Journal of Computer Science

Research Article

CRISP-DM-Based Mobile Application for Predicting High-Crime Areas in Metropolitan Lima

Hugo Vega Huerta[1], Javier Vilca Velasquez[1], Nicolas Anicama Espinoza[1], Luis Guerra Grados[1], Jorge Pantoja Collantes[1], Oscar Benito Pacheco[1], Juan Carlos Lázaro Guillermo[2] and Rubén Gil Calvo[1]

[1]Department of Computer Science, Universidad Nacional Mayor de San Marcos (UNMSM), Lima, Peru
[2]Department of Basic Sciences, Universidad Nacional Intercultural de la Amazonia (UNIA), Ucayali, Peru

Article history
Received: 24-03-2025
Revised: 28-10-2025
Accepted: 29-10-2025

Corresponding Author:
Hugo Vega Huerta
Department of Computer Science, Universidad Nacional Mayor de San Marcos (UNMSM), Lima, Peru
Email: hvegah@unmsm.edu.pe

Abstract: The city of Lima, Peru, has been facing a serious climate of citizen security that has risen extremely high in recent years. The objective of this work is to identify and predict areas of high crime incidence through a mobile application based on historical data on criminal incidents recorded by users. The mobile application has been implemented using the CRISP-DM methodology, which includes the stages of business understanding, data understanding, data preparation, modeling, evaluation, and implementation. The main machine learning algorithms used were Random Forest and Gradient Boosting; likewise, visualization techniques such as heat maps were used to represent criminal events. The results obtained in the prediction of the occurrence of crimes were: Using the Random Forest algorithm, an accuracy of 87% was achieved and using Gradient Boosting 84%, These findings allow people who use the mobile application to know in real time which zones or areas are of high crime incidence therefore dangerous in this way they will be able to opt for prevention behaviors and that these technologies can help address the Security challenges in the city of Lima.

Keywords: Crime Prediction, High Crime Incidence Areas, Crisp-DM, Heat Maps

Introduction

Citizen security is a constant concern in large cities, and Metropolitan Lima is no exception. Theft and crime are major challenges that affect the quality of life of citizens and put their well-being at risk. To address this problem, we propose the implementation of a predictive surveillance system based on machine learning, with the objective of improving citizen security and reducing robbery recidivism in the districts of Metropolitan Lima. Fig. 1 shows the reports ordered by the district.

The implementation of this system offers an innovative and promising approach to address security challenges (Williams et al., 2016). This approach focuses on analyzing large volumes of historical and real-time data to identify hidden patterns, correlations, and trends that can be used to predict areas or times of increased theft risk (Wang et al., 2020; Zhang et al., 2020).

Predictive monitoring, one of the key facets of this system, uses machine learning algorithms to analyze collected data (Bennett Moses and Chan, 2018). These algorithms are trained on historical data and can identify complex relationships between variables, enabling them to predict potential crime incidents.

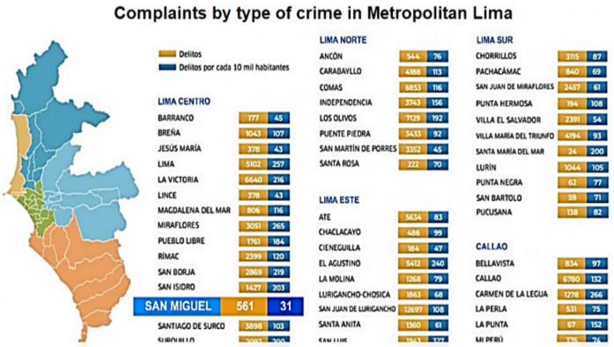


Fig. 1: Complaints by type of crime in Metropolitan Lima (IDL, 2021)

This predictive capability allows citizens to identify high-risk areas and avoid them or take preventive measures (Weisburd et al., 2010). By reducing opportunities for criminals and anticipating risky situations, the machine learning-based security management system can help deter criminals and ultimately reduce the recurrence of robberies in the districts of Metropolitan Lima. In addition, analyzing the behavioral patterns of known offenders, such as the areas where they will offend or their forms of crime, can help identify potential repeat offenders and enable early interventions to prevent new crimes (Gerber, 2014). However, it is important to emphasize that the implementation of this system cannot be considered a single solution to address citizen security. It would be of great importance to complement it with other measures, such as the increase in police officers to safeguard order, prevention and awareness campaigns, and reorganization of urban infrastructure. These combined strategies can lay the foundation for a safer environment and improve the quality of life for the inhabitants of the Lima metropolitan area.

Related Work

Lie (2017) patented a public safety application that uses a tracking device and a client device that has multiple security functionalities, as seen in Fig. 2.

Daniel (2022) patented a method to extract safety data from multiple data sources and, from the analysis of this data, generate an action to be performed by the autonomous data machine. Meijer and Wessels (2019) mentions that predictive policing consists of advanced hotspot identification models and risk terrain analysis to forecast where criminal activity is most likely to occur. On the other hand, Chainey et al. (2008) specify that there are predictive surveillance techniques such as point mapping, thematic mapping of geographical areas, spatial ellipses, thematic mapping of grids, and kernel density estimation. In addition, he indicated that CrimeRank yields better results in the analysis of the PAI (Predictive Accuracy Index) compared to other algorithms such as Random Forest, Hawkes, and GLM (Mohler et al., 2020).

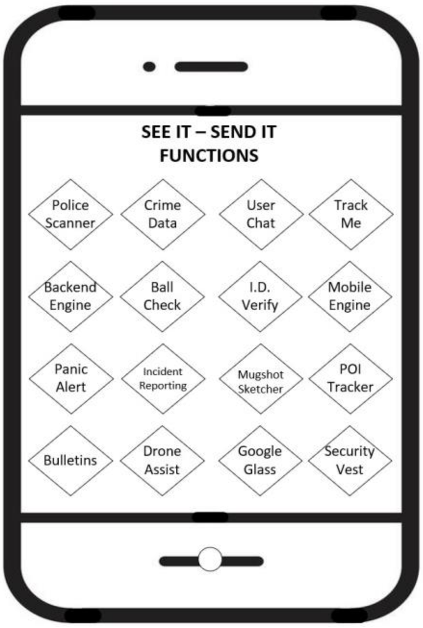


Fig. 2: App Features (Lie, 2017)

According to Martinez-Plumed et al. (2021), Brzozowska et al. (2023); Bokrantz et al. (2024), the CRISP-DM (Cross-Industry Standard Process for Data Mining) model is the predominant methodological standard in data engineering for structuring data mining and data science projects. Its sequential and iterative approach, based on six phases from business understanding to implementation has proven to be highly adaptable to different domains, including complex urban environments. In this work, CRISP-DM provides the structural framework for the development of a smart mobile solution aimed at geospatial prediction of areas with high crime rates in Metropolitan Lima, as shown in Fig. 3.

