Exploración del POLE dataset y aplicación de aprendizaje relacional

John Ureña

12 de enero de 2024

Resumen

This study analyzes the POLE dataset from the Manchester Police using Neo4j and Cypher, exploring its use in crime resolution and police resource allocation. It also examines the dataset's applicability in marketing security products. Using Scikit-learn and Random Forest, the study predicts individual criminal involvement based on centrality measures. Ethical considerations of data analysis in crime prevention are briefly discussed.

1. Introduction

Este trabajo presenta un análisis del dataset POLE, proporcionado por Neo4j, utilizando la base de datos relacional Neo4j y el lenguaje de consulta Cypher. Se explora la utilidad de la información extraída en la resolución de crímenes y la optimización de recursos policiales. Además, se examina cómo estos datos pueden aplicarse en estrategias de marketing para productos de seguridad personal, como cámaras de seguridad y sistemas de alarma. Utilizando Scikit-learn y un modelo de Random Forest, se intenta predecir la implicación criminal de individuos basándose en medidas de centralidad, ampliadas con parámetros específicos del contexto del problema. Finalmente, se discuten brevemente las consideraciones éticas y morales del uso de tecnologías de análisis de datos en contextos de seguridad y prevención del crimen.

2. El conjunto de datos POLE

2.1. Estructura

El conjunto de datos denominado POLE, acrónimo de Personas, Objetos, Localizaciones y Eventos, es una estructura habitualmente empleada en investigaciones policiales. Para nuestra investigación específica, utilizaremos un conjunto de datos de ejemplo incluido en Neo4j, accesible en: [dataset POLE de Neo4j]. Este conjunto de datos se originó a partir de registros de la policía de Manchester. Sin embargo, ha sido modificado para excluir identificadores personales, garantizando así la privacidad. Para una descripción más en detalle del dataset, recomiendo consultar el blog del creador: [Investigaciones POLE con Neo4j] [Dep18].

El conjunto de datos abarca un total de 61,522 registros, de los cuales 28,762 corresponden a crímenes. Sin embargo, solo 55 de estos registros cuentan con un perpetrador asociado. Esta limitada cantidad de datos vinculados a los autores de los crímenes representa un desafío significativo para la predicción de patrones delictivos. Un ejemplo de esta dificultad es la tarea de determinar si una persona nueva en el grafo está involucrada en un crimen. A pesar de la escasez de datos en este aspecto, nos proponemos explorar este caso como una muestra de lo que es posible alcanzar en el análisis de conjuntos de datos con características similares.

Un aspecto notable del conjunto de datos es la alarmante prevalencia de crímenes relacionados con violencia sexual, como se detalla en la tabla 1. Este fenómeno destaca como un área crítica que requiere atención urgente para reducir su incidencia. Otro patrón que llama la atención es el relacionado con los robos. Estos crímenes tienden a seguir un patrón que se podría analizar mediante un proceso de Hawkes. La agrupación de robos en términos espaciotemporales a menudo se debe a la tendencia de los delincuentes a reincidir en propiedades previamente atacadas o en sus inmediaciones, según

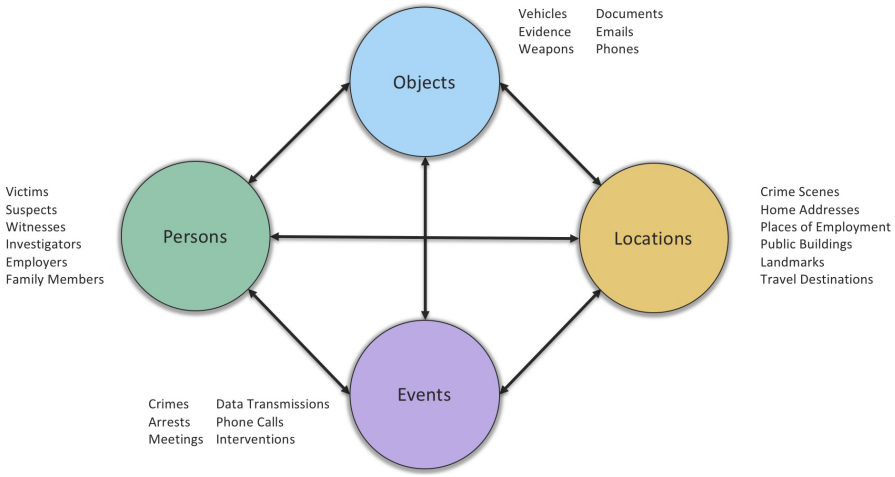

Figura 1: Estructura de un dataset tipo POLE [Dep18].

lo documentado en [MBB+21]. Sin embargo, en esta investigación, no profundizaremos en este tipo de análisis. El enfoque de nuestro trabajo es el aprendizaje relacional, y el patrón de robos podría abordarse de manera no relacional utilizando un Proceso de Hawkes Neuronal.

Crime type	**Total**
Violence and sexual offences	8765
Public order	4839
Criminal damage and arson	3587
Burglary	2807
Vehicle crime	2598
Other theft	2140
Shoplifting	1427
Other crime	651
Robbery	541
Theft from the person	423
Bicycle theft	414
Drugs	333
Possession of weapons	236

Cuadro 1: Total de crimenes por tipo

2.2. Aprovechando el poder de las bases de datos relacionales en la investigación de crímenes.

Digamos que la comunidad quiere dirigir una campaña publicitaria en contra de la violencia doméstica y tiene recursos limitados, así que desea investigar cuáles son las áreas con más personas propensas a sufrir violencia doméstica. En primer lugar, se selecciona una zona vulnerable debido a la cantidad de personas que viven en dicha área y tienen una conexión de tres saltos o menos con amigos que han incurrido en crímenes del tipo de violencia sexual.

