Informatics and Software Engineering
p-ISSN/e-ISSN: 2988-2818/2988-2222
Homepage: https://sanscientific.com/journal/index.php/ise
2**(2) 76-88 (2024)**
https://doi.org/10.58777/ise.v2i2.328

Research Article

Linear Regression Analysis to Predict the Percentage of Smoking in the Population Age 15 Years and Over

Muhammad Alfata Riziq Shihab[1*], Musriatun Napiah[2]
[1,2] Fakultas Teknik dan Informatika Universitas Bina Sarana Informatika

Received: October 16, 2024; Revision: October 30, 2024;
Accepted: November 10, 2024; Available Online: November 15, 2024;

Abstract
Smoking is a serious public health problem in many countries, including Indonesia, as it can cause diseases such as lung cancer, heart disease and respiratory disorders. According to data from the Ministry of Health of the Republic of Indonesia, the prevalence of smoking among the population aged 15 years and above is still high. This study uses secondary data from the Central Bureau of Statistics (BPS) that records the percentage of smoking in the population aged 15 years and above by age group from 2019 to 2023. With this data, a linear regression algorithm was applied using RapidMiner to predict the percentage of smoking in 2024. The analysis showed that out of 11 age groups, 6 age groups experienced an increase in smoking percentage from the previous year: 15-19, 20-24, 25-29, 30-34, 55-59, and 60-64. Meanwhile, the other 5 age groups experienced a decrease: 35-39, 40-44, 45-49, 50-54, and 65+. Evaluation of the prediction model using root mean squared error (RMSE) resulted in a value of 0.4 +/- 0.000. This RMSE value indicates that the model has a low error rate, making it reliable for predicting the percentage of smoking by age group in Indonesia.

Keywords: Smoking, Prediction, Linear Regression, RapidMiner

Abstrak
Merokok adalah masalah kesehatan masyarakat yang serius di banyak negara, termasuk Indonesia, karena dapat menyebabkan penyakit seperti kanker paru-paru, penyakit jantung, dan gangguan pernapasan. Menurut data Kementerian Kesehatan Republik Indonesia, prevalensi merokok di kalangan penduduk usia 15 tahun ke atas masih tinggi. Penelitian ini menggunakan data sekunder dari Badan Pusat Statistik (BPS) yang mencatat persentase merokok penduduk usia 15 tahun ke atas berdasarkan kelompok umur dari tahun 2019 hingga 2023. Dengan data ini, algoritma regresi linear diterapkan menggunakan RapidMiner untuk memprediksi persentase merokok pada tahun 2024. Hasil analisis menunjukkan bahwa dari 11 kelompok umur, terdapat 6 kelompok umur yang mengalami peningkatan persentase merokok dari tahun sebelumnya: 15-19, 20-24, 25-29, 30-34, 55-59, dan 60-64. Sementara itu, 5 kelompok umur lainnya mengalami penurunan: 35-39, 40-44, 45-49, 50-54, dan 65+. Evaluasi model prediksi menggunakan root mean squared error (RMSE) menghasilkan nilai 0.884 +/- 0.000. Nilai RMSE ini menunjukkan bahwa model memiliki tingkat kesalahan yang rendah, sehingga dapat diandalkan untuk memprediksi persentase merokok berdasarkan kelompok umur di Indonesia.

Kata Kunci: Merokok, Prediksi, Regresi Linear, RapidMiner

How to cite: Muhammad Alfata Riziq Shihab, Musriatun Napiah. Linear Regression Analysis to Predict the Percentage of Smoking in the Population Age 15 Years and Over. *Informatics and Software Engineering*, *2*(2), 76–89. https://doi.org/10.58777/ise.v2i2.328

[*]Corresponding author: Muhammad Alfata Riziq Shihab (alfatariziq@gmail.com)

1. Introduction

In many countries, including Indonesia, smoking is a significant public health problem. Smoking can cause lung cancer, heart disease and respiratory problems. According to data from the Ministry of Health of the Republic of Indonesia, the prevalence of smoking among the population aged 15 years and above is still quite high.

In previous research in a journal with the title "Analysis of the Influence of Socio-Economic Factors, Income and Cigarette Prices on Cigarette Consumption in Indonesia." This study aims to evaluate the impact of socio-economic conditions, individual income, and cigarette costs on cigarette consumption patterns in Indonesia. The study adopted a quantitative method, utilizing multiple linear regression techniques. The study findings indicated that socio-economic aspects such as gender and age, the amount of individual income, and the price of cigarette products did not show a significant impact on the level of cigarette consumption in Indonesia (Marianti & Prayitno, 2020).

Many journals examine the factors that lead to smoking and its impact on public health. However, there needs to be more literature examining how specific age groups influence smoking trends over time. Data analysis using statistical and machine learning methods is becoming increasingly important in understanding smoking patterns and predicting future smoking percentages. One of the widely used methods is the linear regression algorithm, which can help in identifying smoking percentages.

In today's digital era, data analysis using statistical and machine learning methods is becoming increasingly important in understanding smoking patterns and predicting future smoking percentages. One method that is widely used is the linear regression algorithm. Linear regression is a prediction method that uses a straight line to describe the relationship between two or more variables (Novianty et al., 2021). We utilize data from 2019 to 2023; this study aims to explore the effectiveness of linear regression algorithms in predicting the percentage of smoking by age group; this study will use RapidMiner as an analysis tool to perform the prediction.

