OPEN ACCESS
J. OF ICT, VOL. 5, NO. 1, PP.89-100, JUNE 2023

ISSN : 2686-1089 (ONLINE)
DOI: 10.52661

Journal of Informatics and Communications Technology (JICT)

Prediction of Liver Disease Using a Linear Regression Algorithm

Deny Haryadi [*1], Dewi Marini Umi Atmaja [2], Arif Rahman Hakim [3]

Prodi S1 Teknologi Informasi Kampus Jakarta, Universitas Telkom
Jln. Daan Mogot KM 11, Cengkareng, Jakarta Barat Indonesia

[*1] denyharyadi@telkomuniversity.ac.id

Prodi Sarjana Bisnis Digital, Universitas Medika Suherman
Jln. Raya Industri Pasir Gombong, Cikarang Utara-Bekasi, Jawa Barat Indonesia

[2] dewi@medikasuherman.ac.id
[3] arif@medikasuherman.ac.id

Abstract

The liver is the most essential organ in the human body. Hepatitis is one such disorder affecting the liver and is a global health issue, including in Indonesia. Liver disease is an inflammatory condition of the liver that can be triggered by genetic factors, viral infections, alcohol consumption, and the use of certain drugs. In principle, prevention of hepatitis or liver disease can be done by adopting a healthy lifestyle. In addition, early detection is also very important in preventing death in those affected by this disease. One method for early detection is through the application of data mining, which can help predict and reduce mortality in patients affected by this disease. Linear regression is a data mining technique utilized to predict the dependent variable or outcome based on the independent variable or predictor. The study conducted tests on this algorithm and obtained a Root Mean Squared Error of 0.349 +/- 0.000. This indicates the presence of a correlation or functional relationship (cause and effect) between the dependent variable (criterion) and the independent variable (predictor). The purpose of this testing process is to detect liver disease using the linear regression algorithm.

Keywords: Liver Disease, Prediction, Data Mining, Linear Regression Algorithm.

I. INTRODUCTION

THE liver is the most essential organ in the human body. It is the largest organ and is often referred to as an exocrine gland because it secretes bile into the intestines. With its various functions and benefits for the body, the liver is also susceptible to liver diseases. Hepatitis is one such disorder affecting the liver and is a global health issue, including in Indonesia. There are five different types of hepatitis, namely hepatitis A, hepatitis B, hepatitis C, hepatitis D, and hepatitis E [1]. Liver disease is an inflammatory condition of the liver that can be triggered by genetic factors, viral infections, alcohol consumption, and the use of certain drugs. According to a global report from the World Health Organization (WHO), the death rate from the liver disease continues to increase every year, with 1.75 million people newly infected with the hepatitis virus or especially hepatitis C. will have a bad impact on the health of the sufferer, if you have been exposed but don't want to be examined at a health service, it is feared that the hepatitis virus or liver disease C will have a worse/chronic impact, there

are several stages of liver damage [2]. Along with the development of science and information technology, data mining has become an interesting topic in the field of information systems, especially in processing health data. In principle, prevention of hepatitis or liver disease can be done by adopting a healthy lifestyle. In addition, early detection is also very important in preventing death in those affected by this disease. One method for early detection is through the application of data mining, which can help predict and reduce mortality in patients affected by this disease. Data mining technology has emerged because of decision-making systems and supporting technology infrastructure in the health sector. The use of data mining has benefits in providing solutions to decision-makers in health to carry out early prevention [3]. In data analysis, linear regression is utilized to understand how the dependent or outcome variable can be predicted based on the independent or predictor variables. Regression analysis plays a crucial role in determining whether changes in the dependent variable can be attributed to increases or decreases in the independent variable. It helps to determine if an increase in the dependent variable can be achieved by increasing the independent variable, and vice versa. Regression analysis aids in making informed decisions regarding the relationship and impact of variables on each other [4]. This research will be a reference in conducting prediction analysis of liver disease. Through data processing, it is hoped that information and knowledge will be obtained that can reveal new potential and increase our understanding. In addition, this research can identify new opportunities or strategic plans in data processing, as well as serve as a guide in making decisions in prevention efforts.

