ISSN 2087-3336 (Print) | 2721-4729 (Online)

TEKNOSAINS: Jurnal Sains, Teknologi dan Informatika
Vol 11, No. 2, July 2024, page. 245-256
http://jurnal.sttmcileungsi.ac.id/index.php/tekno
DOI: 10.37373

Prediction of the level of crime cases using multiple linear regression in the city of Pontianak

Fadillah Bergas, Sucipto*, Asrul Abdullah

*Universitas Muhammadiyah Pontianak, Indonesia, 78123, Jl. Jenderal Ahmad Yani No.111, Bangka Belitung Laut, Kec. Pontianak Tenggara, Kota Pontianak, Kalimantan Barat

*✉ sucipto@unmuhpnk.ac.id

Submitted: 14/01/2024 Revised: 04/03/2024 Accepted: 11/03/2024

ABSTRACT

Using secondary data from the POLRESTA Pontianak Criminal Investigation Unit, this study intends to create a multiple linear regression predictive model for the crime rate in the Police Resort Area of Kota (POLRESTA). By employing descriptive statistical methods and data visualization, pertinent aspects that enhance the information in the model can be found. The evaluation's findings show that Kota Pontianak's crime rates can be accurately predicted by this model as well as modeled. The model exhibits its ability to predict known data even in the presence of differences in error rates between training and testing data. The testing findings also show that there are differences in the testing dataset between the Mean Absolute Percentage Error (MAPE) values for each crime category: MAPE for "heavy" rises to 12.91%, MAPE for "Medium" to 30.11%, and MAPE for "light" to 26.59%. As a result, this study finds that the Multiple Linear Regression method has the potential to be a useful tool for making decisions and developing plans to stop criminal activity in Kota Pontianak.

Keywords: Criminality; data mining; multiple linear regression; MAPE

1. INTRODUCTION

Low-income countries are those that are going through the development process [1], restricted facilities for infrastructure, and have a poorer human development index in comparison to other nations. This stage of development presents several of obstacles for countries, and improving them will need a significant investment of time and energy. The possibility of crimes or criminal activities occurring is one of these issues. It is important to keep in mind that criminal behavior is not natural or inherited; rather, it can be perpetrated by men and women of all ages, including young adults, adults, and the elderly.

The word "crimen," which denotes an act of crime, is where the word "criminal" originates. Crime encompasses a range of conduct and acts that have the potential to inflict financial and psychological damage, in addition to breaching relevant Indonesian laws and social and religious standards [2] [3].

In Indonesia, crime is on the rise right now, particularly in the major cities. According to information provided by the Indonesian National Police (Polri), there were fewer crimes or criminal acts committed in Indonesia between 2018 and 2020. There were 294,281 documented crime occurrences in 2018. After that, the number dropped to 269,324 incidents in 2019 and 247,218 incidents in 2020 [4].

Serious crimes of all kinds, including drug use, theft, murder, and corruption, are committed in this nation. Legally speaking, every criminal act has a varied degree of gravity, which is established by the trial process and the provisions found in the Criminal Procedure Code (KUHAP) [5].

There are various categories of crimes in the context of crime, such as crimes against state property, international crimes, conventional crimes, and crimes with contingent consequences. In Indonesia, there were 165,918 criminal cases overall in 2018. Conventional crimes comprised 134,462 cases, or almost 81% of all crime cases, out of all these categories. With 19,380 cases, or almost 14% of all conventional

crime cases, aggravated theft was the most common type of offense in the conventional crime category [6].

5,275 conventional crime cases were reported in West Kalimantan, particularly in Pontianak City, between 2019 and 2023, according to data from the Pontianak police. However, the Pontianak Police stressed that all criminals will be held accountable for the articles that have been decided; this is meant to serve as a deterrent to those who commit crimes. Crime rates can occasionally change from year to year, going down or up.

Several scholars have undertaken prior studies to evaluate the effectiveness of multiple linear regression. Dahlan Abdullah of Malikussaleh University's informatics engineering study program's "prediction of drug user levels using web-based multiple linear regression methods" is one of them [7]. Aside from that, this technique has also been employed in research titled "Crime Level Prediction System Using the Triple Exponential Smoothing Method" by Arwin Datumaya Wahyudi Sumari from the Master of employed Electrical Engineering Study Program at Malang State Polytechnic [8]. The study "Prediction of Crime Rates using the single moving average method" by Mustopa Husein Lubis from the Putra Indonesia University YPTK Padang's information systems and technology study program is another preliminary study that makes use of the multiple linear regression method [9].

