Plan Overview

A Data Management Plan created using DMPTool

Title: Análisis del Crimen en Estados Unidos

Creator: Julio Solano

Affiliation: Universidad de Los Andes (uniandes.edu.co)

Principal Investigator: Camila Lenis, Ronaldo Ballesteros

Data Manager: Camila Lenis

Project Administrator: Oscar Camilo Álvarez

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Project abstract:

Este proyecto se centra en el despliegue de una solución analítica integral utilizando el conjunto de datos de Estadísticas de Crimen en los Estados Unidos, como parte de la materia 'Despliegue de Soluciones Analíticas'. El objetivo del proyecto es desarrollar un modelo predictivo capaz de analizar patrones de crimen a partir de un conjunto de datos que abarca diferentes tipos de delitos, desde ofensas violentas como homicidios y asaltos hasta crímenes contra la propiedad como robos y allanamientos.

La metodología seguida incluye varias etapas esenciales para el desarrollo del modelo. Primero, se lleva a cabo un proceso de limpieza y preparación de datos, garantizando que los atributos clave (como la fecha y hora del crimen, la ubicación geográfica, las características de la víctima, y el tipo de delito) estén en un formato adecuado para el análisis. A continuación, se realiza un análisis exploratorio para identificar tendencias temporales y variaciones geográficas.

Posteriormente, se aplica un enfoque de aprendizaje supervisado para la clasificación y predicción de tipos de delitos, utilizando algoritmos como Random Forest y XGBoost, para identificar los factores más influyentes en la ocurrencia de ciertos delitos. Además, se incluye un análisis de

agrupamiento no supervisado para descubrir patrones ocultos, como zonas de alta incidencia o tipos de delitos relacionados.

El despliegue del modelo se lleva a cabo utilizando servicios en la nube, particularmente AWS, asegurando la escalabilidad y la capacidad de manejar grandes volúmenes de datos. También se utiliza GitHub para el versionamiento del código, lo cual facilita la colaboración y gestión de cambios.

La solución se desplegará en un entorno donde los resultados predictivos puedan ser consultados por distintas autoridades nacionales o internacionales responsables de políticas públicas, ofreciendo conocimientos accionables que pueden ayudar en la toma de decisiones para la prevención del crimen a nivel mundial.

Start date: 10-18-2024

End date: 11-26-2024

Last modified: 10-20-2024

Análisis del Crimen en Estados Unidos

Data Collection

What data will you collect or create?

Los datos son de dominio público, administrados por la plataforma Kaggle. El set de datos es del tipo estructurado y contiene a la fecha 983,000 observaciones y 28 columnas. Este conjunto de datos ofrece una descripción detallada de las estadísticas delictivas en los Estados Unidos. Incluye varias categorías de delitos, desde delitos violentos como homicidios y agresiones hasta delitos contra la propiedad como robos y allanamientos.

Los datos son del tipo batch, en formato Excel.

How will the data be collected or created?

Los datos fueron recolectados por las agencias gubernamentales estadounidenses entre 2020 y 2023. El proyecto contara con un repositorio en la nube. Para garantizar la coherencia y la calidad del manejo de los datos, se utilizará la metodología de Ciencia de Datos ASUM-DM (Analytics Solutions Unified Method for Data Mining) desarrollada por IBM. Para facilitar la interoperabilidad se usarán formatos abiertos y estándar como CSV o JSON.

Se utilizará una estructura en carpetas donde en la carpeta raíz utilizaremos una carpeta destinada a los datos originales sin procesar, otra para datos limpios y procesados, otra para scripts de análisis y preprocesamiento. Una carpeta para documentación de referencia, otra para resultados (modelos, visualizaciones, informes). El versionamiento se hará atraves de git. Los nombres de los archivos serán de tipo descriptivo que que incluyan fechas o versiones.

El control de versiones con Git para rastrear cambios en los archivos de datos, scripts y documentos. Cada actualización importante puede estar asociada a un commit con un mensaje descriptivo.

La validación de datos incluye la implementación de controles de calidad en los datos para identificar valores atípicos, datos faltantes o inconsistencias. Se documentará el flujo de trabajo y de las decisiones metodológicas, lo que facilita reproducir y verificar el trabajo.

Documentation and Metadata

What documentation and metadata will accompany the data?

Para esto se crea un diccionario de datos, en cual se incluirá aparte del nombre, una descripción del dato con su naturaleza, rango y un ejemplo del formato del dato. Los metadatos a considerar y a acompañar la información consisten en anexos de informes de avance donde se documentarán análisis exploratorios de los datos incluidos, así como lo relativo a las metodologías empleadas y resultados de los análisis en cada una de las fases o iteraciones del proyecto. El titulo del archivo contendrá las fechas de corte y generación de los datos.

