CRIME MAPPING AND PREDICTION USING MACHINE LEARNING APPROACH

[1]Quinto A. Tan Jr.,[2]Jocelyn B. Barbosa
College of Information Technology and Computing, University of Science and Technology of Southern Philippines, Republic of the Philippines
For Correspondence; Tel. +63 955 854 7882, Email: [1]quinto.tan@ustp.edu.ph, [2]jocelyn.barbosa@ustp.edu.ph

ABSTRACT: *Crime incidents, whether index or non-index may happen in any place at any time. These unlawful incidents give people a feeling of fear and uncertainty. As such, crime reporting is considered essential as it enables the monitoring of peace and order within the community. Traditionally, crime reports are sent via phone through text, or call. However, this method of crime reporting is unreliable as the salient details relevant to the crime may be inadequate or inaccurate. In this study, we present a method in crime mapping and analysis using the geo-tagged incident reports (e.g., images or videos, date, time, location, etc.) sent via mobile or web-based applications. Furthermore, we built a predictive model (i.e., using linear regression) based on the historical data that predicts crime-prone areas within the city, and consequently increase society's awareness of the likelihood of crime occurrence. Our system serves as a basis for the authorities in terms of making a decision on police visibility and which areas to prioritize in deploying electronic devices such as CCTV cameras given the limited resources. Furthermore, it paves the way for law enforcement officers to efficiently address the issue of resolving crime incidents. Experiments show that the model is efficient in terms of the predictive model's accuracy and signifies its effectiveness based on the usability and functionality tests conducted.*

Keywords: crime reports, crime mapping and prediction, crime prediction,geo-tagging, data mining, machine learning, supervised learning, linear regression

1. INTRODUCTION

The occurrence of criminal offenses has become commonplace and has resulted in a number of crimes reported on a daily basis. Crime reporting is a task that law enforcers perform to monitor peace and order within the community. This responsibility also takes much of their time and attention. In the Philippines, for example, according to PNP spokesperson Col. Ysmael Yu [1] there are 178, 751 registered incidents in the first nine (9) months of the year 2020 as shown in Table 1. Index crimes are crimes against persons and property while non-index crimes are all other illegal acts, including violations of special laws or local ordinances.

Table 1: Number of Crimes in the Philippines in 2020

INDEX	NON-INDEX
30, 775	**147, 976**
Murder	Gambling
Homicide	Liquor Laws
Physical Injury	Driving under the influence of illegal drugs, alcohol, etc.
Theft	Sex Offenses
Robbery	Forgery
Rape	Counterfeiting
Vehicle Theft	Vandalism

Traditionally, crime reporting is done manually or via phone through text or call. However, this approach may hinder informants or reporters from getting the salient and accurate details (e.g., location, incident type, date, time, etc.) as this implies subjectivity, which may, in turn, result in unreliable reports. Moreover, this may result in issues on prioritization in terms of deployment of resources given the budget constraints. These include police visibility and electronic devices deployment (e.g., police personnel, CCTV camera, etc.) in different areas in the city for crime monitoring. These may result in a delay in responses of the police officers to the reported crime incidents.

There are a number of studies relative to the mapping and monitoring of crime incidents. Some of the studies presented involved computerization of the manual process of reporting incidents utilizing a cloud server to save the data and developed a web-based application that provides a platform for the informant to report incidents to police station control and management [2, 3, 4, 5, and 6]. Shih *et a*l. [7] introduced an approach to protect the security of informants using symmetric and asymmetric keys.

Other studies [8, 9, 10, 11, 12, and 13] were conducted integrating machine learning techniques to further address the problem on crime incidents mapping and monitoring. However, most of them are focused solely on mapping of crimes and developing a predictive model and not much on the integration of the application that allows submission of objective crime reporting based on the captured scene with geotagging, which serves as inputs for a dynamic approach in building predictive models. Daglar *et. al* [8] highlights crime mapping and geographical information systems (GIS) in crime analysis that may serve as input for policy recommendations. Furthermore, several types of research were conducted on crime prediction and monitoring and used machine learning techniques (e.g., kNN, Naive Bayes, k-means clustering, SVM, DT, etc.) and visualization of crime-prone areas [8, 9, 10]. Bandekar *et al* [9] obtained an accuracy of 78%, while the study conducted in [10] got good results with the kNN classifier as claimed by the authors. On the other hand, [11] conducted a grid-based crime prediction using geographical features and tested different types of algorithms such as Random Forest, SVM, and others. Other studies [11, 13] leveraged neural networks. Chun et al. [13] investigated the possibility of using such techniques to make predictions of criminal behavior based on the history of the arrest bookings.

