Una señal de restaurante sobre la banqueta

Descripción generada automáticamente con confianza media

**MÁSTER EN DATA SCIENCE**

CAPSTONE: PLAN DE PROYECTO

Diseño de un modelo de *Machine Learning* para la predicción de pujas en entorno de Publicidad programática

Logotipo, nombre de la empresa

Descripción generada automáticamente

**Alumnos:**

**Javier Alejo Álvarez**

**Marco Duque García**

**Marta Pérez Romero**

**Mentora: Elena Abril Medina**

**Edición 2021-22 (Octubre/Abril)**

# INDICE

[PROPUESTA DE TÍTULO 3](#_Toc101105806)

[PALABRAS CLAVE 3](#_Toc101105807)

[RESUMEN DE LA PROPUESTA 3](#_Toc101105808)

[JUSTIFICACIÓN 4](#_Toc101105809)

[OBJETIVOS 5](#_Toc101105810)

[ESTRUCTURA DE LA MEMORIA 5](#_Toc101105811)

[METODOLOGÍA 6](#_Toc101105812)

[RIESGOS Y PLAN DE CONTINGENCIA 7](#_Toc101105813)

[a) Incidencias técnicas 7](#_Toc101105814)

[b) Incidencias personales 8](#_Toc101105815)

[ANÁLISIS DAFO 9](#_Toc101105816)

[PLANIFICACIÓN 9](#_Toc101105817)

[Calendario de trabajo 9](#_Toc101105818)

[Hitos 11](#_Toc101105819)

[Planificación de tareas 11](#_Toc101105820)

[MEDIOS Y MATERIALES 15](#_Toc101105821)

[BIBLIOGRAFÍA PROVISIONAL 16](#_Toc101105822)

# PROPUESTA DE TÍTULO

Como primer apartado del plan de proyecto del Capstone, se definirá su título respetando los requisitos: completo, claro, preciso y referido al tema principal.

De este modo, el título, que refleja el contenido del proyecto será:

“Diseño de un modelo de *Machine Learning* para la predicción de pujas en entorno de Publicidad programática”.

# PALABRAS CLAVE

Con las siguientes palabras clave es posible resumir el trabajo:

* Machine Learning.
* Modelos predictivos.
* Inteligencia artificial.
* Publicidad programática.
* Despliegue en AWS.

# RESUMEN DE LA PROPUESTA

Hoy en día, la Ciencia de Datos es algo que forma parte de nuestra vida cotidiana. Aspectos como Inteligencia Artificial, modelos de entrenamiento, creación de *chatbots*, etc. ya no resultan tan nuevos, pese a que todavía tienen por delante un enorme potencial de crecimiento, en cantidad y calidad.

El *data mining*, concepto que existe desde hace décadas, ingeniería de datos, Big Data, *Business Intelligence*, etc., son términos cada vez más comunes y alrededor de ellos hemos experimentado un cambio en el paradigma empresarial con perspectivas como DevOps o MLOps. Todo ello, fuertemente relacionado con el desarrollo en la nube, con *Amazon Web Services, Microsoft Azure o Google Cloud Platform* como abanderados.

El *Machine Learning* puede aplicarse a distintos campos y con muy diversos propósitos. En este capstone se hará uso de ello aplicado al mundo de la publicidad por internet.

En este punto entra en juego la segunda parte de la propuesta. Desde los inicios de la publicidad online, con el modelo de pago por clic (CPC) de Espotting, MIVA, Overture (comprada posteriormente por Yahoo!), Yahoo!, Google o Microsoft, las formas de publicidad por internet han ido aumentando, refinándose, segregando el mercado en distintas soluciones. Una de ellas es la publicidad programática. Como sabemos, se basa en la existencia de anunciantes y *publishers*, los primeros desean publicitar su producto o servicio, mientras que los *publishers* (que suelen elegir qué anuncios quieren, con qué precio y qué anunciantes aceptan o no) venden su espacio publicitario.

En este contexto, las empresas de publicidad online buscan la manera de optimizar sus sistemas para obtener la mayor rentabilidad. En concreto, se contará con la ayuda de la empresa Kimia (<https://kimiagroup.com/es/index.html>), que facilitará un *dataset* con datos reales (ninguno de ellos sensible ni personal), a partir del cual se podrá realizar la aplicación de *Machine Learning* a este entorno, que dé respuesta a una problemática real, llegando hasta la fase final de despliegue. La problemática consiste en que la empresa envía a su red de *publishers* los anuncios disponibles. Las redes o *publishers* valoran si les interesa entrar en una subasta y pujar por esos anuncios o no. Esta decisión depende de factores como el tipo de anuncio, su segmento, el precio del que parte, etc. La empresa Kimia tiene un análisis de qué redes o *publishers* son más rentables y adecuados para el negocio (otorgan tráfico de calidad). De este modo, lo ideal es que la empresa envíe los anuncios únicamente a aquellas redes que tengan más probabilidad de aceptar ese anuncio y den mayor rentabilidad, en lugar de enviarlo a todas las redes y encontrarse posteriormente con tráfico de mala calidad y con clics fraudulentos.

# JUSTIFICACIÓN

La justificación principal de este proyecto es, por un lado y tras una experiencia profesional de más de 15 años en negocios digitales y publicidad online, entender cuáles son las actuales tendencias en este campo. Por su parte, la propuesta se justifica también por el deseo de adquirir un conocimiento sustancial de los distintos modelos predictivos que existen en *Machine Learning* e Inteligencia Artificial. Con ello, se diseñará un modelo que, dado un *dataset* con datos reales de la industria, optimice la gestión de redes que son candidatas para pujar por un determinado contenido publicitario.

