International Journal of Science, Technology, Engineering and Mathematics
Volume 2 Issue 1 March 2022
DOI: https://doi.org/10.53378/352875

Building Model for Crime Pattern Analysis Through Machine Learning Using Predictive Analytics

[1]Jonard R. Asor, [2]Francis F. Balahadia, [1]Gene Marck B. Catedrilla & [3]Mia V. Villarica

Abstract

Crime has a big impact in both the human lives and the society's growth, which needs to be addressed and controlled. Machine learning algorithms as the fanciest technology to assist decision makers in policy making has proven its reliability in showing unseen patterns in crime. This research aims to examine the capability of trees and ensemble trees in classifying crime through model development. Experiments were done to enhance the capability of the ensembles in both classification and regression. Feature extraction like synthetic minority oversampling technique was applied in order to address the problem in the imbalanced data. Different metrics relevant to classification and regression were considered in evaluating the performance of each model used. With the use of different metrics, Gradient boosted tree was found to have better classification capability in crime dataset after outperforming decision tree and random forest in both classification and regression problem. Furthermore, random forest was also found to have a promising capability in classification by regression. Therefore, it is highly recommended that this ensemble algorithm be further examined and considered in developing model in other datasets.

Keywords: *Crime incidents, crime report, crime patterns, Laguna, Decision Tree algorithm, KDD*

Received: January 13, 2022 **Revised:** February 4, 2022 **Accepted:** March 6, 2022

Suggested Citation: Asor, J.R., Balahadia, F.F., Catedrilla, G.B. & Villarica, M.V. (2022). Building Model for Crime Pattern Analysis Through Machine Learning Using Predictive Analytics. *International Journal of Science, Technology, Engineering and Mathematics*, Volume 2 Issue 1, pp. 61- 73. DOI: https://doi.org/10.53378/352875

About the authors:

[1]Instructor I, Laguna State Polytechnic University, Los Baños, Laguna, Philippines

[2]Assistant Professor II, Laguna State Polytechnic University, Siniloan, Laguna, Philippines

[3]Assistant Professor I, Laguna State Polytechnic University, Sta. Cruz, Laguna, Philippines

1. Introduction

Crime is a severe problem of any country that needs to be addressed and controlled by the community and the world itself, for it affects not only the people, but the community's growth as well (Almaw & Kadam, 2018). To address the social dilemma, technology integration is one of the most effective and efficient tools in supporting social peace and order. For instance, the introduction of machine learning algorithm in crime analysis and prediction brings a whole new perspective in one's security agency. This intelligent approach in crime eradication outstands the human approach in analysis by mimicking the human concept (Shah et al, 2021). The usage of machine learning algorithms can be helpful to the crime analysts to their fight against crimes and in saving humanity (ToppiReddy et al., 2018).

Predictive analytics is a branch of data analysis used for the advantage of policy maker in empowering their decision-making. This approach becomes prominent and auspicious in crime analysis through the use of machine learning algorithm (Ippolito & Lozano, 2020). Machine learning algorithms such as k-NN, naïve bayes and decision tree were used in classifying crime in a small amount of data (Wibowo & Oesman, 2019). In an experiment by Iqbal et al. (2013), it was found that decision tree was a better performing machine learning compare to naïve bayes after gaining a much higher accuracy, precision and recall. It was proven to give a high accuracy, which can still be improved by integrating ensemble methods or application of different feature selection (Aldossari, et al., 2020). Similarly, the decision tree was found as reliable predicting algorithm when integrated in computer systems (Ahishakiye et al., 2017). Furthermore, other classification techniques were used to improve the performance of decision tree like regression (Sapin et al., 2021).

Regression machine learning algorithms have been used to predict crime (Ajagbe et al., 2020). Linear regression is found to be more effective in terms of handling the randomness of test samples than decision trees. Further, the precision of machine learning in predicting crime to slow down crime occurrences is well worth through knowledge discovery (McClendon & Meghanathan, 2015). Random forest as regressor show a promising result compared to other regression algorithms with equal hyperparameters in crime prediction including linear regression (Kadar et al., 2016). Moreover, the performance of each machine learning algorithms depends on the amount of data. It was already proven that the size of the dataset to be used in model development using machine learning really matters (Althnian et al., 2021).