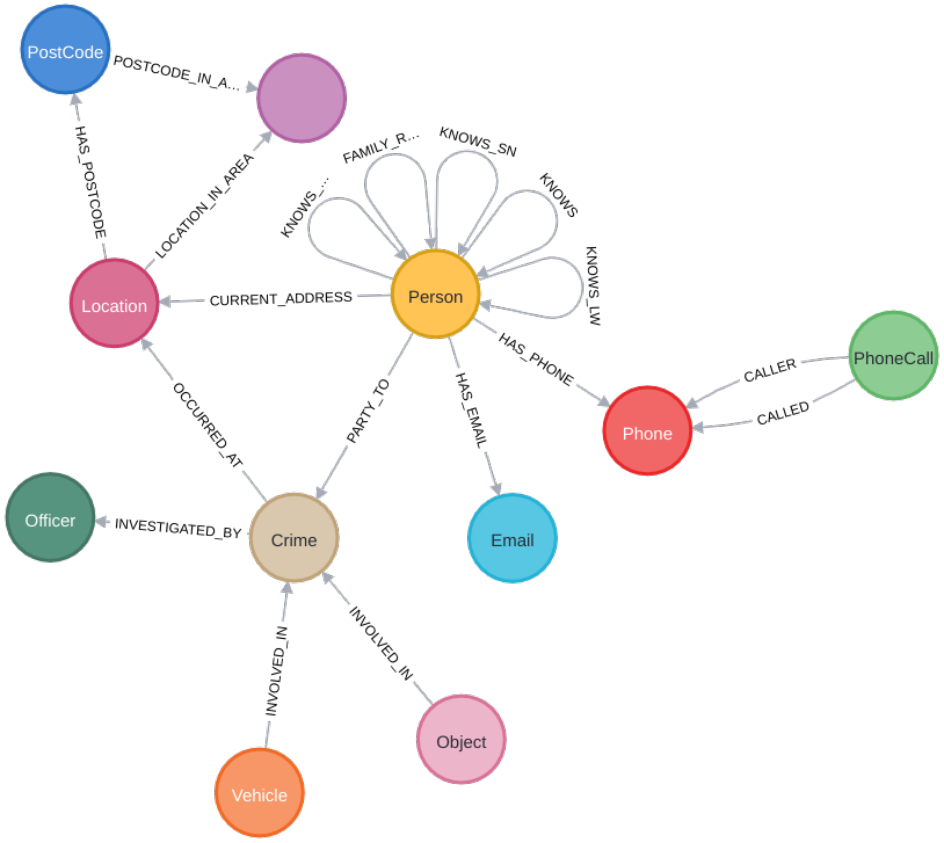

Figura 2: Estructura del dataset investigado.

```
MATCH  (a:Area)-[:LOCATION_IN_AREA]-(l:Location)<-[:CURRENT_ADDRESS]-(
   p:Person)-[:KNOWS*1..3]-(cp:Person)-[:PARTY_TO]->(:Crime{type: "
   Violence and sexual offences"})
WHERE NOT (p:Person)-[:PARTY_TO]->(:Crime)
RETURN a.areaCode AS area_code, count(distinct cp) AS dangerous_people
ORDER BY dangerous_people DESC
LIMIT 5;
```

Codigo 1: Query para encontrar areas con mas personas propensas a realizar crimenes domesticos.

Otra forma de lograr el mismo objetivo es ver las zonas donde ocurren crimenes sexuales con mayor frecuencia de los ultimos 6 meses.

Codigo de area	**Total**
M1	249
BL1	242
M40	234
BL3	223
BL9	204

Cuadro 2: Areas con mayor cantidad de crimenes sexuales de ultimo mes.

Tambien podemos hacer uso de la visualización en Neo4j para extraer conocimiento, las crimenes de drogas siempre suelen ocurrir dentro de cierta determinada estructura jerárquica, al igual que los robos de auto. Pudieramos ver el grafo de estos crimenes y como se relacionan las personas en ellos para entender mejor la red.

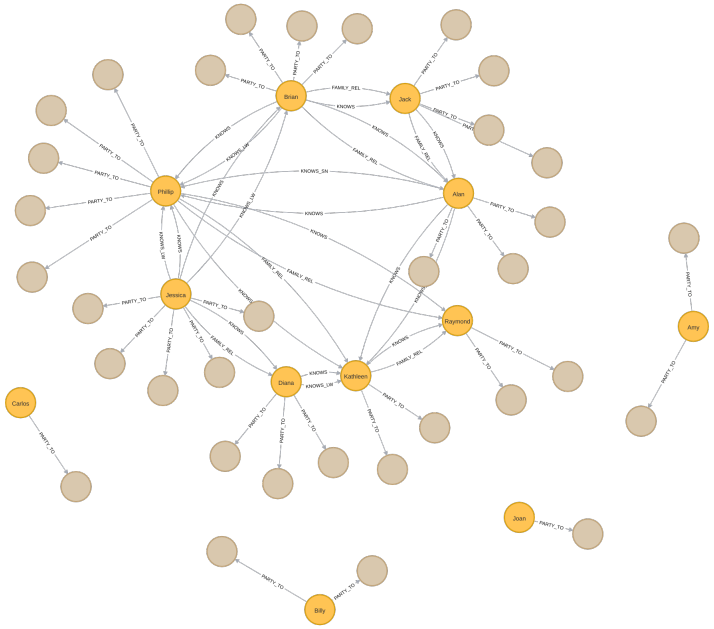

Figura 3: Conexión entre criminales de drogas y carros.

Podemos observar cómo hay una red bien interconectada entre 8 personas con la mayor cantidad de crímenes del tipo drogas y robo de vehículos. Esto puede ser un gran indicativo de una red criminal; es un patrón muy fácil de identificar visualmente después de realizar la consulta correspondiente y puede ayudar a dirigir los recursos de una investigación de una manera más eficiente. Además, vemos que quedan 4 crímenes que están bastante aislados; podemos ampliar la búsqueda a los amigos directos de estas personas para ver si existen conexiones.

