CURSO BIG DATA

DIA 1 y 2 – Tecnología I: Transformación digital

* Búsquedas avanzadas en Google: uso de comandos INFO, SITE, OR, AND, “”, INTEXT, INTITLE
* <https://www.xataka.com/basics/25-codigos-funciones-trucos-para-buscar-google-exprimiendo-al-maximo-su-motor-busqueda>
* (AUSENCIA POR ENFERMEDAD)

DIA 3 – BUSINESS INTELLIGENCE

* Por el momento NO se está utilizando al 100%. No se tienen bien definidos los perfiles necesarios. Las empresas todavía lo están implementando.
* Modelos de datos
* Data Warehouse: almacenaje de datos a lo “bruto”
* Data Mart: almacenaje de datos más específicos
* Bases de datos transaccionales (las habituales) Rápido de procesar, pero lento para analizar estructura OTPL
* Estructura OLAP (también denominada *Cubo OLAP*): análisis multidimensional de datos de forma veloz e interactiva. Formatos Estrella y Copo de Nieve
* ETL: extracción (datos a lo “bruto”), transformación y carga.
* Existen procesos ELT: “más baratos” a priori.
* Herramientas BI: Power BI (líder), Tableau (2ª), Qlik. Google ha comprado Looker y lo ha integrado (Google Cloud Platform) crecimiento potencial.
* Quejas por problemas y fallos en Looker (lo hemos sufrido en prácticas).
* La *minería de datos* es el proceso final para interpretar y evaluar.
* Clustering

EJERCICIO

**E.2 Descarga el dataset de Kaggle ‘Netflix Movies and TV Shows’. Cárgalo en Data Studio e intenta responder a las preguntas que plantea.**

* Utilizar archivos CSV (utilizar previamente una hoja de cálculo de google para unificar formatos).
* Fuente de datos: <https://www.kaggle.com/> Buscar datos con buena valoración. Utilizaremos para la práctica la base de datos de Netflix (<https://www.kaggle.com/datasets/shivamb/netflix-shows>)
* Conectar con <https://lookerstudio.google.com/datasources>
* Crear un informe genérico en Looker Studio
* Trastear añadiendo gráficos y controles

**Conexión entre tablas (Google Studio)**

* Ver video
* Se pueden conectar hasta 5
* Revisar conceptos SQL: LEFT JOIN, RIGHT JOIN, INNER JOIN (intersecciones) <https://programacionymas.com/blog/como-funciona-inner-left-right-full-join>

**Ejercicio: Big Data en Restauración (corrección al día siguiente).**

* Punto de partida
* Objetivos
* Recopilación de datos
* Análisis de datos
* KPIs
* Buyer Persona
* Estrategias de Marketing
* Soluciones

DIA 4 - Tecnología II: Manipulación de datos.

* Formatos de datos CSV, XML, JSON, etc.
* Fundamentos de programación.
* Expresiones regulares: <https://regexr.com/> <https://regex101.com/>
* *Javascript*: conocer para manipular datos, acceder y recogerlos
* <https://www.json.org/json-es.html>
* <https://jsfiddle.net/>
* Ejercicio *json* “juego de tronos”
* APIs: capa intermedia entre datos (BD) y las aplicaciones web finales (cliente).

EJERCICIOS

SQL – ejercicios

* JOIN [https://www.tutorialesprogramacionya.com/mysqlya/temarios/descripcion.php?cod=58 HYPERLINK "https://www.tutorialesprogramacionya.com/mysqlya/temarios/descripcion.php?cod=58&punto=64&inicio"& HYPERLINK "https://www.tutorialesprogramacionya.com/mysqlya/temarios/descripcion.php?cod=58&punto=64&inicio"punto=64 HYPERLINK "https://www.tutorialesprogramacionya.com/mysqlya/temarios/descripcion.php?cod=58&punto=64&inicio"& HYPERLINK "https://www.tutorialesprogramacionya.com/mysqlya/temarios/descripcion.php?cod=58&punto=64&inicio"inicio](https://www.tutorialesprogramacionya.com/mysqlya/temarios/descripcion.php?cod=58&punto=64&inicio)=

Looker Studio. Ejercicios

* Funciones <https://support.google.com/looker-studio/table/6379764?hl=es>
* Expresiones regulares [https://support.google.com/looker-studio/answer/10496674?hl=es HYPERLINK "https://support.google.com/looker-studio/answer/10496674?hl=es&ref\_topic=7570421"& HYPERLINK "https://support.google.com/looker-studio/answer/10496674?hl=es&ref\_topic=7570421"ref\_topic=7570421](https://support.google.com/looker-studio/answer/10496674?hl=es&ref_topic=7570421)
* Ejercicio: regiones del mundo
* Añadir campo editor de fórmulas y fórmula CASE
* Ejercicio: nuevas métricas
* Añadir métrica editor de fórmulas y fórmula SUM

DIA 5 – Big Data, IA y Machine Learning

* Big Data IA Machine Learning Deep Learning
* Es importante no caer en el *hype* del marketing y la innovación asociado al término de Big Data
* Vs del Big Data: velocidad, variedad, volumen VALOR Viabilidad, Veracidad, Validez, Volatilidad.
* Repaso de unidades 1 y 2
* Ejemplos de aplicaciones y de su crecimiento: Tik Tok, LinkedIn, etc.
* Algoritmos supervisados y no supervisados. Ejemplos.
* Ejemplos de publicidad programática: la compra programática nos permite el uso de data de los usuarios (histórico y en tiempo real) y nos capacita para abordar la personalización del mensaje publicitario. Esto se traduce en dos grandes ventajas: Posibilidad de segmentar de manera muy precisa el público objetivo al que queremos impactar.
* Data Lake: para almacenar datos de forma no estructurada (ej: texto, imágenes, artículos, correos, etc.). Repositorio central.
* Ventajas: más flexibles, no diseñados de antemano.
* Ejemplo: Amazon S3
* ¿Ventajas e inconvenientes Data Lake VS Data Warehouse? ¿ETL VS ELT?
* Depende de coste, pero cada vez menos (cruzar con tamaño de la empresa)
* Depende del tipo y cantidad de datos.
* Depende del esfuerzo que tenga que hacer la empresa para transformar esos datos y darles valor.
* En un ELT se depende mucho de la tecnología que se utilice
* Clustering: concepto de *centroide*
* Video aprendizaje supervisado VS no supervisado (VIDEO).
* Concepto de *espacios latentes*.
* El NO supervisado señala un camino muy prometedor
* Cajas negras: algoritmos con matemáticas, estadísticas, etc.
* Videos de Redes Neuronales.
* Explicación de funcionamiento con el ejemplo de gafas VR + nachos

EJERCICIOS

Ejercicio 2: con la ayuda de la extensión ‘Data Miner’, haz un scraping y descarga los datos de las casas en venta en tu ciudad existentes en la web pisos.com. Ahora, haz una regresión múltiple en una hoja de Excel para predecir un precio en base a m2 y/o número de habitaciones.

* Uso de Data Miner en web [www.pisos.com](http://www.pisos.com/)
* Exportar/grabar a Excel
* Realizar gráficos de dispersión

Recursos de aprendizaje

<https://www.codecademy.com/learn/paths/bi-data-analyst> Cursos

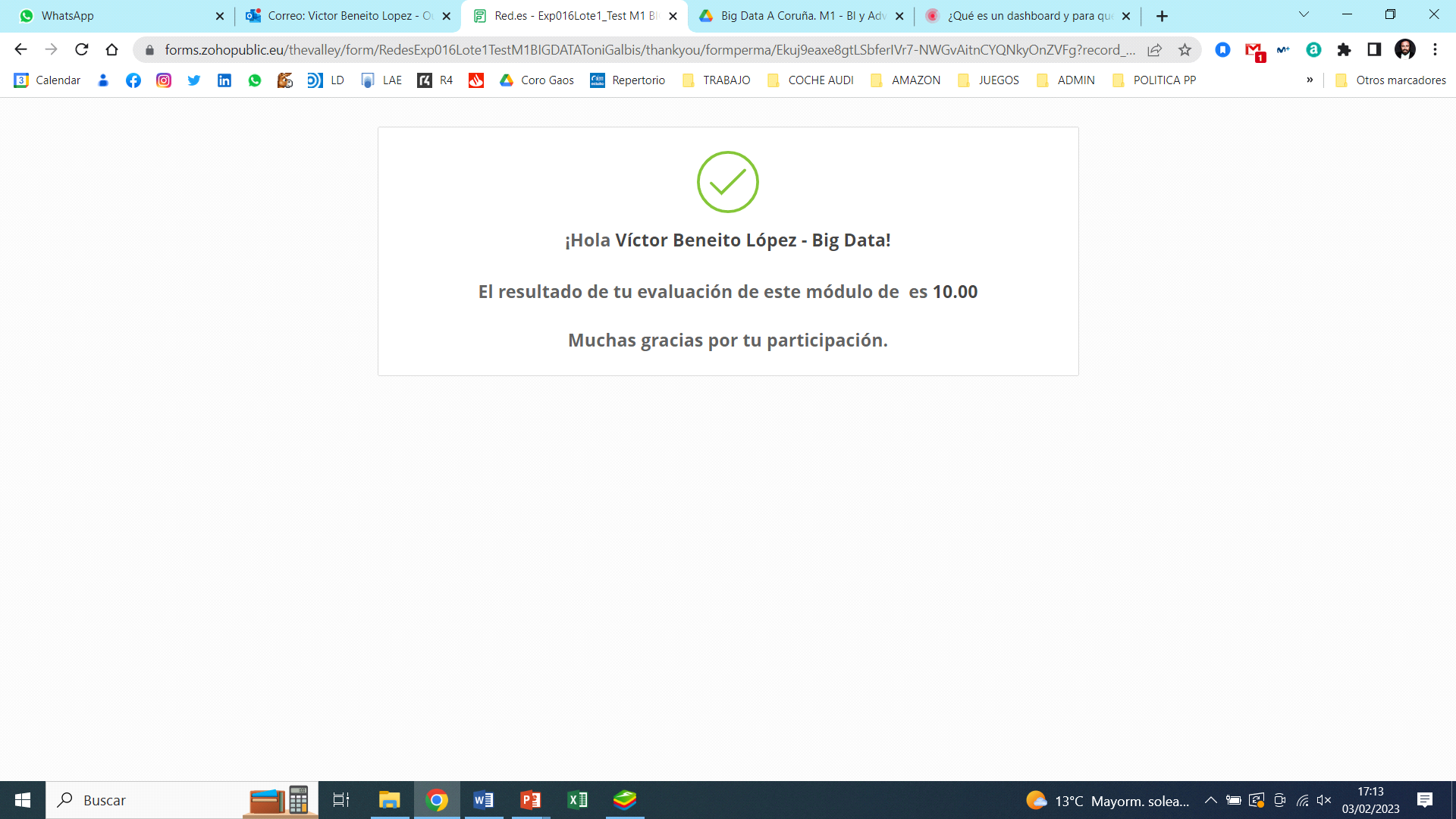
<https://8weeksqlchallenge.com/> Ejercicios SQL

(respuestas en:

<https://github.com/bcamandone/Data_Analysis_SQL/tree/main/8%20Week%20SQL%20Challenge>)

VIDEO ejemplo Red Neuronal: <https://colab.research.google.com/drive/1QH7yhAmklHxBRi1d-dZcI3Y8uN5WNbnF?usp=sharing>

EXAMEN DE VIERNES



DIA 6 – S1 Web Analytics

Profesor: Josema (Alicante). [josemathevalley@gmail.com](mailto:josemathevalley@gmail.com) <https://www.linkedin.com/in/josemarb/>

**Analítica en entornos digitales**

* 3 tipos en base a la finalidad: descriptiva, predictiva o prescriptiva.
* Público objetivo: perfiles generacionales,

**¿PARA QUÉ SIRVE LA ANALÍTICA DIGITAL?**

* Entender el tráfico web que recibe nuestro proyecto
* Definir mejores estrategias para comercializar nuestros productos/servicios
* Comprender el mercado y a nuestra competencia
* Optimizar nuestra estrategia de posicionamiento de marca
* Segmentar el público objetivo
* Tomar decisiones basadas en datos para modificar los procesos del negocio

**EJERCICIO: Crea tu proyecto.**

**PROCESO ANALÍTICA DIGITAL**

* Fase 1. Auditoría
* Fase 2. Estrategia
* Fase 3. Implementación
* Fase 4. Medición
* Fase 5. Optimización

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Competidores** | **Página Web** | **Cantidad Landing Pages** | **\*Keyword en Top3 Google** | **Perfiles Sociales** | **Tráfico web mensual** | **Velocidad carga web** |
|  |  |  |  |  |  |  |

\*Herramientas de medición: Keyword Sheeter detecta qué términos podemos utilizar para realizar las búsquedas/mediciones.

**ENTORNOS ANALÍTICA DIGITAL**

<TopOfMind Brand> lo primero que se viene a la cabeza (ej: hamburguesería – McDonald’s).

Uno de los objetivos principales de la analítica digital es permitirnos entender nuestra actividad en entornos digitales para poder tomar mejores decisiones. Una correcta decisión puede ahorrar enormes cantidades de recursos a una empresa, y permitirle obtener mayores beneficios.

Herramientas: análisis DAFO y benchmarking

**Ejercicio en grupo: DAFO de Samsung**

[**https://drive.google.com/file/d/1eKyZi8GMuthI8srhkJgp4YDrwDN\_5mud/view?usp=share\_link**](https://drive.google.com/file/d/1eKyZi8GMuthI8srhkJgp4YDrwDN_5mud/view?usp=share_link)

**Ejercicio en grupo: DAFO de EPIs y material sanitario.**

**BENCHMARKING**

Podemos utilizar el benchmarking para comparar cualquier tipo de métrica, y resulta muy útil en analítica web para comparar algunos de las siguientes:

* Nº de visitantes únicos.
* Cantidad de enlaces que llevan a nuestra web.
* Tasa de rebote.
* Tráfico orgánico.
* Tráfico referido.
* Ventas mensuales.
* Seguidores en redes sociales.
* Nº de referencias por categoría de producto.

**OBJETIVOS**

* Cuantitativos - Cualitativos
* SMART
* Recursos
* KPIs

**MÉTRICAS Y KPIs**

* Las métricas son aquellos valores numéricos que sirven para analizar el rendimiento de una determinada acción o proceso dentro de un negocio.
* TODO KPI ES UNA MÉTRICA, PERO NO TODA MÉTRICA ES UN KPI

**MÉTRICAS aplicadas al marketing online - Ejemplos**

* Coste de adquisición de cliente (CAC)
* Ciclo de vida del cliente o Life Time Value (LTV) – Gasto acumulado del cliente en 4 meses
* Retorno de la inversión (ROI): (Ingresos – Inversión) / Inversión
* Tasa de conversión: % de acciones exitosas
* Ingresos
* Leads o clientes potenciales: potenciales que nos dejan datos.
* Coste por lead (CPL)
* Número de visitas
* Tasa de clics o Click Trought Rate (CTR)
* Engagement: compromiso que se establece entre una marca y su audiencia en las distintas comunicaciones que producen entre sí.
* Tasa de rebote (Bounce Rate): porcentaje de visitantes que acceden a un sitio y salen sin acceder a nuevas páginas o interactuar con contenido, haciendo apenas una visualización de página. Entre mayor sea la tasa de rebote, peor es la interacción de los visitantes.
* Tasa de abandono de carritos.

**Debate en grupo: opinión acerca de la monitorización en páginas web y en dispositivos conectados como SmartTV.**

DIA 7 – S2 Web Analytics

**Insights**: claves/conclusiones para accionar el negocio: por qué el usuario decide realizar una acción u otra.

**CMS**: acrónimo de Content Management System. Sistema que nos permite crear, administrar y gestionar espacios we Se manera sencilla, sin necesidad de saber cómo programar utilizando lenguajes como php. Es, en definitiva, una “herramienta informática” que nos facilita el diseño y creación de páginas web gracias a una interfaz amigable y a un entorno con el que es muy fácil familiarizarse.

**PRÁCTICA**

* Crea tu propio e-commerce “SX” en la plataforma Ecwid: [http://www.ecwid.com](http://www.ecwid.com/)
* Cómo enlazarlo con GA4: [Acceso enlace con Google HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100449-Enabling-Google-Analytics-for-your-Ecwid-store"Analytics HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100449-Enabling-Google-Analytics-for-your-Ecwid-store" 4](https://support.ecwid.com/hc/en-us/articles/207100449-Enabling-Google-Analytics-for-your-Ecwid-store)
* Cómo añadir Google Ads: [Ecwid HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking" HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking"Help HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking" Center - Google HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking"Ads](https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking), [Conversion HYPERLINK "https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking" Tracking](https://support.ecwid.com/hc/en-us/articles/207100609-Google-Ads-conversion-tracking)

**Sesiones**: una sesión comienza cuando se accede a una página con el código de seguimiento de Google Analytics. Y termina tras 30 minutos de inactividad. Al volver a la página tras terminar una sesión, se inicia otra.