On the case of limited features for prediction, gradient boosting found to be more effective based on the accuracy rather than other machine learning that uses regression (Lamari et al., 2020). In the issue of imbalanced data, machine learning algorithms such as support vector machine (SVM), neural networks and ensemble algorithms like random forest and gradient boosting were found to perform well in terms of prediction (Nguyen et al., 2017). A crime analysis was conducted using boosted decision tree and k-NN and proven that gradient boosted decision tree is more effective than the other algorithms (Kim et al., 2018).

This research work aims to develop a model for crime prediction through ensemble and trees machine learning algorithm in crime dataset. An experiment to enhance the performance of trees and ensembles is done to assure its viability once integrated as model for intelligent system.

2. Methodology

2.1. Data Collection

Table 1 presents the dataset that contains the attributes and description of the crime records of the Philippine National Police (PNP) in Laguna.

Table 1

Crime records of PNP – Laguna

Attribute	Description	Data Type
Date	The date when the crime was committed.	Date
Time	Exact time when the crime was committed.	Time
Address	Location where the crime happened.	String
Violation	Type of crime that is being committed.	String
V_Sex	Sex of the victim.	String
V_Age	Age of the victim.	Numerical
V_Nationality	Nationality of the victim.	String
S_Use	Weapon or device used by the suspect to commit the crime.	String
S_Sex	Sex of the suspect.	String
S_Age	Age of the suspect.	Numerical
S_Nationality	Nationality of the suspect.	String
Action_Taken	Appropriate legal action that the PNP was conducted.	String
Status	The status of the case whether it was filed or not.	String
Remarks	The status of the filing of the case in the court level.	String

Crime records of the PNP in Laguna, Philippines are used as dataset in this study. It contains different attributes that may or may not contribute to the performance of the model once

developed. These attributes were the so called ‘features’ which play the most vital and crucial part in model development (Barnadas, 2016).

2.2. Data Pre-processing

The dataset undergone different processing to assure the reliability and its efficacy when used for model development. Since the dataset contains some duplicated value in the ‘violation,’ which is used as the label or class to be predicted, a modification was done with the help of a criminologist. Civil code inputted in the dataset was converted into the actual crime that is committed, rather than keeping the instances with ‘RA’ or crime code, it was classified by the criminologist and converted into the actual crime like theft, physical injury, among others. Upon finishing this process, each line with missing ‘violation’ value was removed while those lines with ‘violation’ value, but missing other values were kept and still considered in the model development procedure. Further, all the remaining data in the dataset were converted into lowercase to assure that there is no bias or noise in terms of instance meaning. Conversion of the dataset into numerical form was also done for classification by regression procedure. Lastly, some attributes are dropped to enhance the performance and run-time in classification of every classifier that were used.

Table 2

Preprocessed Crime records of PNP – Laguna

Attribute	Description	Data Type
Date	The date when the crime was committed.	Date
Time	Exact time when the crime was committed.	Time
Address	Location where the crime happened.	String
Violation	Type of crime that is being committed.	String
V_Sex	Sex of the victim.	String
V_Age	Age of the victim.	Numerical
S_Sex	Sex of the suspect.	String
S_Age	Age of the suspect.	Numerical
Action_Taken	Appropriate legal action that the PNP was conducted.	String

2.3. Model Development

Figure 1

Model Development Process

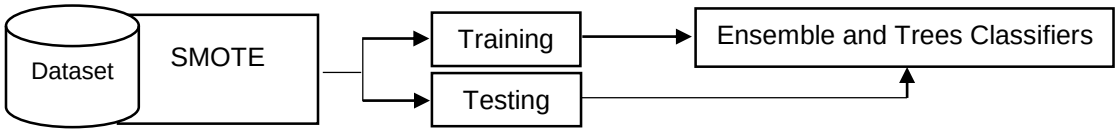


Figure 1 represents the procedure followed in this study to develop a model using ensemble and trees classifiers. In this study, the dataset was found to have an imbalanced data. Imbalanced data has a high impact in deteriorating the performance of even the most prestige machine learning. It drastically decreases the reliability of a particular model (Somasundaram & Reddy, 2016). To address this problem, synthetic minority oversampling technique (SMOTE) was applied. SMOTE was found to be effective in creating new and reliable data which can be used for balancing the dataset by oversampling the minority in a dataset (Peng et al., 2019; Sapin et al., 2021). Moreover, the dataset provided was separated into two (2) parts, the training dataset and the testing dataset. This step was done for validating the performance of the trained model, wherein the 20% of the dataset is used as testing dataset while the remaining 80% is the training dataset (Birba, 2020).