According to Mohler et al. (2020); Wheeler and Reuter (2020), the Prediction Accuracy Index (PAI) is calculated by dividing the percentage of the hit rate by the percentage of the area. On the other hand, Alves et al. (2018) mention that the Random regressor forest allows predicting crime and quantifying the influence of urban indicators on homicides.

One of the most effective approaches between Lexicon and Machine Learning was determined by Sudar et al. (2024). Such an approach includes Random Forest, Logistic Regression (LR), Naive-Bayes, and Support Vector Machine (SVM), demonstrating that Naive-Bayes was the best among the evaluated models. According to Putro and Sensuse (2022), there are 13 threats and 16 vulnerabilities identified in the security principles of critical information infrastructures, and only 25% of the principles have all the security characteristics. Deep neural networks facilitate efficient planning and optimally address a wide range of scenarios (Maquen-niño et al., 2023a-b; DelaCruz-VdV et al., 2023).

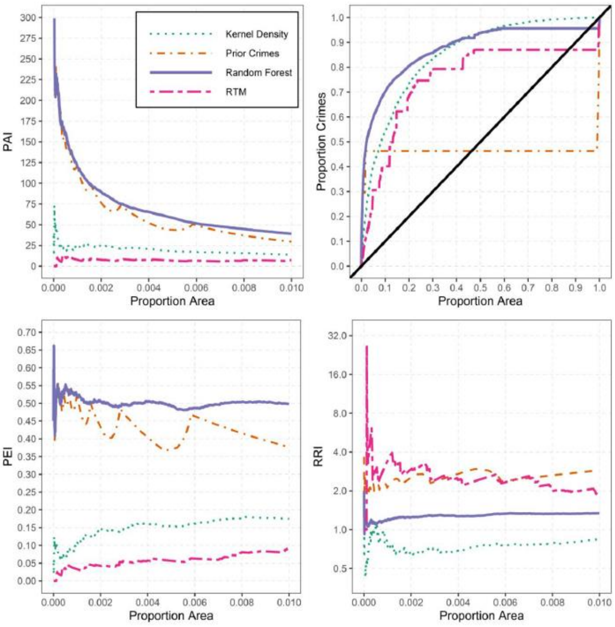


Fig. 3: Accuracy metrics for each of the different prediction models (DelaCruz-VdV et al., 2023)

The Random Forest (RF) algorithm builds multiple decision trees and aggregates their outputs to improve predictive accuracy (Schonlau and Zou, 2020). Its robustness in handling high-dimensional, nonlinear data has been successfully applied in geospatial modeling (Mutale et al., 2024) and geotechnical predictions (Zhang et al., 2021). In the context of our CRISP-DM-based mobile application for crime hotspot prediction in Lima, RF provides an ideal framework for managing heterogeneous urban crime data while enhancing generalization in spatial risk modeling.

Gradient boosting is a machine learning technique that has proven to be very effective in selecting features to optimize data engineering in the agricultural industry. For example, gradient boosting algorithms such as CatBoost and LightGBM have been used to select optimal wavelengths in Near-Infrared (NIR) spectroscopy to predict moisture and protein in multi-country corn kernels via NIR spectroscopy (Zheng et al., 2024). In addition, the extreme gradient boosting (XGBoost) algorithm has been successfully applied in predicting the Remaining Lifespan (RUL) of lithium-ion batteries in energy storage systems. It has been observed that, unlike physical models of failures, XGBoost is more flexible and non-linear, offering greater accuracy and efficiency in predicting large-scale battery degradation using the extreme gradient boosting algorithm (Brillianto Apribowo et al., 2024). This approach could strengthen research on crime prediction in Lima by allowing for a more robust feature selection for the model and could be explored in the study to improve the flexibility and accuracy of the crime prediction model, outperforming traditional linear models.

The most important factor for the success of an application is the optimal organization of its data (Yauri et al., 2023 Arrieta-Espinoza et al., 2023; DeLaCruzVdv et al., 2023; Velez-villanueva et al., 2023; Sánchez-tello et al., 2023; Melgarejo Solis et al., 2023; Rivera-Alvino et al., 2023)

Recent studies highlight that the KDD methodology complements CRISP-DM by integrating citizen participation and real-time geospatial data analysis. This combination enhances predictive accuracy and promotes sustainable public safety management, allowing decision-makers to identify risk areas proactively and implement effective urban security strategies (Vega-Huerta et al., 2025).

Methods

The CRISP-DM methodology has been considered due to its systematic structure and well-defined approach to the data mining process. This methodology provides a step-by-step guide that allows us to effectively organize the project, from understanding the business to implementing the solution (Poh et al., 2018), which is fundamental in our project of identifying areas of high incidence of crime in Metropolitan Lima. This methodology is carried out in several stages, as can be seen in Fig. 4.

Below, the activities carried out in the project will be broken down for each of the phases of the CRISP-DM methodology.

Understanding the Business

In the capital Lima, over the years, there has been an alarming increase in the incidence of criminal acts, which represents a growing concern for the community and local authorities, which is why the objective of the app is efficiency in the detection of the most dangerous areas of the districts of Metropolitan Lima.

Understanding the Data

The dataset used in this article was extracted from an INEI dataset, which contains information on criminal acts that occurred in Peru during 2022 and 2024.

The dataset consisted of a single source of information, the data from the National Institute of Statistics and Informatics (INEI), as it represents an official and reliable source for identifying frequent crimes committed in Metropolitan Lima. The INEI data guarantees quality standards, temporal consistency, and geographic coverage that are essential for the development of reliable predictive models. The use of latitude and longitude coordinates from this source allows for precise georeferencing of actual criminal events, which constitutes a guarantee of methodological rigor and reliability of the results obtained.

In Table 1, we see the most important columns of the dataset and their respective descriptions.

The following categories of crimes were also considered, which can be seen in Table 2.

Once the data has been described, the content of the data is explored using graphs. Fig. 5 shows the number of crimes in some districts of Metropolitan Lima, with San Juan de Lurigancho being where the most criminal acts occur during 2022-2023.