# OBJETIVOS

Partiendo de una motivación personal por la Inteligencia Artificial y, en concreto, por su aplicación a campos como la publicidad programática, se establecen los siguientes puntos como los auténticos objetivos marcados a nivel de grupo:

* Utilizar un *dataset* real, con datos verídicos.
* Ser capaz de construir las variables de entrada para el modelo, que serán vectoriales (cada elemento de esos vectores es una variable categórica que se deberá convertir en numérica con los diccionarios.
* Saber cómo balancear un *dataset*.
* Ejercitarse en las tareas de limpieza, exploración y visualización de los datos con Python y sus librerías.
* Abordar el estudio de posibles algoritmos y modelos, y seleccionar y aplicar el más adecuado.
* Saber evaluar y hacer test de un modelo para comprobar la validez de las predicciones.
* Aprender a crear y utilizar la máquina en AWS, y hacer el despliegue de la aplicación.

# ESTRUCTURA DE LA MEMORIA

A continuación, se ofrece un resumen de lo que se expondrá en cada uno de los capítulos de la memoria, aunque este aspecto se encuentre sujeto a posibles modificaciones:

* **Capítulo 1**: El primer bloque del proyecto se dedicará al Plan de Trabajo, a una introducción y presentación de este, incluyendo un resumen de la propuesta, el marco de trabajo, las contribuciones, la metodología que se va a seguir y una planificación temporal de las distintas tareas requeridas.
* **Capítulo 2**: En el segundo bloque se hablará del estado del arte tanto en *Machine Learning* como en publicidad online. Se repasarán los principales modelos y tendencias del aprendizaje automático, lenguajes que se pueden utilizar, etc. Por su parte, se realizará un estudio completo de las actuales formas de publicidad por internet, incluyendo la evolución del modelo CPC (Pago por Clic) a la publicidad programática.
* **Capítulo 3**: El tercer bloque incluye el trabajo sobre el *dataset*, es decir, definición de variables, limpieza, exploración, visualización, con diferentes librerías de Python. Investigar y seleccionar el modelo de *Machine Learning* más adecuado, junto a la fase de prueba.
* **Capítulo 4**: En el cuarto bloque se hará un estudio comparativo de las distintas soluciones de despliegue del proyecto, analizando ventajas e inconvenientes de utilizar *AWS, Azure* o *GCP*. Se procederá, al final de este análisis, a la implementación del modelo en la nube. Se finalizará con un análisis de los resultados.
* **Capítulo 5**: El último bloque cierra el capstone con las conclusiones extraídas de todo el trabajo realizado, incluyendo reflexiones personales y un detalle de futuras líneas de trabajo.
* **Anexos**: Se incluirá en esta sección todo el código comentado, detalles sobre la comparativa de modelos de *Machine Learning* que se pueden aplicar, y un presupuesto de puesta en marcha y despliegue del modelo en un entorno real.

# METODOLOGÍA

Como se puede extraer de todo lo comentado hasta la ahora, la metodología será una mezcla de teoría, que posteriormente se aplicará a la práctica.

En concreto, hay tres puntos de atención: modelos de *Machine Learning*, publicidad programática y despliegue en la nube.

De forma paralela, pero no simultánea, se estudiará el estado del arte de cada componente citado, y con ello, se tomarán decisiones sobre las soluciones que se adoptarán.

Tras este ejercicio teórico se trabajará en la aplicación práctica en los tres núcleos de acción:

* *Machine Learning*: recepción del *dataset*, análisis de este, aplicación de limpieza, exploración y visualización, aplicación del modelo elegido, entrenamiento y *testing*.
* Publicidad programática: sin entrar a fondo en ello, dado que se encuentra fuera del alcance del proyecto, se abordará el análisis de rentabilidad de las redes y, en conjunto, el modo de funcionamiento de redes-anuncios-subastas.
* Despliegue en la nube: decidida la plataforma en la que hacer el despliegue, y tras un estudio teórico práctico de su funcionamiento, se abrirá una cuenta de prueba y se seguirá el proceso adecuado para implementar el modelo y ponerlo en funcionamiento.

Como último paso, habrá que evaluar los resultados y explicarlos con claridad en la sección de conclusiones y próximas líneas de acción.

# RIESGOS Y PLAN DE CONTINGENCIA

Es de esperar que a lo largo del *capstone* puedan surgir eventualidades no previstas. Estas incidencias se pueden clasificar en dos grupos, y para cada una de ellas se añade su plan de contingencia:

#### Incidencias técnicas

* Problemas con los recursos de *hardware*: es posible que alguno de los equipos no responda como se espera. Por ello se cuenta con tres ordenadores independientes. En caso de que los tres se estropeen, se dispone de soluciones alternativas que incluyen el uso de terminales de familiares.
* Incompatibilidad de los recursos de *software*: se puede dar el caso de que no se consiga desarrollar el proyecto completo sobre una única plataforma por cuestiones de incompatibilidad. Es por ello por lo que se dispone de los tres sistemas operativos principales, Windows, OSX y Linux, de modo que todos los programas podrán funcionar en uno u otro.
* Pérdida de datos: ante la posibilidad de que se pierdan datos, algo no deseable, se hará una copia de seguridad diaria de la carpeta de trabajo, que se guardará tanto en una localización en la nube como en un disco duro portátil y en un dispositivo USB. Se garantiza la sincronización de los datos.
* Dificultades en la utilización del *software* especializado: al ser varias librerías será necesario superar una cierta curva de aprendizaje. Para ello se han previsto unas horas de lectura de documentación que ayudarán a avanzar más rápido cuando se trabaje con ellas.

