Comparing Regression Models to Predict Property Crime in High-Risk Lima Districts

Maria Escobedo[1], Cynthia Tapia[2], Juan Gutierrez[3], Victor Ayma[4]
Student Member, Universidad de Lima, Lima, Peru[1, 2]
Professor Member, Universidad de Lima, Lima, Peru[3]
Professor Member, Universidad del Pacifico, Lima, Peru[4]

***Abstract*—Crime continues to be an issue, in Metropolitan Lima, Peru affecting society. Our focus is on property crimes. We recognized the lack of studies on predicting these crimes. To tackle this problem, we used regression techniques such as XGBoost, Extra Tree, Support Vector, Bagging, Random Forest and AdaBoost. Through GridsearchCV we optimized hyperparameters to enhance our research findings. The results showed that Extra Tree Regression stood out as the model with an R2 value of 0.79. Additionally, error metrics like MSE (185.43) RMSE (13.62) and MAE (10.47) were considered to evaluate the model's performance. Our approach considers time patterns in crime incidents. Contributes, to addressing the issue of insecurity in a meaningful way.**

Keywords—Supervised techniques; machine learning; regression; crime; prediction

I. Introduction

Numerous sources concur that crimes are a constantly growing phenomenon, which detrimentally affects the economy and the overall quality of life [1, 2, 3]. Crimes are prevalent in every social system, with their impact experienced across different continents, countries, and regions.

Numbeo [4] conducted a global ranking, assessing crime rates in 142 countries, where Peru was placed eleventh with a crime rate of 68% and a safety rate of 32%. One of the primary issues in Peru is crime, with reports of crimes seeing a steady increase.

In 2022, the department of Lima recorded the highest number of complaints for the commission of crimes, accounting for 34% of all complaints. Many of these crimes occurred within the Lima metropolitan area, where 44,879 crimes were reported during the first half of the year.

Most of the crimes accounting for 74% of the count were related to property offenses. These offenses involve actions that violate the belongings or possessions of individuals or businesses. When we refer to property in this context we mean any item, with worth. These crimes can be further classified into types based on their nature such as assault involving vehicles, theft severe forms of theft like nighttime burglary and burglaries in occupied houses unauthorized use of property, vehicle theft, attempted thefts, robbery cases armed robbery cases, with aggravated circumstances involved, gang related robberies and attempted robberies [5].

The absence of a higher level of education and the presence of job instability is crucial factors that have a negative impact on both the crime rate and the country’s economy. Regarding the profile of the detainees, 90% of them are men, aged between 18 and 59 years, with a basic educational background, and facing unstable employment conditions [6].

In recent years, there has been an alarming increase, in property crimes in the high-risk districts of Metropolitan Lima. This has caused concerns for the safety and well-being of residents [7]. According to INEI [8] these crimes have reaching effects beyond the immediate victims. They also affect the economy and societal trust in the areas. With a focus, on measures this research aims to use regression models as a predictive tool to aid law enforcement agencies in making informed decisions and effectively addressing the urgent problem of property crimes.

The motivation we use regression models is because they can analyze the relationship, between temporal factors that affect crime rates. These models use a time window structure. Include variables through data organization and feature selection techniques providing a systematic approach to understanding and predicting crime patterns. Our main goal is not to make predictions but also to give law enforcement agencies valuable insights that can help them distribute resources and plan intervention strategies effectively.

It is important to recognize that traditional methods of addressing citizen insecurity in Peru have faced challenges when it comes to integrating with life and promoting collaboration [9]. Using regression models our aim is to bridge this gap by developing a solution that does not consider the experiences of residents but also encourages their active participation, in improving community safety.

After conducting a comparison of regression models such, as XGBoost, Extra Tree, Support Vector, Bagging, Random Forest and AdaBoost this study aims to determine the most effective model based on key evaluation metrics like Mean Absolute Error (MAE) Mean Squared Error (MSE) Root Mean Squared Error (RMSE) and R squared. The chosen best model will be a resource for law enforcement in their efforts to reduce property crimes and create a safer environment for the residents of Metropolitan Lima.

The research is divided into sections. Section II will analyze the state of research in crime prediction. In Section III we will provide a description of the techniques we will use. Section IV will outline the steps and experimental processes undertaken in this study. After explaining our method in detail, we will present the results in Section V. Finally, we will offer a discussion and conclusion in Section VI and Section VII respectively.

