TECNOLOGÍAS EMERGENTES Y DATOS ABIERTOS: ANALÍTICA PREDICTIVA.

GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
datos.gob.es
Iniciativa aporta

Colección Awareness, Inspire, Action

Tecnologías emergentes y datos abiertos: Analítica Predictiva

INTRODUCCIÓN **4**

METODOLOGÍA **5**

AWARENESS **6**

Conceptos clave 7

Analítica predictiva. Un lío de palabras 8

Un poco de historia 9

INSPIRE **10**

Industria: detección de anomalías y mantenimiento predictivo 10

Mercados: predicción de precios y demanda 12

Salud: diagnóstico aumentado de imagen médica 13

ACTION **14**

Introducción al conjunto de datos 14

Metodología seguida en el análisis 15

EDA - Analítica exploratoria de los datos 16

Los datos 17

Técnicas de analítica predictiva utilizadas 21

CONCLUSIONES **24**

PRÓXIMA PARADA... ..... **25**

Libros 25

Tutoriales y cursos 26

Algunos casos de uso muy interesantes 26

Contenido elaborado por Alejandro Alija, Experto en Transformación Digital y datos abiertos.

Este estudio ha sido desarrollado en el marco de la Iniciativa Aporta, desarrollada por el Ministerio de Asuntos Económicos y Transformación Digital, a través de la Entidad Pública Empresarial Red.es. Los contenidos y los puntos de vista reflejados en esta publicación son responsabilidad exclusiva de su autor. El equipo Aporta no garantiza la exactitud de los datos incluidos en el estudio. El uso de este documento implica la expresa y plena aceptación de las condiciones generales de reutilización referidas en el aviso legal que se muestra en: http://datos.gob.es/es/aviso-legal.

INTRODUCCIÓN

Toda buena historia empieza por el principio. **En este informe, analizamos el tema de la analítica predictiva de datos** (en adelante, tan sólo analítica predictiva). La analítica de datos se ha consolidado como un término general para una variedad de diferentes iniciativas relacionadas con la inteligencia empresarial (BI - Business Intelligence) y las aplicaciones digitales. Para algunos, es el proceso de analizar información de un dominio en particular, como el análisis de sitios web. Para otros, es la capacidad genérica de analizar datos de cualquier dominio con el objetivo de extraer información de valor que mejore el rendimiento de un proceso (ventas, cadena de suministro o la gestión de una pandemia sanitaria cómo el Covid-19) o un negocio completo (desde grandes cadenas de distribución o la generación de energía eléctrica en una central hidroeléctrica).

Antes de centrarnos en la analítica predictiva, veamos una descripción general de la analítica de datos. Cada vez más, el término analítica se utiliza para describir

> **el análisis de datos estadísticos y matemáticos que agrupa, segmenta, clasifica y predice qué escenarios son más probables que sucedan.**

Independientemente de los casos de uso, la analítica se ha integrado de forma natural en la lengua nativa de cualquier negocio u organización. En los últimos años, la analítica atrae un interés creciente por parte de todos los profesionales de TI (Tecnologías de la Información) y las áreas de negocio de empresas e instituciones de todo el mundo. Son numerosos los ejemplos de empresas que han creado y siguen creando departamentos, áreas, divisiones y hubs alrededor de la analítica. Estas organizaciones han apostado por el conjunto de tecnologías alrededor de la analítica para explotar grandes cantidades de datos generados internamente y disponibles externamente.

En este informe:

- Repasamos los conceptos clave de las técnicas de analítica predictiva en la sección Awareness.
- En la sección Inspire analizamos en detalle algunos de los principales casos de uso.
- Finalmente, en la sección de Action realizamos un caso práctico con datos (abiertos) reales y código.

Tres, dos, uno, ¡Arrancamos!

METODOLOGÍA

Este informe se enmarca dentro de una colección más amplia de recursos sobre tecnologías emergentes y datos abiertos, cuyo objetivo es introducir en la materia al lector mediante el empleo de casos de uso prácticos, sencillos y reconocibles. Al mismo tiempo, se pretende facilitar una guía de aprendizaje práctica para aquellos lectores con conocimientos más avanzados, que, mediante el desarrollo de un caso práctico, puedan experimentar de forma autodidacta con herramientas reales para el análisis y explotación de datos abiertos.

Para conseguir este doble objetivo, el informe se estructura en tres partes bien diferenciadas: Awareness, Inspire y Action, que pueden ser abordadas de forma independiente en cualquier momento y sin necesidad de haber realizado una lectura previa de las otras secciones.

La primera sección, Awareness, sirve de introducción al tema en cuestión (en este informe, la analítica predictiva de datos). Este apartado está indicado para aquel lector que se inicia en el tema por primera vez y trata de abordar la temática de forma sencilla, clara y sin el uso de tecnicismos que dificulten la lectura.

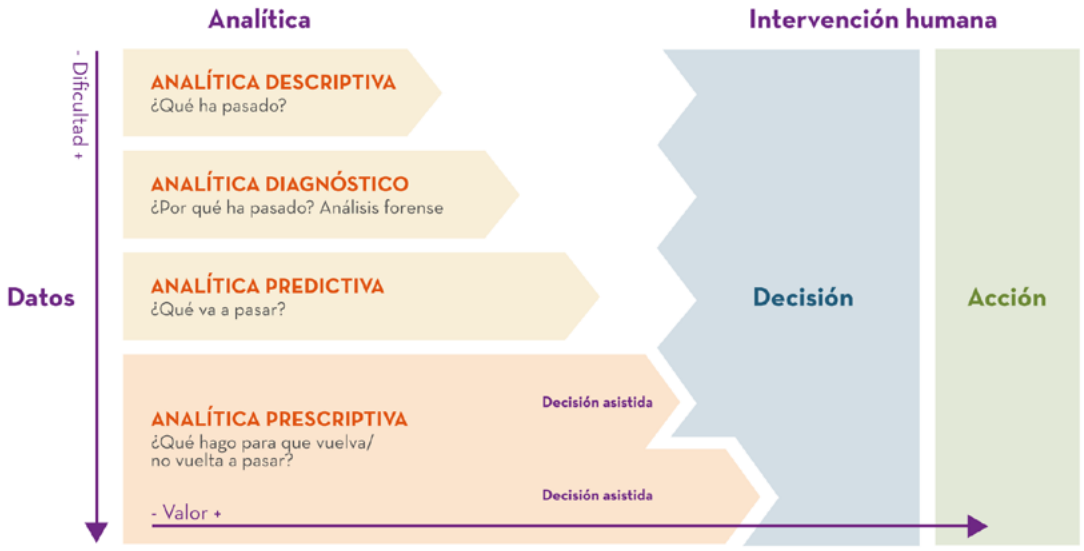
La segunda sección, Inspire, pretende servir de inspiración a aquellos lectores que se han iniciado en la materia y que se preguntan cómo les afecta a ellos en su vida diaria o en su trabajo el tema que se aborda. La forma de identificarse con una tecnología, un campo de la ciencia o cualquier otra materia es verse reflejado en ella. De esta forma, la sección Inspire contiene ejemplos y casos de aplicación de una cierta tecnología en situaciones cotidianas, que facilitan que el lector se identifique y comience a pensar en dicha tecnología como algo que a él también le afecta.

Por último, la sección Action selecciona alguno de los casos de usos explicados en la sección Inspire y lo desarrolla de forma práctica, utilizando para ello, datos y herramientas tecnológicas reales. El ejemplo, desarrollado en Action, se pone a disposición del lector en forma de código y datos abiertos para que éste pueda experimentar y desarrollar con sus propios medios el caso de uso que se aborda en la sección Action.

