Universidad de Lima

Facultad de Ingeniería

Carrera de Ingeniería de Sistemas

COMPARATIVA DE MODELOS DE REGRESIÓN A FIN DE PREDECIR EL CRIMEN EN ZONAS DE ALTO RIESGO DE LA CIUDAD DE LIMA

Tesis para optar el Título Profesional de Ingeniero de Sistemas

Maria Cielo Escobedo Neyra

Código 20172103

Cynthia Lizet Tapia Aquino

Código 20171524

Asesor

Juan Manuel Gutierrez Cardenas

Lima – Perú

Setiembre de 2024

COMPARATIVA DE MODELOS DE REGRESIÓN A FIN DE PREDECIR EL CRIMEN EN ZONAS DE ALTO RIESGO DE LA CIUDAD DE LIMA **(Escobedo, et al., 2024)**

Maria Cielo Escobedo Neyra
20172103@aloe.ulima.edu.pe
Universidad de Lima

Cynthia Lizet Tapia Aquino
20171524@aloe.ulima.edu.pe
Universidad de Lima

Resumen: La delincuencia sigue siendo un problema en Lima Metropolitana, Perú, que afecta a la sociedad. Este artículo tiene como objetivo analizar los delitos contra la propiedad y reconocer la falta de estudios para predecir estos crímenes. Para solucionar este problema, se utilizan técnicas de regresión como Extra Tree, XGBoost, Bag, AdaBoost, Support Vector y Random Forest. Mediante GridSearchCV se optimizan los hiperparámetros para mejorar los resultados de la investigación. El modelo de Extra Tree Regression muestra un coeficiente de determinación (R2) de 0,79, y se evalúan métricas de error como el error cuadrático medio de la raíz (MSE), el error cuadrático medio (RMSE) y el error absoluto medio (MAE). Este enfoque considera patrones temporales de delincuencia para resolver la incertidumbre y combatir la inseguridad ciudadana.

Palabras clave: aprendizaje de máquina; modelos de regresión; crímenes; predicción.

Abstract: In Metropolitan Lima, Peru, crime is still an issue that has an impact on society. The purpose of this paper is to examine property crimes and acknowledge the paucity of research on their predictiveness. Regression approaches including Extra Tree, XGBoost, Bag, AdaBoost, Support Vector, and Random Forest are utilized to overcome this issue. The study outcomes are enhanced by optimizing the hyperparameters with GridSearchCV. Error measures including mean absolute error (MAE), root mean square error (RMSE), and root mean square error (MSE) are assessed for the Extra Tree Regression model, which has a coefficient of determination (R2) of 0.79. This strategy resolves ambiguity and counteracts citizen uneasiness by taking temporal trends of crime into account.

Keywords: machine learning; regression models; crime; prediction.

Línea de investigación IDIC – ULIMA

Área: Productividad y Empleo
Línea: Innovación: tecnologías y productos.

Área y Sub-áreas de Investigación:

Área: Aplicaciones en inteligencia artificial
Línea: Aprendizaje Automático

Objetivo (s) de Desarrollo Sostenible (ODS)

Industria, Innovación e Infraestructura

1. PLANTEAMIENTO DEL PROBLEMA

Diversas fuentes coinciden en que la delincuencia es un fenómeno en constante crecimiento que afecta negativamente a la economía y a la calidad de vida en general (Adel et al., 2016; Bogomolov et al., 2014; Kawthalkar et al., 2020). La delincuencia está presente en todos los sistemas sociales y sus efectos se reflejan en distintos continentes, países y regiones.

Según Numbeo (2023), que realizó estimaciones globales de las tasas de criminalidad en 142 países, Perú ocupa el undécimo lugar con un 68 % de criminalidad y un 32 % de seguridad. Uno de los mayores problemas de Perú es la delincuencia, y el número de denuncias aumenta constantemente.