DIA 8 – S3 Web Analytics

* Si hacemos una campaña en Google Adds, se integran los informes y se pueden ver en Google Analytics (Google integra todo “lo suyo”).

Caso de uso

Vamos a analizar los informes de Usuarios con la cuenta de demostración de Google Merchandise Store. Utilice el periodo comprendido entre el 1 de marzo del 2022 y el 31 de marzo del 2022 para responder a las siguientes preguntas:

1. Informe Resumen de datos demográficos, ¿qué grupo de idioma aportó el máximo número de usuarios? Inglés 56.014

2. Informe Información geográfica > Ubicación correspondiente a la India, ¿qué región aportó el máximo número de usuarios? Maharashtra 1431

3. Informe Resumen de dispositivos móviles, ¿qué porcentaje de sesiones procedió de los dispositivos móviles? Mobile 39,4%

1. En el informe adquisición de tráfico> Fuente/Medio, al filtrar por tráfico referral, ¿qué sitio web envió el mayor número de usuarios? analytics.google.com / referral 1068

2. En el informe adquisición de usuarios > Canales, ¿qué tres canales tuvieron el mayor número de usuarios nuevos? Direct 20.709, Organic Search 18.094, Paid Search 7.400

3. En el informe adquisición de tráfico > en el caso de los usuarios que proceden de la fuente "google.com", ¿Cuál tuvo el máximo de sesiones? analytics.google.com / referral 1.686

1. En el informe de interacción > Paginas & pantallas, ¿cuál de estos títulos de página tiene más vistas de página? Shopping Cart, Home, Store Search results? Home 65.045 (22,51%)

2. En el informe Interacción > Páginas & pantallas, si nos centramos en el **grupo de contenido**, ¿qué páginas de destino tienen el porcentaje de vistas de página más alto? Men’s Unisex (13.566)

No: not set. /Google+Redesign/Apparel/Mensv (10.694).

3. En el informe monetización > Compras en comercio electrónico, ¿cuál es el mayor porcentaje carrito/visualizaciones por **marca** del artículo? Chrome Dino Collectible Figurines

No: Google (38.612)

**¿Qué es una Métrica?** Recuerda: En GA4 Medimos Interacciones = Eventos Dato numérico exacto que dentro de una escala predefinida nos indica el volumen alcanzado de una serie de elementos en un periodo de tiempo.

**Práctica: busca las métricas en la cuenta de demostración de GA4.**

Explorar – Iniciar una nueva exploración (en blanco) – Variables – Metricas + -

**Identificación de usuarios**: USER ID, Google Signals, DEVICE ID.

**Métricas más importantes:**

* Número de eventos,
* Conversiones (Número de eventos + Conversión: todos nuestros KPIs van a crearse uniendo un evento a una conversión),
* Valor del evento,
* Sesiones (cuando cierras página o no interactúas en 30 min).
* Sesiones con interacción,
* Usuarios activos,
* Usuarios nuevos,
* Usuarios activos en N días
* Añadir al carrito, compras, ingresos
* Reembolsos, transacciones
* Visitas
* Salidas
* Clics de GAds, Coste GAds, Coste por clic GAd, impresiones GAds, Visualizaciones de video GAds

**Debate**: cómo lograr atraer tráfico web para tu Ecommerce SX.

* Ejemplo de TESCO en Corea el Sur fotos con QR en el metro para venta online

Módulo 3

**Analítica avanzada**

Su enfoque es usar de manera precisa los datos de las distintas fuentes que tengamos para predecir posibles eventos y/o comportamientos que nos ayuden a afrontar cambios en el negocio

* Mejorar nuestra toma de decisiones al utilizar los datos de forma más concreta
* Automatizar procesos, ahorrándonos un tiempo valioso en tareas recurrentes además de minimizar costes
* Conseguir una mayor eficiencia al centrarnos en procesos de mayor importancia para nuestro negocio.

El uso de la analítica avanzada es esencial en la transformación digital de muchas empresas, debido a que, gracias al exhaustivo análisis de datos, podremos basar las decisiones que hagamos en información en tiempo real y no en suposiciones, instinto o puntos de vista.

**Dimensiones**: cada sistema de clasificación de los datos que segmenta el total de los datos capturados en pequeños grupos a los que pone nombre. Las **dimensiones** son únicas para cada DATO. Dimensión + Métrica = Tabla de resultados.

Dimensiones más relevantes de un evento:

* Nombre del evento
* Día, Semana, Mes, Año
* Fuente, Medio, Campaña
* Campañas a nivel de usuario / Campañas a nivel de sesión
* País, Región, Ciudad
* Edad, Intereses, Sexo
* Dimensiones por geolocalización GA4
* Dimensiones por Google Signal GA4
* Dimensiones por Tecnología GA4
* Dimensiones por Página / Pantalla
* Dimensiones por Comercio Electrónico

Caso de Uso - Práctica

Analizar las dimensiones y las métricas que aparecen en los diferentes informes de GA4 <https://support.google.com/analytics/answer/9143382?hl=es>

* Desarrolla 3 conclusiones para cada informe de Life Cycle.
* Adquisición
* Usuarios activos por semana en forma de gráfico:
* Pico de usuarios el día 31 de enero
* Máximo de usuarios en un día: 3.600
* País con más usuarios. EE.UU (43.000€)
* Usuarios nuevos por país y dispositivo
* En España, 24 y sólo desktop
* Interacción
* Monetización
* Numero de compras por semana: decreciendo
* Retención
* Usuarios recurrentes por ciudad: New York.
* Desarrolla 3 conclusiones para informe de Usuario.
* Grupos demográficos
* Tecnología
* Usuarios recurrentes por sexo y dispositivo: más hombres y por el ordenador.
* ¿Cuántos usuarios de Canadá con iPhone accedieron a la tienda de Google, durante noviembre y diciembre de 2022? 858 y 1261

**Comparaciones:** Sólo se pueden comparar DIMENSIONES.

**Valor del tiempo de vida del cliente (TVC):** promedio de ingresos procedente de los usuarios nuevos durante sus primeros 120 días.

**Informes**: dimensiones en filas y/o columnas, métricas en valores.

**Filas anidadas**: siempre se anida por la primera dimensión.

**Tablas dinámicas:** cruzar con filas y columnas.

**Looker Studio con Google Analytics:**

* (Fallo: error de cuota)
* Importar datos para realizar consultas

DIA 9 – S4 Web Analytics

Práctica

Realizar búsquedas de informes de Google Merchandise Store con el buscador inteligente para conocer:

* Ventas realizadas (ingresos) ingresos totales 16.824,69 US$
* Nivel de tráfico 16.412 usuarios
* Primeras visitas durante la última semana 13.395
* Nuevos usuarios adquiridos desde España 4
* Compras realizadas (transacciones/operaciones) 1587

**Métricas predictivas**: detección de anomalías.

**Seguimiento de campañas**: diferenciar entre medio (¿cómo?) y fuente (¿De quién? ¿Dónde?)

* Todo el tráfico que diriges hacia tu web debe etiquetarse.
* Directo / none: desconocido
* Orgánic: buscador
* Referral: resto webs
* CPC/DFP: campañas (de Google)
* Resto Medios (social media, etc.): resto de etiquetados

**Atribución de campañas**: agregar parámetros a las URL para identificar las campañas que refieren tráfico.

* Cuando agrega parámetros a una URL, siempre debe usar utm\_**source**, utm\_**medium** y utm\_**campaign**.
* utm\_**term** y utm\_**content** son opcionales.
* utm\_**es** simplemente el prefijo requerido para estos parámetros.

**Tipos de campañas de marketing:**

* Anuncios de texto en resultados de búsqueda
* Anuncios de banner en sitios web de editores estratégicos
* Lanzar campañas en redes sociales
* Por correo que muestran su marca y productos a los clientes

DIA 10 – S5 Entornos Analítica Web

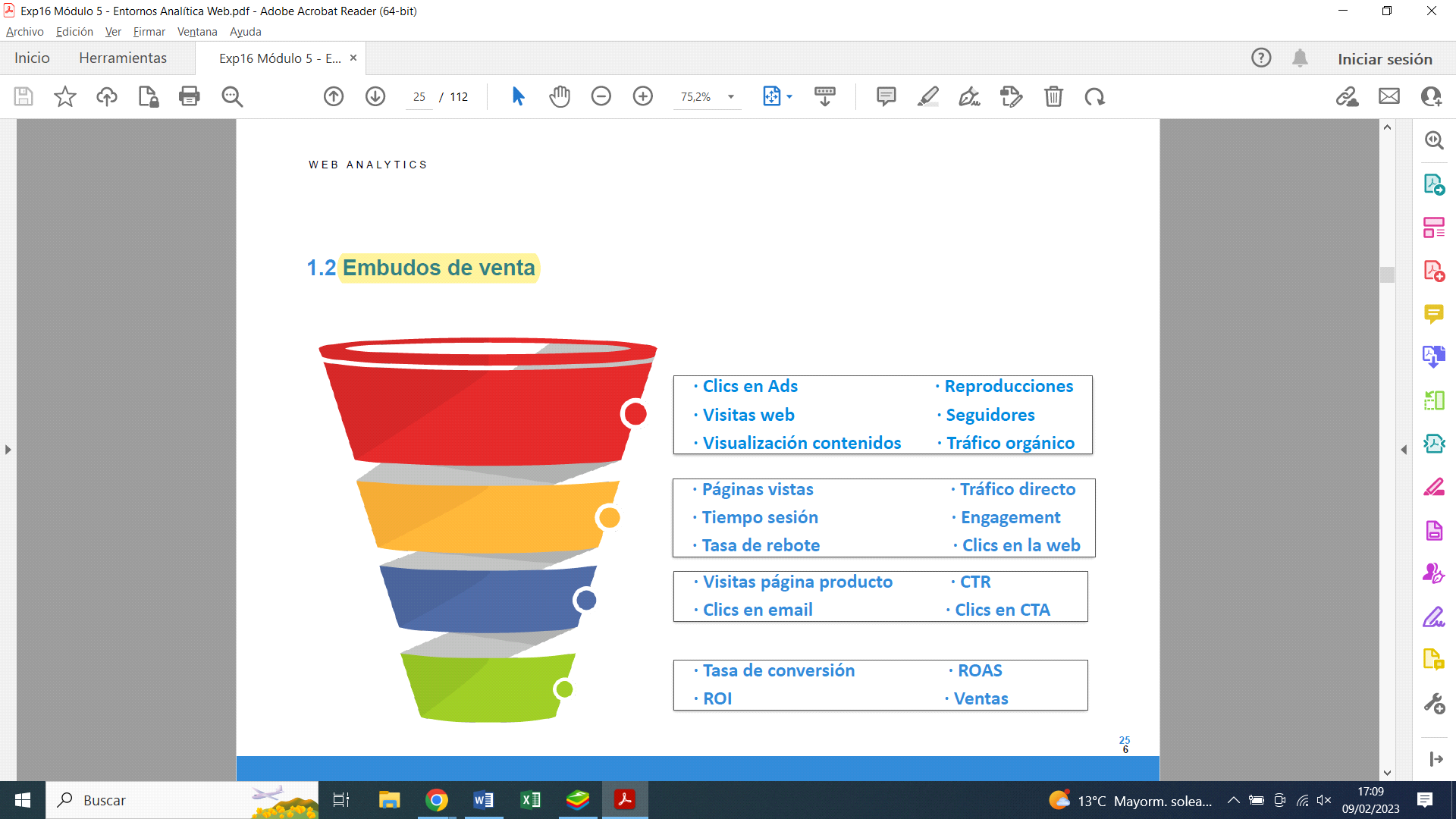
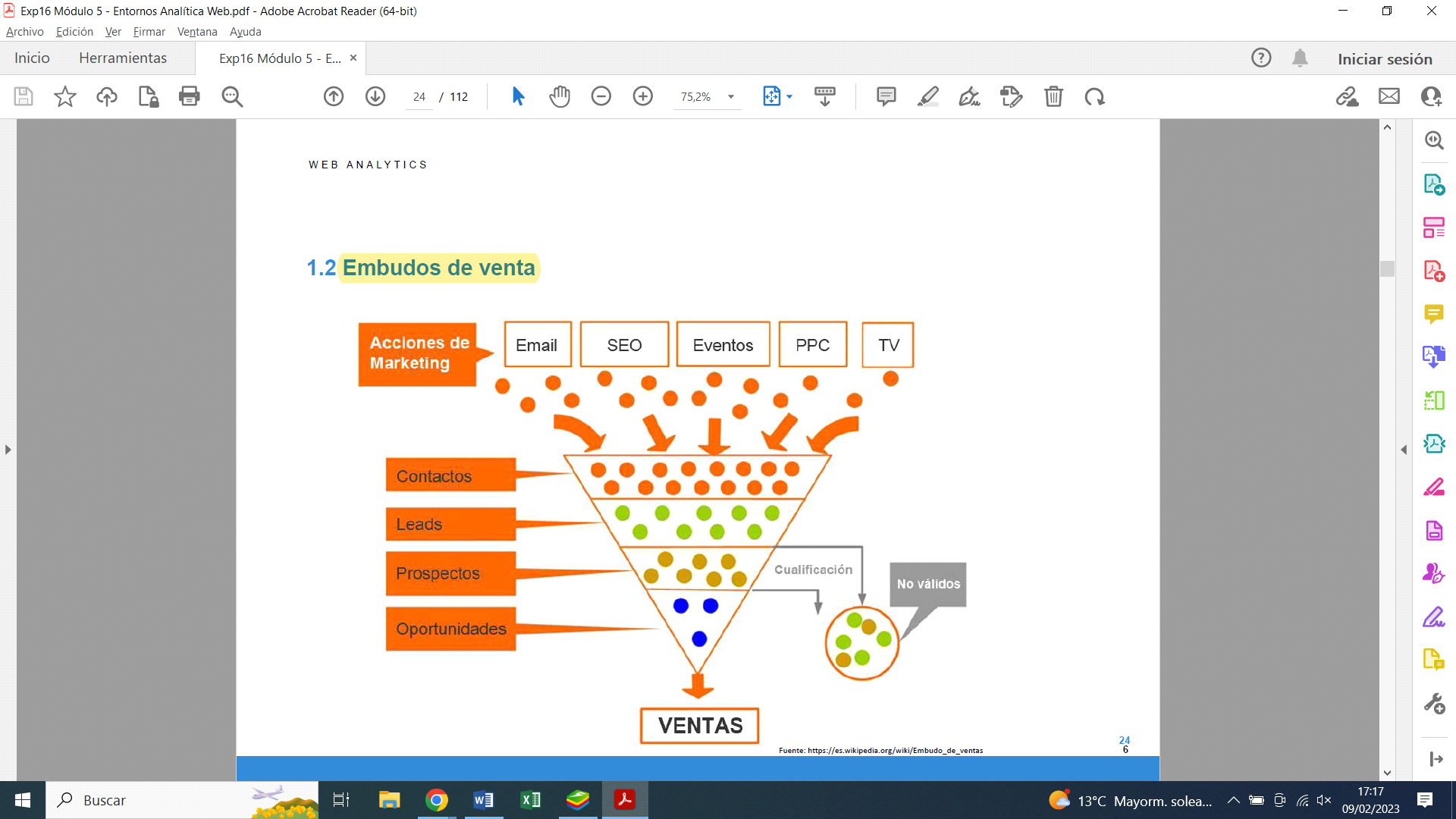
Entornos analíticos

**Estrategias de captación de tráfico**

* Necesidad (SEO, SEM, Shopping) VS Deseo (RR.SS., Influencers)
* Proceso conversión usuario cliente: primer contacto tráfico web Registro Remarketing

**Embudos de venta**

* 3 Fases: TOFU, MOFU, BOFU (top, middle, bottom).
* Diferencias en tipo de contenido, segmentación, formato, objetivo y KPIs.
* Estrategias pull



Práctica

Diseño un embudo de venta para tu propio proyecto SX, utilizando todas las herramientas y recursos disponibles.

* **TOFU**: generación de tráfico. Campañas de contenido:
* Artículos blogs: blog especializados en ocio de lujo con enlace a nuestra web
* Publicaciones promocionadas: webs de organización de eventos (empresa, bodas, viajes, etc.)
* SEM y SEO: palabras: eventos, lujo, caro, exclusivo, etc.
* Redes Sociales: Facebook, Instagram: fotos y videos de eventos redirigiendo a nuestras RRSS
* Influencers: moda y similares, contando experiencia de compra
* **MOFU**: campañas de interacción:
* Comentarios de clientes para recomendar y comentar (dándose previamente de alta en la web con datos básicos): “Cuéntanos cómo fue tu evento soñado”.
* Videos: “¿Te imaginas ser tú? Déjanos tus datos y te contaremos cómo hacerlo realidad.”
* Imágenes: coches, yates resto vehículos de lujo. Botón más información, pidiendo datos.
* RRSS: agregar perfiles que sigan/recomienden nuestras publicaciones
* **BOFU**: campañas de persuasión
* Primera toma de contacto gratuita / primer presupuesto gratuito: formulario de datos del cliente.
* Descuento de un % en el primer servicio al realizar X compra.
* Campaña de fidelización: plata, oro, platino

**SEO / SEM**

El objetivo esencial al hacer SEO será lograr que Google nos muestre entre los primeros resultados para determinadas búsquedas.

