SOLUCIÓN DE INTELIGENCIA DE NEGOCIO

Universidad Pablo de Olavide, Inteligencia de Negocio, 2018

Juan Antonio Rodríguez Rodríguez
Alberto Cárdenas Jiménez
Manuel Ridao Pineda

ÍNDICE

1. INTRODUCCIÓN

Mediante la introducción de las nuevas tecnologías en ámbitos donde tradicionalmente se ha gestionado la información en papel, se ha logrado que el volumen de la información y la velocidad con la que se accede a ellos se multiplique. No obstante, aumentar el volumen de datos captados y almacenados no es suficiente; también es necesario hacer una traducción de los datos para transformarlos en conocimiento, y presentarlos para una interpretación rápida y asequible por parte de las personas que los necesiten y no sean conocedoras de la tecnología.

Con esta premisa, este trabajo pretende coleccionar e interpretar los datos recogidos por el departamento de policía de la ciudad de Nueva York (referidos a partir de ahora como NYPD, del inglés *New York Police Department*)[1]. Se trata del cuerpo de policía más grande en Estados Unidos, que actualmente cuenta con 36 000 agentes y 18 000 empleados, cuya responsabilidad es la de proteger una cuidad con casi 9 millones de habitantes. Su labor ha sido uno de los principales responsables en el descenso de crímenes en la cuidad, convirtiéndola en una de las más seguras del país.

Figure 1: Tasa de homicidios en NYC 1985-2012[2]

La labor del NYPD ha generado una considerable cantidad de datos, que se encuentran disponibles como *open data* en la página web de la ciudad[3]. Es necesario hacer una interpretación de estos datos para presentar al personal encargado de gestionar la seguridad de la urbe. Este proyecto se centrará en analizar los datos de los crímenes principalmente según los tipos de ofensas cometidas y su situación geográfica dentro de la ciudad, haciendo uso de los datos proporcionados en la web y de herramientas de análisis de datos como Pentaho Data Integration o KNIME Analytics, así como de su presentación a través de una aplicación web.

[1] https://www1.nyc.gov/site/nypd/about/about-nypd/about-nypd-landing.page
[2] https://mypolicyviews.wordpress.com/2014/01/09/stop-and-frisk-and-the-new-york-city-murder-rate/
[3] https://opendata.cityofnewyork.us/

2. PLAN

2.1. OBJETIVOS

Los objetivos identificados para el trabajo se dividen en objetivos principales y objetivos secundarios. Estos se exponen a continuación:

2.1.1. OBJETIVOS PRINCIPALES

OBJETIVO P-01	**Optimizar la ubicación de las comisarías de policía**
Descripción	Se pretende optimizar la ubicación geográfica de las comisarías de policía en la ciudad en relación con los focos de delincuencia.
Autores	Juan Antonio Rodríguez, Alberto Cárdenas, Manuel Ridao
Versión	1.0

2.1.2. OBJETIVOS SECUNDARIOS

OBJETIVO S-01	**Visualizar las zonas conflictivas y más seguras de la cuidad**
Descripción	Mediante un mapa de calor, se mostrarán las zonas con mas prevalencia de delitos, así como las más seguras. Este mapa deberá ser interactivo y accesible a través de una aplicación web con un navegador.
Autores	Juan Antonio Rodríguez, Alberto Cárdenas, Manuel Ridao
Versión	1.1

OBJETIVO S-02	**Visualizar los datos en tiempo real en un cuadro de mandos**
Descripción	Los datos se tomarán directamente de la web NYC Open Data y se procesarán de manera automática, para así poder mantenerlos actualizados. Además, estos se visualizarán en un cuadro de mandos interactivo que presente la información de manera accesible mediante gráficos desde un navegador.
Autores	Juan Antonio Rodríguez, Alberto Cárdenas, Manuel Ridao
Versión	1.1

OBJETIVO S-03	**Diseñar un modelo predictivo para detectar la ofensa cometida**
Descripción	Se intentará predecir el tipo de delito cometido para un registro mediante el diseño y desarrollo de un modelo predictivo y técnicas de selección de atributos.
Autores	Juan Antonio Rodríguez, Alberto Cárdenas, Manuel Ridao
Versión	1.0

2.2. PLAN DE TRABAJO

Para alcanzar cada todos los objetivos, se dividirá el proyecto en paquetes de trabajo según los objetivos, y estos se repartirán entre los miembros del equipo de trabajo. Los miembros que no participen en un objetivo tendrán la responsabilidad de dar su visto bueno al trabajo realizado por los compañeros.

Las tareas se han dividido de la siguiente forma:

Nombre de tarea	Duración	Comienzo	Fin
Lanzamiento	**1 día**	**lun 18-04-16**	**lun 18-04-16**
Redacción de objetivos	1 día	lun 18-04-16	lun 18-04-16
Redacción de la documentación	**6 días**	**jue 18-04-26**	**lun 18-05-21**
Plan de Proyecto	1 día	jue 18-04-26	jue 18-04-26
Análisis	1 día	vie 18-04-27	vie 18-04-27
Diseño	2 días	lun 18-04-30	mar 18-05-01
Implementación	1 día	vie 18-05-18	vie 18-05-18
Despliegue	0.5 días	lun 18-05-21	lun 18-05-21
Conclusiones	0.5 días	lun 18-05-21	lun 18-05-21
Búsqueda de conjunto de datos	**7 días**	**mar 18-04-17**	**mié 18-04-25**
Búsqueda y adquisición de datos	1 día	mar 18-04-17	mar 18-04-17
Búsqueda de información de tecnologías	1 día	mié 18-04-18	mié 18-04-18
Aprendizaje de tecnologías	2 días	jue 18-04-19	vie 18-04-20
Preprocesado de los datos	3 días	lun 18-04-23	mié 18-04-25
Objetivo P-01	**7 días**	**mié 18-05-09**	**jue 18-05-17**
Búsqueda de herramientas	1 día	mié 18-05-09	mié 18-05-09
Aprendizaje de las herramientas	2 días	jue 18-05-10	vie 18-05-11
Implementación de la solución	3 días	lun 18-05-14	mié 18-05-16
Pruebas de la solución	1 día	jue 18-05-17	jue 18-05-17
Objetivo S-01	**5 días**	**mié 18-05-02**	**mar 18-05-08**
Búsqueda de herramientas	1 día	mié 18-05-02	mié 18-05-02
Aprendizaje de las herramientas	1 día	jue 18-05-03	jue 18-05-03
Implementación de la solución	2 días	vie 18-05-04	lun 18-05-07
Pruebas de la solución	1 día	mar 18-05-08	mar 18-05-08
Objetivo S-02	**5 días**	**mié 18-05-02**	**mar 18-05-08**
Búsqueda de herramientas	1 día	mié 18-05-02	mié 18-05-02
Aprendizaje de las herramientas	1 día	jue 18-05-03	jue 18-05-03
Implementación de la solución	2 días	vie 18-05-04	lun 18-05-07
Pruebas de la solución	1 día	mar 18-05-08	mar 18-05-08
Objetivo S-03	**5 días**	**mié 18-05-02**	**mar 18-05-08**
Búsqueda de herramientas	1 día	mié 18-05-02	mié 18-05-02
Aprendizaje de las herramientas	1 día	jue 18-05-03	jue 18-05-03
Implementación de la solución	2 días	vie 18-05-04	lun 18-05-07
Pruebas de la solución	1 día	mar 18-05-08	mar 18-05-08

Los paquetes de trabajo se han distribuido de la siguiente manera:

Paquete de tareas	Duración	Responsable
Lanzamiento	1 día	Todo el grupo
Redacción de la documentación	6 días	Todo el grupo

Búsqueda de conjunto de datos	7 días	Todo el grupo
Objetivo P-01	7 días	Todo el grupo
Objetivo S-01	5 días	Juan Antonio Rodríguez
Objetivo S-02	5 días	Alberto Cárdenas
Objetivo S-03	5 días	Manuel Ridao

La planificación semanal acordada es la siguiente:

Paquete de tareas	**Semana**					
Nombre	**16-04**	**23-04**	**30-04**	**07-05**	**14-05**	**21-05**
Lanzamiento	X					
Redacción de la documentación		X	X	X	X	X
Búsqueda de conjunto de datos		X				
Objetivo P-01					X	
Objetivo S-01			X	X		
Objetivo S-02			X	X		
Objetivo S-03			X	X		

2.3. ANÁLISIS DE LA VIABILIDAD

El desarrollo total del proyecto ha tenido con coste total en tiempo asociado a las diferentes fases por las que ha pasado el desarrollo del sistema, que en total han sido seis semanas, desde el inicio de la identificación de los objetivos hasta la ejecución de las pruebas del sistema. También ha tenido costes de personal, el cual ha trabajado en el proyecto desde su inicio hasta su conclusión. El equipo de trabajo ha sido conformado por tres personas.

