**Análisis de negocios para ciencias de datos**

**¿Qué es ciencia de datos y Big data? ¿Cómo afectan a mi negocio?:**

Big Data es un gran volumen de información, las empresas almacenan muchos datos de todo tipo que no sirven de nada si no se utilizan, así nacen las personas especializadas en datos, que convierten la información en acciones.

La información puede ser numéricas, que normalmente se utiliza para disminuir costos y aumentar eficiencia, pero también abra información de categoría. Big data sirve para crear comunicación entre los clientes y la empresa, para saber que se necesita y que se debe mejorar. El trabajo de un científico de datos es crear una solución matemática para un problema de negocio.

**¿Qué tipo de información podemos analizar?:**

Existen diferentes tipos de datos: datos sobre **personas** que generamos nosotros al interactuar con los sitios y que quede registro de esta interacción; el siguiente tipo de datos son las **transacciones** que pueden ser monetarias o no monetarias, las primeras en cuando involucran dinero como cuando se compra algo con una tarjeta de crédito, y las no monetarias pueden ser por ejemplo las compañías telefónicas, que identifican la interacción con la empresa; los siguientes datos son los de **navegación web**, lo que en otras palabras son las cookies, que recopilan información de aquellos que entren a la página; los siguientes datos son **machine 2 machine**, como el GPS y la localización móvil; finalmente están los datos biométricos, que identifican a una persona como un ser único como la huella dactilar.

**¿Cómo crear empresas y culturas data-driven?:**

Hay que seguir 8 pasos:

1. Entender y crear la cultura de datos: hacer que las personas tomen decisiones basadas en datos
2. Recolectar: mientras más información se recolecta, más se entiende y se procesa
3. Medir todo: saber que a que tanto cambia la información
4. Datos relevantes y precisos: saber que datos son más relevantes para el objetivo, pensar que va a ser relevante para la empresa
5. Testear y crear hipótesis: hacer una pregunta específica para un estudio más focalizado (intuicion)
6. Desde los insights de datos a las acciones: por ejemplo, hacer campañas de marketing especifico (confirmación)
7. Cumplir las regulaciones de datos: no tomar decisiones sesgadas
8. Automatizar: pues no hacer la misma investigación a cada rato

**¿Qué es inteligencia artificial y machine learning?:**

IBM desarrollo hace 50 años un algoritmo para retar al mejor jugador del mundo de ajedrez, la primera vez el jugador fue el que gano, pero eventualmente lo hizo el algoritmo, porque podía replicar, entender y predecir las personas. Esto quiere decir que la maquina se perfecciona mientras más juegas. Después se quiso perfeccionar la manera en cómo la maquia realizaba procesos, y nació el Machine learning, posteriormente nació el Deep learning, porque ya podemos sacar datos de imágenes y videos.

El aprendizaje de máquinas tiene unas funciones principales como: detectar fraudes, como en las ventas virtuales, de manera que detecta quien hizo una transacción fraudulenta y bloquean a quienes tienen el potencial de cometer el mismo fraude; la búsqueda web, como cuando buscas algo en Google y después te aparecen anuncios de eso🙄; los anuncios a tiempo real es cuando hacemos una compra, y después hacemos otra cuando ya finalizo la necesidad de consumo, así como los pack de Amazon que aparecen al final de la página del producto😐; el análisis de textos es cuando se trata información categórica y finalmente están los Next Best Action que es saber que está haciendo el cliente, que en fase esta y cuál va a ser su siguiente movimiento.

**¿Qué es deep learning? Análisis de imagen, audio y video:**

Las maquinas ya están configuradas para interpretar, captar y transformar en información las imágenes, a diferencia de las personas que tienen que ir poco a poco. Mientras más imágenes interpreta una máquina, más exacta será su interpretación, esta mecánica también funciona con videos y con música o audios, como en la aplicación Shazam.