#### Incidencias personales

* Simultaneidad con el trabajo: para evitar que ambos puedan interferir en el desarrollo del *capstone* se ha aprovechado el tiempo previo al inicio oficial de este para comenzar la planificación y lectura de bibliografía, y se han establecido sesiones diarias de trabajo de 3 horas, lo cual permite simultanear el proyecto con otras actividades.
* Dificultades en la realización: este proyecto supone un reto, puesto que gran cantidad de los contenidos son nuevos. Es posible que haya puntos a los que se deba dedicar más tiempo del previsto. Para que ello no suponga un problema existe un margen de un 10% de horas que se podrán utilizar para recuperar tiempo en caso de que sea necesario.
* Enfermedad o indisposición: este 10% extra de horas contemplado en el plan será también de utilidad en caso de que alguna enfermedad obligue a replantear los horarios.
* El plan de contingencia principal para evitar cualquier riesgo se basa en hacer un seguimiento continuado del plan de proyecto. A través de ese ejercicio se puede comprobar día a día el grado de cumplimiento y, de ser necesario, hacer algún tipo de restructuración o aplicar alguna medida correctora.

# ANÁLISIS DAFO

Como último ejercicio previo al desarrollo del proyecto y, más específicamente, a su planificación, conviene reflexionar sobre cuáles son los puntos fuertes y débiles, tanto internos como externos, lo cual se ha resumido en una tabla a modo de análisis DAFO:

|  |  |
| --- | --- |
| **ANALISIS INTERNO** | |
| **Fortalezas** | **Debilidades** |
| Interés y motivación plenos para aprender sobre *Machine Learning* y nuevas tendencias de publicidad online. | Desconocimiento de algunas librerías de *Machine Learning.* |
| Experiencia profesional previa en dirección de proyectos técnicos  Experiencia profesional previa y actual en publicidad online. | Conocimiento únicamente a nivel teórico de servicios en la nuble, como *Amazon Web Services*. |
| Perseverancia |  |
| Buen margen de maniobra para gestionar imprevistos |  |
| **ANALISIS EXTERNO** | |
| **Oportunidades** | **Amenazas** |
| Riqueza de recursos en Internet, tanto teóricos como prácticos | Riesgo de necesitar más tiempo del estimado |
| Disposición de la mentora para hacer seguimiento y servir de guía | Riesgo de enfermedad durante el proyecto |
| Posibilidad de hacer diversos modelos de ML antes de seleccionar uno. | Imposibilidad de prever fallos técnicos de los equipos |

# PLANIFICACIÓN

### 

### Calendario de trabajo

El *capstone* se desarrollará durante y después del *Máster en Data Science* (edición 2021-2022), cuyas fechas oficiales van desde el 23/02/2022 (inicio) hasta la fecha de presentación, estimada a efectos del desarrollo del trabajo en la segunda semana de junio (fecha sujeta a modificación una vez confirmada la oficial).

No obstante, puesto que durante el curso se está trabajando por cuenta ajena (y propia), se ha optado por comenzar la realización del proyecto unas semanas antes. De este modo, se dispone en realidad de aproximadamente tres meses y medio.

Para el correcto seguimiento del *capstone*, se ha dividido en cuatro bloques fundamentales: Bloque I (introducción, con la contextualización del proyecto y el plan de trabajo), Bloque II (estado del arte y trabajo con el *dataset*) y Bloque III (soluciones de despliegue, puesta en producción y pruebas funcionales), además de conclusiones y líneas futuras de investigación. Cada uno de los bloques se irá trabajando de acuerdo con un calendario.

Visualmente, este calendario se estructurará de la siguiente manera:

Una captura de pantalla de un celular

Descripción generada automáticamente con confianza media

Tabla

Descripción generada automáticamente

Por su parte, se ha estimado una carga de trabajo de unas 125 horas.

Teniendo en cuenta lo anterior, la disponibilidad diaria será de 2 horas, tal y como queda reflejado en la tabla inferior. El número de días se ha calculado desde el 21/03/22 hasta el 12/06/22. La suma total corresponde a la suma de los días por el número de horas.

No se trabajará en días de Semana Santa (marcados en rojo) ni los viernes y sábados en los que tengamos clases del máster.

Como se puede apreciar, las horas disponibles superan a las horas calculadas. Esto permite un margen de acción suficiente en caso de que otros motivos obliguen a modificar el calendario de dedicación.

|  |  |  |  |
| --- | --- | --- | --- |
| **Día Semana** | **Nº Días** | **Nº Horas** | **Horas totales** |
| **Lunes** | 11 | 2 | 22 |
| **Martes** | 11 | 2 | 22 |
| **Miércoles** | 11 | 2 | 22 |
| **Jueves** | 11 | 2 | 22 |
| **Viernes** | 8 | 2 | 16 |
| **Sábado** | 7 | 2 | 14 |
| **Domingo** | 9 | 2 | 18 |
|  |  | **Total** | **136 horas** |

### Hitos

Los hitos corresponden a las fechas clave de finalización de cada uno de los bloques y *feedback* correspondiente por parte de la mentora. Se añaden también como hitos la realización de la pre propuesta y la fecha de presentación ante el Tribunal:

|  |  |  |
| --- | --- | --- |
| **Grupo** | **Hito** | **Fecha** |
| Bloque I | Llamada inicial con mentora | 02/02/2022 |
|  | Aprobación del proyecto | 23/02/2022 |
|  | Plan de proyecto | 19/03/2022 |
| Bloque II | Obtención del dataset | 21/03/2022 |
|  | Validación del modelo | 28/04/2022 |
| Bloque III | Despliegue de la solución | 28/05/2022 |
| Presentación | Preparación del pptx | 01/06/2022 |
|  | Evaluación del Tribunal | 22/06/2022 |

### Planificación de tareas

Considerando los contenidos del proyecto, el calendario, las horas de trabajo disponibles y las fechas de entrega, se ha desglosado el desarrollo del trabajo en las tareas y actividades que pueden consultarse en las tablas siguientes. Se muestran las fechas de realización, y su duración estimada en horas.

Igualmente, se incluye a nivel global la carga de trabajo por cada una de las tareas, en función del tiempo requerido. Es conveniente indicar que se han distribuido las tareas de manera coherente, y que los porcentajes cuadran a nivel global con la temporización general y con las fechas de entrega.