El coste de la infraestructura ha sido gratuito, ya que cada una de las personas que ha intervenido en el proyecto lo ha hecho con sus propios recursos, sin tener que afrontar ningún coste en equipamiento para el desarrollo. El software y los datos que nos han servido para llevar a cabo los objetivos del sistema son abiertos, por lo que no ha tenido un coste asociado a su uso.

Se ha tomado la decisión de emplear lenguajes de programación y herramientas que ya eran conocidas por el equipo para reducir la curva de aprendizaje, que forzaría a tener un periodo de desarrollo más largo, ya que primero habría que aprender a utilizar los nuevos lenguajes y herramientas. Además, para el desarrollo se ha requerido utilizar ciertos lenguajes y herramientas por el uso de APIs externas que requieren el de los mismos.

Algunos de los lenguajes de programación y herramientas usados en el desarrollo han sido: JavaScript, jQuery, API Google Maps, OpenStreetMap, API Google Charts, Bootstrap, PHP, AJAX, KNIME Analytics o Pentaho Data Integration.

2.4. RIESGOS

La falta de algunos objetivos podría perjudicar al desarrollo de otros, ya que algunos basan su funcionalidad en la funcionalidad de otros. También podría influir en la nota del proyecto que se está desarrollando, pudiendo no legar a los mínimos establecidos.

Las tecnologías utilizadas han sido las siguientes:

- Para la ETL de los datos se ha usado la herramienta Pentaho Data Integration, que se conecta a la API de NYPD y descarga los datos a través de una URL.
- La visualización y reportes se han llevado a cabo a través del navegador web, diseñando la presentación con Bootstrap.
    - Para los gráficos se ha empleado la API de Google Charts, que se controla a través de JavaScript y jQuery.
    - Para los mapas se han usado las APIs de Google Maps y OpenStreetMaps, que también se controlan a través de JavaScript y jQuery.
    - Para el acceso a datos desde el navegador se ha empleado PHP y AJAX
- Para el clustering de los delitos y la minería de datos se ha usado la herramienta KNIME Analytics.
- El intercambio de información entre todos los subsistemas se realiza a través de ficheros CSV.

3. ANÁLISIS

3.1. ESTABLECIMIENTO DE LOS REQUISITOS DEL SISTEMA

Los requisitos funcionales de la solución son los siguientes:

RF-01	**Carga de datos automatizada desde la API de NYPD**
Autores	Juan Antonio Rodríguez, Alberto Cárdenas, Manuel Ridao
Fuentes	Necesidad del usuario
Objetivos asociados	• **OBJ P-01**: Optimizar la ubicación de las comisarías de policía • **OBJ S-01**: Visualizar las zonas más conflictivas y más seguras de la ciudad • **OBJ S-02**: Visualizar los datos en tiempo real con un cuadro de mandos • **OBJ S-03**: Diseñar un modelo predictivo para detectar la ofensa cometida
Descripción	El sistema deberá contemplar las siguientes funcionalidades: • Conexión desde Pentaho a la API de NYPD • Petición de datos a esta API mediante REST • Transformación de los datos para su uso posterior • Carga de datos procesados en varios ficheros

RF-02	Cálculo y muestra de ubicaciones óptimas de comisarías
Autores	Juan Antonio Rodríguez, Alberto Cárdenas, Manuel Ridao
Fuentes	Necesidad del usuario
Objetivos asociados	• **OBJ P-01**: Optimizar la ubicación de las comisarías de policía
Descripción	El sistema deberá contemplar las siguientes funcionalidades: • Clustering de ofensas según su ubicación geográfica. • Cálculo de los centroides de los clusters • Visualización de los resultados en un mapa • Mostrar el mapa a través del navegador

RF-03	Visualización de un mapa de calor de delincuencia
Autores	Juan Antonio Rodríguez, Alberto Cárdenas, Manuel Ridao
Fuentes	Necesidad del usuario
Objetivos asociados	• **OBJ S-01**: Visualizar las zonas más conflictivas y más seguras de la ciudad
Descripción	El sistema deberá contemplar las siguientes funcionalidades: • Visualización a través de un mapa de calor de las zonas más conflictivas de la ciudad • Mostrar dicho mapa a través del navegador

RF-04	Modelo predictivo para el tipo de ofensa cometida
Autores	Juan Antonio Rodríguez, Alberto Cárdenas, Manuel Ridao
Fuentes	Necesidad del usuario
Objetivos asociados	• **OBJ S-03**: Diseñar un modelo predictivo para detectar la ofensa cometida
Descripción	El sistema deberá contemplar las siguientes funcionalidades: • Construir un modelo predictivo para que se pueda predecir el tipo de ofensa cometida a través de los datos de entrada • Visualizar la fidelidad de dicho modelo a través de un navegador