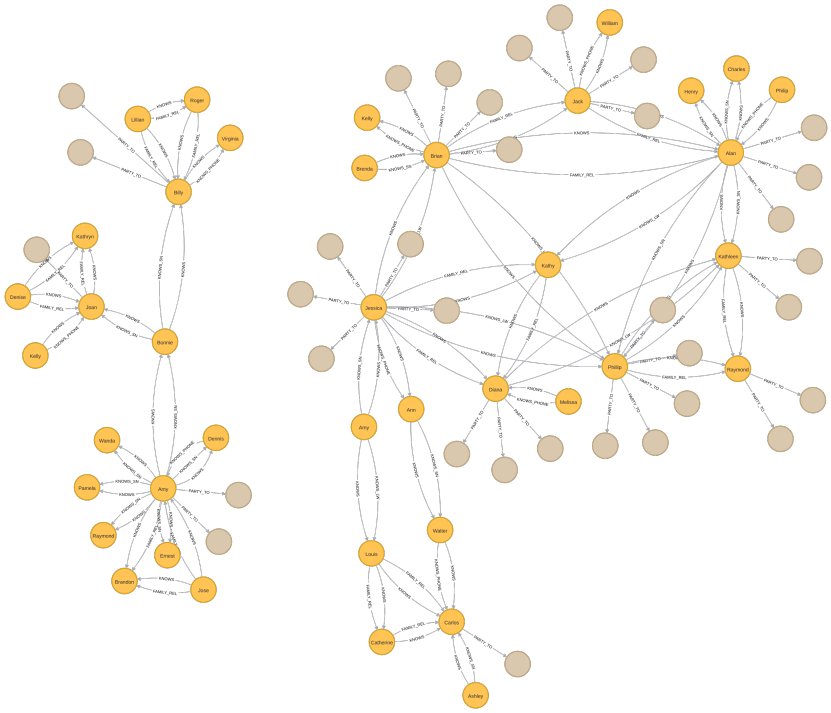

Figura 4: Conexión entre amigos directos de criminales de drogas y carros

Se puede observar claramente que al agregar amigos directos, la cantidad de ciminales individuales disminuye notablemente. En esta situación, las medidas de centralidad pueden resultar extremadamente útiles para la asignación eficiente de recursos. Un criminal que está fuertemente conectado a otros criminales a través de una red de alto grado tiene el potencial de ser un punto focal estratégico. La captura de un individuo con un alto grado de conexiones o un alto autovalor, puede aumentar sig-

nificativamente nuestras posibilidades de capturar a otros delincuentes, ya que es probable que posea información valiosa que podría ser crucial para las investigaciones posteriores.

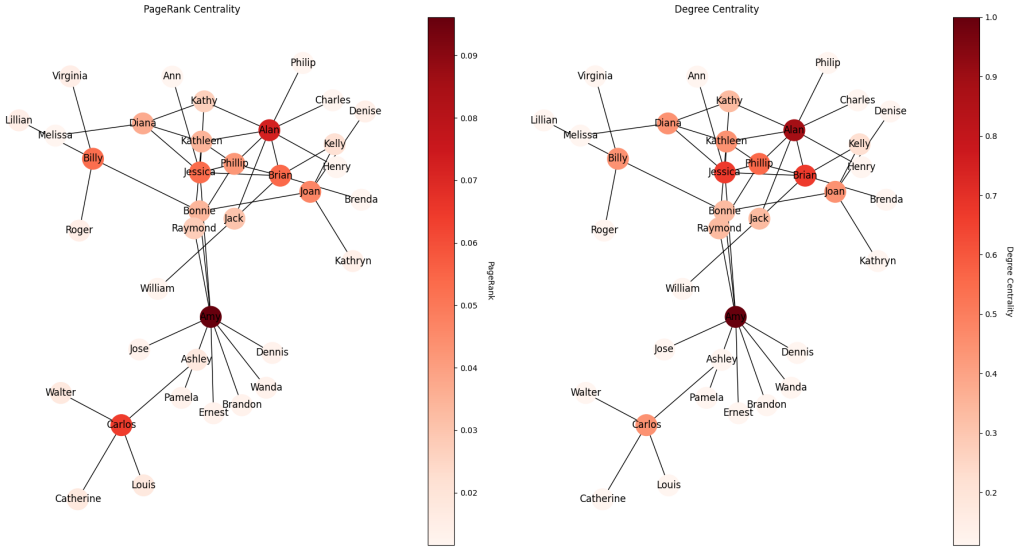

Figura 5: Medidas de centralidad de personas relacionadas con drogas o robo de vehiculos.

2.3. Aprovechando los datos para el comercio

Si un empresario está considerando la mejor ubicación para promover una campaña de ventas de cámaras, estos datos están disponibles de forma pública y, aunque sean anónimos, pueden servirnos para determinar, por ejemplo, las ubicaciones más propensas a robos. También se pueden determinar las ubicaciones más propensas a robos de vehículos para promover la venta de alarmas para los mismos. Esto se puede lograr fácilmente cambiando el texto en color amarillo en la siguiente consulta:

```
MATCH (a:Area)-[:LOCATION_IN_AREA]-(l:Location)-[:OCCURRED_AT]
-(c:Crime {type: "Burglary"})
RETURN a.areaCode, count(distinct(c)) AS crime_count
ORDER BY crime_count DESC LIMIT 10;
```

Codigo 2: Query para encontrar areas con mas personas propensas a realizar crimenes de robo.

3. Aprendizaje automatico relacional

3.1. Metodología

En esta exploración de datos, nos enfocaremos en modelos explicativos debido a la naturaleza de las predicciones, las cuales pueden tener un impacto significativo en las personas. Especificamente, utilizaremos RandomForest. Realizaremos análisis predictivos que, aunque en algunos casos puedan carecer de aplicabilidad práctica o ser considerados moralmente cuestionables. Estos ejemplos sirven para ilustrar las posibilidades que ofrece este conjunto de datos junto con el aprendizaje automático relacional, destacando tanto su potencial como sus limitaciones éticas, que discutiremos mas adelante.