II. LITERATURE REVIEW

A. Hepatitis or Liver Disease

Hepatitis, also known as liver disease, is characterized by liver inflammation resulting from various causes such as toxins (chemicals, drugs) or infectious agents like viruses [5]. Hepatitis is classified into two types based on the cause: infectious and non-infectious hepatitis. Non-infectious hepatitis refers to liver inflammation caused by factors other than infection sources, including chemicals, excessive alcohol consumption, and drug abuse. Non-infectious hepatitis, including drug-induced hepatitis, is not classified as an infectious disease, because hepatitis is caused by inflammation and not by infectious agents such as fungi, bacteria, micro-organisms, and viruses. This disease is found in almost all countries in the world. Hepatitis is not a direct cause of death, but hepatitis causes problems in productive age [6]. Hepatitis that persists for more than 6 months is referred to as "chronic hepatitis," while hepatitis that lasts for a shorter duration is known as "acute hepatitis." Hepatitis can be caused by either viral or non-viral factors. Non-viral causes include alcohol abuse and drug consumption. On the other hand, viral causes encompass various types such as hepatitis A, B, C, D, and E viruses. Additionally, other viruses like the mumps virus, rubella virus, cytomegalovirus, Epstein-Barr virus, and herpes virus can also contribute to the development of hepatitis [7].

B. Data Mining

Data Mining is a crucial stage in the process of knowledge discovery in databases (KDD). It involves analyzing data to uncover hidden patterns or relationships that were previously unknown. These patterns and relationships provide valuable knowledge that was not previously available or recognized. Data mining is an interdisciplinary field that combines various computer science disciplines, aimed at discovering novel patterns from vast datasets. It encompasses techniques derived from artificial intelligence, machine learning, statistics, and database systems to extract valuable insights from large data warehouses. Data mining involves the extraction of new information from extensive datasets, contributing to decision-making processes. It is often referred to as knowledge discovery and involves techniques such as building models based on existing data to identify patterns do not present in the stored database [8]. Data mining is also utilized for prediction purposes. Additionally, data grouping or clustering can be performed as part of the data mining process. The objective is to uncover the underlying patterns within the available data, which can be applicable universally. It is crucial to identify anomalies in transactional data to determine appropriate subsequent actions. These efforts are aimed at supporting the operational activities of a company and ultimately achieving the company's overall objectives. Data mining is a process that utilizes one or more machine learning techniques to automatically analyze and extract valuable knowledge from data.

C. *Prediction*

Prediction is the initial part of the process of planning. In every decision-making that concerns the situation in the future, there will be predictions that underlie the decision. Forecasting or forecasting/prediction is an attempt to predict conditions in the future through testing conditions in the past. In the prediction, the data that is processed is historical data which is used as reference material data coupled with simulation data which can be changed according to the possibilities that can occur. Prediction knows the estimated value of an item in the future. Or prediction is the need for predictions to increase in line with management's desire to provide a fast and appropriate response to opportunities and changes in the future [9]. Qualitative predictions are predictions based on a party's opinion (judgment forecast) and quantitative predictions are predictions based on past data (historical data) and can be made in the form of numbers commonly referred to as time series data. Quantitative predictions are nothing but predictions while qualitative predictions are forecasts, forecasts are seen as a process of predicting future variables based on the data of the variables concerned in the past. Past data is systematically combined through certain methods and processed for conditions in the future. The objective of forecasting is to provide decision-makers and policymakers with insights into potential uncertainties and risks that may arise, allowing them to incorporate such factors into their planning processes [10].

D. *Linear Regression*

Linear regression is employed to determine the relationship between dependent variables or criteria and independent variables or predictors, individually [11]. The utilization of regression analysis helps in determining whether changes in the dependent variable can be achieved by increasing or decreasing the independent variable. It allows decision-makers to assess if increasing the independent variable leads to an increase or decrease in the dependent variable, and vice versa. This analysis provides valuable insights into the impact and relationship between variables, enabling informed decision-making. Least square method (least square method): the most popular method for establishing simple linear regression equations.