AWARENESS

Entre los años 2013 y 2017, la consultora Gartner publica una serie de informes en los que analiza los diferentes tipos de analítica desde el punto de vista de las aplicaciones empresariales. Algunas de las figuras empleadas en esos informes se han convertido en estándar para clasificar los diferentes tipos de analítica según el objetivo que se persiga.

CLASIFICACIÓN DE LOS TIPOS DE ANALÍTICA



Fuente original Gartner

Cómo se puede ver en la figura anterior, **la analítica predictiva incluye técnicas que nos permiten anticipar lo que va a ocurrir:**

- ¿sufriremos una rotura de stock?
- ¿Bajará la cotización de una determinada acción?
- ¿Tendrá esta máquina una avería inesperada?
- ¿Aumentará el número de accidentes en carretera el mes que viene?
- ¿Nos visitarán más turistas el próximo puente?
- ¿Hay riesgo de que un empleado nos vaya a dejar?

A este tipo de cuestiones las denominamos preguntas de negocio. Sea cual sea el sector o la actividad a la que te dediques, si tienes responsabilidad sobre el desempeño de una actividad o un proceso, seguro que te haces preguntas como estas en tu día a día. **Sin la utilización de técnicas de analítica predictiva, la respuesta a estas preguntas depende de la experiencia y la intuición subjetiva de la persona que controla el proceso**. Este hecho plantea evidentes desventajas desde el punto de vista organizativo e importantes riesgos para la continuidad del negocio. Las personas expertas en nuestra organización son de vital importancia, pero están sujetas a las limitaciones humanas (cansancio, enfermedades, disponibilidad, etc.). Es por esto, que una buena estrategia pasa por complementar a las personas expertas con herramientas, aplicaciones y capacidades de analítica predictiva que puedan aumentar las capacidades humanas allí donde las personas tenemos limitaciones. Esto último nos hace avanzar hacia la analítica prescriptiva y la decisión asistida y automática.

Conceptos clave

Cómo hemos visto en los párrafos anteriores, la analítica predictiva es un caso particular de analítica, cuyo objetivo es anticipar hechos relevantes de un proceso que ayuden a un humano con la toma de decisión. Así, **la analítica predictiva se caracteriza por los siguientes atributos o aspectos diferenciales:**

Atributos de la analítica predictiva

1. **Pone el énfasis en la predicción.** A diferencia de la analítica descriptiva cuyo objetivo es la exploración de los datos desde diferentes perspectivas (agrupaciones, segmentaciones, etc.), la analítica predictiva pone el foco en modelos que permitan predecir valores futuros de las variables de interés.

2. **Se enfoca en la relevancia para el negocio de los conocimientos resultantes**. La analítica predictiva persigue resultados concretos dentro de un proceso determinado. No tiene un propósito tan general cómo la analítica descriptiva. La analítica predictiva pretende obtener un modelo válido de los datos que obtenga la mayor precisión posible en la predicción.

3. Cada vez más, la analítica predictiva, **tiende a la democratización para extender su uso más allá de los usuarios especialistas y científicos de datos**. Puesto que su foco es resolver preguntas de negocio, debe de ser accesible a usuarios de negocio (los cuales cada vez están más formados y preparados para aplicar técnicas de analítica predictiva). Así, los fabricantes de software y herramientas de analítica predictiva, vuelcan sus esfuerzos en el desarrollo de interfaces y asistentes (sin necesidad de codificación) que permitan al usuario de negocio, aplicar modelos avanzados de predicción.

Analítica predictiva. Un lío de palabras

Dado el creciente interés en la analítica predictiva en los últimos años, la comunidad de profesionales de los datos, genera de forma continua nuevos términos y conceptos para referirse a ella. Dado que es una disciplina científico-técnica, el inglés predomina cómo lenguaje preferido en el sector. De este modo, podemos encontrar en la bibliografía (libros, Internet, conferencias, etc.) referencias a analítica predictiva cómo: machine learning, deep learning, inteligencia artificial, statistical learning, analítica avanzada, modelización de datos y otros muchos términos diferentes.

Independientemente del término, existe cierto consenso entre la comunidad especializada en que lo importante es que este grupo de técnicas **persiguen el objetivo de *aprender de los datos*.**

Para aprender de los datos, se han desarrollado **modelos matemáticos** que son aplicables en determinadas situaciones y sobre algunos conjuntos de datos. Estos modelos nos permiten predecir una variable futura o clasificar (entender) la forma en la que los datos se organizan. A estos modelos matemáticos nos referimos tan sólo cómo modelos. Existe una clasificación relativamente estándar para estos modelos y la presentamos en el cuadro a continuación:

En la sección de Action de este informe realizamos nuestro caso práctico utilizando técnicas de regresión lineal sobre series temporales enmarcadas dentro del grupo de aprendizaje supervisado.

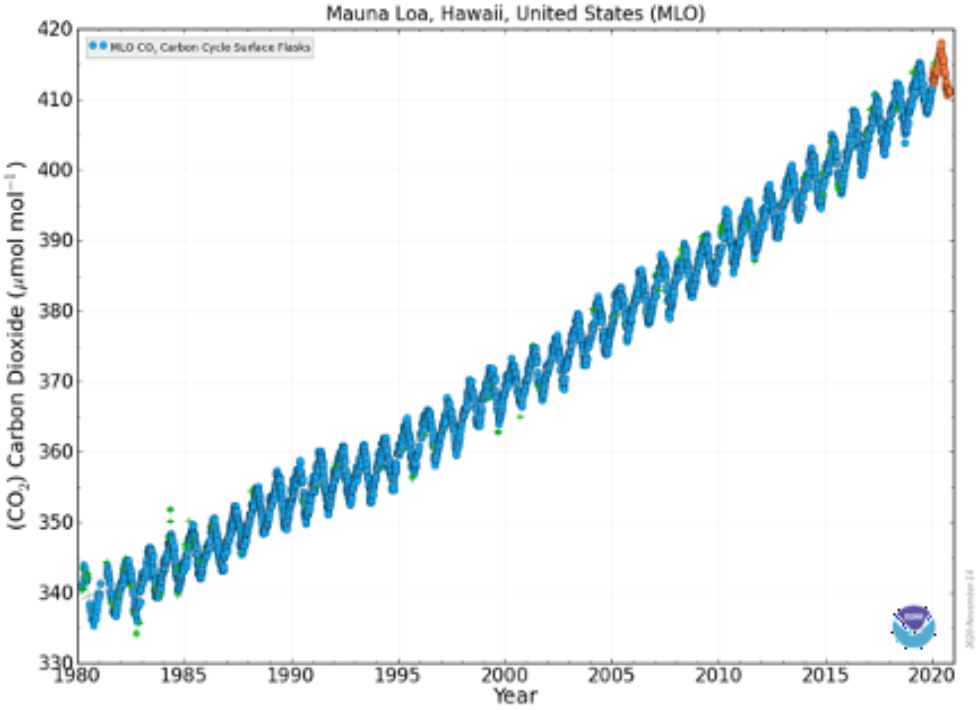
Un poco de historia

Para entender algunos de los hitos más importantes de la analítica predictiva podemos citar algunos ejemplos sobre cómo la analítica predictiva ha ayudado al ser humano y la sociedad

El protocolo de Kioto

El Protocolo de Kioto es un protocolo de la Convención Marco de las Naciones Unidas sobre el Cambio Climático (CMNUCC), y un acuerdo internacional que tiene por objetivo reducir las emisiones de seis gases de efecto invernadero (GEI). Este documento comprometió a los países industrializados signatarios a estabilizar las emisiones de GEI. Estructurado en función de los principios de la Convención, el protocolo establece metas vinculantes de reducción de las emisiones para 37 países y la Unión Europea (UE), reconociendo implícitamente que, en 1997 eran los principales responsables de los elevados niveles de emisiones de GEI en la atmósfera.
El éxito de la firma del Protocolo se basó, en parte, en el análisis de datos de la serie temporal de emisiones de gases y a los modelos de predicción que pronosticaban un incremento insostenible de los gases y el efecto invernadero derivado de estas emisiones.