* Cuándo un cliente busque nuestra marca o producto/servicio
* Hay que entender la necesidad del cliente y cómo busca información
* Es imprescindible generar contenido de calidad alineado a las búsquedas de estos usuarios
* Ser creativos y persuasivos en nuestra web para generar ventas

Elección de **palabras clave**: baja competencia y gran número de búsquedas (Google keyword planner).

SEO **Onpage**

* lectura del código fuente de la página (etiquetas, etc.) para indexar
* Snippet:
* Limitación en cuanto a caracteres (70 para el título 156 para la descripción).
* Escueto, original, que incite al click y que contenga las palabras clave a posicionar.
* Imágenes en línea: para Google, mejor en formatos .png (más calidad) y .webp.
* Crawl Budget: presupuesto de rastreo (tiempo limitado, dependiendo de la “autoridad” de la web).
* Evitar poner vídeos directos.
* No abusar del enlazado

Práctica SEO Onpage

Componer los snippets para la *home* de nuestra web, teniendo en cuenta las directrices para escribir unos buenos snippets y un par de palabras clave a trabajar en SEO. Para ello usar cualquiera de estos 2 generadores:

* Merkle Mangools
* [SERP Simulator](https://www.serpsimulator.com/)

[**Disfruta de la vida sin límites... sueña y lo haremos realidad.**](https://www.serpsimulator.com/)

<https://tuhedonismo.company.site/>

Diseñamos experiencias exclusivas en España para disfrutar verdaderos momentos de lujo y hedonismo. Imagina tu mejor momento y lo haremos realidad.

[**Disfruta de la vida sin límites. Sueña con** HYPERLINK "https://www.serpsimulator.com/"lujo **HYPERLINK "https://www.serpsimulator.com/" y lo haremos realidad.**](https://www.serpsimulator.com/)

[https://tu**hedonismo**.company.site/](https://tuhedonismo.company.site/)

Diseñamos **experiencias** exclusivas en España para disfrutar verdaderos **momentos** de **lujo** y **hedonismo**. Imagina tu mejor **momento** y lo haremos realidad.

[**Disfruta de tu** HYPERLINK "https://www.serpsimulator.com/"hedonismo **HYPERLINK "https://www.serpsimulator.com/" sin límites. Sueña con** HYPERLINK "https://www.serpsimulator.com/"lujo **HYPERLINK "https://www.serpsimulator.com/" y lo haremos realidad.**](https://www.serpsimulator.com/)

[https://tu**hedonismo**.company.site/](https://tuhedonismo.company.site/)

Diseñamos **experiencias** exclusivas en España para disfrutar verdaderos **momentos** de **lujo** y **hedonismo**. Imagina tu mejor **momento** y lo haremos realidad

SEO Offline

* Se debe “ser enlazados” desde otras páginas para adquirir reputación/autoridad.
* Distinguir entre enlaces “follow” y “nofollow: enlaces que interesen o no (deben de tener relación de contenido). Se puede “pedir” a Google que no te enlacen.
* **Linkbuilding**: generar una red de enlaces externos que apunten hacia una de nuestras páginas web. (ej.: mala práctica – trampear con páginas ficticias. Ej.: bloggers, periódicos, influencers).
* **DA - Domain Authority**: refleja en una escala de 0 a 100 el nivel de autoridad de un dominio para predecir cómo se posicionará entre los resultados de búsqueda. Se compara con respecto a los competidores del sector.
* **PA - Page Authority**: refleja un valor entre 0 y 100 para determinar el nivel de autoridad de una página de nuestro sitio web y trata de predecir cómo de bien se posicionará entre los resultados de búsqueda para una serie de palabras clave. Se compara con uno mismo para ir mejorando.
* **Linkbaiting**: generar contenidos de altísima relevancia para lograr que otras páginas externas enlacen hacia nuestro sitio web. (Ej.: Wikipedia).
* ¿Cómo de bien se ha indexado nuestra página web? SATURACIÓN = [nº páginas indexadas / nº páginas creadas] x 100. \*Saturación = 100 ✅
* [SemRush](https://es.semrush.com/) es la mejor herramienta (prueba gratuita)

Práctica de enlaces

Diseña los siguientes enlaces desde uno de los contenidos de tu ecommerce a uno de los siguientes destinos:

* Enlace del blog a la página de inicio:

<a href=“tuhedonismo.company.site/">Regresar al inicio</a>

* Enlace desde la página de inicio a un producto

<a href=“tuhedonismo.company.site/Coches-de-lujo-y-deportivos-p529835504">Coches de lujo y deportivos</a>

* Enlace desde el blog a un producto de una web de terceros

<a href=“www.bershka.com” rel=“nofollow”> Descubre lo que venden en Bershka</a>

* Enlace desde un producto a tu política de privacidad.

<a href=“tuhedonismo.company.site/pages/privacy-policy">Declaración de privacidad</a>

Práctica de Linkbuilding

Averigua la autoridad de dominio y de página de dos competidores de un mismo sector. Puedes utilizar herramientas como:

* MOZ Analytics
* Website SEO Checker [https://websiteseochecker.com/#arearesult](https://websiteseochecker.com/)
* Small SEO Tools

Website Authority

|  |  |
| --- | --- |
| Moz Website Authority | |
| URL | [https://tuhedonismo.company.site/](https://semrush.sjv.io/c/3986542/1208110/13053) |
| Domain Authority (DA) | [77](https://semrush.sjv.io/c/3986542/1208110/13053) |
| Page Authority (PA) | [36](https://semrush.sjv.io/c/3986542/1208110/13053) |
| Total Backlinks | [0](https://semrush.sjv.io/c/3986542/1208110/13053) |
| Quality Backlinks | [0](https://semrush.sjv.io/c/3986542/1208110/13053) |
| Percentage of quality backlinks | [0%](https://www.semrush.com/analytics/overview/?ref=15200283&q=company.site%2F&searchType=domain) |
| MozRank | [4/10](https://semrush.sjv.io/c/3986542/1208110/13053) |
| [1000+ HYPERLINK "https://semrush.sjv.io/c/3986542/1208110/13053"keywords HYPERLINK "https://semrush.sjv.io/c/3986542/1208110/13053" lead to HYPERLINK "https://semrush.sjv.io/c/3986542/1208110/13053"company.site](https://semrush.sjv.io/c/3986542/1208110/13053) | [More Info](https://semrush.sjv.io/c/3986542/1208110/13053) |

Website Authority

|  |  |
| --- | --- |
| Moz Website Authority | |
| URL | [https://luxurymomentsgroup.com/](https://semrush.sjv.io/c/3986542/1208110/13053) |
| Domain Authority (DA) | [11](https://semrush.sjv.io/c/3986542/1208110/13053) |
| Page Authority (PA) | [19](https://semrush.sjv.io/c/3986542/1208110/13053) |
| Total Backlinks | [225](https://semrush.sjv.io/c/3986542/1208110/13053) |
| Quality Backlinks | [2](https://semrush.sjv.io/c/3986542/1208110/13053) |
| Percentage of quality backlinks | [1%](https://www.semrush.com/analytics/overview/?ref=15200283&q=luxurymomentsgroup.com%2F&searchType=domain) |
| MozRank | [2/10](https://semrush.sjv.io/c/3986542/1208110/13053) |
| [1000+ HYPERLINK "https://semrush.sjv.io/c/3986542/1208110/13053"keywords HYPERLINK "https://semrush.sjv.io/c/3986542/1208110/13053" lead to luxurymomentsgroup.com](https://semrush.sjv.io/c/3986542/1208110/13053) | [More Info](https://semrush.sjv.io/c/3986542/1208110/13053) |

**Las principales diferencias entre SEO y SEM**

 Como hemos visto, tanto el SEO como el SEM buscan mejorar el posicionamiento de un sitio web en los motores de búsqueda, como Google. Pero hay varias diferencias entre ellos:

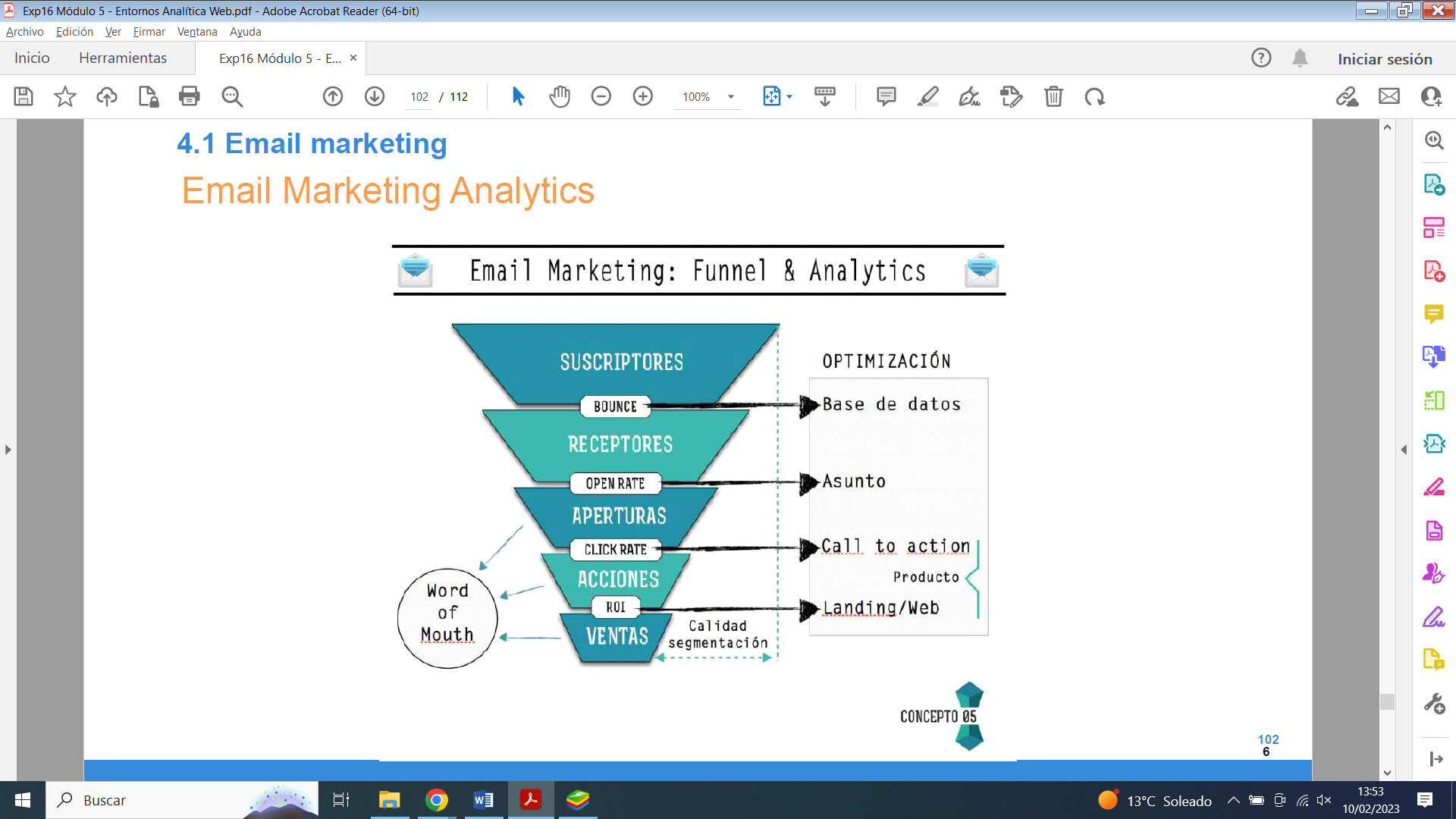
* La más evidente es **el tipo de inversión que requieren**. Existe el estereotipo de que "el SEO es gratis", pero esto no es exactamente así: conseguir una página web capaz de ocupar los primeros puestos en los resultados de búsqueda requiere tiempo, esfuerzo y, en última instancia, presupuesto. Pero a excepción de algunos gastos de mantenimiento, se trata sobre todo de una inversión inicial que da resultados en el futuro. En cambio, si usas SEM, tendrás que seguir pagando por cada clic en tus anuncios.
* El **lugar** que ocupará tu sitio web en las páginas de resultados también es diferente. Los anuncios de Google Ads ocupan una banda en la parte superior y una columna a la derecha de la página, mientras que los resultados orgánicos de búsqueda copan el espacio central.
* El **tipo de contenido** con el que trabajamos es diferente entre ambas estrategias. El SEO se basa en crear contenidos de calidad, generalmente en formatos más largos, mientras que en el SEM cuentas con anuncios con un número muy reducido de caracteres y landing pages diseñadas para conseguir el máximo impacto con los mínimos elementos.
* Por último, la **temporalidad** también es distinta: con el SEM se busca conseguir resultados a corto/medio plazo, mientras que el SEO es una apuesta a medio/largo plazo.

Analítica en Social Ads

* Redes Sociales muchas publicidad digital hipersegmentada.
* Segmentar el público es una variable fundamental para el éxito de una campaña.
* 3 tipos de estrategia, muy orientadas según el embudo de ventas:
* PPC - MOFU
* Visibilidad - TOFU
* Leads Ads - BOFU
* Métricas
* Alcance – personas
* Impresiones – veces que
* Clics
* CTR (Click Trought Rate)
* Tasa de conversión
* Leads

Otros entornos

Email marketing

Esta **estrategia** cuenta con algunas ventajas:

* Mayor flexibilidad
* Menor coste
* Inmediatez de ejecución
* Resultados en un menor tiempo
* Mayor capacidad de segmentación

**Variables / elementos de la estrategia**

* BBDD: alquilar, comprar, construir (más tiempo/recursos pero mejor).
* Contenido / asunto
* Envío
* Optimización

**Métricas clave**

* 📩 Emails entregados
* 📩 Tasa de rebote
* 📩 Emails marcados como spam
* 📩 Tasa de apertura
* 📩 CTR / Tasa de clics
* 📩 Tasa de conversión
* 📩 Bajas de suscriptores

**Herramientas**: mailchimp, sendinblue, etc.

Usabilidad / UX

Experiencia de usuario (o User eXperience) hace referencia a:

* Proceso en el cual un usuario interactúa con una marca a través de su web, puntos de venta y/o canales de comunicación.
* Poner al usuario en el centro de toda estrategia, de forma que cada diseño, contenido y decisión tome como punto de partida la satisfacción de los usuarios.

**Aspectos**:

* ✅ Usability
* ✅ Learning
* ✅ Memory
* ✅ Efficiency
* ✅ Satisfaction
* ✅ Accesibility

**Métricas** que necesitamos analizar:

* Duración de la sesión en la web
* Tasa de rebote
* Clics
* CTAs
* Páginas visitadas
* Conversiones
* Errores

**Herramientas**: hotjar (mapa de calor y pantallazos)

DIA 11 – S1 Data Management

Profesor/formador: Alejandro Arranz

Datos - BBDD

**Definición:** cifra, letra o palabra que se suministra a la computadora como entrada y la máquina almacena en un determinado formato. Mínima cantidad de información. Un [dato](https://es.wikipedia.org/wiki/Dato) es una representación simbólica de un atributo o variable cuantitativa o cualitativa. Los datos describen hechos empíricos, sucesos y entidades.

**Tipos de datos:** texto/string, fecha/hora, número/integer, boolean. Depende del tratamiento/finalidad que se vaya a dar al dato (ej.: teléfono – string, CP – string).

**Bases de datos:** definición, ventajas, etc. Ejemplo: Excel NO es una BBDD.

* **Relacionales**: tablas con datos almacenados en filas y columnas relacionadas entre sí.
* Elementos: registros (filas), variables (columna), celda/valor/dato (intersección).
* Integridad y estructura, almacenamiento, seguridad, escalabilidad.

**Tipos de relaciones:**

* Relación 1:N
* Es la más habitual.
* Ejemplo: Clientes – Pedidos, Productos – Pedidos.
* Relación 1:1
* Tablas con limitaciones.
* Ejemplo: Cliente – NIF
* Relación N:N
* Se da a través de tablas intermedias.
* Ejemplo: Clientes – Productos, con tabla intermedia Pedidos.

**Claves/Keys:**

* Primary Key (PK): en muchos casos suele ser un identificador (ID), no nulo y con valoración incremental. Clave primaria. No se puede repetir. NO todas las tablas tienen que tener un PK. Puede estar compuesta por varios campos.
* Foreign Key (FK): cuando una PK se utiliza en otra tabla (ej.: clientes, pedidos). Clave foránea o secundaria. Se puede repetir en tablas destino.

SQL

Lenguaje para hacer consultas (técnicamente NO es un lenguaje de programación).

Utilizaremos sistemas de gestión de bases de datos (SGBD), algunos con GUI otros con CLI.

Práctica

Separación en tablas y relacionarlas entre sí. Justificar campos, claves, etc.

DIA 12 – S2 Data Management

Ejercicio base de datos biblioteca.

SQL Segmento 1

* Métricas y dimensiones (variables numéricas y categóricas).