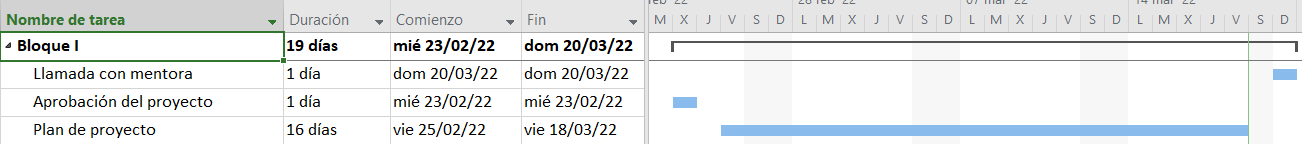
Por último, este plan estima unas 125 horas de trabajo, algo que es posible teniendo en cuenta que en el calendario de disponibilidad tenemos un total de 136 horas. En la sección de “

Incidencias personales” veremos que se han reservado deliberadamente para disponer de un margen de tiempo en caso de ser necesario.

En las próximas páginas podemos encontrar la lista completa de tareas en el calendario. Conviene destacar que seguiremos un modelo de gestión de proyectos en cascada, aunque determinadas tareas de diferentes bloques se solapen en el tiempo.

Tabla

Descripción generada automáticamente



Gráfico

Descripción generada automáticamente con confianza media

Gráfico

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

# MEDIOS Y MATERIALES

Para la elaboración del proyecto se empleará una serie de recursos de *hardware* y *software*, que en el momento de la redacción de este plan de trabajo se encuentran ya disponibles. La lista de aplicaciones está sujeta a revisión en caso de ser necesario.

Bloque de investigación:

* Bibliografía específica sobre *Machine Learning*, incluyendo manuales en soporte papel, artículos en revistas científicas y publicaciones en Internet.
* Realización de ejercicios prácticos con *Amazon Web Services*.

Hardware:

El recurso principal es un portátil donde están instalados todos los programas. También se ha habilitado un equipo de apoyo, que servirá de alternativa en caso de que el equipo principal sufra cualquier eventualidad:

* Equipo principal: portátil HP con procesador Intel Core i7 a 2,80GHz, 16Gb de memoria RAM y SDD de 500Gb. El sistema operativo principal es Windows 10, instalado en una de sus particiones y Ubuntu 16.04LTS a través de máquina virtual.
* Equipo de apoyo: portátil MacBook Air con procesador M1 a 2,5GHz, 8Gb de memoria RAM y SSD de 512Gb. El sistema operativo principal es macOS Big Sur.

Software y herramientas:

Los equipos disponen de los siguientes programas:

* Microsoft Project 2016: utilizado para la planificación del proyecto y su seguimiento.
* Suite de Office 365: para la redacción de documentos y elaboración de presentaciones.
* Capa gratuita de AWS.
* Repositorio grupal en Github.
* Anaconda: para implementar el entorno de Python, con Visual Studio Code y PyCharm..
* Diversas librerías de Python para las fases de limpieza, exploración, visualización, etc.
* Docker y Kubernetes para la implementación del modelo en Amazon Web Services.

# 2. ESTADO DEL ARTE

## Publicidad por internet

Dentro del sector de la publicidad, la compra programática hace referencia al modelo de compraventa de espacios publicitarios en tiempo real. Este modelo está adquiriendo gran relevancia año tras año.

La compra programática ha supuesto un gran avance en los métodos tradicionales de hacer publicidad, una revolución realmente. Por ejemplo, tenemos que los espacios publicitarios son adquiridos a través de un sistema de pujas en tiempo real gracias a los modernos desarrollos tecnológicos y a internet. La automatización de procesos garantiza una inversión más eficiente, ya que, a diferencia de la compra tradicional dónde se compra un paquete concreto de audiencia en un determinado soporte, en la compra programática se compran las impresiones una a una, por individuo. Esto permite una mayor individualización de la publicidad, impactando solo a usuarios que son más afines al producto o servicio ofertado.

La efectividad de la compra programática se basa en la agilidad en el proceso y la capacidad de impactar en individuos particulares que puedan estar mas interesados en el producto que se oferta, a diferencia de los métodos tradicionales que tienen un mensaje dirigido a audiencias en masa.

Basándonos en los resultados del estudio de la inversión publicitaria en España en 2022 (Infoadex 2022) la inversión en publicidad online o digital ocupa por tercer año consecutivo la primera posición por volumen de inversión dentro de los llamados medios controlados o convencionales con un porcentaje de la misma del 45,6% (y un 21,39% del total de la inversión publicitaria). Como podemos ver en el siguiente gráfico, salvo en 2020 que bajó (bajó en general, no solo para digital, siendo digital el que sufrió menos caída de inversión en porcentaje), la inversión en digital no deja de subir haciéndose más grande la diferencia con la inversión de publicidad en televisión.

|  |
| --- |
| Tabla  Descripción generada automáticamente |
| Tabla 1. Evolución de la inversión de España en publicidad de 2013 a 2017. Fuente: Estudio Infoadex de la inversión publicitaria en España 2022. |

La digitalización y los nuevos hábitos de consumos están haciendo que cada vez más empresas vuelvan sus ojos hacia la publicidad digital. Hace unos pocos años no muchas empresas apostaban por el marketing digital o una estrategia de posicionamiento online, pero el paso del tiempo y la pandemia han acelerado el proceso y echo dar el salto a las empresas.

Podemos decir que hay un cambio de paradigma. Las empresas están optando por un tipo de publicidad más eficiente, donde por ejemplo una campaña online te reporta datos muy precisos como el número de veces que se ha mostrado tu anuncio, cuantos usuarios han hecho clic en él, cual es el tipo de usuario que ha terminado comprando tu producto, etc. Datos estos que son reales y no estimados como en el caso de la publicidad tradicional.