Three different algorithms concerning trees were used to develop the model. Decision tree (DT), random forest (RF) and gradient boosted trees (GBT) were found to be a promising algorithm in developing models for crime dataset. As discussed by many researchers, a tree learned by splitting the source into subsets based on the independent variables or the attributes inside the dataset. The result of these splits is one of the most understandable machine learning approaches for human interpretation (Singh & Pal, 2020). On the other hand, ensemble tree is the concept of bagging and boosting a tree. Bagging creates an ensemble of trees to create new training set from the actual training set, this new training sets were called bags (Banfield et al., 2007; Nagpal, 2017; Singh & Pal, 2020).

In the decision tree, the terminal nodes are considered as the class/label attribute which is known as dependent variable or the one to predict represented by Y in mathematical expression. Whereas, the non-terminal nodes, including internal and root nodes are the independent variables

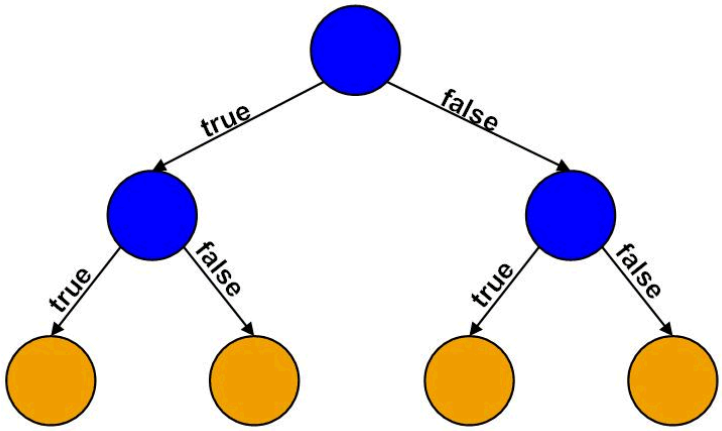
or the attributes to be split to have a pattern and new knowledge, commonly represented by *x* in mathematical representation. In mathematical formulation, decision tree has the following form:

$$3. \quad (x, Y) = (x_{1,} x_{2,} x_{3,} \dots \dots \dots . x_n, Y) \tag{1}$$

The external node or dependent variable Y is the target value to be classified or predict. Whilst, the independent variable or vector *x* will be the input variable, $x_{1,} x_{2,} x_{3,}$ which will be used to do the task. This approach will develop a tree that can be interpreted by *if-else* approach to understand the pattern that is being considered by the decision tree. Figure 2 portrays the architecture of a decision tree:

Figure 2

Decision tree architecture

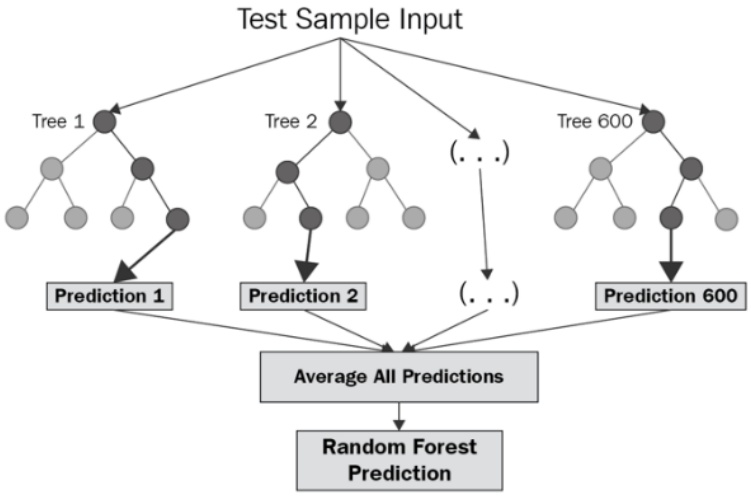


Source: https://tinyurl.com/fnp33822

Random forest is basically a collection of trees that works together in finding the most relevant independent variable in classifying the dependent variable through mode. This machine learning algorithm consists of numerous structured-trees $\{h(x, \Theta_k), k = 1 \dots\}$ where the $\{\Theta_k\}$ are the independent random but identical vectors distributed to cast votes in finding the most common class for each *x*. Figure 3 is the representation of how random forest finalized its prediction based on the result of the votes of every tree inside it.