Uso de SQLiteonline. <https://sqliteonline.com/>

Statements & Clauses Básicos

* **Select**: qué campos (columnas) serán extraídos, separados por coma
* **From**: especifica qué tabla/s.
* **Limit**: restringe el número de filas a ser incluidas en el resultado final
* **As:** visualizar con un alias las columnas que especifiquemos.
* **Distinct:** fuerza valores únicos para el total de campos de UNA columna (hay que especificarla o utiliza la primera).
* **Where:** condicional/comparación para filtrar; NO separadas por coma, sino por operador lógico.
* *Between* NO lleva paréntesis. *IN* SÍ lleva paréntesis.
* <https://learn.microsoft.com/es-es/sql/t-sql/language-elements/operator-precedence-transact-sql?view=sql-server-ver16>
* **Order by**: ordenación. Ascendente por defecto. DESC – descendente. Se pueden utilizar varias columnas. Se puede poner un número como criterio, que corresponderá al orden de la columna en cuestión declarada dentro de la sentencia SELECT.

Ejercicios varios de querys. [Práctica](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\PR%25C1CTICA%20SQL%2014022023.docx) del PDF

DIA 13 – S3 Data Management

Estructura Case When Estatment

El CASE WHEN statement permite crear campos con condiciones específicas, donde el valor que adquiera cada fila dependerá del criterio establecido. El statement se invoca utilizando la siguiente sintaxis (entre el *SELECT* y el *FROM*):

*CASE*

*WHEN* condición\_1 *THEN* valor\_si\_true

*WHEN* condición\_2 *THEN* valor\_si\_true

…

*ELSE* valor\_si\_ninguno\_arriba\_es\_true

*END* (*AS)* nombre\_campo\_nuevo

CASE

WHEN…

…

END (AS) nombre\_campo\_nuevo\_2

* El nombre\_campo\_nuevo se podrá utilizar en sentencias WHERE, ORDER BY, etc.
* Se pueden generar las sentencias CASE WHEN que se quieran (generalmente seguidas). No se permite la anidación (al menos en SQLite).

Práctica sentencias SQL

DIA 14 – S4 Data Management

Funciones de agregación

* Realizan cálculos que permiten resumir información.
* Podemos elegir agrupar observaciones en una o más variables categóricas
* NO podemos utilizar funciones de agregación sobre otras funciones de agregación.
* Debemos utilizar funciones específicas. Siempre devuelven valor de UN ÚNICO campo/columna\*.

\*En caso de querer agrupar observaciones, debemos operar utilizando el **GROUP BY** statement y devolverá diferentes registros según su contenido. Se pueden utilizar diversos campos. Cuando agrupa, ordena de forma ascendente.

Funciones básicas (devuelven un número):

* **Count**(\*), **Count**(field), **Count**(distinct field)
* **Sum**(field)
* **Max**(field)
* **Min**(field)
* **AVG**(field)

Para filtrar sobre el resultado, NO se puede utilizar ~~WHERE~~ sino **HAVING** siempre DESPUÉS de **GROUP BY**. El **WHERE** es para filtrar sobre los datos originales – es decir, aplicará el filtrado ANTES de realizar la función de agregación - y NO sobre el resultado de una función de agregación. NO podemos utilizar funciones de agregación dentro de WHERE.

Ejemplos:

SELECT country, COUNT(movie\_title) AS Num\_Peli

FROM imdb\_movies

GROUP BY country/1

HAVING Num\_Peli > 10

ORDER BY Num\_Peli/2 DESC;

**Ejercicios**

Agregaciones condicionales

Restringimos la variable sobre la cual buscamos operar al cumplimiento de una o más condiciones. Realizamos un

**Ejemplo**

SELECT

CASE

WHEN gender LIKE 'Action%' THEN 'Action Movies'

ELSE 'Rest of Movies'

END AS peliculas,

COUNT(movie\_title),

ROUND(AVG(imdb\_score),2)

FROM imdb\_movies

GROUP BY 1

Ejemplo 2

SELECT

CASE

WHEN artist LIKE '%god%' THEN 'GOD'

WHEN artist LIKE '%death%' THEN 'DEATH'

WHEN artist LIKE '%black%' THEN 'BLACK'

END AS artist\_keyword,

COUNT(DISTINCT artist)

FROM rolling\_top\_albums\_1

WHERE artist\_keyword NOT NULL

GROUP BY 1

ORDER BY 2 DESC

Funciones de fecha

Uso para campos con formato DATE.

**STRFTIME**(*date\_part,date\_field*): permite extraer un parte de la **fecha** y la devuelve como **texto**

* STRFTIME(‘%m’,date\_field): devuelve el mes
* STRFTIME(‘%W’,date\_field): devuelve la semana
* STRFTIME(‘%d’,date\_field): devuelve el día
* STRFTIME(‘%Y’,date\_field): devuelve el año

**EXTRACT**(*date\_part* FROM *date\_field*): la devuelve como número

* EXTRACT(year FROM date\_field): devuelve el año
* EXTRACT(month FROM date\_field): devuelve el mes

**Práctica 10: Funciones de agregación condicionales**

SELECT

channelgrouping,

SUM(CASE WHEN devicecategory = 'mobile' THEN sessions END) AS mobile\_sessions,

SUM(CASE WHEN devicecategory = 'desktop' THEN sessions END) AS desktop\_sessions,

SUM(CASE WHEN devicecategory = 'tablet' THEN sessions END) AS tablet\_sessions,

SUM(sessions) AS total\_sessions

FROM google\_analytics\_formatdate

WHERE STRFTIME('%m',date) = '10' AND STRFTIME('%Y',date) = '2019'

GROUP BY channelgrouping

order by 5 DESC

DIA 15 – S5 Data Management

Corrección de práctica anterior

Subqueries

Realizar una consulta sobre el resultado de otra consulta, ya sea por la naturaliza de la misma o por las limitaciones del lenguaje.

La subconsulta se almacena en una tabla temporal para poder operar después con ella.

Para implementar una subquery, utilizamos el WITH statement con la siguiente sintaxis:

WITH

Temp\_table1 AS (SELECT…FROM),

Temp\_table2 AS (SELECT…FROM)

SELECT…FROM Temp\_table1

**Ejemplo**

WITH my\_subquery AS (SELECT Sector, SUM(Revenue) AS total\_revenue FROM fortune GROUP BY 1)

SELECT AVG(total\_revenue) AS avg\_sector\_revenue FROM my\_subquery

Uniones

UNION ALL

Combina dos tablas de firma vertical (uniendo filas). Incluye TODAS las filas, sin importar duplicados

UNION

Combina dos tablas de forma vertical (uniendo filas). Excluye duplicados del resultado final (aplica la cláusula DISTINCT automáticamente)

**Ejercicio**

Crear 2 tablas en Excel con formato .CSV con delimitador de “,”:

* Tabla 1: una con 3 registros y 2 columnas
* Tabla 2: otra con 2 registros y 2 columnas

JOIN

Combinar dos tablas de datos de forma horizontal (juntando columnas). Existen distintas formas de realizar joins, estas difieren en el tratamiento de las observaciones que no coinciden.

Una *join* de tipo INNER mantiene solo las observaciones que coinciden en tambas tablas (LEFT y RIGHT).

SELECT Orders.OrderID, Customers.CustomerName, Orders.OrderDate

FROM Orders

**INNER JOIN** Customers **ON** Orders.CustomerID=Customers.CustomerID;

Una *join* de tipo LEFT mantiene todas las observaciones de la tabla principal (LEFT) e integra las observaciones que coinciden de la tabla secundaria (RIGHT).

SELECT Customers.CustomerName, Orders.OrderID

FROM Customers

**LEFT JOIN** Orders **ON** Customers.CustomerID = Orders.CustomerID

ORDER BY Customers.CustomerName;

Una *join* de tipo RIGHT es el opuesto a una LEFT join. Mantiene todas las observaciones de la tabla secundaria (RIGHT) e integra aquellas observaciones de la tabla principal (LEFT) cuyos valores coincidan.

SELECT Customers.CustomerName, Orders.OrderID

FROM Customers

**RIGHT JOIN** Orders **ON** Customers.CustomerID = Orders.CustomerID

ORDER BY Customers.CustomerName;

Una *join* de tipo FULL mantiene todas las observaciones de ambas tablas (LEFT y RIGHT), sin importar si coinciden o no.

SELECT Orders.OrderID, Customers.CustomerName, Orders.OrderDate

FROM Orders

**FULL JOIN** Customers **ON** Orders.CustomerID=Customers.CustomerID;

**Práctica**

Extraer consolas no discontinuadas y su fecha de lanzamiento (first retail availability). Agregar ventas globales de videojuegos publicados desde el año 2000. Ordenar según ventas totales de videojuegos (descendiente).

* Table 1: videogames\_games
* Table 2: videogames\_consoles

SELECT

T2.console\_name,

T2.first\_retail\_availability,

SUM(T1.na\_sales)+SUM(T1.eu\_sales)+SUM(T1.jp\_sales)+SUM(T1.other\_sales) AS global\_sales\_games

FROM videogames\_games T1

JOIN videogames\_consoles T2 ON T1.platform\_code = T2.platform\_code

WHERE

T2.discontinued = '' AND T1.year >= 2000

GROUP BY T1.platform\_code

ORDER BY 3 DESC

BIG DATA

El almacenamiento barato y el aumento de capacidad de las redes han hecho que el concepto Big Data haya ido variando a lo largo de los últimos años.

Tipos de datos:

* No estructurado: imágenes, documentos de textos, audios,
* Son aquellos que carecen de estructura clara o interpretable, por lo que su tratamiento digitala costumbra a ser más complejo, o requiere un mayor procesamiento.
* Semiestructurado: redes sociales, html, xml,
* Presentan cierta estructura, pero esta no es fija, pudiendo variar para diferentes registros.
* Estructurado: excels, bbdd,
* Son aquellos que poseen una estructura completamente definida, con un número de atributos (o columnas) fijos y con tipos de datos prestablecidos.

Tratamiento de datos

* Generalmente, un dato por sí mismo no proporciona información, o esta es mínima.
* Por tanto, surge un interés en «extraer información» de un conjunto de datos, lo cual generalmente requiere de algún tipo de proceso o tratamiento sobre los mismos.
* El tratamiento es la serie de procesos a los que sometemos los datos para convertirlos en información relevante.

Ciclo de vida del dato**:** recolección ingesta almacenamiento procesado análisis visualización publicación archivado/eliminación.

Ingesta de datos

* Proceso por el cual se recolectan datos de varias fuentes o bases de datos y se incorporan a un entorno unificado para su posterior procesamiento.
* Las 3 V: volumen, velocidad, variedad VALOR
* Limpieza y transformación de datos
* Durante la ingesta, estos pueden venir con formatos diversos que puede resultar conveniente convertir o dotar de estructura.
* Además, pueden contener errores o anomalías que deben ser corregidas.
* **Filtrado**: en ocasiones, podemos querer ignorar ciertos datos que no cumplen determinadas condiciones, o que no son relevantes para nuestro sistema.
* **Identificación**: es importante dotar a los datos ingeridos de un identificador único (ya sea propio o dependiente de la fuente de datos).
* **Revisión**: deberían ajustarse al esquema (dominio) especificado, cumpliendo con las reglas de integridad y coherencia impuestas. Si no lo hacen, pueden omitirse, subsanarse o marcarse como inválidos.
* **De**-**duplicación**: cuando incorporamos datos de varias fuentes (o varias consultas), es fácil que nos encontremos con datos duplicados, incluso si su estructura no es totalmente idéntica. En este caso, es conveniente eliminar los duplicados.
* **Transformación**: cuando disponemos de datos estructurados o semiestructurados, puede ser conveniente transformar su estructura a una fija con el fin de unificar las diferentes fuentes de datos.
* **Estructuración**: cuando se dispone de datos no estructurados, puede resultar interesante tratar de dotarlos de cierta estructura para facilitar su posterior análisis (en ocasiones, el «machine learning» puede ayudar a esto).

Proceso ETL

ELT es la filosofía Extract—Load —Transform, que plantea realizar la carga de los datos tras su extracción (en «crudo»), transformándolos cuando sea necesario y del modo que resulte más apropiado en cada momento.

**Datos en crudo:** debido al abaratamiento de los costes de almacenamiento, puede ser interesante almacenar no solo el dato procesado, sino también el original o «crudo», por si fuera útil en el futuro.

Paradigmas de procesado de datos

**Migración (lote único):** proceso de realizar un procesado puntual para transformar unos datos en otros, o para tratarlos de algún modo.

**Bach (por lotes):** En una aproximación batch o por lotes, grandes cantidades de datos se procesan de golpe, generalmente mediante algún enfoque distribuido. Este proceso puede repetirse periódicamente, según se dispone de nuevos datos.

**Batch: MapReduce:** Es un paradigma de procesado de datos por lotes que permite transformar y agregar datos de forma distribuida.

**Streaming (tiempo real):** En una aproximación en streaming, los datos llegan de forma continua y a gran velocidad, y este flujo de datos se va transformando y procesando a medida que va llegando al sistema.

**Streaming Lambda:** Es una arquitectura que permite el procesado de datos tanto por lotes como en tiempo real, permitiendo que ambas capas se retroalimenten, y proporcionando una salida conjunta.

**Elección**: ejemplos de elección.

Gestión y documentación de datos

**Ventajas**:

Facilita la comprensión de los procesos de datos a los nuevos empleados de la compañía, y permite la continuidad en caso de renovación del equipo de gestión de datos.

Permite la revisión sistemática de los procesos de ingesta y tratamiento de datos, asegurándose de que las fuentes de datos permanecen correctamente conectadas y la ingesta se realiza satisfactoriamente

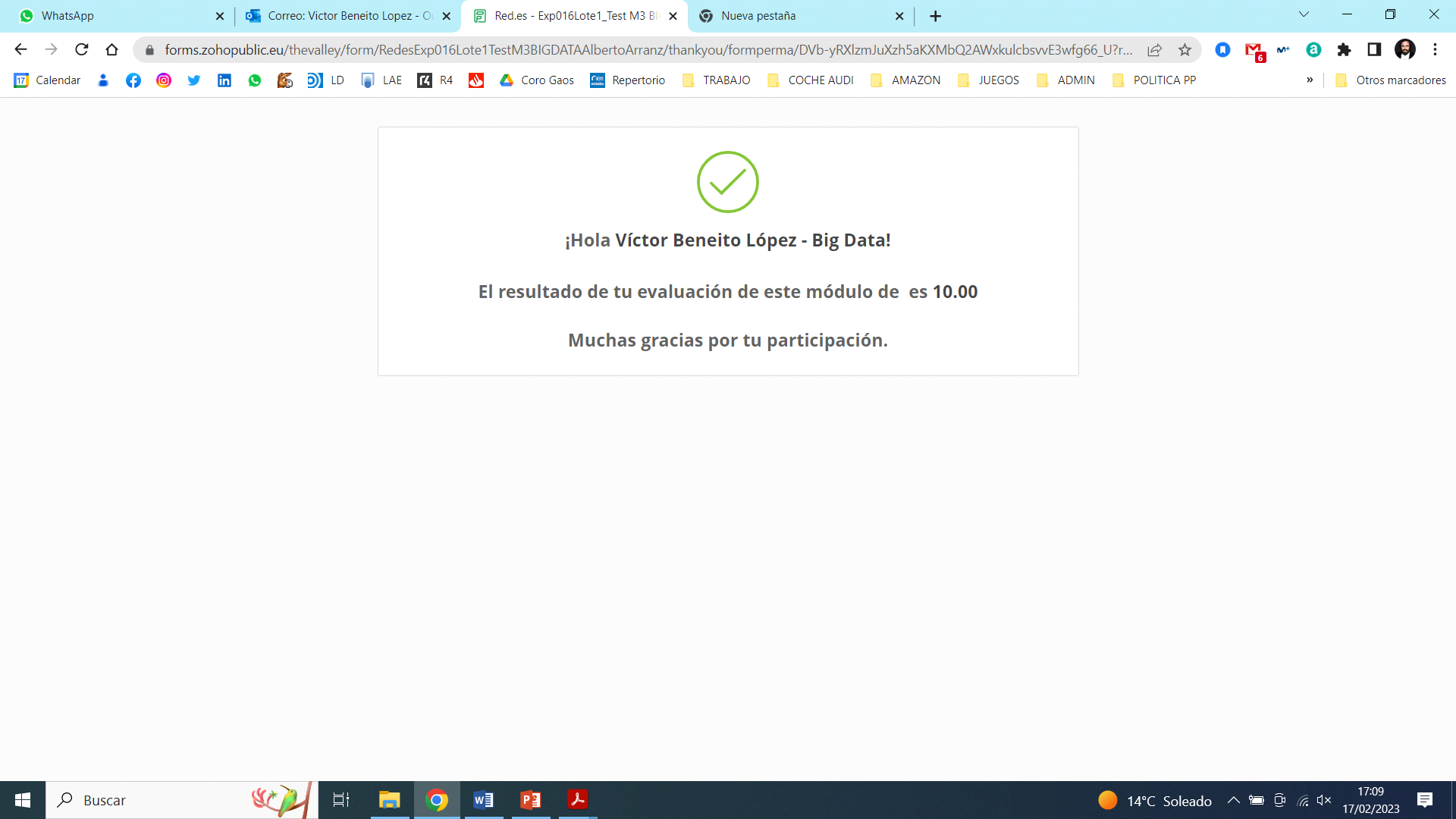
Facilita la auditoría de los datos, tanto interna como externa, ya sea con fines técnicos o regulatorios.