Información Principal del Delito:

- DR_NO: Un identificador único para cada delito. Formato simulado: número entero de 7 dígitos.
  Ejemplo: 1234567
  Date Rptd: Fecha en que se reportó el delito a las autoridades. Formato: fecha en formato YYYY-MM-DD, rango de fechas entre 2010-01-01 y 2023-12-31.
  Ejemplo: 2022-11-15
  DATE OCC: Fecha en que ocurrió el delito. Formato: fecha en formato YYYY-MM-DD, rango entre 2010-01-01 y 2023-12-31.
  Ejemplo: 2022-11-10
  TIME OCC: Hora en que ocurrió el delito. Formato: hora en formato de 24 horas HH:MM (rango 00:00 a 23:59).
  Ejemplo: 14:35
  AREA: Código del área geográfica o comisaría donde ocurrió el delito. Formato: número entero de 1 a 5 dígitos.
  Ejemplo: 15
  AREA NAME: Nombre descriptivo del área. Formato: cadena de texto alfanumérica con un máximo de 50 caracteres.
  Ejemplo: Central Precinct
  Rpt Dist No: Número del distrito que reporta el delito. Formato: número entero de 3 dígitos.
  Ejemplo: 123

Clasificación del Delito:

- Part 1-2: Indica si el delito es de la Parte 1 (grave) o Parte 2 (menos grave). Formato: entero 1 o 2.
  Ejemplo: 1
  Crm Cd: Código o número de clasificación del delito. Formato: número entero de 3 a 4 dígitos.
  Ejemplo: 487
  Crm Cd Desc: Descripción del código del delito. Formato: cadena de texto de hasta 100 caracteres.
  Ejemplo: Robo con fuerza en establecimiento comercial
  Mocodes: Código que representa las motivaciones o circunstancias del delito. Formato: cadena de texto alfanumérica de hasta 6 caracteres.
  Ejemplo: M0045
  Part 1-2: Indica si el delito es de la Parte 1 (grave) o de la Parte 2 (menos grave).
  Crm Cd: Código o número de clasificación del delito.
  Crm Cd Desc: Descripción del código del delito.
  Mocodes: Motivaciones o circunstancias relacionadas con el delito.

Información de la Víctima:

- Vict Age: Edad de la víctima. Formato: número entero entre 0 y 100.
  Ejemplo: 34
  Vict Sex: Sexo de la víctima. Formato: M (masculino), F (femenino) o X (no especificado).
  Ejemplo: F
  Vict Descent: Origen racial o étnico de la víctima. Formato: código alfanumérico de 1 a 3 caracteres.
  Ejemplo: W (White), B (Black), H (Hispanic), etc.

Ubicación y Contexto:

- Premis Cd: Código del tipo de lugar donde ocurrió el delito. Formato: número entero de 2 a 3 dígitos.
  Ejemplo: 123
  Premis Desc: Descripción del tipo de lugar. Formato: cadena de texto de hasta 50 caracteres.
  Ejemplo: Residencia unifamiliar
  Weapon Used Cd: Código del arma utilizada (si aplica). Formato: número entero de 2 dígitos.
  Ejemplo: 13
  Weapon Desc: Descripción del arma utilizada. Formato: cadena de texto de hasta 50 caracteres.
  Ejemplo: Pistola

Información Adicional:

- Status: Estado actual del caso (ej. abierto, cerrado). Formato: cadena de texto de hasta 10 caracteres.
  Ejemplo: Abierto
  Status Desc: Descripción del estado del caso. Formato: cadena de texto de hasta 50 caracteres.
  Ejemplo: Investigación en curso
  Crm Cd 1, 2, 3, 4: Códigos adicionales de delitos si aplican. Formato: número entero de 3 a 4 dígitos.
  Ejemplo: 487, 502, 101, 345
  LOCATION: Ubicación general del delito. Formato: cadena de texto con la dirección o punto de referencia, hasta 100 caracteres.
  Ejemplo: 123 Main St, Los Ángeles
  Cross Street: Intersección o calle cercana. Formato: cadena de texto de hasta 50 caracteres.
  Ejemplo: Calle 1 y Calle 2
  LAT, LON: Coordenadas de latitud y longitud. Formato: flotante para la latitud y longitud con precisión de 6 decimales.
  Ejemplo: 34.052235, -118.243683

Ethics and Legal Compliance

How will you manage any ethical issues?

Dado que estos datos son de uso público y no contienen restricciones éticas específicas, no se requiere obtener consentimiento individual de los participantes. Los datos han sido recopilados y puestos a disposición por las autoridades, generalmente con el propósito de facilitar el acceso público a información de interés social, como la transparencia en los informes de delitos y el análisis de tendencias criminales.