3.2. ANÁLISIS DE CASOS DE USO

Los casos de uso de la aplicación se recogen en el siguiente diagrama

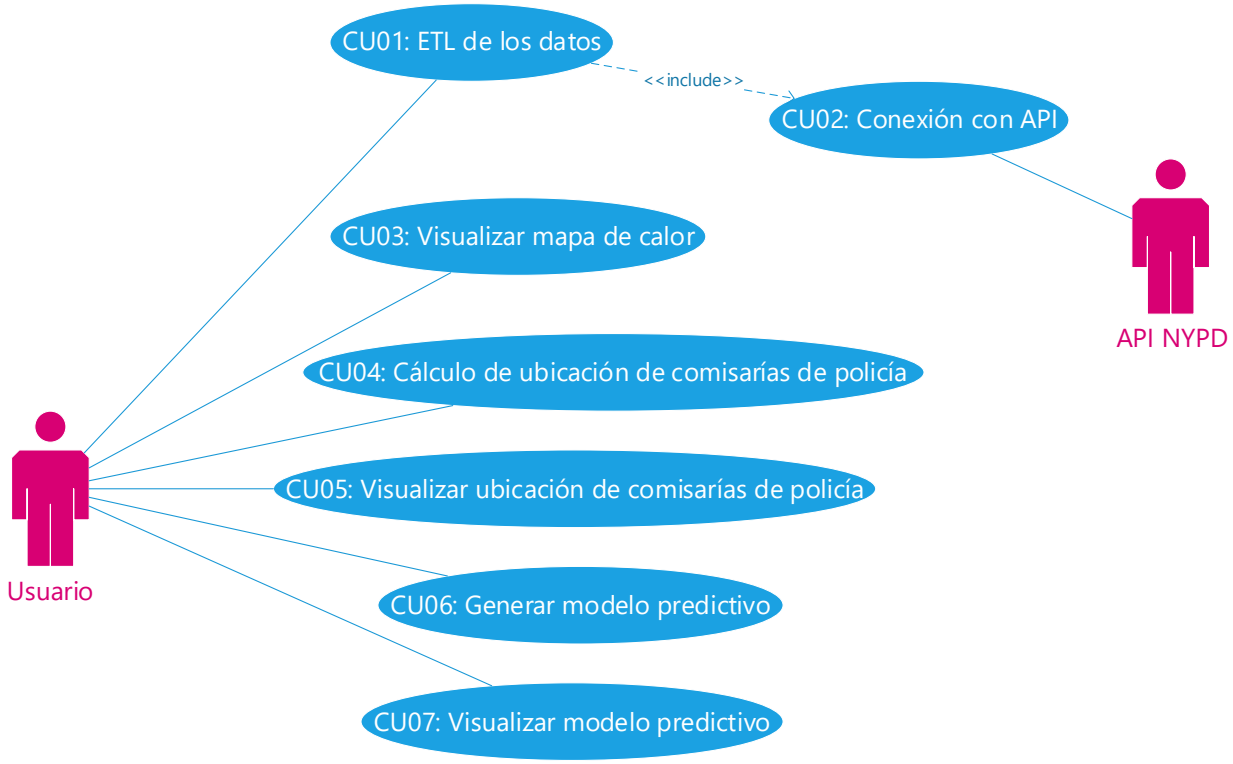


Figure 2: Diagrama de casos de uso

3.3. ESPECIFICACIÓN DEL PLAN DE PRUEBAS

Para las pruebas de integración e implantación, se probarán en sucesión las funcionalidades de la aplicación.

Pruebas de Integración e Implantación					
Id	**Pasos que seguir**	**Datos de entrada**	**Salida esperada**	**OK ?**	**Observaciones**
1	Lanzar la ETL con Pentaho	Ninguno	Los ficheros CSV: • list_of_offense_types • report_map_dataset • offense_groupby_boro • offense_groupby_boro_type • complete_dataset	OK	Se cargan los datos de la API y se guardan en ficheros. El proceso es lento porque hay muchos datos

2	Lanzar el servidor web y acceder a la URL de la aplicación	Ninguno	Ninguno	OK	-
3	Lanzar la construcción del modelo predictivo	Los ficheros CSV: • complete_dataset	Los ficheros CSV: • confusión_matrix • cluster_centroids	OK	El proceso es lento porque hay muchos datos
4	Lanzar el clustering de ofensas	Los ficheros CSV: • complete_dataset • nypd_precincts	Los ficheros CSV: • clustered_results • cluster_centroids	OK	El proceso es lento porque hay muchos datos
5	Visualizar los datos en el navegador	Todos los ficheros CSV	Ninguna	OK	El proceso es lento porque hay muchos datos

Pruebas del sistema				
ID	**Descripción**	**Resultado**	**OK?**	**Observaciones**
1	Probar el sistema en distintos equipos	El sistema se comporta del mismo modo en distintos equipos	OK	-
2	Probar la parametrización del sistema	La parametrización funciona correctamente	OK	KNIME Analytics no permite la parametrización de rutas, por lo que es necesaria su introducción de manera manual

4. DISEÑO

4.1. CAPACIDAD DE MEMORIA DE LA ORGANIZACIÓN

Por un lado, los datos principales han sido extraídos mediante un servicio web REST que proporciona el NYPD. Los datos recogidos se proporcionan en formato JSON y son recogidos directamente por la herramienta de transformación ETL Pentaho Data Integration. Una vez los han sido recogidos los datos por la herramienta, ésta hace una trasformación de los mismo ofreciendo los datos ya tratados. Solo se ha realizado una transformación ETL, que realiza varios procedimientos y genera varias salidas.

Por otro lado, se han tomado datos y creado de manera manual los siguientes ficheros. Estos datos no se recogen de manera automática ya que no se prevén cambios próximos en sus valores:

1. Datos de población de NYC por barrio[4]: Fichero CSV "nyc_pop_by_boro" que recoge datos de población y densidad de población por cada barrio de la ciudad.
2. Número de precintos policiales de la ciudad[5]: Fichero CSV "nypd_precincts" que recoge el número de comisarías de policías actualmente en la ciudad.
3. Agrupación de ofensas: Fichero de referencia que recoge como se han agrupado las 63 ofensas originales de los datos de NYPD en 10 ofensas más genéricas

4.2. CAPACIDAD DE INTEGRACIÓN DE INFORMACIÓN

Una vez que la herramienta Pentaho Data Integration ha recolectado los datos mediante el servicio web, empieza la etapa de transformación de los datos.

Partiendo de todos los datos recogidos, la primara trasformación de los datos es quitar aquellas filas que tienen algún campo vacío. A partir de aquí se abren tres flujos distintos:

1. Se sustituye el valor del atributo "Offense" de cada registro, poniendo el valor de este atributo uno más genérico que abarca distintos valores del atributo "Offense". Por ejemplo, los delitos "Theft", "Larceny" y "Burglary" se han agrupado como "Theft". Genera un fichero "complete_dataset" con extensión CSV que será usado por la herramienta de minería de datos KNIME.

2. Se selecciona sólo las columnas que nos interesan en este flujo (Boro, Offense y Law Cat). Este, a su vez, se divide en otros dos flujos distintos:
   2.1. Hace una ordenación por barrio y realiza una suma de la cantidad de delitos cometidos en cada barrio. Genera un fichero "offenses_groupby_boro", con extensión CSV con dos columnas: Nombre del barrio y número de delitos cometidos en él.
   2.2. Hace una ordenación por barrio y realiza una suma de la cantidad de delitos cometidos en cada barrio según el tipo de delito cometido. Genera un fichero "offenses_groupby_boro_type", con extensión CSV con tres columnas: Nombre del barrio, tipo de delito cometido y la suma total de los delitos cometidos.

3. Se selecciona sólo las columnas que nos interesan en este flujo (Boro, Offense, Law Cat, Latitude y Longitude) y, además, sustituye el valor del atributo "Offense" de cada registro, poniendo el valor de este atributo uno más genérico que abarca distintos valores del atributo "Offense". Éste a su vez se divide en otros dos flujos distintos:
   3.1. Realiza una ordenación por el campo "Offense" y agrupa todos los registros en función de este campo, seleccionado solamente éste campo como salida para el fichero. Genera un fichero "list_of_offense_types" con extensión CSV con la columna "Offense".
   3.2. Genera un fichero "report_map_dataset" con extensión CSV con las columnas Boro, Offense, Law Cat, Latitude y Longitude.

[4] https://www.citypopulation.de/php/usa-newyorkcity.php
[5] https://www1.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page

4.3. CAPACIDAD DE CREAR CONOCIMIENTO

NOMBRE DEL OBJETIVO	DATOS RESULTANTES
OBJETIVO P-01 OPTIMIZAR LA UBICACIÓN DE LAS COMISARÍAS DE POLICÍA	Se mostrarán las ubicaciones idóneas para situar las comisarías de policía a lo largo de la ciudad de Nueva York en función de los delitos cometidos en las distintas localizaciones de la ciudad, para ello, se realizará un clustering de las ofensas empleando la técnica k-Means y hallando su centroide.
OBJETIVO S-01 VISUALIZAR LAS ZONAS CONFLICTIVAS Y MÁS SEGURAS DE LA CUIDAD	Se mostrarán las ubicaciones donde se han cometido delitos en el último año a lo largo de la ciudad de Nueva York, mostrándose claramente las zonas más conflictivas y más seguras de la ciudad.
OBJETIVO S-02 VISUALIZAR LOS DATOS EN TIEMPO REAL EN UN CUADRO DE MANDOS	Se mostrará información resumida de los datos que se han estudiado, mostrando algunas estadísticas como qué barrio es donde se cometen más delitos en relación con la población residente, que delitos son los más cometidos en la ciudad, entre otros.
OBJETIVO S-03 DISEÑAR UN MODELO PREDICTIVO PARA DETECTAR LA OFENSA COMETIDA	Se mostrarán los datos relativos a la predicción, pudiendo ver a simple vista la eficacia del modelo predictivo, viéndose en términos porcentuales y numéricos.

