ICEMD
INSTITUTO ECONOMÍA DIGITAL | ESIC

Autor: 135430

Análisis y predicción de la contaminación del aire en Madrid en base al tráfico y meteorología

Trabajo de fin de máster en Big Data Management 2018/2019
por Alexander Balseiro Ramos

Este trabajo se entrega en cumplimiento parcial de los requisitos para la obtención del título Máster en Big Data Management

Índice

Índice 1

Prólogo 3

Introducción 4

Estado del arte del Big Data 5

¿Qué es el Big Data? 5

Historia del Big Data 6

Situación en España 7

Situación de la contaminación del aire en Madrid 8

Desarrollo 10

Objetivo 10

Tipología y estructura de las fuentes de datos 11

Datos de tráfico 11

Datos de contaminación del aire 11

Datos de observación 12

Datos de predicción 13

Arquitectura 15

Esquema de la arquitectura 15

Flujo de los datos 15

Python scripts de ingesta 16

Flume 17

HDFS 17

Spark 17

H3-Uber 17

Python H3 17

Hive 18

Elasticsearch 18

Kibana 18

Colab 18

Ingesta y procesamiento 19

Visualización **21**
Capa de tráfico 21
Capa de contaminación del aire actual 22
Modelo predictivo **24**
Datos 24
Objetivo 27
Primera fase del modelado 28
Conclusión 34
Recursos de infraestructura **36**
Data Governance **37**
Linaje de los datos 37
Datos de tráfico 37
Datos de contaminación del aire 39
Datos de observación 40
Datos de predicción 42
Data legal 44
Seguridad 44
Planificación del proyecto **44**
Casos de uso y monetización **47**
Explotación y seguimiento por parte del ayuntamiento de Madrid 47
Real Estate 48
Transporte privado 48
Retorno de la inversión **49**
Conclusiones **51**
Trabajo futuro **52**
Bibliografía **53**

Prólogo

Todo el código, ejecutables y documentación relacionada con el proyecto puede consultarse en el siguiente enlace del GitHub de Alexander Balseiro Ramos.
Proyecto Análisis y predicción del tráfico y la calidad del aire en Madrid

Introducción

El tráfico en la ciudad y la contaminación es una problemática que preocupa mucho. En la ciudad de Madrid ha sido un tema muy recurrente de desde un tiempo a la actualidad. Es sabido que el tráfico influye en la contaminación y las restricciones de esta en momentos de un alto porcentaje también influyen de manera inversa en el tráfico.

Por todo ello, existe la necesidad de poder analizar y predecir de qué manera evolucionan ambos visualmente en la ciudad de Madrid y ver si la meteorología pudiera influir en ellos.
Este estudio podría contribuir:

- En la gestión que realiza el ayuntamiento de Madrid y la comunidad con el transporte público y el control de emisiones tanto de vehículos como de industria
- En el sector del Real State y cómo influye en la valoración de inmuebles teniendo en cuenta también la calidad de vida que ofrecen.
- A la gestión de transporte privado y poder ofertar alternativas ecológicas en puntos de alta congestión de tráfico y contaminación.

Antecedentes

Estado del arte del Big Data

¿Qué es el Big Data?

El término Big Data se puede entender por la ingesta, procesamiento y análisis de grandes volúmenes de datos a gran velocidad (incluso en tiempo real), con la intención de generar toma de decisiones para infinitas finalidades. Siendo este proceso escalable al volumen que sea necesario sin que ello afecte al rendimiento, gestión o al coste de un modo exagerado.

Cierto es que haya donde se busque este término existirá una definición diferente (hablarán de diferentes tecnologías o finalidades como direccionado a cliente, automatización de procesos, reducción de riesgos, etc.), pero esta acerca una idea general para cualquiera que desconozca el término y pregunte por él.

En el Big Data cambia el concepto de los datos y con ello también las bases de datos relacionales como Oracle y My SQL entre otros. Ya que estos volúmenes, dejan de ser manejables de manera eficiente y rápida. A raíz de esto entre otras cosas, surgen las bases de datos no relacionales donde se puede guardar y consultar mucho mejor y con mayor variedad (datos clave-valor, documentos, textos de redes sociales, etc.). Esto genera una evolución de las herramientas tradicionales del Business Intelligence que empiezan a quedarse atrás en frente a nuevas que van surgiendo para saciar la necesidad generada.

Dicha evolución afecta como no a las figuras que hasta ahora estaban en el Business Intelligence para dar cabida a unas nuevas como por ejemplo, Data Engineer o Data Scientist.

Todo ello con el objetivo de cumplir las principales necesidades del Big Data recogidas en las tres Vs. Es probable que haya más o diferentes, pero estas son sin lugar a dudas donde la mayoría converge: Volumen, Variedad y Velocidad.

El Big Data se ha encontrado con un nivel de popularidad en auge en el que muchas de las grandes empresas ven una oportunidad real a la que unirse. Ya que ofrece un gran poder, con la posibilidad analizar cantidades ingentes de datos de distintas procedencias como documentos, logs, emails, redes sociales, imágenes, ubicaciones, etc. De este modo generar tomas de decisiones mucho más rápidas y eficientes de lo que se venía haciendo y además adoptar otras muchas más, que hasta ahora no eran posibles, ampliando el campo de oportunidades.

Historia del Big Data

La historia de los datos lleva ya un largo recorrido pero no así tanto con el concepto de Big Data de la mano.

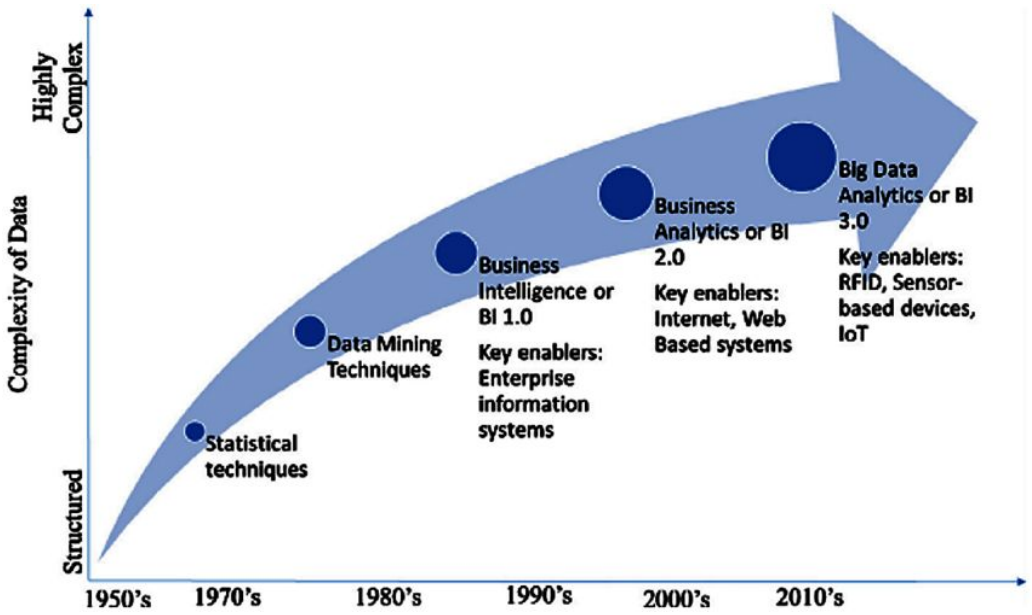


Figura 1: Evolución del Big Data

Se puede decir que todo empieza en 1997 con dos investigadores de la NASA, tras entender que existía un problema al que había que darle solución, "El problema del Big Data". Donde reconocen que existe una problemática con el aumento de los datos y los sistemas informáticos del momento.

Entre finales de los 90 y principios de los 00 se empiezan a ver los primeros avances en el "Cloud Computing" con los servicios SaaS (Software as a Service) de la mano de Salesforce y los primeros desarrollos de Amazon con AWS (Amazon Web Service) ofreciendo una serie de servicios en la nube entre ellos el almacenamiento.

En 2001 se publica un artículo llamado "3D Data Management: Controlling Data Volume, Velocity and Variety" en el que se habla de las tres Vs, hasta el día de hoy, los pilares comúnmente aceptados por el Big Data: Volumen, Velocidad y Variedad.

En 2003 y 2004 Google publica GFS (Google File System, un sistema de archivos distribuidos en cloud) y MapReduce (un framework para procesar grandes volúmenes de datos de forma paralela), dos de los pilares fundamentales de las tecnologías Big Data. Más tarde Google da a

conocer “Bigtable” un servicio de base de datos NoSQL, la cual muchos consideran el inicio de otras bases de datos NoSQL que hoy en día se usan en entornos Big Data.

2006 nace Hadoop, una solución de código abierto para un entorno Big Data basado en los dos pilares (GFS y MapReduce). Este es un método que permite el almacenamiento y procesamiento de enormes cantidades de datos en paralelo, distribuido en servidores que permiten escalar sin limite bajo un coste económico. En ese año también se empiezan a ver soluciones cloud de Google con “Google Docs” o EC2 (Elastic Compute Cloud) de Amazon con un servicios de infraestructura en la nube.

El término “Big Data” comienza a utilizarse cada vez con más frecuencia en el sector tecnológico en 2008. Entre 2009 y 2011 entran en juego en el mercado, Cloudera y Hortonworks, dos de los proveedores que ofrecen la posibilidad de almacenar, centralizar y gestionar de forma segura los datos, basado en Apache Hadoop.

En 2011 El Big Data ya está considerada una de las principales tendencias emergentes. En el artículo de “Big Data: The Next Frontier for Innovation, Competition, and Productivity”, de ese mismo año, se habla de las grandes cantidades de datos que ya empiezan a manejar las grandes empresas americanas.

En 2014 surge el término “Location Intelligence” para describir la utilización de los datos geolocalizados para el análisis y así facilitar la toma de decisiones. Ese mismo año Gartner elimina como tecnología emergente al Big Data, ya que empieza a conocerse y comprenderse mejor más allá de la moda, y empiezan a aparecer como emergentes otras partes del Big Data como “Machine Learning”, “Data Science”, “Internet Of Things”,...

“Smart Cities” o Ciudades Inteligentes (2015), las ciudades empiezan a tener una cantidad altísima de dispositivos conectados generando grandes cantidades de datos para analizar en tiempo real y así poder mejorar la gestión de las ciudades (servicios urbanos, iluminación LED, cerraduras electronicas, estudios de contaminación,...).

Situación en España

El Big Data no lleva muchos años en España. Actualmente está empezando a liderar una transformación digital basada en el dato. Las grandes empresas de sectores como finanzas o telecomunicaciones son las que están en cabeza de esta transformación y aprendiendo el uso y gestión de los grandes volúmenes de datos que guardan y generan en tiempo real. Adaptándose, por el camino, a nuevas reglas como la de protección de datos (GDPR, la nueva ley que ha entrado en vigor en Europa y que protege mucho más la privacidad de los datos de los usuarios) que ha salido en 2018.

Las empresas españolas se encuentran en un periodo de madurez bajo-medio por lo general en la adopción del Big Data. Los dos motivos principales que ven las empresas para abrir carpeta con Big Data son la relación con el cliente y la automatización de procesos.

Muchas de ellas ya utilizan cloud de forma completa o parcial para el almacenamiento de datos, es decir usan soluciones puramente cloud o soluciones híbridas entre cloud y on-premise (que usan bases de datos, herramientas y servidores propios).

La mayoría de empresas basan sus análisis y modelos Big Data en datos internos, no tanto en datos externos o fuentes de datos más desestructuradas como redes sociales, imágenes, vídeos, etc.

A futuro cercano, se prevé que el Big Data se vaya consolidando. Ocupando un lugar más importante en la ayuda de toma de decisiones, siendo empresas más “Data Driven” o dirigidas por los datos, y con sistemas automatizados haciendo foco en el cliente.

Situación de la contaminación del aire en Madrid

Los últimos años la contaminación del aire en Madrid a tomado un lugar importante en la vida de los habitantes y no habitantes de esta ciudad. Todo aquel que interacciona con la ciudad se preocupa o se interesa medianamente por el estado del aire. Ya no solo sus habitantes sino diversos organismos públicos como el ayuntamiento o la Comunidad de Madrid llevan un seguimiento del mismo. También se pueden encontrar diversas aplicaciones que te informan sobre el estado del mismo e incluso te dan explicaciones más detalladas.

El uso de los mapas por puntos de localización o coloración de áreas en base al estado, es algo que puede llegar a encontrarse fácilmente. Existe alguna herramienta que facilita la predicción de la contaminación del aire, directrices con respecto a la matrícula del coche. Son muchas opciones que te permiten saber en muchos de los casos el estado en tiempo real, el histórico, datos de cada magnitud de medida, etc.

Está de actualidad el cambio climático, estos últimos días está siendo noticia diversas publicaciones de científicos, movilizaciones humanas y hay países, organizaciones e incluso empresas privadas que están direccionando sus estrategias a paliar de algún modo esto. Tecnológicamente se están invirtiendo en I+D+I en prototipos de coches eléctricos, en la obtención y manejo de los datos para poder realizar las decisiones más adecuadas a través del Big Data y Data Science.

Madrid está tratando de virar hacia una estrategia que se adapte mejor a la nueva situación del clima, tratando de tomar las mejores decisiones tanto en la gestión de la comunidad o ciudad

de Madrid por parte pública como empresas privadas que ya ofrecen transportes alternativos no contaminantes.