En 2022, el departamento de Lima concentró la mayor cantidad de denuncias de delitos, representando el 34 % del total, con un total de 44 879 delitos en el área metropolitana. La mayoría de los delitos corresponden a delitos contra la propiedad, representando el 74 %. Estos incluyen hurtos y daños a bienes. Asimismo, pueden clasificarse en tipos en función de su naturaleza, como asaltos con vehículos implicados, hurtos graves como robos nocturnos y robos en casas ocupadas sin autorización, robo de vehículos, tentativas de robo, casos de robo a mano armada con circunstancias agravantes implicadas, robos relacionados con bandas y tentativas de robo (Instituto Nacional de Estadística e Informática [INEI], 2019-2020).

La carencia de educación superior y la inestabilidad laboral son factores clave que afectan negativamente tanto al crimen como a la economía nacional. El 90 % de los detenidos son hombres de entre 18 y 59 años, con educación básica y condiciones laborales inestables (INEI, 2020).

En los últimos años, los delitos contra la propiedad se han incrementado sobre manera en zonas de riesgo del área metropolitana de Lima. Esto generó preocupaciones sobre la seguridad y el bienestar de los residentes (Jaitman, 2017). Según el Instituto Nacional de Estadística e Informática (INEI, 2010), estos crímenes tienen un impacto que va más allá de las víctimas inmediatas. También afectan a la economía y a la confianza de los ciudadanos en las regiones. La investigación se centra en métricas y busca usar modelos de regresión para ayudar a la policía a tomar decisiones informadas y abordar eficazmente el problema de los delitos contra la propiedad.

2. OBJETIVO

Después de comparar modelos de regresión como, AdaBoost, Extra Tree, XGBoost, Bagging, Support Vector y Random Forest, se propuso identificar el modelo más eficaz basado en métricas de evaluación como fueron el error cuadrático medio (MSE), error absoluto medio (MAE), el error cuadrático medio (RMSE) y el R^2. El modelo mejor seleccionado podría ayudar a los organismos encargados de hacer cumplir la ley en sus esfuerzos por reducir los delitos contra la propiedad y crear un entorno más seguro para los residentes de la ciudad metropolitana de Lima.

3. JUSTIFICACIÓN

La razón para utilizar modelos de regresión es que permiten analizar la relación entre factores temporales que afectan las tasas de criminalidad. Estos modelos utilizan una estructura de ventana de tiempo e incluyen variables mediante técnicas de organización de datos y selección de características que proporcionan un enfoque sistemático para comprender y predecir patrones delictivos. El principal objetivo no es hacer predicciones, sino proporcionar a las fuerzas de seguridad información valiosa que les ayude a distribuir los recursos y a planificar eficazmente las estrategias de intervención.

Es importante subrayar que los métodos tradicionales de solución de problemas sociales en Perú han enfrentado retos al integrarse a la vida y promover la cooperación (INEI, 2021). El objetivo es reducir esta brecha mediante el uso de modelos retrospectivos para desarrollar soluciones que no solo se centren en las experiencias de los residentes, sino que también fomenten su participación en la mejora de la seguridad social.

4. DISEÑO METODOLÓGICO

En este estudio, se utiliza un enfoque para examinar y analizar datos criminales empleando los métodos de investigación que se muestran en la Figura 1. La primera fase se centra en recopilar antecedentes penales de fuentes confiables y convertir estos registros en un formato adecuado para su uso en el conjunto de datos. Este importante paso garantiza la calidad y confiabilidad de los datos analizados.

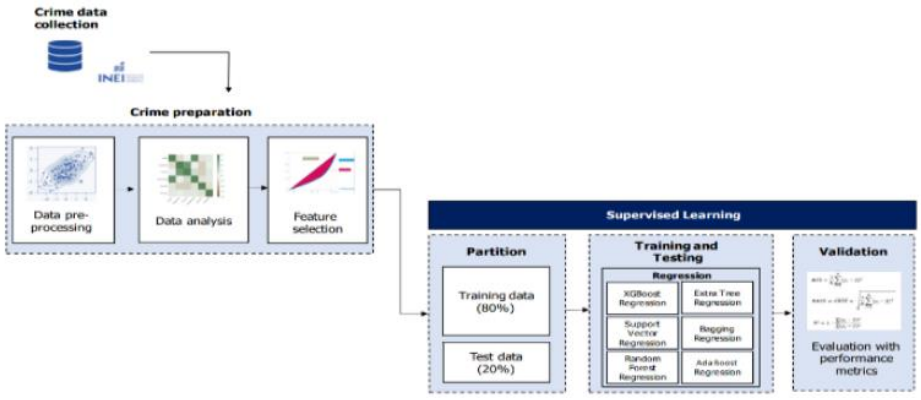

Figura 1. Metodología de investigación para la predicción de delitos contra el patrimonio en Lima (Escobedo et al., 2024)