4.4. CAPACIDAD DE PRESENTACIÓN

La presentación de la solución de inteligencia de negocio desarrollada se muestra mediante un cuadro de mando con una interfaz web que realiza la visualización de los datos bajo demanda. El cuadro de mando dispone de un menú donde se podrá seleccionar el tipo de información a mostrar en cada momento dependiendo de las necesidades del usuario.

Cada elemento del menú ha sido asociado a cada uno de los objetivos del sistema, mostrando la información de la siguiente manera en cada uno de ellos:

NOMBRE DEL OBJETIVO	VISUALIZACIÓN DE LA INFORMACIÓN
OBJETIVO P-01 OPTIMIZAR LA UBICACIÓN DE LAS COMISARÍAS DE POLICÍA	Se visualizará sobre un mapa de la ciudad de Nueva York, y cada una de las comisarías estará representada por un icono que sitúa geográficamente la posición exacta de cada una de ellas.
OBJETIVO S-01 VISUALIZAR LAS ZONAS CONFLICTIVAS Y MÁS SEGURAS DE LA CUIDAD	Se visualizará en forma de un mapa de calor sobre la ciudad de Nueva York, tomando tonos más fuertes en las zonas donde hay una mayor aglomeración de delitos y más suaves donde se cometen menos delitos en términos de distancia entre ellos.

OBJETIVO S-02 VISUALIZAR LOS DATOS EN TIEMPO REAL EN UN CUADRO DE MANDOS	Se visualizará mediante gráficos de diverso tipo, dependiendo de los datos que muestren. Se usará un gráfico de barras para representar la cantidad de delitos cometidos en un barrio en función del tipo de delito, de tarta para representar la cantidad total de delitos por barrio y de burbujas para representar la relación de delitos y habitantes por cada uno de los barrios.
OBJETIVO S-03 DISEÑAR UN MODELO PREDICTIVO PARA DETECTAR LA OFENSA COMETIDA	Se visualizará en dos tablas diferenciadas, la primera mostrará mediante una matriz de confusión la precisión del modelo predictivo, especificando los tipos de resultados obtenidos, y la segunda, mostrará en términos porcentuales la precisión del modelo predictivo, especificando los tipos de resultados obtenidos.

5. IMPLEMENTACIÓN

La implementación de la solución de IN se ha dividido en tres fases: fase de obtención de datos, fase de análisis de datos, y fase de presentación.

5.1. FASE DE OBTENICIÓN DE DATOS

Los datos se han obtenido directamente desde la API de NYPD a través de la herramienta de Pentaho, tal y como se describe en el punto 4.1. Para ello han sido claves tres nodos incluidos en esta herramienta:

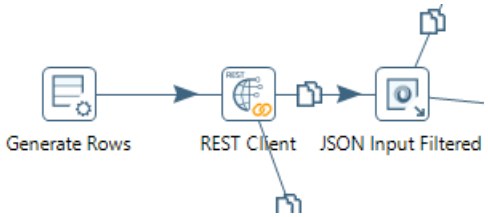


Figure 3: Nodos de entrada de datos

5.1.1. Generate rows

Permite introducir valores directamente a la transformación. Este nodo se emplea para proporcionar la URL a la que realiza la petición REST. El NYPD permite descargas directamente de sus datos mediante una petición a la siguiente URL:

https://data.cityofnewyork.us/resource/7x9x-zpz6.json?$where=lat_lon%20is%20not%20null&$limit=999999

5.1.2. REST Client

Nodo que implementa un cliente RESTful. En este caso, realiza una petición GET a la URL anterior, que devuelve los datos en formato JSON.

5.1.3. JSON Input

Transforma los datos obtenidos en formato JSON a formato tabla tradicional, desde la cual se realizan todas las transformaciones.

Estos datos son transformados como se explica en el apartado 4.2. Este proceso general varios ficheros que se usan en pasos posteriores. El flujo de información entre aplicaciones puede consultarse en el punto 6.

5.2. FASE DE ANÁLISIS DE DATOS

En esta fase, se emplea el programa KNIME Analytics para realizar dos análisis de datos, el clustering y la minería de datos.

5.2.1. Clustering de ofensas

Para hallar la ubicación óptima de las comisarías, se ha realizado un clustering de las ofensas y hallado su centroide. Así, las comisarías caen en el centro de masa de cada clúster. Esto se ha implementado mediante el algoritmo k-Means, para el que KNIME incorpora un nodo, agrupando por latitud y longitud.

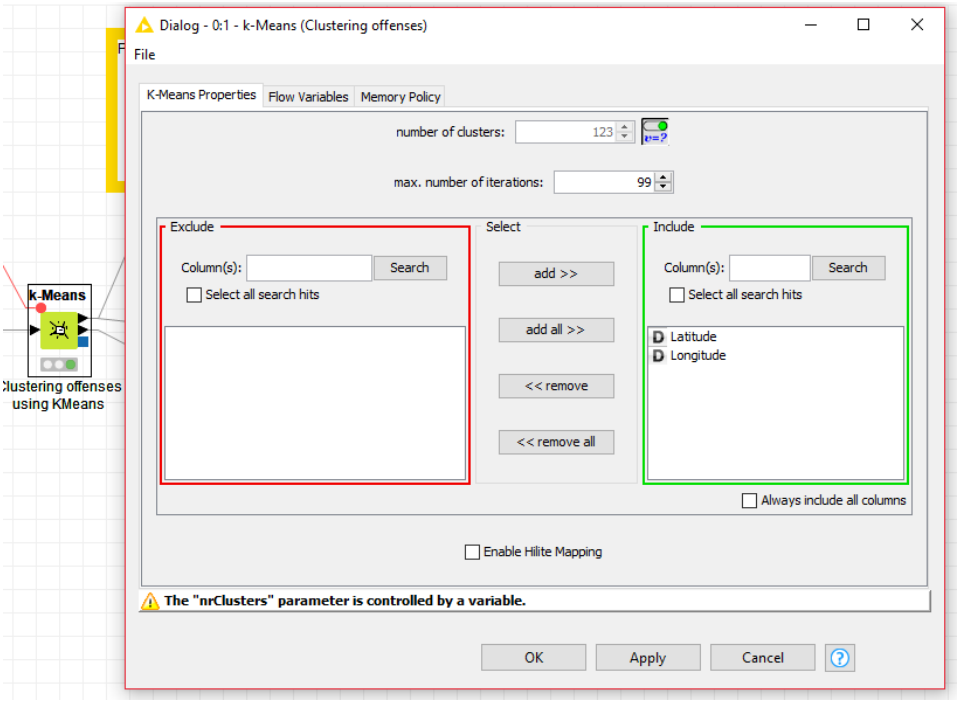


Figure 4: Nodo k-Means y parametrización

5.2.2. Minería de datos

Se ha diseñado un modelo predictivo para averiguar el tipo de delito (por ejemplo: robos, tráfico, drogas, entre otros) según los datos proporcionados por la policía. Para ello, se ha seguido el siguiente procedimiento:

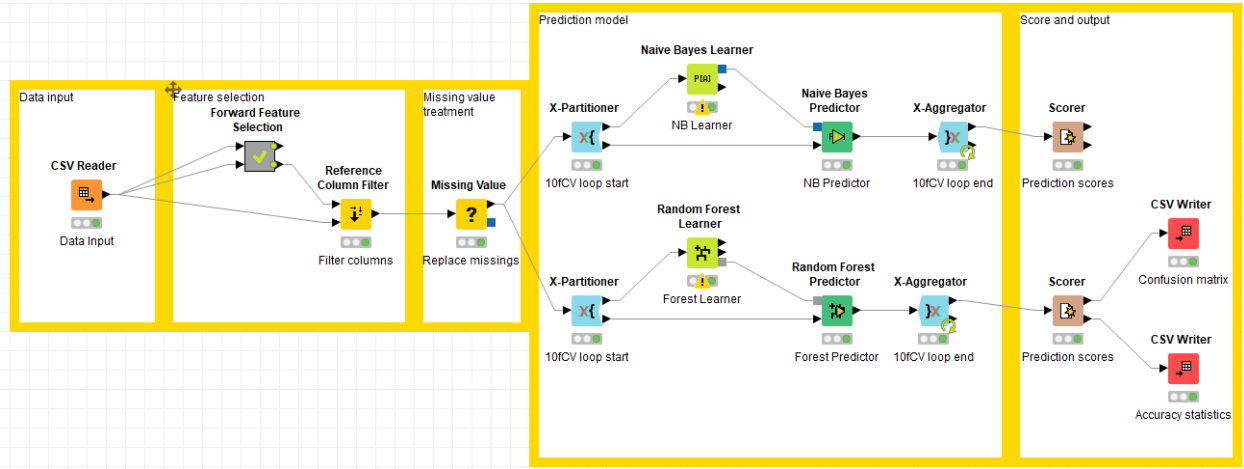


Figure 5: Modelo predictivo de KNIME

El funcionamiento del modelo es el siguiente:

Paso	Nodos	Descripción
Carga de datos	CSV Reader	Carga datos desde un CSV
Selección de atributos	Forward Feature Selection (metanodo)	Realiza una selección de los atributos más relevantes según el valor de clase que se proporcione
	Reference Column Filter	Filtra las columnas que devuelve el nodo anterior, quitando las no relevantes del dataset.
Valores perditos	Missing Value	Corrige los valores perdidos aplicando medias a los valores numéricos y modas a los categóricos
Modelo predictivo: Naive-Bayes	X-Partitioner y X-Aggregator	Implementan los bucles de validación cruzada para entrenar el modelo
	Naive Bayes Learner	Nodo que crea un modelo predictivo mediante Naive-Bayes con el conjunto de entrenamiento
	Naive Bayes Predictor	Evalúa el modelo predictivo con el conjunto de test
Modelo predictivo: Random Forest	X-Partitioner y X-Aggregator	Implementan los bucles de validación cruzada para entrenar el modelo
	Random Forest Learner	Nodo que crea un modelo predictivo mediante árboles de decisión aleatorios con el conjunto de entrenamiento
	Random Forest Predictor	Evalúa el modelo predictivo con el conjunto de test
Resultados y output	Scorer	Evalúa los resultados del bucle de aprendizaje, proporcionando estadísticas de precisión y matriz de confusión
	CSV Writer	Escribe los resultados del Scorer en un fichero CSV

Se ha optado por recoger los resultados del Random Forest, ya que ofrece los mejores resultados, pero se han dejado visibles los resultados de Naive-Bayes para realizar comparaciones. El motivo de esto es que el metamodelo de Random Forest incopora muchos modelos de árboles de decisión. Al combinar los resultados de varios árboles, el resultado general siempre es mejor que el que un solo árbol sería capaz de generar.

No obstante, el rendimiento del modelo es bastante mediocre. La precisión del modelo predictivo apenas supera 0.55. Esto puede deberse, entre otros motivos, a la alta población y densidad de población de la ciudad, que no permite aislar focos de criminalidad ni de tipos de crimen, y a la espontaneidad e irregularidad de la obtención de datos, ya que no todos los delitos pueden registrarse. En conclusión, no se considera viable la predicción del tipo de delitos obtenido mediante esta tecnología, y será necesario un análisis más en profundidad, en coordinación con las fuerzas policiales y jurídicas.

5.3. FASE DE PRESENTACIÓN

La presentación de resultados se ha realizado a través de una aplicación web visible en el navegador. El desarrollo de la aplicación se ha llevado a cabo a través de NetBeans[6].

Componente de la aplicación	Tecnología empleada	Motivo
Front-End	JavaScript + jQuery	Realizar peticiones directamente a las API de Google y OpenStreetMaps, lo que permite visualizar los datos a través de mapas y gráficos comodamente
Back-End	Ajax + PHP	Con Ajax se pueden realizar peticiones a datos sin recargar la página. El acceso a datos se ha llevado a cabo mediante ficheros con PHP

6. DESPLIEGUE

El despliegue de la aplicación se lleva a cabo según el siguiente diagrama:

[6] https://netbeans.org/

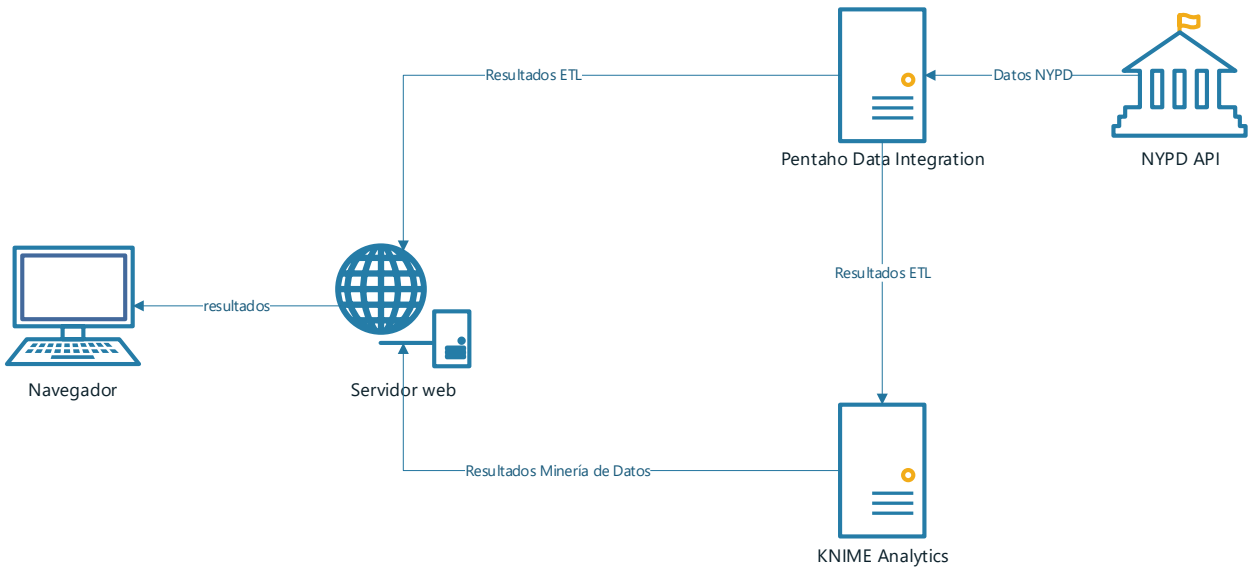


Figure 6: Diagrama de despliegue

En primer lugar, Pentaho Data Integration se conecta a la API ofrecida por NYPD para descargar los datos y realizar las transformaciones. Parte de los resultados de las transformaciones se emplean en KNIME Analytics para construir un modelo predictivo y realizar los clusterings. Finalmente, los datos de KNIME y el resto de los datos de las transformaciones se cargan mediante un servidor web para ser mostrados a través del navegador.