Debido a esto, es un gran momento para poder ofrecer un despliegue exhaustivo con las herramientas necesarias para poder hacer las tomas de decisiones más oportunas, teniendo en cuenta múltiples factores. Este proyecto viene a abrir un primer paso para ese camino, tratando de servir información de valor teniendo en cuenta los múltiples factores como puede ser: la calidad del aire; el tráfico, la situación de los puntos que pueden afectar a la calidad del aire y también importante saber cuales no están afectados por esto; o como la meteorología incide en la calidad del aire directa o indirectamente.

Desarrollo

Objetivo

El objetivo de negocio como ya se ha planteado, es dar una solución a la contaminación del aire de Madrid en relación al tráfico y meteorología. Esto llega a generar varios casos de uso de negocio que podrían ser explotados y para ello ha de plantearse un objetivo para la arquitectura que facilite el análisis y predicción.

Esta arquitectura inicialmente va a optar por un producto mínimo viable para que salga en producción lo antes posible y luego posteriormente ir iterando y mejorando. Por lo que inicialmente se trata de resolver una arquitectura que pueda llegar a ingestar datos de diferentes fuentes, procesarlos y disponibilizarlos con una periodicidad variable en la ingesta y horaria en el procesamiento. Por otro lado, este proyecto trata de resolver un modelo de visualización, inicialmente con actualización diaria y en futuros pasos horaria, donde la visualización de nuevos datos sea automática sin que un usuario deba cargar manualmente ficheros, donde puedan verse datos en diferentes periodos de tiempo y que visualmente sea muy sencillo y comprensible.

Este desarrollo incluye una solución para poder tener previsiones de los contaminantes y que luego el negocio interesado en el producto pueda tomar decisiones teniendo en cuenta datos de valor. Para ello inicialmente se va a tratar un modelado más sencillo pero más rápido de implementar para poder estar en producción en los próximos pasos más cercanos. Posteriormente se plantean modelos más complejos y efectivos en sus predicciones como pueden ser series temporales o redes neuronales.

Una vez conocidos los pormenores de los objetivos a cumplir para el desarrollo del proyecto, se procede a presentar las diferentes fases del mismo.

Tipología y estructura de las fuentes de datos

Para este estudio las fuentes de los datos van a ser diversas. Por un lado desde el portal de Open Data de la ciudad de Madrid, se obtendrán los datos de contaminación del aire y los del tráfico.

Datos de tráfico

Estos datos permiten conocer el estado del tráfico en Madrid y se obtienen a través de un xml que se disponibiliza mediante una url. Son unos datos con una periodicidad de actualización de 5 minutos y ofrecen la siguiente información:

- **idelem**: Identificador del punto de medida.
- **descripción**: Denominación del punto de medida.
- **accesoAsociado**: Código de control relacionado con el control semafórico para la modificación de los tiempos.
- **intensidad**: Intensidad de número de vehículos por hora.
- **ocupación**: Porcentaje de ocupación del punto de control por los vehículos.
- **carga**: Parámetro de carga del vial en función de la intensidad, ocupación y características de la infraestructura.
- **nivelServicio**: Nivel de servicio
- **intensidadSat**: Intensidad de saturación de la vía en vehiculo/hora y que se corresponde con el máximo número de vehículos que pueden pasar en el acceso a la intersección manteniéndose la fase verde del semáforo.
- **error**: Código de control de la validez de los datos del punto de medida.
- **subarea**: Identificador de la subárea de explotación de tráfico a la que pertenece el punto de medida.
- **st_x**: Coordenada X UTM del centroide que representa al punto de medida en el fichero georreferenciado
- **st_y**: Coordenada Y UTM del centroide que representa al punto de
- medida en el fichero georreferenciado.

Datos de contaminación del aire

El Sistema Integral de la Calidad del Aire del Ayuntamiento de Madrid permite conocer en cada momento los niveles de contaminación atmosférica en el municipio y se obtiene en csv a través de una url que se disponibiliza. Son unos datos con una periodicidad de actualización de una hora. Ofrecen la siguiente información:

- **provincia**: Código de la provincia.
- **municipio**: Código del municipio.
- **estación**: Código de la estación.
- **magnitud**: Código de la magnitud que se está midiendo.
- **punto_muestreo**: Incluye el código de la estación completo (provincia, municipio y estación) más la magnitud y la técnica de muestreo
- **ano**: El año en el que se está haciendo la medida.
- **mes**: El mes en el que se está haciendo la medida.
- **dia**: El día en el que se está haciendo la medida
- **H00**: El valor del dato de la magnitud por cada hora.
- **V00**: El código de validación.

Por otro lado se necesitarán saber las observaciones y predicciones meteorologicas y ver cómo estas influyen. Para ello se va a hacer uso de los datos que pone a disposición AEMET a través de su API.

Datos de observación

Datos de observación proporcionados por el servicio de Open Data de Aemet, en horarios de las últimas 24 horas (actualización cada hora). Estos datos se disponibiliza mediante api con api key en formato json.

- **alt**: Altitud de la estación en metros
- **dmax**: Dirección del viento máximo registrado en los 60 minutos anteriores a la hora indicada por 'fint' (grados).
- **dv**: Dirección media del viento, en el período de 10 minutos anteriores a la fecha indicada por 'fint' (grados).
- **fint**: Fecha hora final del período de observación, se trata de datos del periodo de la hora anterior a la indicada por este campo (hora UTC).
- **hr**: Humedad relativa instantánea del aire correspondiente a la fecha dada por 'fint' (%)
- **idema**: Indicativo climatológico de la estación meteorología automática.
- **lat**: Latitud de la estación meteorológica (grados).
- **lon**: Longitud de la estación meteorológica (grados).
- **prec**: Precipitación acumulada, medida por el pluviómetro, durante los 60 minutos anteriores a la hora indicada por el período de observación 'fint' (mm, equivalente a l/m2)
- **pres**: Presión instantánea al nivel en el que se encuentra instalado el barómetro y correspondiente a la fecha dada por 'fint' (hPa).
- **rviento**: Recorrido del viento durante los 60 minutos anteriores a la fecha indicada por 'fint' (Hm)
- **ta**: Temperatura instantánea del aire correspondiente a la fecha dada por 'fint' (grados Celsius)

- **tamax**: Temperatura máxima del aire, valor máximo de los 60 valores instantáneos de 'ta' medidos en el período de 60 minutos anteriores a la hora indicada por el período de observación 'fint' (grados Celsius).
- **tamin**: Temperatura mínima del aire, valor mínimo de los 60 valores instantáneos de 'ta' medidos en el período de 60 minutos anteriores a la hora indicada por el período de observación 'fint' (grados Celsius).
- **ubi**: Ubicación de la estación. Nombre de la estación
- **vmax**: Velocidad máxima del viento, valor máximo del viento mantenido 3 segundos y registrado en los 60 minutos anteriores a la hora indicada por el período de observación 'fint' (m/s).
- **vv**: Velocidad media del viento, media escalar de las muestras adquiridas cada 0,25 ó 1 segundo en el período de 10 minutos anterior al indicado por 'fint' (m/s).

Datos de predicción

Datos de predicción proporcionados por el servicio de Open Data de Aemet, de hora en hora hasta 48 horas próximas. Estos datos se disponibiliza mediante api con api key en formato json/xml.

- **estadoCielo**: Descripción del estado del cielo.
- **fecha**: Período de validez de la Predicción.
- **humedadRelativa**: Valor de la humedad relativa.
- **nieve**: Cantidad total de nieve que se prevé que caiga durante la hora anterior.
- **ocaso**: Hora del atardecer.
- **orto**: Hora del amanecer.
- **precipitacion**: Cantidad total de precipitación durante la hora anterior.
- **probNieve**: Valor de la probabilidad de precipitación de nieve.
- **probPrecipitacion**: Valor de la probabilidad de precipitación.
- **probTormenta**: Valor de la probabilidad de tormenta.
- **sensTermica**: Valor de la sensación térmica.
- **temperatura**: Valor de la temperatura.
- **viento.direccion**: Dirección del viento.
- **viento.velocidad**: Velocidad del viento.
- **rachaMax**: Valor de la Racha máxima.

Estas cuatro fuentes de datos estan pensadas para cubrir todas las necesidades que requiere el proyecto. Por un lado con los datos de la contaminación del aire y del tráfico se usan para poder monitorizar la situación actual y hacer un seguimiento histórico a estos dos fenómenos. Por otro estos mismos datos sumados a los datos meteorológicos son para la parte de darle inteligencia al proyecto y poder hacer predicciones basadas en conocimiento histórico. Por ello se va a usar los datos del aire, tráfico y observación meteorológica para entrenar los modelos

de predicción y se utilizarán los valores de predicción meteorológica como entrada del modelo para que genere dichas predicciones.

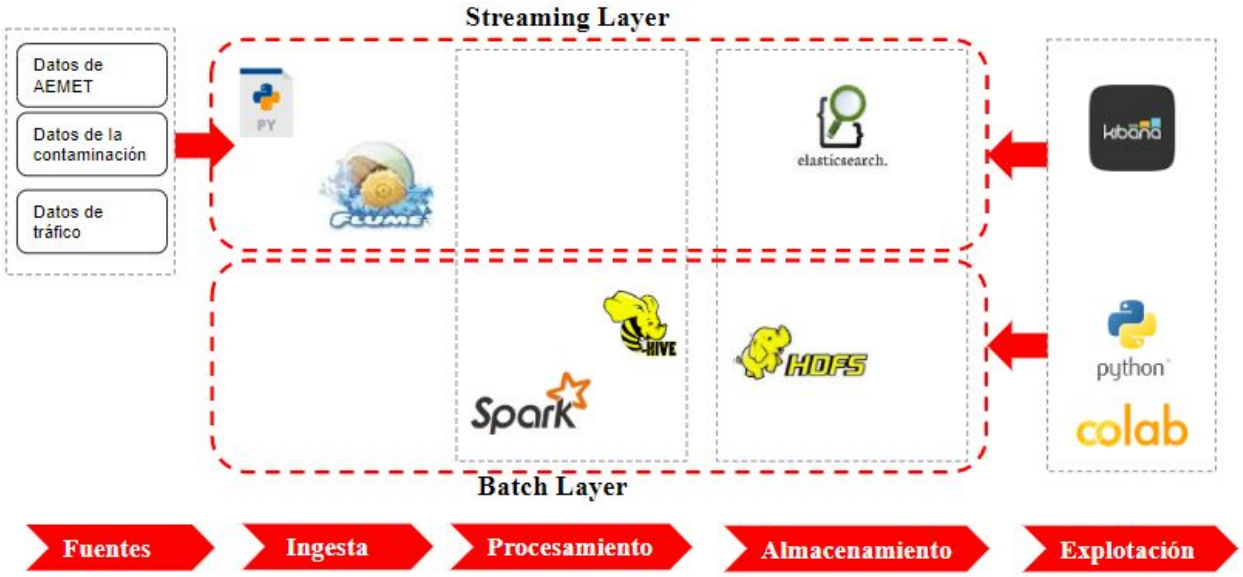
En futuros pasos también se valorará visualizar los datos de observación meteorológica para poder hacer la monitorización y seguimiento en conjunto con el tráfico y la contaminación del aire.

Arquitectura

Esquema de la arquitectura

La arquitectura siguiente a plantear ha de tratar de cubrir las necesidades del caso de uso planteado, “Análisis y predicción de la contaminación y el tráfico en Madrid”.

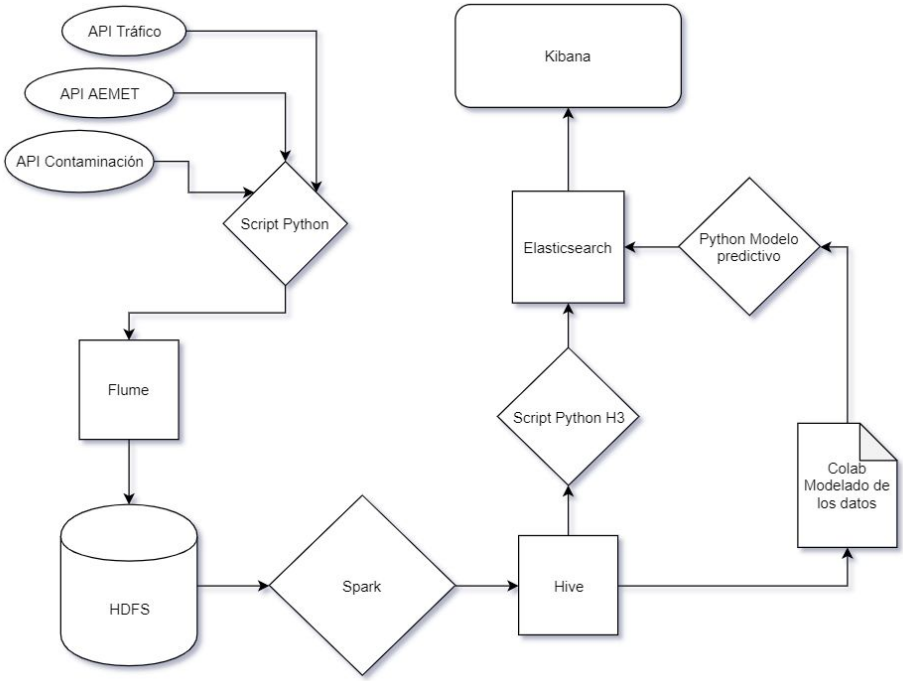
Para ello se ha de tener en cuenta que hay tres orígenes de información, datos meteorológicos, datos del aire de Madrid y datos del tráfico de Madrid. Estos tienen una periodicidad de 1h los dos primeros y de 5 minutos el último. Una vez almacenados han de ser procesados y dispuestos para su consumo tanto para modelar como para visualización. Por lo tanto, teniendo en cuenta estas premisas la arquitectura que se plantea es la siguiente:



Las herramientas que van a formar parte de esta arquitectura están pensadas para cumplir una función en el flujo de datos planteado para el caso.