II. Related Work

Many studies have explored the domain of crime management, suggesting various methods employing machine learning algorithms for forecasting. Highlighting the crucial role of characteristic choice and data origins, a huge part of this investigation has been centered in nations with elevated crime rates like India, Bangladesh [10, 11, 12, 13], Brazil [14], and Colombia [15], with specific attention on urban areas. Curiously, regions in Asian countries like China display decreased crime rates, molding the outlook of its inhabitants [16]. It is vital to recognize that the forecasting effectiveness of crime models is impacted by the location setting.

Various regression techniques have been employed in these studies, including Random Forest, Support Vector Machine, KNN Regression, LASSO Regression, Ridge Regression, Linear Regression, Polynomial Regression, Gradient Boosting Regression, AdaBoost R2, Additive Regression, Extreme Gradient Boosting Regression, Bagging, Iterated Bagging and Multivariate Adaptive Regression Splines [11, 13, 15, 16, 17 20]. It is, worth noting that Random Forest consistently delivered results with an adjusted R2 of 80% [14] which explains its widespread adoption, in numerous research studies [11, 13, 15, 17, 18, 20, 21].

The collection of crime records relies on data sources from police departments and governmental agencies of the respective countries [11, 12, 13, 19]. Time and location-related variables appear as the primary features, with time variables including year, month, day, and hour [19, 20]. The time window method plays a crucial role in predicting crime rates over short (monthly) and long (annual) data periods, revealing significant variations [23].

Concerning location variables, latitude, longitude, and city are commonly used [23, 20, 24]. More specific variables, such as cab flow [16], gross domestic product, household income, unemployment [14], and socio-economic indicators [21], are explored in some studies but are not as often employed in crime prediction models.

The output variable in these studies varies, aiming to predict crime rates at different geographical levels—countries [13], cities [19], regions [10], municipalities [15], and neighborhoods [17]. Furthermore, certain research studies are dedicated to forecasting crime rates based on their categories [11, 12 17, 23 24, 25] while others aim to enhance the distribution of law enforcement personnel, in localities [19].

The results, from studies show R2 values. The Linear Regression model had the R2 of 0.99 when socioeconomic data was used [20]. It is worth mentioning that the Random Forest Regression model consistently yielded R2 values, between 0.77 and 0.98 because it can handle both categorical data [13, 14, 15, 19, 21, 22].

In recent times, research on new approaches to improve crime prediction models has been very active. For example, Briz developed a spatial-temporal logistic regression model to address the complex challenge of temporal uncertainty in crime data analysis. Then, they decided to advance further by developing a Bayesian approach that would allow them to address temporal uncertainty even more effectively. Wanting to show the impressive quality of their new model, they evaluated both fictitious information and real data on residential burglaries in Valencia, Spain. Next, they conducted several tests to validate their work. In one of them, they analyzed only "perfect" cases, excluding uncertain events. They applied different techniques to fill in missing data in another [30].

Moreover, Hu et al. [29] collected the number of daily crimes in each region and store it in a historical matrix. Additionally, they apply a time window which slides through the matrix of crime occurrences. It was considered a length of 15 days to generate the data samples. To determine the optimal time to analyze, they closely observed the patterns in the data for approximately 15 days. They concluded that this two-week time span was the most appropriate to capture the significant trends and patterns present in the data collected. This choice was not arbitrary but was based on empirical evidence obtained through a meticulous process of monitoring and preliminary analysis of the data.

In conclusion, the field of crime prediction research is constantly evolving by incorporating a diversity of data sources and advanced modeling techniques. The use of dynamic functions that consider temporal windows shows promising potential for research focused on improving the accuracy and adaptability of predictive models of crime in different geographic settings. These advances allow for a better understanding of crime patterns and can lead to more effective strategies to prevent crime and increase safety in our communities.

III. Methodology

The study uses a research methodology as shown in Fig. 1 to examine and analyze crime data using an approach. The first phase focuses on the collection of crime records from a reliable source, followed by the conversion of these records into respective formats suitable for dataset use. This critical first step ensures the quality and reliability of the data under investigation.

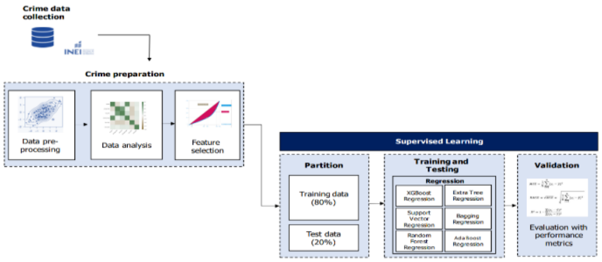

Fig. 1. Research methodology for predicting crimes against patrimony in Lima.

