Mining the crime data using naïve Bayes model

Lourdes M. Padirayon1, Melvin S. Atayan2, Jose Sherief Panelo3, Carlito R. Fagela, Jr4
1,3College of Information & Computing Sciences, Cagayan State University Sanchez Mira, 3518, Philippines
2,4College of Criminal Justice Education, Cagayan State University Sanchez Mira, 3518, Philippines

ABSTRACT
A massive number of documents on crime has been handled by police departments worldwide and today's criminals are becoming technologically elegant. One obstacle faced by law enforcement is the complexity of processing voluminous crime data. Approximately 439 crimes have been registered in Sanchez Mira municipality in the past seven years. Police officers have no clear view as to the pattern crimes in the municipality, peak hours, months of the commission and the location where the crimes are concentrated. The naïve Bayes model is a classification algorithm using the Rapid miner auto model which is used and analyze the crime data set. This approach helps to recognize crime trends and of which, most of the crimes committed were a violation of special penal laws. The month of May has the highest for index and non-index crimes and Tuesday as for the day of crimes. Hotspots were barangay centro 1 for non-index crimes and barangay centro 2 for index crimes. Most non-index crimes committed were violations of special law and for index crime rape recorded the highest crime and usually occurs at 2 o'clock in the afternoon. The crime outcome takes various decisions to maximize the efficacy of crime solutions.

Keywords:
Classification algorithm
Crime data
Mining
Naive Bayes classifier

1. INTRODUCTION
Global protection in the present world is an aspect in which it aimed in reducing the incidence of crime [1], a significant issue and a prominent reality of the current situation [2]. A problem that limits the freedom of citizens to engage in society. Crime as a social phenomenon could be both destructive and distractive and can generate substantial fear within the community national crime prevention institute (NCPI) [3] and could hamper faster social development.

To combat violence for society to feel safe and protected. It needs multiple public policy and regulatory interventions coupled with a more proactive community of people to help. Yet the crime issue it self is still persistent amidst all efforts of the government. Global patterns show that the volume of crimes has been on the rise. Policymakers are constantly trying to look for possible solutions [4], for a well-informed society is of advantage in combatting the crime problem [5]. More so, it is widely claimed that community participation is indispensable in addressing crimes that lead to increase crime records.

As stated in the PNP annual report (2016) crime volume is based in the revised penal code that contains the general penal laws of the Philippines. The code included the classification of crimes as crimes against persons and crimes against property which could be further classified as index crimes and non-index crimes. Index crimes, as defined by the Philippine national police (PNP), are those crimes against persons
such as murder, homicide, physical injury, and rape, and crimes against property such as robbery, theft, carnapping or carjacking, and cattle rustling. Non-index crimes, on the other hand, are violations of special laws such as illegal logging or local ordinances kinds of crimes are reported to the police.

Since 2010, the PNP has adopted the national crime reporting system (NCRS) that provides readily available data on crime incidences in a given location. In 2013, there were about 9334 index crime incidents across Region 2 of the Philippines comprising of the provinces of cagayan, isabela, quirino, nueva vizcaya, and batanes [6] and has a population of around 3.36 million [7].

In the past seven years 2013 to 2019, approximately 439 crimes were reported in the municipality of Sanchez mira. Of these crimes, police officers have no clear view on the trend of prevalent crimes in the municipality as well as the peak hours and months of crime commission and the place where crime incidence is concentrated. There is knowledge from the crime data in the municipality. A knowledge that can be used to provide a better basis for formulating courses of action and to prevent the commission of the crime in a given place in a given time.

Further, a massive number of documents on crime have been handled by police departments worldwide and are becoming even more complicated. Difficulty in processing a large amount of data involved in crime is one obstacle faced by intelligence and law enforcement agencies [8]and could adversely affect the effectiveness and efficiency of police performance.

Besides, there is a need to analyze the increasing number of crimes [9], so that agencies will know the most appropriate technique to apprehend the criminal and become more advantageous over them. Such an important domain has been accepted among many experts and specialists in criminal justice and law enforcement [10] who must be guided with empirical data that could be of help in policymaking and law enforcement actions [11] and to have practical measures and help police officers and investigators to enhance their crime solution efficiency.