Pequeñas y medianas empresas también se ven beneficiados del hecho que una publicidad programática y muy dirigida no implica un valor altísimo de inversión que sería inviable para ellos, como si sucede por ejemplo con los espacios publicitarios tradicionales, por ejemplo, la televisión. Es una publicidad mucho más económica.

Dentro de la publicidad online o digital, el RTB (*Real-time Bidding*) o sistema de subasta en tiempo real o de puja en tiempo real, es un método de compraventa de publicidad en línea basado en subastas instantáneas de impresiones de publicidad realizadas programáticamente. Esta es una vía muy prometedora dentro de la publicidad digital por el momento que vivimos, la era del Big Data. Basado en el análisis de cantidades masivas de datos de cookies generadas por los usuarios de internet, la publicidad RTB tiene el potencial de identificar en tiempo real las características e interés del público objetivo para cada impresión de anuncios, sirviendo automáticamente los anuncios más indicados para cada individuo y optimizando sus precios a través de un esquema de compra programática basado en subastas. El RTB ha cambiado significativamente la publicidad en línea, evolucionando desde el patrón tradicional de "compra de medios" y "compra de espacios publicitarios" a "compra de público objetivo", y se espera que sea el modelo comercial estándar para la publicidad online en el futuro.

Desde un punto de vista científico, la mayor demanda de automatización, integración y optimización en RTB abre nuevas oportunidades de investigación en campos como la recuperación de información (IR por sus siglas en ingles, Information Retrieval), la minería de datos (DM), el machine learning y la economía. Los investigadores de IR, por ejemplo, se enfrentan al desafío de definir la relevancia de las audiencias subyacentes dado un objetivo de campaña y, en consecuencia, desarrollar técnicas para encontrarlas y filtrarlas en el flujo de datos de solicitudes de ofertas en tiempo real. En la minería de datos, una tarea fundamental es identificar patrones repetidos sobre la transmisión de datos a gran escala de solicitudes de ofertas, ofertas ganadoras e impresiones de anuncios. Desde el punto de vista del machine learning, un problema emergente es como decirle a una máquina que reaccione a un flujo de datos, es decir, que aprenda como actuar sobre el sistema de forma inteligente en nombre de los anunciantes y las marcas para maximizar las conversiones y mantener los costes al mínimo.

El proceso comercial típico del RTB se ilustra en la siguiente figura:

|  |
| --- |
| Diagrama  Descripción generada automáticamente |
| Fuente: Yuan et al., 2012. |

Supongamos que un usuario de Internet está navegando en el sitio web de un editor:

A través del análisis de la cookie, la plataforma de gestión de datos (DMP) puede identificar los intereses y características del usuario. Cuando este usuario abre una página web, se activará una subasta una vez que ingrese la URL y presione enter: el editor enviará la información del usuario a la plataforma que gestiona el lado de la oferta (SSP), quien reenvía la información al Ad Exchange (ADX).

El AdX envía la información del usuario a las plataformas elegibles del lado de la demanda (DSP), decimos elegibles porque solo va a enviar la información del auction o subasta a las plataformas que tengan anunciantes interesados en publicitar su producto a un usuario de esas características. Esto se lleva a cabo cuando los DSPs preguntan al DMP y obtienen una respuesta, porque ejemplo que el usuario es fanático de los coches. Entonces, cada DSP envía la información del usuario a aquellos de sus anunciantes con un producto relacionado con los coches (vertical que los anunciantes ya han definido previamente en su configuración como de su interés) y comienza una subasta, donde los anunciantes que pueden presentar ofertas por la oportunidad de mostrar anuncios al usuario (es decir, la impresión del anuncio).

El ganador de cada subasta de DSP ingresará a la subasta de segunda ronda en ADX. El mejor postor de entre todos los DSP finalmente obtiene la impresión, y su anuncio se mostrará al usuario en las páginas web del editor. El proceso comercial, incluida la identificación de la audiencia, la subasta y la visualización de anuncios, terminará en un tiempo de entre 10 a 100 milisegundos, y por lo tanto se denomina "oferta en tiempo real".

Dos de los principales problemas a los que se enfrenta el RTB son las regularizaciones como consecuencia de la protección de datos y el control del fraude.

En RTB necesitamos la recopilación, acumulación y difusión de datos sobre los usuarios y sus actividades para operar el proceso de licitación, perfilar a los usuarios para "enriquecer" las solicitudes de licitación y operar funciones auxiliares como la detección de fraude. Como consecuencia, el RTB ha dado lugar a una serie de problemas de privacidad (Weinberg,Ggabriel, 2019 y Veale, Michael; Zuiderveen Borgesius, Frederik, 2022) y ha atraído la atención de las autoridades de protección de datos (DPA). Según la DPA del Reino las empresas involucradas en RTB "estaban recopilando e intercambiando información como raza, sexualidad, estado de salud o afiliación política" sin el consentimiento de los usuarios afectados. Simon McDougall de ICO informó, en junio de 2019, que "compartir los datos de las personas con cientos de empresas, sin evaluar y abordar adecuadamente el riesgo de hacer esto, genera dudas sobre la seguridad y la retención de estos datos".

En 2019, 12 ONG se quejaron de RTB ante una variedad de reguladores de la Unión (Herbrich, Tilman; Niekrenz, Elisabeth, 2021-10-01) lo que llevó a una decisión en febrero de 2022 en la que la Autoridad de Protección de Datos de Bélgica encontró una serie de ilegalidades en aspectos de un sistema utilizado para autorizar gran parte de RTB en la UE bajo el GDPR (General Data Protection Regulation). A raíz de todo esto se aboga desde muchos sectores por una regularización del RTB que permita ser autorizado por todos los organismos competentes en materia de protección de datos.