La segunda fase, preparación de la delincuencia, comprende tres etapas: preprocesamiento de datos, análisis de datos y selección de características. El preprocesamiento de datos ayuda a limpiar el conjunto de datos eliminando filas en blanco y utilizando técnicas de normalización. Luego, el análisis de datos es fundamental para comprender los patrones y tendencias de los datos sobre delitos. Las técnicas de análisis de datos exploratorios ayudan a encontrar valores atípicos y otros factores importantes que pueden influir en las actividades delictivas. Además, durante esta fase se lleva a cabo la selección de características con el fin de mostrar las variables más relevantes

para construir modelos sólidos de aprendizaje supervisado. Para dicho propósito, se evalúa la importancia de cada característica en la predicción de la delincuencia, lo que aumenta la precisión y la eficacia del modelo. Por otro lado, la incorporación de métodos de ventanas móviles y pronósticos semanales añade más profundidad temporal al análisis.

Tras la preparación completa del conjunto de datos, la investigación pasa a la tercera fase, que se centra en el aprendizaje supervisado. Esta se divide en tres pasos fundamentales: partición, entrenamiento y pruebas, y validación. En el primer paso, el conjunto de datos se divide en un 80 % de entrenamiento y un 20 % de pruebas, lo que constituye una base sólida para la evaluación del modelo. El segundo paso es entrenar y probar modelos de regresión supervisados, incluidos Extra Tree, Support Vector, XGBoost, Bagged, AdaBoost y Random Forest, siguiendo las referencias de Silva et al. (2020), Wang et al. (2020), Belesiotis et al. (2018), Cavadas et al. (2015) y Khan y Salim (2019). Se realiza una optimización de hiperparámetros de cada modelo para mejorar el rendimiento general.

En el tercer paso, las predicciones del modelo de regresión se evalúan minuciosamente utilizando métricas de rendimiento, como el MSE, el RMSE, el R2 y el MAE (McClendon y Meghanathan, 2015; Ingilevich y Ivanov, 2018; Saltos y Cocea, 2017). Este meticuloso proceso de evaluación tiene como objetivo decidir qué modelo es el más eficaz para proporcionar información a las fuerzas y cuerpos de seguridad a la hora de abordar sus actividades.

AGRADECIMIENTOS

La culminación de esta tesis ha sido un viaje lleno de aprendizajes, retos y satisfacciones compartidas, y no habría sido posible sin el apoyo y la colaboración de numerosas personas a quienes queremos expresar nuestro más sincero agradecimiento.

En primer lugar, queremos agradecer a nuestro director de tesis, al Dr. Juan Gutiérrez, por su invaluable guía y apoyo a lo largo de este proceso de forma constante. Su experiencia y conocimiento han sido pilares fundamentales para el desarrollo de esta investigación.

También queremos agradecer al Dr. Víctor Ayma, por sus valiosos sugerencias y comentarios que han contribuido significativamente a mejorar la calidad de esta tesis.

REFERENCIAS

Adel, H., Salheen, M., & Mahmoud, R. (2016). Crime in relation to urban design. case study: the greater cairo region. Ain Shams Engineering Journal, 7, 925–938.

Belesiotis, A., Papadakis, G., & Skoutas, D. (2018). Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Trans. Spatial Algorithms Syst.*, *3*(4), https://dl.acm.org/doi/10.1145/3190345.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). Once upon a crime: towards crime prediction from demographics and mobile data. *ICMI*, 427–434. https://doi.org/10.1145/2663204.2663254.

Cavadas, B., Branco, P., & Pereira, S. (2015). Crime prediction using regres- sion and resources optimization. *EPIA* , 513–524.