Furthermore, data mining was used to analyze the pattern from a huge amount of crime records or knowledge discovery in databases (KDD) to gain some information. It is the exact field applicable to a high-volume crime dataset that can discover hidden knowledge, unexpected patterns, and new rules from large amounts of crime data. Likewise, it is one of the important applications with many tasks that perform classification, association, clustering and each of them has its significance [12], and the knowledge obtained from data mining will surely help police officers in their complex tasks.

In the data mining field, classification is a commonly used technique [13]. It is an unavoidable task by which the data could be classified by the previously recognized class labels [14]. Classification is a supervised activity of machine learning that generates a model based on labeled data [15]. The model is used for class determination and there are several kinds of algorithms for classification, such as naïve Bayes, SVM, k-NN, neural network, logistic regression, decision tree, random fores [16].

The rapid miner auto model is used in the study and suggests the best learning technique which is the naïve Bayes classifier. It is a predictive classifier that can be used to make a decision based on the data. As a technique used to identify crime patterns and improve the efficiency of crime solutions in the range of law enforcement [17], with crime categories such as barangay where crimes are committed, day and month, hotspots, hot place, seasonal, or frequency. The crime patterns to be extracted will be the basis for law enforcement agencies to develop crime prevention programs comprehensively.

Thus, the primary goal of the study is to identify the pattern of committed crime incidents that are the most prevalent in the municipality based on the revised penal code of the Philippines. Specifically, to determine the month, the time, the day, and the barangay of Sanchez Mira significantly occurred the most prevalent crime. To be able to obtain the crime trends, a naïve Bayes model which is a classification algorithm using the Rapid miner auto model is used to analyze the set of crime data [18]. Further, the result basis in proposing an intervention in an online system or an AI environment, an ideal crime analysis tool to be able to identify crime patterns quickly and in an efficient manner for future crime pattern detection and action, which is hoped to be beneficial for the municipality of sanchez mira.

Conceptual framework
The framework of the study was based on the knowledge discovery process (KDP) illustrated by [19]. The KDP figure was modified to suit the objectives of the study. The modified version was presented in Figure 1 following the steps from preprocessing wherein noisy and irrelevant data were removed, selection and transformation where data relevant to the analysis task were retrieved from the database and further transformed or consolidated into forms appropriate for mining, data mining where naïve Bayes classifier was applied to extract data patterns, interpretation, and evaluation where the truly interesting patterns representing knowledge based were identified and knowledge presentation where visualization and knowledge presentation techniques were used to present the mined knowledge to the user.
2. RESEARCH METHOD
2.1. Data set
The dataset used in this study was collected from the result being compiled in the office of the investigation and operation section of the unit crime periodic report system result of the Philippine national police office-sanchez mira from the year 2013-2019. Data collected were entered into the Microsoft Excel Application for data cleansing or attributes needed to be transformed into nominal data. Standardizing the data involved the following steps: 1) removing extra spaces, 2) filling all blank cells with ‘0’, 3) converting numbers stored as text into numbers, 4) removing duplicate values from the data set, 5) changing text to lower/upper/proper case for consistency, and checking the spelling.

The standardized data set is transformed by saving the data into CSV (comma separated values) format. After data cleansing, these were entered into the rapid miner application for pre-processing. The result can be seen in Table 1, it shows that the data set arranged according to month, crime type, classification, barangay, time and day.

Table 1. Sample UCPER data

| No | Month    | Crime Type          | Classification     | Barangay  | Time       | Day       |
|----|----------|---------------------|--------------------|-----------|------------|-----------|
| 1  | January  | Index               | Crime against person | Dacal     | 23:24      | Wednesday |
| 2  | February | Non-Index           | Crime against property | Masisit   | 1:30 pm    | Thursday  |
| 3  | March    | Index               | Crime against property | Centro 02 | No time    | Tuesday   |
| 4  | March    | Index               | Crime against property | no place  | 18:30      | Saturday  |
| 5  | April    | Non-Index           | Crime against property | San Andres | 1:30       | Tuesday   |
| 6  | April    | Index               | Crime against person  | San Andres | 20:50      | Thursday  |

2.2. Naïve Bayes classifier
In this analysis, the naïve Bayes classification method was certain to examine the dataset mined from the UCPER (unit crime periodic report) forecast population data result. The classification discusses data from various classes in one of the two parts of supervised learning. The training dataset trains the model to predict the unspecified population [20]. Entirely these algorithms have their style of implementation and different methods of classification [21].