En nuestra busqueda, consideraremos diversas predicciones potenciales. Como aplicación inicial, nos proponemos evaluar si es posible predecir si una persona ha cometido un delito basándonos en las medidas de centralidad dentro de un grafo social, es decir, analizando las conexiones y relaciones entre individuos. Es crucial enfatizar que, desde un punto de vista ético, no es adecuado juzgar si alguien ha cometido un crimen únicamente por sus asociaciones personales. Sin embargo, es un hecho reconocido que el entorno ejerce cierta influencia en las personas.

Por supuesto, esto implica una reducción en el tamaño de nuestro conjunto de datos, ya que nos centraremos exclusivamente en el subgrafo que incluye a las personas que han cometido un crimen y

a sus conocidos. En la figura 6, presentamos una visualización de este grafo social. Además, hemos incorporado un nodo representativo de un crimen como ejemplo ilustrativo. Esto nos permite mostrar de manera clara la variable objetivo de nuestra predicción: un valor booleano que indica si una persona dentro de este grafo específico ha cometido o no un crimen

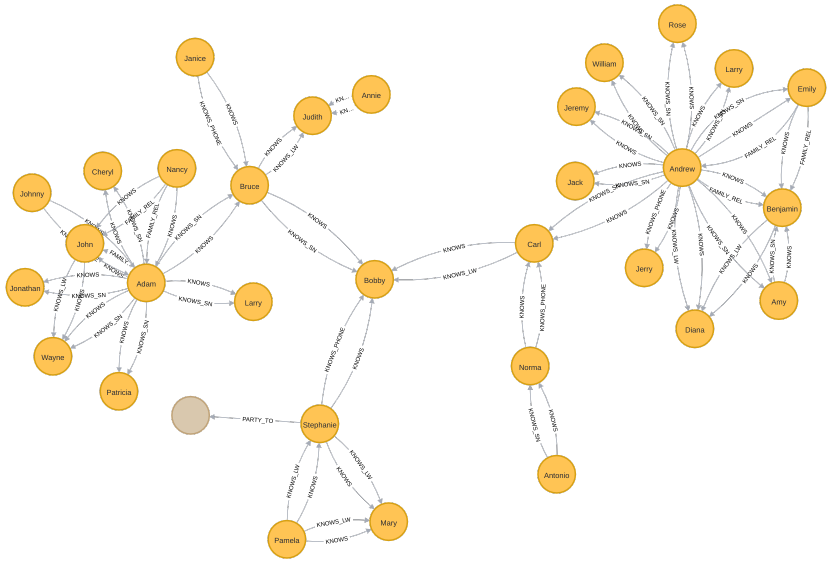

Figura 6: Grafo social limitado a 30 registros.

El conjunto de datos utilizado para la predicción se construye con el número de identificación de la persona y los siguientes parámetros de centralidad:

- Pagerank(Aproximación al autovalor): Medida de centralidad que indica cuán bien conectado está un nodo con otros nodos que, a su vez, están bien conectados.
- Betweenness: Esto indica si un nodo sirve de intermediario en el camino más corto entre muchos otros nodos.
- Closeness: Medida que indica cuán cerca está un nodo del resto de los nodos en la red.
- Degree: Esta medida indica a cuántos nodos está directamente conectado nuestro nodo.
- Clustering: Es una medida que indica cuán bien conectados están los vecinos de un nodo entre sí.
- Triangle Count: Calcula el número de ciclos de longitud 3 en los que un nodo participa.

Vamos a entrenar un Random Forest en búsqueda de los parámetros que mejor se ajusten. Exploramos múltiples combinaciones sin repetición de estas métricas. Luego, se añaden metricas especificas en base al conocimiento del problema:

- Length to criminal: La distancia más corta que conecta a una persona con un criminal.
- connected criminal count: La cantidad de criminales que una persona conoce a través de un máximo de 3 niveles de conexiones, donde un nivel representa la amistad de un amigo.
- Lives with criminal: Indica si una persona vive con algún criminal.
- total crimes count: Representa la cantidad de crímenes cometidos por los criminales conocidos a través de un máximo de 3 niveles de conexiones.

Procedemos a explorar los resultados y los comparamos con las expectativas en relación a la estructura de datos que poseemos. Finalmente, discutimos las implicaciones morales del tema.

4. Resultados

4.1. Analizis basado en medidas de centralidad

Disponemos de un conjunto de datos que incluye 368 registros, de los cuales solo 29 son ejemplos positivos. Esta significativa desproporción resulta en un marcado desequilibrio en el dataset, lo que representa un desafío notable para la predicción eficaz de casos positivos. Es razonable anticipar ciertas dificultades en los resultados debido a este desbalance.

Existen técnicas para mitigar problemas de desequilibrio, como la generación de datos sintéticos. Sin embargo, en nuestro contexto específico, que implica el análisis de comunidades criminales, la aplicación de estas técnicas podría no ser óptima. Las dinámicas y patrones en dichas comunidades son complejos, variados y potencialmente únicos. Por lo tanto, existe el riesgo de que los datos sintéticos no capturen la esencia de estos patrones complejos. Además, si los casos positivos en nuestro dataset representan anomalías (outliers) o no abarcan un espectro suficientemente amplio de situaciones, la capacidad de generalización de cualquier modelo entrenado con estos datos podría verse seriamente comprometida.

Es crucial tener en cuenta estos factores y se deben interpretar los resultados con precaución entendiendo la limitación en el aprendizaje del modelo por la cantidad de datos de entrenamiento y pruebas.

En la Figura 7, se presenta la matriz de dispersión, y se puede apreciar que, a simple vista, los parámetros no muestran una separación clara de los datos que facilite su clasificación.

