Using_Multinomial_Logistic_Regression_and_K_Nearest_Neighbor.pdf

by

Submission date: 17-Apr-2023 06:08AM (UTC+0700)
Submission ID: 2066307819
File name: Using_Multinomial_Logistic_Regression_and_K_Nearest_Neighbor.pdf (317.55K)
Word count: 2671
Character count: 13805

Modeling Public Crime Type Using Multinomial Logistic Regression and *K*-Nearest Neighbor: Pre-and During-Pandemic COVID-19

Riswan Efendi[1,2,*], Yaumil Isnaini[2], Sri Widya Rahayu[2], Rohaidah Masri[1], Noor Azah Samsudin[3], Rasyidah[4]

[1]Mathematics Department, Faculty of Sciences and Mathematics, Universiti Pendidikan Sultan Idris
35900 Tanjung Malim, Perak, Malaysia
[2]Faculty of Science and Technology, UIN Sultan Syarif Kasim
28293 Pekanbaru, Riau, Indonesia
[3]Faculty of Computer Science and Information Technology
Univerisiti Tun Hussein Onn
84600 Batu Pahat, Johor, Malaysia
[4]Information Technology Department, Politeknik Negeri Padang
25164 Padang, Indonesia
riswanefendi@fsmt.upsi.edu.my

Abstract. There are common factors that characterize individuals who commit crime. Statistically, relationship between factors could be measured and formed using regression models. While modeling crime rates was widely approached using the ordinary regression model. However, this model is not much capable for categorical response variables such as crime type. In this paper, we are interested to classify crime type using some independent variables using multinomial logistic regression and *K*-Nearest Neighbor (*K*-NN) models. While both are powerful models for classification purposes. The secondary crime data was collected from 2019 (pre-pandemic) and 2020 (during a pandemic) in the police office of Payakumbuh Region, West Sumatra, Indonesia. The results indicate that the crime type was influenced by employment status (unemployed persons) and time occurring (daytime) for both years periods. In the testing data phase, the average of accuracy levels of multinomial logistic regression and *K*-NN are 66.81% and 73.86%, respectively. In this case study, *K*-NN model is better approach to be used for the prediction and classification of crime type if compared with multinomial logistic regression. Both models could be considered for supporting the police divisions on decision making and prevention strategy.

Keywords: Crime type, Logistic regression, K-Nearest Neighbor, Decision making, Pandemic era

1 Introduction

Statistics Indonesia declared that crime number has been decreased since 2017-2019. While this number was increased again during the pandemic COVID-19. It is caused by some factors such as termination of employment, working from home, isolation or lockdown, others. Consecutively, the unemployment rate was also increased rapidly. This rate will stimulate the crime rates and criminal actions in this country such as theft, plunder, robbery, online or offline deceptions. All criminal actions are continuously collected, monitored, and projected by police-crime divisions, government, and other counterparts anytime. Indonesia is very well-known with multi ethnics, cultures, and very big population around the world. The population explosion is a serious issue in this country because it the not an easy task to control and push the people to follow the rules and norms in daily life activities. Therefore, crime rate and type are very indispensable areas to be concerned and projected in supporting environment, health, and safety levels and minimizing the crime rate [1].

Many existing studies have been investigated crime rate prediction and forecasting using machine learning and data mining approaches [6, 10] such as ordinary regression, K-NN [3-5, 7], logistic regression [2, 11], Naïve Bayes [3], decision tree algorithm [3], and others. These approaches can predict the association between the crime response variable and its independent variables. In certain cases, most crime data sets are represented in categorical types such as crime type, preparator demographics, and time occurring. Motivated by data types, we are interested to investigate the crime type and its factors using multinomial logistic regression and K-NN models. Interestingly, pre-pandemic and during-pandemic crime data sets are also compared using both models, respectively.

In this paper, we explore the implementation of multinomial logistic regression and K-NN algorithms in projecting crime types based on external factors. The crime type prediction will be derived by using some stages such as data preparation, multinomial logistic regression model, K-NN model, evaluation, and comparison models.

2 The Basic Concepts and Methodology

2.1 Multinomial Logistic Regression (MLR) Model

MLRM is a straightforward generalization of the binary model, and both models depend mainly on logit analysis or logistic regression. Logit analysis in many ways is the natural complement of ordinary linear regression whenever the response is a categorical variable. When such discrete variables occur among the explanatory variables, they are dealt with by the introduction of one or several (0, 1) dummy variables, but when the response variable belongs to this type, the regression model breaks down. Logit analysis provides a ready alternative. Mathematically, MLR model is written as [8, 9]:

$$\pi(x) = \frac{\exp\left(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p\right)}{1+\exp\left(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p\right)}. \quad (1)$$

From Eq. (1), $\pi(x_i)$ is an occurred probability for event-i, $(x_1, \ldots, x_p)$ are independent factors or variables, and $(\beta_1, \ldots, \beta_p)$ are slopes MLR, respectively.