RapidMiner is an open-source software or application with an AGPL (GNU Affero General Public License) license that is useful for Data Mining processing and was created by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at the Artificial Intelligence Unit of the University of Dortmund. RapidMiner is an application that is a solution for Data Mining analysis, text mining, and also predictive analysis. RapidMiner utilizes various descriptive and predictive techniques to provide knowledge to users so that the results obtained can later be used for better and more precise decision-making (Dahlia & Andri, 2020).

This research, using RapidMiner, is expected to provide accurate and useful results in understanding smoking trends by age group. A deeper understanding of this trend could help the government and health agencies formulate more effective policies to reduce smoking prevalence. In addition, this study can also contribute to the field of public health, especially in efforts to prevent and overcome smoking problems in Indonesia.

2. Literature Review

2.1 Smoking

According to Hamdan's view, smoking activity can be defined as the act of inhaling and exhaling smoke containing an addictive substance called nicotine. This behavior is a form of addiction that is formed through a complex process and is reinforced by various factors. These factors include personal habits that have been formed, positive perceptions of cigarettes, and support from the surrounding environment that allows or even encourages smoking activities. In addition, the lack of determination or strong desire to stop smoking also plays a role in maintaining this addictive behavior. All of these aspects contribute to the difficulty of breaking away from smoking (Gita Kanya Paramitha & Stephani Raihana Hamdan, 2022).

When someone first tries smoking, some of the initial symptoms that may be experienced include coughing, tartness on the tongue, and nausea in the stomach. Despite this, many beginners tend to ignore these uncomfortable sensations. They often continue smoking, which then develops into a regular habit

and eventually evolves into a form of dependence that is difficult to break. This dependence is psychologically perceived as a source of pleasure that gives smokers deep satisfaction. This phenomenon can be explained through the concept of "tobacco dependency." This concept asserts that smoking behavior is initially perceived as a pleasurable activity, but over time, it can shift into an activity that is obsessive and difficult to stop (Susilaningsih, 2022).

The majority of cigarette consumers in Indonesia started their smoking habit in the age range of 15-19 years. Based on data obtained from Basic Health Research, it was recorded that 52.1% of total smokers admitted to first trying cigarettes when they were between 15 and 19 years old. Furthermore, the percentage of smokers in the 15-19 age group has increased in 2020. Data shows that 10.61% of the population aged 15-19 years were recorded as active smokers in 2020. This figure shows an increase from 10.54% recorded in the previous year, 2019. This phenomenon illustrates an increasing trend in the number of young smokers in Indonesia, which certainly requires serious attention from various related parties (Susilaningsih, 2022).

2.2 Data Mining

Data mining is the process of in-depth analysis of data sets with the aim of discovering previously unforeseen relationships. The process involves summarizing data with innovative methods that differ from conventional approaches, resulting in new insights that are useful to the data owner. As an interdisciplinary field, data mining integrates techniques from several disciplines, including machine learning, pattern recognition, statistics, database management, and data visualization. This integration is aimed at overcoming the challenges of extracting valuable information from large databases. Furthermore, data mining utilizes various advanced techniques such as statistics, mathematics, artificial intelligence, and machine learning. The goal is to extract and identify valuable information and hidden knowledge from various large-scale databases. In other words, data mining serves as a tool to uncover hidden knowledge in databases, using artificial intelligence, machine learning, statistics, and mathematics to extract and identify important information and knowledge from complex and large data sets (Utomo & Mesran, 2020).

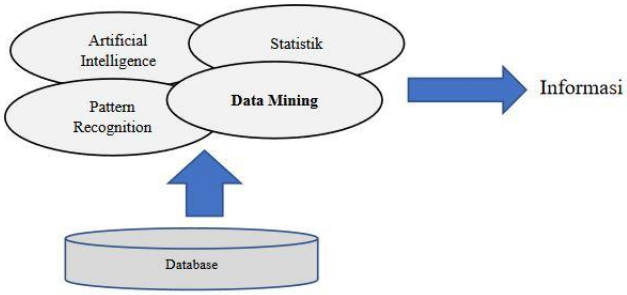

Figure 1. The Roots of Data Mining Source

Several terms have similar or closely related meanings in the context of data mining. One term often used interchangeably with data mining is Knowledge Discovery in Database (KDD). Both data mining and KDD have an identical goal: to utilize the data available in the database to be processed to produce new, useful, and meaningful information.

Besides KDD, several other terms are often used to describe processes similar to data mining. These terms include knowledge extraction, pattern or data analysis, data archaeology, and data mining. Each of these terms emphasizes a particular aspect of the process of processing and analyzing data to generate new insights.

Interestingly, people vary in how they perceive the relationship between data mining and related terms. Most understand data mining as a synonym for Knowledge Discovery from Data (KDD), while others view it as just one important stage in a broader knowledge discovery process. This difference in perspective reflects the complexity and breadth of the field of modern data analysis (Utomo & Mesran, 2020).