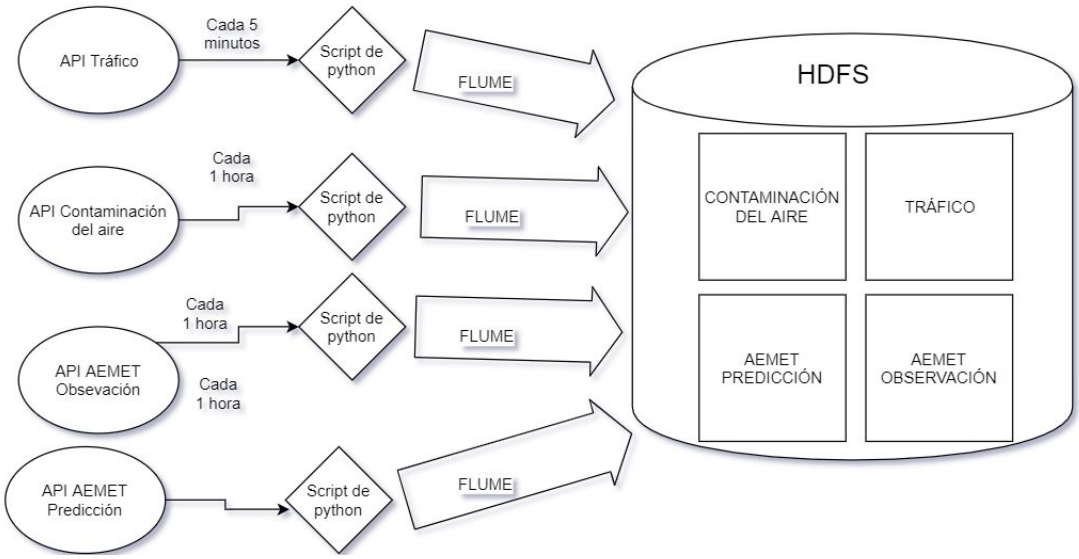
Flujo de los datos

Una vez visto cual es el cometido de cada herramienta dentro del flujo de datos la presentación visual del mismo será la siguiente:



Python scripts de ingesta

Este pequeño proceso es el encargado de obtener en cada periodo de tiempo (1h o 5 min) los datos y depositarlos en el nodo frontera para que sean consumidos por Flume. Inicialmente se ha planteado hacer la llamada mediante este script porque es un volumen de datos muy pequeño sin vistas a que vaya aumentar demasiado y es un modo más sencillo cubrir la necesidad en un primer paso. En próximos pasos se buscará un modo más eficiente de llevar a cabo el proceso.



Flume

Esta herramienta está pendiente del nodo frontera para trasladar toda información nueva que se deposite debidamente ordenada y clasificada en HDFS. De este modo también se logra que nadie ajeno al cluster acceda directamente dentro.

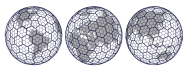
HDFS

Este tipo de almacenamiento utiliza para persistir toda la información en bruto que se vaya recopilando, por un lado para ser posteriormente procesada y por otro lado a modo de backup para futuros casos de uso o necesidades de este mismo. Además se habilita una carpeta donde se deposita el último dato ingestado para poder posteriormente procesarlo con Spark.

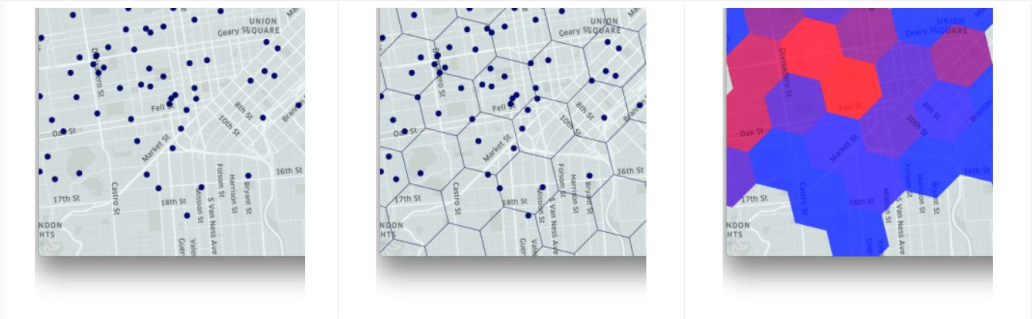
Spark

Es la herramienta que es para procesar los datos ingestados con en lenguaje de programación scala, nativo de la herramienta. Estos datos se obtienen de la carpeta “datosSpark” donde Flume deja los datos divididos en carpetas por cada origen.

H3-Uber



Es una librería que aplica un sistema de cuadrícula hexagonal a cada punto del mapa, pudiendo regular el diámetro de la misma. Es muy útil para tratar grandes cantidades de datos geoespaciales y poder no solo visualizarlos de una forma más cómoda si no poder tratarlos y procesarlos mejor. Es de código abierto y está disponible en GitHub y en la web con toda la documentación para diferentes lenguajes de programación.



Python H3

Por incompatibilidades con la librería de Uber, H3, se ha decido, para un producto mínimo viable, aplicar esta librería con python que sí tiene compatibilidad y enviar los datos a Elasticsearch instalado en local.

El proceso de agregación de esta librería es muy importante ya que se le aplica una parte importante de la lógica del caso de uso.

Esta es una cuadrícula hexagonal se le aplicada al mapa de Madrid donde por cada cierta área le corresponde un hexágono. A cada hexágono se le asignan los valores que le correspondan por cada tipología de datos. De este modo, se logra tener una visión más general de los mismos en un mapa, pudiendo trasladarlos a un modelo de datos que además de visualizarlos puedan ser tratados para analítica. Esta parte del agregado se trata con una librería open ya mencionada de Uber, H3, para varios lenguajes de programación entre los cuales no esta Scala pero sí Python. Por lo que este proceso el cual inicialmente estaba planteado para hacerlo en Spark, se realiza con un script de python. Una vez aplicada esta librería, estos datos se envían a Elasticsearch. Este proceso inicialmente se realiza por batch cada hora.

Hive

Bajo esta herramienta es el modo en el que se disponibiliza los datos limpios y agregados para que poder tratarlos en el modelado, en su posterior tratamiento con H3 y visualización con Kibana u otros intereses analíticos. Se disponen debidamente ordenados por tablas con cada tipología de datos y ordenados por fecha y hora.

Elasticsearch

Este motor de búsqueda es donde se van a almacenar los datos ya agregados, limpios y con la librería de H3 aplicada, para poder visualizarlos con Kibana de manera dinámica dentro del flujo de datos. En este sentido serán los datos de tráfico y calidad del aire los que se incluyen y visualizan con Kibana. Además se enriquecerán con los datos de predicción de los modelos aplicados para tener tanto la visión real como de lo predicho.

Kibana

Es la herramienta de visualización del ecosistema ELK de Elastic el cual conecta perfectamente con el motor de búsqueda Elasticsearch. Es la misma que se va a utilizar para la visualización dentro del flujo de datos para poder ver tanto la evolución de los mismos como en futuros pasos las predicciones por capas en la herramienta que tiene de mapa. De este modo se logra conectar directamente en flujo de los datos con la visualización pudiendo ser explotados con la herramienta.

Colab

Es la herramienta de notebook que disponibiliza en cloud Google y sobre la que se va a trabaja el estudio y actualización del modelado de los datos. Una vez se tenga en modelo listo, se genera el debido codigo para ponerlo en producción e insertarlo en Elasticsearch.

Ingesta y procesamiento

Esta primera parte de adquisición de los datos y ponerlos a disposición es muy importante ya que de ellos va a tratar el proyecto y de que los posteriores procesos funcionen.

En primer lugar esta la parte ingesta de los datos donde toman parte varios procesos o aplicaciones. Por un lado, hay 4 procesos de python corriendo en el nodo frontera del cluster que periódicamente llaman a la API correspondiente, convierten los datos en csv y los depositan en carpetas del nodo frontera, una por tipo de dato.
Hay cuatro flumes configurados, uno por tipo de dato, para que observen las carpetas mencionadas y envíen los datos por un lado a historificar en HDFS y por otro a una carpeta donde Spark lee para procesarlos posteriormente.

En segundo lugar, una vez finalizada la ingesta comienza el procesamiento. Esta parte esta dividida en varias fases debido a que hay incompatibilidad de librerías en Spark con H3 para centralizar todo el proceso en uno.

En Spark se siguen los siguientes procesados con los datos:

- A los datos de calidad del aire se eliminan los nulos y las columnas que indican esto. Además se agregan los datos de ubicación de las estaciones y los nombres de las magnitudes de medida. Se le aplica un formato concreto al campo fecha.

- En los datos de tráfico, se eliminan los datos nulos y, se agregan los datos de ubicación de las estaciones. Además se aplica un formato concreto al campo fecha.

- En los datos de predicción meteorológica se completan algunos columnas concretas que vienen medio vacías unificando los datos y se eliminan algunas variables que no aportan.

- Una vez procesados los datos se envían a carpetas donde leerá Hive para cada una de las tablas de los datos origen y se elimina la carpeta “datosSpark” para prepararla para los nuevos datos que se ingestarán y haya que procesar.

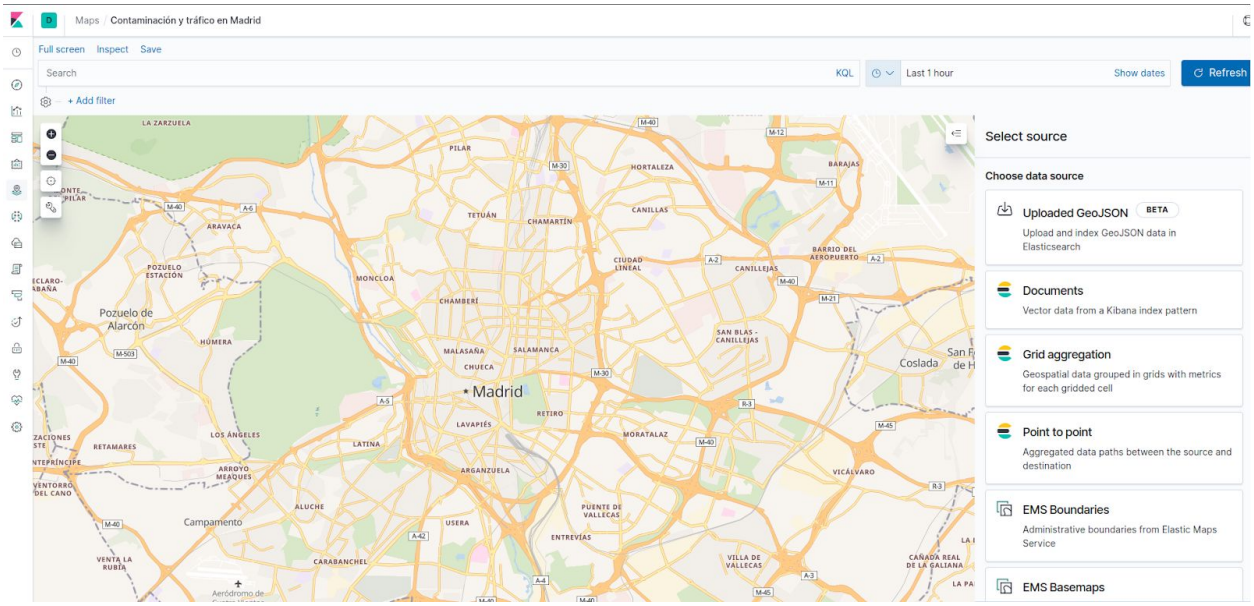
En Hive para cada tipología de dato se ha configurado previamente una tabla que se almacena en carpetas ordenadas por fecha.

En la segunda parte del procesamiento es donde con el uso de la librería PyHive en dos procesos Python (uno para el aire y otro para el tráfico) se leen los datos del día anterior ya completos, se les añade la cuadrícula hexagonal de H3 correspondiente a cada localización y se ingesta en Elasticsearch para su visualización en Kibana.

Esta parte final, inicialmente se ha preparado para que sea una ingesta diaria nocturna pero en próximos pasos está pensado llevarla a horaria. En Elasticsearch se han configurado dos índices, uno para el tráfico y otro para el aire, pero al igual que la ingesta horaria en próximos pasos la idea es incluir un índice más para la predicción de los datos de contaminación que proporcionará el modelo preentrenado y preconfigurado que se ha elaborado.

Visualización

La visualización e interpretación de los datos se realizará como anteriormente se ha mencionado con la herramienta Kibana. Está obtendrá toda la información directamente de Elasticsearch ya que tienen una conexión nativa automática y si fuera necesario se podría obtener los datos en tiempo real una vez estuvieran en el motor de búsqueda.



El manejo de la misma es muy sencillo, se ha dividido en diferentes capas, cada capa es un tipo de dato (tráfico y contaminación del aire actual). Estas capas pueden activarse y desactivarse muy fácilmente e incluso aplicar filtros que se vean necesarios, por ejemplo para cambiar de magnitud de medida de contaminación para focalizar el estudio.

