.334 + (0.009.X_1) + (-0. 001.X_2) + (0.001.X_3), where the X_1 variable is Age, variable X_2 is BMI, and variable X_3 is Glucose.

The multiple linear regression equation will then be implemented in predicting the testing data. In general, the application of the Linear Regression equation is applied to predict 3 dataset records as testing data that has been determined. From the calculations carried out manually and compared to the process in the Rapid Miner application, the results shown do not have a significant difference, in other words, both manual calculations and those processed in the application show similar results. Below is a comparison table between the results of manual calculations and those in the Rapid Miner application on the Y variable.

Table 5. Comparison of manual and rapid miner

Age	BMI	Glucose	Cholesterol (Y) Manual	Cholesterol (Y) Rapid miner
49	29.62	60	0	0
67	25.75	87	0	0
48	23.62	68	0	0

E. Analysist of Results

Meanwhile, the comparison of the actual Y variable value from the testing (observation) data with the predicted Y variable value in the Rapid Miner application can be seen through the following graph.

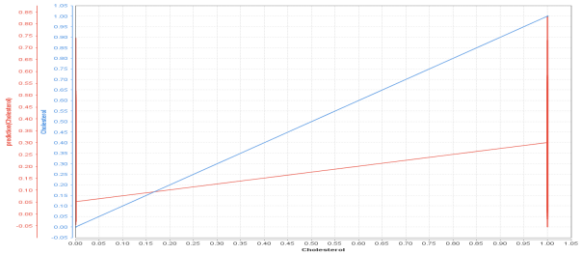

Figure 9. Graph of comparison of the value (Y) of observations with (Y) predictions

In general, in figure 9. it can be seen through the graphic image, the actual value of the Y variable (observation) is marked with a blue line, while the predicted value of the Y variable is marked with a red line. This graph will later be evaluated using the RMSE method to see how big the error value is. Furthermore, the evaluation of this model is to find the root mean squared error or RMSE, through the RapidMiner application, the RMSE value is 0.347 with a standard deviation of +/- 0.000.

CONCLUSION

Based on the results of the tests that have been carried out in this study, a conclusion can be drawn, namely: In this study by utilizing some data of patients with cholesterol disease that has been stored in the database using several attributes, namely age, BMI, glucose, and cholesterol. So by applying a linear regression algorithm can be done a prediction in the identification of cholesterol diseases based on functional relationships on the attributes in the data. The process of classifying cholesterol disease data using linear regression algorithms starts from the data selection stage (attributes used and determination of training data and data testing), the algorithm testing stage (linear regression), and the RMSE test stage (using split validation) so that it can provide information in the identification and prevention of cholesterol diseases. The process of processing cholesterol disease data using linear regression algorithms starts from modeling (manual calculation of linear regression algorithms), the process of testing data on rapidminers, and comparing the value of the results of manual calculations with the results of the data testing process on the rapidminer. Based on the results of the tests that have been carried out that the variables or attributes used in this study (age, BMI, glucose, and cholesterol) have a significant effect on this study, it is proven that using a linear regression algorithm is able to give good results with a Root Mean Squared Error value: 0.347 + /- 0.000. This is due to a correlation or functional relationship (cause and effect) between one variable (dependent or criteria) and other variables (independent or predictor). This testing process is carried out to identify cholesterol disease with a linear regression algorithm.

REFERENCE

[1] F. J. P. Putri, "Pengukuran Kadar Kolesterol Pada Jaringan Kulit Tiruan (Phantom) Menggunakan Teknik Diffuse Reflectance Spectroscopy," p. 120, 2017, [Online]. Available: http://repository.its.ac.id/47486/.

[2] B. Sofiandi, J. Raharjo, and K. Usman, "Identifikasi Pola Citra Iris Mata Untuk Mendeteksi Kelebihan Kadar Kolesterol Menggunakan Metode Gray Level Co-occurrence Matrix (glcm) Dan Decision Tree," *eProceedings …*, vol. 6, no. 3, pp. 10242–10249, 2019.

[3] A. Saputra, W. Broto, and L. B. R., "Deteksi Kadar Kolesterol Melalui Iris Mata Menggunakan Image Processing Dengan Metode Jaringan Syaraf Tiruan Dan Gray Level Co-Occurrence Matrix (Glcm)," vol. VI, pp. SNF2017-CIP-65-SNF2017-CIP-74, 2017, doi: 10.21009/03.snf2017.02.cip.09.

[4] M. A. SIDDIK, L. NOVAMIZANTI, and I. N. A. RAMATRYANA, "Deteksi Level Kolesterol melalui Citra Mata Berbasis HOG dan ANN," *ELKOMIKA J. Tek. Energi Elektr. Tek. Telekomun. Tek. Elektron.*, vol. 7, no. 2, p. 284, 2019, doi: 10.26760/elkomika.v7i2.284.