Moving forward, the second phase of crime preparation involves three sub-phases: data pre-processing, data analysis and feature selection. In the first sub-phase, the elimination of empty rows and the application of normalization techniques contribute to refining the dataset. The next sub-phase, data analysis, is a crucial step in understanding patterns and trends within the crime data. Exploratory data analysis techniques help in finding outliers, and other significant factors that may influence criminal activities. Furthermore, feature selection is undertaken during this phase to show the most relevant variables for building robust supervised learning models. This involves assessing the importance of each feature in predicting crime occurrences, enhancing the model's accuracy and efficiency. Additionally, the incorporation of the sliding window method, forecasting on a weekly basis, adds temporal depth to the analysis.

Following the meticulous preparation of the dataset, the research advances to the third phase, which revolves around supervised learning. This phase is further subdivided into three key sub-phases: partition, training and testing, and validation. In the first sub-phase, the dataset is partitioned into an 80% training set and a 20% testing set, proving a robust foundation for model evaluation. The second sub-phase entails the training and testing of various supervised regression models, including XGBoost, Extra Tree, Support Vector, Bagging, Random Forest and AdaBoost, as named in references [15, 16, 17, 18, 20]. Optimization of hyperparameters for each model is conducted to enhance overall performance.

In the stage the regression model's predictions are thoroughly assessed using performance metrics, like Mean Squared Error (MSE) Root Mean Squared Error (RMSE) Coefficient of Determination (R2) and Mean Absolute Error (MAE) as mentioned in references [19, 23, 24]. This meticulous evaluation process looks to decide the model that performs best providing insights, for law enforcement agencies as they tackle and curb activities. The structured and systematic approach outlined in this methodology ensures a robust and data-driven foundation for the development of effective crime intervention strategies.

IV. Experimental Settings

A. Crime Data Collection

The dataset used for the development of this investigation has crimes against property registered in police stations in the Metropolitan Lima area, this was compiled from the official page of the National Institute of Statistics and Informatics. It is important to highlight that the crimes included in this dataset occurred during the years 2015, 2016, and 2017 [26]. This dataset had 490 916 reported crimes. It can be obtained through the following link https://github.com/Cielo12019/Thesis_ML/blob/main/Delitos_Final_2017_2016_2015.xlsx.

The set of crime records is made up of sixty attributes. These attributes describe the crime, considering three aspects: location, time, and type of crime.

- Regarding the location, this allows to show where the criminal incident took place. The related attributes are district, presumed place of occurrence, latitude, and longitude. Likewise, these allow the identification of a criminal pattern based on the area where the criminal act occurred.
- In terms of time, this allows us to conduct an analysis based on years, months, and days to find similar criminal incidents and seasonal patterns that indicate that in certain months or days of the week there is a greater probability of crime. Also, the hour and minute attributes are useful to find in which parts of the day there is a greater number of criminal acts.

In relation to the type of crime, this allows to know the type of event that occurred. The variables that require it are generic crime, specific crime, and modality crime.

B. Crime Data Preparation

In the preparation of the data, the pre-processing of the data is conducted, an exploratory analysis and finally, the selection of characteristics that will be used by machine learning techniques. Each of the sub-phases is presented below.

1) Data pre-processing: During the data pre-processing phase, we filtered the generic type to consider crimes related to property. Within specific crime we carefully examined records about robbery and theft. In the crime modality, various modalities were considered, including assault, vehicle theft, robbery, aggravated robbery, aggravated nighttime robbery, aggravated burglary, burglary, vehicle theft, frustrated robbery, robbery, aggravated robbery, armed robbery, aggravated gang robbery, and attempted robbery. This comprehensive approach allowed us to analyze aspects of activity in relation, to property crimes.

Furthermore, to raise the model's recognition ability, qualitative attributes such as the type of road, the specific crime, and criminal organization were factorized. Moreover, place of occurrence, the medium used, and the instrument used have all been made into the categorical data type also. Additionally, day, month, hour, and minute variables, when thought to be float values, changed to integer data type.

Moreover, we used the district list from the INEI website to match district names, with the location codes in the dataset. We repeated this process for each year which helped us gain an understanding of where these events took place.

Relating the variable, for types of crimes we combined categories with each other. As an example, we labeled all robberies as robbery and grouped all other thefts as theft. Additionally, we classified cellphone robberies under the category of robbery. Adjusted aggravated robberies that occurred at night or in areas to be categorized as aggravated robbery during night. These changes were made to not make the representation of activities consistent but also to simplify the analysis of the dataset afterwards.