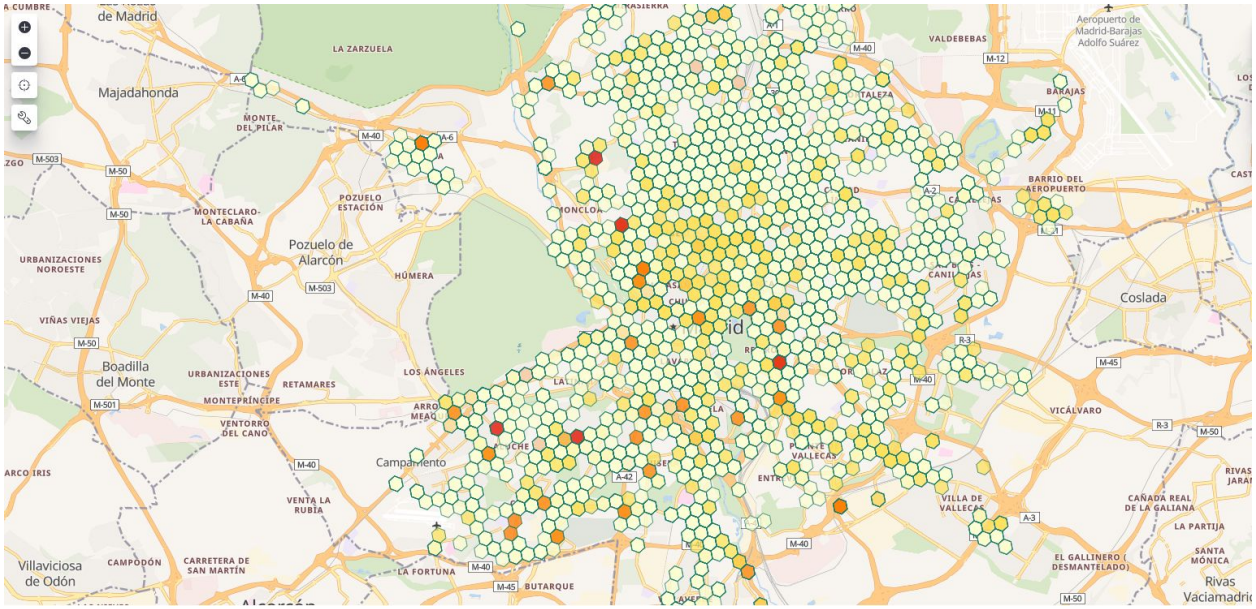
La capa de contaminación predicha, se incluirá en el siguiente paso del proyecto cuando se productivice el modelo de Random Forest seleccionado. Y funcionará igual que la capa de contaminación del aire actual.

Capa de tráfico

Esta es la capa más sencilla de entender. Con solo filtrar por fecha (arriba a la derecha del cuadro de mando) y hora tanto absoluta como relativa, se visualizarán los datos en el mapa. El mismo se divide en hexágonos localizados por zona geográfica y cada uno de ellos indica el nivel de congestión según su color de la siguiente manera:

- Tráfico fluido: color amarillo claro
- Tráfico lento: color amarillo oscuro

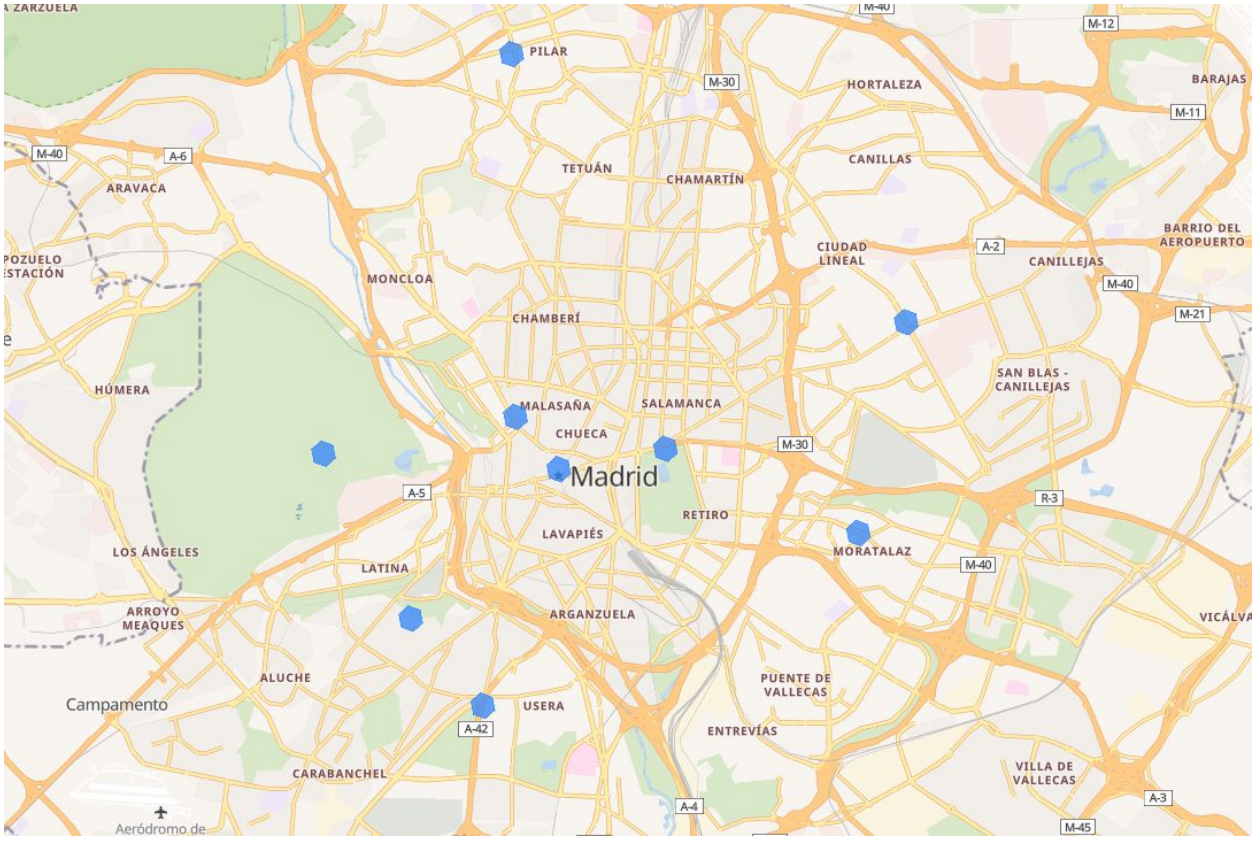
- Tráfico con retenciones: color naranja
- Tráfico congestionado: color rojo



Capa de contaminación del aire actual

En esta capa se han hecho dos divisiones. Por un lado una capa de contaminación del aire genérica donde se pueden ver mediante filtros cada uno de los contaminantes. Y por otro se ha hecho una capa por cada uno de los contaminantes más relevantes seleccionando bien los umbrales límite de contaminación. Como ejemplo está el contaminante de Monóxido de Carbono (CO) donde se pintan las ubicaciones de las estaciones que miden este contaminante en el mapa, coloreado en base a las diferentes franjas o umbrales. En la imagen inferior se ven todas en azul que es el umbral más bajo de este contaminante. Estas franjas serían:

- Muy bajo: Azul
- Bajo: Verde
- Medio: Amarillo
- Moderado: Naranja
- Alto: Rojo
- Muy alto: Negro


PILAR
M-30
HORTALEZA
BARAJAS
M-12
M-11
M-40
A-6
ARAVACA
TETUÁN
CHAMARTÍN
CANILLAS
BARRIO DEL AEROPUERTO
CIUDAD LINEAL
A-2
CANILLEJAS
M-40
M-21
MONCLOA
CHAMBERÍ
HÚMERA
SAN BLAS - CANILLEJAS
MALASAÑA
SALAMANCA
CHUECA
M-30
Madrid
A-5
R-3
RETIRO
LAVAPIÉS
MORATALAZ
LOS ÁNGELES
LATINA
M-40
ARROYO MEAQUES
ARGANZUELA
ALUCHE
PUENTE DE VALLECAS
Campamento
USERA
A-42
ENTREVÍAS
CARABANCHEL
VILLA DE VALLECAS
A-4
A-3
M-45
Aeródromo de

Modelo predictivo

Se deja disponible el notebook a través de este enlace con el detalle de todo el trabajo realizado en el modelado y análisis exploratorio.

Para el planteamiento de predecir los valores de contaminación se ha planteado en varias fases. En una primera fase la idea es tener un modelo predictivo que se aproxime lo máximo posible pero que no consuma mucho tiempo a la hora de plantearlo para tener el producto en producción lo antes posible. Este modelo es una regresión lineal. En una segunda fase se tratará de mejorar la predicción teniendo en cuenta la tipología de los datos a predecir, con series temporales multivariantes o redes neuronales como LSTM.

Datos

Como datos de entrada para el análisis, selección del modelo y entrenamiento del mismo, se han cogido los históricos de septiembre de 2019 de los datos de observación meteorológica, de tráfico y de contaminación del aire. Estos tres datos de entrada se conforman en un dataset con las siguientes variables:

Datos de contaminación & trafico & meteo Septiembre 2019	**Descripción**
PROVINCIA	Código de la provincia.
MUNICIPIO	Código del municipio.
ESTACIÓN	Código de la estación
PUNTO_MUESTREO	Incluye el código de la estación completo (provincia, municipio y estación) más la magnitud y la técnica de muestreo
NOMBRE_ESTACION	Nombre de la estación que ha hecho la medición
LONGITUD	Coordenada que representa la ubicación de la estación de medida del aire en el eje de la longitud
LATITUD	Coordenada que representa la ubicación de la estación de medida del aire en el eje de la latitud
MAGNITUD	Nombre de la magnitud que se está midiendo
FECHA	Fecha y hora del momento de la medición
LOCALIZACION	Punto de la ubicación de la estación de medida del aire

HORA	Hora del momento de la medición
VALOR	El valor del dato de la magnitud por cada hora.
hex_id	Id de el hexágono de la rejilla de H3, correspondiente al punto de ubicación de la estación de medida del aire
GEOM	Poligono de puntos de ubicación geográfica que delimitan y conforman el hexágono correspondiente a la ubicación de la estación de medida del aire
trafico_intensidad	Intensidad de número de vehículos por hora
trafico_ocupación	Porcentaje de ocupación del punto de control por los vehículos
trafico_carga	Parámetro de carga del vial en función de la intensidad, ocupación y características de la infraestructura
dmax	Dirección del viento máximo registrado en los 60 minutos anteriores a la hora indicada por 'fint' (grados).
dv	Dirección media del viento, en el período de 10 minutos anteriores a la fecha indicada por 'fint' (grados).
hr	Humedad relativa instantánea del aire correspondiente a la fecha dada por 'fint' (%)
prec	Precipitación acumulada, medida por el pluviómetro, durante los 60 minutos anteriores a la hora indicada por el período de observación 'fint' (mm, equivalente a l/m2)
pres	Presión instantánea al nivel en el que se encuentra instalado el barómetro y correspondiente a la fecha dada por 'fint' (hPa).
rviento	Recorrido del viento durante los 60 minutos anteriores a la fecha indicada por 'fint' (Hm)
ta	Temperatura instantánea del aire correspondiente a la fecha dada por 'fint' (grados Celsius)
tamax	Temperatura máxima del aire, valor máximo de los 60 valores instantáneos de 'ta' medidos en el período de 60 minutos anteriores a la hora indicada por el período de observación 'fint' (grados Celsius).

tamin	Temperatura mínima del aire, valor mínimo de los 60 valores instantáneos de 'ta' medidos en el período de 60 minutos anteriores a la hora indicada por el período de observación 'fint' (grados Celsius).
vmax	Velocidad máxima del viento, valor máximo del viento mantenido 3 segundos y registrado en los 60 minutos anteriores a la hora indicada por el período de observación 'fint' (m/s).
vv	Velocidad media del viento, media escalar de las muestras adquiridas cada 0,25 ó 1 segundo en el período de 10 minutos anterior al indicado por 'fint' (m/s).
inso	Duración de la insolación durante los 60 minutos anteriores a la hora indicada por el período de observación 'fint' (horas)
stdvv	Desviación estándar de las muestras adquiridas de velocidad del viento durante los 10 minutos anteriores a la fecha dada por 'fint' (m/s)
stddv	Desviación estándar de las muestras adquiridas de la dirección del viento durante los 10 minutos anteriores a la fecha dada por 'fint' (grados)
vis	Visibilidad, promedio de la medida de la visibilidad correspondiente a los 10 minutos anteriores a la fecha dada por 'fint' (Km)
ts	Temperatura suelo, temperatura instantánea junto al suelo y correspondiente a los 10 minutos anteriores a la fecha dada por 'fint' (grados Celsius)
pres_nmar	Valor de la presión reducido al nivel del mar para aquellas estaciones cuya altitud es igual o menor a 750 metros y correspondiente a la fecha indicada por 'fint' (hPa)
tpr	Temperatura del punto de rocío calculado correspondiente a la fecha 'fint' (grados Celsius)

Este dataset conformado por estas variables es el resumen por horas cada valor de cada contaminante con la meteorología y tráfico asociado a la ubicación correspondiente según la cuadrícula hexagonal de H3.

Objetivo

Como objetivo es la predicción de los valores de contaminación teniendo como entrada unos valores de tráfico y meteorología. Como dichos valores son diferentes según el contaminante, se decide separar tanto el análisis como la predicción por contaminante. Estos contaminantes son:

Contaminante Dióxido de Azufre:
Es un gas incoloro con un característico olor irritante. El dióxido de azufre es el principal causante de la lluvia ácida ya que en la atmósfera es transformado en ácido sulfúrico. Es liberado en muchos procesos de combustión ya que los combustibles como el carbón, el petróleo, el diésel o el gas natural contienen ciertas cantidades de compuestos azufrados.