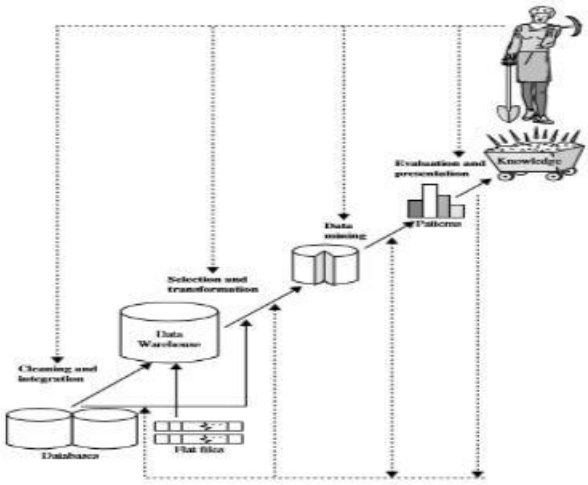

Figure 2. Stages of the Knowledge Discovery in Database (KDD) Process

2.3 Forecasting

Forecasting, also known as forecasting, is a method or technique used to estimate or predict future events or conditions. This process is done in a systematic and pragmatic way, meaning that it uses a structured and result-oriented approach that is practical and applicable (Muhartini et al., 2021).

Success in prediction or forecasting depends on several key factors. First, an in-depth technical knowledge of the methods of collecting data or information from the past is essential. This includes understanding the right data sources, accurate sampling techniques, and efficient collection methods (Nurmanto & Islami, 2024).

The types in the quantitative method forecast are grouped into two, namely :

1. Time Series Model

A Time Series Model is an approach or methodology in quantitative forecasting. This model is specifically designed to analyze and project data arranged sequentially based on time. In other words, the Time Series Model utilizes a series of chronologically arranged historical data to identify temporal patterns and trends, which are then used as a basis for making predictions or estimates about future values with a greater degree of accuracy.

2. Causal Model

A method that carefully and systematically considers various variables or crucial factors that have the potential to influence or significantly impact the amount or value being forecasted. This approach relies on the principle of causality as its basic assumption. By analyzing the complex interactions between various independent and dependent variables, this method seeks to produce more accurate and comprehensive predictions, allowing for a deeper understanding of the underlying dynamics of the phenomenon being studied or forecasted.

2.4 Time Series Data

Time series data is widespread and can be found in many aspects of everyday life. Its presence is very common and can be observed in a variety of contexts, such as in a stock price movement chart that shows fluctuations in the value of an investment over time, a weather forecast that predicts future atmospheric conditions based on historical patterns, or an analysis of population growth that reflects the demographic changes of a region over a certain period (Anis et al., 2022).

2.5 Linear Regression

Linear regression is a form of regression analysis that shows a linear relationship, aiming to find the correlation between the target or dependent variable and the predictor or independent variables. This method works most effectively when the data set used has linear characteristics, forming a straight-line relationship on the plot. However, in the practical application of linear regression, the process of collecting and selecting data samples is a very critical stage and determines the success of the analysis.

The quality and accuracy of the data sample used have a significant influence on the accuracy and reliability of the linear regression results produced (Maulana et al., 2024).

Linear regression algorithms have the main function of measuring and evaluating the strength of the relationship between the two variables being analyzed. In addition, this algorithm also plays an important role in describing and explaining the flow or pattern of the relationship formed between the dependent variable, which is the variable being affected or predicted, and the independent variable, which is the variable that affects or is used as a predictor (Aqsho Ramadhan et. Al. 2023).

Linear regression, in general, does have two main types that are often used in data analysis :

1. Simple Linear Regression

Simple linear regression is used to measure the strength of the causal relationship between the causal factor variable (X) and the effect variable. The causal factor is usually represented by x, which is also referred to as the predictor, while the effect variable is represented by y which is also referred to as the response (Trianggana, 2020).

2. Multiple Linear Regression

We make predictions with multiple linear regression involving two or more variables: the influencing variable and the influenced variable. These variables have a causal or interrelated relationship (Maharadja et al., 2021).

3. Methods

3.1 Research Process and Steps

1. Research Framework

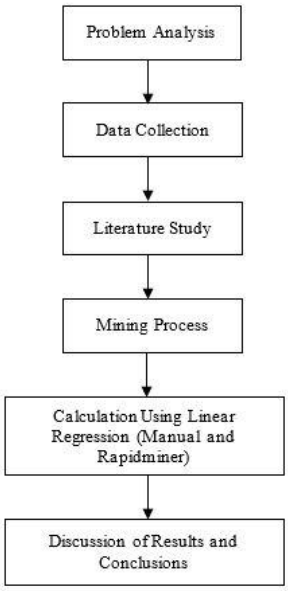

Figure 3. Research Methods

Figure 3 shows the steps of the research method, starting with problem analysis to identify the main issues that need to be resolved. After that, data collection relevant to the research topic is carried out. The next step is a literature study to understand the context of the research and find appropriate theories and methods. Next, the mining process is carried out to extract important information from the data that has been collected. This process is followed by calculations using linear regression with manual calculations and RapidMiner software. The results of this analysis are then discussed in the discussion of the results section, ending with a conclusion that summarizes the main findings of the research.

2. Data Source

Source data regarding the percentage of smoking in the population aged 15 years and over from 2019 to 2023 comes from the official website Badan Pusat Statistik (BPS).

Table 1. Percentage of Smoking in the Population Age ≥15 Years 2019-2023

Age Criteria	2019	2020	2021	2022	2023
15-19	10,54	10,61	9,98	9,36	9,62
20-24	28,77	28,65	26,97	25,99	26,95
25-29	32,79	31,81	32,32	31,55	32,12
30-34	34,71	34,2	34,66	33,83	33,65
35-39	35,28	35	35,55	34,81	35,21
40-44	34,36	34,23	35,13	34,57	35,1
45-49	32,58	32,45	33,7	33,03	33,83
50-54	31,32	31,41	31,97	30,85	31,87
55-59	30,03	29,69	30,04	29,31	29,35
60-64	28,76	27,31	28,05	26,92	27,46
65+	21,61	21,18	21,9	21,29	21,86

Table 1 shows 11 age groups and smoking percentages from 2019-2023 that the research will use to predict the percentage of smoking in 2024.