Serie temporal (1980-2020) de la emisión de Gases de Efecto Invernadero (GEI). Estimación de crecimiento en rojo.



Fuente original NOAA

Las series temporales de datos sobre emisiones de GEI están disponibles cómo datos abiertos en la sección de cambio climático en el sitio web de las Naciones Unidas https://di.unfccc.int/time_series

Fuga de clientes

Cuando tus ingresos recurrentes se basan en contratos de suscripción mensual o anual, cada cliente que se va impacta en tu flujo de caja. Altas tasas de fidelización de clientes son vitales para la supervivencia de estos modelos de negocio cada vez más extendidos y populares. Por ejemplo, el negocio de un influencer que proporciona servicios a través de un canal de youtube, depende íntegramente del número de suscriptores de ese canal (y por lo tanto del número de reproducciones). Una definición más formal para la fuga de clientes se conoce en inglés cómo *customer attrition*. En los últimos años se ha popularizado un tipo particular de modelos de analítica predictiva cuyo objetivo es el de anticipar la fuga de clientes. Con el acceso a los resultados de estos modelos, los gestores del negocio pueden poner en marcha acciones y campañas para contrarrestar, en parte, la potencial fuga. Los modelos matemáticos de fuga de clientes analizan un periodo de tiempo con tasas conocidas de fuga, expresadas en %, y tratan de encontrar aquellas variables (factores) cómo edad, periodo temporal, comentarios en redes, reclamaciones de clientes, caducidad de medios de pago, etc. que más pesan a la hora de ajustar el modelo de fuga.

INSPIRE

Ya hemos entendido los diferentes tipos de analítica de datos que existen, y además, hemos analizado las principales características de la analítica predictiva de datos. Es momento, ahora, de desarrollar con más detalle algunos de los **casos de uso por sector** con más relevancia en la actualidad. En esta sección veremos, con cierta profundidad, tres casos de uso de aplicación de la analítica predictiva en tres sectores muy diferentes: industria, mercados (distribución y energía) y salud.

Industria: detección de anomalías y mantenimiento predictivo

La industria pesada se caracteriza típicamente por un uso intensivo de activos (todo tipo de máquinas y sistemas) necesarios para asegurar la producción o fabricación de materiales y productos. La industria - en general - depende intensivamente de estos activos para producir y generar beneficios. Los activos mecánicos, electrónicos y, actualmente, con gran peso, los sistemas digitales (servidores, sistemas de comunicaciones, etc.) son sistemas críticos cuyo mantenimiento es fundamental para garantizar la operación del negocio y la rentabilidad de éste. Así, la introducción de la analítica predictiva en este sector ha irrumpido de forma notable en los últimos años, especialmente en aquellas aplicaciones que sirvan para mejorar las operaciones de mantenimiento y el estado de salud de estos activos. Los casos de uso más representativos en la industria son los de **mantenimiento predictivo** (en inglés PdM, Predictive Maintenance) y la **detección de anomalías**. Para todos aquellos interesados en profundizar en el tema, existen incontables recursos en Internet sobre ambos casos. Algunos informes introductorios sobre el tema pueden encontrarse aquí y aquí. Pero, veamos un ejemplo práctico sobre estos casos de uso:

Mantenimiento predictivo: análisis de las vibraciones en equipos rotativos

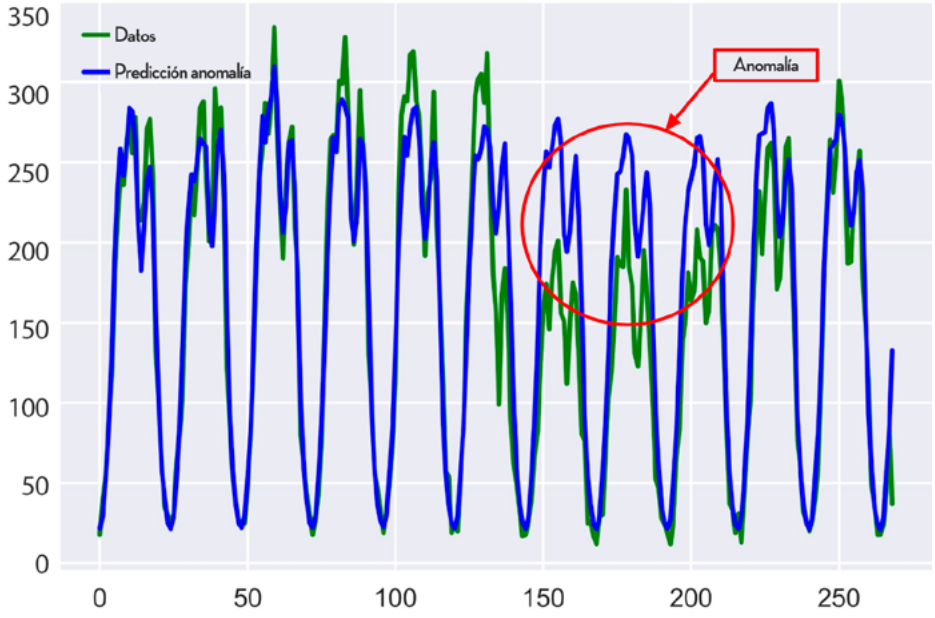
En la industria, existe una forma de clasificar los activos de producción en dinámicos y estáticos. De forma sencilla, los activos dinámicos son todos aquellos que se mueven cómo bombas, compresores, ventiladores,

cintas transportadoras, etc. Por el contrario, los activos estáticos son aquellos que no se mueven, pero que son igual de relevantes para la producción, cómo las tuberías, los tanques, hornos, etc. Los activos dinámicos se mueven fundamentalmente por el uso de motores. Los motores industriales, cómo cualquier otro motor (cómo el de nuestro coche o nuestra lavadora) sufren desgaste por su uso. En ocasiones, este desgaste se acentúa por diferentes factores cómo los ambientales, eléctricos, mecánicos, etc. Una forma efectiva de medir el desempeño de los motores es mediante la monitorización de las vibraciones que éstos producen. Medir las vibraciones en un motor es como hacer un electrocardiograma de nuestro corazón cuando estamos enfermos. El patrón de vibraciones de un motor aporta mucha información sobre lo que le ocurre. Cuando aplicamos **técnicas de monitorización en continuo** mediante el uso de **sensores** (Internet of Things) combinadas con **analítica predictiva** somos capaces de estimar indicadores útiles para predecir la vida útil hasta el próximo fallo o próximo mantenimiento programado. Los algoritmos optimizados para este tipo de analítica también son capaces de clasificar el tipo de fallo potencial (desalineamiento, problema en los rodamientos, etc.). Sin embargo, podemos ir un paso más allá y acoplar el análisis de vibraciones con sistemas predictivos que sean capaces de calcular la disponibilidad de stock de piezas de repuesto, el tiempo de suministro en caso de tener que ejecutar una compra. Últimamente, también existen sistemas que son capaces de comparar escenarios de compra de repuestos frente a fabricarlos por impresión 3D.

Detección de anomalías

En el campo de la analítica predictiva, la detección de anomalías cobra un papel fundamental. Especialmente en el espacio del análisis de las **series temporales**, donde las anomalías en la serie pueden significar **eventos de interés por sí mismas** o pueden, de lo contrario, **introducir ruido en la serie** y por tanto **disminuir la precisión de los modelos** de predicción de valores futuros de la serie. En el caso de anomalías en las series temporales de datos de carácter industrial (cómo monitorización de activos), las anomalías tienen sentido por sí mismas y pueden ayudar a **anticipar fallos en máquinas**.