During the process of cleaning and filtering the data we dropped rows that had values of ninety-nine in the month day and hour columns. We did this to improve the quality and reliability of the dataset by getting rid of instances that may have unusual information.

To generate the dataset, we excluded features that were not relevant to the crime. Also features with than 70% missing values, such as the location of the police station where the complaint was filed, and the time of registration were left out. We executed this procedure using a notebook on the Deepnote web application. Exported the dataset, in CSV format.

2) Exploratory data analysis: Fig. 2 illustrates the distribution of crimes throughout the weeks of the year. The horizontal axis stands for the weeks while the vertical axis stands for the corresponding number of crimes. A visible cyclic pattern in crime rates can be seen. The number of crimes tends to rise during the weeks of each year reaching its peak in week twenty. Afterward it continuously declines until the end of the year. Nonetheless there is also variation in the data. The range, between the minimum and maximum number of crimes exceeds 100%. These seen temporal patterns offer a chance to improve predictive modeling by including trends, within time periods.

In Fig. 3, the bar graph illustrates the frequency of crimes occurring on each day of the month. The x-axis shows the days, and the y-axis is the number of crimes. It should be noted that on the 31st the number of reported crimes is recorded. Nevertheless, it is important to note that not all months have 31 days; only January, March, May, June, August, October, and December have 31 days. On the contrary third day has the recorded number of crimes. Nonetheless there are no deviations, from days except, for days 19th, 29th and 31st.

In Fig. 4, the x-axis is the 24-hour clock of the day, while the y-axis shows the frequency of crimes that happened each hour. This figure illustrates depicts how the number of committed crimes varies throughout the day. Notably, the highest concentration of crimes is observed during the nighttime, specifically at 20:00 PM. Contrariwise, the lowest number of crimes is recorded during the early morning, at 2:00 AM. Based on this pattern, we considered a new variable as a Time Range created by segmenting the hours into four ranges: 0:00-6:00 AM (early morning), 6:00-12:00 AM (morning), 12:00-18:00 PM (afternoon), and 18:00-24:00 PM (night). This segmentation would provide a more generalized representation of the time of day, helping further analysis based on these time intervals.

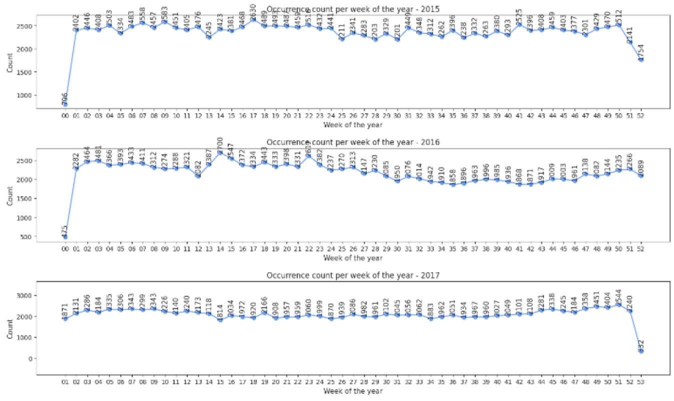

Fig. 2. Crimes per week.

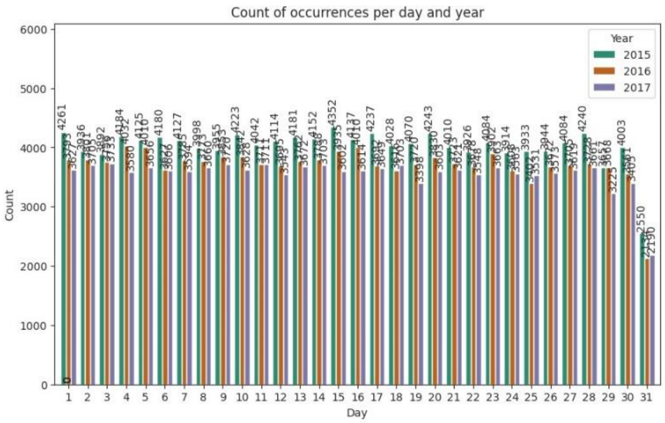

Fig. 3. Crimes per day.