La mayoría de los fraudes de clics los llevan a cabo bots. En un artículo del New York Intelligencer en 2018 Max Reed aseguraba que "*menos del 60 por ciento del tráfico web es humano*".

El fraude publicitario digital afecta entre el 10 y el 60 por ciento de los diferentes tipos de publicidad digital, según numerosos estudios.

Debido a la naturaleza de bajo riesgo del fraude de clics impulsado por bots, los ciberdelincuentes y las bandas del crimen organizado se han involucrado más en el uso de configuraciones sofisticadas para lavar las ganancias del crimen o simplemente para desviar grandes sumas de dinero.

Según martechseries.com, las pérdidas globales por fraude publicitario en 2020 ascendieron a un total de 35 mil millones de dólares, mientras que un informe de *Campaign Asia* publicado en diciembre encontró que el fraude publicitario es un 20% del gasto mundial en publicidad online.

Y se espera que la cifra aumente. *Juniper Research* estima que la industria pierde aproximadamente $51 millones por día debido al fraude publicitario y que para 2023 esa cifra aumentará a $100 mil millones anuales.

¿Qué acciones se están tomando? La Autoridad de Mercados y Competencia del Reino Unido, en un estudio sobre plataformas en línea y publicidad digital, confirmó que "*el fraude publicitario se reconoce como un desafío de toda la industria y un problema que requiere soluciones de toda la industria para abordarlo*". Cada vez están surgiendo más empresas dedicadas a ofrecer en exclusiva soluciones antifraude, así como algunos DSPs y SSPs ofrecen también algunas herramientas para combatirlo. Las empresas que tecnológicamente tienen la capacidad técnica para llevar a cabo soluciones antifraude también las aplican a nivel interno. El problema es que al igual que el RTB está en constante evolución los tipos de fraude llevados a cabo también, de manera que cuando un tipo de fraude se ve eliminado del ecosistema por alguna regularización enseguida aparece otro capaz de burlarla. Hasta ahora el fraude ha sido un mal endémico del RTB.

En la publicidad basada en RTB, existen oportunidades de investigación fructíferas. En un ámbito más amplio, la publicidad RTB se ha convertido en un importante campo de batalla para los grandes datos. Dado que las transacciones publicitarias se agregan en los sitios web en tiempo real, la industria de la publicidad RTB tiene una oportunidad única de comprender el tráfico de Internet, los comportamientos de los usuarios y las transacciones en línea. El papel del machine learning y el tratamiento de datos en la industria de la publicidad programática tiene un papel central, encontrar algoritmos que ayuden a hacer más eficiente el sistema, así como para mejorar la detección del fraude será la clave para el siguiente salto cualitativo y cuantitativo del RTB. A medida que se propongan algoritmos más específicos, la optimización de la publicidad se volverá más parecida a la del comercio del mercado financiero y tiende a estar impulsada por las ganancias de marketing y el retorno de la inversión (ROI). Es decir, existe un objetivo de campaña explícito y medible de adquirir nuevos usuarios y obtener ventas de los usuarios adquiridos. Por lo tanto, uno de los próximos desafíos es ofertar adecuadamente por una impresión de anuncio para impulsar las ganancias y el ROI, todo esto a base de algoritmos.

En resumen, la publicidad mediante RTB es un campo de juego importante y desafiante para la investigación en campos como: la recuperación de información, la ciencia de datos, el aprendizaje automático y la economía.

* 1. **MACHINE LEARNING**

La habilidad del ser humano de poder interpretar información y extraer conclusiones a partir de ella lo es lo que nos aporta la mayor ventaja frente a los demás seres vivos del planeta. Cuando hablamos de la inteligencia artificial y el aprendizaje automático contemplamos el reto de diseñar sistemas capaces de replicar esta habilidad.

Las respuestas que actualmente existen ante este reto llevan madurando durante años y se han visto enriquecidas por los conocimientos de un abanico disciplinas distintas (1). Desde los avances computacionales y matemáticos hasta las más profundas reflexiones filosóficas y éticas, la inteligencia artificial sigue nutriéndose del conocimiento humano y no muestra indicios de parar.

Parejo, las soluciones actualmente disponibles ya juegan un papel importante en muchas de las interacciones que tenemos en nuestro día a día. En este apartado, vamos a explorar algoritmos y aplicaciones de uso de la actualidad. Entre ellos se diferencian los modelos supervisados de los no supervisados en función de si existe o no una variable objetivo a predecir por el modelo.

Dentro de los algoritmos **supervisado**, donde hay una variable a predecir cabe destacar la diferencia entre los algoritmos de regresión, donde la variable de salida es cuantitativa frente a los y los de clasificación donde la variable a predecir el cualitativa.

Empezando por **regresión lineal**. Se trata de uno de los algoritmos más sencillos pero que nos permite empezar a conceptualizar herramientas que servirán para modelos más complejos.

El caso más sencillo dentro de la regresión lineal es la **regresión lineal simple**. Este tiene como objetivo la predicción de una variable continua a partir de una única variable de entrada. A partir de un conjunto de observaciones de estas variables, el método identifica cual es la relación lineal que mejor se ajusta a los datos. Así, ante una observación nueva de la variable de entrada, se puede aplicar la relación obtenida para predecir el valor de la variable de salida.

La forma más común de realizar el ajuste del modelo es mediante el error cuadrático medio. Esta medida da una idea de cómo de separada está el valor de la variable de salida real frente al valor predicho por el modelo.

Empleando la misma función de error, y de nuevo con el objetivo de minimizarla, el modelo es ampliable a un número de variables de entrada indefinido.

Mientras que antes ya era importante preparar los datos para sacar el mayor provecho del modelo, ahora más aún.