3. Type of Data

The type of data collected is Quantitative Data, which is numerical information that includes the percentage of smoking by age group.

4. Data Collection Technique

a Data Source Identification (observation)

The main source of data to be used is BPS, which provides comprehensive data on various demographic and socio-economic aspects of Indonesia.

b Literature Study

This research draws heavily on various international journals, national journals, local journals, and books as reference sources. This literature will help in solving problems related to the research topic.

5. Data Collection Procedure

a Access BPS Data: Visit the official BPS website at https://www.bps.go.id/id.

b Search for relevant datasets such as Demographic and Social Statistics, Health.

c Download Dataset: Download the relevant dataset from the BPS website. Document the data source and the year the data was collected.

3.2 Data Processing and Data Analysis Methods

1. Data Preprocessing

The first stage in data processing is data preprocessing, which prepares the raw data for analysis. Data preprocessing includes cleaning, removing duplicates, and handling missing values.

2. Data Grouping

Once the data has been preprocessed, the next step is to group the data by age. This grouping aims to facilitate the analysis of smoking trends in different age groups.

3. Linear Regression

The next step was linear regression analysis using RapidMiner. Linear regression was used to model the relationship between the independent variable (age) and the dependent variable (smoking percentage). This process involved several steps :

a Using the linear regression module in RapidMiner to create a predictive model that fits the data. RapidMiner automatically determines the model parameters based on the available data.

b RapidMiner allows users to perform model testing using test data.

c Evaluation in RapidMiner to assess model performance. Evaluation metrics such as Root Mean Squared Error (RMSE) were used to assess how well the model could predict the percentage of smoking by age group.

4. Analysis Tool

This research utilizes RapidMiner software for data processing and analysis. RapidMiner provides the various tools and techniques needed to perform all the stages mentioned, from data preprocessing to

model evaluation. The use of RapidMiner simplifies the analysis process and allows researchers to focus on interpreting the results.

4. Results

4.1 Manual Calculation

This study applies a multiple linear regression algorithm to predict the percentage of smoking in the population aged 15 years and above by age group in 2019-2023. Analysis model :

a Smoking Percentage in 2019 (X1)

b Smoking Percentage in 2020 (X2)

c Smoking Percentage in 2021 (X3)

d Smoking Percentage in 2022 (X4)

e Smoking Percentage in 2023 (Y)

With the following linear regression equation : $Y = a + b_1X_1 + b_2X_2 + ... + b_nX_n$

Furthermore, a summary analysis was carried out based on the values of X1, X2, X3, X4 and Y which have been presented in Table 1. The results of these calculations can be summarized as follows:

Kelompok Umur	Y	X1	X2	X3	X4	X1Y	X2Y	X3Y	X4Y	X1X2	X1X3	X1X4	X2X3	X2X4	X3X4	X1^2	X2^2	X3^2	X4^2
15-19	9,62	10,54	10,61	9,98	9,36	101,4	102,1	96,01	90,04	111,8	105,2	98,65	105,9	99,31	93,41	111,1	112,6	99,6	87,61
20-24	26,95	28,77	28,65	26,97	25,99	775,4	772,1	726,8	700,4	824,3	775,9	747,7	772,7	744,6	701	827,7	820,8	727,4	675,5
25-29	32,12	32,79	31,81	32,32	31,55	1053	1022	1038	1013	1043	1060	1035	1028	1004	1020	1075	1012	1045	995,4
30-34	33,65	34,71	34,2	34,66	33,83	1168	1151	1166	1138	1187	1203	1174	1185	1157	1173	1205	1170	1201	1144
35-39	35,21	35,28	35	35,55	34,81	1242	1232	1252	1226	1235	1254	1228	1244	1218	1237	1245	1225	1264	1212
40-44	35,1	34,36	34,23	35,13	34,57	1206	1201	1233	1213	1176	1207	1188	1202	1183	1214	1181	1172	1234	1195
45-49	33,83	32,58	32,45	33,7	33,03	1102	1098	1140	1117	1057	1098	1076	1094	1072	1113	1061	1053	1136	1091
50-54	31,87	31,32	31,41	31,97	30,85	998,2	1001	1019	983,2	983,8	1001	966,2	1004	969	986,3	980,9	986,6	1022	951,7
55-59	29,35	30,03	29,69	30,04	29,31	881,4	871,4	881,7	860,2	891,6	902,1	880,2	891,9	870,2	880,5	901,8	881,5	902,4	859,1
60-64	27,46	28,76	27,31	28,05	26,92	789,7	749,9	770,3	739,2	785,4	806,7	774,2	766	735,2	755,1	827,1	745,8	786,8	724,7
65+	21,86	21,61	21,18	21,9	21,29	472,4	463	478,7	465,4	457,7	473,3	460,1	463,8	450,9	466,3	467	448,6	479,6	453,3
Total	317	320,8	316,5	320,3	311,5	9790	9664	9802	9547	9753	9887	9628	9758	9503	9640	9882	9627	9897	9390

Figure 4. Manual Calculation

Figure 4 shows the result of the manual calculation of X1, X2, X3, X4, and Y. Furthermore, to predict the percentage of smoking in the population aged 15 years and over by age group in 2024, the manual calculation is carried out using the matrix method. This process aims to determine the values of b0, b1, b2, b3 and b4 using a system of four equations, which can be described as follows :