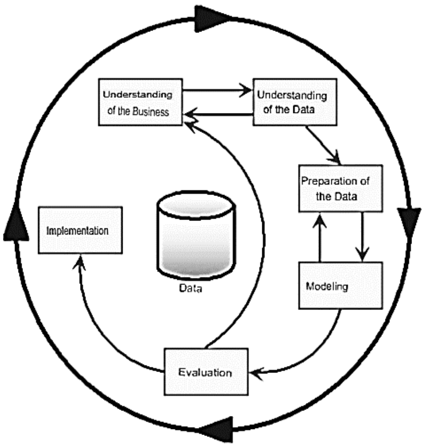


Fig. 4: Methodology CRISP-DM (Cortina, 2015)

Table 1: Columns in the collected data

Column	Description
X	It is the latitude of where the criminal act occurred
Y	It is the longitude of where the criminal act occurred
GENERICO	Type of crime committed
NOMBDEP	Department in which the criminal act occurred
NOMBPROV	The province in which the criminal act occurred
NOMBDIST	District in which the criminal act occurred

Table 2: Categories of crime

Category	Description
ROBBERY	The act of seizing someone else's property, usually with violence or intimidation
THEFT	The act of seizing someone else's property without using violence or intimidation
ATTEMPTED MURDER	Attempt to cause the death of another person
HOMICIDE	The act of deliberately killing another person

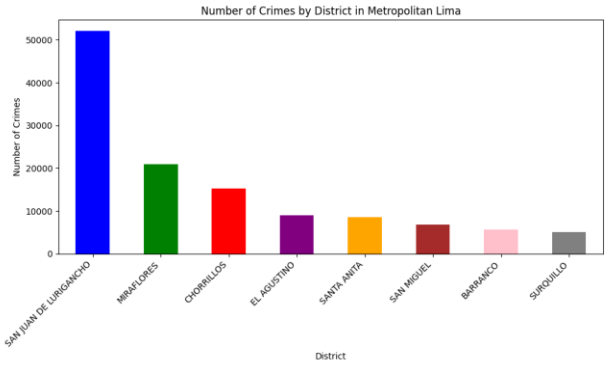


Fig. 5: Number of Crimes by District in Lima

Data Preparation

At this stage, the columns of interest were chosen; in this case, 2 were selected, which can be seen in Table 3.

As can be seen in Figure 5, the cases with the highest incidence were registered in the district of San Juan de Lurigancho, whose coordinates are X: -77.009587, Y: -11.9970785.

As could be seen in the Verify data quality stage, no nulls were found in the columns of interest, so we will not preprocess the data.

In the data, there are no cases in which there was no crime, so random latitudes and longitudes were generated within the range of Metropolitan Lima, validating that they are not the same as those obtained for criminal acts. Once these latitude and longitude values have been obtained for the cases in which there was no crime, a column will be added, which will represent whether or not there was a crime, which will take values of 1 and 0, respectively. By selecting these as simulation parameters, we will be directly addressing the elements necessary to model and predict crime incidence based on geographic location and crime occurrence.

Table 3: List of columns of interest

Column
X (Latitude)
Y (Longitude)

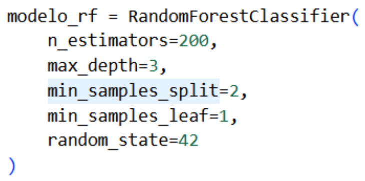
Modeling

The classification algorithms of Random Forest, Support Vector Machine (SVM), Gradient Boosting, and XGBoost were selected based on the reviewed literature. For the first algorithm, Random Forest, several parameters were configured: n_estimators, which defines the number of decision trees comprising the ensemble max_depth, which controls the maximum depth of each tree to prevent overfitting; min samples split, which specifies the minimum number of samples required to split an internal node; min_samples_leaf, which determines the minimum number of samples needed to form a leaf; and random state, which ensures the reproducibility of the results by setting a constant random seed, as shown in Fig. 6.

For the second algorithm, the implemented Support Vector Machine (SVM) model employs an RBF kernel and data standardization through StandardScaler. To optimize its performance, GridSearchCV was implemented, an exhaustive search technique that systematically evaluates hyperparameter combinations defined in param_grid: C (regularization), gamma (kernel coefficient), and class weight (class balancing). The process performed 3-fold cross-validation, optimizing the F1 Score rather than accuracy to address the class imbalance inherent in crime datasets. GridSearchCV automatically identified the optimal configuration, allowing selection of the model with superior balance between precision and recall for crime prediction, as illustrated in Figure 7.

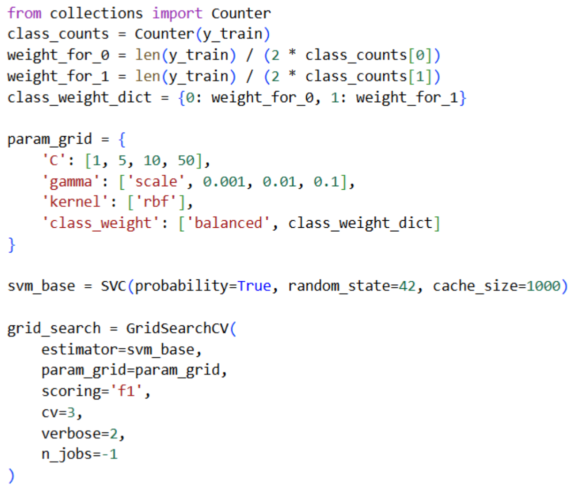
For the third algorithm, the following parameters were configured in the Gradient Boosting (Gradient Boosting Classifier) model: N estimators, which defines the number of decision trees trained sequentially to improve the model's performance; max depth, which controls the maximum depth of each tree to prevent overfitting; and random_state, which ensures the reproducibility of results by setting a constant random seed. These parameters contribute to optimizing the learning process and maintaining the model's stability, as shown in Figure 8.

Finally, Figure 9 presents the parameters configured for the XGBoost (XGBClassifier) classification model: n_estimators, which determines the number of decision trees used in the boosting process; max depth, which controls the maximum depth of each tree to prevent overfitting; use label encoder, which disables the old label encoder and avoids unnecessary warnings; eval_metric, which sets the evaluation metric using logarithmic loss, suitable for binary classification problems; and random_state, which sets a constant random seed.