Uno de los factores nuevos que entran en juego es la selección de las variables de entrada y la interpretación que hacemos del modelo sobre como esas variables afectan a nuestra variable de salida (2). Por ejemplo, si seleccionásemos variables de entrada que no estén relacionadas con nuestra variable de salida estaríamos introduciendo ruido al modelo. Otra situación desfavorable se puede dar si incluimos variables que estén fuertemente correlacionadas entre sí.

Incluir variables de entrada altamente correlacionadas puede desestabilizar nuestro modelo, incitar que el modelo memorice (*overfitting*) y dificultar la interpretabilidad de este. Este será un reto al que nos enfrentaremos en muchos otros modelos. Una de las herramientas más potentes para afrontarlo es la regularización.

La regularización consiste en modificar ligeramente la función de error mediante la cual aprende nuestro modelo de tal forma que penalice la cantidad de variables que considera (L1) y el peso de estas (L2) (3). Hasta este punto, el modelo que minimizaba el error cuadrático medio se podía determinar mediante algebra lineal, pero al definir una función de coste (error del modelo) más compleja, se genera un reto añadido. Para minimizar la función de coste se emplean algoritmos de descenso de gradiente.

Los algoritmos de minimización por descenso de gradiente inicializan los parámetros del modelo en un punto arbitrario. A partir de este estado, hacen uso del gradiente de la función de coste para determinar que modificación de los parámetros causa el mayor descenso de la función de coste. Mediante una serie de etapas, el algoritmo tiende a minimizar el valor de dicha función. El descenso de gradiente debe ser ajustado para optimizar su comportamiento. Para ello, se cuenta principalmente con tres parámetros:

La tasa de aprendizaje: Es el factor multiplicativo del gradiente. Determina como de bruscas o sutiles son las etapas del algoritmo.

La tolerancia: Es la diferencia mínima de la función entre una etapa y la siguiente antes de que el algoritmo concluya que se ha alcanzado un mínimo.

El lote: Para casos de uso con muchas observaciones el algoritmo resulta muy costoso computacionalmente. Para ello el algoritmo puede realizar sub-etapas de descenso en las que en vez de tener en cuenta como las modificaciones en el modelo afectan a la función de coste para todas las observaciones, se tiene en cuenta un lote de ella a la vez.

La **regresión logística** adapta la base de la regresión lineal para predecir variables de salida es categóricas. Esto se logra mediante una utilidad nueva, la función de activación. La función de activación transforma la variable de salida, del intervalo de los números reales, a un intervalo (0, 1). El resultado de esta función se puede interpretar como la certeza que tiene el modelo de que la muestra pertenezca a la categoría 1. A su vez, nos vemos forzados a adaptar la función de coste del modelo. Una de las más populares es la función log loss. Se define de tal forma que favorezca las predicciones extremas (cercanas a 0 o 1) acertadas.

Además de la regresión logística, existe otra gran variedad de modelos de clasificación. El algoritmo de **Naïve-Bayes** es uno de ellos. Se basa en el teorema del que toma nombre y puede aportar buenos resultados cuando las hipótesis en las se sustenta se cumplen. El algoritmo **K-Nearest Neighbor** es otro ejemplo de algoritmo de clasificación. En contraste con el de bayes, este algoritmo tiene una interpretación fuertemente espacial. Como su nombre sugiere, predice la categoría de una muestra mediante una función de cercanía de la muestra a las observaciones de aprendizaje. Mientras que hasta aquí había sido conveniente la estandarización de las variables de entrada, para este modelo puede resultar indispensable.

**Los árboles de decisión** son capaces de incorporar clasificación y a posteriori regresión. La forma en la que funcionan es ramificando las entradas en sub-arboles. Estos sub arboles son determinados mediante el “feature” que mejor es capaz de hacer particiones de los datos. Para ello se usa la entropía de la partición. Es importante limitar el número de particiones explícitamente o poniéndole un límite a la función de partición para evitar que el modelo memorice en su totalidad los datos de aprendizaje. Cuando se da una de esas situaciones, se alcanza una hoja del árbol en la que se asigna una categoría determinada a los datos que hayan seguido esa ramificación. Se puede aplicar un modelo de regresión lineal a cada una de las hojas del árbol para expandir la idea de árbol de clasificación a árbol de regresión.

Para poder exprimir el potencial de estos algoritmos, una de las técnicas comúnmente aplicadas es Random Forest. Basándose en los árboles de decisión, de forma similar al concepto de lote en los algoritmos de descenso por gradiente, el árbol realiza cada partición (descenso), con una muestra aleatoria de datos de entrenamiento en vez de con la totalidad de los datos. En combinación con esta técnica, se puede generar una serie de árboles y categorizar la muestra en función de la categoría más frecuente entre todos los árboles.

Hasta este punto las técnicas que hemos visto tenían como objetivo predecir una variable objetivo. A continuación, comentaremos algunas técnicas dentro de las llamadas **no supervisadas**.

**La reducción de dimensionalidad** es una técnica no supervisada con un potencial enorme. Permite transformar las variables de entrada del modelo a un conjunto de variables de menor dimensión. De esta forma permite combatir los problemas generados por datos de grandes dimensiones.

Tomando como base el algebra lineal, los métodos de análisis de componentes principales reducen las dimensiones de la transformación lineal que presentan los datos a una transformación lineal definida por los auto vectores con mayores auto valores. Para decidir el número de dimensiones ideal para cada caso, nos podemos apoyar en la perdida de información sufrida debida a la reducción de dimensionalidad.

Mientras que este método es adecuado para problemas donde las variables de entrada presentan una relación lineal con la variable de salida, pierde aplicabilidad para casos donde no es así. Para estos casos se emplean métodos de *manifold learning* como T-SNE y UMAP. Estos métodos se basan en la teoría de que los datos reales yacen en una variedad (4) de dimensión reducida con respecto a los datos reales.

**Los modelos de clusterización** pretenden tomar los datos y agruparlos en un número determinado de grupos. Para ello también existe una amplitud de métodos cada uno con su aplicabilidad.