Naïve bayesian classifier is a simple probabilistic classifier that works by applying the Bayes’ theorem along with Naïve assumptions about feature independence. Despite its independence assumption, the Naïve Bayesian classifier is proved to be quite useful in modeling the conditions of the complex real-world problem [14] for the theorem based on the Bayes's theorem that gets the probability of the event based on the knowledge already known about the conditions related to the event. Bayes’ theorem is stated mathematically as [22].

\[
A(B/C) = \frac{A(C/B)A(B)}{A(C)} \quad \text{--- 1, where } B \text{ and } C \text{ are the events and } A(C) \neq 0, \quad A(B/C) \text{ – the likelihood of event } B \text{ occurring given that } C \text{ is true, } A(C/B) \text{ - the likelihood of event } C \text{ occurring given that } B \text{ is true, } A(B) \text{ and } A(C) \text{ are probabilities of observing } B \text{ and } C \text{ independently of each other.}
\]

2.3. Naïve Bayes in rapid miner’s auto model
As [23] used a long procedure or the black boxes that start with the blank process, create process model by putting different operators like retrieve, numerical to polynomial, set role, validation rule, and naïve Bayes classifier. to plot the crime data and to see a different statistical chart of crime in different states just to establish that how efficiently naïve Bayes algorithm. Likewise [24] used a model based on the naïve Bayes classifier but encourages further studies on the criminal prediction problem with its new methodologies for both crime dataset generation and decision-making with higher accuracy rates.
Auto model is an extension to rapid miner's auto model extension that speeds up the process of building and validating models. It also makes it easier to modify or bring into production since no black boxes are used. Auto model addresses three large classes of problems: prediction, clustering, outlier. Within the prediction category, it can solve both classification and regression problems. Once the calculations are complete, auto model gives a list of models that are suitable for a problem. Model types are naïve bayes, generalized linear model, logistic regression, deep learning, decision tree, random forest, gradient boosted trees (XGBoost). Auto model evaluates the data, provide and compare appropriate model results for the solution of the problem. The auto model recommends the best machine learning methods and with the naïve Bayes model being the best for the data crime dataset.

Naive Bayes is a low-variance, high-bias classifier that can create a good model even with a limited collection of data. In this study the classification issue is divided into two categories: crime “index” and “non index,” with the findings based on a mixture of incidents. It denotes a grouping of events as attributes or features of each observation in machine learning. The events are used as attributes in the crime data collection, and the Naive Bayes algorithm makes an assumption that the algorithm treats all of the attributes equally, assuming that attributes are all similarly important and statistically independent. This implies the worth of one attribute has no bearing on the worth of another. The naïve Bayes algorithm counts the number of times each combination of an attribute value with each of the possible classes occurs, then converts the counts into probabilities. Simply divide each count by the number of observations in each class to get the right probabilities.

To begin the data analysis in rapid miner auto model, prepare target crime dataset in comma-separated values (CSV) format with different parameters like a month, crime type, classification, barangay, time, and day for the year 2013 to 2019 entered into the rapid miner application from the source hard drive. By selecting the data location, specify data format, format columns, and choose where to store the data.

The naïve Bayes model is given by rapid miner auto model to map the crime data to understand the various crime statistical charts. Rapid miner offers various types of processing scenarios like similar churn modeling [25], demand based analysis, and outlier detection. But in this study, the researchers use the rapid miner auto model like, in Figure 2. In rapid miner, the entered data exists under the 'local repository' option. For data plotting, an auto model method is provided by the rapid miner. Methods such as load data, select tasks, select inputs model types, and results. Among the models given by the auto model, on top of it was the naïve Bayes operator thus, the best model for the problem was the naïve Bayes model.

The goal of the analysis is to supply the input to the naïve Bayes classifier along with the validation rule [15] and see the results in Figure 3. The performance assessment of the naïve Bayes classifier will be given by this setup. To set up the procedure and obtain the desired result using naïve Bayes classifier, the following steps are to be implemented in rapid miner.