Figure 3

Random Forest architecture

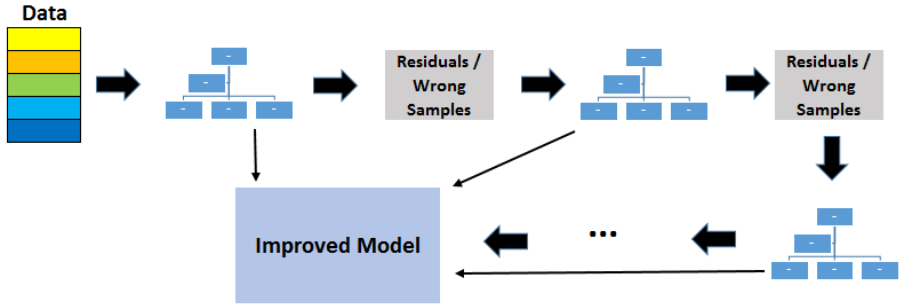


Source: https://tinyurl.com/2fu4ws5a

Gradient boosting is the process of buffering weak to enhance their performance. In this work, gradient boosted tree was used to further enhance the performance of decision trees. Figure 4 represents the top-view of the gradient boosted tree:

Figure 4

Gradient boosted tree ensemble architecture



Source: https://tinyurl.com/ukcd5p48

In the finalizing and assuring the reliability and avoid biases in the result, the experiment was conducted. Similarities of hyperparameter were observed considering their applicability in each machine learning algorithm used. Aside from k-fold validation, shown in table 2 is the hyperparameters used in this study.

Table 2

Hyperparameters used in the first experiment

Hyperparameter	Decision Tree	Random Forest	Gradient Boosted Tree
Pruning	True	True	n/a
Pre-pruning	True	True	n/a
Reproducible	n/a	n/a	True
Minimal Gain	0	0	n/a
Number of trees	n/a	50/100/200	50/100/200
Learning rate	n/a	n/a	0.01

Pruning and pre-pruning is applied in both DT and RF to assure that no overfitting happens in the model development, thus, it is not appropriate or not a parameter in GBT. Nevertheless, reproducible is an alternative to pruning that is only done in GBT. Minimal gain is a parameter for calculating node splitting, a higher value of its result to a not split at all. Number of trees are only applicable to ensemble algorithms which is used to determine how many trees to generate in the ensemble. In this paper, three values for the number of trees (50, 100 & 200) were used to see the veracity of ensembles in terms of repetition. The learning rate is the capability of the model to interrelate each vector in the dataset for classification, hence, it is better to be in a lower value to assure the reliability of the model's performance and to address the overfitting issue.

Moreover, classification by regression was also done. Since all the machine learning algorithms stated were designed to solve problems through the nominal label, a conversion to classification into regression was done to have another substantial result in performance enhancement of tree and ensembles. This process is a nested operator that creates a subprocess which generates classification model through regression learning.

2.4. Validate Model

As stated in model development phase, the dataset was divided into two for validation–the training and testing dataset. With this, necessary validation metrics are produced in order to have a reliable validation of the performance of each machine learning algorithm. Moreover, together with k-fold validation other necessary metrics were calculated in both classification and regression techniques. For classification, metrics such as accuracy, kappa, recall, precision, specificity, false

positive rate (FPR), false negative rate (FNR), Matthew's coefficient correlation (MCC) and f-score were used to evaluate the algorithms' performance while root mean square error (RMSE), mean square error (MSE) and R-square were used for evaluating regression technique. Both the metrics for classifications and regression were used to understand how accurate the classifiers or algorithms are, by examining the numbers of correct and incorrect classified data.

3. Results and Discussion

Using the hyperparameters in table 2, results of the experiments are shown in this section. As shown in table 3, GBT got the highest accuracy with a total score of 80.43%, thus it also outperformed the DT and RF in different viability metrics. Further, the RF outperformed the GBT in regression problem in terms of R^2 evaluation with 0.034 difference, however, it is noticeable that GBT has a better classification capability after getting a better RMSE and MSE score compare to the other two algorithms.