In this study, we build a predictive model that determines the likelihood of having a certain number of crimes in a particular area leveraging the strength of the linear regression (LR) model. LR model has also been reported to be an effective supervised-learning technique for areas where pattern analysis and prediction are necessary. We utilized the dataset extracted from the historical records in the city of Cagayan de Oro Police Office to build a model based on the geo-tagged reports submitted by the informants identified by the office. In addition, we developed a mobile-based application that allows timely and easy submission of objective crime reports based on the capture images with geographical information that is automatically retrieved from the mobile phone. The predicted values of crime incidents per barangay, do not only provide an objective basis for resource deployment prioritization given the insufficiency of resources (e.g., police hiring and deployment, electronic device / CCTV acquisition, etc.) or budget constraints for the acquisition of such resources. This can also be the basis for policy formulation such as intervention or rehabilitation programs that the local government unit may implement. Building a predictive model to detect the possible number of crimes per area or determining which area is prone to crime incidents can be utilized for planning some interventions to prevent the recurrence of crimes. In addition, it can also be used as inputs for reform and rehabilitation programs for offenders or criminals at an early phase, thereby, preventing them from committing severe crimes.

2. MATERIALS AND METHODS

Our work basically performs incident reporting and predicting of crime incidents in areas covered during the training and testing of the dataset. Information needed by police officers is available as it allows concurrent access to the central repository, thereby increasing the productivity of the organization. Figure 1 shows the system architecture of the study. Users of the system include the informant and police personnel. The informant uses the mobile-based application to send crime details directly to the server. With the integration of the Global Positioning System (GPS), the location details where the incident happens can be automatically detected or identified and sent along with the other details and images directly to the head office and routed to the assigned station.

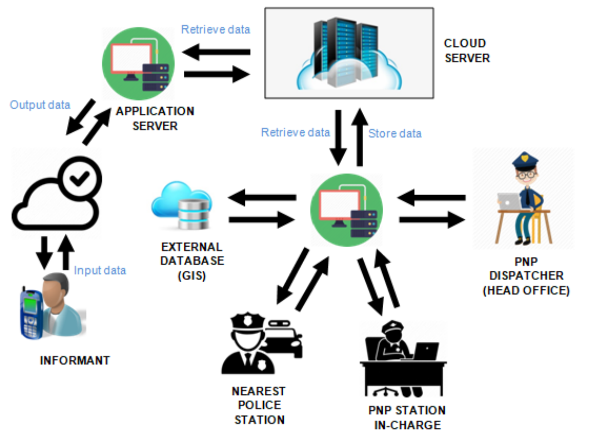


Figure 1: **System Architecture**

In addition, a web-based application was also developed for the administration and management of the system. Apart from report generation, the building of a predictive model using machine learning techniques is also considered to predict the number of future crimes per area. The system uses a cloud server as the primary repository of the system, while the web and mobile-based application have direct access to the webserver to store data and performs the query. The system also uses API's on OpenStreetMap (OSM) for the mapping and geographical information of the system.

2.1. Overview of Geotagging

Geotagging is the process of adding geographical identification of metadata to various media such as geotagged or video, websites, SMS messages, QR Codes or RSS feeds and is a form of geospatial metadata. This data usually consists of latitude and longitude coordinates, though they can also include altitude, bearing, distance, accuracy data, and place names, and a timestamp. This can help users find a wide variety of location-specific information from a device. For instance, someone can find images taken near a given location by entering latitude and longitude coordinates into a suitable image search engine. Geotaggingenabled information services can also potentially be used to find location-based news, websites, or other resources.