Matriks A

$$\begin{pmatrix} 11 & 320,75 & 316,54 & 320,27 & 311,51 \\ 320,75 & 9882,4 & 9752,9 & 9886,5 & 9627,9 \\ 316,54 & 9752,9 & 9627,1 & 9758,3 & 9503,3 \\ 320,27 & 9886,5 & 9758,3 & 9897,4 & 9639,8 \\ 311,51 & 9627,9 & 9503,3 & 9639,8 & 9389,5 \end{pmatrix}$$

DetA = 17339,88

Matriks H

$$\begin{pmatrix} 317,02 \\ 9790,1 \\ 9663,7 \\ 9801,8 \\ 9546,8 \end{pmatrix}$$

Matriks A1

$$\begin{pmatrix} 317,02 & 320,75 & 316,54 & 320,27 & 311,51 \\ 9790,1 & 9882,4 & 9752,9 & 9886,5 & 9627,9 \\ 9663,7 & 9752,9 & 9627,1 & 9758,3 & 9503,3 \\ 9801,7 & 9886,5 & 9758,3 & 9897,4 & 9639,8 \\ 9546,8 & 9627,9 & 9503,3 & 9639,8 & 9389,5 \end{pmatrix}$$

DetA1 = -2045,66

Matriks A2

$$\begin{pmatrix} 11 & 317,02 & 316,54 & 320,27 & 311,51 \\ 320,75 & 9790,1 & 9752,9 & 9886,5 & 9627,9 \\ 316,54 & 9663,7 & 9627,1 & 9758,3 & 9503,3 \\ 320,27 & 9801,7 & 9758,3 & 9897,4 & 9639,8 \\ 311,51 & 9546,8 & 9503,3 & 9639,8 & 9389,5 \end{pmatrix}$$

DetA2 = -5239,69

Matriks A3

$$\begin{pmatrix} 11 & 320,75 & 317,02 & 320,27 & 311,51 \\ 320,75 & 9882,4 & 9790,1 & 9886,5 & 9627,9 \\ 316,54 & 9752,9 & 9663,7 & 9758,3 & 9503,3 \\ 320,27 & 9886,5 & 9801,7 & 9897,4 & 9639,8 \\ 311,51 & 9627,9 & 9546,8 & 9639,8 & 9389,5 \end{pmatrix}$$

DetA3 = 4417,736

Matriks A4

$$\begin{pmatrix} 11 & 320,75 & 316,54 & 317,02 & 311,51 \\ 320,75 & 9882,4 & 9752,9 & 3790,1 & 9627,9 \\ 316,54 & 9752,9 & 9627,1 & 9663,7 & 9503,3 \\ 320,27 & 9886,5 & 9758,3 & 9801,7 & 9639,8 \\ 311,51 & 9627,9 & 9503,3 & 9546,8 & 9389,5 \end{pmatrix}$$

DetA4 = 17207,82

Matriks A5

$$\begin{pmatrix} 11 & 320,75 & 316,54 & 320,27 & 317,02 \\ 320,75 & 9882,4 & 9752,9 & 9886,5 & 9790,1 \\ 316,54 & 9752,9 & 9627,1 & 9758,3 & 9663,7 \\ 320,27 & 9886,5 & 9758,3 & 9897,4 & 9801,7 \\ 311,51 & 9627,9 & 9503,3 & 9639,8 & 9546,8 \end{pmatrix}$$

DetA5= 933,1502

$b0 = \frac{detA1}{detA}$ $b1 = \frac{detA2}{detA}$ $b2 = \frac{detA3}{detA}$ $b3 = \frac{detA4}{detA}$ $b4 = \frac{detA5}{detA}$

b0 = -0,11797 b1 = -0,30218 b2 = 0,254773 b3 = 0,992384 b4 = 0,053815

Based on the analysis that has been carried out, a multiple linear regression model is obtained as follows :
Y = -0.11797 + (-0.30218X1) + (0.254773X2) + (0.992384X3) + (0.053815X4)
After obtaining the multiple linear regression equation through manual calculation, the next step is to make predictions based on the equation that has been formulated, as shown in table 2.

Table 2. Manual Calculation Prediction Result

Age Criteria	2019	2020	2021	2022	2023	Prediction(2024)
15-19	10,54	10,61	9,98	9,36	9,62	9,623
20-24	28,77	28,65	26,97	25,99	26,95	28,419
25-29	32,79	31,81	32,32	31,55	32,12	32,564
30-34	34,71	34,2	34,66	33,83	33,65	34,543
35-39	35,28	35	35,55	34,81	35,21	35,131
40-44	34,36	34,23	35,13	34,57	35,1	34,182
45-49	32,58	32,45	33,7	33,03	33,83	32,347
50-54	31,32	31,41	31,97	30,85	31,87	31,048
55-59	30,03	29,69	30,04	29,31	29,35	29,718
60-64	28,76	27,31	28,05	26,92	27,46	28,409
65+	21,61	21,18	21,9	21,29	21,86	21,036

Table 2 shows the data from the prediction of the percentage of smoking in 2024 for each age group.

4.2 Rapidminer Calculation

This step starts in the operator view section by typing Read Excel, as the training data used is a file.

