**Introducción**

Para estar mejor preparado para responder a cualquier actividad delictiva y predecirlos, es importante comprender los patrones para que se produzca un delito. En nuestro proyecto, analizamos los datos delictivos de la ciudad de San Francisco, extraídos de un conjunto de datos públicamente disponible.

Objetivo principal

El objetivo principal del proyecto es clasificar la información de la actividad criminal de San Francisco dado el tiempo y la localización. Realizar un análisis y obtener un conjunto de modelos de representación de la información de éstas actividades; y que a partir de dichos modelos podamos consultar información con mejor enfoque y con valores más discretos. Para ello introduciremos las tecnologias MongoDb, Cassandra y Neo4j y obtener diferentes perspectivas de proceso y manejo de la información de forma respectiva.

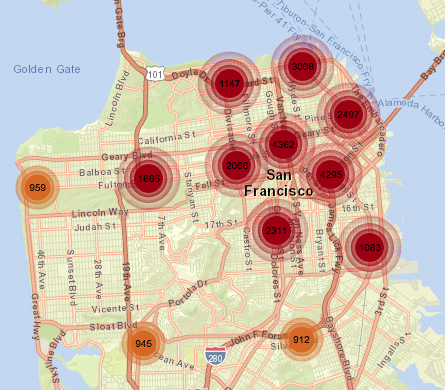
Finalmente, realizaremos una pequeña investigación para encontrar el mejor modelo de aprendizaje que se ajuste a la información que hemos procesado y poder realizar predicciones basado en el proyecto realizado.

En el resultado final la información disponible no solo se basa de la información en bruto obtenida a partir de un volumen masivo de datos correspondientes a incidencias desde el 2013 hasta la actualidad; sino que ademas al ajustar un modelo de predicción se podrá encontrar datos futuros; y la mejor manera es visualizar dichas predicciones sobre la información presente.

Introducción al dataset. ¿Qué es? ¿De donde proviene?

Dataset

Nuestro dataset es publicamente dispinible desde el portal del departamento Policial de *San Francisco*. Contiene datos de todas las incidencias generadas por actividades criminales producidas desde el año 2003 hasta la actualidad. De forma concreta, esta información se encuentra expuesta de forma pública desde el sistema de actualización diaria del *SFPD Crime Incident Reporting* (sistema de reportes del departamento policial) a traves de su plataforma Socrata.



La información esta representada por un conjunto de 2 millones de filas, que contiene información detallada de actividades criminales (o incidencias) representada por: *tipo de delito, resolución, distrito, coordenadas y dirección de la zona, nº de incidencia vinculada*; todos producidos entre el 2003 hasta la actualidad.

· Dates: fecha y hora del incidente del crimen

· Category - categoría del incidente criminal

· Descript - descripción detallada del incidente del crimen

· dayoftheweek de la semana: el día si la semana

· PdDistrict- nombre del distrito del departamento de policía

· Resolution: cómo se resolvió el incidente del crimen

· address: dirección aproximada del incidente delictivo

· X - longitud

· Y - latitud

La X y la Y dan esencialmente el parámetro de ubicación. Es un dato interesante que veremos aprovechar mediante tecnologia que proporciona MongoDb con respecto a geolocalizaciones.

La fecha se agrupa como una fecha, hora y día de la semana. Esto se divide para su uso posterior en las siguientes secciones.

<tabla con algunas filas del dataset>

Representación de las soluciones propuestas

A partir de la estructura de datos definida para el dataset se llevará a cabo una serie de procesos y fases, que nos permitan almacenar dicha información desde distintas perspectivas de almacenaje y representación masiva de datos

Mongodb, como manejo de información a traves de documentos

Cassandra, con potencial residente en la arquitectura para el volumen de datos

Neo4j, como orientación en la busqueda de relaciones entre los datos

A. Análisis de la estructura de datos

B. Preprocesamiento

C. Análisis de datos

D. Representación de datos

**Bibliografia**