Facilita la revisión de mecanismos de almacenamiento y procesado para asegurarnos de que siempre responden a las necesidades de la compañía.

**Data Management Plan (DMP)**

* Describir el propósito de los datos
* Describir los orígenes de datos
* Especificar los esquemas de datos
* Describir los mecanismos de ingesta
* Describir los almacenes de datos
* Describir los mecanismos de tratamiento y procesado de datos
* Establecer procesos de revisión y actualización de los mecanismos
* Establecer políticas de publicación, reutilización y archivado

**Examen 17/02/2023**



DIA 16 – S1 Data Fundamentals con Python

**Actividad por grupos:** pensar y redactar distintos problemas (al menos 3) que podamos automatizar o resolver con un ordenador. Transformar la solución en pseudocódigo.

**Google Colab / Python**

* No permite trabajar de forma simultánea – hay que tener precaución al trabajar en equipo. Se debe de tener un control de las versiones/cambios de los archivos.
* Los nombres de las variables deben ser significativos y tener sentido.
* El lenguaje se llama MarkDown.
* Python es un lenguaje de tipado débil no hay que definir el tipo de las variables ni tampoco inicializarlas.
* Si en la última línea de código se pone una instrucción, entiende que hay que mostrar el resultado de la misma en pantalla.
* Python es sensible a minúsculas y mayúsculas.
* Muy a tener en cuenta con valores True y False
* Python es un lenguaje tabulado (ej.: estructura If)