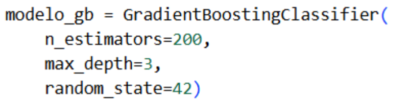
```
modelo_rf = RandomForestClassifier(
    n_estimators=200,
    max_depth=3,
    min_samples_split=2,
    min_samples_leaf=1,
    random_state=42
)
```

Fig. 6: Random Forest Model



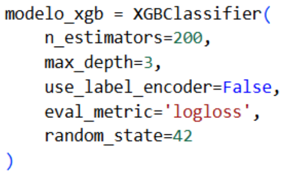
```
from collections import Counter
class_counts = Counter(y_train)
weight_for_0 = len(y_train) / (2 * class_counts[0])
weight_for_1 = len(y_train) / (2 * class_counts[1])
class_weight_dict = {0: weight_for_0, 1: weight_for_1}

param_grid = {
    'C': [1, 5, 10, 50],
    'gamma': ['scale', 0.001, 0.01, 0.1],
    'kernel': ['rbf'],
    'class_weight': ['balanced', class_weight_dict]
}

svm_base = SVC(probability=True, random_state=42, cache_size=1000)

grid_search = GridSearchCV(
    estimator=svm_base,
    param_grid=param_grid,
    scoring='f1',
    cv=3,
    verbose=2,
    n_jobs=-1
)
```

Fig. 7: Support Vector Machine Model



```
modelo_gb = GradientBoostingClassifier(
    n_estimators=200,
    max_depth=3,
    random_state=42)
```

Fig. 8: Gradient Boosting Model



```
modelo_xgb = XGBClassifier(
    n_estimators=200,
    max_depth=3,
    use_label_encoder=False,
    eval_metric='logloss',
    random_state=42
)
```

Fig 9: XGBoosting Model

Evaluation

In the Random Forest model, better results were obtained. This can be observed in its confusion matrix, for which it correctly detected 1297 cases in which there were no crimes and 32 false positives, and correctly detected 1,448 crimes, with 422 false negatives. This is observed in Figure 10.

In the Support Vector Machine (SVM) model, optimized through GridSearchCV, superior results were obtained compared to Random Forest. This improvement can be observed in its confusion matrix, for which it correctly detected 1237 cases in which there were no crimes and only 2 false positives, and detected 1500 cases in which there were crimes and 437 false negatives, as shown in Figure 11.

In the Gradient Boosting model, good results were also obtained. This can be observed in its confusion matrix, for which it correctly detected 1252 cases in which there were no crimes and 77 false positives, and detected 1617 cases in which there were crimes and 253 false negatives. This is observed in Figure 12.

Finally, in Figure13 the results obtained by the XGBoost model are visualized in its confusion matrix, where 1248 cases in which there were no crimes were correctly detected, and 81 false positives; in addition, 1655 cases in which crimes did occur were identified, and 215 false negatives.

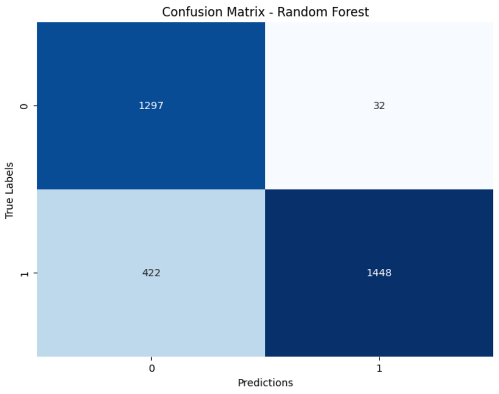


Fig. 10: Random Forest Confusion Matrix

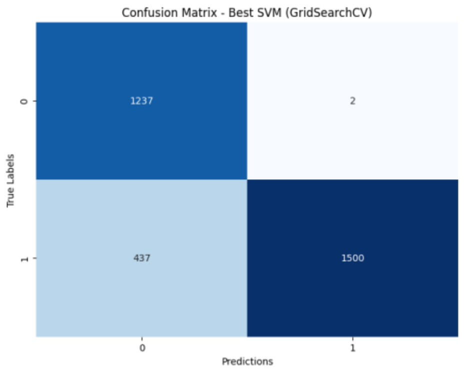


Fig. 11: SVM Confusion Matrix

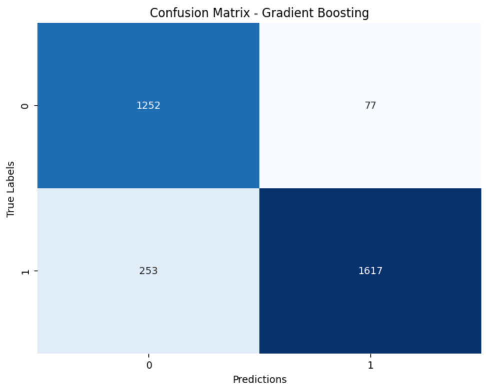


Fig. 12: Gradient Boosting Confusion Matrix

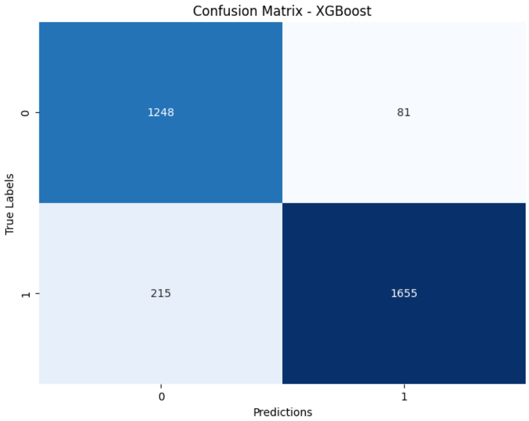


Fig. 13: XGBoosting Confusion Matrix

Deployment

The solution has multiple functionalities: Permissions of two types. Figure 14 shows in detail the use cases of the implemented system.

Figure 15 shows the technological architecture used by Node for communication with servers, Postman for the development and testing of APIs, and Python for artificial intelligence algorithms.

Figure 16 shows the deployment model. The Android device acts as the host for the server, which is divided into two parts: The mobile application and the database. In addition, within the mobile app, an external API is integrated for additional functions and remote data access.

Figure 17 shows the prototype of the mobile app's start and main menu, which offers an intuitive and accessible interface for users. The clear design and user-friendly navigation elements enhance the user experience and make it easier to access the app's various functionalities.

In the prototype shown in Figure 18, the "View Map" function is emphasized through a clear arrangement of elements that allow users to visually explore and understand the geographical distribution of criminal acts. The inclusion of zoom controls and interaction with markers on the map improves the user experience by providing detailed information about specific locations.