As a result, gathering data regarding the number of criminal episodes in the Pontianak City Area is crucial to make it available to the general public or the community. Police personnel will find it simpler to make decisions and gauge the number of criminal incidences in Pontianak City as a result. Finding a multiple linear regression model to forecast the volume of criminal cases in the Pontianak City resort police area (POLRESTA) is the primary goal of this study. When developing measures to lower monthly crime incidences in Pontianak City, police officers in the Pontianak City Resort Police Area (POLRESTA), West Kalimantan Province, will consider the research's findings.

A data analysis technique called linear regression is used to forecast the value of unknown data by comparing it to values of related and known data. From a mathematical perspective, this approach uses a linear equation to represent the relationship between unknown or dependent variables and known or independent variables [10].

The kind of dependent variable determines how logistic regression and linear regression differ from one another. When the dependent variable is numerical, linear regression is employed; when the dependent variable is categorical or dichotomous, logistic regression is used [11].

To determine the level of crime cases, the author is interested in developing a system that uses the Multiple Linear Regression method and references data from secondary data taken from Bareskrim, which is located at the Pontianak Police Station. This system would be used to model predictions of the level of crime cases.

2. METHOD

The research method employed in this study is depicted in the research flowchart in Figure 1.

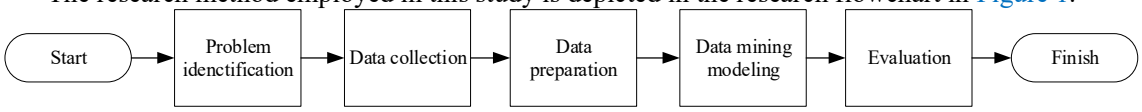

Figure 1. Research flow diagram.

Identification of the problem, this study attempts to tackle two major issues. First, let's look at how multiple linear regression can be used to forecast the volume of criminal cases in Pontianak. A model that connects independent variables to crime rates will be developed using information from the Pontianak Police Criminal Investigation Unit. In Pontianak, this model is frequently employed to forecast crime cases. Second, this study assesses how well the multiple linear regression approach predicts Pontianak crime cases [12]. This method's ability to forecast Pontianak crime rates will be assessed by contrasting model predictions with real data.

Gathering of data. Gathering information from the Pontianak Police Criminal Investigation Department is the first phase, particularly about conventional offenses from 2019 to 2023. To assure its dependability in further analysis, this data will thereafter be explained using typical analysis techniques including descriptive statistics and data quality assessment [13].

The Pontianak Police's criminal investigation section provided the majority of the data used in this study's data gathering, particularly for information on conventional crimes that happened between 2019 and 2023. This information is crucial for examining the features of crimes committed during that time.

Table 1. Dataset attributes and description.

No	Attribute	Information
1	Year	This feature gives details about the year the offense was committed.
2	Month	This characteristic gives details on the month the offense was committed.
3	Serious Crimes	This characteristic indicates a very serious crime. Murder, armed robbery, and other major crimes are examples of serious crimes.
4	Moderate Crime	This characteristic represents moderately serious offenses. Aggravated theft, extortion, and other offenses that fall short of major crimes are examples of moderate crimes.
5	Minor Crimes	This characteristic represents less serious offenses. Petty theft, fraud, and other offenses that fall short of being classified as serious or moderate crimes are examples of minor crimes.

The characteristics of the dataset obtained from the Pontianak Police's criminal investigation unit are described in Table 1. While the "Month" and "Year" properties provide information on the month of the occurrence and the year the crime happened, the "Serious Crimes," "Moderate Crimes," and "Light Crimes" attributes each indicate distinct crime types with varying degrees of severity. To guarantee these data's dependability for additional analysis, descriptive statistics and data quality assessment will be used in the analysis process.

Data Gathering. Several crucial actions will be made during the data preparation phase to create the final dataset that will be utilized in modeling [14]. Here's a more thorough breakdown of these procedures:

a. Data cleaning.
b. The data will be analyzed in this step to find and fix any missing, duplicate, or incorrect values. Before being used in modeling, this phase attempts to guarantee the data's integrity and cleanliness. We will fix or remove any inaccurate or lacking data to ensure that the analysis's findings are unaffected [15].
c. Data transformation. To satisfy certain modeling or analysis requirements, data transformation entails modifying the format or structure of data. This could involve addressing outliers, correcting the data's scale, or recoding categorical variables into numerical variables. Data transformation ensures that the data is prepared for use in the upcoming modeling [16].

The final dataset will be prepared for usage in the modeling phase by completing the aforementioned processes. To guarantee the caliber and precision of the outcomes of the analysis and modeling that is done, careful and comprehensive data preparation is essential [17]. models for data mining. A technique in data mining called multiple linear regression Modeling is used to create models or prediction correlations between several independent variables (features) and the dependent variable (target) [18]. Remember that in this context, "mining" refers to the process of trying to wrest useful information from vast quantities of raw data. As a result, data mining has a lengthy history in several scientific domains, including databases, machine learning, artificial intelligence, and statistics [19]. Knowledge discovery in databases, or KDD, is an iterative process that includes data mining as a crucial stage [20].