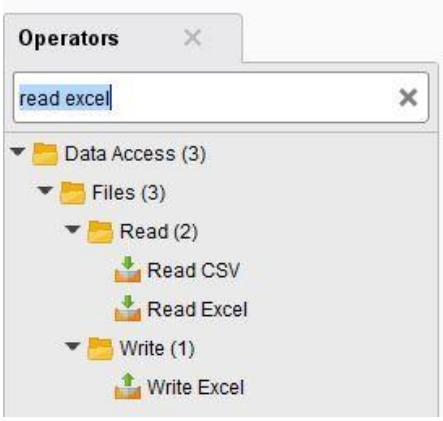

Figure 5. Add Read Operator in Excel

Figure 5 shows the first step in the calculation process using RapidMiner. It starts with typing "Read Excel" in the operator display section because the training data used is in the form of Exel, and then selects the Read Excel operator.

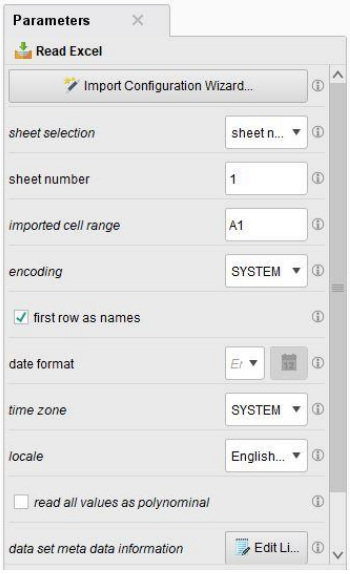

Figure 6. Training File Data Selection

Figure 6 shows the display of the Read Excel operator added. The next step is to select the Excel file of training data to be used. Click the Import Configuration Wizard button, then select the desired Excel file.

Figure 7 is a view of the selected training data file. Next, assign the role of id and label attributes. By properly assigning the role of id and label attributes, RapidMiner can correctly use the required information in every stage of data analysis, from processing to model building and evaluation.

Figure 8 shows the formation of a linear regression model by adding the linear regression operator. After adding the linear regression operator, the next step is to connect it to the "tra" socket of the linear regression operator with the "out" socket of the Read Excel operator. Next, connect the "mod" socket of the linear regression operator with the "res" socket to display the resulting model.

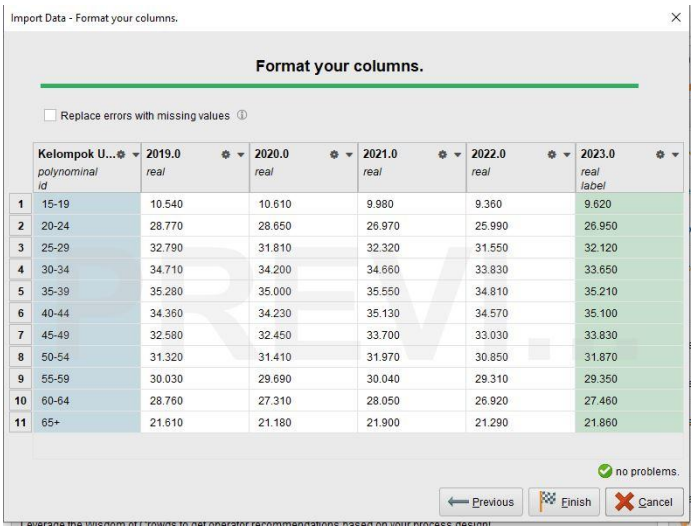

Figure 7. Select the ID and Label attributes

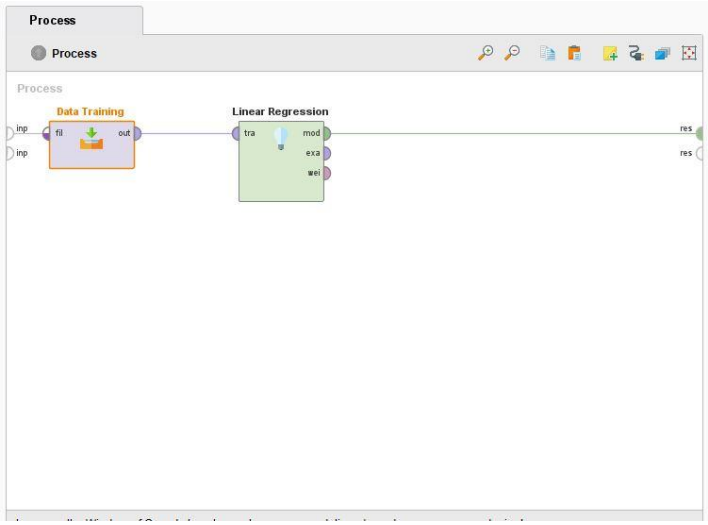

Figure 8. Linear Regression Model Formation

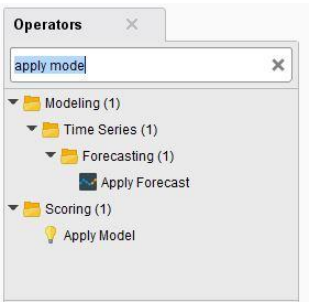

Figure 9. Operator Location Apply Model

Figure 9 shows how to find the location of the apply model operator by typing the apply model in the operator view section. The apply model function evaluates the model's performance by looking at the prediction results and comparing them with the actual values in the test data.