Additionally, some stages of MLR such as estimation parameters, parameters testing, model validation, and classification accuracy as presented in Figure 1.

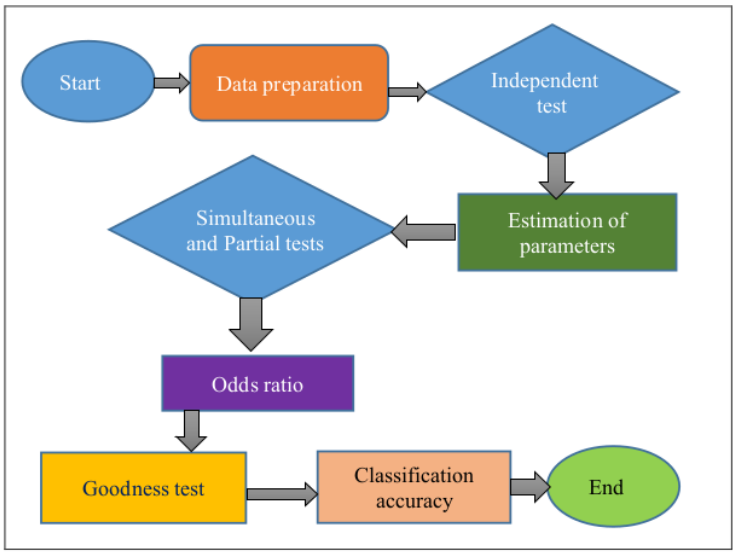

Figure 1. Stages of MLR

2.2 *K*-Nearest Neighbor (*K*-NN)

The main objective of *K*-NN is to predict the label of a query instance based on the labels of *K* closest instances in the stored data, assuming that the label of an instance is similar to that of its *K*-NN instances. *K*-NN is simple and easy to implement but is very effective in terms of prediction performance. *K*-NN makes no specific assumptions about the distribution of the data. Because it is an instance-based learning algorithm that requires no training before making predictions, incremental learning can be easily adopted. For these reasons, *K*-NN has been actively applied to a variety of supervised learning tasks including both classification and regression tasks [4]. Mathematically, the label's prediction is calculated by using steps given as follows:

Step 1: Data preparation. In this step, all categorical data should be transformed to numerical types.
Step 2: Select the number *K* of the neighbors.
Step 3: Calculate the Euclidean distance of *K* number of neighbors using mathematical formula as below:

$$d_{euclidean} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \quad (2)$$

Step 4: Take the K nearest neighbors as per the calculated Euclidean distance.
Step 5: Among these K neighbors, count the number of the data points in each category.
Step 6: Assign the new data points to that category for which the number of the neighbor is maximum.

K-nearest neighbors is a non-parametric method used for classification and regression. It is one of the easiest machine learning techniques used. Besides, this method is also a lazy learning model, with local approximation. Based on our review, K-NN and MLR are appropriate models to be considered for classification and prediction purposes, especially categorical variable types. While K-NN is a non-parametric approach.

3 Modeling Crime Type Using MLR and K-NN

Section 3 demonstrates how can MLR and K-NN be used to model and classify crime types and their comparative performances. In the beginning stage, descriptive statistics regarding crime rates from 2019 and 2020 in the Payakumbuh Region of West Sumatra is presented in Table 1 and Figure 2, respectively.

Table 1. Descriptive statistics for crime rate from 2019 to 2020 in Payakumbuh

Independent Variable (Factor)	2019 Freq.	2019 (%)	2020 Freq.	2020 (%)
Age (x_1)				
≤18-year-olds	7	8.61%	4	7.14%
>18-year-olds	74	91.35%	52	92.85%
Employment status (x_2)				
Unemployed	43	53.08%	37	66.07%
Employee	38	46.91%	19	33.92%
Time occurring (x_3)				
Morning	26	32.09%	20	35.71%
Lunch time	22	27.16%	19	33.92%
Evening	13	16.04%	10	17.85%
Night	20	24.69%	7	12.50%

Table 1 indicates the criminal activities are frequently occurred by persons aged more than 18-year-olds in pre-and during the pandemic era. While the highest percentage occurred in 2020. Additionally, the unemployed were the dominant category in these activities if we compared with the persons who have jobs, especially during COVID-19. On average (2019 and 2020), the criminal activities were done in the morning and lunchtime at Payakumbuh Region.