A statistical analysis technique called linear regression is used to model the relationship between multiple variables as an explicit linear equation. A linear equation with only one variable included is known as an explicit linear equation [21].

The link between one dependent variable, or response (Y), and two or more independent variables, or predictors (X1, X2,..., Xn), can be explained using an equation model called multiple linear regression. When the values of the independent or predictor variables (X1, X2,..., Xn) are known, the goal of a multiple linear regression test is to predict the value of the dependent or response variable (Y). In addition, the direction of the link between the dependent and independent variables can be ascertained using this test. Equation 1 contains the formula for multiple linear regression [22].

$$Y = a + b1 \times 1 + b2 \times 2 + b3 \times 3 \quad (1)$$

Where:

Y	= Variable dependent
x1, x2, … xn	= Variable independent
a	= Konstanta
b1, b2, … bn	= Koefisien regresi

Finding the linear equation that most accurately captures the relationship between the dependent and independent variables is the goal of multiple linear regression. Multiple linear regression modeling involves the following steps [23]:

a. Data preparation.
The data that will be used in the modeling process must first be prepared. This entails sanitizing the data, choosing pertinent characteristics, and dividing the data into independent and dependent variables.

b. Selection of independent variables.
Independent variables that might be related to or have an impact on the dependent variable are chosen at this point. Building an accurate model requires careful consideration of the independent variable selection.

c. Model adjustment.
Using the available data, a multiple linear regression equation is constructed in this stage. For every independent variable in multiple linear regression, a linear equation is defined with coefficients.

d. Coefficient estimation.
Methods like the least squares approach. Finding the coefficient values that result in the model that best fits the data is the aim.

We can forecast the value of the dependent variable based on the value of the independent variable and comprehend the link between the dependent and independent variables by applying multiple linear regression modeling. This facilitates enhanced decision-making and a more profound comprehension of an issue or circumstance using the carried-out data analysis [24].

Assessment. An assessment technique called MAPE (mean absolute percentage error) gauges how frequently a prediction is used [25]. We can determine the difference between the real and anticipated values using MAPE [26]. Equation 2's MAPE calculation formula is as follows.

$$MAPE = \frac{100}{n} \sum_{i}^{n} = 0 \left| \frac{\hat{y}i - \hat{y}}{\hat{y}} \right| \quad (2)$$

Where:

$\hat{y}i$	= Prediction results
yi	= Actual value.
n	= The amount of data tested.

As a percentage of the average absolute error rate value for the real data period, MAPE will calculate the average absolute error. The accuracy value increases with decreasing MAPE value, as explained by the MAPE value's criterion. **Error! Reference source not found.** displays the requirements for the MAPE value.

Table 2. MAPE criteria.

No	MAPE Value	Criteria	Information
1	< 10%	Very Good	With an error rate of less than 10%, projections and real data accord remarkably well.
2	10% - 20%	Good	The prediction accuracy is high, with an error rate of 10% to 20%.
3	20% - 50%	Enough	Predictions with an error rate of 20% to 50% can still be utilized despite errors.
4	> 50%	Bad	Forecasts are inaccurate for making decisions because of their extremely high error rate, which often exceeds

50%.

3. RESULTS AND DISCUSSION

The findings of this study's discussion align with the problem formulation and research methods outlined in the preceding section.

3.1 Data exploration results

To comprehend the dataset, this study used data visualization and descriptive statistics. While data visualization makes graphs to show data trends, descriptive statistics provide a summary of the data, including averages and variation. It is anticipated that this method will highlight significant data from the dataset and expand it.

The fundamental properties of a data collection can be summed up and described using descriptive statistics. The objective is to enhance comprehension of current data by elucidating, characterizing, and analyzing the data. A variety of data-centering measures, including mean, median, and mode, as well as data dispersion measures, including range, standard deviation, and quartiles (see Table 3 are used in descriptive statistics.

Table 3. Descriptive statistics data describe.

Atribut	Count	Mean	Std	Min	25%	50%	75%	max
Year	55	2020.82	1.35	2019	2020	2021	2022	2023
Weight	55	40.96	11.67	11	32	39	49	69
Medium	55	12.31	3.63	5	10	12	14	21
Mild	55	42.64	24.76	10	23	36	59	94
Total	55	95.91	33.99	48	68.5	83	126.5	165
Average	55	31.97	11.33	16	22.83	27.67	42.17	55

Between 2019 and 2023, a period of five years, our dataset comprises fifty-five entries/data. The average year that has been recorded is roughly 2020.82, while the standard deviation is roughly 1.35. This indicates that the bulk of the data is centered around 2020–2021. 2019 was the lowest recorded year in the dataset, and 2023 had the highest figure. Features of crime:

Table 4. Descriptive statistics data information.