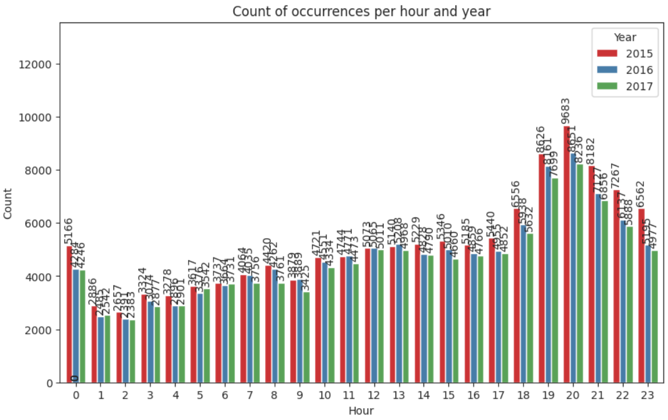

Fig. 4. Crimes per hour.

3) Feature selection: In the characteristic selection stage, it was divided into three sub-phases. In the first subphase, it was decided to select the variables that were going to add value to the model. Therefore, variables that have been used in most research were selected, which are: year, month, day, time range, number of crimes and district in which the event occurred.

From there, an aggregation was performed based on the attributes of year, month, day, time interval and district to determine the number of crimes. This aggregation was stored in a data frame, from which only the column referring to the crime count was extracted.

In the second subphase, a time window approach was applied, as can be seen in Table I, in which the values of the data frame are combined by shifts using different step numbers with 1, 2, 3, 4, 5, 6, 7 and 8. The initial shift shifts the values 8 steps, followed by the second shift with 7 steps, and so on, as each 7 blocks are taken as predictors and the next block as the variable to be predicted. It is important to mentioned that block is a quarter of day. Furthermore, it should be noted that this is done for both the training and testing stages, without losing the time scale.

TABLE I. NUMBER OF CRIMES PER WEEK

BLOCK	1	2	3	4	5	6	7	8	9	10	11	…
TRAINING	91	57	78	48	49	49	78	**35**	52	79	103	…
		57	78	48	49	49	78	35	**52**	79	103	…
			78	48	49	49	78	35	52	**79**	103	…
				48	49	49	78	35	52	79	**103**	…
					…	…	…	…	…	…	…	…
TEST	117	45	90	93	118	44	64	**80**	147	53	75	…
		45	90	93	118	44	64	80	**147**	53	75	…

These shifted data frames are then concatenated along the column axis, resulting in the new data frame, as shown in Table II. The columns of the data frame are renamed with the designations "t-7", "t-6", "t-5", "t-4", "t-3", "t-2", "t-1", "t". The rows originating from the offsets are removed from the new data frame. Starting with row eight, attributed to the maximum offset of eight, and extending to the end of the data frame.

TABLE II. SLINDING WINDOW SCHEME

t-7	t-6	t-5	t-4	t-3	t-2	t-1	t
91	57	78	48	49	49	78	35
57	78	48	49	49	78	35	52
78	48	49	49	78	35	52	79
48	49	49	78	35	52	79	103
49	49	78	35	52	79	103	76
49	78	35	52	79	103	76	42
78	35	52	79	103	76	42	58
…	…	…	…	…	…	…	…
52	79	103	76	42	58	117	45

In the third sub-phase, it was chosen which variables were to be part of the predictors and which were to be part of the goal. Since the purpose was to predict the crime rate, the predictors will be "t-7", "t-6", "t-5", "t-4", "t-3", "t-2", "t-1" and the target variable will be t.

C. *Supervised Learning*

Regression models will be evaluated with training and test data. In addition, its hyperparameters will be perfected with the GridSearchCV technique to improve its performance and metrics such as MSE, RMSE, R2 and MAE will be used to confirm its performance.

1) Partition: Based on the research developed by [27]., the data set was divided into two, 80% was for training and 20% for testing. The crime data set had 35,008 records of crimes against property corresponding to the Metropolitan Lima area. Of that total, 24,504 registrations were designated to train the models. For testing these, the remaining 10,504 records were used.

2) Training and testing: To test the regression models, two scenarios were performed. In the first training and testing scenario, the default parameters were considered to test the R2 of the model. Table III shows the results obtained with the training and test data.

TABLE III. SCENARIO 1 WITH DEFAULT HYPERPARAMETERS

Model	Training data	Test data
XGBoost Regression	0.97	0.75
Extra Tree Regression	1	0.79
Support Vector Regression	0.72	0.71
Bagging Regression	0.95	0.76
Random Forest Regression	0.97	0.78
AdaBoost Regression	0.7	0.69

To improve the performance of the models, a second scenario was carried out where their hyperparameters were analyzed. GridSearchCV was used to fine-tune the hyperparameters of each of the models. The results indicated which values to consider to obtain a better prediction.