**Comentarios**:

* # (hashtag) una línea o tras una instrucción
* ``` comentario ``` (tres comillas) varias líneas. Permite que dentro vayan ‘ y “

Crear **variable**: nombre = valor no hay que declarar el tipo

**Funciones**: nombre de la función(a1,a2) paréntesis y argumentos separados por coma

* La función *print* no se puede asignar a una variable

La **clase** *tuple* en Python es un tipo contenedor, compuesto, que en un principio se pensó para almacenar grupos de elementos heterogéneos, aunque también puede contener elementos homogéneos. Junto a las clases *list* y *range*, es uno de los tipos de secuencia en Python, con la particularidad de que son inmutables.

* Tupla: variable = (valor1, valor2,…,valorN)
* Lista: variable = [valor1, valor2,…, valorN]
* Rango: lista inmutable de números enteros en sucesión aritmética. Variable = rango(0,N)

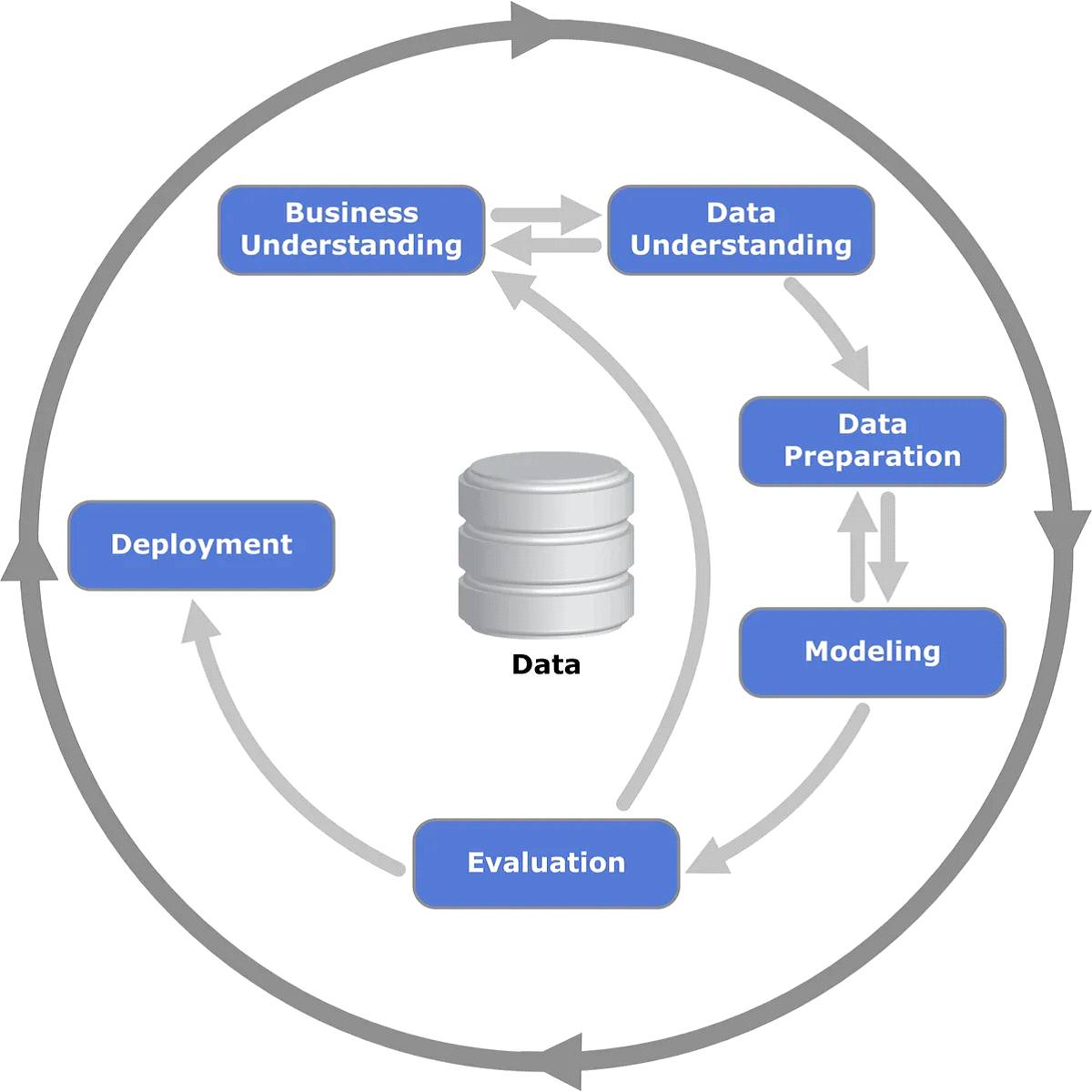
**Librerías**

* Hay librerías preinstaladas que se pueden utilizar directamente: librería.funcion()
* Para importarlas: import nombre\_librería

Instrucción condicional: if condición1: elif condicion2: else:

DIA 17 – S2 Data Fundamentals con Python

CRISP-DM

* [Estándar de estructuración de datos.](https://www.datascience-pm.com/crisp-dm-2/)
* **Primeros pasos**
* ¿Cuál es el Objetivo?
* ¿Qué tipo de problema es?
* ¿Qué queremos predecir?
* ¿Tenemos datos suficientes?
* ¿Calidad de datos mínima?
* ¿Qué criterio de aceptación tenemos?
* ¿Funciona como esperábamos?
* **Fases**: business understanding data understanding data preparation modeling evaluation deployment / \*vuelta al inicio \*
* Primero debemos entender el modelo de negocio.
* Después debemos entender los datos que tenemos y/o con los que trabajaremos: cantidad, estructura, exploración/captación, calidad, etc.
* Posteriormente se pasará a la preparación y modelado de los datos.
* Evaluación de resultados, revisión de procesos, próximos pasos.
* Si NO se ha llegado a lo esperado vuelta al inicio.
* Deployment o implementación cuando se sigue todo lo anterior.
* Uso de metodologías AGILE o WATERFALL (cascada).

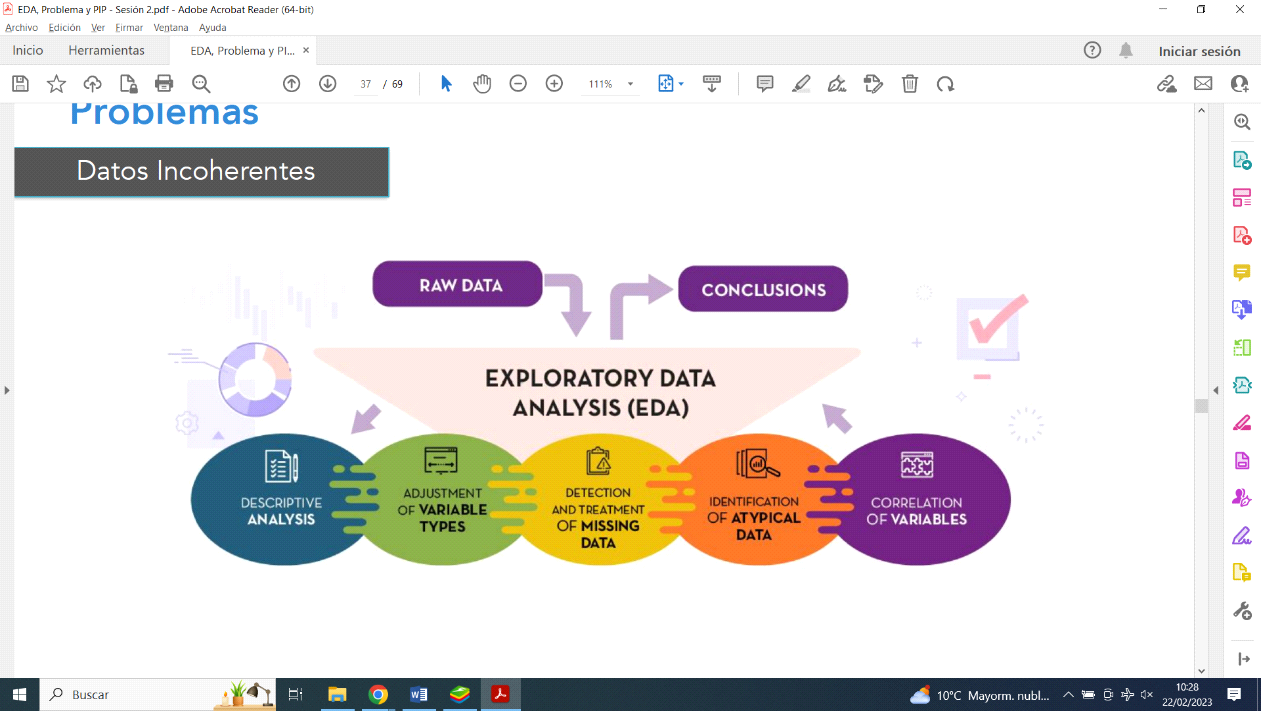
RIRO-GIGO

* Rubbish In Rubbish Out
* La **limpieza** de datos es fundamental. Da igual el modelo si los datos no son correctos.
* ¿Tenemos datos suficientes?
* ¿Calidad de datos mínima?
* ¿Qué tipo de problema es?
* ¿Qué criterio de aceptación tenemos?
* ¿Funciona como esperábamos?

Problemas

El sesgo

* Aunque las tres respuestas son técnicamente correctas, por alguna razón tenemos un sesgo para preferir una de ellas.
* No todas las personas compartirán ese sesgo; lo que percibimos y cómo respondemos está influenciado por nuestras normas.
* **Equidad**: evaluar un modelo de aprendizaje automático requiere hacer más que solo calcular las métricas de pérdida. Antes de implementar un modelo en producción, es fundamental auditar los datos de entrenamiento y evaluar las predicciones para determinar si existen sesgos.
* Diferentes tipos de sesgos cognitivos humanos se pueden manifestar en los datos de entrenamiento proporcionar estrategias para identificarlos y evaluar sus efectos.
* Modelado Data Science: varianza (variance) y sesgo (bias).

Datos incoherentes

* Exploratory Data Analysis (EDA)
* Conceptos: histograma, distribuciones, campaña de gauss, datos atípicos, etc.

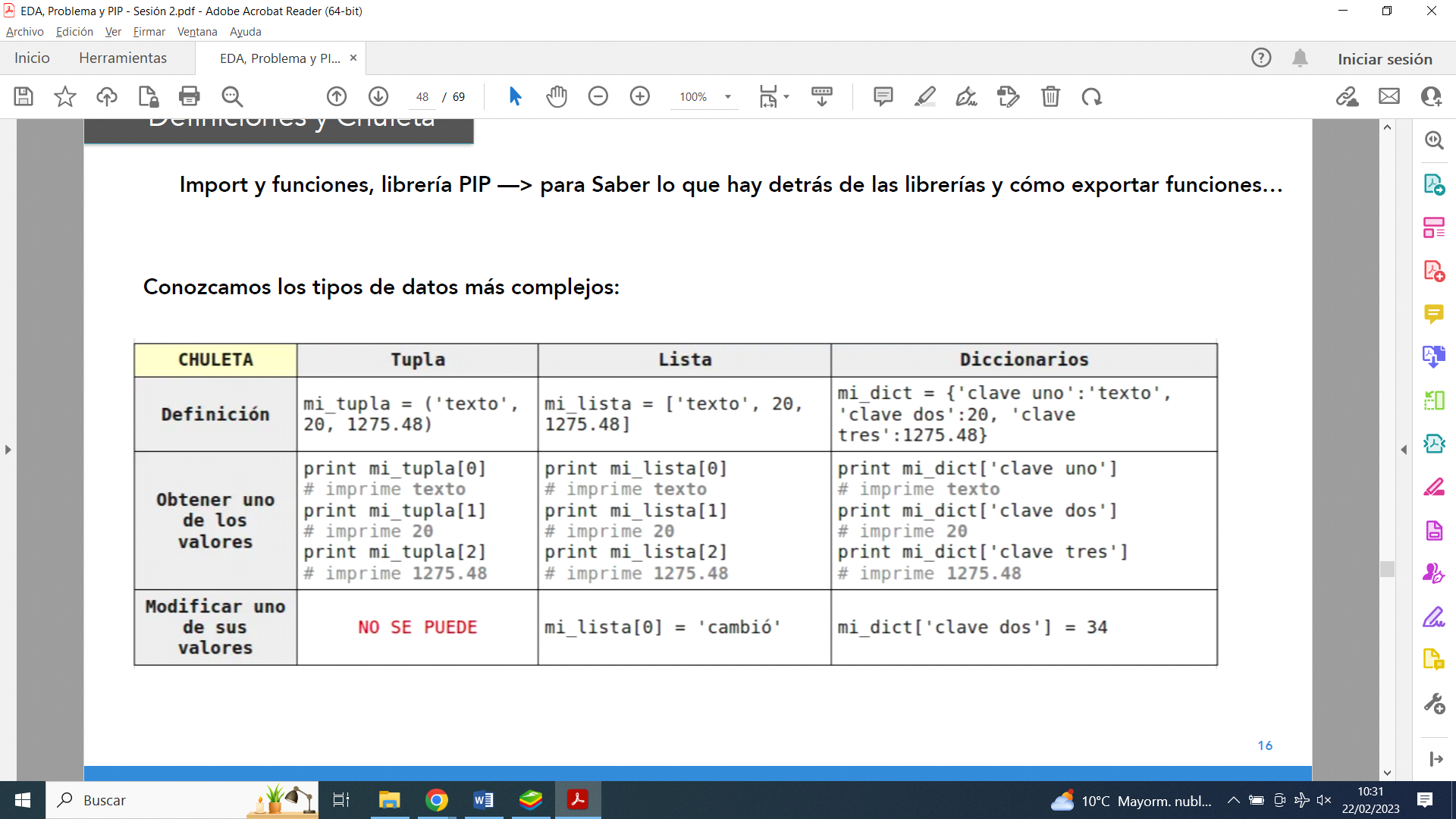
Datos incompletos y/o insuficientes

* Los datos reales tienen restricciones: coste, protección/privacidad, leyes, etc.
* Datos sintéticos: generados artificialmente.
* Más fácil de generar a partir de datos reales.
* Más baratos.
* Mantiene la estructura de los originales, pero NO es la misma.

Librerías y PIP

* Import y funciones, librería PIP para saber lo que hay detrás de las librerías y cómo exportar funciones…

Tipos de datos complejos



* Las claves de los diccionarios son de tipo texto.
* Las posiciones en diccionarios NO importan, ya que el orden lo da la clave.

Funciones

* Te permite definir un bloque de código reutilizable que se puede ejecutar muchas veces dentro del programa.
* Una de las grandes ventajas de usar funciones en el código es que nos permitirá reducir el número total de líneas de código en el proyecto.
* **Nativas**: el intérprete de Python tiene una serie de funciones y tipos incluidos en él que siempre están disponibles. <https://docs.python.org/es/3.9/library/functions.html>
* Los paréntesis van pegados a la función para su ejecución/invocación.
* NO es imprescindible que tengan argumentos o que devuelvan algún valor.

Librerías PIP

* <https://pypi.org/> Instalar librerías Repositorio Python Package Index
* Para crear datos:
* Jupyter <https://jupyter.org/>
* Pandas <https://pandas.pydata.org/docs/>
* NumPy <https://numpy.org/doc/stable/>
* Plotly <https://plotly.com/>
* Las librerías pueden tener dependencias de otras librerías y/o versiones..
* <https://www.anaconda.com/products/distribution>

Python Prácticas

* Strings y cadenas
* %... para no traducir. Pasarlo directamente al servidor/ordenador
* %pip list listado de librerías
* Código para montar la carpeta del Google Drive
* from google.colab import drive
* drive.mount('/content/drive')
* Instrucción bucle ***FOR*** *variable* ***IN*** *cadena* recorre cada valor/posición de la cadena
* Función strip() uso en strings para quitar espacios al principio y al final del texto

DIA 18 – S3 Data Fundamentals con Python

NumPY y Pandas

NumPY

* Nos ofrece objetos array n-dimensionales (vectores/matrices).
* Herramientas para la integración de código C/C++ y Fortran.
* Funciones de álgebra lineal, transformadas de Fourier, y funciones de aleatoriedad.

También puede utilizarse como un contenedor multidimensional de datos generales, lo cual permite que se conecte de forma sencilla y rápida a distintas bases de datos.

Muy útil. Aconsejable importarlo habitualmente.

Panda

Proporciona estructuras de datos rápidas, flexibles y expresivas, diseñadas para que trabajar con datos relacionales sea fácil e intuitivo.

* Es como NumPY pero con más elementos, más potente.
* Fallo o aspectos mejorables: son poco interactivos, poco intuitivos.

Jupyter

Nos permite la creación de documentos y programas con mayor versatilidad, pero… ¡YA LO HEMOS UTILIZADO!

Ventaja: crear archivos jupyter notebooks (ej: Google Colab).

Estructuras de datos

Ver apuntes de días anteriores.

Operaciones con listas

* lista.*append*(elemento) añade un elemento al final de la lista
* lista.*insert*(posición, elemento) inserta un elemento en la posición y desplaza el resto de elementos.
* list.*extend*(elemento1,elemento2,…,elementoN) añade lista/tupla de elementos al final de la lista.
* list.*index*(elemento) busca y muestra las primera posición donde coincida el elemento.
* list.*remove*(elemento) busca y elimina la primera posición donde coincida el elemento. NO devuelve valor.
* list.*pop*(posición) saca el elemento – lo devuelve - y lo elimina de la lista.

Operaciones con diccionarios

Recordatorio: el diccionario se ordena mediante sus claves. Muy útiles para crear JSON.

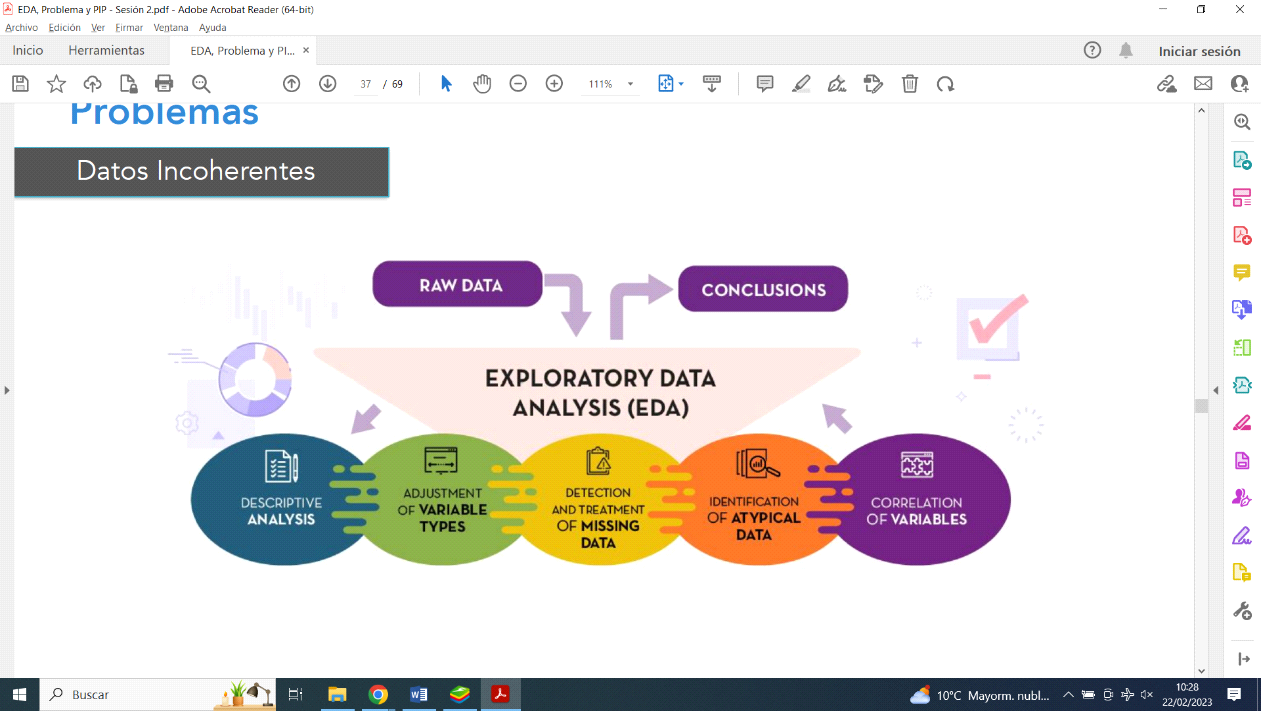
* Añadir: diccionario[nueva clave] = valor
* Actualizar: diccionario[clave] = nuevo valor
* Eliminar: *del* diccionario[clave]

Operaciones con arrays

* Shape: (valor\_dimension1, valor\_dimension2, valor\_dimension3,…, valor\_dimensionN)
* Ejemplo: shape(4 filas, 2 columnas, 3 profundidades)

EDA (Exploratory Data Analysis)

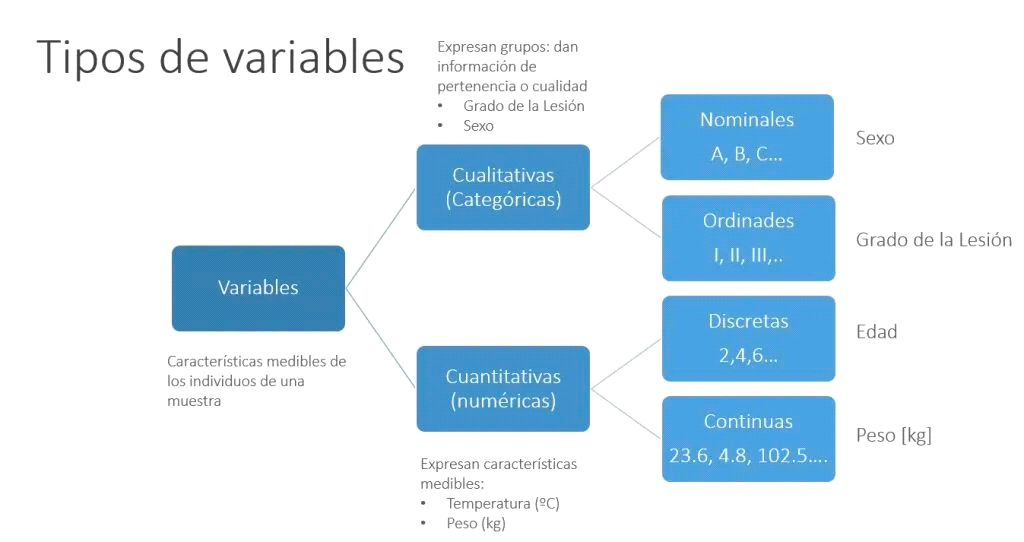
Empezamos preguntándonos... “¿de qué se trata?”, patrones y distribuciones. Nos permitirá:

* Conocer los Datos y entenderlos
* Identificar patrones e incoherencias
* Fijarnos en los Outliers o valores atípicos e identificarlos
* NO se manipulan datos.

**Preguntas básicas:**

* ¿Cuántos registros hay?
* ¿Son demasiado pocos?
* ¿Son muchos y no tenemos capacidad (CPU+RAM) suficiente para procesarlo?
* ¿Están todas las filas completas ó tenemos campos con valores nulos?
* En caso que haya demasiados nulos: ¿queda el resto de información inútil?
* ¿Cuáles parecen ser *features* importantes? ¿Cuáles podemos descartar?
* ¿Hay correlación entre *features* (características/campos/columnas)?

Pandas

* ¿Qué datos son discretos (se pueden medir por unidades) y cuáles son continuos?
* ¿Siguen alguna distribución?
* Obtenemos datos de casi cualquier fichero y podemos volcarlo a otro fichero.
* Creación de columnas a partir de otras.
* “Graficación” de datos sencilla implementada.
* Selección de datos específicos.

Ejercicios con estructuras de datos

Práctica Kaggle

Análisis de datos de un fichero CSV.

Tipos en Pandas: series (un conjunto índice/elemento) y dataframes (tablas y similares).

* <https://www.analyticslane.com/2021/07/15/pandas-cambiar-los-tipos-de-datos-en-los-dataframes/>

DIA 19 – S4 Data Fundamentals con Python

Continuar con la práctica de ayer.

<https://datacarpentry.org/python-ecology-lesson-es/04-data-types-and-format/>

Data Manipulation

Ver ejemplos y prácticas anteriores.

Data Wrangling

Hay más contenido teórico-práctico para manejar mejor los datos:

* Data Cleansing / Cleaning
* Data Preparation

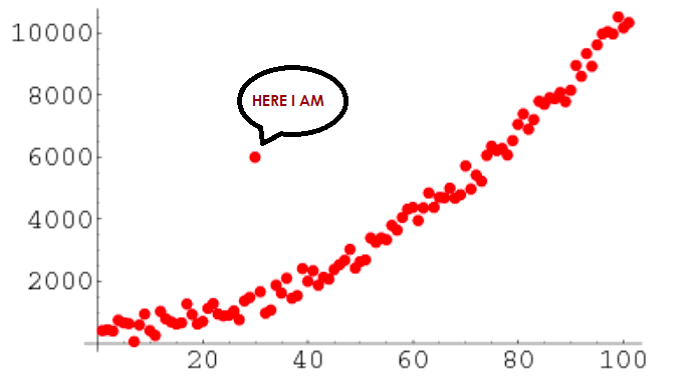
Una vez tenemos bien explorados los datos, debemos identificar los valores:

* Incompletos
* Nulos
* Incoherentes
* Inexistentes

*Incompletos y/o nulos*

* Tenemos que identificarlos y “marcarlos”.
* -1, Null, 0, Símbolos… Tenemos que saber dónde están y colocar una señal.
* Habrá veces que no tendremos buena distribución de los datos, cuando esto ocurra diremos que están no balanceados / desbalanceados o que no tienen dispersión suficiente.

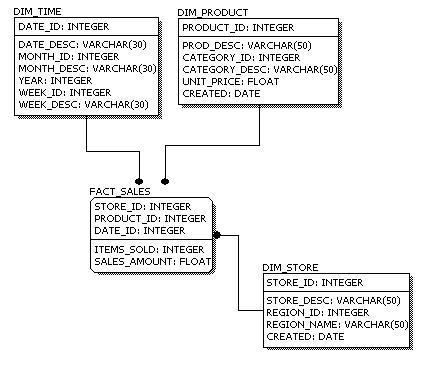
*Valores Atípicos*

* Valores fuera de la normal general *OUTLIERS*
* Utilizamos gráficas de dispersión
* Medidas de posición: cuartiles, deciles y percentiles
* Depende de la [distribución](https://es.wikipedia.org/wiki/Distribuci%C3%B3n_de_probabilidad) de los datos. Generalmente suponemos que es normal.

*Incoherentes*

* ¿Cómo son los datos?
* ¿Cuál es el rango de valores posible?
* ¿Existe un orden?

¡Un Diccionario de Datos es la respuesta! análisis exploratorio de datos.

**Ejercicios**:

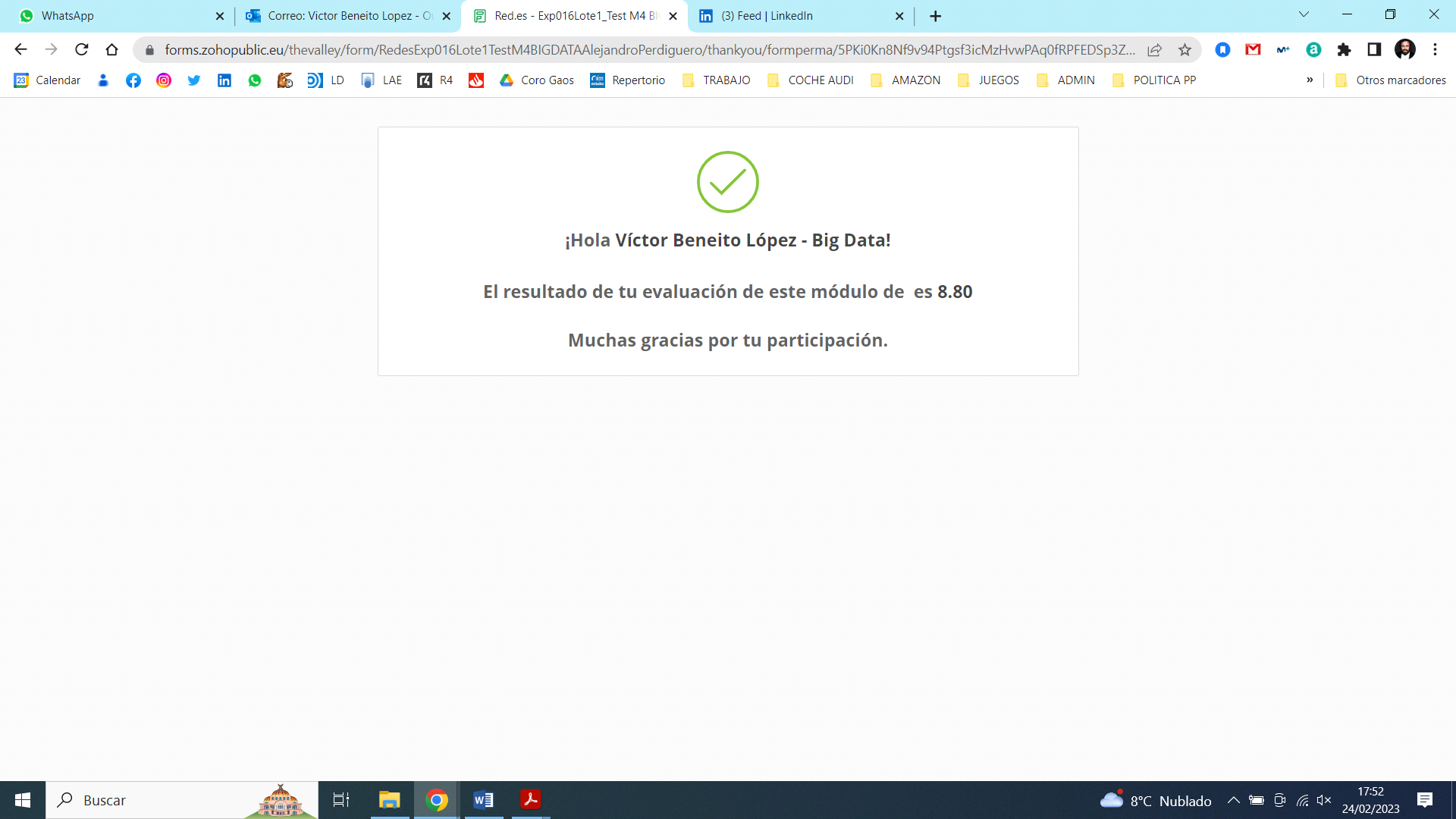
* calcular percentiles / cuartiles con la función .*describe()*
* recuperar datos de tabla con la función .*iloc(f,c)*
* suma de datos con la función .*sum()*
* comprobación de nulos con la función .*isnull()*
* recuperar datos no nulos con la función .*notnull()*

**Filtros**

* serie = tabla[‘índice’] <,>,==,etc valor de comparación tabla[serie]
* tabla[tabla['campo']<,>,==,etc. valor de comparación]

**Plotear**

* import plotly.express as px
* px.box(tablaSurveys, y='weight')



DIA 20 – S5 Data Fundamentals con Python

Corrección del examen de ayer

* Fallos: problema de corrección

Revisión de enlaces de interés del documento [Links de interés](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\MODULO%204\Links%20de%20Inter%25E9s.pdf)

<https://www.youtube.com/watch?v=YUAUwUt9ggg>

<https://informationisbeautiful.net/>

Uso de [Pandas\_Practica.ipynb](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\MODULO%204\Pandas_Practicas.ipynb) y estudiar las diferentes operaciones.

DIA 21 – S1 Data Science y Machine Learning

Profesora: Aurora Cobo Aguilera <https://www.linkedin.com/in/aurora-cobo-aguilera-9a8823117/>

<https://drive.google.com/drive/folders/1VaF-JxNvb5uBK8M5medu2ZnWuOWSS7YA?usp=share_link>

Contextualización

Diferentes roles:

* Big Data Architect / Developer
* Data Engineer
* Data Analyst
* Data Scientist Data Science & Machine Learning