#	Column	Non-Null Count	Dtype
0	Year	55 non-null	Int64
1	Month	55 non-null	Object
2	Weight	55 non-null	Int64
3	Medium	55 non-null	Int64
4	Mild	55 non-null	Int64
5	Total	55 non-null	Int64
6	Average	55 non-null	Float64
7	Crime Category	55 non-null	object

The data frame in Figure 3 contains eight columns with various sorts of data. Table 4 there are 55 entries/rows in this data frame, which includes information about several variables such as year, month, weight, intensity, total, average, and crime type. These columns could provide facts or statistical information about a certain occurrence or phenomenon.

A method of visually representing data to comprehend its distributions, correlations, and patterns is called data visualization. It is typically applied to statistical data, including anomalies, trends, and comparisons. Bar graphs, lines, circles, histograms, and scatter plots are a few examples of methods.

We may extract patterns and conclusions from complicated data by using data visualization. The distribution of the weight features is displayed in Figure 4.

The total number of claims broken down by month and category is displayed in Figure 2 bar graph. The categories are mild, medium, and heavy. For the heavy category, the biggest number of claims happened in March and September; for the moderate category, it happened in June and November. The

months with the most claims for the light category were January and August. Generally speaking, the months of January through March and October through December see the greatest amount of claims.

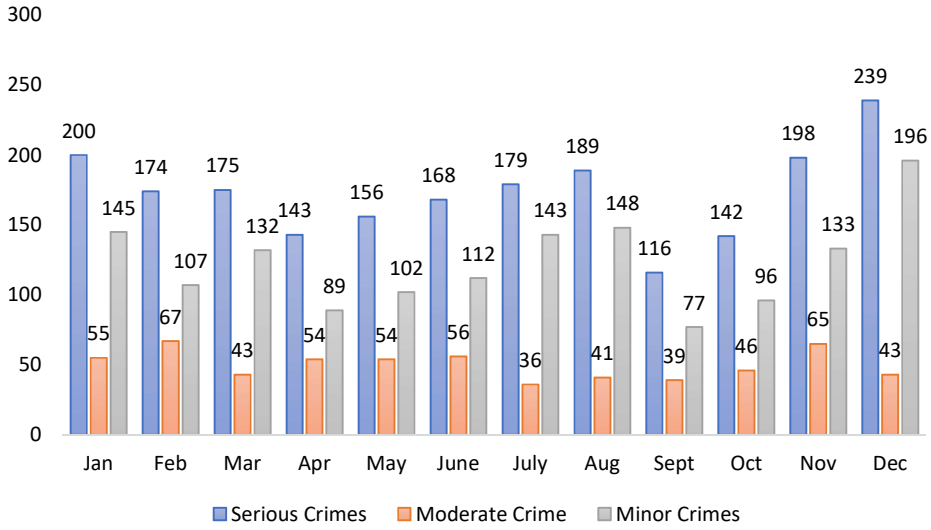

Figure 2. Graph of total crime months and categories.

The association between characteristics and crime categories is displayed in the heatmap findings. Correlation values vary from -1 to 1. Perfect positive correlation is represented by a value of 1, perfect negative correlation is represented by a value of -1, and no correlation is represented by a value of 0. "Years," "Severe," and "Total" features have very substantial positive connections with crime categories.

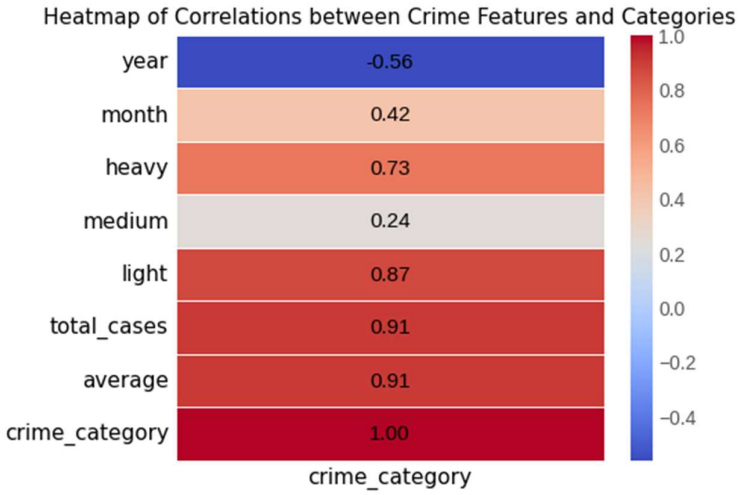

Figure 3. Correlation of features with crime categories.