Los campos como edad, sexo y origen étnico se presentan de forma agregada y no permiten identificar a personas individuales. Sin embargo, en caso de que alguna variable adicional pudiera revelar indirectamente la identidad de un individuo, se podrían aplicar técnicas adicionales de anonimización, como la supresión o generalización de datos sensibles.

Los datos serán almacenarse en sistemas seguros con acceso controlado, utilizando herramientas de encriptación tanto para el almacenamiento como para la transferencia si es necesario.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

1. La propiedad de los datos generalmente recae en la entidad gubernamental o agencia pública que recopiló y publicó los datos. En el caso de los datos de crímenes, es probable que pertenezcan a una agencia de seguridad o departamento de policía local, que es responsable de su recopilación y administración. Aunque los datos son de uso público, la agencia propietaria sigue siendo la titular oficial.

2. La reutilización de los datos públicos suele estar sujeta a las condiciones establecidas por la agencia o institución que los publica. Normalmente, estos datos están bajo una licencia abierta, como Creative Commons (CC-BY o CC0) o alguna forma de licencia de datos públicos (por ejemplo, Open Government License). Esto permite que los datos sean reutilizados, adaptados y distribuidos libremente, siempre y cuando se cumplan las condiciones especificadas (por ejemplo, atribución al propietario de los datos, si es requerido).

3. En general, los datos de uso público no suelen tener restricciones significativas en cuanto a su reutilización, siempre que se respeten las condiciones de la licencia bajo la cual fueron publicados. Sin embargo, es importante verificar si el conjunto de datos contiene alguna porción o componente proporcionado por terceros, en cuyo caso podría haber restricciones adicionales. Si los datos incluyen información de terceros o conjuntos de datos fusionados, se debe garantizar que el uso de esa información también cumpla con las licencias asociadas.

4. Dado que estos datos son de uso público y su objetivo principal es la transparencia y el acceso abierto para análisis e investigación, no es común que su intercambio se vea aplazado o restringido. No obstante, si los datos van a ser utilizados en una investigación que busque publicaciones o patentes, puede haber consideraciones en cuanto a la temporalidad de la publicación, pero esto generalmente depende más del contexto académico o de investigación, no de los datos públicos en sí. En estos casos, es posible que los investigadores decidan posponer la divulgación de sus propios resultados, pero no del acceso a los datos en sí.

Storage and Backup

How will the data be stored and backed up during the research?

La administración de los archivos del proyecto se realizara en un servidor en la nube, el cual tendrá a Git como herramienta de administración y control DVC (Data Versión Control), sistema que permite el manejo de versiones de los datos y del pre procesamiento realizado sobre estos. La etapa de exploración y desarrollo de la solución se adelantará con una muestra de los datos, buscando reducir el impacto de los costos de almacenamiento. El proveedor del servidor en la nube será Amazon (AWS).

El equipo del proyecto será el responsable será el responsable de los respaldos, para lo cual como equipo tendremos backups incrementales del repositorio DVC y de manera semanal se realizará una copia de seguridad que sirva como respaldado en una infraestructura diferente.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

La retención y preservación de datos de crímenes debe seguir un enfoque que equilibre las necesidades legales, contractuales y regulatorias con el potencial de investigación futura. Los datos relacionados

con investigaciones en curso o aquellos exigidos por la ley deben ser conservados durante el tiempo estipulado, mientras que los datos sensibles o irrelevantes deben ser destruidos cuando ya no sean necesarios.

Los usos previsibles de los datos incluyen el análisis de patrones delictivos, la evaluación de políticas de seguridad, investigaciones académicas, y estudios de justicia social. Los datos pueden ser conservados a largo plazo si se consideran valiosos, con una revisión periódica para evaluar su necesidad.

La administración del repositorio se encuentra a cargo de uno de los miembros del equipo, el cual es el responsable de todos los temas de infraestructura y servidores. El acceso al repositorio estará limitado al equipo del proyecto en primera instancia.

Data Sharing

How will you share the data?

El servicio utilizado para el almacenamiento y control de los datos, al ser parte de los servicios de AWS, tiene un costo considerable. Por esta razón, una vez finalizado el proceso de investigación, los datos y los resultados de la investigación serán alojados en un repositorio privado en GitHub, que no requiere ningún costo para su mantenimiento.

El proyecto inicial será de carácter privado, y no se contempla compartirlo más allá de los propósitos definidos en el alcance del proyecto. En este caso, se otorgará acceso al repositorio correspondiente.

Responsibilities and Resources

Who will be responsible for data management?

En este caso, todos los miembros del equipo del proyecto compartimos la responsabilidad de implementar el Plan de Gestión de Datos (DMP), asegurando que sea revisado y actualizado según sea necesario. Cada miembro será responsable de diferentes actividades de gestión de datos según su rol en el proyecto, lo que incluye la recopilación, almacenamiento, análisis y preservación de los datos.