2.2. Context Diagram of the Proposed System

As the study includes the creation and manipulation of the database for the reported crime incidents, we create a context diagram (shown in Figure 2) that represents the bird's eye view of the whole system reflecting the entities that serve as the source and destination of the data. The four (4) entities involved include the informant, police station office, head office, and the external application. Informant sends crime information using the mobile-based application. PNP Head Office acknowledges every reported crime. Police stations may have different roles (i.e., assigned station and nearest police station). In cases like the reported incident happens to be at the border of the two barangays, it will be the nearest police station that should respond rather than the assigned one. The external application is another application being used to request some information from the database.

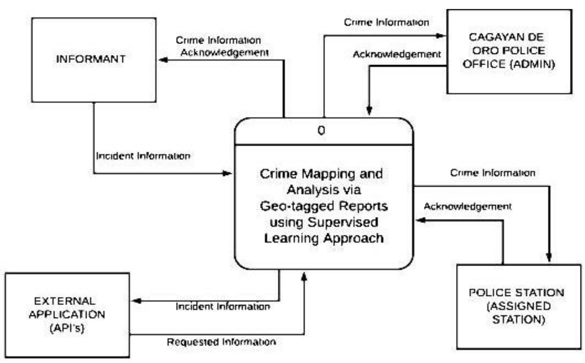


Figure 2. Context Diagram of the Study

2.3 Framework of the System

Figure 3 shows how the transactions take place, with an emphasis on the data mining process. The tool executes the program set by the system, and it will generate statistical reports and build a predictive model for finding patterns in order to predict trends of incidents in the succeeding year. With the real dataset gathered from police stations, we employ at least three machine learning techniques, which we prospect to be appropriate techniques to address this problem, as literature reveals that these models perform well. These include a) generalized linear model; b) support vector machine (SVM); and c) linear regress (LR). The models do have the capability of finding patterns, thereby roviding aid to the police officers of which barangays or communities are prone to crime incidents.

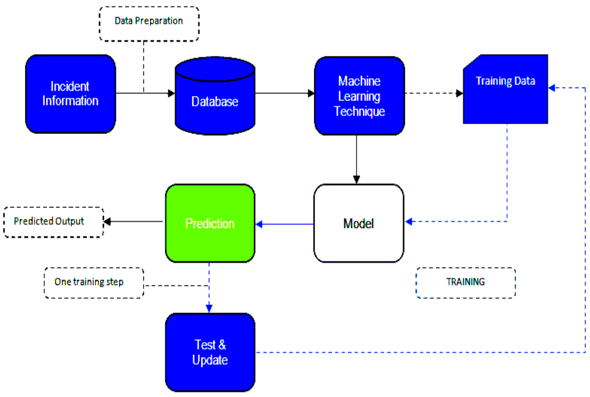


Figure 3. **The framework of the System**

The generalized linear model is an advanced statistical modeling technique, a model that encircled many other models, which allows the response variable y to have an error distribution other than a normal distribution [14]. Another model used is the support vector machine (SVM), which is a supervised learning method that can be used in both classification or regressions problems, SVM is simple and flexible for solving a range of classification problems, SVMs distinctively afford balanced predicted performance, even in studies where sample sizes is limited [15].On the other hand, Linear Regression (LR) is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope or predicting quantity[16]. It is also widely used in machine learning for linear problems. In particular, linear regression is often used for feature selection and has been shown to have good generalization performance in the presence of many irrelevant features. The equation has the form y = a + bx, where y is the dependent variable (that's the variable that goes on the y axis), x is the independent variable (i.e. it is plotted on the x axis), b is the slope of the line and a is the y-intercept.

$$\text{Where: } a = \bar{y} - b\bar{x} \qquad b = r\frac{Sy}{Sx}$$

And the correlation coefficient is

$$r = \frac{\Sigma\left((x-\bar{x})(y-\bar{y})\right)}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}}$$

In this study, y = number of incidents (dependent variable) and x = number of incidents in each barangay in a specific month and year based on the historical records (independent variable).

Since the dataset we used in the experiments is not very large, we use k-fold cross-validation to test the efficiency of the model and select the model that appears to be superior among the three techniques.