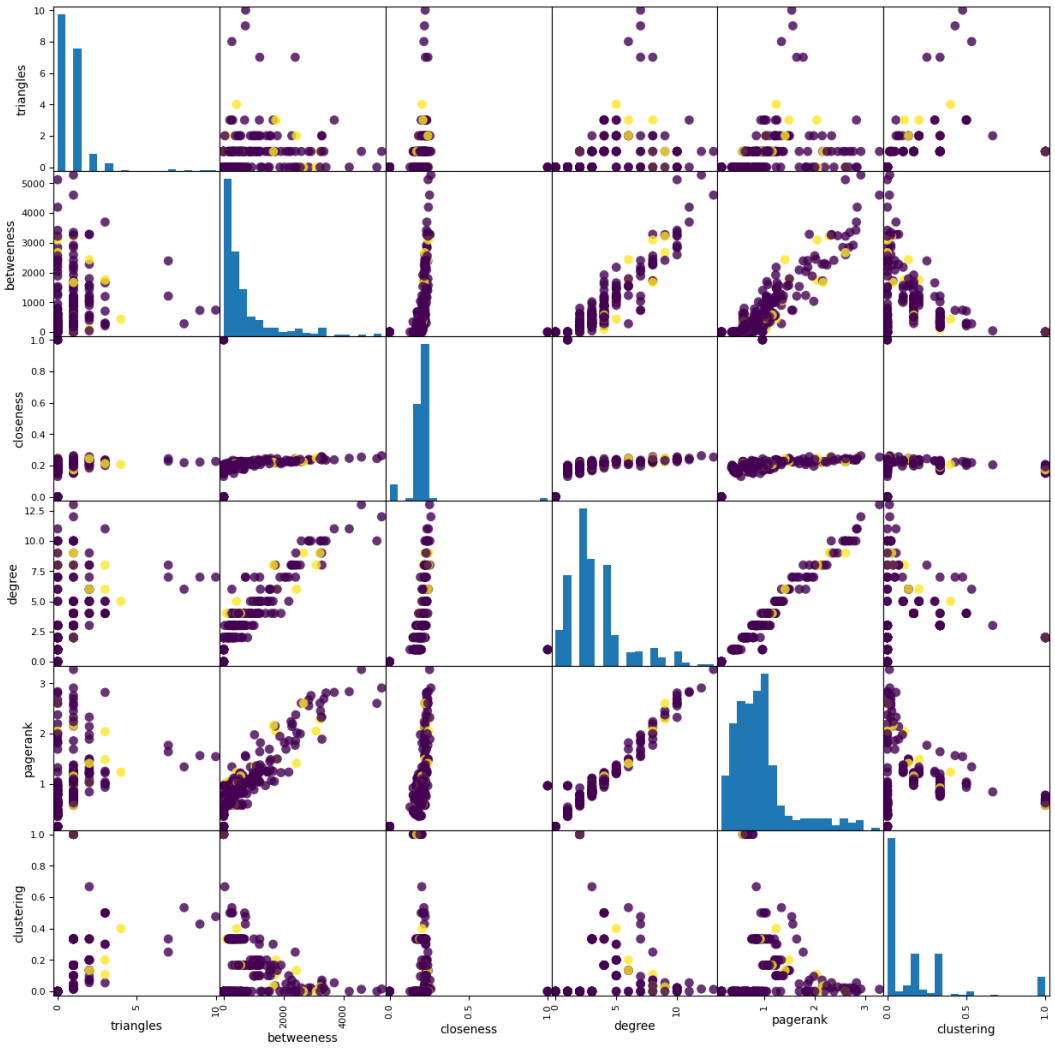

Figura 7: Matriz de dispersión del dataset basado en medidas de centralidad

La matriz de correlación de Pearson, presentada en la Figura 8, revela una notable correlación entre la medida de PageRank y el grado de los nodos en nuestro grafo. Esto sugiere que los nodos con un alto número de conexiones tienden a estar interconectados, lo que podría indicar una estructura de red centralizada o la presencia de subgrupos densamente conectados. Además, se observa una fuerte correlación entre el grado y la centralidad de intermediación (betweenness), lo que implica que los nodos con muchas conexiones frecuentemente desempeñan un papel crucial como intermediarios en la red. Estos nodos actúan como puntos clave en la transmisión de información o recursos dentro de la estructura del grafo, conectando diferentes segmentos o comunidades.