Table 3

Classifier performance with default metrics

	Classification									Regressor		
Classifier	Accuracy	Kappa	Recall	Precision	Specificity	FPR	FNR	MCC	Fscore	RMSE	R^2	MSE
DT	48.07	0.423	0.48	0.5	0.90	0.1	0.52	0.37	0.43	1548.00	0.442	2154.994
GBT	80.43	0.783	0.8	0.8	0.97	0.03	0.2	0.77	0.79	1353.00	0.47	2014.696
RF	54.73	0.497	0.55	0.63	0.92	0.08	0.45	0.48	0.5	1509.00	0.504	2127.675

After the first experiment, it is shown that ensemble trees are better at classifying crime than the decision tree itself. Hence, another experiment was done to further evaluate the performance of the two highest performing ensembles where their common hyperparameters were constantly modified. Reflected in table 4 are the performance of the GBT and RF in both classification and regression with a hundred trees as main parameter. Still, GBT has outperformed RF in classification with 82.72.% vs 54.5% differences in accuracy. Nevertheless, both GBT and RF have a total R^2 score of 0.53 for regression, hence, GBT outperformed RF after having a better RMSE and MSE score. These show that GBT has a better performance than RF in both classification and regression problem when there are more trees.

Table 4

Ensemble's performance with 100 trees

	Classification									Regressor		
Classifier	Accuracy	Kappa	Recall	Precision	Specificity	FPR	FNR	MCC	Fscore	RMSE	R^2	MSE
GBT	82.72	0.803	0.82	0.81	0.98	0.02	0.18	0.79	0.81	1479.00	0.53	2106.419
RF	54.5	0.494	0.55	0.63	0.92	0.08	0.46	0.47	0.49	1488.00	0.53	2112.818

Table 5 shows the result of the last experiment done in this study. For this experiment, the number of trees was raised to 200. It is noticeable that GBT still outperforming the RF in classification problem after acquiring a total accuracy of 83.03% and a Fscore of 82%. Also in regression, GBT still shows a promising classification capability after outperforming RF in all the regression metrics used in this experiment.

Table 5

Ensemble's performance with 200 trees

	Classification									Regressor		
Classifier	Accuracy	Kappa	Recall	Precision	Specificity	FPR	FNR	MCC	Fscore	RMSE	R^2	MSE
GBT	83.03	0.811	0.83	0.82	0.98	0.02	0.17	0.8	0.82	1266.00	0.54	1948.846
RF	54.73	0.497	0.55	0.63	0.92	0.08	0.45	0.48	0.5	1482.00	0.50	2108.554

4. Conclusion

In this research paper, decision trees and ensemble trees were used to develop a model in crime reports in the province of Laguna, Philippines. Synthetic minority oversampling technique (SMOTE) is used to address the issue of imbalanced data. It was found that ensembles like random forest and gradient boosted tree were better in classifying law violation or crime than decision tree. Furthermore, after the experiments, it was found that gradient boosted tree was more effective in both classification and regression than the random forest especially when the number of trees were more than a hundred. Nonetheless, the random forest shows a promising capability in regression after outperforming the gradient boosted tree in the first experiment where it gains a higher R^2 score.

For future studies, the result of this research paper can be a model for intelligent system in predicting crime for a strategic planning basis for the PNP in Laguna. Random forest shows an

unexpected result during classification by a regression process which implicates for a better classification capability, therefore, a study about random forest enhancement must be considered. In this study, a validation for current crime occurrence was not done yet, hence, it is recommended that the model developed must be test in current crime record to have clearer findings in terms of variance accurateness. Moreover, similar study must be done in a larger data set.

References

Ahishakiye, E., Omulo, E. O., Taremwa, D., & Niyonzima, I. (2017). Crime Prediction Using Decision Tree (J48) Classification Algorithm. *International Journal of Computer and Information Technology*, 188-195.

Ajagbe, S. A., Idowu, I. R., Oladosu, J. B., & Adesina, A. O. (2020). Accuracy of Machine Learning Models for Mortality Rate Prediction in a Crime Dataset. *International Journal of Information Processing and Communication, 10*(1&2), 150-160.

Aldossari, B. S., Alqahtani, F., Alshahrani, N. S., Alhammam, M. M., Alzamanan, R. M., Aslam, N., & Irfanullah. (2020). A Comparative Study of Decision Tree and Naive Bayes Machine Learning Model for Crime Category Prediction in Chicago. *2020 The 6th International Conference on Computing and Data Engineering* (pp. 34-38). Senya: ACM. doi:10.1145/3379247.3379279

Almaw, A., & Kadam, K. (2018). Survey Paper on Crime Prediction using Ensemble Approach. *International Journal of Pure and Applied Mathematics, 118*(8), 133-139. Retrieved from https://www.acadpubl.eu/jsi/2018-118-7-9/articles/8/18.pdf

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. B., Alzakari, N., . . . Kurdi, H. (2021). Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Applied Science, 11*(2), 1-18. doi:10.3390/app11020796

Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2007). A Comparison of Decision Tree Ensemble Creation Techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 173-180. doi:10.1109/tpami.2007.250609.

