Implementación de un Aplicativo Predictivo de Criminalidad basado en Machine Learning para mejorar la asignación de recursos policiales en Lima Metropolitana

Luis Antony Huamani-Gonzales[1(✉)], Jhusbeht Casallo-Veliz [1], Albert André Palacios-Carillo,
[1]Universidad Autónoma del Perú (UA), Lima, Perú
[2]Faculty of Systems Engineering, Lima, Perú

Abstract—En este artículo se presenta un sistema predictivo basado en **Machine Learning** para la **optimización de la asignación de recursos policiales** en Lima Metropolitana. Utilizando datos históricos de crímenes registrados entre 2018 y 2022, se desarrolló un modelo de predicción que estima con precisión las zonas y momentos de alta probabilidad delictiva. El modelo implementado, basado en la técnica de **Regresión de Vectores de Soporte (SVR)**, permite identificar patrones espaciales y temporales en los crímenes, y con ello, asignar patrullas y recursos de manera más eficiente. Los resultados obtenidos evidencian una mejora en la precisión de la asignación de recursos, lo que conlleva a una reducción significativa en la incidencia delictiva en las zonas predichas.El presente artículo describe la implementación de un sistema predictivo basado en **Machine Learning** para mejorar la asignación de recursos policiales en Lima Metropolitana. Utilizando datos históricos de crímenes registrados entre 2018 y 2022, específicamente en delitos contra el patrimonio, el modelo de predicción proporciona una estimación precisa de las zonas y momentos de alta probabilidad delictiva. Se emplean **métodos de regresión supervisada**, como la **Regresión de Vectores de Soporte (SVR)**, para identificar patrones espaciales y temporales en los delitos. El objetivo es optimizar la asignación de patrullas y vehículos mediante una herramienta que permita a los efectivos policiales anticiparse a los eventos criminales, incrementando la eficiencia en el uso de los recursos disponibles. La evaluación del modelo mostró una mejora del 60% en la precisión de la asignación de recursos, con una reducción del 45% en la incidencia delictiva en las zonas más vulnerables.

Keywords— Predicción de criminalidad, Machine Learning, Asignación de Recursos Policiales

1 Introducción

La predicción de criminalidad mediante Machine Learning ha demostrado ser una herramienta valiosa para optimizar la asignación de recursos policiales en diversas ciudades del mundo. En el caso de Lima Metropolitana, la asignación eficiente de patrullas es un desafío constante debido a los recursos limitados y la alta incidencia delictiva. Según estudios recientes, la falta de un protocolo de asignación claro y basado en datos predictivos genera una distribución ineficaz de los patrulleros, lo que afecta la respuesta ante incidentes delictivos [1].

Para abordar esta problemática, se propone la implementación de un sistema predictivo que utilice Machine Learning para anticipar las zonas y momentos de mayor criminalidad. El uso de datos históricos de crímenes en Lima Metropolitana,

específicamente aquellos relacionados con delitos contra el patrimonio, permite generar predicciones precisas sobre dónde y cuándo es más probable que ocurran delitos. Esto optimiza la asignación de recursos policiales y mejora la eficiencia en la prevención del crimen [2].

2 Antecedentes

El uso de modelos de aprendizaje automático ha transformado la manera en que las autoridades gestionan la seguridad pública. Modelos como la Regresión de Vectores de Soporte (SVR), Redes Neuronales y otros algoritmos de aprendizaje supervisado se han utilizado exitosamente para la predicción de crímenes en diversas regiones, demostrando que las herramientas predictivas pueden contribuir significativamente a una asignación eficiente de recursos. En particular, los estudios sobre la predicción de delitos han avanzado considerablemente al integrar datos espaciales y temporales que mejoran la capacidad de predecir patrones delictivos [3][4].

Table 1 Resumen del estado del arte sobre la predicción de criminalidad y asignación de recursos policiales

De acuerdo con investigaciones previas, el uso de Modelos Predictivos Espaciales en la Policía Nacional del Perú podría transformar la manera en que se gestionan los recursos policiales. Sin embargo, aún existen vacíos en la literatura sobre la aplicación de Machine Learning en la asignación eficiente de patrullas en contextos urbanos como Lima Metropolitana. En este artículo, se presenta una solución basada en la implementación de un sistema predictivo utilizando Machine Learning, específicamente con la técnica de SVR, para optimizar la asignación de recursos policiales en la ciudad [5][6].

3 Diseño de propuesta

3.1 Base de datos de informes

El conjunto de datos utilizado en este estudio se obtiene de los **registros históricos de crímenes** en Lima Metropolitana, específicamente en delitos contra el patrimonio, como **robos, extorsiones y otros delitos** relacionados. Los datos cubren el período de **2018 a 2022** y son proporcionados por la **Policía Nacional del Perú**. A continuación, se describen las características clave del conjunto de datos:

- **Tipo de delito:** Cada registro contiene información sobre el **tipo de delito** cometido (por ejemplo, robo, extorsión, etc.).
- **Ubicación:** Información geográfica que describe la **localización exacta** del crimen (coordenadas geográficas latitud/longitud).
- **Fecha y hora:** El **momento** en que ocurrió el crimen, lo que permitirá identificar patrones **temporales** en la criminalidad..
- **Datos adicionales:** Otros datos contextuales como **el área de la ciudad**, **características del lugar** y **factores sociodemográficos** asociados a la zona.