Otros casos de uso donde la detección de anomalías juega un papel fundamental, son aquellos donde es necesario **identificar y clasificar producto defectuoso**. Por ejemplo, en el sector primario, la selección de productos cómo fruta y verdura, con el objetivo de seleccionar sólo aquellos productos de alta calidad que satisfagan los estándares para su comercialización es un caso de uso de detección de anomalías. En estas situaciones, normalmente un sistema de visión artificial (cámaras industriales sincronizadas con otros activos cómo cintas transportadoras y sistemas de aire comprimido) acoplado con algoritmos de detección de producto anómalo (una fruta en mal estado o por debajo de la talla esperada) son capaces de detectar y clasificar el producto y, en consecuencia, expulsarlo del proceso o apartarlo con el objetivo de rechazar o llevarlo por otro camino en el proceso de producción.



Mercados: predicción de precios y demanda

Un ejemplo ilustrativo de este grupo de casos de uso es aquel en el que un gran distribuidor, cómo una cadena de supermercados, intenta **predecir las ventas semanales de sus productos por departamento**. Podemos imaginar la gran ventaja competitiva que adquiere una compañía cuándo es capaz de predecir la demanda de sus productos y por lo tanto tomar las mejores decisiones para gestionar su cadena de suministro a todos los niveles.

La cadena de supermercados estadounidense Walmart es conocida por plantear desafíos recurrentes en la plataforma de científicos de datos Kaggle. Con el objetivo de predecir las ventas semanales de sus productos, la cadena de supermercados facilita datos históricos de ventas de varias tiendas localizadas en diferentes regiones de Estados Unidos. Cada tienda, tiene un número de departamentos. Los conjuntos de datos facilitados, contienen, además, información sobre la **temperatura media de la región**, la **tasa de desempleo**, el **calendario vacacional**, los **precios del combustible**, el **índice de precios al consumo**, etc. Para este tipo de casos de uso, se utiliza analítica predictiva, en particular técnicas de **deep learning**, que emplean **algoritmos de redes neuronales** para predecir las ventas por departamento en función de una serie de factores (también llamadas características o *features* en inglés).

Pero veamos un ejemplo diferente de predicción de la demanda. En este caso en un sector tan importante como el de la **energía**. Las compañías energéticas llevan años enfrentándose al problema de amortiguar los

picos de demanda de energía (en particular electricidad) con el objetivo de ajustar correctamente la generación de electricidad a la demanda. ¿Por qué? Muy sencillo, por naturaleza la electricidad no se puede almacenar en grandes cantidades. Los únicos grandes reservorios de energía son los recursos naturales. Por ejemplo, una presa para generar energía hidroeléctrica. En el sector esta práctica se denomina *energy demand-response (DR)*. El panorama de predicción de la demanda se ha complicado mucho en los últimos años gracias a lo que se denomina generación distribuida. Con la generalización de las energías renovables (autogeneración de energía fotovoltaica en hogares residenciales) y las nuevas maneras de consumir energía (puntos de recarga de vehículos eléctricos) la predicción de la demanda resulta mucho más compleja hoy. Las técnicas de analítica predictiva empleadas en los, cientos sino miles de, programas de DR a lo largo del mundo son muy numerosas. Desde simples métodos de regresión lineal hasta sofisticadas redes neuronales se emplean para modelar y predecir la demanda futura de energía por grupos de consumo.

Salud: diagnóstico aumentado de imagen médica

La idea de aplicar técnicas de analítica predictiva a conjuntos de datos de imágenes médicas es un área fascinante y en rápido crecimiento. Algunos de los ejemplos más conocidos es la fuerte apuesta por el sistema Watson de IBM para el análisis automatizado de imágenes de resonancia magnética nuclear (del inglés MRI). En imágenes médicas, el diagnóstico y/o la evaluación precisa de una enfermedad depende, tanto de la adquisición de imágenes como de la interpretación de imágenes. La adquisición de imágenes ha mejorado sustancialmente en los últimos años, con dispositivos que adquieren datos a velocidades más rápidas y mayor resolución. Sin embargo, el proceso de interpretación de imágenes tan solo hace unos años que se ha empezado a beneficiar de la evolución tecnológica en el campo de la analítica predictiva. La mayoría de las interpretaciones de imágenes médicas las realizan médicos. Es evidente que, la interpretación de imágenes por humanos es limitada por varios factores: la subjetividad del que interpreta, las grandes variaciones entre intérpretes, la fatiga y el cansancio en largas jornadas.

Cómo en otros muchos procesos de predicción de variables futuras, previamente a la aplicación de modelos de analítica avanzada, es necesario ejecutar procesos de limpieza de datos y detección de anomalías. Las herramientas de análisis de imágenes y aprendizaje automático son los habilitadores clave para mejorar el diagnóstico, facilitando la identificación de los hallazgos que requieren tratamiento y apoyando el flujo de trabajo del experto. En este caso, a diferencia de otras aplicaciones en las que conviven diferentes técnicas de analítica predictiva (más o menos sofisticadas), las técnicas de deep learning más modernas dominan fundamentalmente en el análisis de imagen médica.

ACTION

En esta nueva sección de ACTION ilustraremos algunos de los conceptos introducidos en este informe de manera práctica con un ejemplo real sobre un conjunto de datos abiertos disponible en datos.gob.es.

Introducción al conjunto de datos

En esta ocasión hemos escogido un conjunto de datos amplio y rico relacionado con el registro de accidentes de tráfico en la ciudad de Madrid. Tal y cómo dice la descripción oficial del conjunto de datos:

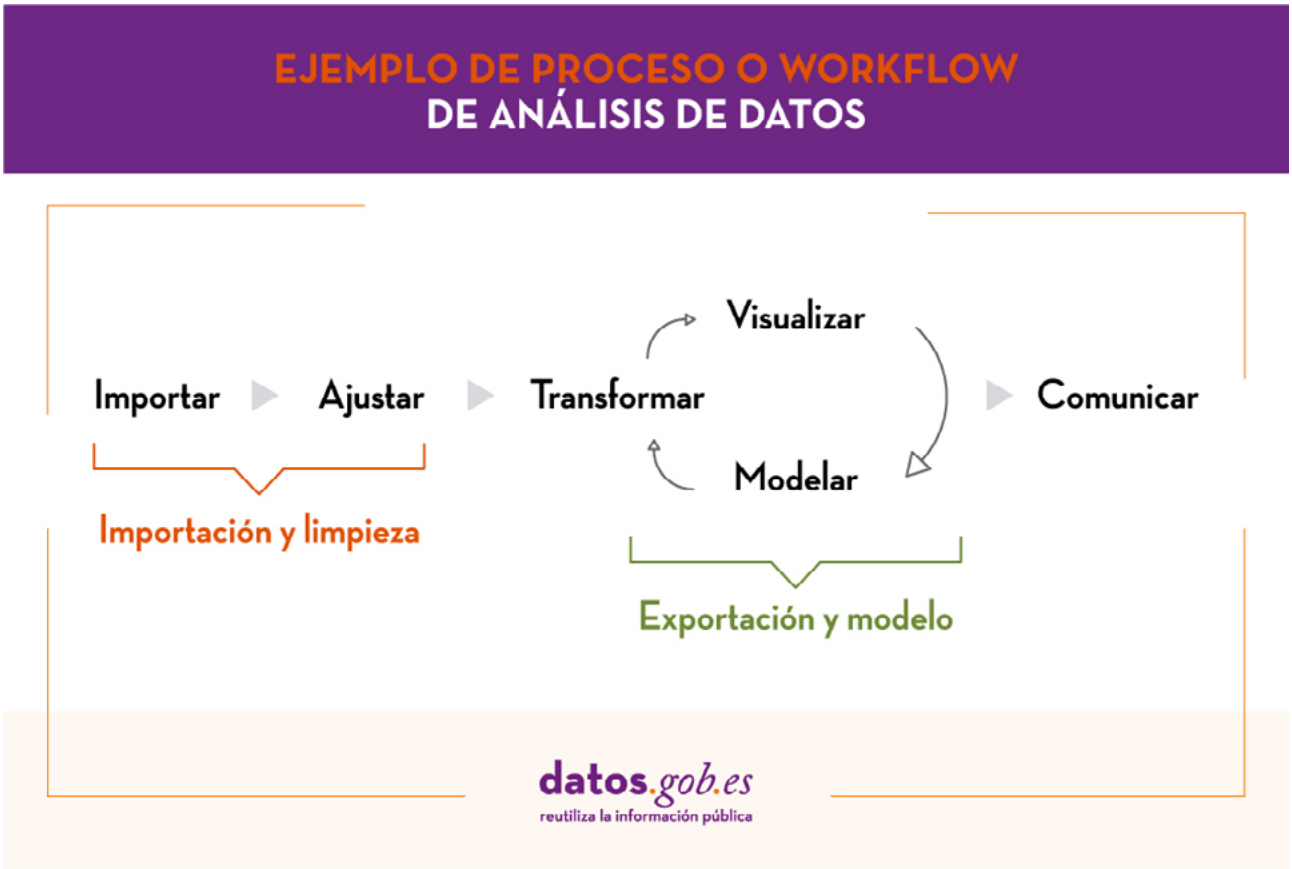
> *Accidentes de tráfico en la Ciudad de Madrid registrados por la Policía Municipal. IMPORTANTE: Se incluye un registro por persona implicada en el accidente. En el año 2019 y posteriores: La estructura de datos varía respecto a los años anteriores. El detalle de estas estructuras está disponible en el apartado 'Documentación Asociada'. No se incluyen registros de testigos.* ***Los ficheros de 2010 a 2018 solo registran los accidentes con heridos o con daños al patrimonio municipal****. Los datos publicados son provisionales hasta seis meses después del año vencido. Actualmente no se dispone de datos por barrio. Este portal también ofrece dos conjuntos de datos con información relacionada como son: Datos estadísticos de actuaciones de la Policía Municipal (incluye datos de atestados/partes de accidente por distrito y año) y Accidentes de tráfico con implicación de bicicletas.*

Cabe destacar, cómo explica la descripción del conjunto de datos, que existen inconsistencias en los datos antes y después del año 2019. Esto es una situación muy común en los casos reales de ciencia de datos en los que el científico o analista se enfrenta a situaciones en las que **los datos no tienen una estructura homogénea a lo largo de los años** (el número de columnas o campos descriptivos será diferente en número y tipo), y/o los datos están sucios (huecos en los datos, datos duplicados, datos mal formateados, etc.) en el sentido del análisis de datos. Así es el mundo real del análisis de datos, y lo cierto es que, de forma general, se dedica mucho más tiempo a la limpieza y preparación de los datos que al análisis en sí mismo.

Metodología seguida en el análisis

En este ejemplo práctico vamos a ilustrar las diferentes etapas del proceso de análisis, desde la **importación y la limpieza de los datos**, pasando por una fase de **analítica más exploratoria** para terminar con un ejemplo de **analítica predictiva**.

Una metodología muy aceptada en el sector de la analítica de datos es la propuesta por Hadley Wickham, científico jefe en RStudio, y que nos propone el siguiente flujo lógico:



Por sencillez para el lector no especialista, el código aquí mostrado no está diseñado para su eficiencia sino para su fácil entendimiento.

El objetivo de la sección no es explicar el código en detalle sino describir los principales bloques y su funcionalidad principal.

Durante el análisis, de forma indistinta iremos mostrando diferentes periodos temporales con el objetivo de resaltar características del conjunto de datos. Es importante tener en cuenta que, cómo se dice en la descripción del dataset, algunos periodos pueden no ser comparables pues la estructura de los datos varía con el tiempo. Cómo hemos explicado en anteriores ocasiones, el lector que desee reproducir este análisis podrá hacerlo utilizando el código fuente suministrado con este informe.

Para realizar este análisis, nosotros hemos utilizado código R y el entorno de programación RStudio. Nuestra configuración puede verse a continuación:

```
R version 4.0.3 (2020-10-10)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Catalina 10.15.7

Matrix products: default
BLAS:   /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.
framework/Versions/A/libBLAS.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
```

La forma de suministrar el código fuente es a través de un documento RMarkdown, de forma que una vez cargado en el entorno de programación, con tan solo ejecutar Knit en R el análisis se reproducirá automáticamente[1].

EDA - Analítica exploratoria de los datos

Pre-requisitos

Para desarrollar este análisis **necesitamos instalar una serie de paquetes de R adicionales** a la distribución base. Con el siguiente fragmento de código descargamos, instalamos y cargamos los paquetes necesarios:

```
#Installing dependencies
## First specify the packages of interest
packages = c("tidyverse", "dplyr",
             "ggplot2", "plotly", "readr",
             "lubridate", "tibbletime",
             "timetk", "modeltime",
             "tidymodels", "data.table")

## Now load or install&load all
package.check <- lapply(
  packages,
  FUN = function(x) {
    if (!require(x, character.only = TRUE)) {
      install.packages(x, dependencies = TRUE)
      library(x, character.only = TRUE)
    }
  }
)
```

Ahora fijamos el directorio de trabajo:

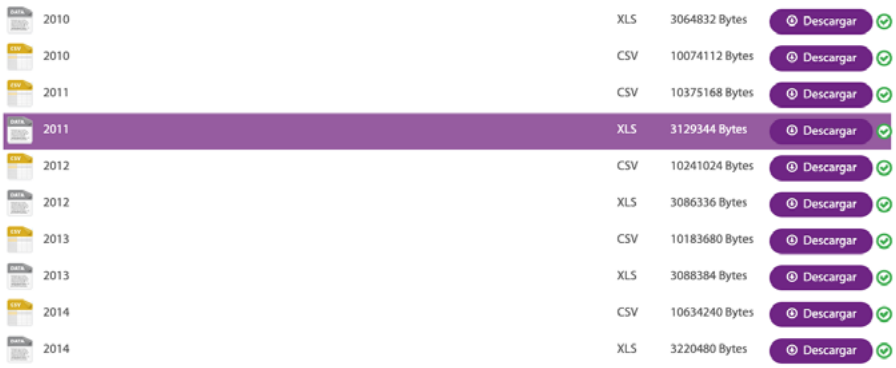
```
#Setting the working directory
## Important! Set here your own working directory ##
setwd("/Users/alija.alejandro/Documents/RProjects/REDES_Report_1_2020")
```

[1] Importante en la línea 42 del documento RMarkdown el usuario deberá de fijar su propio directorio de trabajo. En caso contrario el script generará el error correspondiente.

Los datos

Lo primero que vamos a ilustrar es la forma de importar los conjuntos de datos a nuestro análisis. En el portal de datos abiertos datos.gob.es, los datos de accidentalidad se encuentran divididos por ficheros (en csv o Excel) anuales. Así, si queremos realizar un análisis de varios años debemos descargarlos y unificarlos en un único conjunto de datos.