E. *Rapid Miner*

RapidMiner is software developed by the Blanchardstown Institute of Technology and Ralf Klinkenberg from rapid-i.com. It features a user-friendly Graphical User Interface (GUI) that simplifies the software's usability. This open-source software is built using Java programming language and operates under the GNU Public License, making it compatible with various operating systems [12]. With RapidMiner, users do not require extensive coding skills as it provides a wide range of built-in functionalities. The software is primarily designed for data mining purposes and offers comprehensive models such as Bayesian models, tree induction, and neural networks. RapidMiner encompasses numerous methods including classification, clustering, association, and more.

III. RESEARCH METHOD

A. *Research Stages*

To carry out analysis and find data patterns that can facilitate research and achieve the desired goals, a series of steps is made in the research stage which will be carried out as follows:

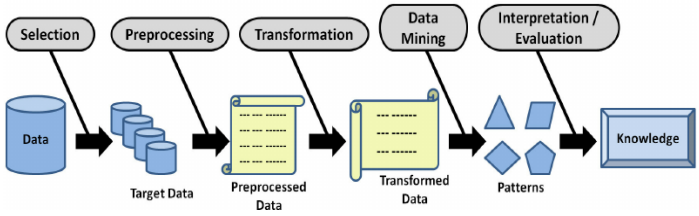

Fig. 1. Research Stages

B. Data on Liver Disease (Hepatitis)

In this study, data regarding liver disease or hepatitis were collected using various methods, which involved attributes or variables and the value of the object being studied. This activity includes certain variations determined by the researcher so that it can be studied and then conclusions drawn. The data collection technique used in this study was taken from the Kaggle.com data source. To carry out analysis and look for patterns that can facilitate research and achieve the desired goals, the steps in the data processing and dataset compilation are arranged. This can be explained as follows:

1. Data Selection

The selection phase involves choosing specific data from an existing operational dataset before proceeding to the data mining stage. During this phase, a series of steps will be undertaken. Data samples are chosen by applying attribute parameters to liver disease or hepatitis data, selecting the dataset with the largest amount of data. The aim is to ensure that the selected data is appropriate for use in modeling or testing data processes. After this dataset is grouped, it is found that the amount of data on liver disease or hepatitis will be processed using a linear regression algorithm. The attributes to be used and analyzed are selected after selection, considering that there are several irrelevant attributes in the initial data.

2. Preprocessing

In this study, the data can be processed and structured into datasets, which are then partitioned into training data and testing data based on research requirements. The objective is to develop a model with strong generalizability for data classification. The training data is employed to train predictions and execute algorithm functions in alignment with the research objectives. On the other hand, testing data is utilized to assess the accuracy or performance of the data. Moreover, during the data cleaning phase, the data to be utilized undergoes a procedure to ensure its cleanliness. This involves eliminating any missing values or duplicated entries, as well as verifying data consistency and rectifying any errors. The data cleaning process is typically performed manually, employing spreadsheet software like Microsoft Excel.

3. Data Transformation

Data transformation involves converting the original data format into a standardized format that can be easily interpreted by the algorithm implemented in the program or device being used.

C. *Test Evaluation*

The test method is used to check the results of calculations that have been analyzed and evaluate the performance of the function. After performing manual calculations, the data is tested using the RapidMiner tool to ensure that the calculation results can provide accurate predictions and produce RSME values that match the data testing. Evaluation is done by observing and analyzing the results of the algorithm used, to verify that the test produces results that are consistent with the discussion. Each attribute value and model that has been created is examined, and evaluation is carried out by observing and analyzing the results of the algorithm, to ensure that the test produces an RSME value that is accurate and by the discussion. The purpose of the test is to measure the level of accuracy of each model that has been proposed.

IV. RESULTS AND DISCUSSION

A. Linear Regression Algorithm Modeling

In this research, the Linear Regression algorithm was employed to detect the presence of liver disease (hepatitis) and produce Root Mean Square Error (RMSE) values, along with predictions that can aid in making decisions related to the disease status in patients. The data utilized for this study was derived from historical

records sourced from the website Kaggle.com. The dataset involved multiple attributes or variables, such as age, total cholesterol (totChol), glucose levels, and liver condition.