This demonstrates that the likelihood of a crime occurring increases with the feature's value. "Average" is the trait that has the strongest inverse relationship with the crime category. This demonstrates that the likelihood of crime occurring decreases with increasing average value. There was a slight correlation between the crime categories and the "moderate" and "mild" traits. This implies that the likelihood of a crime occurring is not significantly impacted by this attribute. Organizations can utilize this information to better understand the causes of crime and create strategies for prevention.

3.2 Preprocessing

Preprocessing or pre-processing is a series of steps performed on data before the data can be used for further analysis or modeling.

a. Data cleaning results.

To guarantee the quality of the dataset, data cleansing is crucial. To prevent duplicates and missing values, precautions were taken. Consequently, there are neither duplicates nor missing values. This first stage is crucial to guaranteeing valid interpretations and findings in the context of the crime data being studied, as well as accurate and trustworthy analytical outcomes.

b. Feature engineering results (feature development).

c. A key strategy in this research is feature engineering, which is taking current data and using it to create new features. There have been three feature engineering phases completed:
   - Total Crimes: the number of serious, medium, and light crimes is combined in a new column called "total". This gives an accurate picture of how many offenses there were in each case.
   - Average Intensity: To determine the average intensity of heavy, medium, and light, an "average" column was added. This offers a closer look at the general level of crime.
   - Crime severity categorization: a new column called "crime category" was created, describing crime severity groups according to the total number of offenses. Three levels of data are distinguished: high, medium, and low. It gives details regarding the seriousness of the offense in every instance.

d. These phases add new dimensions to the dataset, enhancing its understanding of the quantity, severity, and intensity categories of crimes. This enhances the data that may be analyzed further, making it easier to spot patterns, trends, and other traits that might not be apparent in the raw data.

e. One-hot encoding results
   One-hot encoding is a method for transforming categorical variables into binary data so that machine learning or statistical models can analyze them more easily. To illustrate, the variable "month" is converted into two additional columns: one for category matches and zero otherwise. This method facilitates the recognition of correlations or patterns in data by algorithms. Models or analyses that use this data perform better as a result of One-Hot Encoding.

f. Scaler results (scaling)
   A data processing tool called a scaler is used to change numerical information on a specific scale so that machine learning algorithms may use or interpret them more easily. The min-max scaling approach is one type of scaler methodology used in this study. Each numerical feature value is converted by min-max scaling into a scale between two given values, typically 0 and 1.

g. Data separation results.
   For testing and training models, data must be divided into distinct subsets. The data is split into two primary sections for this process: the training portion (training data) and the testing portion (testing data). In this study, the data was split up into 80% for training and 20% for testing, based on the relatively small amount of data (55 in total).

3.3 Results of linear regression modeling.

We describe the linear regression modeling results in this section. There are two parts to this process: testing and model training. To maximize its capacity to identify patterns in data, the model is trained using training data during the training phase. Next, the model is evaluated using test data that has never been seen before. The mean absolute percentage error, or MAPE, approach was used to assess the performance of the model. Table 5 contains the MAPE and multiple linear regression calculations:

Table 5. Example data for multiple linear regression calculations.

Year	Month	Weight	Currently	Mild
2019	January	30	15	90
2019	February	60	19	80
2019	March	21	20	80
2019	April	12	16	73
2019	May	23	9	89
2019	June	40	25	50
2019	July	50	26	78

Step 1: The same notation as previously described is utilized. "Mild" will be the label (dependent variable) that we employ. Equation 1 expresses the multiple linear regression model for the given data.

Step 2: Estimate regression coefficients in step two. You need to use a statistical technique like the least squares method to estimate regression coefficients. But in this instance, I'll provide you with a useful example of the coefficient estimation findings. Assume the following outcomes of the coefficient estimation:

$$b0 = 10, b1 = 0.2, b2 = 0.5, b3 = 0.3$$

Step 3: Compute the prediction Based on the available data, we can determine the expected value of y (criminal case rate) using the coefficient estimation results above. Let's determine a prediction for every data entry:

1. January 2019 : y_pred = 10 + 0.2 × 30 + 0.5 × 15 + 0.3 × 90 = 48.5
2. February 2019 : y_pred = 10 + 0.2 × 60 + 0.5 × 19 + 0.3 × 80 = 53.9
3. March 2019 : y_pred = 10 + 0.2 × 21 + 0.5 × 20 + 0.3 × 80 = 43.5
4. April 2019 : y_pred = 10 + 0.2 × 12 + 0.5 × 16 + 0.3 × 73 = 46.5
5. May 2019 : y_pred = 10 + 0.2 × 23 + 0.5 × 9 + 0.3 × 89 = 47.3
6. Juny 2019 : y_pred = 10 + 0.2 × 40 + 0.5 × 25 + 0.3 × 50 = 43.5
7. July 2019 : y_pred = 10 + 0.2 × 50 + 0.5 × 26 + 0.3 × 78 = 48.6