Distribuciones




```
#Following the pattern you can add more files for this analysis.

if (dir.exists(".files") == FALSE)
  dir.create("./files")

setwd("./files")

datasets <- c("https://datos.madrid.es/egob/catalogo/300228-12-accidentes-trafico-detalle.csv",
              "https://datos.madrid.es/egob/catalogo/300228-13-accidentes-trafico-detalle.csv",
              "https://datos.madrid.es/egob/catalogo/300228-14-accidentes-trafico-detalle.csv",
              "https://datos.madrid.es/egob/catalogo/300228-15-accidentes-trafico-detalle.csv",
              "https://datos.madrid.es/egob/catalogo/300228-16-accidentes-trafico-detalle.csv",
              "https://datos.madrid.es/egob/catalogo/300228-17-accidentes-trafico-detalle.csv",
              "https://datos.madrid.es/egob/catalogo/300228-18-accidentes-trafico-detalle.csv",
              "https://datos.madrid.es/egob/catalogo/300228-19-accidentes-trafico-detalle.csv",
              "https://datos.madrid.es/egob/catalogo/300228-21-accidentes-trafico-detalle.csv"
)

dt <- list()
for (i in 1:length(datasets)){
  files <- c("traffic2012",
             "traffic2013",
             "traffic2014",
             "traffic2015",
             "traffic2016",
             "traffic2017",
             "traffic2018",
             "traffic2019",
             "traffic2020")
  #Uncomment the following line if you want donwload the files (e.g if this is the first time
you execute the notebook)

  #download.file(datasets[i], files[i])
  filelist <- list.files(".")
  print(i)
  dt[i] <- lapply(filelist[i], read_delim, ";", escape_double = FALSE,
                  locale = locale(encoding = "WINDOWS-1252"),
                  trim_ws = TRUE)
}

traffic<-rbindlist(dt, use.names=TRUE, fill=TRUE)
traffic <- setDT(traffic)
```

Una vez cargado el conjunto de datos debemos de **re-ajustar algunos formatos** para poder realizar las agregaciones y gráficos necesarios.

```
#Formating the Date and some other data types

traffic$FECHA <- dmy(traffic$FECHA)
traffic$`TIPO ACCIDENTE` <- as.factor(traffic$`TIPO ACCIDENTE`)
traffic$`TIPO VEHÍCULO` <- as.factor(traffic$`TIPO VEHÍCULO`)
traffic$`TIPO PERSONA` <- as.factor(traffic$`TIPO VEHÍCULO`)
traffic$`ESTADO METEREOLÓGICO` <- as.factor(traffic$`ESTADO METEREOLÓGICO`)
traffic$`RANGO EDAD` <- as.factor(traffic$`RANGO EDAD`)
traffic$SEXO <- as.factor(traffic$SEXO)
traffic$SEXO <- toupper(traffic$SEXO)
```

¡Perfecto! en este punto tenemos **cargado nuestro conjunto de datos** listo para empezar a analizar. Veamos que tenemos cómo materia prima.

```
print(summary(traffic))
str(traffic)
```

La salida total de estos comandos se omite por sencillez en la lectura. Sin embargo, las **principales características del conjunto de datos son:**

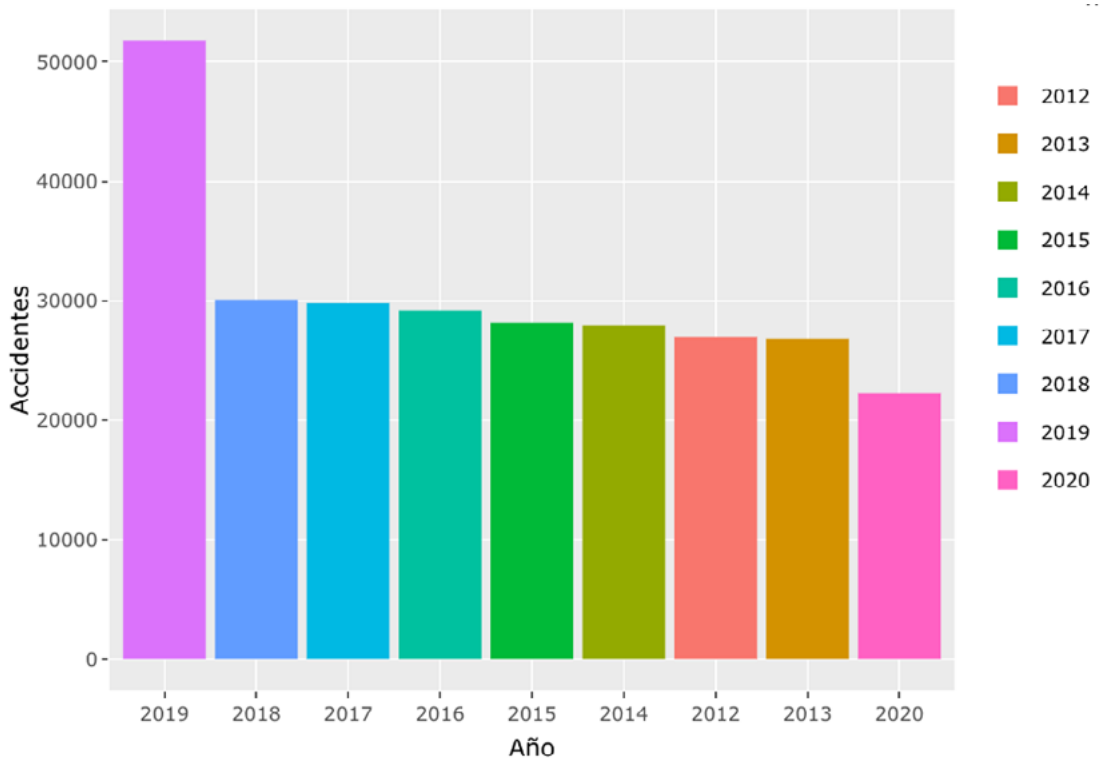
- El rango temporal abarca desde el 01-01-2012 hasta el 30-09-2020.
- El número total de filas del conjunto de datos son 273.147 observaciones.
- El número de columnas son 40 en total.
- La mayoría de las variables son de tipo categórico.
- No todas las variables disponibles se han tenido en cuenta en este análisis.

Veamos ahora algunos **resúmenes descriptivos del análisis.**

```
traffic$YEAR <- factor(year(traffic$FECHA))
trafficcount <- traffic[, .(count = .N), by= year(FECHA)]
trafficcount <- trafficcount[order(-count)]
ggplot(trafficcount[order(count)], aes(x=reorder(year, -count))) +
  geom_bar(aes(y=count, fill=as.factor(year)), stat = "identity") +
  xlab("Año") +
  ylab("Accidentes") -> baseplot

ggplotly(baseplot)
```

En la siguiente figura observamos el **ranking de accidentes por años**. Podemos hacer varias consideraciones al respecto:

- 2020 solo contiene datos hasta el mes de octubre. Independientemente de disponer de un menor histórico, la reducción drástica del número de accidentes es debido al confinamiento domiciliario derivados de la crisis del covid-19.
- En 2019 los datos son significativamente mayores que el resto de años debido al cambio de cuantificación desde 2019 en adelante. De 2010 a 2018 solo se registraron accidentes con heridos o con daños al patrimonio municipal.