We select the best among machine learning techniques for regression, by cross-validation.

2.4. Haversine formula

The Haversine formula calculates the distance between two locations on a sphere given their coordinates (latitude and longitude). In order to get the distance of two locations on earth, you need the coordinates of point A and point B as shown in the equation below [17]. This method has been utilized in this study to get the nearest police stations from the incident location area.

*Distance, d = 3963.0 * arccos[(sin(lat1) * sin(lat2)) + cos(lat1) * cos(lat2) * cos(long2 – long1)]*

Where:

lat1 = source latitude (incident location coordinates)
long1 = source longitude (police station coordinates)
lat2 = destination latitude (incident location coordinates)
long2 = destination longitude (police station coordinates)

3. RESULTS AND DISCUSSION

In this section, we present some sample screenshots of our mobile and web-based application, which include modules for crime reporting and mapping as well the transaction where the predicted number of crimes is highlighted.

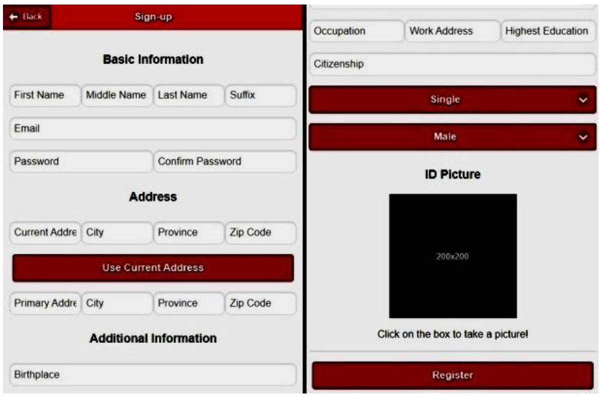
3.1 Design of the Application

A. Mobile Application

Figure 4 shows the login page of the mobile application where they can use their credentials (username and password) to log in to the system. If the informant/reporter/user has not registered yet, the "Register" button will bring them to the registration page of the application. This also shows the dashboard of the application where the user can choose from the menu. "Report Incident" will bring the user to the Crime Reporting page of the application. The "Status" button will let them view the status of the crime they reported. The "Edit Profile" button will bring the user to the editing profile page of their information. The "Other" button will show the other features of the system. The "Logout" button will sign out you from the application.

Figure 4. Mobile Login Page and Dashboard

Figure 5 shows the registration form of the mobile application. This page lets the non-registered user to sign-up before using the application. All the information gathered is based on the "Incident Form" of the police officers of the city to ensure that all the salient information is captured. On the other hand, Figure 6 shows the page where the user can send a report on crime incidents seen in a certain barangay. This information will be redirected to the assigned station. Coordinates of the location (i.e., longitude and latitude) are automatically detected when the user opens the page via GPS. In case the accuracy or location detected is poor, the "Get barangay" button will let one manually pinpoint the exact location on a map where the crime happens.



s.

Figure 5. Informants Registration Page

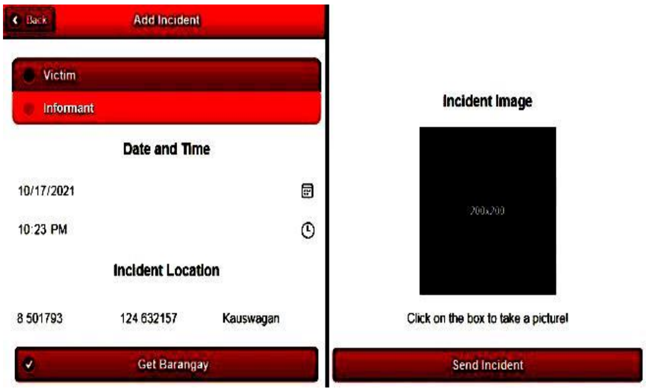


Figure 6:.Send Crime Page

B. Web Application

Figure 7 shows the login page for the admin or police stations, where the user needs to input their credentials such as username and password. It also presents the home page or dashboard of the web-based application. This page contains the following menu: Incidents, will bring you to the master list of all crimes; Personnel, will show the master list of personnel. Barangays will bring you to the master list of barangays; Station will show you the list of all police station details in Cagayan de Oro and their Areas of Responsibility (AOR), Reports will display all necessary reports provided by the system.