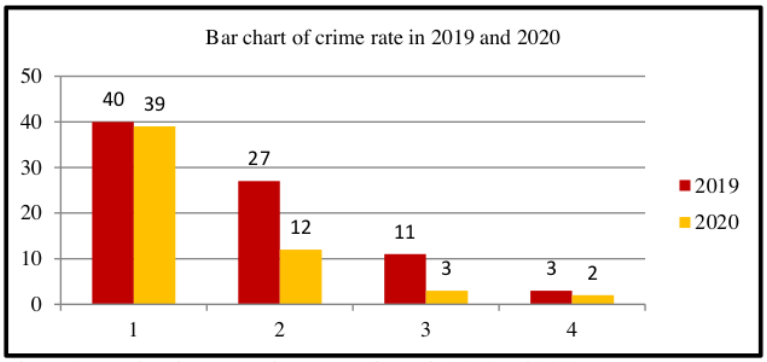

Figure 2. Bar chart between crime rate and type in 2019 and 2020

Based on Figure 2, theft crime cases are the highest frequency in 2019 and 2020 among other cases. It is caused by the persons who do not have jobs as mentioned in Table 1 also. Moreover, how significant of the independent variables above influence the crime type. By using MLRM stages in Section 2, the association between dependent and independent variables is measured as follows:

Step 1: Measure the association between variables using the Chi-square test. It is presented in Table 2.

Table 2. Association between crime type and its independent variables

Year	Dependent variable	Independent variable	Chi-square test
2019	Crime type (y)	Age (x_1)	No association
		Employment status (x_2)	Significant association
		Time occurring (x_3)	Significant association
2020	Crime type (y)	Age (x_1)	No association
		Employment status (x_2)	Significant association
		Time occurring (x_3)	Significant association

Table 2 shows that employment status and time occurring variables have a significant association with the crime type at pre-and during pandemic COVID-19, respectively. Thus, both independent variables should be handled seriously by police divisions, the regional authority, and the societies in this region.

Step 2: Build MLR model for public crime type (pre-and during-pandemic) using stages as described in Section 2.1 as presented in Table 3.

Table 3. MLR for public crime (pre-pandemic and during-pandemic)

MLR (Pre-Pandemic, 2019)	MLR (During-Pandemic, 2020)
g_1 (x) = 17.976 – 18.692 x_3 (2) - 18.303 x_3 (3)	g_1 (x) = 18.865 – 18.116 x_3 (2) – 21.595 x_3 (3)
g_2 (x) = 19. 851 – 18.819 x_3 (2) – 18.541 x_3 (3)	g_2 (x) = 19.898 – 19.264 x_3 (2)
g_3(x) = 19.090 – 18. 904 x_3 (2)	

Based on Table 3, the time occurring of crime type is the most significant independent variable if compared with perpetrator age and employment status in pre-pandemic (2019) and during-pandemic (2020). In 2019, odds ratios indicate the theft and fraud cases tend to have occurred during the daytime about 7.62 and 6.169 times if compared with the evening time and at night. While the persecution case tends to have occurred about 8.867 times in the evening time if compared with daytime and at night. These ratios are very important to be considered by the police division, crime division, or other counterparts in improving the residential safety and prevention purpose in the Payakumbuh region of West Sumatra. In this decade, most public crimes occurred in the morning and daytime in Indonesia, because most house owners were working in the office or outdoor activities during crime events occurred. A little bit different during-pandemic 2020, the theft case tends to have occurred in the evening time about 4.181 times if compared with the daytime and at night. Probably, the house owners were doing something outside in the afternoon until evening because they were bored staying at home during lockdown periods.

Step 3: Build *K*-NN for investigating public crime type (pre-and during-pandemic) using stages as described in Section 2.2 as presented in Tables 4 and 5, respectively.

Table 4. Euclidian distance for crime type (pre-pandemic, 2019) using *K*-NN

Case. No	Crime type	Age	Employment status	Time occurring	Euclidean distance
1	2	2	2	3	1.414214
2	2	1	2	3	1.732051
3	2	2	2	2	1
4	1	2	1	1	1
5	1	2	1	1	1
…	…	…	…	…	…
64	1	2	1	1	1
65	2	1	2	2	1.414214

Table 5. Euclidian distance for crime type (during-pandemic, 2020) using *K*-NN

Case. No	Crime type	Age	Employment status	Time occurring	Euclidean distance
1	1	2	1	1	1
2	1	2	1	1	1
3	1	2	1	2	0
4	1	2	1	1	1
5	1	2	1	1	1
…	…	…	…	…	…
44	1	2	1	1	1
45	1	2	1	2	0

Based on Tables 4 and 5, all distance values are calculated using Euclidian distance formula (Eq. 2). The distance value is calculated between the target data and training data. From both tables, the minimum distances are 0 and 1. Moreover, the accuracy of prediction category can be validated using confusion matrix as presented in Table 6.