Si nos preguntamos cómo se representa el análisis por distritos podemos generar esta **visualización.**

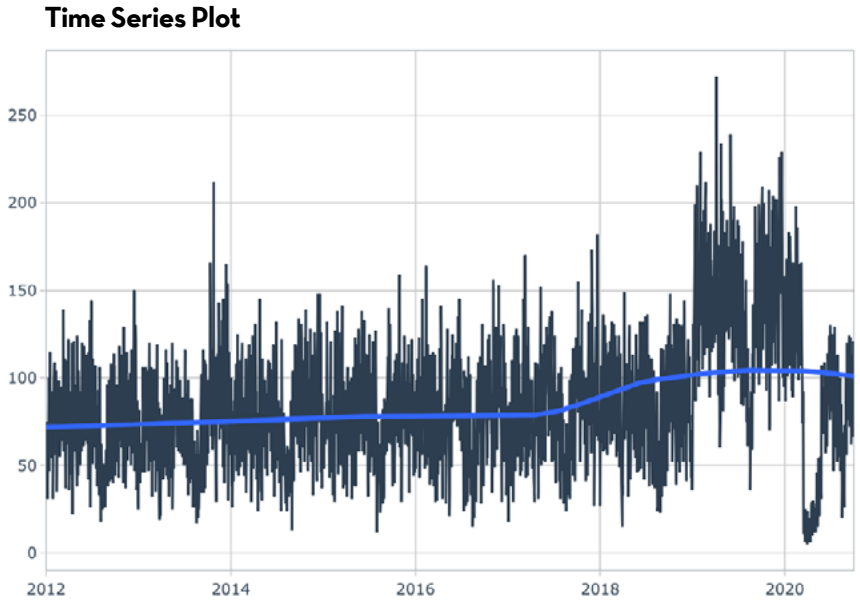
```
baseplot <- ggplot(na.omit(traffic, cols=c("SEXO", "DISTRITO")), aes(x=year(FECHA),
fill=SEXO)) +
  geom_bar(stat = "count") +
  xlab("Año") +
  ylab("Accidentes") +
  facet_wrap(~DISTRITO, ncol = 5, as.table = FALSE)

ggplotly(baseplot)
```

Además, hemos utilizado la categoría **"SEXO"** que representa a la persona que se ha visto involucrada en el accidente.

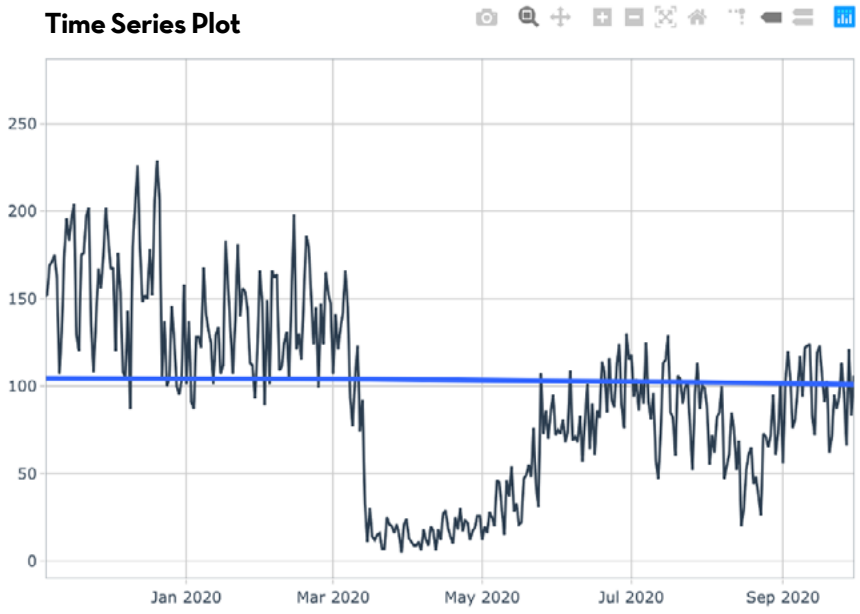
Observamos en esta figura el número de accidentes a lo largo de los años, clasificados por el distrito correspondiente y analizando el efecto del sexo del involucrado en el accidente. Como bien es sabido de todas las estadísticas facilitadas por las autoridades, los hombres cuentan con más siniestralidad de tráfico que las mujeres.

Finalmente vamos a representar la **serie temporal completa** para observar los efectos del confinamiento domiciliario impuesto por la crisis del Covid-19.

```
traffic_agg3 <- traffic[, .(count = .N),
                        by= c("FECHA")]
names(traffic_agg3) <- c("date", "value")
plot_time_series(traffic_agg3, date, value)
```



Cómo vemos, al principio de 2020 se aprecia una súbita bajada del número de accidentes. Dado que en el documento RMarkdown estamos generando visualizaciones interactivas, podemos hacer zoom sobre una determinada región para ver más detalles.



Hasta aquí la sección de **analítica descriptiva**. Evidentemente puede realizarse un análisis mucho más profundo dado que existen muchos más datos a disposición cómo el número de víctimas, el rango de edad de los afectados, el tipo de vehículo, las condiciones climatológicas, etc. Sin embargo, vamos a tratar ahora la sección de analítica predictiva.

Técnicas de analítica predictiva utilizadas

En esta sección de analítica predictiva vamos a utilizar **técnicas de análisis de series temporales para modelar y predecir el número de accidentes en los meses futuros** desde octubre de 2020.

No es el objetivo de este informe, ni de esta sección, explicar en detalle los modelos utilizados ni el código en detalle. Basta con decir que el flujo general que se va a seguir es el siguiente:

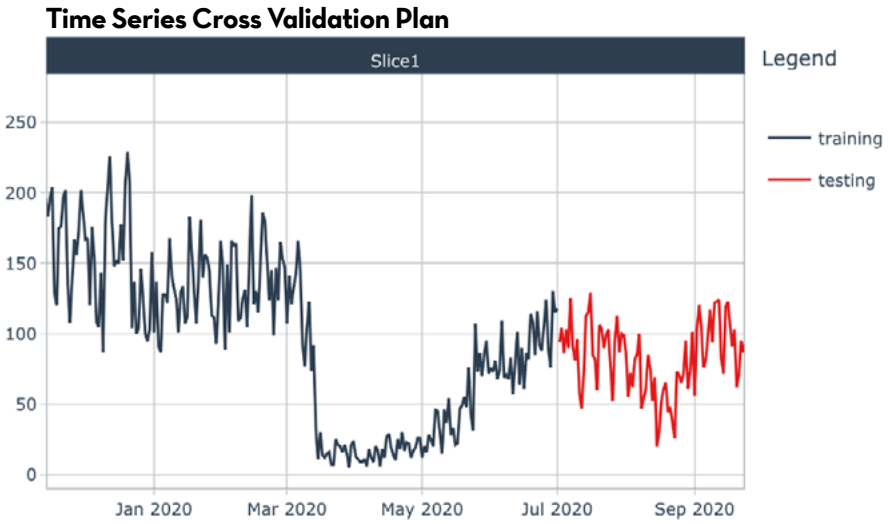
1. Partimos del conjunto total de datos y lo dividimos en una parte para entrenamiento y otra para test.
2. Aplicamos un modelo muy sencillo de regresión lineal.
3. Vamos a graficar el resultado del ajuste del modelo para ver de forma visual la calidad del ajuste.
4. Finalmente, pronosticamos los accidentes futuros en los meses de octubre, noviembre y diciembre de 2020.

Empecemos.

Paso 1. Lo primero que vamos a hacer es dividir el conjunto de datos:

```
splits <- traffic_agg3 %>%
  time_series_split(assess = "3 months", cumulative = TRUE)

splits %>%
  tk_time_series_cv_plan() %>%
  plot_time_series_cv_plan(date, value)
```

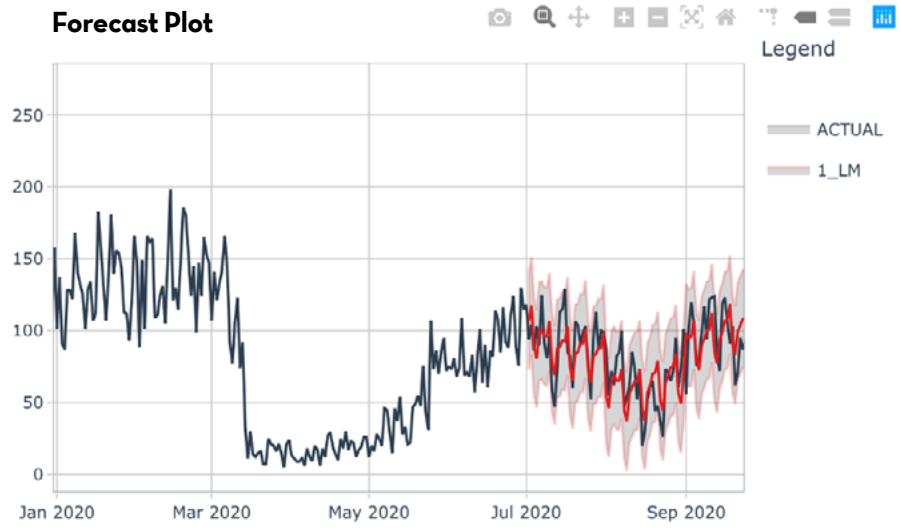
En este caso vamos a tomar 3 meses para el testing del modelo (línea roja) y el resto del conjunto para el *training* (línea azul).