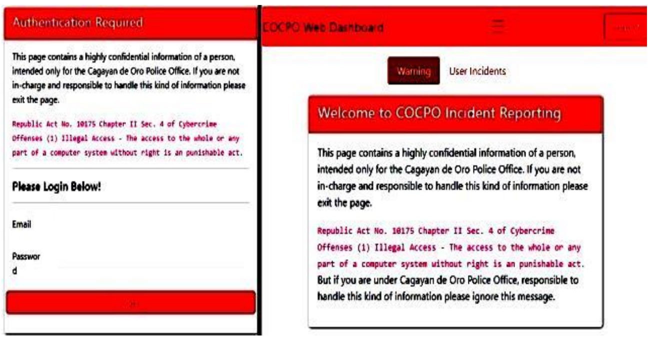


Figure 7: The login and home page of the web-app

Figure 8 shows the location of police stations in a city. Highlighted blue is the area under the Cagayan de Oro Police office. On this page, we can also view details of each police station once it is clicked. It also shows the barangays under them and plots or maps incidents reported on that certain month and year.

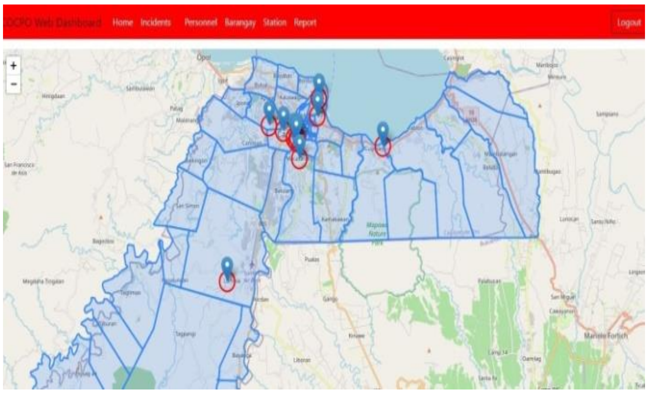
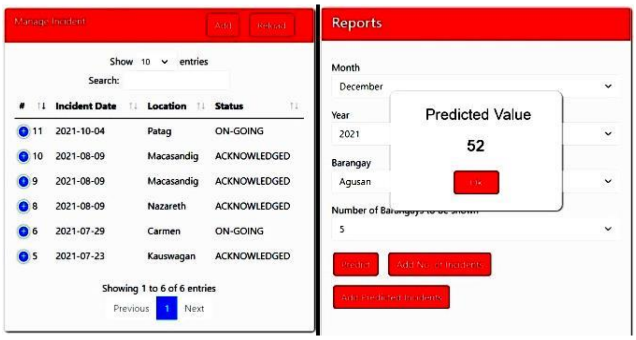


Figure 8:.Crime map based on the data collected

Figure 9 presents the newly recorded crime. It is the dispatcher/head office who controls or accepts the newly reported crime. After one clicks the "Accept" button, the crime details reported will go directly to the assigned station. The informants/reporter details together with the reported crime information can also be displayed. It also shows the prediction page where the user utilizes the option to perform a prediction for the barangay. One can choose which month, year and barangay to perform. The user can view the top 5, 10 barangays that have the highest predicted incident number for that particular month and year.



(a) (b)

Figure 9: (a) **Sample screenshots of recorded reports; and (b) Reports and Prediction Page**

3.2 System Usability Evaluation

In order to measure the usability of the proposed system, we utilize the System Usability Scale (SUS) instrument. SUS is a reliable tool for measuring the usability of software, mobile devices, websites and applications. As defined by ISO 9241-11 with references in over 1300 articles and publications, SUS provides an overall usability assessment measurement that becomes an Industry Standard.SUS is composed of some characteristics like effectiveness, efficiency and satisfaction. Effectiveness tells users how they successfully achieve their objectives and efficiency, at the same time, how much effort and resource are expended in achieving those objectives; while satisfaction determines whether the experience was satisfactory [18].