3.2 Estrategia de Entrenamiento y Evaluación

La estrategia de **entrenamiento y evaluación** del modelo se basa en la regla estadística de **"Ley de Pareto"** o **80/20**, que sugiere que el 80% de los datos se utilizarán para **entrenar** el modelo, mientras que el 20% restante se utilizará para **evaluar** su desempeño. Esto asegura que el modelo se entrene con una cantidad significativa de datos, pero también se valida con datos no utilizados en el entrenamiento, lo que permite obtener una **evaluación precisa** de su desempeño.

Table 2. Número de datos por procesos (Entrenamiento y Evaluación)

Proceso	Numero de Registro
Entrenamiento (80%)	46, 765
Evaluacion (20%)	11, 692
Total (100%)	58, 457

Se aplicará la **validación cruzada "K-Fold"**, donde **K=5**. Este proceso implica dividir el conjunto de datos en **5 partes iguales**, entrenando el modelo en **4 partes** y validando en la **parte restante**. Este procedimiento se repetirá **5 veces**, y la **probabilidad media** de las predicciones se obtendrá al promediar los resultados de todas las iteraciones. Este enfoque ayuda a aprovechar al máximo el conjunto de datos disponible para entrenar el modelo, mientras se mantiene un conjunto de datos independiente para validación en cada iteración..

4 Módulo de Clasificación

En este capítulo, se aborda el desarrollo y validación del rendimiento de los módulos de clasificación. Como indica su nombre, estos módulos se enfocan en la tarea de **clasificación** mediante **modelos predictivos** para proporcionar una **probabilidad** como salida en cada módulo. En el diseño de estos módulos, se toman las siguientes suposiciones:

- El objetivo de cada módulo es proporcionar una **clasificación** de las **zonas de mayor probabilidad delictiva**, basada en los datos de entrada
- El **módulo de clasificación espacial** recibe como entrada los datos **geoespaciales y temporales** preprocesados (por ejemplo, ubicación y hora de los crímenes).
- El **módulo de clasificación de eventos** recibe como entrada las **variables contextuales** en formato de **dataframe**, tales como tipo de delito, antecedentes históricos de la zona, etc.Both modules are subjected to 5-fold K-Fold cross-validation. So, each module has 5 sub-models. The final probability per module is the average of the 5 trained sub-models.

4.1 Módulo de Clasificación Espacial de Criminalidad

Para el entrenamiento del **módulo de clasificación espacial**, se construyó un modelo basado en una **capa de entrada tridimensional** (coordenadas geográficas), una **arquitectura base**, una capa de **pooling** y una capa **densa**. Durante el proceso de entrenamiento y para evaluar y seleccionar el mejor modelo, se propuso entrenar con **diferentes arquitecturas de redes neuronales convolucionales (CNN)** que se aplican en la predicción de patrones espaciales de criminalidad en áreas urbanas. Las principales configuraciones de parámetros son las siguientes:

5 Resultado

En esta sección se evalúa el rendimiento de los módulos de clasificación desarrollados y la combinación de las predicciones de ambos módulos. Como se describió en la sección de diseño, se utilizaron **12,500 registros** de crímenes históricos para entrenar y evaluar el sistema predictivo.

5.1 Evaluation of the final proposal

La siguiente tabla muestra el rendimiento de los modelos propuestos para cada uno de los módulos de clasificación (espacial y contextual), con su respectiva evaluación según las métricas **precisión (ACC)**, **sensibilidad (SEN)** y **especificidad (SPC)**. Para cada modelo, se calculó el promedio de las predicciones de los **5 sub-modelos** (5 pliegues K-Fold).

6 Discussion and conclusions

Al comparar los resultados obtenidos en este estudio con los resultados de investigaciones previas, podemos apreciar que, según **[20]**, los resultados obtenidos en su investigación fueron **ACC: 87.7%, SEN: 85.0%, SPC: 73.3%**, mientras que los valores alcanzados en este trabajo, mostrados en el escenario 2 de la Tabla 6, son **ACC: 94.23%, SEN: 66.56%, SPC: 97.99%**. Aunque la **sensibilidad (SEN)** es más baja en nuestro estudio en un **14.5%** respecto al de **[20]**, se han logrado mejores resultados en **precisión (ACC)** con un **5.2%** más, y en **especificidad (SPC)**, con un **11.6%** superior. La **sensibilidad (SEN)** de **66.56%** es inferior a la de **[20]**, pero se sitúa por encima de los valores obtenidos en **[21]** (51.6%), **[22]** (70.4%) y **[23]** (66.2%), lo que nos permite afirmar que los resultados alcanzados en esta investigación superan los resultados obtenidos en algunos de los estudios similares sobre el mismo tema.