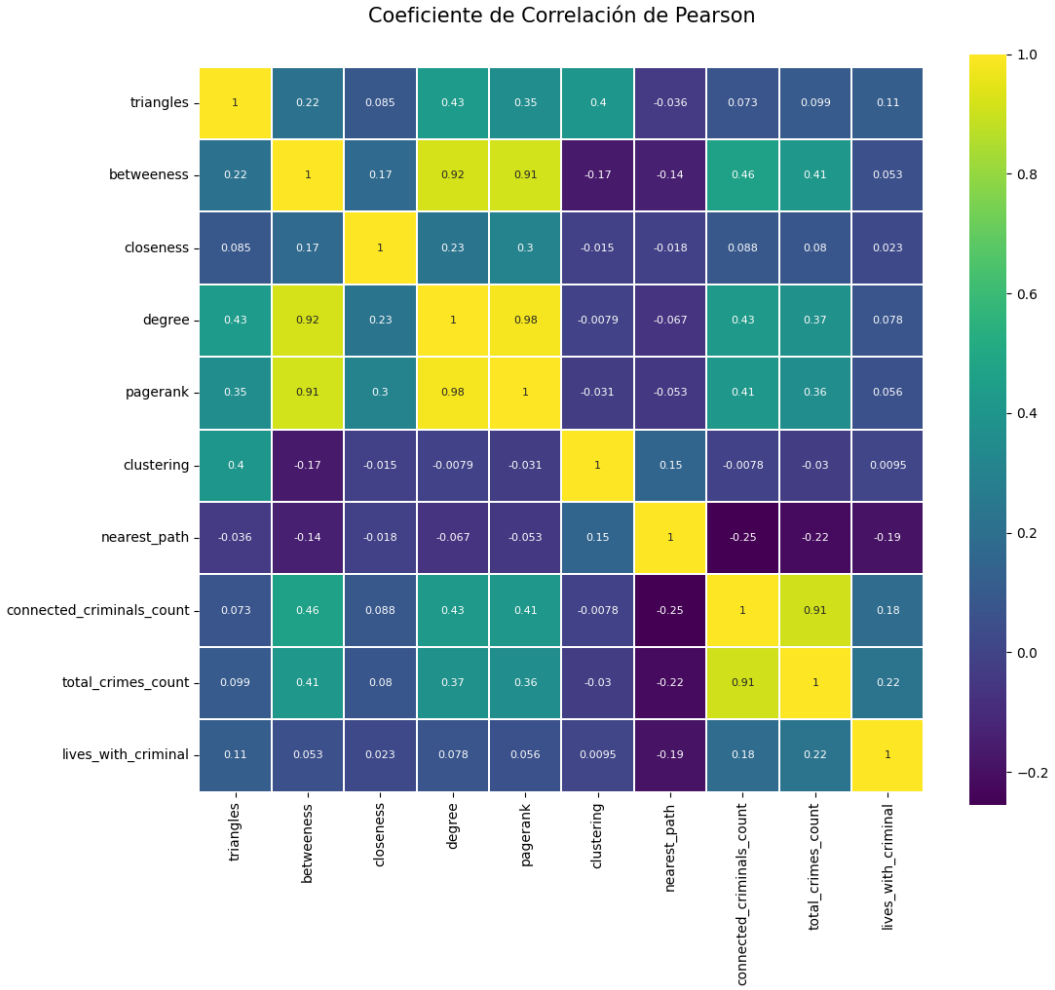

Figura 8: Matriz de correlació de Pearson del dataset basado en medidas de centralidad

Los resultados del entrenamiento se muestran en el apendice A, muestra un desempeno decente, sin embargo, es de esperarse que este resultado no sea muy significativo de la clasificacion en ambas clases debido al imbalance en los datos, esto se refleja en la figura 9.

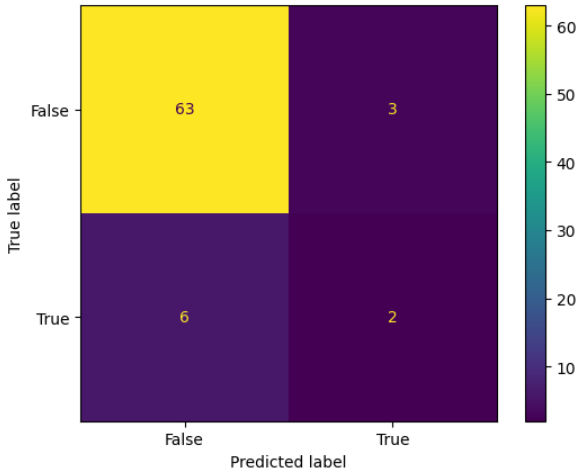

Figura 9: Matriz de confusión del dataset de pruebas basado en medidas de centralidad

La matriz de confusión revela que nuestro modelo tiene dificultades para clasificar correctamente los ejemplos positivos. Esta limitación puede atribuirse en gran medida al desequilibrio en los datos y a la escasa cantidad de ejemplos positivos disponibles. No obstante, buscamos mejorar nuestras métricas introduciendo otros parámetros que podrían ofrecer una discriminación más efectiva de los datos, con la esperanza de lograr una separación más evidente.

5. Análisis híbrido

En esta sección, exploraremos el rendimiento de la red utilizando una combinación de medidas de centralidad y características que hemos extraído manualmente, basándonos en nuestro conocimiento del problema. El conjunto de datos es idéntico al caso anterior, pero ahora incluye cuatro atributos adicionales, los cuales se presentan en la siguiente figura:

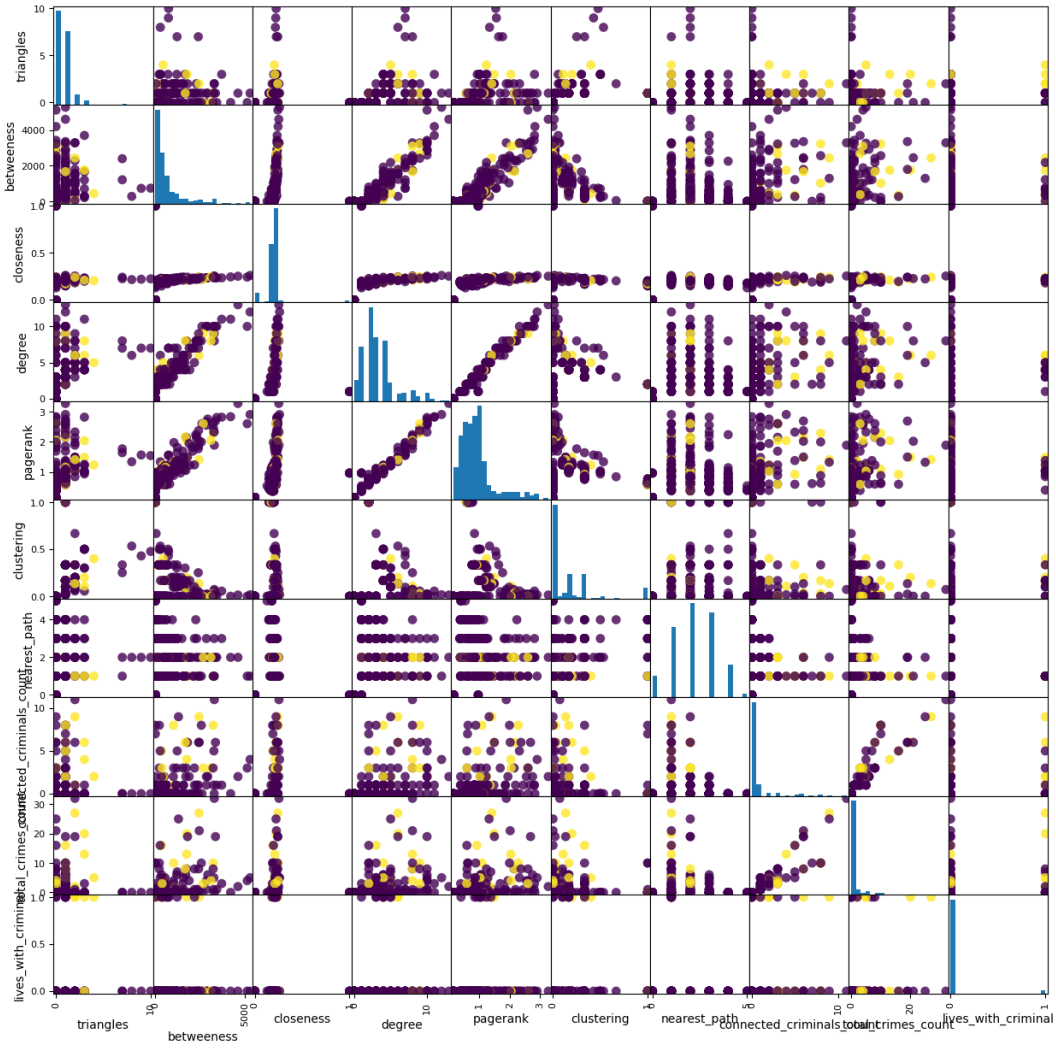

Figura 10: Matriz de dispersión del dataset híbrido.

Nuevamente, no se ve una separacion a simple vista de la cual podamos concluir que separar. Por otro lado, en esta matriz actualizada, resalta una correlación significativa entre la cantidad de criminales a los que un individuo está conectado y el número total de delitos cometidos por estas conexiones. Este hallazgo es coherente con las expectativas, dado que una mayor red de contactos criminales podría implicar una mayor exposición a actividades delictivas.

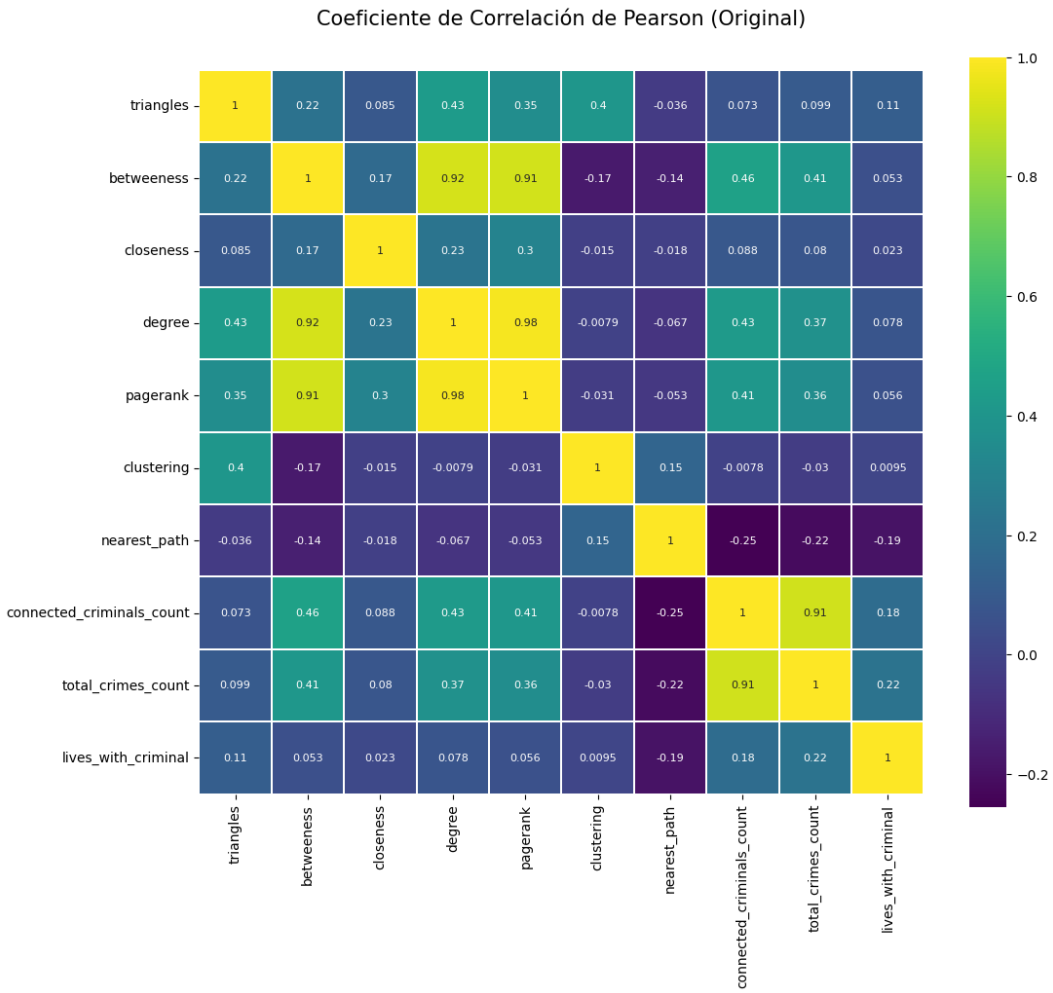

Figura 11: Matriz de correlación de Pearson del dataset híbrido.