Algunas de las ideas en las que se basan estos modelos coinciden con nociones empleadas en la categorización de los modelos supervisados. Los métodos de K-means y K-mediods son solo dos de los muchos basados en medidas de distancia. Los métodos jerárquicos también muestran un fuerte paralelismo con los arboles de decisión supervisados. De mayor coste computacional, pero a cambio de mayor potencial y versatilidad están los modelos basados en densidad.

Para determinar la calidad de nuestra clusterización, se emplean muchos métodos. El elbow method por ejemplo se fija en la forma en la que añadir un nuevo cluster modifica la suma de las distancias de cada uno de los puntos a su correspondiente cluster. Por otro lado, para poder ganar perspectiva sobre cómo de separados están los clusters entre sí, tiene gran utilidad el avarage silouhete method.

Estos métodos tienen gran aplicabilidad para la detección de datos anómalos. Aquellos datos que antes tienden a diferenciarse del resto y formar sus propios cluster conviene examinar de forma individual.

Dentro de los métodos no supervisados también tienen gran aplicabilidad **los modelos de asociación**. Estos modelos tienen como objetivo identificar sucesos que ocurren de forma conjunta. La interpretación de esta información puede ser muy útil para tratar de provocar y predecir eventos.

La fortaleza de cada asociación se determina teniendo en cuenta el soporte (probabilidad observada) de un suceso y aquel de los sucesos que lo componen para generar indicadores, la confianza y el lift. Es importante tener los dos en cuenta a la hora de generar conclusiones.

Al tener que realizar estos cálculos para cada una de las combinaciones posibles de sucesos estos modelos pueden llegar a ser computacionalmente muy costosos. Para mitigar este inconveniente, se pueden aplicar varias técnicas. El pruning consiste en excluir el cálculo de todos aquellos indicadores asociados a sucesos con soporte pequeño ya que la intersección de dichos sucesos con otros a lo máximo tendrá dicho soporte.

Finalmente hablaremos los modelos que más cercanamente replican el funcionamiento del cerebro humano, **Deep Learning**. Con una evolución que origina en 1946 (5), los modelos de Deep Learning son aplicables de forma supervisada y no supervisada. A partir de una abstracción del funcionamiento de las neuronas los modelos Deep son capaces de combinar un gran número de ellas formado redes complejas.

Las redes funcionan por capas de tal forma que las entradas son interpretadas por una o más capas que a su vez transmiten una interpretación de los datos a otras capas. Al final de la cadena de capas se define una capa cuya excitación es interpretable por nosotros.

En la construcción de estos modelos se emplean muchos conceptos ya comentados para otros modelos entre ellos la normalización, la función de activación, la función de coste, el descenso por gradiente o la regularización. Se emplean además herramientas avanzadas para ajustar las capas para cada caso de uso.

En efecto, todos los modelos a nuestra disposición requieren ser optimizados e interpretados para aprovecharlos al máximo. A pesar de que la inteligencia artificial es una disciplina con muchos años de desarrollo, todavía queda mucho por explorar y avanzar.

# CREACIÓN REPOSITORIO GITHUB

# OBTENCIÓN DEL DATASET

# DEFINICIÓN DE LA VARIABLE

# LIMPIEZA DEL DATASET

# ANÁLISIS EXPLORATORIO

# ANÁLISIS VISUAL

# PRESELECCIÓN DE MODELOS

# ANÁLISIS DE MODELOS

# DEFINICIÓN DEL MODELO FINAL

# ENTRENAMIENTO Y TEST

# VALIDACIÓN

# IMPLEMENTACIÓN EN AMAZON WEB SERVICES (AWS)

## Estudio de solución AWS

## Provisión de medios AWS

## Pruebas previas

## Implementación: Docker y Kubernetes

# BIBLIOGRAFÍA PROVISIONAL

* Rob J Hyndman and George Athanasopoulos - Forecasting: principles  
  and practice (2º edition)
* George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M.  
  Ljung - Time Series Analysis. Forecasting and Control (5th edition)
* Jerome H. Friedman, Rober Tibshirani and Trevor Hastie, The  
  Elements of Statistical Learning.
* Chistopher Bishop, Pattern Recognition and Machine Learning.
* Aurélien Géron, O’Reilly - Hands-On Machine Learning with  
  Scikit-Learn, Keras and TensorFlow: Concepts, Tools, and Techniques  
  to Build Intelligent Systems
* Andreas C. Müller & Ssarah Guido, O’Reilly - Introduction to Machine  
  Learning with Python
* Alice Zheng & Amanda Casari, O’Reilly - Feature Engineering for  
  Machine Learning
* Peter Bruce, Andrew Bruce & Peter Gedeck - O’Reilly - Practical  
  Statistic for Data Scientist (Second Edition)
* Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong -  
  Mathematics for Machine Learning
* Jason Brownlee - Basics of Linear Algebra for Machine Learning
* Charu C. Aggarwal - Linear Algebra and Optimization for Machine  
  Learning: A Textbook
* Jay Dawani - Hands-On Mathematics for Deep Learning: Build a solid  
  mathematical foundation for training efficient deep neural networks
* Essentials of Business Analytics. An Introduction to the   
  Methodology and its Applications. Bhimasankaram Pochiraju
* Mathematics for machine learning. Marc Peter Deisenroth
* Probability and Statistics for Computer Science. Springer. David   
  Forsyth
* Data Mining Concepts and Techniques. Third Edition. Jiawei   
  Han.
* Improved Density Based Spatial Clustering of Applications of Noise   
  Clustering Algorithm for Knowledge Discovery in Spatial Data.
* Recommender Systems – An Introduction. Cambridge. Dietmar Jannach
* Aprende Machine Learning en español. Teoría + Práctica Python.   
  Juan Ignacio Bagnato