B. Split Validation

In this validation approach, the data will be randomly partitioned into two subsets: the training data and the testing data. The Split Validation technique will be employed, where training experiments will be conducted based on predetermined split ratios. The portion of data not utilized for training will serve as the testing data.

TABLE I
LIVER DISEASE DATASET

age	**totChol**	**glucose**	**TenYearCHD**
39	195	77	0
46	250	76	0
48	245	70	0
61	225	103	1
46	285	85	0
43	228	99	0
63	205	85	1
45	313	78	0
52	260	79	0
43	225	88	0
….	….	….	….
68	176	79	1
50	313	86	1
51	207	68	0
48	248	86	0
52	269	107	0

C. Linear Regression Count

TABLE II
LINEAR REGRESSION CALCULATION

No	**Aged (X_1)**	**Tot Chol (X_2)**	**Glucose (X_3)**	**Ten Year CHD (Y)**	**$X_1 Y$**	**X2 Y**	**X3 Y**	**X1 X2 X3**	**X_1^2**	**X_2^2**	**X_3^2**
1	39	195	77	0	0	0	0	585585	1521	38025	5929
2	46	250	76	0	0	0	0	874000	2116	62500	5776
3	48	245	70	0	0	0	0	823200	2304	60025	4900
4	61	225	103	1	61	225	103	1413675	3721	50625	10609
….	….	….	….	….	….	….	….	….	….	….	….
100	63	273	56	0	0	0	0	963144	3969	74529	3136
Total	4961	23708	8407	18	954	4286	1554	100003046	253871	5771656	777301

$$\boldsymbol{b_1} = \frac{n\sum(x_1 y) - (\sum x_1)(\sum y)}{n(\sum x_1)^2 - (\sum x_1)^2}$$

$$\boldsymbol{b_1} = \frac{100(954) - (4961)(18)}{100(253871) - (4961)^2}$$

$$\boldsymbol{b_1} = \frac{95400 - 89298}{25387100 - 24611521}$$

$$\boldsymbol{b_1} = \frac{6102}{775579} = \mathbf{0,0078676705}$$

$$\boldsymbol{b_2} = \frac{n\sum(x_2y) - (\sum x_2)(\sum y)}{n(\sum x_2)^2 - (\sum x_2)^2}$$

$$\boldsymbol{b_2} = \frac{100(4286) - (23708)(18)}{100(5771656) - (23708)^2}$$

$$\boldsymbol{b_2} = \frac{428600 - 426744}{577165600 - 562069264}$$

$$\boldsymbol{b_2} = \frac{1856}{15096336} = \mathbf{0,0001229437}$$

$$\boldsymbol{b_3} = \frac{n\sum(x_3y) - (\sum x_3)(\sum y)}{n(\sum x_3)^2 - (\sum x_3)^2}$$

$$\boldsymbol{b_3} = \frac{100(1554) - (8407)(18)}{100(777301) - (8407)^2}$$

$$\boldsymbol{b_3} = \frac{155400 - 151326}{77730100 - 70677649}$$

$$\boldsymbol{b_3} = \frac{4074}{7052451} = \mathbf{0,0005776715}$$

$$\boldsymbol{a} = \frac{\sum y - b_1\sum x_1 + b_2\sum x_2 + b_3\sum x_3}{n}$$

$$\boldsymbol{a} = \frac{18 - \left((0{,}0078676705) * (4961)\right) + \left((0{,}0001229437) * (23708)\right) + ((0{,}0005776715) * (8407))}{100}$$

$$\boldsymbol{a} = \frac{18 - 39{,}0315133505 + 2{,}9147492396 + 4{,}8564843005}{100}$$

$$\boldsymbol{a} = \frac{18 - 46{,}8027468906}{100}$$

$$\boldsymbol{a} = \frac{-28{,}8027468906}{100} = \mathbf{-0,2880274689}$$

TABLE III
PREDICTION OF LIVER DISEASE

age	TotChol	Glucose	TenYearCHD
42	250	79	?
40	290	73	?
66	278	84	?