Escobedo, M., Tapia, C., Gutiérrez, J., & Ayma, V. (2024). Comparing regression models to predict property crime in high-risk lima districts. *(IJACSA) International Journal of Advanced Computer Science and Applications, 15*(3), 62-68.

Ingilevich, V., & Ivanov, S. (2018). Crime rate prediction in the urban environment using social factors. *Procedia Computer Science, 136*, 472–478.

Instituto Nacional de Estadística e Informática [INEI]. (2010). *Victimización en el Perú 2010-2019. Principales Resultados.* https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1730/Libro.pdf

Instituto Nacional de Estadística e Informática [INEI]. (2019-2020). *Principales indicadores de seguridad ciudadana a nivel regional.* https://www.inei.gob.pe/media/MenuRecursivo/boletines/estadisticas-de-seguridad-ciudadana-regional-nov19-abr20.pdf

Instituto Nacional de Estadística e Informática [INEI]. (2021). *Estadísticas de las tecnologías de información y comunicación en los hogares.* https://www.inei.gob.pe/media/MenuRecursivo/boletines/01-informe-tecnico-tic-iv-trimestre-2020.pdf

Instituto Nacional de Informática y Estadística [INEI]. (2020). *Informe técnico - Estadísticas de seguridad ciudadana.* http://m.inei.gob.pe/media/MenuRecursivo/boletines/ boletin-de-seguridad-ciudadana.pdf

Jaitman, L. (2017). *Los costos del crimen y de la violencia. Nueva evidencia y hallazgos en América Latina y el Caribe.* Banco Interamericano de Desarrollo [BID].

Kawthalkar, I., Jadhav, S., Jain, D., & Nimkar, A. (2020). A survey of predictive crime mapping techniques for smart cities," 2020 National Conference on Emerging Trends on Sustainable Technology and Engi- neering Applications. *NCETSTEA, 2,* https://ieeexplore.ieee.org/document/9119948.

Khan, S., & Salim, F. (2019). *Crime Rate Prediction with Region Risk and Movement Patterns.* https://arxiv.org/pdf/1908.02570

McClendon, L., & Meghanathan, N. (2015). Using machine learning algo- rithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ), 2*(1), 1–12.

Numbeo. (2023). *Crime index by country 2023*. https://www.numbeo.com/crime/rankings

Saltos, G., & Cocea, M. (2017). An exploration of crime prediction using data mining on open data. *International Journal of Information Technology & Decision Making*, *16*(05), 1155–1181.

Silva, J., Romero, L., González, R., Larios, O., Barrantes, F., Lezama, O., & Manotas, A. (2020). *Algorithms for crime prediction in smart cities through data mining.* https://repositorio.cuc.edu.co/handle/11323/7743

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://www.sciencedirect.com/science/article/pii/ .

ANEXOS

Datos del artículo publicado

- Nombre del artículo: Comparing Regression Models to Predict Property Crime in High-Risk Lima Districts
- Autores: Maria Escobedo, Cynthia Tapia
- Co autor(es): Juan Gutierrez, Victor Ayma

Publicación en revista

- Nombre de la revista: International Journal of Advanced Computer Science and Applications (IJACSA)
- Volumen: 15
- Número: 3
- Año: 2024
- Pp:3
- Enlace web donde se encuentra publicado el artículo (identificador DOI, ISBN, ISSN o equivalentes): https://dx.doi.org/10.14569/IJACSA.2024.0150307

articulo

INFORME DE ORIGINALIDAD

13%	12%	3%	5%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.ulima.edu.pe Fuente de Internet	2%
2	Submitted to Universidad de Lima Trabajo del estudiante	2%
3	ichi.pro Fuente de Internet	1%
4	repositorio.continental.edu.pe Fuente de Internet	1%
5	w.thescipub.com Fuente de Internet	1%
6	Submitted to Universidad Cesar Vallejo Trabajo del estudiante	1%
7	eprints.ucm.es Fuente de Internet	1%
8	www2.uca.edu.ar Fuente de Internet	1%
9	ohchr.org Fuente de Internet	1%