Contaminante Monóxido de Carbono:
Es un gas incoloro y altamente tóxico. Puede causar la muerte cuando se respira en niveles elevados. Se produce por la combustión deficiente de sustancias como gas, gasolina, queroseno, carbón, petróleo, tabaco o madera.

Contaminante Monóxido de Nitrógeno:
Es un gas sin color, y difícilmente soluble en agua, que constituye uno de los contaminantes de la atmósfera que forma parte de la lluvia ácida y en muy pequeñas cantidades, también lo se puede encontrar en los mamíferos, a pesar de ser considerado como un agente tóxico.

Contaminante Dióxido de Nitrógeno:
Es un compuesto químico formado por los elementos nitrógeno y oxígeno. Se forma como subproducto en los procesos de combustión a altas temperaturas, como en los vehículos motorizados y las plantas eléctricas.

Contaminante Óxidos de Nitrógeno:
Se aplica a varios compuestos químicos binarios gaseosos formados por la combinación de oxígeno y nitrógeno. Los óxidos de nitrógeno son liberados al aire desde los tubos de escape de vehículos motorizados (sobre todo diésel y de mezcla pobre), de la combustión del carbón, petróleo o gas natural, y durante procesos tales como la soldadura por arco, galvanoplastia, grabado de metales y detonación de dinamita.

Contaminante Partículas < 2.5 µm:
Partículas en suspensión de menos de 2,5 micras. Un mejor indicador de la contaminación urbana. En buena medida provienen de las emisiones de los vehículos diesel en la ciudad.

Contaminante Partículas < 10 µm:
Pequeñas partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera, y cuyo diámetro aerodinámico es menor que 10 µm.

Contaminante Ozono:
Es una sustancia cuya molécula está compuesta por tres átomos de oxígeno. Es un gas que desprende olores fuertes. Tiene uso industrial como precursor en la síntesis de algunos compuestos orgánicos, pero principalmente como desinfectante depurador y purificador de aguas minerales.

Contaminante Tolueno:
Se trata de un hidrocarburo de tipo aromático de olor agradable que se produce a partir del benceno. Esta sustancia puede hallarse en la naturaleza en árboles del género Myroxylon y en el petróleo crudo. Además, puede obtenerse mediante distintos procesos industriales.

Contaminante Benceno:
Es un hidrocarburo aromático de fórmula molecular. Se encuentra en la lista de los 20 productos químicos de mayor volumen de producción. El benceno se utiliza como constituyente de combustibles para motores, disolventes de grasas, aceites y pinturas; en el grabado fotográfico de impresiones.

Contaminante Etilbenceno:
Es un líquido inflamable, incoloro, de olor similar a la gasolina. El uso principal del etilbenceno es para fabricar otro producto químico, estireno y por la polimerización de este, se obtiene el poliestireno.

Contaminante Hidrocarburos totales:
Los términos hidrocarburos totales de petróleo (abreviados TPH en inglés) se usan para describir una gran familia de varios cientos de compuestos químicos originados de petróleo crudo.

Contaminante Metano:
Es el hidrocarburo alcano más sencillo. El 60 % de las emisiones en todo el mundo es de origen antropogénico. Proceden principalmente de actividades agrícolas y otras actividades humanas.

Contaminante Hidrocarburos no metánicos:
Aparecen otros grupos de hidrocarburos en atmósferas urbanas.(Etano, Hexano, Benceno, Tolueno, etc.).

Primera fase del modelado

De inicio se ha tratado de entender los datos. En una primera vista, se entiende que puede aportar valor tener como variables el dia de la semana (de 0 a 6), si es día de labor y agregar

los datos de las horas en franjas horarias. Por lo que en el tratamiento previo de los datos se han agregado esas dos columnas. Para el análisis y posterior modelado se procede a dejar el dataset con las columnas numéricas que pueden aportar valor en la predicción sin repetir la información, además del nombre del contaminante. Así se lleva a cabo un análisis exploratorio de las mismas con el fin de entender más el dataset. Este solo tiene nulos en la variable *tpr*, se mantiene en el análisis exploratorio pero no se ve decisiva y recalcular los nulos podría alterar el dataset por lo que se elimina.

En este análisis exploratorio, se pueden entender diversos puntos. Por un lado que se encuentra una clara estacionalidad en el tráfico sobre todo en las franjas de la tarde y la noche. Aunque sorprendentemente aquellos contaminantes que deberían verse influenciados por el tráfico teóricamente, no se ha visto así. Se entiende que habría que seguir observando como va evolucionando el mismo en próximos meses para tener una idea más firme.

En cambio si se ha visto que la meteorología puede llegar a influir en los contaminantes, como por ejemplo la falta de precipitación. Teóricamente esto es cierto, pero en este punto también sería oportuno seguir analizando cómo evoluciona. Ya que se intuye que durante el periodo en el que se ubica el dataset no hubo mucha precipitación en carácter general.

Una vez se conocen los datos, y como se ha mencionado la tipología del modelo a seleccionar es una regresión lineal, ya que debido a que se quiere predecir un valor exacto en base a unas variables de entrada, de inicio como producto mínimo viable es lo indicado. Las variables de entrada que se escogen son aquellas que se han visto que no aportan información duplicada y pueden llegar a ser relevantes en la predicción:

```
features = ['HORA', 'DIASEMANA','trafico_intensidad', 'trafico_ocupacion',
'trafico_carga', 'dmax', 'dv', 'hr', 'inso', 'prec', 'pres', 'rviento',
'ta', 'vis', 'vmax']
```

Y como variable objetivo `'VALOR'`, que indica el valor del contaminante. Además aquí se incluye la fecha con objetivo de luego plantear un gráfico en eje temporal para comparar la predicción con el valor real.

Al tener que utilizar modelos supervisados de regresión lineal se debe dividir además de en variables de entrada y objetivo, en dataset de entrenamiento y de test. Esto es debido a que esta tipología de modelos deben previamente entrenarse y aprender de los datos para poder predecir. Y se reserva una parte del dataset llamada test para evaluar el aprendizaje. Por todo ello se divide el dataset entre entrenamiento y test además de entre variables de entrada y objetivo.

Como ya se ha mencionado uno de los métodos para evaluar los modelos es el gráfico donde salen las dos líneas de tiempo, real y predicha, de los valores de contaminación. Aquí se

evalúa visualmente lo cerca y similar que se comporta la predicción en comparación al real. Además como métrica característica para evaluar lo óptimo que son este tipo de modelos se usa el error cuadrático medio. Este mide el promedio de los errores al cuadrado, es decir, la diferencia entre el valor predicho y el real. La diferencia se produce debido a la aleatoriedad o porque el predictor no tiene en cuenta la información que podría producir una estimación más precisa. En los modelos de regresión se usa como la media de las desviaciones al cuadrado, de las predicciones de los verdaderos valores. Por lo que la selección del modelo indicado se realiza teniendo en cuenta estas dos formas de evaluar.

Inicialmente se hará la prueba con un contaminante y luego se aplicará el modelo al resto ya que disponen de las mismas variables y el comportamiento a la hora de predecir será exportable al resto.

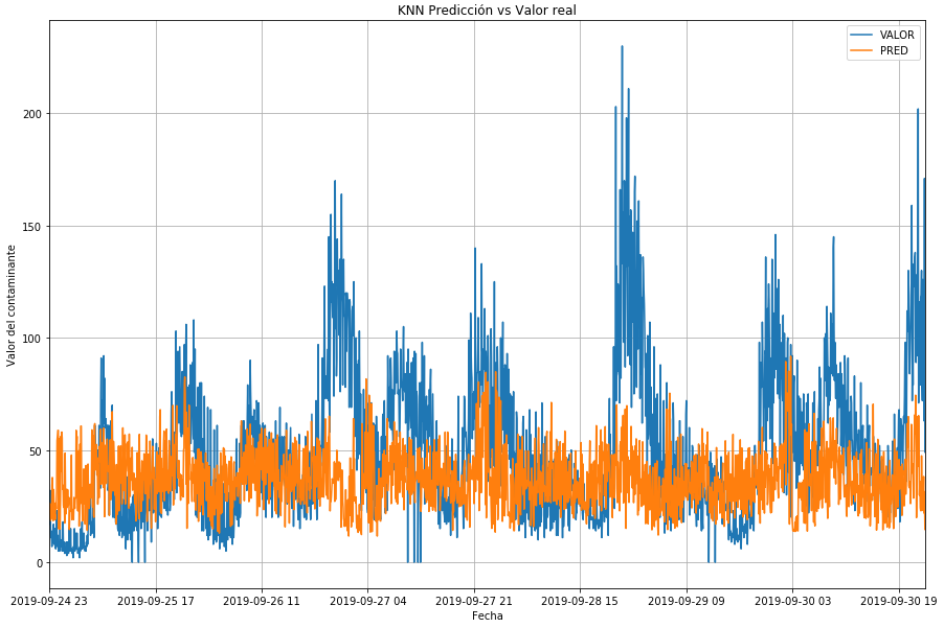
Los modelos que se han seleccionado para el objetivo que se plantea son:

KNeighbors

El modelo de los k vecinos más cercanos utiliza un método de entrenamiento mediante ejemplos cercanos en el espacio de estos.

Tras entrenar el modelo y evaluarlo, el resultado es el siguiente:

```
Mean squared error: 1235.04
```

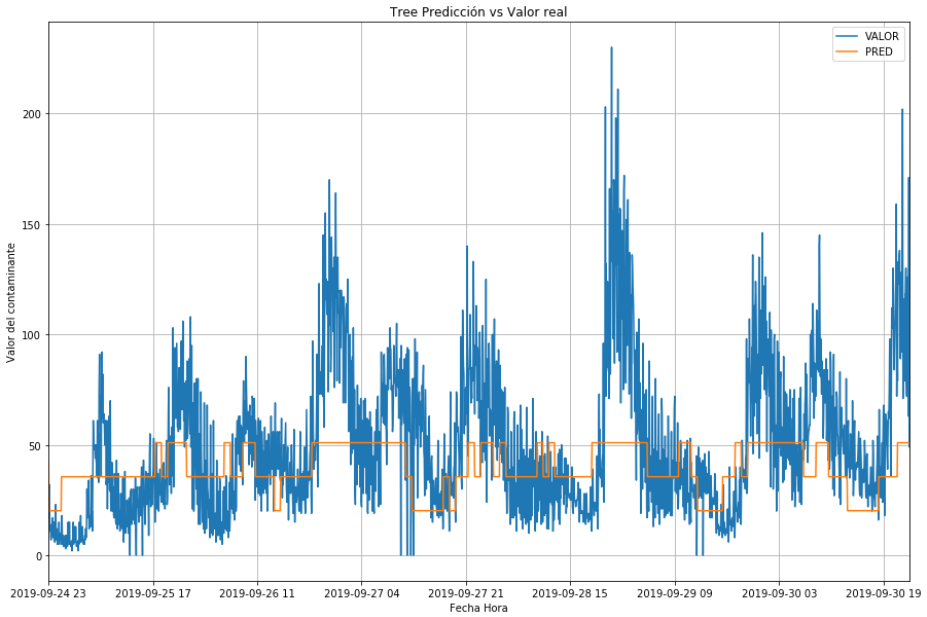


Decision Tree

Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva.

Tras entrenar el modelo y evaluarlo, el resultado es el siguiente:

```
Mean squared error: 998.60
```

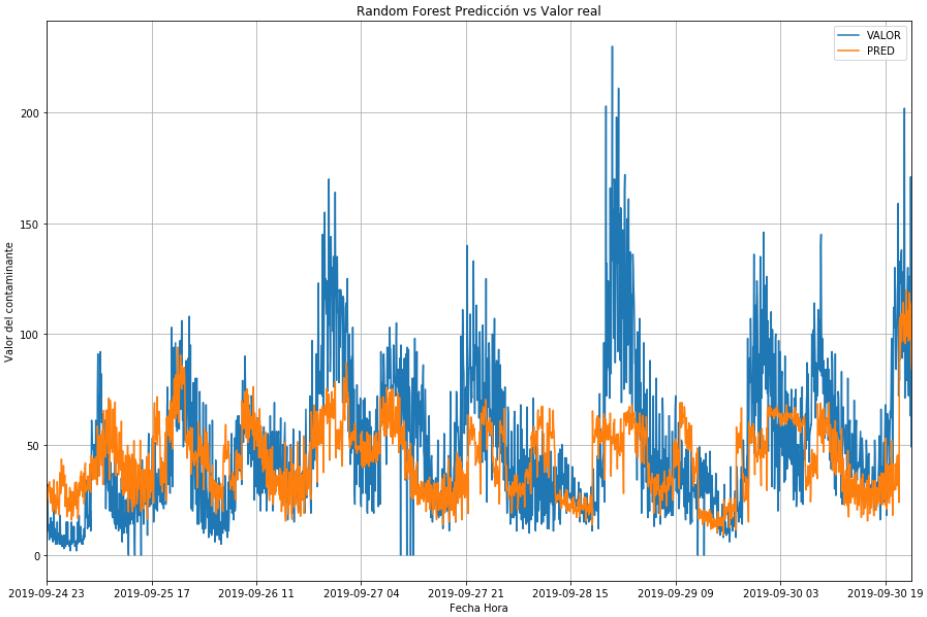


Random Forest

Es un promedio de árboles predictores generados de forma aleatoria.