Subsequently, the training and the respective test were accomplished, it could be observed that the R2 of the regression models improved, the results are presented in Table IV.

TABLE IV. SCENARIO 2 WITH HYPERPARAMETERS CHOSEN BY GRIDSEARCH CV

Model	Training data	Test data
XGBoost Regression	0.85	0.77
Extra Tree Regression	0.94	0.79
Support Vector Regression	0.83	0.76
Bagging Regression	0.82	0.77
Random Forest Regression	0.85	0.78
AdaBoost Regression	0.71	0.69

3) Validation: To validate the performance of the regression models, the metrics MSE, RMSE, R2 and MAE on the second scenario, which obtained better results. Table V shows the results of the performance metrics.

TABLE V. COMPARISON OF PERFORMANCE METRICS FOR SUPERVISED ALGORITHMS

Model	MSE	RMSE	R2	MAE
XGBoost Regression	200.11	14.15	0.77	11.04
Extra Tree Regression	185.43	13.62	0.79	10.47
Support Vector Regression	211.57	14.55	0.76	11.31
Bagging Regression	206.91	14.38	0.77	11.07
Random Forest Regression	195.63	13.99	0.78	10.72
AdaBoost Regression	270.7	16.45	0.69	13.04

V. RESULTS

From the experimentation conducted, the summary of the results of each model with GridSearchCV hyperparameter setting obtained by each performance metric is shown in Fig. 5. The Extra Tree Regression model achieved better results, obtained the lowest MSE, RMSE, MAE of 185.43, 13.62, 10.47, respectively, and the highest R2 of 0.79, showing that it is the best model with lower error and higher variance reduction for predicting the number of crimes. Next, it was found that the Random Forest Regression, XGBoost Regression, Support Vector Regression and Bagging Regression models obtained similar values for the R2, above 0.76 and close average errors, which shows that there is no significant difference in their metrics. Likewise, it was shown that the model that obtained the lowest R2 of 0.69 and the lowest MSE, RMSE, MAE of 270.7, 16.45, 13.04 was the AdaBoost Regression.

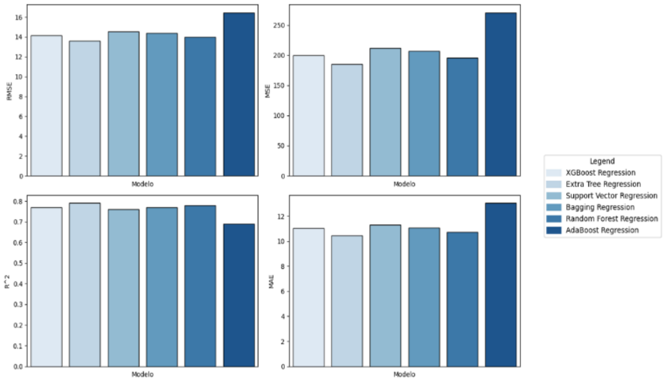

Fig. 5. Metrics for regression models.

VI. Discussion

Although several rigorous research and solutions have been conducted, these studies have not been adopted from a criminological practice where predictive analysis using machine learning techniques is involved. For this reason, in this research supervised machine learning algorithms such as XGBoost Regression, Extra Tree Regression, Support Vector Regression, Bagging Regression, Random Forest Regression and AdaBoost Regression were analyzed and implemented. The models were evaluated with the set of crimes against patrimony in Metropolitan Lima from 2015 to 2017.

To evaluate the supervised regression models, two scenarios were considered. In the first scenario, the highest R2 of 0.79 was obtained with the Extra Tree Regression model. In the second scenario, to improve the results, hyperparameter optimization was performed with GridSearchCV for each of the models. The results showed the best regression model, this was achieved an R2 of 0.79 and error related metrics of 185.43, 13.62, 13.47 for MSE, RMSE, MAE. For the AdaBoost Regression, Random Forest Regression and Bagging Regression models, similar investigations have been conducted as [15], obtained an R2 of 61% with AdaBoost Regression, for which they used numerical variables. It should be noted that they performed hyperparameter optimization using Grid Search CV to improve prediction accuracy. Also, it was found that the Random Forest Regression model obtained an R2 of 0.78 close to the Extra Tree Regression model and the MSE, RMSE, MAE of 195.63, 13.99 and 10.72.