Step 4: Determine the MAPE after that, we may compute the mean absolute percentage error, or MAPE, for every forecast we had previously made. Assuming real values that are unknown to us, for instance: We'll make use of the previously determined estimates:

$$y\_pred = [48.5, 53.9, 43.5, 46.5, 47.3, 43.5, 48.6]$$

We assume actual value:

$$y\_actual = [55, 60, 42, 45, 48, 44, 52]$$

a. Determine the difference in percentage and absolute terms. For every month, we will compute the absolute difference and absolute percentage.
b. Determine the MAPE, which is the mean of all data's absolute percentages. Regarding the MAPE formula using Equation 2.
    1) January 2019:
       AbsDiff = | 55 - 48.5 | = 6.5 AbsPercent = 6.5 / 55 × 100 = 11.82%
    2) February 2019:
       AbsDiff = | 60 - 53.9 | = 6.1 AbsPercent = 6.1 / 60 × 100 = 10.17%
    3) March 2019 :
       AbsDiff = | 42 - 43.5 | = 1.5 AbsPercent = 1.5 / 42 × 100 = 3.57%
    4) April 2019:
       AbsDiff = | 45 - 46.5 | = 1.5 AbsPercent = 1.5 / 45 × 100 = 3.33%
    5) May 2019:
       AbsDiff = | 48 - 47.3 | = 0.7 AbsPercent = 0.7 / 48 × 100 = 1.46%
    6) Juny 2019:
       AbsDiff = | 44 - 43.5 | = 0.5 AbsPercent = 0.5 / 44 × 100 = 1.14%
    7) July 2019:
       AbsDiff = | 52 - 48.6 | = 3.4 AbsPercent = 3.4 / 52 × 100 = 6.54%
c. Calculate average MAPE

$$MAPE = (11.82 + 10.17 + 3.57 + 3.33 + 1.46 + 1.14 + 6.54) / 7 \approx 5.88\ \%$$

Thus, for the data you just gave, the average MAPE of the multiple linear regression forecasts is roughly 5.88%. It shows the typical mistake rate between the model's predictions and the real data. The model's ability to predict outcomes is better when the MAPE value is lower.

3.4 Training

The results of the MAPE test show how much the trained multiple linear regression model's prediction error level has increased. The average percentage difference between the actual value from the test data and the value predicted by the model is measured using the evaluation method known as MAPE (mean absolute percentage error). As seen in Table 6, the model performs better when making predictions when the MAPE value is lower.

Table 6. Training accuracy.

MAPE	Akurasi
Weight	10.60%
Medium	14.32%
Mild	12.61%

In the context of the test results in Table 6, the MAPE results for each crime category are as follows:

a. MAPE Weight : 10.60%
b. MAPE Medium : 14.32%
c. MAPE Mild : 12.61%

About 10.60% of the heavy crime category, 14.32% of the medium crime category, and 12.61% of the light crime category are predicted incorrectly by the model. The model's accuracy in forecasting the quantity of crimes in each category is gauged by MAPE. The predicted and actual values are more closely aligned when the MAPE is smaller. These findings show how well the Multiple Linear Regression model estimates the number of crimes about other models. The low error rate suggests the model's strong ability to forecast based on training data, even while the MAPE values vary.

3.5 Testing

The multiple linear regression model's performance on test data is assessed using the MAPE test results that are displayed. The average percentage difference between the expected value from the model and the actual value in the test data is measured using the assessment technique known as MAPE (Mean Absolute Percentage Error). As seen in Table 7, the model's predictive power increases with a decreasing MAPE value.

Table 7. Testing accuracy.

MAPE	**Akurasi**
Weight	12.91%
Medium	30.11%
Mild	26.59%

In the test results in Table 7, the MAPE values for each crime category are as follows:

a. MAPE Weight : 12.91%
b. MAPE Medium : 30.11%
c. MAPE Mild : 26.59%

According to the test data, the model's prediction error rate for the serious crime category is approximately 12.91%, for the medium crime category it is 30.11%, and for the light crime category, it is 26.59%. The model's predictions are more accurate the lower the MAPE. The Multiple Linear Regression model's capacity to generalize data that has never been seen is demonstrated by the results of the MAPE test. The test data has a somewhat larger error rate, but the model performs well in estimating crimes based on the extracted features, as seen by the comparatively low MAPE values. The potential for enhancing model performance through additional changes is indicated by the difference in MAPE values between training and testing data.