Tras entrenar el modelo y evaluarlo, el resultado es el siguiente:

```
Mean squared error: 801.60
```

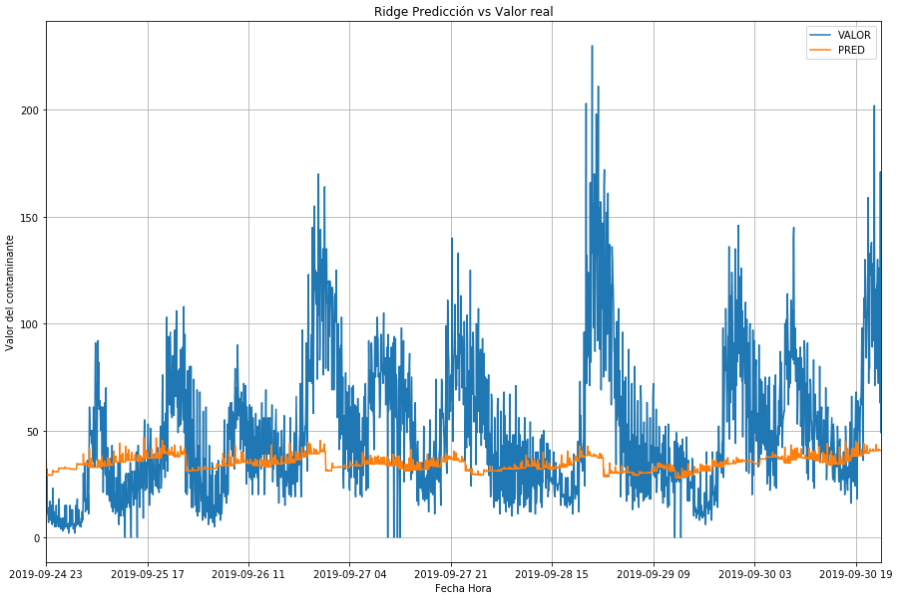


Ridge

Regulariza el modelo resultante imponiendo una penalización al tamaño de los coeficientes de la relación lineal entre las características predictivas y la variable objetivo.

Tras entrenar el modelo y evaluarlo, el resultado es el siguiente:

```
Mean squared error: 1255.27
```

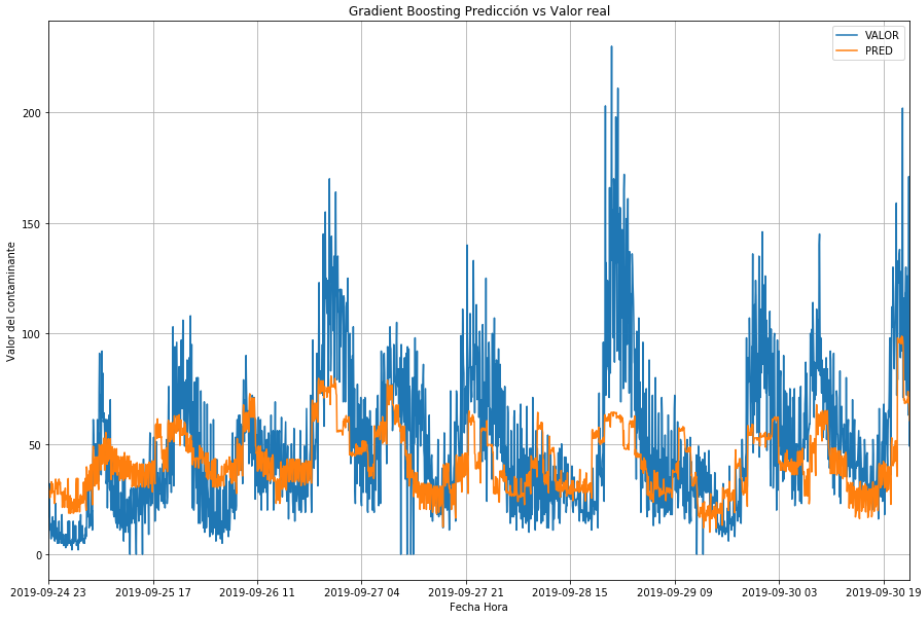


Gradient Boosting

Realiza una optimización arbitraria de una función de pérdida diferenciable recurrente en diferentes árboles de decisiones. Cada uno de los cuales es un modelo en forma de árbol de decisiones y sus posibles consecuencias.

Tras entrenar el modelo y evaluarlo, el resultado es el siguiente:

```
Mean squared error: 757.15
```

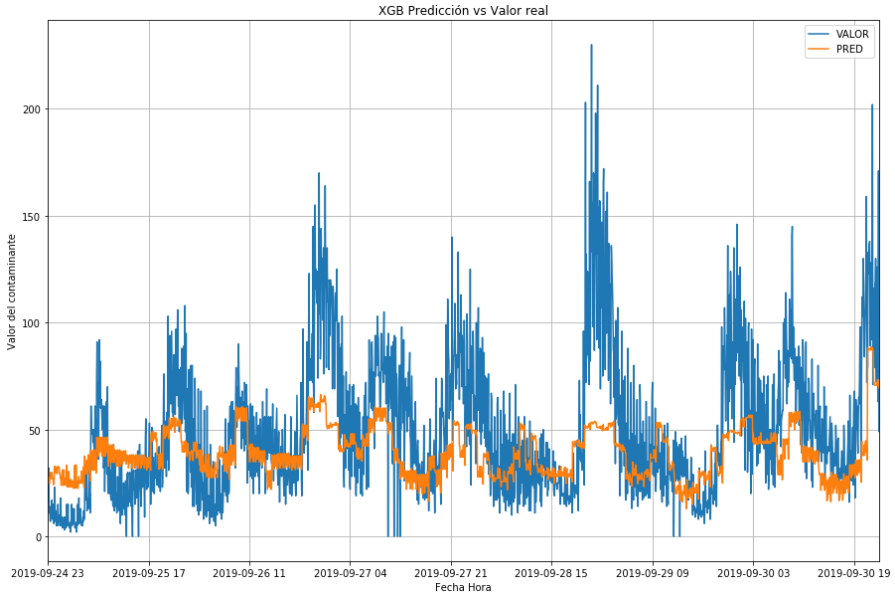


XGBoost

Es un algoritmo que intenta predecir de forma apropiada una variable de destino mediante la combinación de un conjunto de estimaciones a partir de un conjunto de modelos más simples y más débiles.

Tras entrenar el modelo y evaluarlo, el resultado es el siguiente:

```
Mean squared error: 851.06
```

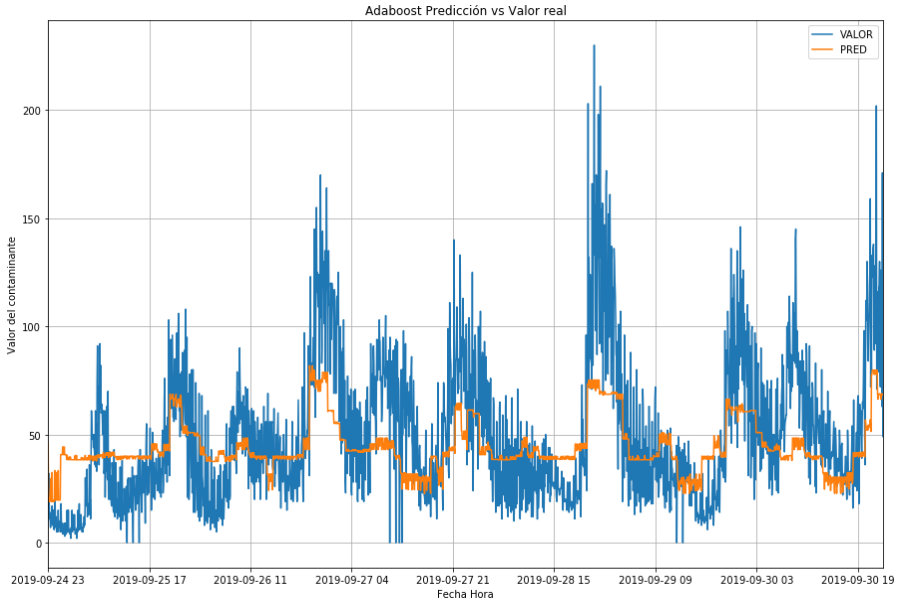


AdaBoost

Basado en la idea de que contar con un grupo de expertos para tomar decisiones es mejor que tener uno solo. Al grupo de expertos se le conoce como ensamble y éste representa un clasificador fuerte, es decir, un clasificador con una precisión muy buena. Por su parte, a los expertos que conforman el ensamble se les denomina clasificadores débiles, es decir, clasificadores con una precisión menor.

Tras entrenar el modelo y evaluarlo, el resultado es el siguiente:

```
Mean squared error: 729.03
```



Conclusión

Después de analizar el dataset y sacar conclusiones, se puede repetir la confirmación que efectivamente la contaminación puede tener influencia por parte de la estacionalidad, meteorología y tráfico. Pero esta afirmación ha de ser confirmada con posteriores análisis con otros periodos temporales ya que este no es concluyente, debido a las conclusiones ya mencionadas en la parte de la exploración.

En cuanto a los modelos, se ha optado como un acercamiento o un producto mínimo viable por utilizar modelos de regresión para obtener la predicción de los valores de los contaminantes en base a la meteorología, tráfico y estacionalidad. Para ello se ha utilizado como ejemplo uno de los contaminantes y con el resultado se aplicará al resto del mismo modo visto que las variables y el objetivo es el mismo. En base a ello se han seleccionado varios modelos para entrenarlos y compararlos entre sí con el error cuadrático medio y el gráfico lineal de los valores reales frente a lo predicho. Este primero da la idea de cuanto se acerca la predicción al valor real para tener una primera métrica a la hora de la toma de decisión. El gráfico ayuda a ver si sigue la línea del comportamiento real de una forma regular o no. Teniendo en cuenta estos dos puntos de decisión, el modelo que mejor a cumplido esto ha sido el Random Forest. Qué aunque no ha tenido el mejor valor del error cuadrático medio, si acerca al máximo al comportamiento de la misma. Al final en los modelos se busca no solo la mejor métrica si no que trate de predecir lo más fiel posible la realidad. Por lo que se procederá a ponerlo en producción.

- Algoritmo seleccionado: Random Forest
- Error cuadrático medio: 801.60
- Importancia de las variables según el modelo:

```
(0.15688167245711773, 'HORA')
 (0.040186539473707227, 'DIASEMANA')
 (0.10767601603633795, 'trafico_intensidad')
 (0.0702555035952913, 'trafico_ocupacion')
 (0.07446310695265812, 'trafico_carga')
 (0.029481805495204922, 'dmax')
 (0.0209553442965493, 'dv')
 (0.07325425445612843, 'hr')
 (0.011834654157470384, 'inso')
 (0.0010828598963042026, 'prec')
 (0.0768512536704821, 'pres')
 (0.036965174726482565, 'rviento')
 (0.07426293787062459, 'ta')
 (0.010772456648729405, 'vis')
 (0.21507642026691187, 'vmax')
```

Añadir el orden de la importancia de las variables que refleja el modelo se alinea con mucho sentido a lo que inicialmente y sobre la teoría entendemos por los contaminantes del aire ya mencionados. Refleja de mayor a menor importancia:

- Viento. Siendo contaminantes del aire, es totalmente lógico que las velocidades máximas del viento afecten directamente a estos contaminantes.
- Estacionalidad con respecto a la hora del día. Tiene sentido que según sea hora punta o no la contaminación del aire incrementa. Variable de HORA
- Intensidad del tráfico. Muchos de los contaminantes son parcial o totalmente generados por las emisiones de vehículos, por lo que se ve normal que la tercera variable más importante sea la intensidad del tráfico.

Por lo tanto tal y como se mencionaba, el modelo refleja claramente la sospecha inicial de donde parte el proyecto, el tráfico y la meteorología pueden influir directamente en la contaminación del aire.

De todos modos mencionar que esta selección del modelado es un primer acercamiento. Se ha podido ver que al fin y al cabo es una serie a la que le puede afectar varios factores. Por lo que se ve como mejora, que en próximos pasos se siga no solo con el análisis, si no con la implementación de series temporales con multivariables o redes neuronales como LSTM que puedan llegar a predecir con mayor exactitud.

Recursos de infraestructura

En esta primera versión del proyecto desarrollada, hay dos requisitos de infraestructura necesarias, por un lado sería el cluster de ICEMD y el otro diversas aplicaciones en local.

La parte del cluster de ICEMD, es un cluster de Hortonworks del cual son tecnologías necesarias para este proyecto, HDFS, Flume, Hive y Spark. La utilización del mismo es gratuita como alumno, tiene alta disponibilidad de uso y servicio de mantenimiento. Por lo que es el lugar idóneo para las necesidades del proyecto.

Por otra parte, se hace uso de recursos locales, una máquina virtual de linux con las librerías necesarias instaladas para ejecutar scripts de python por un lado que agreguen la librería de H3. Por otro tiene instalados el motor de búsqueda de Elasticsearch y la herramienta de visualización de Kibana para la parte final de visualización de los datos. Esta última parte se lleva a cabo en local por falta de permisos de instalación las herramientas en el cluster y ser un modo sencillo y eficiente de llegar a un producto mínimo viable en el proyecto.

Data Governance

Linaje de los datos

Datos de tráfico

Estos datos permiten conocer el estado del tráfico en Madrid y se obtienen a través de un xml que se disponibiliza mediante una url. Son unos datos con una periodicidad de actualización de 5 minutos y ofrecen la siguiente información. Se van a conservar todos los datos de entrada excepto las coordenadas UTM (st_x y st_y), que serán transformadas en longitud y latitud.