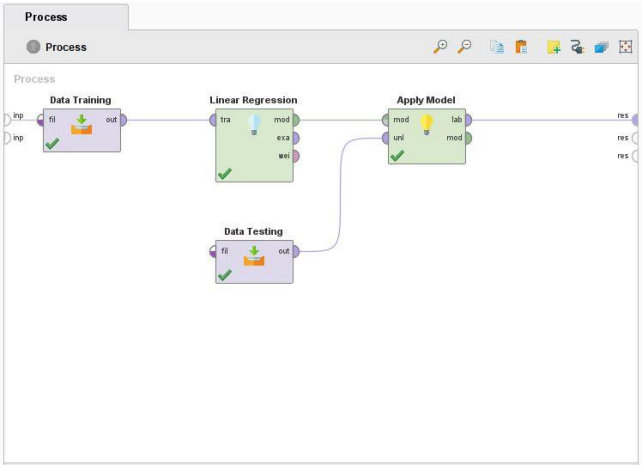

Figure 10. Testing on the Regression Model

Figure 10 shows the testing stage, starting with entering the testing data using a procedure identical to the training data entry process shown in Figures 6 and 7. The last step is connecting the available connectors.

Row No.	Kelompok U...	2023.0	predictio...	2019.0	2020.0	2021.0	2022.0
1	15-19	9.620	9.623	10.540	10.610	9.980	9.360
2	20-24	26.950	28.419	28.770	28.650	26.970	25.990
3	25-29	32.120	32.564	32.790	31.810	32.320	31.550
4	30-34	33.650	34.543	34.710	34.200	34.660	33.830
5	35-39	35.210	35.131	35.280	35	35.550	34.810
6	40-44	35.100	34.182	34.360	34.230	35.130	34.570
7	45-49	33.830	32.347	32.580	32.450	33.700	33.030
8	50-54	31.870	31.048	31.320	31.410	31.970	30.850
9	55-59	29.350	29.718	30.030	29.690	30.040	29.310
10	60-64	27.460	28.409	28.760	27.310	28.050	26.920
11	65+	21.860	21.036	21.610	21.180	21.900	21.290

Figure 11. Prediction Results Against Testing Data

Figure 11 shows the prediction results made by the RapidMiner application against the testing data with the existing linear regression model. The prediction results are in the "prediction" section.

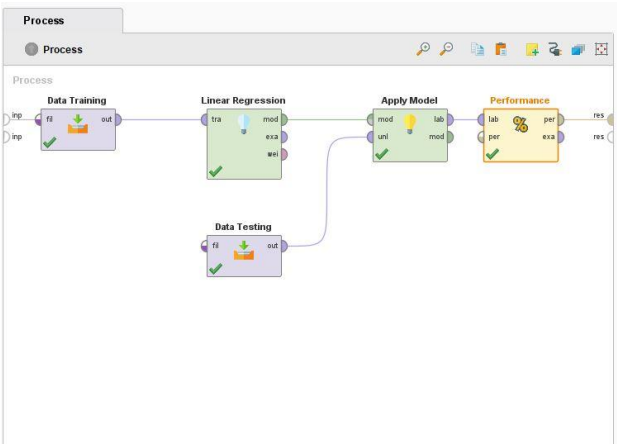

Figure 12. Linear Regression Model Performance Test Process

Figure 12 is an evaluation of the performance of the linear regression model that has been obtained. This

evaluation uses the Regression Performance operator, which is linked to the Apply Model operator. The Performance operator is commonly used for regression tasks. It automatically determines the type of task or method used and measures the general criteria of the task type.

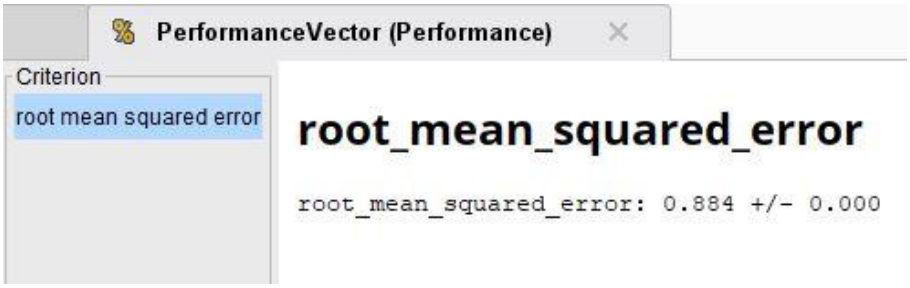

Figure 13. Performance Test Results

Figure 13 is a display of the results of the evaluation of the performance of the linear regression model. The results that have been obtained from the calculation of the RMSE obtained are classified as accurate. It is said to be accurate because there is < 1 prediction error between training data and testing data. RMSE is divided by the average value of the observed value. SI= (RMSE/average observed value)*100%. It will be easier to know whether the model is accurate or not. If SI < 10% is an accurate model, then SI < 5% is a very accurate model.

4.3 Testing Results

After manual calculation and RapidMiner, the same prediction results were obtained for the percentage of smoking in the 15-19 age group of 9.623, 20-24 age group of 28.419, 25-29 age group of 32.564, 30-34 age group of 34.543, 35-39 age group of 35.131, 40-44 age group of 34.184, 45-49 age group of 32.347; age group 50-54 by 31.048; age group 55-59 by 29.718, age group 60-64 by 28.409, and age group 65+ by 21.036. The evaluation results using RMSE with a value of 0.884 +/- 0.000 shows that the linear regression model built using RapidMiner has a low error rate. This indicates that the model can accurately predict the percentage of cigarette consumption according to the observed data from 2019 to 2023. These results provide an overview of the predicted percentage of smoking for each age group based on the data analyzed.