Figure 19 shows an interface focused on the validation of crimes and the analysis of statistics. The simple and clear design provides a comfortable and easy-to-navigate user experience, highlighting the core functionality of the system.

Figure 20 shows the options for configuring alerts and logging reports. The interface offers its functionalities to users in a simple and friendly way.

Figure 21 shows an interface of key visualizations, including scatter plots, histograms, and pie charts, with patterns detected using the predictive model. The prediction of future crimes with 90% accuracy is also appreciated, providing a visual representation of the areas with high probability of crime in the metropolitan Lima area.

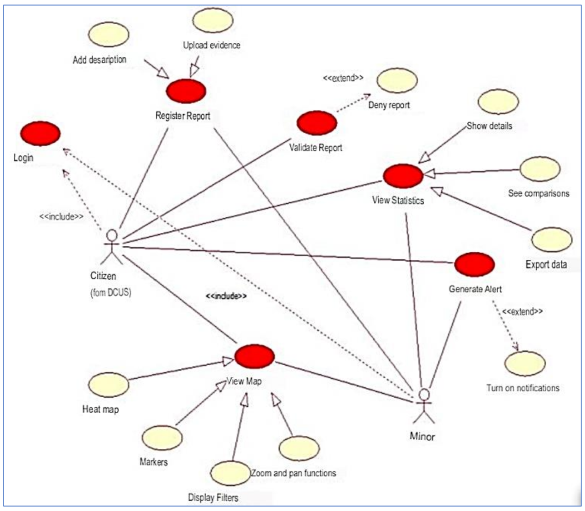


Fig. 14: Prioritized CUS Diagram

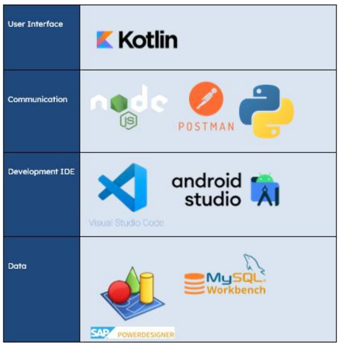


Fig. 15: System Architecture

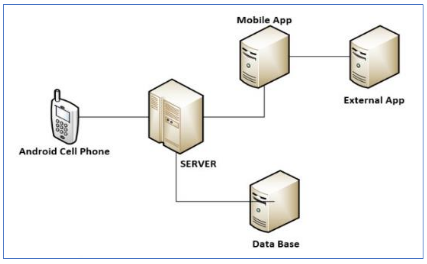
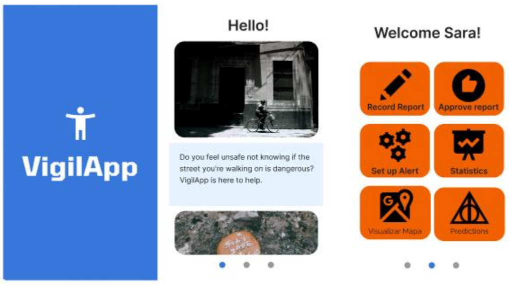


Fig. 16: Deployment Model



Do you feel unsafe not knowing if the street you're walking on is dangerous? VigilApp is here to help.

Fig. 17: Prototype Main and Start Menu

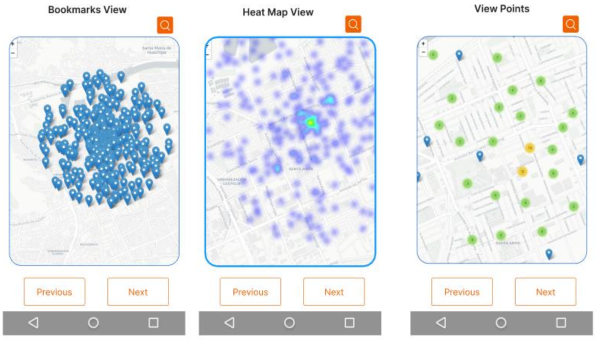
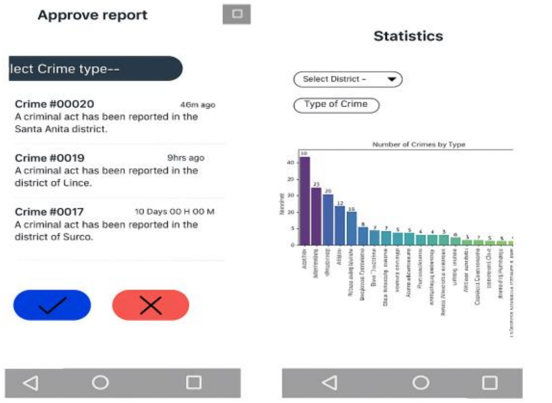


Fig. 18: Prototype Map Display

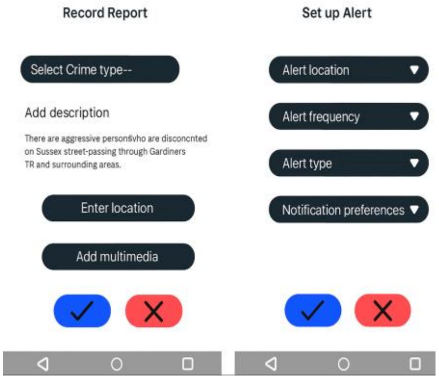


Crime #00020 46m ago
A criminal act has been reported in the Santa Anita district.

Crime #0019 9hrs ago
A criminal act has been reported in the district of Lince.

Crime #0017 10 Days 00 H 00 M
A criminal act has been reported in the district of Surco.

Fig. 19: Prototype of Validate Crime and View Statistics



There are aggressive personšvho are disconcnted on Sussex street-passing through Gardiners TR and surrounding areas.

Fig. 20: Prototype Log Report and Configuration Alert

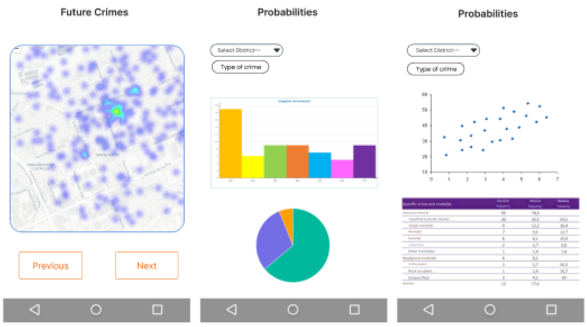


Fig. 21: Prototype predictions

Results

In order to present the results of this research, the four models were compared through their obtained metrics as visualized in Fig. 22. The Random Forest model achieved an Accuracy of 0.86, indicating that 86% of the total predictions were correct; its AUC of 0.92 demonstrates a good capacity to distinguish between positive and negative classes; the F1 Score of 0.86 reflects a balance between precision and recall; however, its Recall of 0.77 was the lowest among the four models, meaning that it only correctly detected 77% of the actual crime cases, leaving 23% of false negatives unidentified.