It is very important to measure the effectiveness and efficiency of the software application most especially if one is trying to optimize the User Experience (UX). SUS used a concise 10-item questionnaire that is provided at the end of a usability test to calculate a website's score. Respondents answer questions using a 5-point scale from 1 as "Strongly Disagree" to 5 as "Strongly Agree".The answers from the respondent are used to generate an extremely reliable overall usability score of your application. We deployed an online survey questionnaire following the 10 item questions of SUS. Eighteen (18) respondents were requested to answer each given question from 1 to 5 by the level of their agreement with the given statement, with 1 being lowest (Strongly Disagree) and 5 as the highest (Strongly Agree).
Figure 10 shows the Odd Items result of the respondents. The survey results reveal that the respondents find the system very usable and are likely to recommend the system.

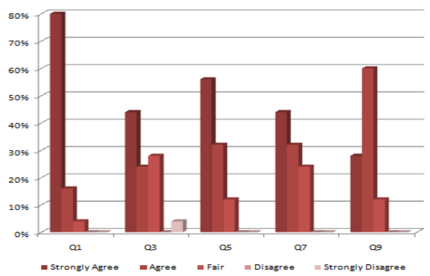


Figure 10. **System Usability Test Results (Odd Items)**

Figure 11 shows the Even item result of the respondents. The survey results demonstrate how the respondents view the effectiveness of the web and mobile-based system. Overall, the SUS final score average, survey scored 72.63. Based on the requirements set by the SUS tool, a score above 68 is being considered above average. Thus, the results demonstrate the effectiveness of the system.

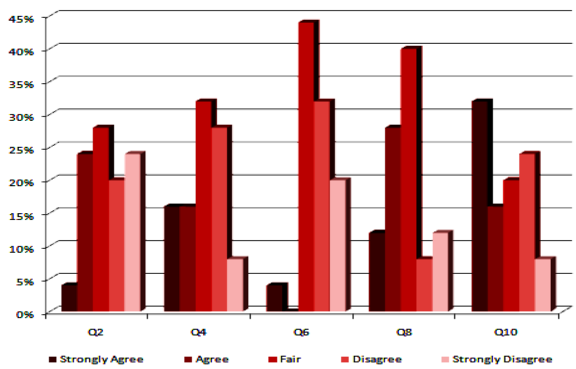


Figure 11:. **System Usability Test Results (Even Items)**

3.3. Comparison of the different Regressions

According to statistics, the average squared difference between the estimated values and the actual value is the mean squared error (MSE) or mean squared deviation (MSD) of an estimator measures the average of the squares of the errors. The MSE is a measure of the quality of an estimator (of a procedure for estimating an unobserved quantity)—it is always non-negative, and *values closer to zero are better*[19]. Figure 12 shows the absolute errors of Linear Regress (LR), Generalized Linear (GL) model and Support Vector Machine (SVM). GL Model has an absolute error of 11, SVM has 10.37, while LR has5.52 absolute error. Among these machine learning techniques used, Linear Regressionappears to be superior among the other models. As such, it is the linear regression model that we integrate into our final web-based system for prediction problems.

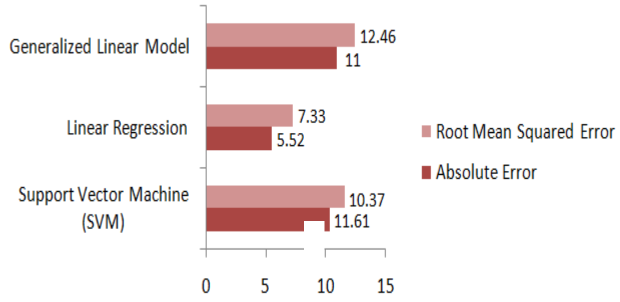


Figure 12: **Absolute Error & Root Mean Squared Error Comparison to other models**

4. CONCLUSION AND RECOMMENDATION

A. CONCLUSION

In this study, we present a method for crime mapping and monitoring using mobile and web-based platforms leveraging the strength of geo-tagging. The geotagged report submitted by the informants was used as inputs for efficient crime mapping and monitoring as well as in building a model for crime prediction. Specifically, we introduce a model that predicts the number of crime incidents in a particular area using the linear regression (LR) model. The predicted values of crime incidents per barangay can provide an objective basis for resource deployment prioritization given the insufficiency of resources (e.g., police hiring and deployment, electronic device / CCTV acquisition, etc.) or budget constraints for the acquisition of such resources. Furthermore, our system can also serve as the basis for the authorities for policy formulation such as intervention or rehabilitation programs that the local government unit may implement.