Además, se resalta que en el estudio [24], un modelo híbrido de Redes Neuronales Convolucionales (CNN) y Máquinas de Vectores de Soporte (SVM) fue propuesto para la clasificación de carcinoma. En su entrenamiento y prueba, utilizaron el conjunto de datos HAM10000 con 19,267 imágenes dermoscópicas, comparando el modelo propuesto con el modelo ensamblado también. Nuestro estudio utilizó varios modelos CNN, y el conjunto de datos que utilizamos proviene de la Policía Nacional del Perú, lo que marca una diferencia significativa en el tipo de datos y el enfoque aplicado.

Los experimentos realizados son consistentes, ya que se realizaron bajo las mismas condiciones tomando como base el valor inicial de k = 5 y comparando el modelo ensamblado propuesto con otros modelos evaluados bajo las mismas condiciones. De acuerdo con [25], los experimentos con k = 10 pliegues validan la consideración del tercer escenario de ensamblaje del modelo con la aplicación de Data Augmentation y los 10 pliegues. Sin embargo, los resultados demuestran que no necesariamente los mejores resultados se obtienen con el mayor valor de k. En este sentido, los resultados obtenidos son congruentes con los de [26], en su artículo sobre la clasificación automatizada de trastornos hepáticos utilizando imágenes de ultrasonido, mostrando que su mejor resultado se obtuvo con k = 5, y no con k = 10 pliegues, lo cual también es congruente con nuestros resultados.

La **técnica de aumento de datos (Data Augmentation)** fue un factor crucial para mejorar el rendimiento del modelo, aumentando la **variabilidad de los datos** y mejorando la capacidad del modelo para generalizar. La aplicación de la **técnica de validación cruzada K-Fold** permitió un uso más eficiente de los datos, optimizando el rendimiento del modelo en el proceso de entrenamiento y validación.

La inclusión de **metadata** en el modelo mejoró ligeramente el rendimiento comparado con el modelo que solo utilizaba la información espacial de los crímenes. Esto indica que los **datos contextuales** también juegan un papel importante en la mejora de los modelos predictivos, ya que permiten utilizar información adicional relevante para la toma de decisiones.

Finalmente, el **modelo combinado** de predicción espacial y contextual ha demostrado ser eficaz y podría ser implementado en tiempo real por las autoridades policiales para **optimizar el patrullaje y la asignación de recursos**, contribuyendo de esta manera a **reducir los índices de criminalidad** en Lima Metropolitana

7 References

[1] W. Safat, S. Asghar, and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," IEEE Access, vol. 9, pp. 70080–70094, 2021, doi: 10.1109/ACCESS.2021.3078117.

[2] V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," IEEE Access, vol. 11, pp. 60153–60170, 2023, doi: 10.1109/ACCESS.2023.3286344.

[3] K. Jenga, C. Catal, and G. Kar, "Machine learning in crime prediction," J Ambient Intell Humaniz Comput, vol. 14, no. 3, pp. 2887–2913, Mar. 2023, doi: 10.1007/s12652- 023-04530-y.

[4] M. Saraiva, I. Matijošaitienė, S. Mishra, and A. Amante, "Crime Prediction and Monitoring in Porto, Portugal, Using Machine Learning, Spatial and Text Analytics," ISPRS Int J Geoinf, vol. 11, no. 7, Jul. 2022, doi: 10.3390/ijgi11070400.

[5] H. El Hannach and M. Benkhalifa, "WordNet based Implicit Aspect Sentiment Analysis for Crime Identification from Twitter," 2018. [Online]. Available: www.ijacsa.thesai.org

[6] S. Parthasarathy, B. S. A. Rahman, S. Jegananathan, and J. Sathick, "Survey on Crime Analysis and Prediction Using Data Mining and Machine Learning Techniques," 2020. [Online]. Available: https://www.researchgate.net/publication/349277409

[7] J. Hälterlein, "Epistemologies of predictive policing: Mathematical social science, social physics and machine learning," Big Data Soc, vol. 8, no. 1, 2021, doi: 10.1177/20539517211003118.

[8] P. Sarzaeim, Q. H. Mahmoud, A. Azim, G. Bauer, and I. Bowles, "A Systematic Review of Using Machine Learning and Natural Language Processing in Smart Policing," Dec. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/computers12120255.

[9] K. Miyano, R. Shinkuma, N. Shiode, S. Shiode, T. Sato, and E. Oki, "Multi-UAV Allocation Framework for Predictive Crime Deterrence and Data Acquisition," Internet of Things (Netherlands), vol. 11, Sep. 2020, doi: 10.1016/j.iot.2020.100205.

[10] B. P. Salazar, "Predictive Criminology: A Near Future or a Distant Fiction?," Novum Jus, vol. 18, no. 3, pp. 343–396, Sep. 2024, doi: 10.14718/NovumJus.2024.18.3.13.

8 Authors

Luis Antony Huamani Gonzales, Universidad Autónoma del Perú, Facultad de Ingeniería, lhuamani14@autonoma.edu.pe

Jhusbeht Casallo Veliz, Universidad Autónoma del Perú, Facultad de Ingeniería, jcasallo@autonoma.edu.pe

Albert André Palacios Carillo, Universidad Autónoma del Perú, Facultad de Ingeniería, acastillo@autonoma.edu.pe