**Figure 2. Rapid miner auto model interface**

![Rapid miner auto model interface](image)

The goal of the analysis is to supply the input to the naïve Bayes classifier along with the validation rule [15] and see the results in Figure 3. The performance assessment of the naïve Bayes classifier will be given by this setup. To set up the procedure and obtain the desired result using naïve Bayes classifier, the following steps are to be implemented in rapid miner.
3. RESULTS AND DISCUSSION

Finally, data analysis was performed to find out that test result’s attributes served as the basis for classifying the various scopes of the crime records. Figure 4 shows the crime types committed in the municipality of Sanchez mira, cagayan, for the calendar year 2013-2019. It can be gleaned in the figure that for index crimes, crimes against persons, and crimes against property got the highest followed by violations of special penal laws.

On the other hand, for non-index crimes, the crimes against chastity, crimes against personal liberty and security, quasi offenses recorded the highest. Most of the crimes committed violated special penal laws. It is followed by crimes against the person. The lowest number of crimes were plotted on crimes against public order.

It can also glanced from the Figure 5 that for index crimes, barangay centro 2 has the highest cases followed by barangay centro 1. Barangay tokitok and barangay nagrangtayan making them the lowest record of index crimes. Barangay centro 1 has the highest non-index crimes committed followed by barangay centro 2, barangay langagan, barangay callungan, and namuac. Barangay tokitok plotted the lowest non-index crime.
The data also show that the most non-index crimes [4] committed were special law specifically RA 9262 otherwise known as violence against women and their children. It is followed by violations of RA 9165 comprehensive dangerous drugs act. The lowest were cases of acts of lasciviousness. For index crime, it is alarming that the crime of rape recorded the highest cases. It is followed by robbery and theft while carnapping registered the lowest.

The data was also analyzed according to the months when the crimes were committed and it yielded that month of May has the greatest number of crimes committed in the municipality. Comparing the data on index and non-index when analyzed by month shows that May has the highest both for index and non-index crimes while January and February are the safest months for their low record of both index and non-index crimes [5]. Analyzing Figure 6 suggests that for index crimes, the months of October and November plotted the highest number of crimes committed while June and December have the lowest index crimes. Looking further on the figure shows that March has the highest number of non-index crimes in the municipality followed by June and December respectively.

Analyzing Figure 7 and it was found out that crime usually occurs at 2 p.m., followed by those committed at 1 o’clock in the afternoon while non-index crimes were low at 11 a.m. and 7:30 in the evening. On the other hand Figure 8 which index crimes are commonly committed in the afternoon as shown by the findings that many of the index crimes committed in the municipality were committed at 2 p.m. and 7 p.m. Index crimes are not common at 10 o’clock to 11 o’clock as evidenced by the recorded cases of two (2) each
only. Non-index crimes are usually committed in the afternoon and early morning as shown in the data that most non-index crimes were committed at 1 o’clock, 2 o’clock, and 8 o’clock respectively. It can also be inferred based on Figure 9 that crimes are common during Tuesdays and Sundays backed by the result that crimes committed during Tuesdays got the highest record. Friday is the safest day of the week as proven by the data that crimes are low on Fridays.

Figure 7. Time (non-index)

Figure 8. Time (index)

Figure 9. Day

4. CONCLUSION

Crimes are inevitable, they display the presence and prevalence of both index and non-index crimes and these are considered as social problems besetting the municipality of Sanchez Mira. Auto model in rapid miner with naïve Bayes classifier technique used in this study. The trend of crime data is generated straightforwardly. Otherwise, which is difficult to achieve in practice, but because of the simplicity of the naive Bayes classifier in the auto model, it is realistic. However, the auto model creates a rapid miner studio mechanism behind the scenes, which prevents researchers from knowing how the model works. The determined crimes are seasonal and could happen in a certain period within the year. There is no specific period crimes are
perpetrated in the municipality. The identified hotspots were centro 1 and centro 2 which were considered as the most densely populated barangays and these speak for criminal demography, that the higher the population in the municipality of sanchez mira, the higher the prevalence of criminality. Therefore, criminality can be equated with the population. Some interventions failed to address their purpose for the reason that there is no continuous program for interventions hence, the greater chance of reoffending others is increased. And, the data needed in this study for both index and non-index crimes are confidential. Recommendations are hereby presented based on the conclusions where coordination between the law enforcers and barangay officials is encouraged. Force multipliers may be reactivated. Crime prevention activities may be conducted including the knowledge of the law by the community are encouraged just at least to control the occurrence of criminality. A crime map may be developed using a geographic information system to be utilized by the PNP in designing the programs and updating the activities to minimize criminality in the municipality.