B. RECOMMENDATION

Although our approach for crime mapping and prediction works well in the given dataset, they may be some areas that need to be considered so as to improve the performance of the system. These include, (a) Data synchronization on sending a report when there is poor to no internet connectivity; and (b) Revisiting and addressing of imbalanced dataset.

REFERENCES

[1] Tupas, E. February 26, 2019. PNP: Total crime volume down in 2018, PhilstarGlobal. Retrieved from "https://www.philstar.com/nation/2019/02/26/1896714/pnp-total-crime-volume-down-2018"

[2] Khan, A., Singh A., Chauhan Ankit, Aiswarya G. April 2019. "Crime Management System", International Research Journal of Engineering and Technology (IRJET), **Vol. : 06.** Issue: 04,

[3] Yugandhar, P., Muni Archana, B.. April 2018. "Crime Reporting System". INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY". **Vol. :04** Issue: 11

[4] Bukohwo, E. M., Michael, A. (n.d.). “The Design of Real Time Crime Reporting System”. American International Journal of Research in Science, Technology, Engineering & Mathematics.

[5] GodlinJesil, S. P., RajatBasant, Pratishvir. (n.d.). “CRIME REPORTING SYSTEM USING ANDROID APPLICATION”. International Journal of Pure and Applied Mathematics, Special Issue.

[6] Jimoh, R. G., Ojulari, K.T., Enikuomehin, O.A. (January 2014). “A Scalable Online Crime Reporting System”, African Journal of Computing & ICT

[7] Shih T F, Chen CL, Syu B Y, Deng Y Y. February 2019. “A Cloud-Based Crime Reporting System with Identity Protection”. Symmetry

[8] Daglar, M, UgurArgun. (2016). “Crime Mapping and Geographical Information Systems in Crime Analysis”. International Journal of Human Sciences**, Vol. :13** Issue: 1 Year

[9] Bandekar, S. R., Vijayalakshmi, C. 2019. “Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India”. The 9th World Engineering Education Forum (WEEF – 20)

[10] Bharati, A., Sarvanaguru, R.A. September 2018. “Crime Prediction and Analysis Using Machine Learning”. International Research Journal of Engineering and Technology (IRJET).

[11] Lin,Y. Y, Yen, M. F., Yu, L. C. July 2018. “Grid-Based Crime Prediction Using Geographical Features”. International Journal of Geo-Information.

[12] Pacot, M. P. January 2019. “Mobile Based Crime Mapping and Event Geography Analysis for CARAGA. Retrieved from: “https://www.researchgate.net/publication/330370752_MobileBased_Crime_Mapping_and_Event_Geography_Analysis_for_CARAGA”

[13] Chun, S. A. 2019. “Crime Prediction Model using Deep Neural Networks”

[14] Ram, P. April 27, 2021. “Generalized Linear Models | What does it mean?”. Retreived from “https://www.mygreatlearning.com/blog/generalized-linear-models/”

[15] Pisner, D.A, Schnyner D.M. 2020. “Support Vector Machine”. Retrieved from “https://www.sciencedirect.com/topics/neuroscience/support-vector-machine”

[16] Wikepedia. Simple Linear Regression. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Simple_linear_regression

[17] GeeksforGeeks. April 2021. Program for distance between two points on earth. Retrieved from: “https://www.geeksforgeeks.org/program-distance-two-points-earth/”

[18] System Usability Scale (SUS). (n.d.). Retrieved from:“https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html”

[19] Wikepedia. (n.d.). Mean Squared Error. Retrieved from: “https://en.wikipedia.org/wiki/Mean_squared_error#cite_note-:0-1”