D. Calculating Linear Regression Equations

Y = a + b 1 X 1 + b 2 X 2 + b 3 X 3

y= -0,2880274689+(0,0078676705 x 42)+(0,0001229437 x 250)+(0,0005776715 x 79) = 0.1187866656

y= -0,2880274689+(0,0078676705 x 40)+(0,0001229437 x 290)+(0,0005776715 x 73) = 0.1045030436

y= -0,2880274689+(0,0078676705 x 66)+(0,0001229437 x 278)+(0,0005776715 x 84) = 0.3139415387

E. Data Testing Process (RapidMiner)

After completing data collection, the next step is to create data modeling in RapidMiner. This modeling will use a linear regression algorithm as the model. The first step is to enter liver disease data into RapidMiner for processing.

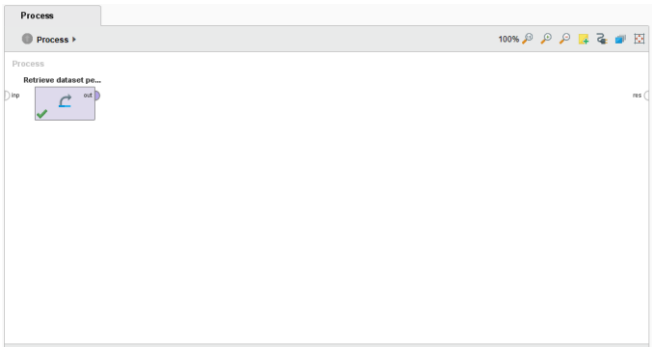

Fig. 2. Process of Entering Data in Rapidminer

In this research, the dataset was split into two parts: 90% for training data and 10% for testing data, utilizing the split validation method.

Fig. 3. Data Sharing Process with Split Validation in Rapidminer

The subsequent stage involves configuring the split validation parameter in RapidMiner to effectively segregate the training and testing data.

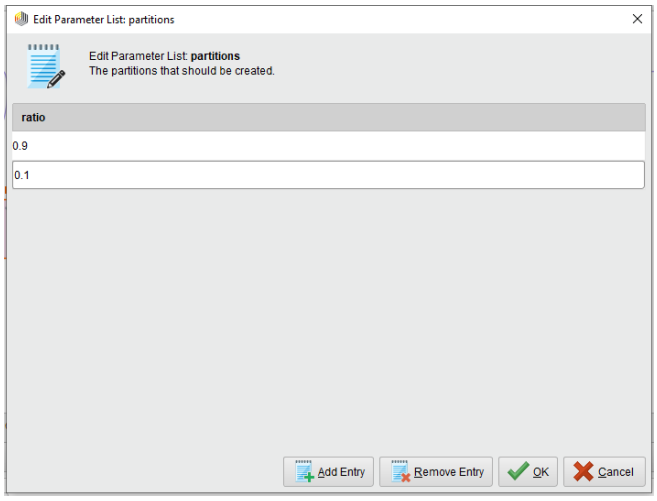

Fig. 4. The Process of Sharing Training Data and Data Testing Using Split Validation

The subsequent procedure involves implementing the linear regression algorithm in RapidMiner to observe the predicted outcomes.

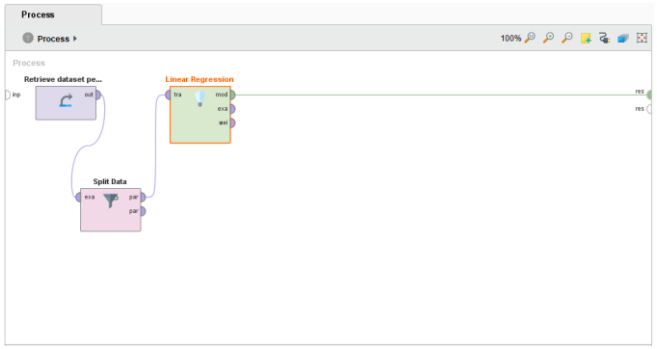

Fig. 5. The Process of Entering the Liner Regression Algorithm in Rapidminer

Performing attribute selection aims to compare the predicted results from RapidMiner with manual calculation results, as well as test results at RapidMiner.