En la aplicación de Shazam interpreta el sonido y lo convierte en unos y ceros, y luego los compara con los que tiene almacenados en la base de datos, de la misma manera puede hacernos recomendaciones de canciones que la maquina interpreta que nos gusta. Otro ejemplo de Deep learning son los carros de tesla, que interpreta todo lo que aparece en frente del vehículo y lo convierte para saber si el carro debe seguir o avanzar.

**Flujo de trabajo en ciencia de datos: fases, roles y oportunidades laborales:**

Roles de datos:

Interfaz de usuario gráfica, Aplicación, Sitio web

Descripción generada automáticamente

El ingeniero de datos es el arquitecto de toda la información que se almacena, esta apersona hace la conexión entre el dispositivo y la base de datos. Los analistas, BI parte de la información almacenada en una base de datos y va a extraer la información que el interesa, y automatiza los procedimientos para que cualquier persona dentro de la empresa interprete la información. Los Data Scientist son los científicos de datos, ellos saben extraer información y saben predecir, explica cuál es el pasado y llegan hasta el futura para saber a dónde va la empresa. Finalmente están los Data Translator, que son los que lideran el equipo, pueden ser ingenieros, o especialistas en negocios, se busca alguien que pueda interpretar los datos y facilite la comunicación, además es experta en las necesidades del negocio.

**Herramientas para cada etapa del análisis de datos:**

En el rol del analista y del ingeniero se utiliza el lenguaje SQL que ayuda a extraer información de una base de datos, hace una síntesis de datos y hacen cuadros de control de la operación. El científico de datos utiliza herramientas exploratorias y deductivas basadas en modelos estadísticos y matemáticos, el primer recurso es r que tiene un enfoque más estadístico y el segundo es Python que tiene un enfoque más ingenieril, con estas herramientas se hace un análisis descriptivo y exploratorio, para interpretar la información y predecir, además ambos lenguajes tienen paquetes de apoyo (ggplot2 para gráficos y dplyr para reorganizar la información)o librerías en el caso de Python (pandas y numpy para interpretar datos y organizar información).

Supongamos que en la empresa donde trabajas caen las ventas, y te piden extraer la información de cuantos clientes hubo ayer y cuantos hoy y que herramienta se utilizaría.

**¿Qué es y cómo usar una base de datos relacional con SQL?:**

Toda la información de la empresa queda almacenada en tablas, que tiene columnas y filas, y todo el conjunto de tablas conforman una base de datos. SQL por su parte es la herramienta para extraer datos de las tablas, y para hacer un comando lo primero que hay que hacer es seleccionas la información con la palabra SELECT seguido de la columna que se quiere seleccionar y se deben separa por comas, lo siguiente es indicar donde testa la información con FROM, y con la palabra WHERE puedes especificar mas la información, otra condición es GROUP BY para hacer una condición agrupada por una de las columnas, y finalmente ORDER BY para ordenar los resultados.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Los operadores lógicos funcionan para unir comandos, AND significa ‘y’, OR significa ‘o’ y NOT es para decirle que no incluya.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Las funciones agregadas ayudan a no hacer cálculos como sumas y promedios, AVG es para sacar el promedio, COUNT es para contar, DISTINCT para que no se repitan usuarios, SUM es para sumar, MAX para los valores máximos y MIN es para los valores mínimos.

Texto

Descripción generada automáticamente

**Cómo estructurar queries en SQL:**

Los queris ayudan a extraer información de una tabla. Se tiene una tabla en Excel con la información que se quiere extraer, y hay filtrar por año y valor cuantas bocinas se han vendido.

Interfaz de usuario gráfica

Descripción generada automáticamente

El día tiene que ser de 1 a 31, el mes tiene que ser del 1 al 12; el año tiene que ser uno en el cual la empresa exista; los productos son multimedios, y la pregunta solo es acerca de las bocinas y su valor; la moneda tiene que ser del país del cual se está analizando.