3.6 Evaluation of the results of multiple linear regression modeling.

A comparison graph of the MAPE scores from the training and testing data, along with the evaluation method for multiple linear regression modeling in weight, medium, and mild variables.

```
# Fitting a multiple linear regression model for the weight variable
weight_model = Linear Regression()
weight_model.fit(X_weight_train, y_ weight _train)

# Fitting a multiple linear regression model for the medium variable
medium_model = LinearRegression()
medium_model.fit(X_medium_train, y_medium_train)

# Fitting a multiple linear regression model for the mild variable
mild_model = LinearRegression()
mild_model.fit(X_mild_train, y_mild_train)
```

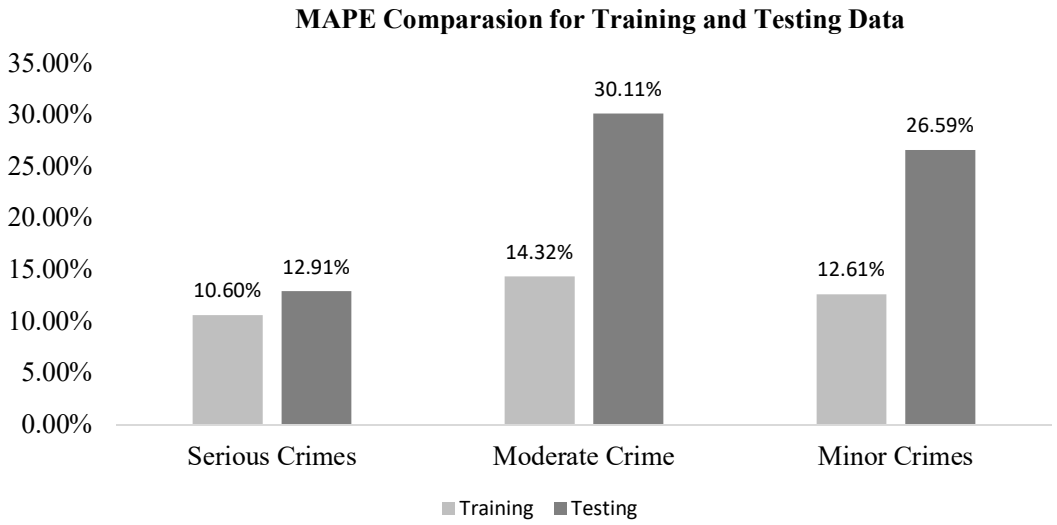

Figure 4. MAPE comparison.

The model has a lower prediction error rate (MAPE) on the training data than on the testing data, as can be seen from the comparison in Figure 4. This suggests that known data can be predicted more accurately than previously unknown data. The test data's moderate and light crime categories had significantly different MAPE values, which suggests that there may be difficulties applying the model to new data. To enhance the model's performance on test data, more work might be needed, such as modifying the model's parameters or including features.

4. CONCLUSION

The study's findings indicate that the multiple linear regression model performs well in forecasting the volume of criminal cases in Pontianak. Approximately 10.60% for the heavy crime category, 14.32% for the medium crime category, and 12.61% for the light crime category are the model's prediction error rates (MAPE) on the training data, according to the study results. However, there are issues with generalizing the model to previously unseen data in the test data, particularly in the moderate and light crime categories. To improve performance on test data, more model modifications or feature additions can be required. However, this study offers valuable insights into the model's potential for understanding and managing crime in Pontianak.

REFERENCES

[1] T. Marini, "Analisis Faktor-Faktor Yang Mempengaruhi Pertumbuhan Ekonomi Dan Tingkat Kemiskinan Di Kabupaten Berau," *J. Ekon. Keuangan, dan Manaj.*, vol. 12, no. 1, pp. 108–137, 2016.

[2] K. Mendome, N. Nainggolan, and J. Kekenusa, "Penerapan Model Arima Dalam Memprediksi