ACKNOWLEDGEMENTS

The authors extend their gratefulness to the Philippine national police of Sanchez mira for supporting the research, providing all required resources, and to Cagayan State University Sanchez Mira for the unending support to this study.

REFERENCES

[1] J. Peterson, I. Sommers, D. Baskin, and D. Johnson, “The Role and Impact of Forensic Evidence in the Criminal Justice Process,” Revised Final Report 6-10-10 National Institute of Justice, 2006-DN-BX-0094, 2019.

[2] V. J. Mujica, A. Choi, R. N. Leong, and F. Co, “Spatial analysis of violent crimes in Metro Manila, Philippines,” International Journal of Comparative and Applied Criminal Justice, vol. 43, no. 1, pp. 29-47, 2019, doi: 10.1080/01924036.2017.1398669.

[3] NATLEX, “Database of national labour, social security, and related human rights legislation,” Revised Penal Code of the Philippines’ International Labour Organization, p. 84, 2012.

[4] E. P. Patulin and R. E. Talingting, “Crime Trend Analysis Using Data Mining Technique,” International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 3, 2019, doi: 10.30534/jatcse/2019/52832019.

[5] A. J. P. Delima, “Applying Data Mining Techniques in Predicting Index and non-Index Crimes,” International Journal of Machine Learning and Computing, vol. 9, no. 4, 2019, doi: 10.18178/ijmlc.2019.9.4.837

[6] R. S. H. Baculino and R. Ceballos, “An analysis on the Location and Type of Index Crimes in the Philippines,” 14th National Convention on Statistics, 2019.

[7] M. J. Sanchez, “Crime rate in the Cagayan Valley Region of the Philippines from 2009 to 2014,” Statista, 2019. [Online]. Available: https://www.statista.com/statistics/1078430/philippines-crime-rate-cagayan-valley/

[8] A. Jaina and V. Bhatnagar, “Crime Data Analysis Using Pig with Hadoop,” Conference on Information Security & Priv, vol. 1, 2019.

[9] G. Dudfield, C. Angel, L. W. Sherman, and S. Torrence, “The ‘Power Curve’ of Victim Harm: Targeting the Distribution of Crime Harm Index Values Across All Victims and Repeat Victims over 1 Year,” Cambridge J. Evidence-Based Polic., vol. 1, no. 1, pp. 38-58, 2017, doi: 10.1097/s41887-017-0001-3.

[10] F. Herrera, R. Sosa, and T. Delgado, “GeoBI and Big VGI for Crime Analysis and Report,” 2015 3rd International Conference on Future Internet of Things and Cloud, 2015, pp. 481-488, doi: 10.1109/FICloud.2015.112.

[11] I. S. Makki and F. Alqurashi “An Adaptive Model for Knowledge Mining in Databases EMO_MINER for Tweets Emotions Classification,” International Journal of Advanced Trends in Computer Science and Engineering, vol. 7, no. 3, pp. 52-60, 2018, doi: 10.30534/jatcse/2018/04732018.

[12] P. Sharma, D. Singh, and A. Singh, “Classification algorithms on a large continuous random dataset using rapid miner tool,” 2015 2nd International Conference on Electronics and Communication Systems (ICECS), 2015, pp. 704-709, doi: 10.1109/ECS.2015.7125003.

[13] A. A. H. Oleiwi and A. O. Adebayo, “Data Mining Application Using Clustering Techniques (K-Means Algorithm) In the Analysis of Student’s Result,” Journal of Multidisciplinary Engineering Science Studies, vol. 5, no. 5, pp. 2857-2893, 2019.