El código empieza con SELECT para seleccionar la información, y se quiere hacer un recuento único de números de id así que se le pone COUNT (DISTINCT id), después se le indica en donde está la información FROM y se le pone la tabla VENTAS\_2020, después se le colocan las condiciones específicas WHERE, y las condicionas separadas por AND. El comando quedaría de la siguiente manera:

SELECT COUNT (DISTINCT id) FROM VENTAS\_2020 WHERE Producto=’Bocina’ AND Valor>600 AND Año>=2019

El resultado de este comando va a ser 2. Otro ejemplo para hacer la suma de todos los auriculares vendidos podría ser:

SELECT SUM Valor FROM VENTAS\_2020 WHERE Producto=’Auriculares’

**Conflictos y retos actuales sobre la ética y tratamiento de datos:**

En la aplicación faceapp existen muchas personas las cuales han brindado su imagen para poder usar los filtros y la aplicación a recolectado información sobre los vectores faciales de muchas personas, gracias a esto la aplicación es capaz de predecir cómo se verán las personas si fueran del género opuesto o si fueran más viejas o jóvenes basándose en los datos almacenados.

Esto genera muchas preguntas como si vulnera o no los derechos de las personas, y si es ético usar esa información, que amenazas puede representar, se pueden publicar esas imágenes y la de los compañeros.

Otra aplicación en la que salen preguntas similares es Tinder, que almacena gustos, género, edades. etc. gracias a las indicaciones que se le da a la plataforma y los datos que esta almacena a través de la interacción de sus usuarios con la aplicación y con otras personas a través de esta, aquí el dilema ético es si la plataforma debería o no mostrar toda la información al momento de mostrar preferencias.

La tecnología avanza a un ritmo mas veloz independientemente de la ley, así que no siempre la tecnología es ética, y las leyes que se hacen al respecto tampoco.

**Aplica técnicas de storytelling para convertir problemas de datos en historias:**

Queremos que todos entiendan y que a todos les importe, por tanto, lo primero es estructurar el problema de negocio, hay que definirlo cual es el problema, que se quiere resolver, cual es la posible solución, y finalmente cual es el alcance.

El problema puede surgir de cualquiera de las áreas del negocio. Imaginemos que se acerca a nosotros una persona de servicio al cliente, y genera la siguiente estructura:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

La solución que se plantea en vista de que no se puede identificar a las personas que llaman, es crear script para saber quienes son los top offenders, de esta manera se entiende porque esta sucediendo esto, y finalmente se definen acciones para prevenir esta tendencia y que se pueda aplicar de forma muy amplia, aunque también a una escala muy macro pueden hacerse distinciones por ciudad o regiones, y esa distinción dependerá de las características y condiciones, si por ejemplo es una empresa de envíos, la distinción dependerá de que tan fácil es llegar a determinada región, dentro del alcance también está la distinción temporal, se quieren ver los ambientes y porque los clientes se quejan. Así que el alcance tendrá distinciones mes a mes y por ciudad, además deberá ser automatizable.

**Cómo estructurar un caso de negocio:**

Ahora vamos a minimizar el problema a través de una estructura de hipótesis. El ‘que’ es el problema de negocio, el ‘porque’ es para sintetizar todos los mensajes en categorías, el ‘como’ va a ser la estrategia de cómo se va a organizar todo el análisis.

Interfaz de usuario gráfica, Aplicación, Sitio web, Teams

Descripción generada automáticamente

**Análisis cuantitativo en un caso de negocio:**

Un análisis cuantitativo es el análisis de los números, lo primero que se hace es identificar las variables numéricas y luego identificar cuáles son relevantes.

La primera pregunta es quienes son los que contactan, de manera que lo primero que hay que hacer es descargar la información de la manera más genérica posible, de manera que se busca cualquier cliente que haya hecho al menos una queja. Después se identifican patrones de comportamiento y variable significativas, en este caso el margen operativo neto mide le rendimiento a partir de todo lo que hace el cliente, esto es lo que vendí multiplicado por el precio de lo que vendí, menos el precio de los mensajes y menos todo el dinero devuelto.