Belesiotis et al. [17] mentions that Bagging Regression and Random Forest Regression algorithm use similarly; however, they have differences because of the hyperparameter fitting rule they employ. In addition, Random Forest Regression offers a higher R2 for cases where the true values of the coefficients of a set is zero or small. Likewise, most of these investigations have worked with data sets from Asia and Europe, which have a better record and quantity of data on reported crimes and have a greater number of variables, which allows a better distribution and analysis of the data to obtain more accurate models; however, it should be noted that according to the INEI [28], only 15.5% of the Peruvian population that has been a victim of a criminal act chooses to report it to the National Police or the Attorney General's Office, because they consider it to be a waste of time. Therefore, the amount of data that is entered into the INEI for investigation purposes is reduced and must be preprocessed before being used.

VII. Conclusion

The main goal of this research was to compare regression models for the prediction of property crime rates in the districts of Metropolitan Lima as a function of space and time. For that reason, supervised models such as XGBoost Regression, Extra Tree Regression, Support Vector Regression, Bagging Regression, Random Forest Regression and AdaBoost Regression were implemented. These supervised models obtained an R2 lower than 0.79 and higher than 0.69. These predictions, in comparison with earlier studies, are within the prediction range that varies between 0.50 and 0.80 of R2 for regression models.

When making predictions with historical data, it would not be possible to obtain values that approximate the real events because when trying to find patterns, the model suffers an overadjustment and does not find new patterns that adapt to the new events. Given that the phenomenon of crime has changing patterns, predictions could be made with information from current events, and thus know the amount of crime that could occur to prevent crimes. Also, with the predictions, police officers can plan their distribution in Metropolitan Lima a week in advance.

As future work, it is suggested to include more data sources related to the geographic space where the crime originated, and data related to criminal activities to find crime hotspots. Likewise, it is important to use hybrid methods that include regression and classification models to have models that are more efficient and help to counteract crime.

References

[1] H. Adel, M. Salheen, and R. A. Mahmoud, “Crime in relation to urban design. case study: The greater cairo region,” Ain Shams Engineering Journal, vol. 7, pp. 925–938, 9 2016.

[2] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, “Once upon a crime: Towards crime prediction from demographics and mobile data,” p. 427–434, 2014. [Online]. Available: https://doi.org/10.1145/2663204.2663254

[3] I. Kawthalkar, S. Jadhav, D. Jain, and A. V. Nimkar, “A survey of predictive crime mapping techniques for smart cities,” 2020 National Conference on Emerging Trends on Sustainable Technology and Engineering Applications, NCETSTEA 2020, 2 2020.

[4] Numbeo, “Crime index by country 2023,” Numbeo, 2023. [Online]. Available: https://www.numbeo.com/crime/rankings by country.jsp

[5] I. N. de Informa´tica y Estad´ıstica, “Principales indi- cadores de seguridad ciudadana a nivel regional,” Instituto Nacional de Informa´tica

y Estad´ıstica, 2020. [Online]. Available: https://www.inei.gob.pe/media/MenuRecursivo/boletines/ estadisticas-de-seguridad-ciudadana-regional-nov19-abr20.pdf

[6] ——, "Informe te´cnico - estad´ısticas de seguridad ciudadana," Instituto Nacional de Informa´tica y Estad´ıstica, 2020. [On- line]. Available: http://m.inei.gob.pe/media/MenuRecursivo/boletines/ boletin-de-seguridad-ciudadana.pdf

[7] B. I. de Desarrollo, "Los costos del crimen y de la violencia: nueva evidencia y hallazgos en ame´rica latina y el caribe," Banco Interamericano de Desarrollo, 2017. [Online]. Available: https://goo.su/S4L2Gk

[8] I. N. de Informa´tica y Estad´ıstica, "Victimizacio´n en el peru´ 2010 – 2019," Instituto Nacional de Informa´tica y Estad´ıstica, 2019.

[9] ——, "Estad´ısticas de las tecnolog´ıas de informacio´n y comunicacio´n en los hogares," Instituto Nacional de Informa´tica y Estad´ıstica, 2020. [Online]. Available: https://www.inei.gob.pe/media/MenuRecursivo/ boletines/informe tic abr-may jun2020.pdf

[10] M. A. Awal, J. Rabbi, S. I. Hossain, and M. M. A. Hashem, "Using linear regression to forecast future trends in crime of bangladesh," pp. 333–338, 2016.

[11] A. A. Biswas and S. Basak, "Forecasting the trends and patterns of crime in bangladesh using machine learning model," pp. 114–118, 2019.