5. Discussion

The finding that the percentage of smoking in the 15-19, 20-24, 25-29, 30-34, 55-59, and 60-64 age groups increases compared to a decrease in other age groups needs to be further analyzed by comparing it with previous relevant studies. The discussion process should not only focus on explaining the results but also on linking the findings with existing theories and research so as to produce representative conclusions in accordance with the research objectives.

For example, the increase in smoking among younger age groups can be attributed to studies showing the increased accessibility of cigarettes and the influence of the social environment in Indonesia. Meanwhile, the decline in older age groups may be due to increased health awareness or age-related health factors. This discussion aims to explore the significance of the research results in a broader context and identify their implications for public health policy without simply repeating previously mentioned results.

6. Conclusion

Based on the research findings, there is a significant variation in smoking consumption trends across age groups in Indonesia from 2019 to 2023, with some groups experiencing an increase and others a decrease. The linear regression model built using RapidMiner demonstrates high accuracy, with an RMSE value of 0.884, confirming its reliability in predicting smoking consumption percentages. These results indicate that using RapidMiner is effective in supporting the analysis of smoking consumption trends by age group, making the resulting model a valuable reference for understanding smoking

patterns in Indonesia.

Recommendation

As for some suggestions that the authors convey for further research :

1. Conduct research and analysis with other factors that may affect cigarette consumption, such as government policies, economic conditions, and local culture.
2. Conduct research using other algorithms and compare them with linear regression algorithms.
3. Perform predictions with other data analysis tools, then compare them with RapidMiner.

Limitations and avenue for future research

1. This study only uses data regarding cigarette consumption by age from 2019 to 2023. Other factors that may affect cigarette consumption, such as government policies, economic conditions, and local culture, will not be analyzed.
2. This research does not include variations or comparisons with other statistical or machine learning methods other than linear regression algorithms for cigarette consumption prediction analysis.
3. Analysis was performed using RapidMiner, and did not compare with other data analysis tools.

References

Anis, P., Sekar, D., & Kharisudin, I. (2022). *Analisis Peramalan dengan Long Short Term Memory pada Data Kasus Covid-19 di Provinsi Jawa Tengah*. *5*, 752–758.

Aqsho Ramadhan, Y., Faqih, A. ., & Dwilestari, G. . (2023). Prediksi Penjualan Handphone di Toko X menggunakan Algoritma Regresi Linear. *Jurnal Informatika Terpadu*, *9*(1), 40–44. https://doi.org/10.54914/jit.v9i1.692

Dahlia, & Andri. (2020). Implementasi Data Mining untuk Prediksi Persediaan Obat pada Puskesmas Kertapati menggunakan Regresi Linier Berganda. *Jurnal Sistem Dan Informatika*, 95–103. https://doi.org/10.30864/jsi.v15i2.331

Gita Kanya Paramitha, & Stephani Raihana Hamdan. (2022). Pengaruh Self-Control terhadap Perilaku Merokok Mahasiswa selama Pandemi COVID-19. *Jurnal Riset Psikologi*, *1*(2), 132–139. https://doi.org/10.29313/jrp.v1i2.559

Maharadja, A. N., Maulana, I., & Dermawan, B. A. (2021). Penerapan Metode Regresi Linear Berganda untuk Prediksi Kerugian Negara Berdasarkan Kasus Tindak Pidana Korupsi. *Journal of Applied Informatics and Computing*, *5*(1), 95–102. https://doi.org/10.30871/jaic.v5i1.3184

Marianti, A., & Prayitno, B. (2020). Analisis Pengaruh Faktor Sosial Ekonomi, Pendapatan dan Harga Rokok Terhadap Konsumsi Rokok di Indonesia. *Economie: Jurnal Ilmu Ekonomi*, *1*(2), 93. https://doi.org/10.30742/economie.v1i2.1126

Maulana, A., Martanto, M., & Ali, I. (2024). Prediksi Hasil Produksi Panen Bawang Merah Menggunakan Metode Regresi Linier Sederhana. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(4), 2884–2888. https://doi.org/10.36040/jati.v7i4.7281

Muhartini, A. A., Sahroni, O., Rahmawati, S. D., Febrianti, T., & Mahuda, I. (2021). Analisis Peramalan Jumlah Penerimaan Mahasiswa Baru Dengan Menggunakan Metode Regresi Linear Sederhana. *Jurnal Bayesian : Jurnal Ilmiah Statistika Dan Ekonometrika*, *1*(1), 17–23. https://doi.org/10.46306/bay.v1i1.2

Novianty, D., Palasara, N. D., & Qomaruddin, M. (2021). Algoritma Regresi Linear pada Prediksi Permohonan Paten yang Terdaftar di Indonesia. *Jurnal Sistem Dan Teknologi Informasi (Justin)*, *9*(2), 81. https://doi.org/10.26418/justin.v9i2.43664

Nurmanto, R., & Islami, H. Al. (2024). *Pemberangkatan Unit Luar Pulau Untuk Memprediksi Jangka Waktu Penerimaan Unit Ac*. *3*(4), 850–864.

Susilaningsih. (2022). Faktor-faktor Penyebab Perilaku Merokok Pada Remaja di Tegalrejo. *Jurnal Keperawatan*, *8*, 46–56.

Trianggana, D. A. (2020). a Peramalan Jumlah Siswa-Siswi Melalui Pendekatan Metode Regresi Linear. *Jurnal Media Infotama*, *16*(2), 115–120. https://doi.org/10.37676/jmi.v16i2.1149

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *Jurnal Media Informatika Budidarma*, *4*(2), 437. https://doi.org/10.30865/mib.v4i2.2080