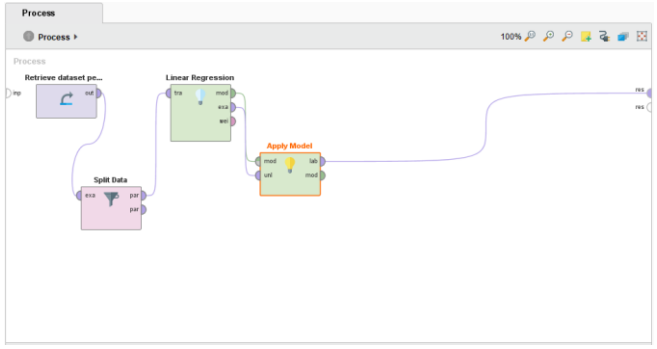

Fig. 6. Model Evaluation Process in Rapidminer

At this stage, input training data and testing data are carried out to produce predictions on class attributes.

Row No.	TenYearCHD	prediction(T...	age	totChol	glucose
1	0	0.042	39	195	77
2	0	0.115	46	250	76
3	0	0.121	48	245	70
4	1	0.288	61	225	103
5	0	0.139	46	285	85
6	0	0.124	43	228	99
7	1	0.269	63	205	85
8	0	0.125	45	313	78
9	0	0.175	52	260	79
10	0	0.104	43	225	88
11	0	0.151	50	254	76
12	0	0.062	43	247	61
13	0	0.105	46	294	64
14	0	0.105	41	332	84
15	1	0.028	38	221	70
16	0	0.121	48	232	72
17	1	0.147	46	291	89
18	0	0.035	38	195	78
19	0	0.039	41	195	65
20	0	0.081	42	190	85

Fig. 7. Prediction Results

Once the predictions have been obtained, the subsequent stage is to evaluate the accuracy of the prediction results.

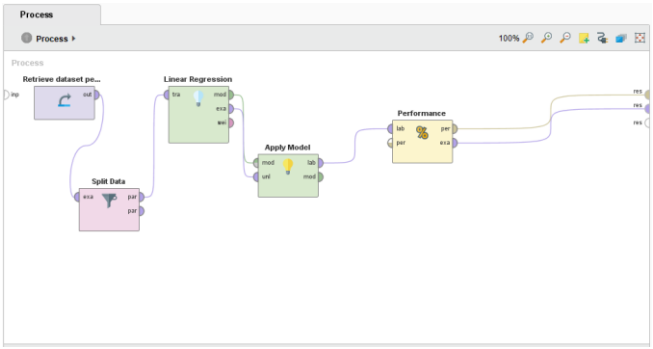

Fig. 8. Search Process Root Mean Squared Error

To facilitate the interpretation of liver data, performance tools are utilized to determine the Root Mean Squared Error (RMSE). Here are the obtained results.

root_mean_squared_error

root_mean_squared_error: 0.349 +/- 0.000

Fig. 9. Root Mean Square Error Test Results

In the second stage, the linear regression algorithm was implemented using the RapidMiner tool. The following are the steps involved in implementing the linear regression algorithm: Predict test data using RapidMiner and produce predicted confidence values. Determine performance by generating output to find Root Mean Squared Error (RMSE). In split validation modeling, there are two primary components: the training section, utilized for

the classification algorithm, and the testing section. The testing section employs the Apply Model feature to apply the model to the test data, along with the Performance feature to display the Root Mean Square Error (RMSE).