Jumlah Tindak Kriminalitas Di Wilayah Polresta Manado Provinsi Sulawesi Utaraklorofil," *J. Mipa*, vol. 5, no. 2, pp. 113–116, 2016.
[3] U. Saipi, H. Kadir, and J. Lantowa, "Kriminalitas Dalam Novel Perjanjian Rahasia Karya Sandra Brown," *J. Bahasa, Sastra, dan Budaya*, vol. 11, no. 1, pp. 61–75, 2021.
[4] Z. Resiana and T. Aditya, "Analitik Geovisual Pengaruh Pandemi COVID-19 Terhadap Pola Dan Kecenderungan Kriminalitas Di Daerah Istimewa Yogyakarta," *JGISE J. Geospatial Inf. Sci. Eng.*, vol. 6, no. 1, p. 24, 2023, doi: 10.22146/jgise.80670.
[5] "STATISTIK KRIMINAL 2021 i," 2021.
[6] C. Angelita, "Kajian Hukum Tindak Pidana Kekerasan Fisik Dalam Rumah Tangga Ditinjau Dari Perspektif Viktimologi Kritis," *JUSTITIA J. Ilmu Huk. dan Hum.*, vol. 9, no. 4, pp. 2008–2019, 2022.
[7] D. Abdullah, M. Maryana, and M. Ani, "Prediksi Tingkat Pengguna Narkoba Dengan Metode Regresi Linear Berganda Berbasis Web," *TECHSI - J. Tek. Inform.*, vol. 13, no. 2, p. 41, 2021, doi: 10.29103/techsi.v13i2.3738.
[8] A. D. W. Sumari, R. Y. A. Pratama, and O. D. Triswidrananta, "Sistem Prediksi Tingkat Kriminalitas Menggunakan Metode Triple Exponential Smoothing: Studi Kasus Pada Polres Kabupaten Probolinggo," *J. Tek. Inform.*, vol. 13, no. 2, pp. 171–178, 2021, doi: 10.15408/jti.v13i2.18128.
[9] M. H. Lubis and S. Sumijan, "Prediksi Tingkat Kriminalitas Menggunakan Metode Single Moving Average (Studi Kasus Polres Asahan Sumatera Utara)," *J. Sistim Inf. dan Teknol.*, vol. 3, pp. 183–188, 2021, doi: 10.37034/jsisfotek.v3i4.63.
[10] I. M. YULIARA, "REGRESI LINIER SEDERHANA," *Br. J. Anaesth.*, vol. 62, no. 4, pp. 429–433, 2016, doi: 10.1093/bja/62.4.429.
[11] T. Theodoridis and J. Kraemer, *Analisis Data Menggunakan Multiple Logistic Regression Tst di Bidang Kesehatan Masyarakat dan Klinis*.
[12] P. Purwadi, P. S. Ramadhan, and N. Safitri, "Penerapan Data Mining Untuk Mengestimasi Laju Pertumbuhan Penduduk Menggunakan Metode Regresi Linier Berganda Pada BPS Deli Serdang," *J. SAINTIKOM (Jurnal Sains Manaj. Inform. dan Komputer)*, vol. 18, no. 1, p. 55, 2019, doi: 10.53513/jis.v18i1.104.
[13] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.
[14] I. N. Abrar and A. Abdullah, "Klasifikasi Penyakit Liver Menggunakan Metode Elbow Untuk Menentukan K Optimal pada Algoritma K-Nearest Neighbor (K-NN)," vol. 12, pp. 218–228, 2023.
[15] F. Gorunescu, *Data Mining: Concepts, models and techniques*. 2011.
[16] M. A. Muslim *et al.*, "Data Mining Algoritma C4.5 Disertai contoh kasus dan penerapannya dengan program computer," *Nucl. Phys.*, vol. 13, no. 1, pp. 104–116, 2019.
[17] G. A. Marcoulides, *Discovering Knowledge in Data: an Introduction to Data Mining*, vol. 100, no. 472. 2005. doi: 10.1198/jasa.2005.s61.
[18] Z. Maisat, E. Darmawan, and A. Fauzan, "Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM Implementation of GridSearchCV Hyperparameter Optimization in Heart Attack Prediction System Using SVM," vol. 13, no. 1, pp. 8–15, 2023.
[19] M. M. Hidayat, "Data Mining Data mining," *Min. Massive Datasets*, vol. 2, no. January 2013, pp. 5–20, 2015.
[20] Kusrini, E. T. Luthfi, and Universitas Amikom, *Algoritma Data Mining - Google Buku*. 2009.
[21] T. Syahputra, J. Halim, and K. Perangin-angin, "Penerapan Data Mining Dalam Memprediksi Tingkat Kelulusan Uji Kompetensi (UKOM) Bidan Pada STIKes Senior Medan Dengan Menggunakan Metode Regresi Linier Berganda," *Sains dan Komput.*, vol. 17, no. 1, pp. 1–7, 2018.
[22] A. Géron, *Hands-on Machine Learning whith Scikit-Learing, Keras and Tensorfow*. 2019.
[23] D. N. A. Janie, *Statistik deskriptif & regresi linier berganda dengan SPSS*, no. April 2012. 2012.
[24] B. Widiyawati, "Abstract Selection of the Best Multiple Linear Regression Model in Multicolinearity Case With Principal Component Regression and Stepwise Regression," p. 6,

2021.

[25] C. V. Hudiyanti, F. A. Bachtiar, and B. D. Setiawan, "Perbandingan Double Moving Average dan Double Exponential Smoothing untuk Peramalan Jumlah Kedatangan Wisatawan Mancanegara di Bandara Ngurah Rai," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2667–2672, 2019.

[26] M. Al Husaini, A. Hermansyah, V. Purwayoga, H. H. Lukmana, and D. Ramadhan, "Aplikasi Cerdas Berbasis Website Prediksi Harga Emas dengan Implementasi Algoritma Smoothing Time Series Forecasting," *Data Sci. Indones.*, vol. 2, no. 2, pp. 30–43, 2022, doi: 10.47709/dsi.v2i2.1888.