Datos: pequeñas pinceladas

* Identificarlos
* Obtenerlos
* Transformarlos
* Almacenarlos
* Consumirlos
* Analizarlos

Business Intelligence: ¿BI vs Big Data / Data Science? Ambas se alimentan de datos masivos, pero …

* Big Data: Captura, almacenamiento y procesamiento de datos masivo
* Data Science: Análisis predictivo y prescriptivo de esos datos
* Business Intelligence: aprovechamiento de los datos para optimizar las decisiones y el reporting de una compañía

Python, ¿Por qué?

* Más fácil
* Multipropósito
* Sencillo de interpretar
* Muchas librerías
* Tipado fuerte y dinámico

Además, muy interesante para:

* IA, para hacer modelos de ML y algoritmos
* Big Data: Transformaciones de datos (ETL) y pequeños scripts

Ejercicio

* **¿Qué es una regresión lineal?**

En estadística, la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, m variables independientes X\_i con ∈ + término aleatorio. Este método es aplicable en muchas situaciones en las que se estudia la relación entre dos o más variables o predecir un comportamiento, algunas incluso sin relación con la tecnología. En caso de que no se pueda aplicar un modelo de regresión a un estudio, se dice que no hay correlación entre las variables estudiadas.

* **¿Y un dato perdido?**

Son aquellos que no constan debido a cualquier acontecimiento, como por ejemplo errores en la transcripción de los datos o la ausencia de disposición a responder a ciertas cuestiones de una encuesta. Los datos pueden faltar de manera aleatoria o no aleatoria.

* **¿Qué significa entrenar un modelo?**

El proceso de entrenamiento de un modelo de ML consiste en proporcionar datos de entrenamiento de los cuales aprender a un algoritmo de ML (es decir, el algoritmo de aprendizaje). El término modelo de ML se refiere al artefacto de modelo que se crea en el proceso de entrenamiento.

Utilizaremos el set de datos de entrenamiento para ejecutar nuestra máquina y deberemos de ver una mejora incremental (para la predicción). Recordar inicializar los “pesos” de nuestro modelo aleatoriamente, los pesos son los valores que multiplican o afectan a las relaciones entre las entradas y las salidas, se irán ajustando automáticamente por el algoritmo seleccionado cuanto más se entrena. Revisar los resultados obtenidos y corregir (por ej. inclinación de la pendiente) y volver a iterar…

* ¿Qué es un algoritmo?

Conjunto ordenado de operaciones sistemáticas que permite hacer un cálculo y hallar la solución de un tipo de problemas. Un algoritmo informático es un conjunto de instrucciones definidas, ordenadas y acotadas para resolver un problema, realizar un cálculo o desarrollar una tarea. Es decir, un algoritmo es un procedimiento paso a paso para conseguir un fin.

* ¿Y minibatch?

El Mini-batch Stochastic Gradient Descent es uno de los métodos más utilizados en el uso de las redes de neuronas profundas durante el Deep Learning, puesto que su proceso práctico resulta efectivo y sencillo en comparación con otros métodos de la teoría de optimización.

Práctica Notebook: 1. Estructuras de Datos en Python

Introducción al Aprendizaje Automático

**Motivación.: la Era del ‘big data’**

Estamos presenciando una explosión de aplicaciones donde el análisis de datos juega un papel fundamental. Tan importantes como las fuentes de datos son los algoritmos que extraen la información relevante.

Ejemplos de aplicaciones:

* Máquinas de búsqueda en Internet
* Biomedicina (clasificación automática de pacientes y
* enfermedades)
* Predicción de mercados financieros / Inversión
* Sistemas de recomendación

**Aprendizaje Automático o Machine Learning (ML):** es un conjunto de métodos que pueden automáticamente detectar patrones en datos, y utilizar los patrones descubiertos para predecir datos futuros, o realizar otros tipos de toma de decisiones bajo incertidumbre.

Utiliza herramientas de teoría de probabilidad. La teoría de la probabilidad se puede aplicar a cualquier problema que involucre incertidumbre. Muy relacionado con estadística.

* ¿Cuál es la mejor predicción sobre el futuro dados algunos datos del pasado?
* ¿Cuál es el mejor modelo para explicar ciertos datos?
* ¿Cuál medida debería ser la siguiente?

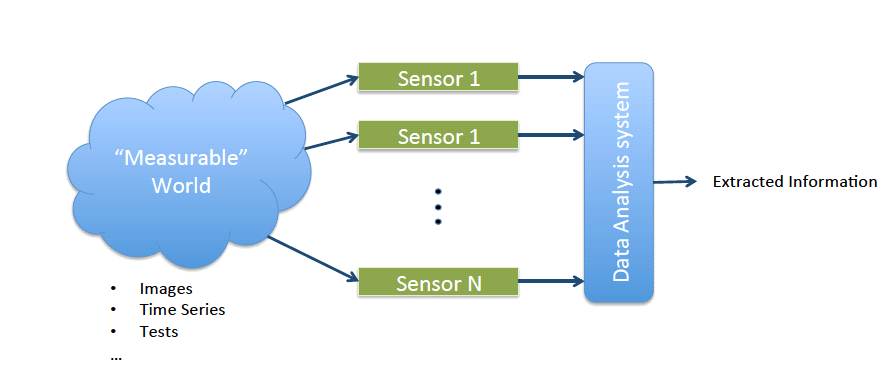
Ejercicio

Diferencias entre… ¿Qué son?

* **Machine Learning**: algoritmos que permiten conseguir el aprendizaje automático a través de los datos. Es el resultado de datos + variables + algoritmos.
* **Artificial Intelligence**: ciencia que tiene como objetivo imitar la mente humana.
* **Deep Learning**: máquina aprende por sí sola a través de un gran conjunto de algoritmos que imitan la red de neuronas del cerebro humano.

Conceptos fundamentales del ML

Elementos básicos

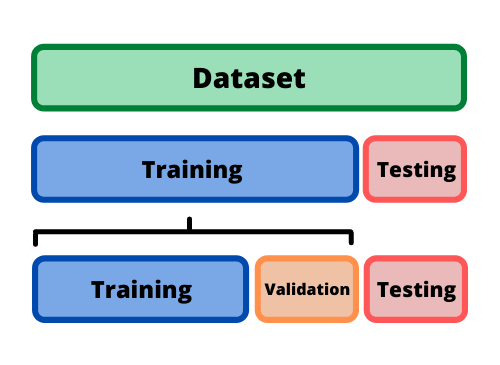
Construir modelos que se ajustan a una colección de datos

* **Modelo**: objeto de una clase (programa informático) o función matemática (modelo estadístico) con parámetros libres.
* **Conjunto de entrenamiento**: Conjunto de ejemplos sacados de la distribución de datos que se pretende modelar.
* **Optimizador**: Método que ajusta los valores de los parámetros libres del modelo (entrena) para que capturen patrones informativos contenidos en los datos.

Objetivo: **Generalización**

* Conseguir que el modelo pueda hacer predicciones más o menos correctas cuando se le presenten datos que no se hayan usado durante el entrenamiento.
* Evitar el *sobreentrenamiento (overfitting)*, que provoca un error pequeño en el entrenamiento, pero elevado en test. <https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/>

Conjuntos de…

**Entrenamiento**: Conjunto de datos usado durante el entrenamiento del modelo.

**Test**: Conjunto de datos que nunca se usa durante el entrenamiento ni durante la optimización de los parámetros, sólo para calcular la puntuación de test tras la finalización de dicho proceso.

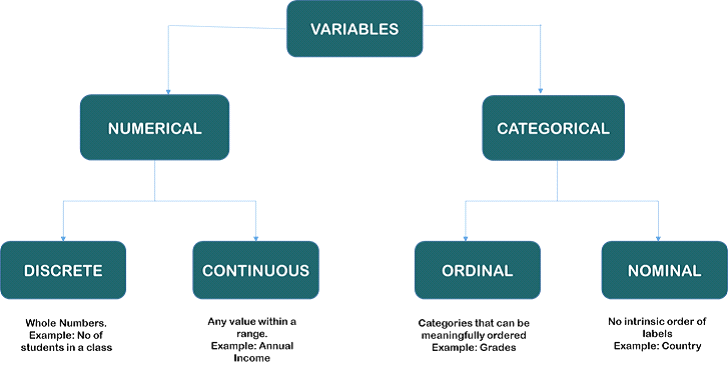
**Validación**: Conjunto de datos usado para evaluar el modelo durante el entrenamiento y elegir el mejor valor para los parámetros, los que obtengan mejor puntuación en dicho conjunto.

Repasemos EDA y Data Preparation

Conocer los datos, identificar patrones y detectar *outliers.*

Definiciones importantes

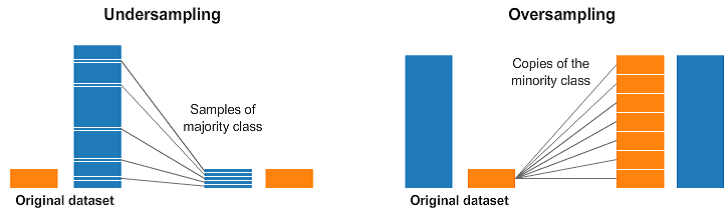
* **Observaciones:** también se les denomina registros. Hace referencia a cada una de las filas de la BD.
* **Características/variables:** también se les denomina *features*. Hace referencia a cada una de las columnas en una BD.
* **Missing data:** Son los valores perdidos de la base de datos. Se les puede llamar nulos. A veces son campos vacíos, NaN, None, 0, -1, …
* **Outliers:** son observaciones de la base de datos que se alejan de la distribución del resto. Es decir, son muestras muy diferentes a las demás.

**Variables**. Tipos de datos:

* Int numérica discreta.
* Float y double numérica continua.
* String categórica nominal u ordinal.
* Bool categórica ordinal.

Data preparation: proceso

Una vez que tenemos los datos explorados, empezamos a detectar:

* **Nulos (missing):** debemos detectarlos y colocar una señal que identifiquemos en nuestro código como nulos. Unificar todos los nulos de la misma forma. Opciones que podemos hacer, dependiendo del tipo y significado de los datos:
* Completar con una constante
* Completar con la media o la mediana
* Completar con el valor más frecuente
* Poner un flag (por ejemplo -999)
* Otro método de imputación
* Borrar las filas o columnas correspondientes (¡no recomendable!)
* **Valores atípicos (*Outliers*):** valores fuera de la norma general / de la distribución de los datos. Opciones que podemos hacer:
* Cambiarlos
* Eliminarlos
* **Datos incompletos:** hay veces que, aun teniendo todo el histórico de los datos, no es suficientemente buena la calidad de los mismos y no es representativa la distribución que tenemos. Ejemplo: datos no balanceados.
* **Datos erróneos:** una mala calidad en los datos también se ve reflejada en errores en las bases de datos.
* ¿Cómo son los datos?
* ¿Cuál es el rango de valores posible?
* ¿Existe un orden?

Práctica

Repasemos Pandas con Python e incorporemos algún concepto nuevo de ML.

[Notebook: 2. HYPERLINK "CURSO%20BIG%20DATA/MODULO%205/Notebooks/2\_PreprocesadoDeUnDatasetConPandas.ipynb"Preprocesado HYPERLINK "CURSO%20BIG%20DATA/MODULO%205/Notebooks/2\_PreprocesadoDeUnDatasetConPandas.ipynb" de un HYPERLINK "CURSO%20BIG%20DATA/MODULO%205/Notebooks/2\_PreprocesadoDeUnDatasetConPandas.ipynb"dataset HYPERLINK "CURSO%20BIG%20DATA/MODULO%205/Notebooks/2\_PreprocesadoDeUnDatasetConPandas.ipynb" con Pandas](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\MODULO%205\Notebooks\2_PreprocesadoDeUnDatasetConPandas.ipynb)

Herramientas para el ML

* Aplicación fundamental en el área de las matemáticas.
* Álgebra lineal, cálculo y probabilidad son los ‘lenguaje’ en los que el ML está escrito.
* El aprendizaje de estas materias ayudará a entender los algoritmos y a modificarlos o crear unos nuevos.
* Forma la columna vertebral de muchos algoritmos de ML.
* **Álgebra**: para el tratamiento de los datos vamos a trabajar con vectores (1D), matrices (2D) y tensores (mayores dimensiones) y operaciones entre ellos.
* **Cálculo**: es imprescindible en el paso de la optimización de una función de pérdidas para ajustar los parámetros del modelo. Optimización de una función de coste/error/pérdidas para ajustar los parámetros del modelo. Derivadas parciales: cómo se altera la función de pérdidas con individuales cambios en cada parámetro. Las derivadas se agrupan en matrices para un cálculo más directo
* **Programación**: es la herramienta para analizar los datos, entrenar los modelos y visualizar los resultados. Usaremos notebooks de python para estudiar todo el módulo, ver la teoría, procesado de datos y aplicación del modelo de ML con ejemplos prácticos. Ya los conocéis del módulo anterior. Google Colab.
* **Teoría de la probabilidad**: es la base teórica de un modelo de ML para cuantificar la incertidumbre.

Ejercicio

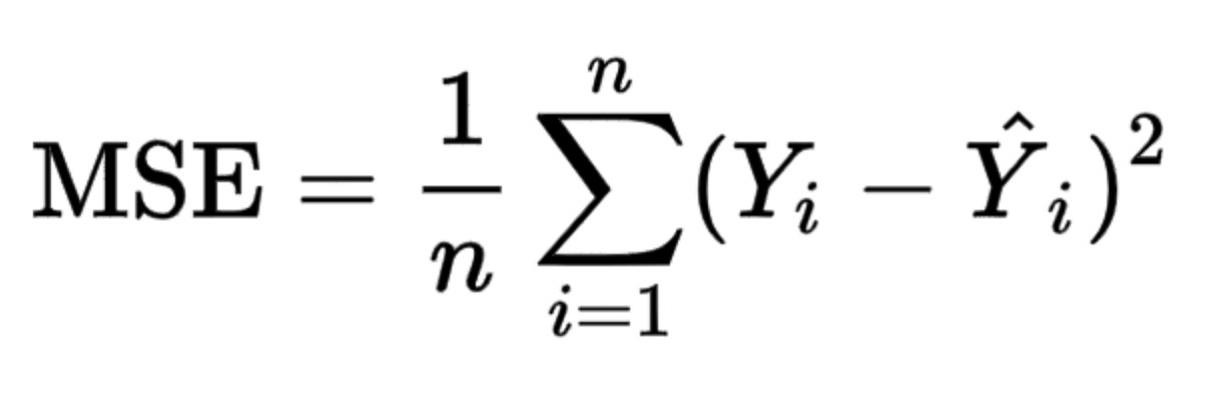
¿Conoces la librería numpy de Python? [Notebook: 3. Librería HYPERLINK "CURSO%20BIG%20DATA/MODULO%205/Notebooks/3\_LibreriaNumpy\_student.ipynb"Numpy](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\MODULO%205\Notebooks\3_LibreriaNumpy_student.ipynb)

DIA 22 – S2 Data Science y Machine Learning

Corrección y repaso del ejercicio anterior.

Cálculo en ML

**Función de pérdidas/coste**

* Es una función que mapea el resultado de un modelo (predicción) en un número real que representa el coste asociado al modelo.
* El problema de optimización busca minimizar dicha función: calcular las derivadas e igualar a cero.
* Si en vez de función de pérdidas tenemos función de ganancias o de probabilidad, se buscará maximizar dicha función.
* Ejemplo: MSE (Mean Square Error o Error cuadrático medio).
* Optimización: buscar los parámetros que minimizan una función de pérdidas.
* Función de pérdidas. En función de 2 parámetros, por ejemplo
* Descenso por gradiente:
* Cuando la optimización de una función no tiene forma cerrada: no podemos despejar los parámetros en la ecuación, hay que hacer uso de algoritmos iterativos que buscan esa solución del mínimo en la función.
* Descenso por gradiente es uno de ello, muy conocido y utilizado en ML.
* Tiene un parámetro muy importante: el ‘*learning rate*’.

Ejercicio

Practiquemos con datasets reales el uso de las librerías más importantes. Leer datasets de ficheros, separar variables de entrada y salida, hacer gráficas e incluso entrenar el primer modelo.