F. Discussion

On the 100 record dataset data training after done calculation to find the values of X12, X22, X32, X1Y, X2Y, X3Y, and X1X2X3 the overall results of each are for the total value of the variable X1 is 4961, the total value of the variable X2 is 23708, the total value of the variable X3 is 8407, the total the value of variable Y is 18, the total value of X12 is 253871, the total value of X22 is 5771656, the total value of X32 is 777301, the total value of X1Y is 954, the total value of X2Y is 4286, the total value of X3Y is 1554, the total value of X1X2X3 is 100003046. The sum of the above sums is used to convert the formula by finding the values of a factor, b1 factor, b2 factor, and b3 factor such that the value of a is 1, the value of b1 is 0, the value of a is 0. the value of b2 is 0, and the value of b3 is 0. These three factor values will be used in applying the equations of the multiple linear regression algorithm. From the factor values above, we can use a simple mathematical model to determine the linear regression algorithm formula using the model Y = a + (b1X1) + (b2X2) + (b3X3), where variable X1 is Age, variable X2 is tot Chol and variable X3 is Glucose. Next, the double Linear Regression equation will be employed to predict the data within the test dataset. This Linear Regression equation will be utilized to predict three specific records within the test data. Upon performing manual calculations and comparing them with the results from the RapidMiner application, no significant discrepancies were observed. In other words, both the manual calculations and the processes conducted within the application yielded similar outcomes. The following table presents a comparison between the results obtained from manual calculations and the results derived from the RapidMiner application for variable Y.

TABLE IV
COMPARISON OF MANUAL & RAPIDMINER

Age (X_1)	**Tot Chol (X_2)**	**Gluclose (X_3)**	**Heart (Y) Manual**	**Heart (Y) Rapid miner**
42	250	79	0.1187866656	0.120
40	290	73	0.1045030436	0.104
66	278	84	0.3139415387	0.309

The graph below illustrates a comparison between the observed actual values of variable Y from the testing data and the predicted values of variable Y generated by the RapidMiner application.

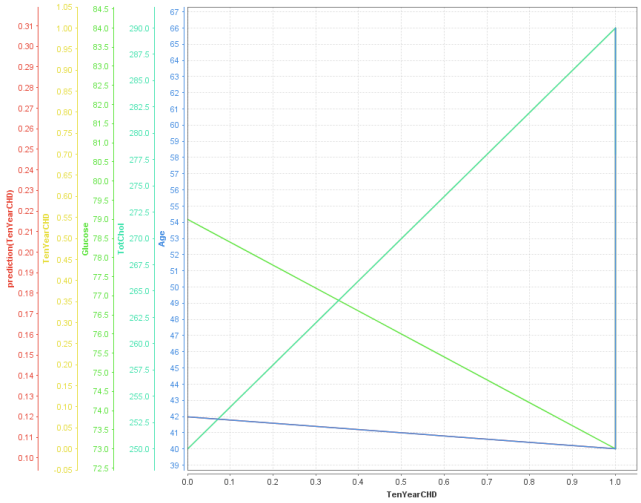

Fig. 9. Graph of Comparison of (Y) Observation and (Y) Predictions

In general, the presented graph depicts the actual (observed) values of variable Y represented by the yellow line, while the red line represents the predicted values of variable Y. Based on these data, it can be concluded that the predicted values fall within a similar range as the actual (observed) values of Y. To assess the model's performance, the root mean squared error (RMSE) method is employed to measure the extent of error in the values. In the evaluation of this model, the RMSE value was calculated using the RapidMiner application, resulting in an RMSE value of 0.349 with a standard deviation of +/- 0.000. Based on the conducted tests, it can be concluded that the variables or attributes utilized in this study, namely age, total cholesterol (totChol), and glucose, have a significant impact on the research outcome. This assertion is supported by the utilization of the linear regression algorithm, which yields a favorable Root Mean Squared Error (RMSE) value of 0.349 +/- 0.000, indicating promising results. These outcomes can be attributed to the presence of a correlation or functional relationship (cause and effect) between the dependent variable or criteria and the independent variables or predictors. The testing process was conducted with the objective of identifying liver disease using the linear regression algorithm.