Al emplear el método disponible en el paquete 'utils' para encontrar la mejor combinación de parámetros, se encontró que la tupla ('closeness', 'connected_criminals_count', 'lives_with_criminal') obtuvo el mejor rendimiento en nuestro conjunto de pruebas, alcanzando una precisión del 91.9 %. Esto representa una ligera mejora en comparación con el modelo sin los parámetros adicionales que fueron derivados a partir del conocimiento del problema, el cual tenía una precisión del 89 % en el conjunto de pruebas. Sin embargo, es importante destacar que se observó una notable reducción en el rendimiento en el conjunto de entrenamiento, pasando del 96 % al 88 %. A continuación la nueva matriz de confusion:

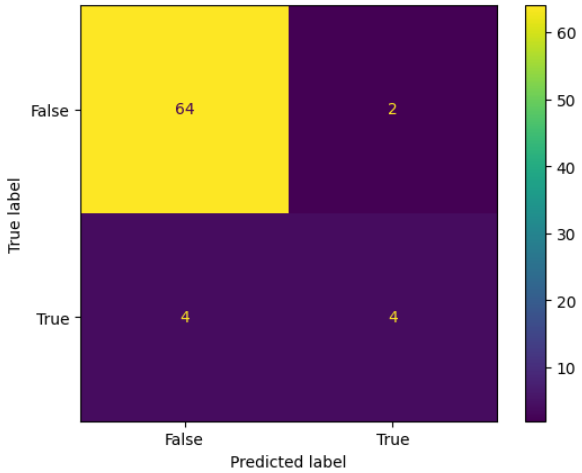

Figura 12: Matriz de confusión del dataset de pruebas basado en medidas de centralidad

6. Implicaciones morales y éticas

Al considerar las implicaciones morales y éticas de utilizar un algoritmo para determinar si alguien es criminal o no, nos adentramos en un terreno complejo y delicado. Es fundamental reconocer que, aunque la tecnología y los análisis de datos pueden ofrecer herramientas poderosas, su uso en contextos tan sensibles como la identificación de comportamientos criminales lleva consigo una serie de consideraciones éticas y morales de gran importancia.

Los algoritmos suelen ser tan imparciales como los datos en los que se entrenan. Este documento ha sido un ejemplo de como el imbalance en los datos puede afectar el resultado, claro, este ejemplo es extremo porque son muy pocos datos y a medida la cantidad de datos aumente sera cada vez mas robusto. Pero, si los datos históricos contienen sesgos, estos se perpetuarán y amplificarán, pudiendo llevar a prácticas discriminatorias, particularmente contra grupos ya marginados o estigmatizados.

Por otro lado, la implementación adecuada de algoritmos puede proporcionar a los investigadores una visión más profunda y matizada de las redes criminales y patrones de comportamiento. Esto podría llevar a investigaciones más eficientes y precisas, ayudando a identificar conexiones y tendencias que de otro modo podrían pasarse por alto.

Pero, mientras los algoritmos ofrecen un potencial significativo para ayudar en la identificación y prevención de actividades criminales, su implementación debe realizarse con una profunda consideración de los principios éticos y morales.

7. Conclusión

El uso de bases de datos relacionales, como Neo4j, ha demostrado ser una herramienta valiosa en la investigación de actividades criminales. Estas bases de datos facilitan el análisis de complejas redes de relaciones, las cuales son características frecuentes en grupos criminales. En comparación con las bases de datos no relacionales, las relacionales ofrecen una mayor eficacia para comprender las conexiones presentes en estos grupos.

El entrenamiento utilizando técnicas de aprendizaje relacional, especificamente, de modelos que hacen uso de medidas de centralidad y otros parámetros derivados del conocimiento específico del problema en cuestión, se perfila como un enfoque prometedor para orientar investigaciones futuras. Sin embargo, es importante reconocer las limitaciones encontradas en este estudio, particularmente en relación con la escasez de datos disponibles y la dificultad inherente a la aproximación de estos mediante métodos de generación de datos sintéticos.

No obstante, subrayar la importancia de considerar las implicaciones morales y éticas en la aplicación de estas tecnologías. Mientras exploramos y aprovechamos las capacidades de las bases de datos relacionales y el aprendizaje automático en el contexto de la investigación criminal, debemos hacerlo con una consideración cuidadosa de las complejidades éticas y humanas involucradas.

Referencias

[Dep18] Joe Depeau. Graph technology is in the pole position to help law enforcement. `https://neo4j.com/blog/graph-technology-pole-position-law-enforcement/`, 2018. Accessed: 11-01-2024.

[MBB+21] M. Mahfoud, W. Bernasco, S. Bhulai, et al. Forecasting spatio-temporal variation in residential burglary with the integrated laplace approximation framework: Effects of crime generators, street networks, and prior crimes. *Journal of Quantitative Criminology*, 37:835–862, 2021.

A. Appendix

Cuadro 3: Mejores metricas para distintas combinaciones de features

$min_{\text{samples\_split}}$	$min_{\text{samples\_leaf}}$	max_{depth}	Test Set Accuracy	Training Set Accuracy	Features
10	4	6	0.892	0.963	['triangles', 'betweenness', 'closeness', 'pagerank']
5	8	6	0.878	0.901	['betweenness', 'closeness']
10	8	6	0.878	0.891	['triangles', 'betweenness', 'closeness', 'pagerank', 'clustering']
10	6	6	0.878	0.912	['triangles', 'closeness', 'degree', 'pagerank']
5	6	7	0.878	0.932	['triangles', 'betweenness', 'closeness', 'degree', 'pagerank']

Cuadro 4: Mejores métricas para distintas combinaciones features híbridos

$min_{\text{samples\_split}}$	$min_{\text{samples\_leaf}}$	max_{depth}	Test Set Accuracy	Training Set Accuracy	Features
5	6	6	0.919	0.881	['closeness', 'connected_criminals_count', 'lives_with_criminal
5	6	6	0.905	0.915	['lives_with_criminal']
5	6	6	0.905	0.857	['pagerank', 'connected_criminals_count', 'lives_with_criminal
5	8	7	0.905	0.895	['betweenness', 'pagerank', 'nearest_path', 'connected_criminals_count', 'live
10	8	7	0.905	0.878	['closeness', 'degree', 'connected_criminals_count']