[Notebook: 4. Librerías para el análisis de datos](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\MODULO%205\Notebooks\4_LibreriasParaElAnalisisDeDatos_student.ipynb)

DIA 23 – S3 Data Science y Machine Learning

Partes imprescindibles de un programa para visualizar datos (ejemplo completo de predicción en un modelo de regresión)

* Cargar datos ( read\_csv() ) - *pandas*
* Separar (si necesario) X (entrada) e Y (salida)
* Seleccionar las columnas
* Dividir entrenamiento y test (train\_test\_split – módulo *sklearn*)
* Xtrain
* Xtest
* Ytrain
* Ytest
* Normalizar datos ( standarScaler() ) - *sklearn.preprocessing.StandardScaler*
* Aplicar el modelo: entrenar, validar, testear
* .train
* .scare
* .predict Ŷ

Fundamentos del Aprendizaje Automático (PDF [S3](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\MODULO%205\Transparencias\S3.pdf))

Diagrama de bloques de un sistema de aprendizaje

* La información deseada no se puede acceder directamente, así que hay que usar una serie de variable estadísticas relacionadas.
* El análisis de datos explota esta información estadística para proporcionar resultados precisos…
* …pero algunos errores son generalmente inevitables.
* Mucha inversión está justificada cuando el sistema se opera intensivamente en diferentes datos.

Tipos de modelos de ML

Aprendizaje supervisado

Aprender un mapeo de entradas x a salidas y, dado un conjunto etiquetado de pares de entrada-salida. Aprende una relación 1 a 1 entre observaciones y targets.

**Regresión**: aprender una función que para cada observación prediga un número real o score. Yi es un número real.

**Clasificación**: aprender una función que para cada observación prediga a qué clase, dentro de un conjunto finito y establecido a priori, pertenece esta observación. Yi es una categórica

Aprendizaje no supervisado

Sólo se dan datos de entrada. Objetivo: Encontrar ‘patrones interesantes’ en los datos. No se dispone de un target/salida.

Existen **otros paradigmas** como el aprendizaje semi-supervisado o el aprendizaje por refuerzo.

Regresión (aprendizaje supervisado)

* La variable de respuesta es continua Yi c R.
* Objetivo: predecir valores (número real)
* Ejemplos: linear vs polinómica
* Lineal: y = W0 + W1\*X1 +…+Wn\*Xn
* Polinómica: y = W0 + W1\*X + W2\*X2+…+WnXn

Práctica

[1 – Regresión lineal](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\MODULO%205\2%20-%20Regresi%25F3n\1_RegresionLineal.ipynb)

[2-Regresión KNN](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\MODULO%205\2%20-%20Regresi%25F3n\2_RegresionKNN_student.ipynb)

[3 - 3\_Homework\_Regresion\_student](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\MODULO%205\2%20-%20Regresi%25F3n\3_Homework_Regresion_student.ipynb)

DIA 24 – S4 Data Science y Machine Learning

Clasificación (aprendizaje supervisado)

Varían los modelos, ya que la Y tiene que ser una clase/categoría y NO un número real.

Aprender un mapeo de entradas a salidas, donde Y € {1,…,C}

* Si C=2, clasificación binaria
* Si C>2, clasificación multiclase
* Si las etiquetas no son mutuamente exclusivas, clasificación multi-etiqueta

Necesidad de predicciones probabilísticas:

* Para manejar casos ambiguos, es deseable retornar una probabilidad.
* Corresponde a la moda de la distribución: MAP (Maximum a posteriori)
* Evaluación del riesgo: importante especialmente en medicina y finanzas.

Aplicaciones del mundo real

* Clasificación de documentos y filtrado de spam
* Bolsas de palabras / bag of words
* Clasificación de flores
* Detección y reconocimiento facial

Prácticas

[1\_Regresión logística](file:///H:\Mi%20unidad\CURSO%20BIG%20DATA\CURSO%20BIG%20DATA\MODULO%205\3%20-%20Clasificaci%25F3n\1_RegresionLogistica_student.ipynb)

DIA 25 – S5 Data Science y Machine Learning

Modelos no supervisados

**Agrupamiento o clustering**

**Recordemos**: en los datos no supervisados NO tenemos los datos de salida Y.

* El agrupamiento es quizá el problema más inmediato que se puede resolver con aprendizaje automático.
* El algoritmo de agrupamiento más referenciado posiblemente sea el Kmedias. Divide los datos en K grupos (K es un parámetro que se debe fijar a priori) encontrando los K centroides o prototipos que mejor representan estos grupos.

**Aprendizaje de densidad de probabilidad**

* Caracterización de los datos. La caracterización más completa desde un punto de vista matemático es mediante la función de densidad de probabilidad que modela el proceso generativo de los datos.
* Estimar densidades de probabilidad es complejo, especialmente si no se dispone de información a priori sobre el proceso real responsable de la generación de los datos.
* Frecuente: mejora considerablemente las prestaciones de los uso de densidades aproximadas en la construcción de los modelos mismos.
* Quizá el método más habitual para aprender densidades de probabilidad que se ajusten a una colección de datos definidos en términos de variables numéricas continuas sea la mezcla de Gaussianas.

**Reducción de dimensionalidad**

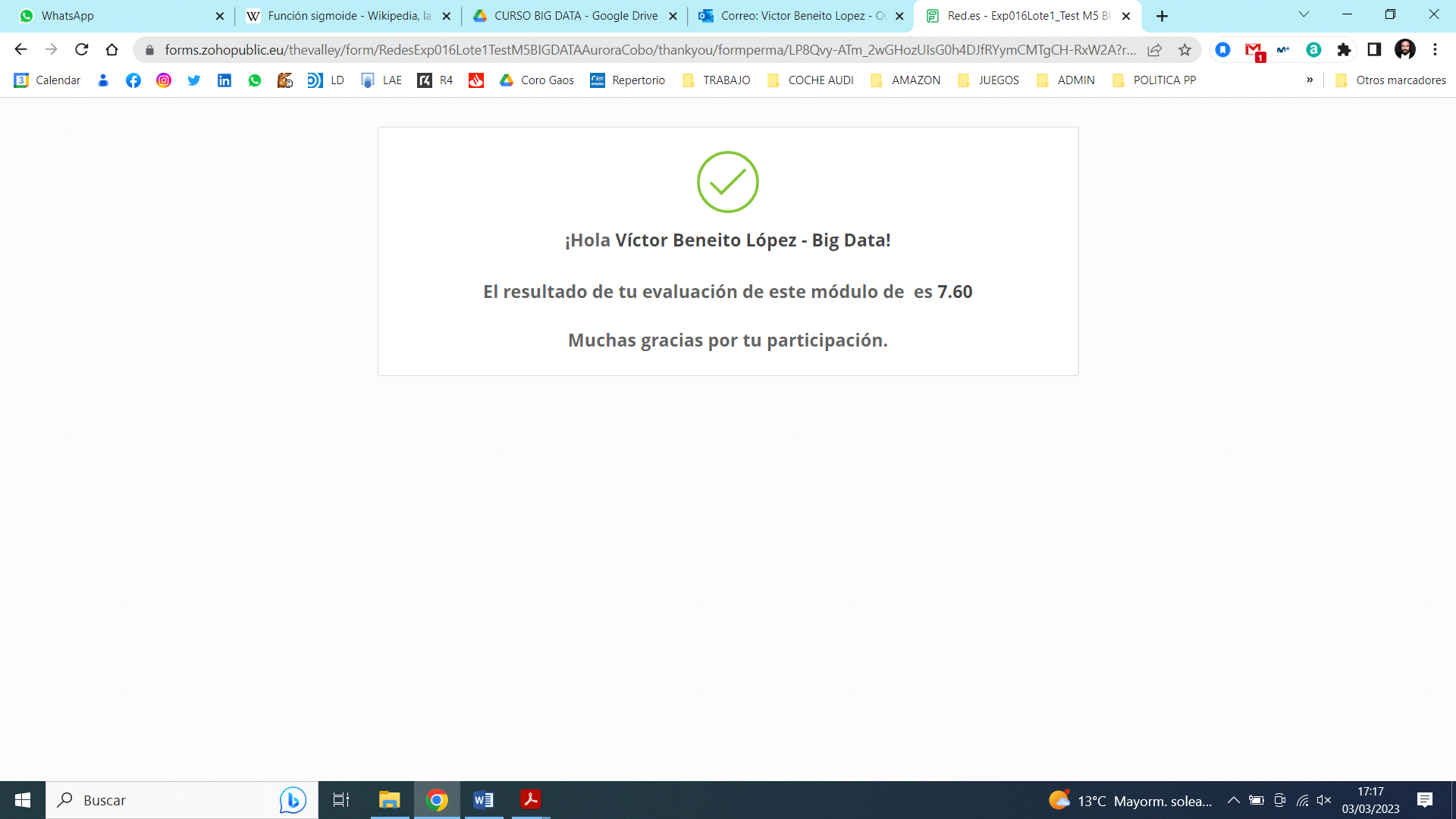
* Simplificar la carga computacional de los optimizadores al emplear datos con menos dimensiones
* Eliminar componentes ruidosas o que no estén alineadas con el patrón que queremos capturar
* El algoritmo más comúnmente empleado para este tipo de problemas es el Análisis en Componentes Principales o PCA.

**Detección de novedad**

* Muy relacionada con el aprendizaje de densidades de probabilidad.
* Evaluar en qué medida cada una de las observaciones puede ser un *outlier*, es decir, es lo suficientemente diferente del resto de observaciones del conjunto de entrenamiento como para considerar que es una observación ruidosa o que no se ha generado con el modelo que suponemos ha generado la mayoría de los datos.
* No hace falta aprender una densidad de probabilidad que represente los datos muy fielmente, sólo nos basta con encontrar una métrica que nos permita decidir si un dato está suficientemente lejos de otras zonas más densamente pobladas de observaciones.

Redes neuronales

**Motivación**

* Posiblemente el método más representativo del aprendizaje automático. Siempre vuelven al foco.
* Son un método general que puede adaptarse a problemas de clasificación, regresión, detección de patrones, etc.
* Inspiración en las redes de neuronas naturales.
* Conexión con neurociencia.
* Una neurona es una unidad elemental de cálculo que recibe unas entradas, las combina linealmente y dispara con una función sigmoide. 

DIA 26 – S1 Arquitecturas Cloud & Big Data

* El lenguaje de **Base de Datos** (relacionales) es **SQL**. Lenguaje de consultas.
* **Python** se utiliza en el 31,2% de los casos como lenguaje de programación.
* Lenguaje de alto nivel, fácil de aprender.
* Expresivo y legible.
* Sintaxis elegante y tipado dinámico y fuerte.
* Multiparadigma.
* Interpretado: se lee línea por línea y se va ejecutando.
* Multiplataforma.
* **IDE**: ¿Qué es un IDE? Software que nos permite desarrollar de una forma más fácil (ej.: Anaconda, Google Colab).
* Autocompletado de código
* Coloración sintáctica
* Navegación de clases, objetos, funciones
* **CRISP-DM**: esquema. 80% del tiempo en Data Preparation
* **Variables**:
* Donde se almacenan y se recuperan los datos de un programa.
* No utilizar palabras reservadas.
* Los nombres tienen que ser significativos.
* En programación, siempre debemos empezar a contar “en 0” (problemática de **índices**).
* **Funciones**:
* Permite definir un bloque de código reutilizable, el cual se puede ejecutar muchas veces dentro del programa.
* Ventaja: reduce el número total de líneas del proyecto.
* Las hay con parámetros o sin parámetros (o argumentos).
* Funciones nativas y propias
* **Librerías**: NumPY, Pandas, Plotly, Jupyter, Matplotlib, etc. Buscador: [www.pypi.org](http://www.pypi.org/)
* Búsqueda de errores para solucionarlos: [www.stackoverflow.com](http://www.stackoverflow.com/),
* Inteligencia Artificial. Machine Learning. Redes Neuronales. ImageNet.
* Google Colab: repaso al contenido de Python
* Operadores aritméticos. Novedad: // para división entera (int), % para el resto de una división.
* Ejercicios básicos

DIA 27 – S2 Arquitecturas Cloud & Big Data

* GitHub: repositorio multifunción. Actualmente se valora como un CV técnico.
* Práctica: crear repositorio en GitHub llamado Curso Big Data y subir todos los ejercicios resueltos, separados por carpetas con un nombre descriptivo del ejercicio. Además, creamos un archivo README.md donde deberemos escribir el lenguaje utilizado y la lista de ejercicios con una descripción acotada
* Ejercicios avanzados
* Ejercicio condicionales

DIA 28 – S3 Arquitecturas Cloud & Big Data

* Colocar archivos en repositorio GITHUB
* Ejercicios avanzados
* Ejercicio Bucles
* Funciones LAMBDA (función anónima). Útil para llamar directamente a las funciones como parámetro.
* Función MAP: aplica una función (1º arg) a todos los elementos de una lista (2º arg). Crea un elemento de tipo "generador", por lo que para su manejo habría que transformarlo a listas.
* Función FILTER: aplica una función de comparación (1º arg) a todos los elementos de una lista (2º arg) filtrando el resultado.

DIA 29 – S4 Arquitecturas Cloud & Big Data

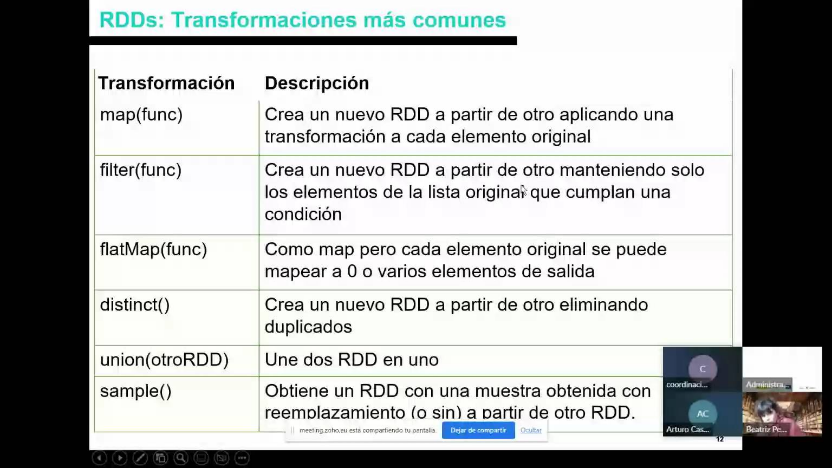
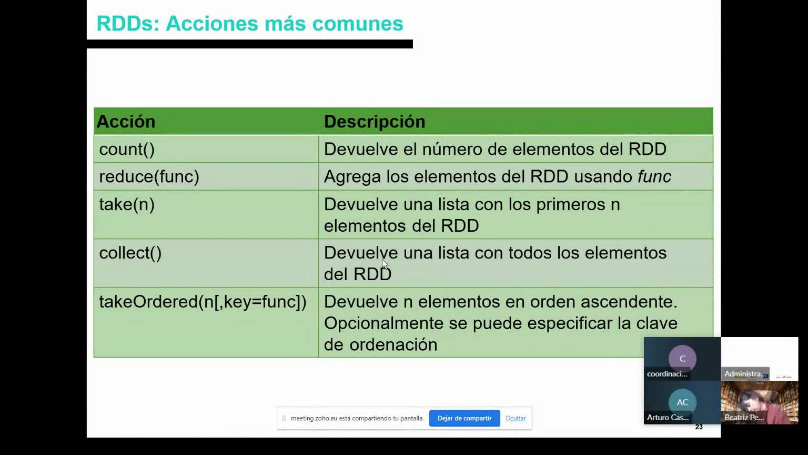
<https://github.com/VicBene/Curso-Big-Data/tree/main/MODULO%206>

* Corrección de la práctica 4
* Concepto FIFO

## Introducción a PySpark: RDDs

<http://ferestrepoca.github.io/paradigmas-de-programacion/paralela/paralela_teoria/index.html>

<https://geekytheory.com/apache-spark-que-es-y-como-funciona/>

* Apache Spark es un motor de código abierto desarrollado específicamente para el procesamiento y análisis de datos a gran escala.
* Objetivo: procesar de manera distribuida grandes cantidades de información.
* *Cluster*: conjunto de ordenadores trabajando entre sí, cooperando para lograr un objetivo.
* Estructura maestro-esclavo (Master Daemon-Worker Daemon). El Master divide la estructura de datos y designa los Workers a cada una de las divisiones mediante variables compartidas (tipo Broadcast)
* **SPARK**: aportaciones
  + Plataforma de computación para clústers
  + Es de propósito general
  + Desarrollo simplificado
  + Trabaja en memoria
  + Rápido
  + Permite trabajo interactivo, streaming
* **DRIVER:** programa que corre el MASTER (main).
* En Apache Spark:
  + Una **sesión** es una conexión de un cliente con el cluster Spark (ej.: una empresa). Sólo se puede tener una sesión abierta en Google Colab.
  + Un **contexto** es un objeto que representa una conexión a un cluster Spark y proporciona un punto de acceso a todas las…
* **RDDs (Resilient Distributed Dataset)**:
  + Colección inmutable y distribuida de elementos.
  + Spark automáticamente distribuye los datos y paraleliza las operaciones.
  + Los RDD realmente cargan colecciones de datos.
  + **Closure**: lo que hay que hacer con los RDDs (acciones a realizar).
* Copia de M6 - 05 - SparkSession Teoría.ipynb
* Copia de M6 - 03- Primer RDD Teoría.ipynb
* Test: 75%
* **Transformaciones**: cambios/tareas que hay que realizar dentro de cada *closure*. Tras aplicar una transformación, obtenemos un nuevo y modificado RDD basado en el original.
* *nuevo\_rdd* = rdd\_ejercicio.map(lambda x : x\*3)
* **Acciones**: cuando el Master les proporciona las transformaciones a realizar a los Workers. Es aplicar una operación sobre un RDD y obtener un valor como resultado, que dependerá del tipo de operación.
* nuevo\_rdd.*collect*() y sc.stop()