[12] P. Gera and D. R. Vohra, "Predicting future trends in city crime using linear predicting future trends in city crime using linear predicting future trends in city crime using linear predicting future trends in city crime using linear regression regression regression regression," International Journal of Computer Science Management Studies), vol. 14, 2014. [Online]. Available: www.ijcsms.com

[13] R. Sridhar and D. Fathimal, "Crime prediction and visualisation using data analytics," International Research Journal of Engineering and Technology, 2020. [Online]. Available: www.irjet.net

[14] L. G. Alves, H. V. Ribeiro, and F. A. Rodrigues, "Crime prediction through urban metrics and statistical learning," Physica A: Statistical Mechanics and its Applications, vol. 505, pp. 435–443, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0378437118304059

[15] J. Silva, L. Romero, R. J. Gonza´lez, O. Larios, F. Barrantes, O. B. P. Lezama, and A. Manotas, "Algorithms for crime prediction in smart cities through data mining," pp. 519–527, 2020.

[16] J. Wang, J. Hu, S. Shen, J. Zhuang, and S. Ni, "Crime risk analysis through big data algorithm with urban metrics," Physica A: Statistical Mechanics and its Applications, vol. 545, p. 123627, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0378437119320229

[17] A. Belesiotis, G. Papadakis, and D. Skoutas, "Analyzing and predicting spatial crime distribution using crowdsourced and open data," ACM Trans. Spatial Algorithms Syst., vol. 3, no. 4, apr 2018. [Online]. Available: https://doi.org/10.1145/3190345

[18] B. Cavadas, P. Branco, and S. Pereira, "Crime prediction using regression and resources optimization," pp. 513–524, 2015.

[19] L. McClendon and N. Meghanathan, "Using machine learning algorithms to analyze crime data," Machine Learning and Applications: An International Journal (MLAIJ), vol. 2, no. 1, pp. 1–12, 2015.

[20] S. K. Rumi, P. Luong, and F. D. Salim, "Crime rate prediction with region risk and movement patterns," CoRR abs/1908.02570, 2019.

[21] G. J. J. and L. Aera, "Crime prediction and socio-demographic factors: A comparative study of machine learning regression-based algorithms," Journal of Applied Computer Science Mathematics, vol. 13, pp. 13–18, 4 2019. [Online]. Available: https://doi.org/10.4316/JACSM.201901002

[22] G. Farrell, W. Sousa, and D. Weisel, "The time-window effect in the measurement of repeat victimization: a methodology for its examination, and an empirical study," Crime Prevention Studies, vol. 13, 01 2002.

[23] V. Ingilevich and S. Ivanov, "Crime rate prediction in the urban environment using social factors," Procedia Computer Science, vol. 136, pp. 472–478, 2018, 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S1877050918315667

[24] G. Saltos and M. Cocea, "An exploration of crime prediction using data mining on open data," International Journal of Information Technology & Decision Making, vol. 16, no. 05, pp. 1155–1181, 2017. [Online]. Available: https://doi.org/10.1142/S0219622017500250

[25] S. Wu, C. Wang, H. Cao, and X. Jia, "Crime prediction using data mining and machine learning," Advances in Intelligent Systems and Computing, vol. 905, pp. 360–375, 8 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-14680-1 40

[26] I. N. de Estad´ıstica e Informa´tica, "Registro nacional de denuncias de delitos y faltas 2018," Instituto Nacional de Estad´ıstica e Informa´tica, 2018. [Online]. Available: https://webinei.inei.gob.pe/anda inei/index.php/catalog/652/study-description

[27] A. G. Pratibha and S. D. Uprant, "L. chouhan," crime prediction and analysis," 2020.

[28] I. N. de Informa´tica y Estad´ıstica, "Informe te´cnico - estad´ısticas de seguridad ciudadana," Instituto Nacional de Informa´tica y Estad´ıstica, 2021. [Online]. Available: https://www.inei.gob.pe/media/ MenuRecursivo/boletines/boletin seguridad nov20 abr21.pdf

[29] Hu, K., Li, L., Liu, J., & Sun, D. (2021). "DuroNet." ACM Transactions on Internet Technology, 21(1), 1–24. https://doi.org/10.1145/3432249

[30] Briz-Redón, Á. (2024). "A Bayesian Aoristic Logistic Regression to Model Spatio-Temporal Crime Risk Under the Presence of Interval-Censored Event Times." Journal of Quantitative Criminology. https://doi.org/10.1007/s10940-023-09580-1