V. Conclusion

Based on the findings obtained from the conducted tests in this study, several conclusions can be drawn, as outlined below: In this study, data from patients suffering from liver disease were used from the database. The data utilizes the attributes of age, total cholesterol, and blood sugar levels. By using a linear regression algorithm, this study can make predictions to identify liver disease based on the functional relationship attributes in the data. The classification of liver disease data using a linear regression algorithm encompasses multiple stages or steps. The first step is data selection, where the attributes to be used and training and test data are determined. The second step is testing the linear regression algorithm on the data. The third step is RMSE testing using disaggregated validation, which provides important information in identifying and preventing liver disease. The processing of liver disease data using a linear regression algorithm involves several stages. The first stage is modeling, the first stage involves manual calculation of the linear regression algorithm. The second stage encompasses testing the data using RapidMiner. Finally, the values obtained from manual calculations are compared with the results derived from the data testing process in RapidMiner.

REFERENCES

[1] N. T. Rahman, "Analisa Algoritma Decision Tree Dan Naïve Bayes Pada Pasien Penyakit Liver," *J. Fasilkom*, vol. 10, no. 2, pp. 144–151, 2020, doi: 10.37859/jf.v10i2.2087.

[2] B. Tri Rahmat Doni, S. Susanti, and A. Mubarok, "PENERAPAN DATA MINING UNTUK KLASIFIKASI PENYAKIT HEPATOCELLULAR CARCINOMA MENGGUNAKAN ALGORITMA NAÏVE BAYES," *J. RESPONSIF*, vol. 3, no. 1, pp. 12–19, 2021.

[3] R. Annisa, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019, [Online]. Available: https://jurnal.kaputama.ac.id/index.php/JTIK/article/view/141/156

[4] D. Haryadi, D. Marini Umi Atmaja, A. Rahman Hakim, and N. Suwaryo, "Identifikasi Tingkat Resiko Penyakit Stroke Menggunakan Algoritma Regresi Linear Berganda," *SNTEM*, vol. 1, no. 1, pp. 1198–1207, 2021.

[5] W. D. Septiani, "Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis," *J. Pilar Nusa Mandiri*, vol. 13, no. 1, pp. 76–84, 2017.

[6] I. Irmawati, K. Widianto, F. Aziz, A. Rifai, and A. Rahmawati, "Implementasi artificial neural network dalam mendeteksi penyakit hati (liver)," *J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 6, no. 1, pp. 193–198, 2022, doi: 10.52362/jisamar.v6i1.694.

[7] T. Septiani Nurfauzia Koeswara, M. Sukrisno Mardiyanto, and M. Abdul Ghani, "Penerapan Particle Swarm Optimization (PSO) Dalam Pemilihan Atribut Untuk Meningkatkan Akurasi Prediksi Diagnosispenyakit Hepatitis Dengan Metode Naive Bayes," *J. Speed – Sentra Penelit. Eng. dan Edukasi*, vol. 12, no. 1, pp. 1–10, 2020.

[8] 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, pp. 28–38, 2019, doi: 10.33021/itfs.v4i1.1181.

[9] A. Mudgal, K. Gopalakrishnan, and S. Hallmark, "Prediction of emissions from biodiesel fueled transit buses using artificial neural networks," *Int. J. Traffic Transp. Eng.*, vol. 1, no. 2, pp. 115–131, 2011.

[10] H. W. Herwanto, T. Widiyaningtyas, and P. Indriana, "Penerapan Algoritme Linear Regression untuk Prediksi Hasil Panen Tanaman Padi," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 364, 2019, doi: 10.22146/jnteti.v8i4.537.

[11] E. Triyanto, H. Sismoro, and A. Dwi Laksito, "Implementasi Algoritma Regresi Linear Berganda Untuk Memprediksi Produksi

Padi Di Kabupaten Bantul," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 4, no. 2, pp. 73–86, 2019, doi: 10.36341/rabit.v4i2.666.

[12] D. Haryadi, D. Marini Umi Atmaja, A. Rahman Hakim, and W. Witanti, "Classification of Drug Effectiveness Based on Patient's Condition Using Text Mining With K-Nearest Neighbor," *9th Int. Conf. ICT Smart Soc. Recover Together, Recover Stronger Smarter Smartization, Gov. Collab. ICISS 2022 - Proceeding*, vol. 9, no. 1, pp. 1–9, 2022, doi: 10.1109/ICISS55894.2022.9915156.