The Support Vector Machine (SVM) model exhibited an Accuracy of 0.86, indicating that 86% of its predictions were accurate, matching the Random Forest model; its AUC of 0.90 demonstrates a solid discriminative capability between criminal and non-criminal events, though slightly lower than the other boosting models; the F1 Score of 0.87 reflects a good balance between precision and recall, and its Recall of 0.77 was equal to Random Forest's, meaning it correctly identified 77% of actual crimes while missing 23% of positive cases.

On the other hand, the Gradient Boosting model showed better performance than the previous ones, obtaining an accuracy of 0.90, that is, 90% of its predictions were correct; its AUC of 0.96 indicates an almost optimal capacity to discriminate between crimes and non-crimes; the F1 Score of 0.91 demonstrates a balance between the precision of positive predictions and the ability to detect all positive cases; and its Recall of 0.86 indicates that it correctly identified 86% of actual crimes, reducing false negatives compared to Random Forest and SVM.

Finally, the XGBoost model presented the best results, achieving an accuracy of 0.91, which means it was correct in 91% of all its predictions; it matched Gradient Boosting with an AUC of 0.96, confirming its classification capability; it stood out with the highest F1 Score of 0.92, evidencing the best balance between precision and thoroughness; and it obtained the highest Recall of 0.89, which implies that it correctly detected 89% of all crimes that occurred, minimizing false negatives to only 11%.

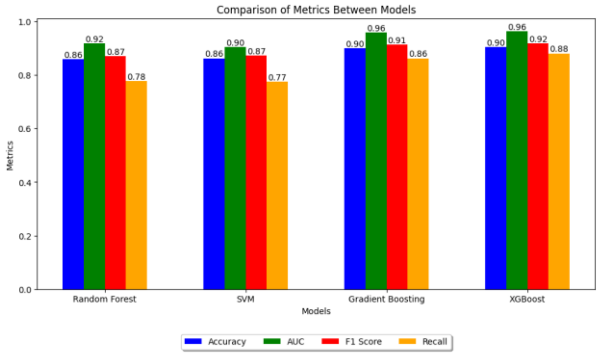


Fig. 22: Comparison of metrics

This last characteristic is particularly relevant in public safety applications, where it is important to identify as many actual crimes as possible to allow effective preventive interventions. Therefore, XGBoost positioned itself as the most efficient and reliable model for crime prediction in this study.

Discussion

The results obtained in this research demonstrate the effectiveness of machine learning models for crime prediction in Metropolitan Lima, particularly when structured under the CRISP-DM methodology. As established by Martinez-Plumed et al. (2021), Brzozowska et al. (2023); Bokrantz et al. (2024), CRISP-DM provides a robust structural framework for data mining projects in complex urban environments, which was validated in this study through the systematic development and evaluation of three models.

The XGBoost model emerged as the most effective approach, achieving the highest metrics across all evaluated parameters: Accuracy of 0.91, AUC of 0.96, F1 Score of 0.92, and recall of 0.89. These results are consistent with the research by Brillianto Apribowo et al. (2024), who found that XGBoost's extreme gradient boosting algorithm provides superior accuracy and efficiency due to its flexibility and non-linear modeling capabilities. The recall of 0.89 is particularly significant, as it means that 89% of actual crimes were correctly identified, minimizing false negatives to only 11%. This characteristic is essential in public safety applications, as emphasized by Lie (2017); Daniel (2022), who patented systems that require accurate security data analysis to determine appropriate preventive actions.

Comparing our results with previous studies, Mohler et al. (2020) reported that CrimeRank yielded better results in Predictive Accuracy Index (PAI) analysis compared to algorithms such as Random Forest. While our study did not use CrimeRank, the superior performance of XGBoost suggests that advanced ensemble methods can achieve comparable or potentially superior results in crime prediction tasks. Additionally, Sudar et al. (2024) found that Naive-Bayes outperformed other machine learning models, including Random Forest, SVM, and Logistic Regression, in their context. However, our findings demonstrate that gradient boosting methods, particularly XGBoost, can surpass traditional algorithms when applied to geospatial crime prediction with properly structured urban data.

Furthermore, the predictive capabilities demonstrated by these models complement the existing crime prediction techniques mentioned by Meijer and Wessels (2019); Chainey et al. (2008), such as hotspot identification models and risk terrain analysis. The high AUC values (0.96 for both Gradient Boosting and XGBoost) indicate excellent discriminatory capacity between high-risk and low-risk areas, which is fundamental for the geospatial prediction objectives of this mobile application.

The results validate that XGBoost represents an efficient and reliable approach for crime prediction in Metropolitan Lima. These findings contribute to the growing body of evidence supporting the use of advanced machine learning methods in public safety applications, while demonstrating the practical applicability of the CRISP-DM methodology in complex urban crime prediction scenarios.

Since the predictive variables are restricted to spatial coordinates, the model functions primarily as a hotspot identification tool rather than a causal predictor of future crimes. Future versions should integrate temporal and socioeconomic indicators to enhance generalization.

The integration of validated and standardized datasets significantly enhances AI predictive performance, improving diagnostic reliability in complex analytical systems (López-Córdova et al., 2025).

Conclusion

This research developed and evaluated a crime prediction system for Metropolitan Lima using machine learning models structured under the CRISP-DM methodology. The comparative analysis of four machine learning algorithms Random Forest, Support Vector Machine, Gradient Boosting, and XGBoost demonstrates that they can effectively predict the occurrence of crimes in urban environments.

The findings conclusively establish that XGBoost is the most effective model for crime prediction in this context, achieving superior performance across all evaluated metrics with an accuracy of 0.91, AUC of 0.96, F1 Score of 0.92, and, most importantly, a recall of 0.89. This detection rate of 89% of actual crimes, with only 11% false negatives, represents an important characteristic for public safety applications where failing to identify actual crimes can have serious consequences for preventive interventions and community safety.