Barnadas, M. V. (2016, September 1). Machine Learning Applied to Crime Prediction. Barcelona.

Birba, D. E. (2020). A Comparative study of data splitting algorithms for machine learning model selection. *Degree Project in Computer Science and Engineering*.

Ippolito, A., & Lozano, A. C. (2020). Tax Crime Prediction with Machine Learning: A Case Study in the Municipality of Sao Paulo. *22nd International Conference on Enterprise Information*

Systems. 1, pp. 452-459. Czech Republic: Science and Technology Publications. doi:10.5220/0009564704520459

Iqbal, R., Panahy, P. H., Murad, M. A., Mustapha, A., & Khanahmadliravi, N. (2013). An Experimental Study of Classification Algorithms for Crime Prediction. *Indian Journal of Science and Technology, 6*(3), 4219-4225.

Kadar, C., Iria, J., & Cvijikj, I. P. (2016). Exploring Foursquare-derived features for crime prediction in New York City. *KDD - Urban Computing WS '16.* San Francisco: ACM. doi:10.1145/1235

Kim, S., Joshi, P., Kalsi, P. S., & Taheri, P. (2018). Crime Analysis Through Machine Learning. *9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).* Vancouver: IEEE. doi:10.1109/IEMCON.2018.8614828

Lamari, Y., Freskura, B., Abdessamad, A., Eichberg, S., & Bonviller, S. d. (2020). Predicting Spatial Crime Occurrences through an Efficient Ensemble-Learning Model. *International Journal of Geo-Information, 9*(645), 1-20. doi:10.3390/ijgi9110645

McClendon, L., & Meghanathan, N. (2015). Using Machine Learning Algorithms to Analyze Crime Data. *Machine Learning and Applications: An International Journal, 2*(1), 1-12. doi:10.5121/mlaij.2015.2101

Nagpal, A. (2017, October 18). *Decision Tree Ensembles- Bagging and Boosting*. Retrieved from towardsdatascience: https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting

Nguyen, T. T., Hatua, A., & Sung, A. H. (2017). Building a Learning Machine Classifier with Inadequate Data for Crime Prediction. *Journal of Advances in Information Technology, 8*(2), 141-147. doi:10.12720/jait.8.2.141-147

Peng, M., Zhang, Q., Xing, X., Gui, T., Huang, X., Jiang, Y.-G., . . . Chen, Z. (2019). Trainable Undersampling for Class-Imbalance Learning. *The Thirty-Third AAAI Conference on Artificial Intelligence* (pp. 4707-4714). Hawaii: AAAI.

Sapin, S. B., Lerios, J. L., Padallan, J. O., Buama, C. A., & Asor, J. R. (2021). Fire incidents visualization and pattern recognition using machine learning algorithms. *Indonesian Journal of Electrical Engineering and Computer Science, 22*(3), 1427-1435. doi:10.11591/ijeecs.v22.i3

Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art, 4*(9), 1-14. doi:10.1186/s42492-021-00075-z

Singh, R., & Pal, S. (2020). Machine Learning Algorithms and Ensemble Technique to Improve Prediction of Students Performance. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(3), 3970-3976. doi:10.30534//ijatcse/2020/221932020

Somasundaram, A., & Reddy, U. S. (2016). Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data . *1st International Conference on Research in Engineering, Computers and Technology* (pp. 28-34). Peru: IEEE.

ToppiReddy, H. K., Saini, B., & Mahajan, G. (2018). Crime Prediction & Monitoring Framework Based on Spatial Analysis. *Internation Conference in Computational Intelligence and Data Science* (pp. 696-705). Ohio: Elsevier. doi:10.1016/j.procs.2018.05.075

Wibowo, A. H., & Oesman, T. I. (2019). The comparative analysis on the accuracy of k-NN, Naive Bayes, and Decision Tree Algorithms in predicting crimes and criminal actions in Sleman Regency. *iCAST-ES 2019. 1450*, pp. 1-6. Bali: Journal of Physics: Conference Series. doi:10.1088/1742-6596/1450/1/012076