[5] M. A. C, "Oleh : DESTRI ARIANTI," 2010.

[6] H. A. D. Rani, E. Supriyati, and T. Khotimah, "DETEKSI IRIS MATA UNTUK MENENTUKAN KELEBIHAN KOLESTEROL MENGGUNAKAN EKSTRAKSI CIRI MOMENT INVARIANT DENGAN K-MEANS CLUSTERING Handini," *Pros. SNATIF*, pp. 287–292, 2014.

[7] M. Busthomi, N. Nafi'iyah, and N. Q. Nawafilah, "Sistem Pakar Diagnosa Penyakit Kolesterol Pada Remaja Dengan Metode Certainty Factor," *J. Process.*, vol. 15, no. 1, p. 23, 2020, doi: 10.33998/processor.2020.15.1.670.

[8] L. Mahanani, "Perbandingan 3 Metode dalam Data Mining untuk Penentuan Kadar Kolesterol di RSUD dr.Moewardi Surakarta," 2016, [Online]. Available: http://eprints.ums.ac.id/43275/.

[9] S. Puspitorini, "Uji Korelasi Dan Analisis Clustering Gula Darah Puasa, Kolesterol Total, Trigliserida, Serta Asam Urat," *FORTECH (Journal Inf. Technol.*, vol. 1, no. 1,

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.28/E/KPT/2019, September 26, 2019. Published by LPPM Universitas Nusa Mandiri

VOL. 8. NO. 1 AUGUST 2022
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v8i1.3328

pp. 49–54, 2017.

[10] M. K. Harahap and N. Khairina, "Jaringan Syaraf Tiruan Perceptron untuk Pengenalan Pola Makanan Sehat Rendah Kolesterol," pp. 209–214, 2018, doi: 10.31227/osf.io/acmj8.

[11] D. Haryadi, D. Marini, U. Atmaja, A. R. Hakim, and N. Suwaryo, "IDENTIFIKASI TINGKAT RESIKO PENYAKIT STROKE MENGGUNAKAN ALGORITMA REGRESI LINEAR BERGANDA," vol. 1, no. November, pp. 1198–1207, 2021.

[12] R. A. Permata, D. Triyanto, and Ilhamsyah, "Aplikasi Penyusun Menu Makanan Untuk Pencegahan Hiperkolesterolemia Menggunakan Algoritma Genetika," *J. Coding Sist. Komput. Untan*, vol. 04, no. 2, pp. 96–106, 2016.

[13] D. Haryadi, "Penerapan Algoritma K-Means Clustering Pada Produksi Perkebunan Kelapa Sawit Menurut Provinsi," *J. ICT (Informatics ...*, vol. 1089, pp. 1–15, 2021, [Online]. Available: http://ejournal.akademitelkom.ac.id/j_ict/index.php/j_ict/article/download/71/57.

[14] D. Marini, U. Atmaja, W. Witanti, and A. I. Hadiana, "Pembangunan Sistem Informasi Biaya Proyek pada PT . Skyline Semesta Menggunakan Metode Earned Value Management (EVM)," *J. Univ. Jendral Achmad Yani*, vol. 2, no. 2, pp. 3–8, 2018.

[15] D. Haryadi and R. Mandala, "Prediksi Harga Minyak Kelapa Sawit Dalam Investasi Dengan Membandingkan Algoritma Naïve Bayes, Support Vector Machine dan K-Nearest Neighbor," *IT Soc.*, vol. 4, no. 1, 2019, doi: 10.33021/itfs.v4i1.1181.

[16] N. Suwaryo, D. Haryadi, D. Marini, U. Atmaja, and A. R. Hakim, "Analisa Data Mining Menggunakan Algoritma Apriori Untuk Mencari Pola Pemakaian Obat," vol. 1, no. November, pp. 1208–1217, 2021.

[17] D. M. U. Atmaja, "Penerapan Algoritma K-Nearest Neighbor Untuk," vol. 1, no. November, pp. 199–208, 2019.

[18] D. M. U. Atmaja, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Penyebaran Covid-19 di Provinsi Jawa Barat," vol. 1, no. November, pp. 1218–1226, Aug. 2021.

[19] D. M. U. Atmaja and R. Mandala, "Analisa Judul Skripsi untuk Menentukan Peminatan Mahasiswa Menggunakan Vector Space Model dan Metode K-Nearest Neighbor," *IT Soc.*, vol. 4, no. 2, pp. 1–6, 2020, doi: 10.33021/itfs.v4i2.1182.

[20] D. Haryadi, "Implementation of K-Medoids Clustering Algorithm for Grouping Palm Oil Exports by Destination Country," pp. 129–134, 2021.

[21] D. Haryadi and D. M. U. Atmaja, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Tingkat Risiko Penyakit Jantung," *J. Informatics Commun. Technol.*, vol. 3, no. 2, pp. 51–66, 2021, doi: 10.52661/j_ict.v3i2.85.