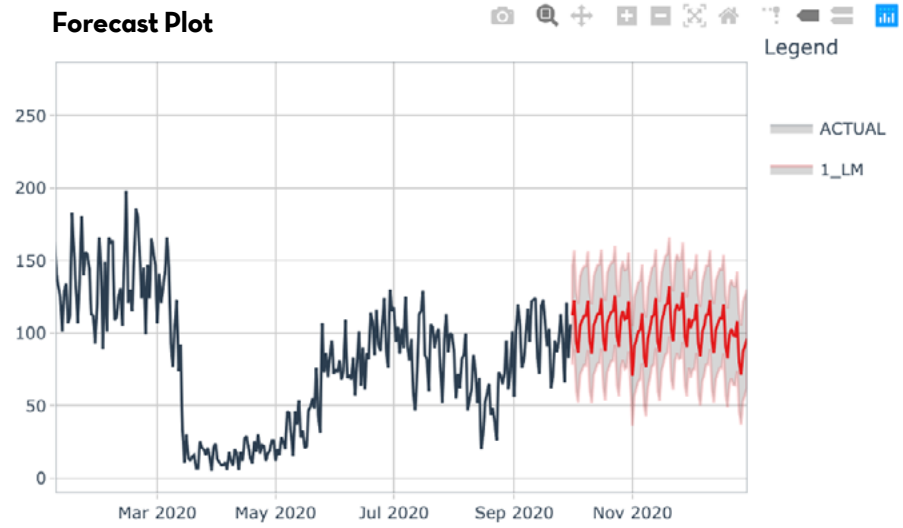
Paso 2. A partir de aquí, preparamos el modelo y ajustamos.

```
# Add time series signature
recipe_spec_timeseries <- recipe(value ~ ., data = training(splits)) %>%
  step_timeseries_signature(date)

bake(prep(recipe_spec_timeseries), new_data = training(splits))

recipe_spec_final <- recipe_spec_timeseries %>%
  step_fourier(date, period = 365, K = 5) %>%
  step_rm(date) %>%
  step_rm(contains("iso"), contains("minute"), contains("hour"),
          contains("am.pm"), contains("xts")) %>%
  step_normalize(contains("index.num"), date_year) %>%
  step_dummy(contains("lbl"), one_hot = TRUE)

juice(prep(recipe_spec_final))

model_spec_lm <- linear_reg(mode = "regression") %>%
  set_engine("lm")

workflow_lm <- workflow() %>%
  add_recipe(recipe_spec_final) %>%
  add_model(model_spec_lm)

workflow_lm

workflow_fit_lm <- workflow_lm %>% fit(data = training(splits))

model_table <- modeltime_table(workflow_fit_lm)

model_table

calibration_table <- model_table %>%
  modeltime_calibrate(testing(splits))

calibration_table

calibration_table %>%
  modeltime_forecast(actual_data = traffic_agg3) %>%
  plot_modeltime_forecast()
```

Paso 3. La línea roja, nuevamente, es el ajuste sobreimpresionado encima de los datos reales. Cómo vemos, la tendencia del ajuste reproduce visualmente bien la curva de accidentes reales.



Paso 4. Finalmente, podemos dar por bueno este modelo y tratar de predecir valores futuros en los tres meses siguientes (octubre, noviembre y diciembre) dentro de un intervalo de confianza:



CONCLUSIONES

La analítica predictiva juega un papel fundamental en todos los procesos de cualquier empresa u organización. Cualquier decisión actual que no se fundamente en el aprendizaje cuantitativo de los datos carece de rigor y fundamento. Las técnicas, herramientas y soluciones para poder aplicar analítica predictiva se han democratizado significativamente en los últimos años. Junto con el crecimiento exponencial de los volúmenes de datos generados por la sociedad - buena parte de ellos datos abiertos - las técnicas de analítica predictiva progresan a un ritmo increíble. En este informe hemos pretendido realizar una introducción asequible, para todos los públicos, sobre los conceptos generales de la analítica predictiva. Hemos repasado los conceptos claves de este dominio de los datos y explorado algunos de los casos de uso más representativos. Hemos finalizado con un caso de aplicación práctica sobre un rico conjunto de datos abiertos. Los lectores más avezados podrán reproducir este caso práctico e incluso intentar nuevas técnicas de análisis y diferentes modelos. Esperamos que os haya gustado este nuevo informe y seguiremos generando contenidos de interés alrededor del apasionante mundo de los datos.

Hasta la próxima.

PRÓXIMA PARADA...

Si no has tenido suficiente ;) y quieres seguir profundizando en el apasionante mundo de la analítica predictiva te sugiero los siguientes recursos que se mencionan a continuación. Cómo hemos explicado a lo largo de este informe, la analítica predictiva es un tipo avanzado de analítica que nos ayuda a tomar decisiones informadas apoyadas en los datos. La analítica predictiva se aplica a multitud de campos y se puede analizar desde múltiples puntos de vista, algunos más técnicos y otros más de negocio. Esperamos que los siguientes enlaces os ayuden a profundizar para convertiros en unos auténticos especialistas en analítica predictiva.

Libros

A continuación, os recomendamos algunos libros que constituyen auténticos pilares en analítica y en particular en analítica avanzada. En concreto, los siguientes libros podríamos considerarlos libros técnicos. Estos libros suelen incluir conjuntos de datos de prueba y ejemplos con código (R o Python) para ilustrar ejemplos de aplicación de modelos predictivos.

- R for Data Science
- The Elements of Statistical Learning
- Forecasting: Principles and Practice
- Learning Predictive Analytics with Phyton

Además, conviene destacar que existen libros con una orientación menos técnica. En este artículo se revisan algunos libros de analítica avanzada para "no-techies" según su autor.

- The 6 Best Data Science Books for Non-Techies

Tutoriales y cursos

Además de leer libros, probablemente la mejor forma de convertirte en un auténtico científico de datos sea practicando y practicando. A continuación te dejamos unos enlaces a tutoriales y cursos on-line con una importante carga de programación práctica en analítica avanzada.

- Data Science Courses for Business
- Time Series Machine Learning
- Introduction to Predictive Analytics using Python
- Predictive Analytics
- Predictive Analytics using Machine Learning
- SAS Statistical Business Analyst Professional Certificate

Algunos casos de uso muy interesantes

Tan importante es ser un experto en las técnicas de analítica predictiva cómo conocer sus aplicaciones en las organizaciones. Saber relacionar las técnicas de análisis con los principales casos de uso de aplicación es fundamental para extraer valor de la tecnología que está a nuestro alcance. Para seguir profundizando en casos de aplicación, además de los mencionados en este informe, os dejamos una selección de buenos artículos sobre el tema.

- How enterprises are using Predictive Analytics to transform historical data into future insights
- Repsol advances in digitalization and development of talent with training on data
- How to plan an optimal tour using Network Optimization in SAS Viya
- From one year to six weeks: Highmark Health teams with IBM to accelerate AI in urgent times
- Deep dive into various configurations with Oracle Weblogic Server- WLS Installation, plugin configurations, JMS, SSL,...

GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

datos.gob.es

Iniciativa aporta