Datos de tráfico	**Descripción**
idelem	Identificador del punto de medida.
descripción	Denominación del punto de medida.
accesoAsociado	Código de control relacionado con el control semafórico para la modificación de los tiempos.
intensidad	Intensidad de número de vehículos por hora
ocupación	Porcentaje de ocupación del punto de control por los vehículos
carga	Parámetro de carga del vial en función de la intensidad, ocupación y características de la infraestructura
nivelServicio	Nivel de servicio
intensidadSat	Intensidad de saturación de la vía en vehículo/hora y que se corresponde con el máximo número de vehículos que pueden pasar en el acceso a la intersección manteniéndose la fase verde del semáforo.
error	Código de control de la validez de los datos del punto de medida.
subarea	Identificador de la subárea de explotación de tráfico a la que pertenece el punto de medida.
longitud	Coordenada que representa la ubicación de la estación de medida de tráfico en el eje de la longitud
latitud	Coordenada que representa la ubicación de la estación de medida de tráfico en el eje de la

	latitud
fecha	Fecha y hora exacta del momento de la medición
localización	Punto de la ubicación de la estación de medida de tráfico
hex_id	Id de el hexágono de la rejilla de H3, correspondiente al punto de ubicación de la estación de medida de tráfico
GEOM	Poligono de puntos de ubicación geográfica que delimitan y conforman el hexágono correspondiente a la ubicación de la estación de medida de tráfico.

Datos origen:

Datos de tráfico	
idelem	nivelServicio
descripción	intensidadSat
accesoAsociado	error
intensidad	subarea
ocupación	st_x
carga	st_y
nivelServicio	Nivel de servicio

Transformaciones:

- Transformación de las variables st_x y st_y coordenadas UTM a longitud y latitud.
- Aplicar la rejilla de la librería de uber H3 en base a la posición.

Datos de contaminación del aire

El Sistema Integral de la Calidad del Aire del Ayuntamiento de Madrid permite conocer en cada momento los niveles de contaminación atmosférica en el municipio y se obtiene en csv a través de una url que se disponibiliza. Son unos datos con una periodicidad de actualización de una hora. Datos filtrados sin errores (V00) con los valores de contaminación, fecha y datos de la medición e ubicación.

Datos de contaminación del aire	**Descripción**

PROVINCIA	Código de la provincia.
MUNICIPIO	Código del municipio.
ESTACIÓN	Código de la estación
PUNTO_MUESTREO	Incluye el código de la estación completo (provincia, municipio y estación) más la magnitud y la técnica de muestreo
NOMBRE_ESTACION	Nombre de la estación que ha hecho la medición
LONGITUD	Coordenada que representa la ubicación de la estación de medida del aire en el eje de la longitud
LATITUD	Coordenada que representa la ubicación de la estación de medida del aire en el eje de la latitud
MAGNITUD	Nombre de la magnitud que se está midiendo
FECHA	Fecha y hora del momento de la medición
LOCALIZACION	Punto de la ubicación de la estación de medida del aire
HORA	Hora del momento de la medición
VALOR	El valor del dato de la magnitud por cada hora.
hex_id	Id de el hexágono de la rejilla de H3, correspondiente al punto de ubicación de la estación de medida del aire
GEOM	Poligono de puntos de ubicación geográfica que delimitan y conforman el hexágono correspondiente a la ubicación de la estación de medida del aire

Datos origen:

Datos de contaminación del aire	
provincia	ano
municipio	mes
estación	dia
magnitud	H00
punto_muestreo	V00

Transformaciones:

- Filtrado de los datos sin errores de medición
- Unir cada estación con la posición correspondiente en longitud y latitud y descripcion de la ubicación.
- Agregar la columna de fecha y hora teniendo en cuenta las columnas ano, mes, día y H00.
- Aplicar la rejilla de la librería de uber H3 en base a la posición.

Por otro lado se necesitarán saber las observaciones y predicciones meteorologicas y ver cómo estas influyen. Para ello se va a hacer uso de los datos que pone a disposición AEMET a través de su API.

Datos de observación

Datos de observación proporcionados por el servicio de Open Data de Aemet, en horarios de las últimas 24 horas (actualización cada hora). Estos datos se disponibiliza mediante api con api key en formato json. Datos de valores de observación meteorológica se mantienen tal cual se obtienen desde el origen.

Datos de observación	**Descripción**
dmax	Dirección del viento máximo registrado en los 60 minutos anteriores a la hora indicada por 'fint' (grados).
dv	Dirección media del viento, en el período de 10 minutos anteriores a la fecha indicada por 'fint' (grados).
hr	Humedad relativa instantánea del aire correspondiente a la fecha dada por 'fint' (%)
prec	Precipitación acumulada, medida por el pluviómetro, durante los 60 minutos anteriores a la hora indicada por el período de observación 'fint' (mm, equivalente a l/m2)
pres	Presión instantánea al nivel en el que se encuentra instalado el barómetro y correspondiente a la fecha dada por 'fint' (hPa).
rviento	Recorrido del viento durante los 60 minutos anteriores a la fecha indicada por 'fint' (Hm)
ta	Temperatura instantánea del aire correspondiente a la fecha dada por 'fint' (grados Celsius)

tamax	Temperatura máxima del aire, valor máximo de los 60 valores instantáneos de 'ta' medidos en el período de 60 minutos anteriores a la hora indicada por el período de observación 'fint' (grados Celsius).
tamin	Temperatura mínima del aire, valor mínimo de los 60 valores instantáneos de 'ta' medidos en el período de 60 minutos anteriores a la hora indicada por el período de observación 'fint' (grados Celsius).
vmax	Velocidad máxima del viento, valor máximo del viento mantenido 3 segundos y registrado en los 60 minutos anteriores a la hora indicada por el período de observación 'fint' (m/s).
vv	Velocidad media del viento, media escalar de las muestras adquiridas cada 0,25 ó 1 segundo en el período de 10 minutos anterior al indicado por 'fint' (m/s).
inso	Duración de la insolación durante los 60 minutos anteriores a la hora indicada por el período de observación 'fint' (horas)
stdvv	Desviación estándar de las muestras adquiridas de velocidad del viento durante los 10 minutos anteriores a la fecha dada por 'fint' (m/s)
stddv	Desviación estándar de las muestras adquiridas de la dirección del viento durante los 10 minutos anteriores a la fecha dada por 'fint' (grados)
vis	Visibilidad, promedio de la medida de la visibilidad correspondiente a los 10 minutos anteriores a la fecha dada por 'fint' (Km)
ts	Temperatura suelo, temperatura instantánea junto al suelo y correspondiente a los 10 minutos anteriores a la fecha dada por 'fint' (grados Celsius)
pres_nmar	Valor de la presión reducido al nivel del mar para aquellas estaciones cuya altitud es igual o menor a 750 metros y correspondiente a la fecha indicada por 'fint' (hPa)
tpr	Temperatura del punto de rocío calculado correspondiente a la fecha 'fint' (grados Celsius)

Datos de predicción

Datos de predicción proporcionados por el servicio de Open Data de Aemet, de hora en hora. Estos datos se disponibiliza mediante api con api key en formato json/xml. Datos de predicción meteorológica y con la fecha para la que se hace la predicción.

Datos de predicción	**Descripción**
periodo	periodo de las mediciones
estadoCielo	Descripción del estado del cielo
fecha	Período de validez de la Predicción.
humedadRelativa	Valor de la humedad relativa.
nieve	Cantidad total de nieve que se prevé que caiga durante la hora anterior.
ocaso	Hora del atardecer.
orto	Hora del amanecer.
precipitacion	Cantidad total de precipitación durante la hora anterior.
probNieve	Valor de la probabilidad de precipitación de nieve.
probPrecipitacion	Valor de la probabilidad de precipitación.
probPrecipitacion_periodo	El periodo en el que se ha tomado la medición de la probabilidad de precipitación
probTormenta	Valor de la probabilidad de tormenta.
probTormenta_periodo	El periodo en el que se ha tomado la medición de la probabilidad de tormenta
sensTermica	Valor de la sensación térmica.
temperatura	Valor de la temperatura.
viento.direccion	Dirección del viento.
viento.velocidad	Velocidad del viento.
rachaMax	Valor de la Racha máxima.
fecha	fecha para la que se hace la predicción

Datos origen:

Datos de predicción	Descripción
estadoCielo	probPrecipitacion
fecha	probTormenta
humedadRelativa	sensTermica
nieve	temperatura
ocaso	viento.direccion
orto	viento.velocidad
precipitacion	rachaMax
probNieve	

Transformaciones:

- Añadir varias columnas con el periodo de la medición (hacer referencia a la hora de la medición), tanto en general como alguna diferente para ciertas medidas.
- Se añade una columna con la fecha a la que se le hace la predicción

Data legal

El contenido de los datos utilizados en este proyecto no contienen ningún tipo de dato personal y provienen todos de APIs de datos abiertos, por lo que inicialmente no hace falta tomar ninguna medida adicional para anonimizar y preservar la privacidad de ningún usuario.

No obstante se usa código y librerías Open Source, como la librería de H3 de Uber. Por lo que todo el proyecto y su código se disponibilizará públicamente en GitHub.

Seguridad

La parte más mayor peso y con uso de más tecnológicas además de el almacenamiento de los datos se realiza en el cluster de Hortonworks que disponibiliza para el alumnado ICEMD. Este tiene todas las medidas de seguridad necesarias acceso al mismo a través de usuario y contraseña únicos por alumno. Además la IP y puertos de acceso a cualquiera de las tecnologías ubicadas en el cluster son privados por lo que no se puede acceder a ellas externamente.

Planificación del proyecto

El objetivo de este proyecto es poder visualizar de calidad del aire y tráfico y generar predicción de la contaminación. El mismo tiene un componente bastante complejo de elaborar todo el flujo y la arquitectura de los datos hasta que llegue a la visualización. Al tener una carga tan grande en la parte de elaborar la arquitectura y el flujo de los datos no se ve viable poder tener un producto mínimo que presentar y sobre el que poder iterar para ir mejorandolo.

De modo que por la complejidad de llegar a un producto mínimo y por razones organizativas se más eficiente que el proyecto se desarrolle en un modelo de waterfall dividido en varias fases.

Por lo que el proyecto tendrá 5 fases:

1. Fase de análisis y preparación del proyecto
2. Fase de elaborar la arquitectura y flujo de los datos
3. Fase de visualización
4. Fase de modelización
5. Fase de documentación

Fase 1: Análisis y preparación	**Descripción**	**Duración**
Tarea 1.1: Investigación de los datos	Buscar e investigar las fuentes de los datos necesarios para el proyecto, su disponibilidad y ver si alguno de ellos puede tener datos personales o de carácter privado.	1 semana
Tarea 1.2: Diseño de la arquitectura	Analizar los requisitos necesarios del proyecto (técnicos y económicos) y diseñar una arquitectura que genere un flujo del dato óptimo para el objetivo del proyecto.	1 semana
Tarea 1.3: Preparación inicio proyecto	Elaborar la documentación a seguir durante el proyecto y preparar los entornos de desarrollo.	1 semana

Fase 2: Desarrollo arquitectura	**Descripción**	**Duración**
Tarea 2.1: Ingesta de los datos	Hacer las llamadas a cada una de las APIs de las fuentes de los datos mediante scripts de python y con Flume depositarlos en HDFS.	2 semanas
Tarea 2.2: Procesamiento de los datos	Utilizando Spark leer los datos de HDFS, limpiarlos y prepararlos para depositarlos en diferentes tablas de Hive. Pensando que el proceso se ejecutará cada hora.	2 semanas
Tarea 2.3: Incluir libreria H3 y almacenar los datos en Elasticsearch	Elaborar diferentes scripts de python para incluir la rejilla de la librería de Uber H3 a los dataset de contaminación y tráfico y que los almacene en Elasticsearch	2 semanas
Tarea 2.4: Verificar flujo de dato	Comprobar el flujo completo del dato, corregir errores y verificar que los datos llegan a Kibana perfectamente	1 semana

Fase 3: Visualización	**Descripción**	**Duración**
Tarea 3.1: Preparar la capa de visualización de los datos de tráfico.	Montar la capa visualización de la congestión de tráfico utilizando los datos almacenados en Elasticsearch.	0.5 semana
Tarea 3.2: Preparar la capa de visualización de los datos de contaminación.	Montar las capas necesarias para visualizar los diferentes magnitudes de medida y sus valores almacenados en Elasticsearch.	1 semana

Fase 4: Modelización	**Descripción**	**Duración**
Tarea 4.1: Preparación y prueba de los datos	Teniendo en cuenta el algoritmo a utilizar, Prophets, preparar los datos (contaminación, tráfico y	0.5 semana

	observación meteorológica).	
Tarea 4.2: Modelado	Entrenar el modelo con los datos preparados y comprobar su efectividad.	0.5 semana
Tarea 4.3: Productivizar	Generar el algoritmo para productivizar e incluirlo en un script de python ejecutarlo dentro del flujo de los datos. Estos datos se almacenarán en Elasticsearch.	1 semana

Fase 5: Documentación	**Descripción**	**Duración**
Tarea 5.1: Comprobación del funcionamiento del proyecto	Ver que todas las partes del flujo de los datos funcionan correctamente y todo se visualiza correcto.	0.2 semana
Tarea 5.2: Documentación de todo el código	Documentar todo el código elaborado, explicarlo, hacerlo entendible y disponibilizar en Github.	0.4 semana
Tarea 5.3: Grabar visualizaciones	Grabar las pruebas de visualizaciones necesarias para añadir al entregable del proyecto.	0.2 semana
Tarea 5.4: Memoria	Elaborar el documento explicativo del proyecto, Memoria, siguiendo las pautas exigidas.	1.8 semanas
Tarea 5.5: Presentación	Elaborar una presentación ejecutiva del proyecto para adjuntar al entregable.	0.2 semana
Tarea 5.6: Entregable	Generar los ejecutables del código. Aunar todo el código y documentación en un .zip entregable con un documento explicativo de cómo proceder con el proyecto y entregarlo.	0.2 semana

Casos de uso y monetización

Explotación y seguimiento por parte del ayuntamiento de Madrid

El caso del ayuntamiento y comunidad de Madrid por las necesidades tan generales de gestión del tráfico, control de calidad del aire y gestión del transporte público, necesita hacer seguimiento de toda la información y con apoyo visual. Por lo que podría plantearse un acceso a los datos procesados y agregados mediante API con una cuota de pago mensual y un apoyo extra si fuera necesario configurarles un almacenamiento y una herramienta de visualización. Otro modo sería la venta o concesión del total del proyecto para su propia gestión.

