PROJECT PROPOSAL Grupo 6: Marcos Gago Rivas, Jorge Sierra Alonso, Pablo Agustín Chicharro Gómez

- **Description of the problem.**

El problema que hemos seleccionado se centra en el análisis de los crímenes ocurridos en Chicago desde 2001 hasta la actualidad. La relevancia de este estudio radica en la notable incidencia delictiva en la ciudad a lo largo de los años. Nos proponemos investigar el tipo de crimen más recurrente en este periodo, identificando patrones y tendencias sobre la franja horaria en la que los ciudadanos de Chicago son más propensos a cometer delitos.

Además, planeamos examinar la evolución de los tipos de crímenes más frecuentes en intervalos de cinco años, analizando posibles cambios en esta dinámica a lo largo del tiempo. También nos enfocaremos en determinar el distrito más problemático de Chicago y calcular el porcentaje de delitos en los cuales se logró realizar un arresto, proporcionando una visión de la eficacia de las acciones legales.

Finalmente, abordaremos un análisis específico de los años de la pandemia, buscando comprender cómo pudo haber influido en los patrones delictivos. Estos enfoques nos permitirán obtener una comprensión profunda de la criminalidad en Chicago, destacando áreas críticas para la intervención y proporcionando información valiosa para la formulación de políticas públicas y estrategias de seguridad.

- **Description of the need for Big Data processing.**

Para llevar a cabo el estudio de los datos explicados anteriormente, es necesario emplear Big Data. Esto se debe a que los datos de crímenes en la ciudad de Chicago a lo largo de estos años, desde 2001 hasta la actualidad, representan una gran cantidad de información de muy diverso tipo(Volumen y Variedad). Además, queremos llevar a cabo un análisis con el mayor rendimiento posible(Velocidad). También realizamos el proyecto con la idea de poder realizar este mismo análisis en otras ciudades(Escalabilidad). Por todo esto, reiteramos la importancia de emplear las técnicas de Big Data aprendidas en clase para poder realizar un análisis y una comprensión correcta.

- **Description of the data: Where does it come from? How did you acquire it? What does it mean? What format is it? How big is it?**

La identificación y adquisición de los datos para nuestro proyecto se llevó a cabo mediante una búsqueda exhaustiva entre diversas fuentes de datos proporcionadas en el material de referencia. Tras una evaluación de varias opciones, se identificó que data.gov era la plataforma que albergaba conjuntos de datos relevantes, particularmente relacionados con Estados Unidos, lo que nos permitió acceder a la información necesaria para llevar a cabo nuestra investigación.

Cabe destacar que los datos que utilizamos provienen directamente del Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system, lo que asegura su origen oficial y su confiabilidad en el contexto del estudio de crímenes en la ciudad de Chicago.

La estructura de los datos se compone de múltiples campos, que incluyen información crucial para nuestro análisis. Estos campos son los siguientes: id, case number, date, block, IUCR, Primary type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI code, X-coordinate, Y-coordinate, Year, Update-On, Latitude, Longitude y Location.

En cuanto al tamaño de los datos, estos ocupan un total de 1.74GB, lo que refleja la magnitud del conjunto de datos y justifica aún más la elección de técnicas de Big Data processing para su procesamiento y análisis. Su formato es .cvs como en los Assignments.

Este enfoque meticuloso en la identificación y selección de datos confiables y representativos sienta las bases para una investigación sólida y detallada sobre el índice de criminalidad en Chicago, permitiéndonos abordar de manera efectiva la detección de patrones y tendencias delictivas en la ciudad.

PROJECT PROPOSAL Grupo 6: Marcos Gago Rivas, Jorge Sierra Alonso, Pablo Agustín Chicharro Gómez

- **Description of needed tools and infrastructures.**

En cuanto a las herramientas e infraestructuras que vamos a emplear para este proyecto, vamos a usar los siguientes:
- Primero hemos creado un repositorio de **GitHub** en el cual subiremos tanto los datasets como el código empleado y los análisis realizados.
- Debido a la capacidad de github en contraposición de nuestros datasets, hemos tenido que subir los datos a una carpeta de **Drive.**
-Además, vamos a emplear **python** como nuestro lenguaje de programación al igual que en los assignments.
-Debido al tamaño de los datos, utilizamos **Google Cloud** para su almacenamiento y ejecución, en la que jugará un papel crucial **Apache Spark** por su capacidad de procesamiento y análisis de datos.
-A la hora de representar los resultados obtenidos, vamos a emplear **Excel.**
-Finalmente utilizaremos **html** y **css** para crear la página web sobre nuestro proyecto.