Los casos de uso de la aplicación y su relación con los dispositivos es la siguiente:

Identificador	Caso de uso	Material
CU01	ETL de los datos	Pentaho Data Integration y NYPD API
CU02	Conexión con API	Pentaho Data Integration
CU03	Visualizar mapa de calor	Servidor Web y navegador
CU04	Cálculo de ubicación de comisarías de policía	KNIME Analytics Platform
CU05	Visualizar la ubicación de comisarías de policía	Servidor Web y navegador
CU06	Generar modelo predictivo	KNIME Analytics Platform
CU07	Visualizar resultados de modelo predictivo	Servidor Web y navegador

7. CONCLUSIONES

Mediante esta solución se permite, de una manera integrada y semi-automatizada, la carga, comprensión y visualización de los datos recolectados por el NYPD. Por un lado, con muy poca preparación técnica necesaria, se ofrece la posibilidad de descargar los datos directamente desde la fuente, así como su tratamiento para su posterior uso. Para estos procesos solo son necesarios dos aplicaciones software. Por otro lado, la visualización se realiza mediante un navegador web, disponible en todos los equipos. Esta visualización se proporciona de manera simple y gráfica, para que pueda ser interpretada por cualquier usuario, independientemente de su formación.

Debido a la creciente integración tecnológica de los servicios públicos, la cantidad y complejidad de datos generados se ha visto aumentada, así como su capacidad de almacenamiento. Por ello, se hace necesario una solución de inteligencia de negocio que permita su interpretación para mejorar el rendimiento de estos servicios y, por lo tanto, la calidad de vida de las personas que los disfrutan.

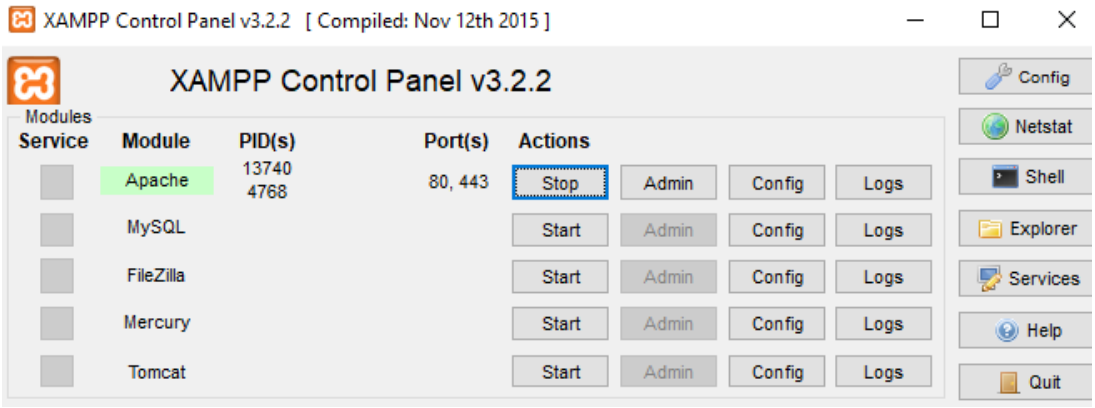
8. MANUAL DE USO

8.1. SOFTWARE NECESARIO Y PUESTA EN MARCHA

Es necesario tener instalado un servidor web[7] para poder acceder al cuadro de mando. Una vez instalado XAMPP, es necesario copiar la carpeta IN_Proyecto completa en la carpeta xampp/htdocs. Se puede acceder al cuadro de mando de forma local a través de la siguiente ruta:

http://localhost/IN_Proyecto/CuadroDeMando/index.php

Debe de estar iniciado el servicio Apache para que la URL anterior funcione. Además, es necesario disponer de una conexión a internet para descargar los datos y acceder a los servicios de Google Charts, Google Maps y OpenStreetMaps.

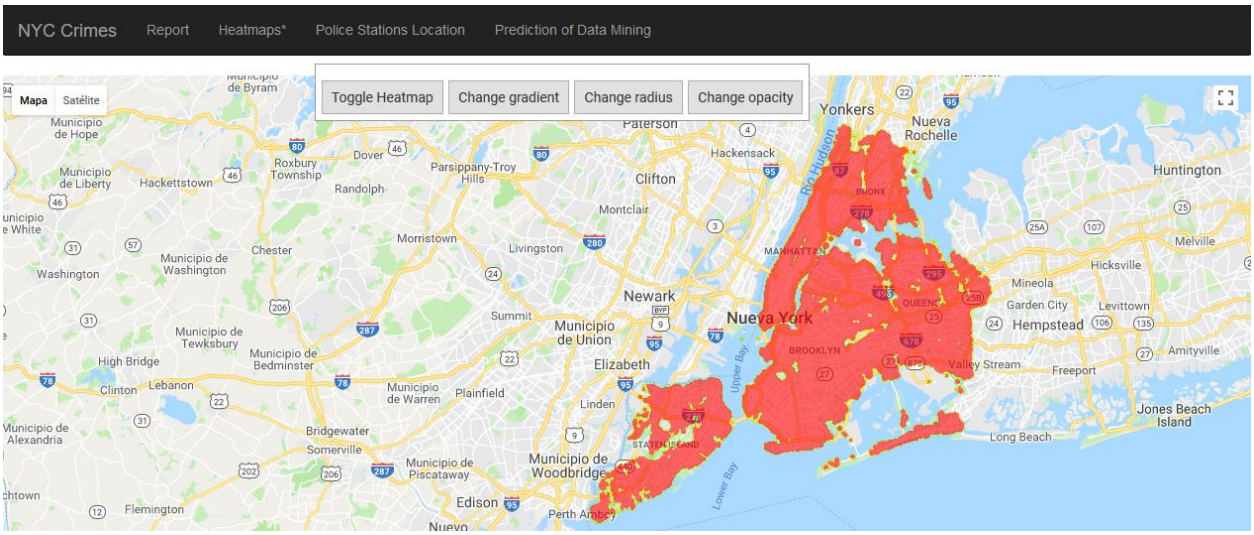


8.2. VENTANAS DEL CUADRO DE MANDO

8.2.1. Ventana reporte (*Report*)

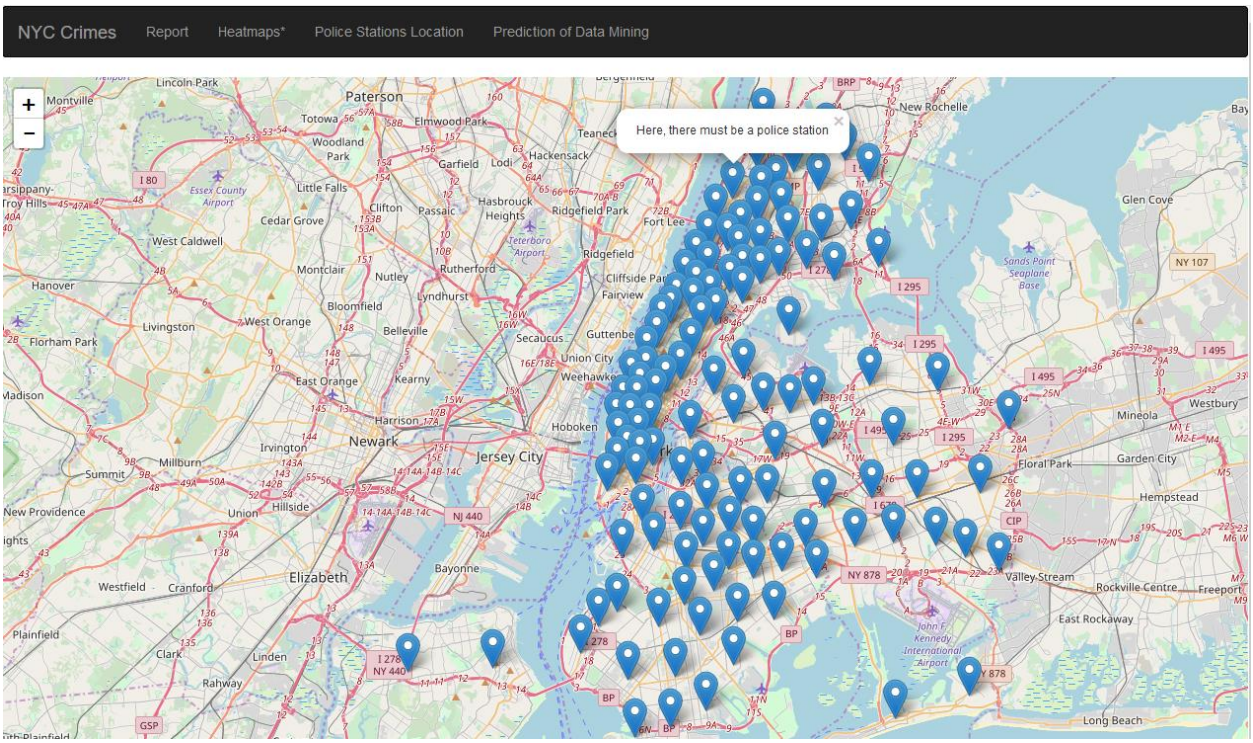
[7] https://www.apachefriends.org/index.html