The performance progression from Random Forest (recall: 0.77) and SVM (recall: 0.77) to Gradient Boosting (recall: 0.86) and XGBoost (recall: 0.89) demonstrates a clear trajectory of improvement, validating the hypothesis that boosting methods provide greater predictive accuracy and reliability. The consistently high AUC values of 0.96 for Gradient Boosting and XGBoost confirm their discriminatory capacity between high-risk and low-risk areas, which is fundamental for effective geospatial crime prediction.

The implementation of the CRISP-DM methodology proved essential for structuring the data mining process, from business understanding to model deployment, ensuring a systematic and reproducible approach to developing the predictive system.

This study validates that XGBoost, combined with the CRISP-DM methodology, represents an efficient, reliable, and practical approach for crime prediction in urban environments. The results demonstrate that data-driven predictive systems can play a vital role in modern crime prevention strategies and that the selection of appropriate algorithms, particularly gradient boosting methods over traditional approaches, significantly impacts the effectiveness of crime detection and prevention efforts.

Limitation and Future Work

Continued collaboration with security and technology experts is recommended to maintain the effectiveness and accuracy of the application. In addition, it is advisable to explore the feasibility of integrating emerging technologies and expanding data sets to improve the predictive capability of the application. Finally, it is suggested to conduct periodic evaluations and updates of the application to ensure its long-term relevance and effectiveness.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors received no financial support for the research, authorship, and or publication of this article.

Author’s Contributions

Hugo Vega Huerta, Javier Vilca Velasquez and Nicolas Anicama Espinoza: Conceptualization, formal analysis, software, resource, writing original draft.

Luis Guerra Grados and Jorge Pantoja Collantes: Methodology, project administration, visualization, writing review and edited.

Oscar Benito Pacheco: Investigation, software, methodology, writing review and editing.

Juan Carlos Lázaro Guillermo and Rubén Gil Calvo: Data curation, validation, supervision, writing review and edited.

Ethics

This manuscript adheres to all the ethical standards outlined in the publication policies of the journal.

References

Alves, L. G. A., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and Its Applications*, *505*, 435–443. https://doi.org/10.1016/j.physa.2018.03.084

Arrieta-Espinoza, E., Moquillaza-Henríquez, S., Vega-Huerta, H., Cámara-Figueroa, A., De-La-Cruz-Vdv, P., & Grados, L. G. (2023). Improvement to the Papeat software of the gastronomy sector based on Intelligence using Tableau. *Proceeding of the 18th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–7. https://doi.org/10.23919/cisti58278.2023.10211723

Bennett Moses, L., & Chan, J. (2018). Algorithmic Prediction in Policing: Assumptions, Evaluation and Accountability. *Policing and Society*, *28*(7), 806–822. https://doi.org/10.1080/10439463.2016.1253695

Bokrantz, J., Subramaniyan, M., & Skoogh, A. (2024). Realising the promises of artificial intelligence in manufacturing by enhancing CRISP-DM. *Production Planning & Control*, *35*(16), 2234–2254. https://doi.org/10.1080/09537287.2023.2234882

Brillianto Apribowo, C. H., Hadi, S. P., Wijaya, F. D., Bambang Setyonegoro, M. I., & Sarjiya. (2024). Early prediction of battery degradation in grid-scale battery energy storage system using extreme gradient boosting algorithm. *Results in Engineering*, *21*, 101709. https://doi.org/10.1016/j.rineng.2023.101709

Brzozowska, J., Pizoń, J., Baytikenova, G., Gola, A., Zakimova, A., & Piotrowska, K. (2023). Data Engineering In Crisp-Dm Process Production Data – Case Study. *Applied Computer Science*, *19*(3), 83–95. https://doi.org/10.35784/acs-2023-26

Chainey, S., Tompson, L., & Uhlig, S. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, *21*(1–2), 4–28. https://doi.org/10.1057/palgrave.sj.8350066

Cortina, V. G. (2015). Aplicación de la metodología crisp-dm a un proyecto de minería de datos en el entorno universitario. *Universidad Carlos III de Madrid*. https://e-archivo.uc3m.es/handle/10016/22198

Daniel, K. (2022). Security and Public Safety Application for a Mobile Device. *United States Patent Application Publication*, Article US20220322086A1 (depending on the exact filing).

DeLaCruzVdv, P., Moquillaza-Henríquez, S., Valeriano-Peña, M., Maquen-Niño, G. L. E., Vega-Huerta, H., & Adrianzén-Olano, I. (2023). Data Mart and key performance indicators to optimize decisions in a medical service clinic. *Proceeding of the 18th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/10.23919/cisti58278.2023.10211492

DelaCruz-VdV, P., Cadenillas-Rivera, D., Vega-Huerta, H., Cancho-Rodriguez, E., Bulnes, M. E. P., Maquen-Niño, G. L. E., & Pantoja-Collantes, J. (2023, June). Diagnosis of brain tumors using a convolutional neural network. In *International Conference in Information Technology and Education* (pp. 45-56). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-5414-8_6

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*, 115–125. https://doi.org/10.1016/j.dss.2014.02.003

IDL. (2021). *Instituto de defensa legal*. Https://Www.Idl.Org.Pe/. https://www.idl.org.pe/

Lie. (2017). Systems and Methods for Security Data Analysis and Display. *Official USPTO Patent Database / Patent Publication*, Article 1st.

López-Córdova, F., Vega-Huerta, H., Maquen-Niño, G. L. E., Cáceres-Pizarro, J., Adrianzén-Olano, I., & Benito-Pacheco, O. (2025). Construction of a New Data Set of Pleural Fluid Cytological Images for Research. *International Journal of Online and Biomedical Engineering (IJOE)*, *21*(07), 138–151. https://doi.org/10.3991/ijoe.v21i07.54323

Maquen-Niño, G. L. E., Bravo, J., Alarcón, R., Adrianzén-Olano, I., & Vega-Huerta, H. (2023a). Una revisión sistemática de Modelos de clasificación de dengue utilizando machine learning. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, *50*, 5–27. https://doi.org/10.17013/risti.50.5-27

Maquen-Niño, G. L. E., Sandoval-Juarez, A. A., Veliz-La Rosa, R. A., Carrión-Barco, G., Adrianzén-Olano, I., Vega-Huerta, H., & De-La-Cruz-VdV, P. (2023b). Brain Tumor Classification Deep Learning Model Using Neural Networks. *International Journal of Online and Biomedical Engineering (IJOE)*, *19*(09), 81–92. https://doi.org/10.3991/ijoe.v19i09.38819