[14] R. Kiani, S. Mahdavi, and A. Keshavarzi, “Analysis and Prediction of Crimes by Clustering and Classification,” Int. J. of Advanced Research in Artificial Intelligence, vol. 4, no. 8, 2015, doi: 10.14569/IJARAI.2015.040802.

[15] V. Nasteski, “An overview of the supervised machine learning methods,” Horizons, pp. 51-62, 2017, doi: 10.20544/HORIZONS.B.04.1.17.P05.

[16] M. Arumaiselvam and R. Anitajesi, “Study of Clustering Methods in Data Mining,” International Journal of Data Mining Techniques and Applications, vol. 7, no. 01, pp. 55-59, 2018.

[17] P. Das, A. K. Das, J. Nayak, D. Pelusi, and W. Ding, “Incremental classifier in crime prediction using bi-objective Particle Swarm Optimization,” Information Sciences, vol. 562, pp. 279-303, 2021, doi: 10.1016/j.ins.2021.02.002.

[18] M. P. Gatpandan and S. C. Ambat, “Mining crime instance records of Philippine National Police District VI Province of Cavite, Philippines: An Exploratory Study to Enhance Crime Prevention Programs,” Journal of Advanced Research in Social Sciences and Humanities, vol. 2, no. 3, pp. 176-187, 2017, doi: 10.26500/JARSSH-02-2017-0303.

Mining the crime data using naïve Bayes model (Lourdes M. Padirayon)
[19] A. Bansal, M. Sharma, and S. Goel, “Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining,” *International Journal of Computer Applications*, vol. 157, no. 6, pp. 35-40, 2017, doi: 10.5120/jica2017912719.

[20] M. Vural and M. Gök, “Criminal prediction using Naive Bayes theory,” *Neural Comput & Application*, vol. 28, pp. 2581-2592, 2017, doi: 10.1007/s00521-016-2205-z. 2017.

[21] S. R. Bandekar and C. Vijayalakshmi, “Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India,” *Procedia Computer Science*, vol. 172, pp. 122-127, 2020, doi: 10.1016/j.procs.2020.05.018.

[22] J. Arunadevi, S. Ramya, and M. R. Raja, “A study of classification algorithms using Rapidminer,” *International Journal of Pure and Applied Mathematics*, vol. 119, no. 12, pp. 15977-15988, 2018.

[23] W. Ahmed, Md. T. Nafis, and S. S. Biswas, “Performance Analysis of Naive Bayes Algorithm on Crime Data using Rapid Miner,” *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, 2017.

[24] M. Kaur and S. Kang, “Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining,” *Procedia Computer Science*, vol. 85, pp. 78-85, 2016, doi: 10.1016/j.procs.2016.05.180.

[25] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui, “A comparative evaluation of outlier detection algorithms: Experiments and analyses, Pattern Recognition,” *Pattern Recognition*, vol. 74, pp. 406-421, 2018, doi: 10.1016/j.patcog.2017.09.037.

**BIOGRAPHIES OF AUTHORS**

**Lourdes M. Padirayon**, Faculty member, College of Information and Computing Sciences. She finished her Master in Information Technology at the University of the Cordilleras, Baguio City, and currently writing her dissertation for Doctor in Information Technology at the University of the Cordilleras, Baguio City.

**Melvin S. Atayan**, Dean, College of Criminal Justice Education at the Cagayan State University Sanchez Mira. He finished a Master in Public Administration major in Police Administration at the University of Northern Philippines. At present, he is pursuing a Doctor of Philosophy in Criminal Justice with a specialization in Criminology at the University of Baguio.

**Dr. Jose Sherief O. Panelo**, Faculty member, College of Criminal Justice Education-Cagayan State University. He is a graduate of Master of Science in Criminology at the University of Cagayan Valley and Doctor of Philosophy in Criminal Justice with specialization in Criminology at the University of Cordilleras, Baguio City.

**Dr. Carlito R. Fagela, Jr**, Faculty member, College of Criminal Justice Education-Cagayan State University. He is also a graduate of Master of Science in Criminology at the University of Cagayan Valley and Doctor of Philosophy in Educational Management at Cagayan State University Sanchez Mira.