8.2.2. Ventana mapa de calor (*Heatmaps*)

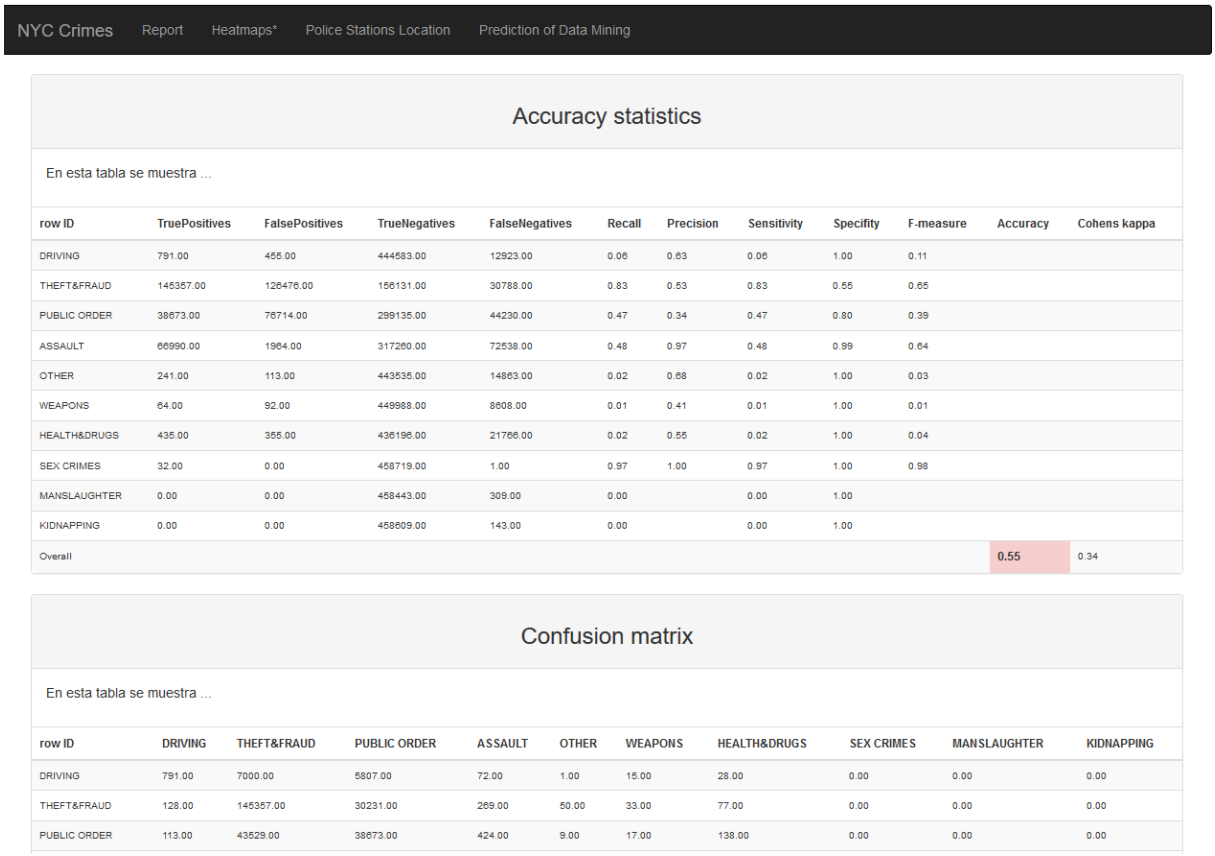


Una vez termina de cargar los datos es necesario pulsar el botón "*Toggle Heatmap*" para que muestre en el mapa los puntos de calor.

8.2.3. Ventana localización de comisarias (*Police Stations Location*)



8.2.4. Ventana predicción de minería de datos (*Prediction of Data Mining*)

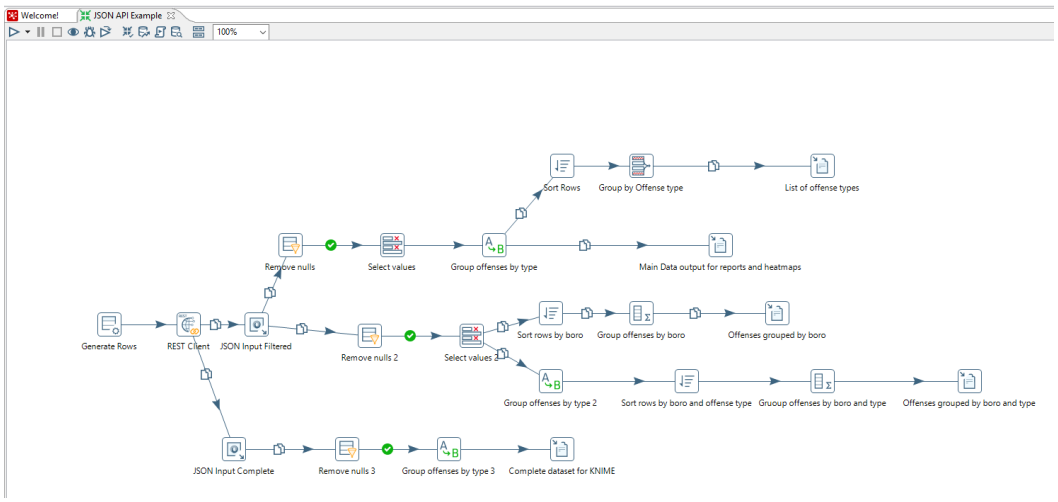


8.3. ACTUALIZACIÓN DE DATOS

Con el proyecto se incluye una serie de ficheros con datos actualizados a la fecha de realización de este, para actualizarlos es necesario generar los ficheros con los datos nuevos usando Pentaho Data Integration[8] y KNIME Analytics[9].

Una vez instalados, es necesario abrir con Pentaho el fichero "*JSON API Example v2.ktr*", ubicado dentro del proyecto en "*CuadroDeMando/csv/JSON API Example v2.ktr*".