Table 6. Confusion matrix for accuracy of prediction public crime type

Classification		2019				2020			
		1	2	3	4	1	2	3	4
Class	1	5	2	3	4	8	1	2	0
Obs.	2	1	0	0	1	0	0	0	0

Step 4: Compare results between MLR and *K*-NN in investigating public crime type (pre-and during-pandemic) as presented in Table 7.

Model	Accuracy of testing data	
	2019	2020
MLR	50%	72.72%
K-NN	75%*	72.73%*

Table 7 shows that *K*-NN can predict the public crime type better than MLR for 2019 and 2020, respectively. In average, the prediction accuracy of *K*-NN is higher than MRM model. It is caused by some components such as the number of training data and independent variable type. Additionally, all independent variables of crime are the categorical type in this study. Thus, these variables are appropriate to be analyzed and predicted using *K*-NN rather than MLR models. While the significant independent variables and the odds ratios by category cannot be determined using *K*-NN model. Therefore, both algorithms have different advantages to help and support decision-making in the public crime type prediction.

4 Conclusion

The investigating public crime type and its determinants is a very interesting study because of many related factors involved. This investigating has been attempted by applying two different models namely multinomial logistic regression (MLR) and *K*-nearest neighbor (*K*-NN). Based on MLR, the employment status and time occurring have a non-linear relationship (logit function) with the public crime type. Additionally, the theft case is the most public crime type that has occurred by unemployed persons during the daytime (working hours) of pandemic COVID-19 (2020). Based on prediction accuracy, *K*-NN is better than MLR in predicting public crime type. Moreover, *K*-NN supports non-linear solutions and can only output the labels. Finally, this paper may help and support the police division, criminal division, and government in improving safety levels or minimizing the crime rates based on historical data from any region in Indonesia. Because data analytics is one of the important components in decision-making in a digital era.

Reference

1. Wibowo, A. H., Wijaya, A. R.: Kajian awal profiling kejahatan dan strategi dalam usaha mencegah terjadinya tindak kriminalitas di Kabupaten Sleman. Proceeding of National Conference IDEC. 1-8 (2018).
2. Putra, A. D., Martha, G. S., Fikram, M and Yuhan, R. J.: Faktor-faktor yang mempengaruhi tingkat kriminalitas di Indonesia tahun 2018. Indonesian J. App. Stat. **3**, 123-131 (2018).
3. Wibowo, A. H and Oesman, T. I.: The comparative on the accuracy of K-NN, Naïve Bayes and decision tree algorithms in predicting crimes and criminal actions in Sleman regency. J. Phys. Conf. Ser. **1450**, 012076 (2020).
4. Kang, S.: K-nearest neighbor learning with graph neural network. Math. **9**, 1-12 (2021).
5. Pednekar, V., Mahale, T., Gadhave, P., and Gore, A.: Crime rate prediction using KNN. Int. J. Recent and Innovation Trend Compt. & Comm. **6**, 124-127 (2018).
6. Mahmud, S., Nuha, M and Sattar, A.: Crime rate prediction using machine learning and data mining. S. Burak *et al.*, (eds), Soft Compt. Tech. and App. Adv. Intel. Syst. Compt. **1248**, (59-69) (2021).
7. Kumar, A., Verma, A., Shinde, G., Sukhdeve, Y and Lal, N.: Crime prediction using K-NN algorithm. Proceeding ic-ETITE (2020).
8. Agresti, A.: Categorical data analysis. 3rd Ed. Wiley. (2012).
9. El-Habil, A. M.: An application on multinomial logistic model. Pak. J. Stat. Oper. Res. **7**, 271-291 (2012).
10. Shah, N., Bhagat, N and Shah, M.: Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Vis. Compt. Industry. Bio. Art. **4** (2021).
11. Kim, K-S and Jeong, Y-H.: A study on crime prediction to reduce crime rate based on artificial intelligence. Kor. J. Art. Intel. **9**, 15-20 (2021).

Using_Multinomial_Logistic_Regression_and_K_Nearest_Neig...

ORIGINALITY REPORT

10%	%	10%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Seokho Kang. "k-Nearest Neighbor Learning with Graph Neural Networks", Mathematics, 2021 Publication	4%
2	Pethuru Raj, D. Peter Augustine, P. Beaulah Soundarabai. "The machine learning algorithms for data science applications", Institution of Engineering and Technology (IET), 2022 Publication	3%
3	Riswan Efendi, Sri Widya Rahayu, Rohaidah Masri, Nor Azah Samsudin, Rasyidah. "Chapter 31 Most Profitable Currency Exchange for ASEAN Countries Using Dijkstra's Algorithm", Springer Science and Business Media LLC, 2022 Publication	3%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		