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/tkde.2019.2962680

Meijer, A., & Wessels, M. (2019). Predictive Policing: Review of Benefits and Drawbacks. *International Journal of Public Administration*, *42*(12), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664

Mohler, G., Porter, M., Carter, J., & LaFree, G. (2020). Learning to rank spatio-temporal event hotspots. *Crime Science*, *9*(1), 3. https://doi.org/10.1186/s40163-020-00112-x

Melgarejo-Solis, R., Antón-Sancho, Á., Vega-Huerta, H., & Vergara-Rodríguez, D. (2023). Impact of COVID-19 on the use of ICT resources among university professors in Peru. *Proceedings of the 21th LACCEI International Multi-Conference for Engineering, Education and Technology (LACCEI 2023): "Leadership in Education and Innovation in Engineering in the Framework of Global Transformations: Integration and Alliances for Integral Development,"* 795. https://doi.org/10.18687/laccei2023.1.1.795

Mutale, B., Withanage, N. C., Mishra, P. K., Shen, J., Abdelrahman, K., & Fnais, M. S. (2024). A performance evaluation of random forest, artificial neural network, and support vector machine learning algorithms to predict spatio-temporal land use-land cover dynamics: a case from lusaka and colombo. *Frontiers in Environmental Science*, *12*, 1431645. https://doi.org/10.3389/fenvs.2024.1431645

Poh, C. Q. X., Ubeynarayana, C. U., & Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach. *Automation in Construction*, *93*, 375–386. https://doi.org/10.1016/j.autcon.2018.03.022

Putro, P. A. W., & Sensuse, D. I. (2022). Review of Security Principles and Security Functions in Critical Information Infrastructure Protection. *International Journal of Safety and Security Engineering*, *12*(4), 459–465. https://doi.org/10.18280/ijsse.120406

Rivera-Alvino, R., Vega-Huerta, H., Guzmán-Monteza, Y., Puelles-Bulne, M., Cancho-Rodríguez, E., Pantoja-Collantes, J., & Cruz-VDV, P. D.-L. (2023). Modelling, design and simulation using BPM to reduce the time of vehicle safety accessories manufacturing process. *RISTI Preprint Repository*, 310–325.

Sánchez-Tello, J., Vega-Huerta, H., De-La-Cruz-Vdev, P., Maquen-Niño, G., Melgarejo-Solis, R., Cámara-Figueroa, A., & Cancho-Rodriguez, E. (2023). Implementation of a chatbot for virtual attention of queries Case: Postgraduate School of the UNMSM. *Proceedings of the 21th LACCEI International Multi-Conference for Engineering, Education and Technology (LACCEI 2023): "Leadership in Education and Innovation in Engineering in the Framework of Global Transformations: Integration and Alliances for Integral Development,"* 787. https://doi.org/10.18687/laccei2023.1.1.787

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https://doi.org/10.1177/1536867x20909688

Sudar, L. Z. S., Imbenay, J. L., Budi, I., Ramadiah, A., Putra, P. K., & Santoso, A. B. (2024). Textual Analysis for Public Sentiment Toward National Police Using CRISP-DM Framework. *Revue d'Intelligence Artificielle*, *38*(1), 63–72. https://doi.org/10.18280/ria.380107

Vega-Huerta, H., Vilca Velasquez, J., Anicama Espinoza, N., Maquen-Niño, G. L. E., Guerra-Grados, L., Pantoja-Collantes, J., Benito-Pacheco, O., Lázaro-Guillermo, J. C., Camara-Figueroa, A., Cabrera-Díaz, J., Gil-Calvo, R., & López-Córdova, F. (2025). Mobile application based on KDD to predict high-crime areas and promote sustainability in citizen security in a district of Lima-Perú. *Frontiers in Computer Science*, *7*, 1585632. https://doi.org/10.3389/fcomp.2025.1585632

Velez-Villanueva, R., Vega-Huerta, H., De-La-Cruz, P., Gamboa-Cruzado, J., Cancho-Rodriguez, E., & Cámara-Figueroa, A. (2023). Support in the Neurorehabilitation for Older People Using a Mobile Application. *Human-Centered Intelligent Systems*, *694*, 85–96. https://doi.org/10.1007/978-981-99-1912-3_8

Wang, J., Hu, J., Shen, S., Zhuang, J., & Ni, S. (2020). Crime risk analysis through big data algorithm with urban metrics. *Physica A: Statistical Mechanics and Its Applications*, *545*, 123627. https://doi.org/10.1016/j.physa.2019.123627

Weisburd, D., Telep, C. W., Hinkle, J. C., & Eck, J. E. (2010). Is problem-oriented policing effective in reducing crime and disorder? *Criminology & Public Policy*, *9*(1), 139–172. https://doi.org/10.1111/j.1745-9133.2010.00617.x

Wheeler, A., & Reuter, S. (2020). Redrawing Hot Spots of Crime in Dallas, Texas. *CrimRxiv*, *24*(2), 159–184. https://doi.org/https://doi.org/10.1177/10986111209 57948

Williams, M. L., Burnap, P., & Sloan, L. (2016). Crime Sensing with Big Data: The Affordances and Limitations of using Open Source Communications to Estimate Crime Patterns. *British Journal of Criminology*, *57*(2), 320–340. https://doi.org/10.1093/bjc/azw031

Yauri, J., Lagos, M., Vega-Huerta, H., De-La-Cruz-VdV, P., Maquen-Ni˜no, G. L. E., & Condor-Tinoco, E. (2023). Detection of Epileptic Seizures Based-on Channel Fusion and Transformer Network in EEG Recordings. *International Journal of Advanced Computer Science and Applications*, *14*(5), 1–7. https://doi.org/10.14569/ijacsa.2023.01405110

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, *12*(1), 469–477. https://doi.org/10.1016/j.gsf.2020.03.007

Zhang, X., Liu, L., Xiao, L., & Ji, J. (2020). Comparison of Machine Learning Algorithms for Predicting Crime Hotspots. *IEEE Access*, *8*, 181302–181310. https://doi.org/10.1109/access.2020.3028420

Zheng, R., Jia, Y., Ullagaddi, C., Allen, C., Rausch, K., Singh, V., Schnable, J. C., & Kamruzzaman, M. (2024). Optimizing feature selection with gradient boosting machines in PLS regression for predicting moisture and protein in multi-country corn kernels via NIR spectroscopy. *Food Chemistry*, *456*, 140062. https://doi.org/10.1016/j.foodchem.2024.140062