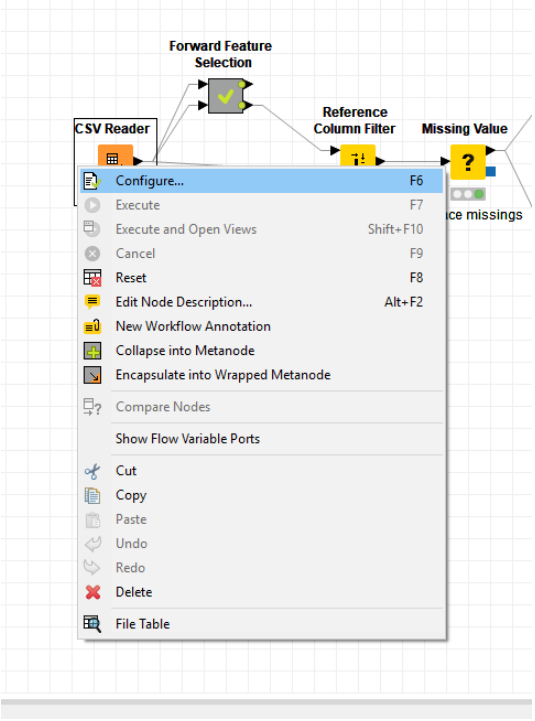
[8] https://community.hitachivantara.com/docs/DOC-1009855
[9] https://www.knime.com/knime-analytics-platform

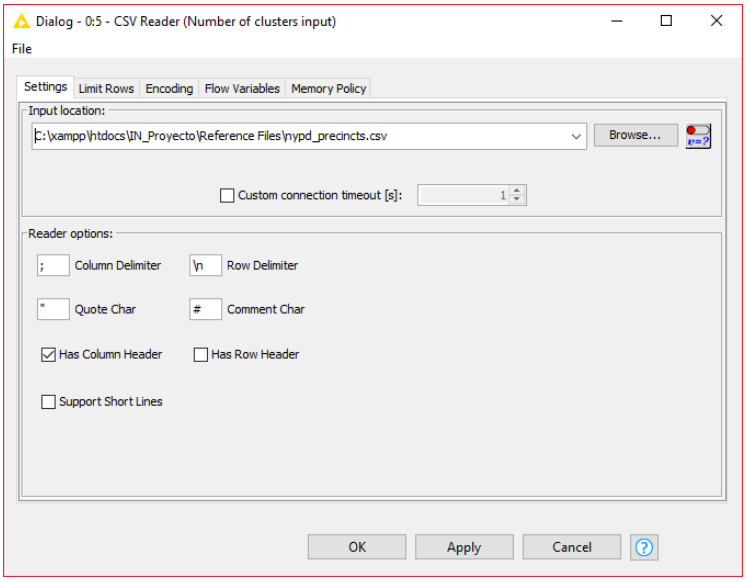


Una vez abierto, se ejecuta mediante el botón "*Play*", lo que genera automáticamente los ficheros con los datos extraídos y tratados desde la API, los cuales serán usados para mostrar información en el cuadro de mando. Este proceso puede demorarse un tiempo debido a la cantidad de datos.

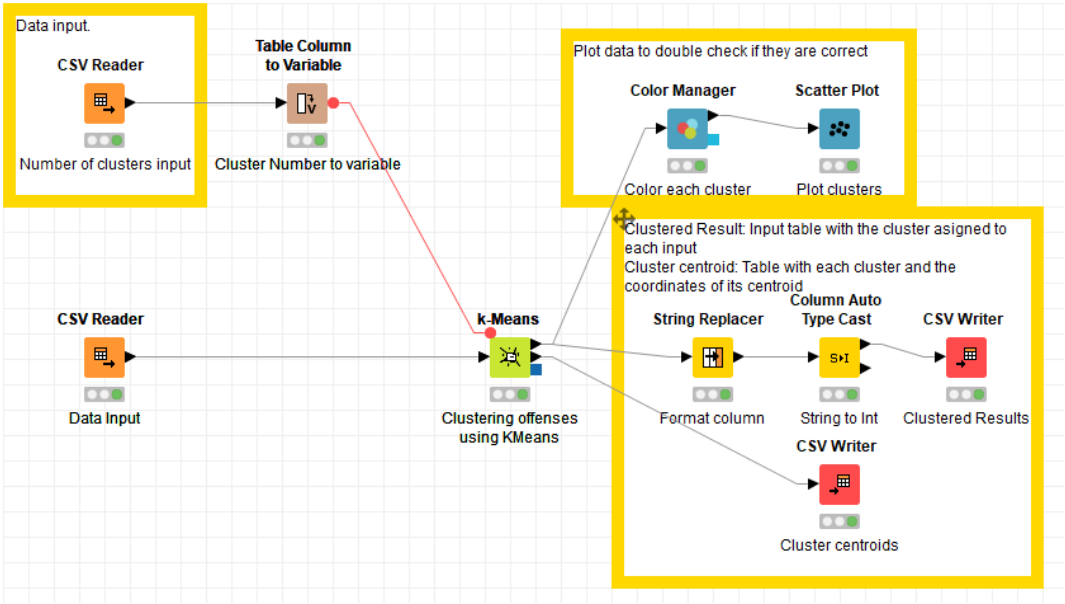
8.4. KNIME ANALYTICS

Con este programa se crean los ficheros necesarios para los objetivos P-01 y S-03. Los workflows necesarios se encuentran en el directorio "KNIME Workflows". Es necesario concretar la ruta de los ficheros de entrada y salida, ya que KNIME no permite parametrizar estos valores. Para ello, se hace clic derecho en los nodos CSV Reader y CSV Writer, y se introduce el valor deseado.



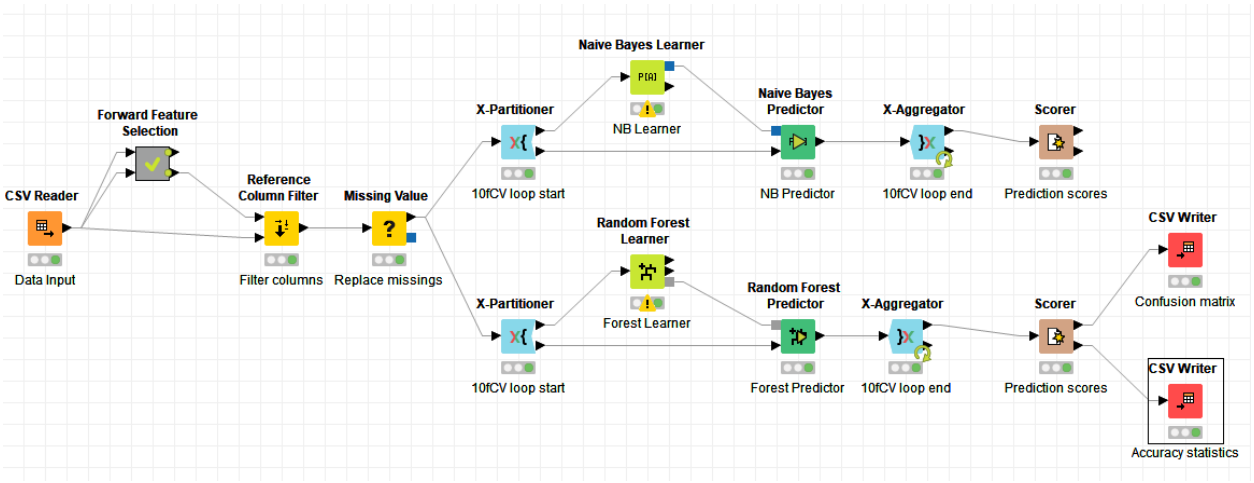


8.4.1. Objetivo P-01: Clustering de delitos



Tipo de nodo	Nombre	Fichero
CSV Reader	Number of clusters input	IN_Proyecto\Reference Files\nypd_precincts.csv
	Data Input	IN_Proyecto\CuadroDeMando\csv\report_map_dataset.csv
CSV Writer	Clustered Results	IN_Proyecto\CuadroDeMando\csv\clustered_results.csv
	Cluster Centroids	IN_Proyecto\CuadroDeMando\csv\cluster_centroids.csv

8.4.2. Objetivo S-03: Minería de datos



Tipo de nodo	Nombre	Fichero
CSV Reader	Data Input	IN_Proyecto\CuadroDeMando\csv\report_map_dataset.csv
CSV Writer	Confusion Matrix	IN_Proyecto\CuadroDeMando\csv\confusion_matrix.csv
	Accuracy statistics	IN_Proyecto\CuadroDeMando\csv\accuracy_statistics.csv

Se puede comprobar la calidad de los modelos predictivos haciendo clic derecho en los nodos "Scorer" y seleccionando *Confusion Matrix* y *Accuracy Statistics*.