El uso que el ayuntamiento y comunidad podrían hacer serían:

- Detección de puntos de alta concentración de tráfico y contaminación. La solución serían medidas de apoyo de transporte público (nuevas líneas de bus o mayor frecuencia en metro o cercanías) o medidas de regulación de la circulación en los puntos y horas concretas en los que suceden.
- Regular las emisiones de CO2 o cualquier otro agente contaminante en la calidad del aire. Con la visualización podrían ver cuales son los puntos de mayor contaminación y ver si son a causa del tráfico o por algún otro motivo como el sector industrial. De este modo podrían tomar las medidas necesarias para atajar de un modo más eficaz dichas emisiones.

Real Estate

Con los datos y conclusiones que se obtienen en este proyecto pueden conocerse las zonas de mayor foco de tráfico o con contaminación del aire más alta. Esto afecta directamente a la calidad de vida de los inquilinos de dichas zonas. Por lo que el sector inmobiliario podría valerse de dicha información para recalcular la compra-venta de pisos y alquileres teniendo en cuenta estos valores.

De este modo se pondría a disposición dos formatos de consumo de la información. En el primero sería al igual que en el caso anterior, la disposición de una API con los datos que luego ellos lo consumiría en su entorno con un apoyo para implantar la parte de visualización, con un pago mensual. El segundo serían una elaboración de informes por la parte del proyecto que se venderían individualmente.

Transporte privado

Este sector está evolucionando en la parte de oferta no contaminante. Por lo que valiéndose de la información obtenida en este análisis y predicción, podrían reubicar opciones de transporte eléctricas como carsharing o patinete eléctrico para desahogar el tráfico sin emitir más contaminación.

Los datos los podrían obtener a través de una API de pago mensual y acompañamiento para implantar el modelo de visualización. De este modo no solo tendrían acceso a cómo se comportan históricamente si no incluso a las predicciones para poder adelantarse a próximos momentos de alta congestión de tráfico.

Retorno de la inversión

Una vez vistos los casos de uso, hace falta contabilizarlo y ver la viabilidad del proyecto con los gastos e ingresos.

Este proyecto se inicia como trabajo de final de máster en ICEMD haciendo uso en los primeros meses del propio cluster que facilita la escuela., Una vez logrado el producto mínimo viable la intención es llevarlo a cloud, inicialmente al Data Hub de Cloudera y con el uso de Elastic Cloud Service. De este modo se ahorran los primeros meses de la plataforma. Del mismo modo la idea es que los perfiles que trabajen en el proyecto sea por teletrabajo, disponibilizando ellos su pc para trabajar. De este modo el gasto de alquiler, luz, agua y material no aplica. Además de ser un formato de trabajo cada vez más en auge que da comodidad y satisfacción al empleado no obligándole a desplazarse para trabajar.

De inicio los perfiles a contratar solo sería un Data Engineer Junior con alguna experiencia. Una vez los 3 primeros meses avanza la arquitectura ya se incorporaría a un Data Scientist para las labores de modelado y análisis de los datos.

Todo ello se resume en la siguiente tabla de inversión:

Inversión (€)	**Mes 1**	**Mes 2**	**Mes 3**	**Mes 4**	**Mes 5**	**Mes 6**	**Mes 7**	**Mes 8**	**Mes 9**	**Mes 10**	**Mes 11**	**Mes 12**	**Total proyecto**
Data Hub Cloudera	0	0	0	0	784	784	784	784	784	784	784	784	**6,274**
Elastic cloud service	0	0	0	0	65	65	65	65	65	65	65	65	**523**
Data Engineer	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	**24,000**
Data Scientist	0	0	0	2,200	2,200	2,200	2,200	2,200	2,200	2,200	2,200	2,200	**19,800**
Seguros sociales (30%)	600	600	600	1,260	1,260	1,260	1,260	1,260	1,260	1,260	1,260	1,260	**13,140**
Gastos de constitución	3,000	0	0	0	0	0	0	0	0	0	0	0	**3,000**
Gastos comerciales	0	0	0	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	**9,000**
Total gastos	**5,600**	**2,600**	**2,600**	**6,460**	**7,310**	**7,310**	**7,310**	**7,310**	**7,310**	**7,310**	**7,310**	**7,310**	**75,737**

Por otro lado es necesario tener ingresos lo antes posible para la viabilidad del proyecto. Inicialmente lo más accesible será la generación de informes a empresas de Real State que quieran añadir valor a sus estudios de mercado. Aunque la idea es extender los informes a todo quien lo solicite. Posteriormente se disponibilizará inicialmente mediante API los datos agregados de las observaciones y las predicciones con el modelado simple inicial. Una vez se haya concluido el segundo y mejor modelado se pasará a disponibilizar de modo independiente los datos mediante API de las predicciones. Y finalmente se perseguirán empresas privadas de

transporte y órganos públicos (Comunidad de Madrid, Ayuntamiento de Madrid, etc.) para tratar de venderles el servicio completo de instalación de la plataforma de visualización y el servicio de mantenimiento, con el beneficio de obtener todas las mejoras futuras sin gastos añadidos. Las tarifas correspondientes serían las siguientes:

- Informe: 500€
- API datos observación agregados: 100€/mes
- API datos de predicción: 100€/mes
- Servicio completo: 5000€/mes

Por lo que la estimación de ingresos a 12 meses sería la siguiente:

Ingresos (€)	**Mes 1**	**Mes 2**	**Mes 3**	**Mes 4**	**Mes 5**	**Mes 6**	**Mes 7**	**Mes 8**	**Mes 9**	**Mes 10**	**Mes 11**	**Mes 12**	**Total proyecto**
Venta completa del servicio + mantenimiento	0	0	0	0	0	0	5,000	5,000	5,000	10,000	10,000	10,000	**45,000**
Disposición de los datos mediante API	0	0	0	0	100	200	400	600	600	900	1,500	1,700	**6,000**
Venta de informes	0	0	0	500	1,000	2,000	2,000	3,000	4,000	4,000	5,000	6,000	**27,500**
Total ganancias	**0**	**0**	**0**	**500**	**1,100**	**2,200**	**7,400**	**8,600**	**9,600**	**14,900**	**16,500**	**17,700**	**78,500**

Por lo tanto el ROI, indicador de la rentabilidad de un proyecto relacionando el beneficio obtenido a partir del mismo y la inversión realizada para su desarrollo, sería del 3.65%. Esto indica que por cada 100 euros invertidos el primer año se obtendría ya un retorno de 3.65 euros.

Conclusiones

Este trabajo culmina en final del máster de Big Data Management. Como tal se ha tratado de escoger un tema que sea de negocio pero a su vez esté alineado con la actualidad. Con él se pretende plasmar todos los conocimientos adquiridos en el transcurso del máster, tratando de incluir los conocimientos adquiridos en cada módulo, cada charla con cada uno de los profesores y experiencias vividas en todos los trabajos realizados. Este trabajo al final es el reflejo de todo ello en forma de una simulación de un proyecto que simula estar en el ámbito real de negocio. Y te obliga a saber ver las cosas como negocio y tratar de monetizar los proyecto. Pero sin olvidar en es vital un proyecto tener paciencia y ser exhaustivo con los desarrollos, como lo es un técnico. Hay que aprender a mezclar ambos.

Por ello se ha dividido en una parte inicial donde se busca información, se plantea la idea y se entiende el entorno. Posteriormente se analizan los datos y se plantea la arquitectura a desarrollar. En este apartado ha sido uno de los que más han reflejado un entorno laboral real con todos los problemas que puedes llegar a encontrar. Ciertamente duro pero enriquecedor, se aprende no solo a lidiar con los problemas sino incluso a lidiar con unos planteamientos iniciales erróneos o de buscar una primera solución que no la mejor pero si viable. Ella permitirá llegar a un producto comercializable y su vez esto dejará seguir mejorandolo. Todo eso no se hubiera aprendido de no ser por lo práctico que ha sido tanto el proyecto como el máster. Finalmente el planteamiento del modelo junto con el conocimiento de los datos, que aunque solo se ha planteado la punta del iceberg, pero abre la puerta a poder seguir trabajando en los datos y tratar de obtener el mejor modelo aplicable al caso que concierne el proyecto.

Los datos inicialmente no han causado excesivos problemas, pero es cierto que han tenido que adaptarse a las necesidades. Al final en la vida real los datos muchas veces no son perfectos y hay que lidiar con ello, en este caso aun y siendo datos reales no ha sido excesivamente difícil tratar con ellos. No obstante la parte del procesado de datos si ha dado más problemas, incompatibilidades de librerías que han obligado a enfocar desde otro prisma el proyecto o no tener acceso a los datos para terminar el flujo. Seguramente si se iniciase hoy el proyecto, el planteamiento se haría algo diferente, todo fruto del aprendizaje vivido.

Estas sensaciones y aprendizaje vivido ha sido gracias a los profesores y sus conocimientos, pero también al apoyo de los compañeros, que al igual que en el mundo real se ha sabido trabajar en grupo cuando ha sido necesario y apoyar a quien necesitaba ayuda.

Por todo ello estas conclusiones toman parte para trasladar una sensación del trabajo bien hecho, del esfuerzo empleado y de la satisfacción de poder realizar un proyecto como el que se presenta en este trabajo de fin de máster.

Trabajo futuro

Una vez finalizado el producto mínimo viable para poner en marcha el proyecto hay diversos focos a trabajar el el futuro.

El primero de todos es lanzar a producción el modelo predictivo desarrollado con Random Forest y conectarlo con la visualización. Esto terminará dando la visión final de la funcionalidad del proyecto completo, pudiendo tomar decisiones en base a no solo valores pasados si no posibles valores futuros. Una vez esté en producción se comienza a preparar el nuevo modelado mucho mejor adaptado a la serie de los datos de contaminación y los factores que pueden afectar. Para ello la idea es tratar de seleccionar modelos de series temporales que admitan múltiples variables (podría ser Prophet de Facebook) o redes neuronales que se adapten al caso (podría ser redes como LSTM).

El segundo foco, es llevar toda la arquitectura a cloud, una vez que se ha probado que el flujo de datos funciona. De esta manera se independiza la arquitectura de la infraestructura de Hortonworks proporcionada inicialmente por ICEMD para el producto mínimo viable. Además se sirve de las herramientas dispuestas por los operadores de cloud para eficientar el flujo de ingesta, procesado y visualización de una forma sencilla y económica teniendo en cuenta la alternativa a hacer una inversión de una infraestructura *on premise*. Incluso podría valerse de las opciones de los proveedores cloud para tener el servicio en alta disponibilidad para clientes mediante APIs.

En paralelo de esta parte más técnica del proyecto, con esta primera versión ya podría empezar a visitar posibles clientes y hacerles una demo del producto, para poder monetizar la inversión realizada lo antes posible.

Una vez estas dos partes, tanto la técnica como la comercial estén encaminadas, habría que seguir investigando los valores contaminantes, el tráfico y posibles variables nuevas que puedan llegar aportar valor al servicio y poder incrementar la monetización del mismo.

Bibliografía

Documentación del Aula Virtual del Master Big Data Management

H3: Uber's Hexagonal Hierarchical Spatial Index

Elastic Stack and Product Documentation

AEMET OpenData

Calidad del aire. Datos en tiempo real

Calidad del aire. Datos horarios años 2001 a 2019

Calidad del aire. Estaciones de control

Tráfico. Datos del tráfico en tiempo real

Tráfico. Histórico de datos del tráfico desde 2013

Tráfico. Ubicación de los puntos de medida del tráfico

datosclima.es -Base de datos Meteorológica-

Calidad del aire y salud (Comunidad de Madrid)

Big Data Analytics: A Necessary Roadmap for Enterprises - Scientific Figure on ResearchGate.
[Figura 1: Evolución del Big Data]

Wikipedia

Estudio ESADE

WinShuttle Historia Big Data

Ideas para tu empresa (vodafone), Historia Big Data

Computación en la nube

El origen de: El Cómputo en la Nube

El origen del Big Data

Bigtable

El impacto del Mobile World Congress en una visualización dinámica de BBVA y CartoDB

BBVA analiza el turismo en España a partir de las transacciones con tarjeta

Designing for the Discovery of Big Data (Carto)

Tourists from all countries (BBVA & Vizzuality)

Cuando Big Data es igual a 'real money': tres casos de éxito de Big Data orientado a negocio

Quién lidera en España la carrera de la Industria 4.0

¿Cómo convertir el Big Data en Smart Data?

El 'Chacho' y varios jugadores de la ACB invierten un millón de euros en la 'start-up' española NBN23

NBN23