La definición es el objetivo, encontrar una segmentación en función de la rentabilidad, o sea si un cliente compra más, dejo que se queje más. También hay que identificar quienes son top offenders, y quienes no, o sea que tipo de clientes son.

Diagrama, Escala de tiempo

Descripción generada automáticamente

Después se realizará un mapeo, siempre hay que tener la referencia de cuál es el comportamiento normal. Se identifica que según el promedio de compras sube, también lo hace el promedio de quejas, pero no lo hace en la misma proporción, la ratio de contacto es la división del número de quejas entre el número de compras:

Tabla

Descripción generada automáticamente

La primera hipótesis formulada es que los usuarios que hacen más compras se quejan menos porque ya entienden cómo funciona el servicio, el siguiente grafico corrobora la hipótesis, porque el eje x es el ratio de contacto, si está en el uno quiere decir que hico una queja y una compra, o sea que se queja el 100 de las veces, es un gráfico de densidad. En el grafico se observa que el grafico de los usuarios viejos casi no se quejan, pero los usuarios viejos se quejan mucho más ya que la gráfica está más pegada al uno. Los top offenders no son usuarios nuevos,

Tabla

Descripción generada automáticamente con confianza media

Lo que se busca es impactar el mínimo de personas posible y el máximo de quejas, y encontrar el número de top offenders de cada categoría. Se busco el mínimo porcentaje de la población, con el máximo de quejas posible

Tabla

Descripción generada automáticamente

**Análisis cualitativo en un caso de negocio:**

Lo primero es identificar las variables cualitativas para el análisis categórico. En el ejerciese seria de la siguiente manera:

Captura de pantalla de un celular con letras

Descripción generada automáticamente

La primera identificación es la clusterizacion de todo el problema en solo 4 categorías, después se hace lo contrario, hay que identificar palabras para ver los motivos de contacto en particular y se hará de manera geolocalizada

Gráfico

Descripción generada automáticamente

Se identifica que la mayoría contacta a servicio al cliente es para buscar un retorno del dinero, y la segunda razón es por preguntas.

**Fusión cuanti-cualitativa en un caso de negocio:**

Hay que resolver de manera conjunta la información cuantitativa y cualitativa para sacar copnclusiones. Hay que ver porque se quejan cada uno de los tipos de Top Offenders.

Imagen que contiene Tabla

Descripción generada automáticamente

Esto es un hitmap o un mapa de calor, el color azul es donde menos hay una concentración de usuarios que se están quejando, los regulares son lo que menos compran y los que mas buscan dinero. La matriz es opuesta, lo que quiere decir que la razón por la cual se queja un regular es por la que menos se quejaría un oro. Así se identifica el comportamiento de cada tipo de usuario.

**¿Qué es minería de texto? ¿Cómo usarla para obtener información adicional?:**

Twitter tiene su base de negocios en mensajes. Hay que leer los mensajes de los clientes.

Escala de tiempo

Descripción generada automáticamente con confianza baja

En el caso de los clientes regulares se encontraron tres preguntas relacionadas con la devolución de dinero, los clientes de bronce tienen un comportamiento muy similar a los regulares, pero aparece el estado del producto, sin embargo, en el caso de los plata aparece el tema de facturas, están solicitando una factura y están comprando mucho, esto sucede porque son una cuenta empresarial, y finalmente los oro tienen un log in, esto es porque no pueden acceder al servicio, ya que son varios empleados por la cuenta empresarial y se les olvida la contraseña.

Con este ejemplo, en el caso de log in es un problema que es culpa del usuario y no de la empresa que presta el servicio.

**Variación de comportamientos a partir de la geolocalización:**

Hay que saber cómo se comportan según las ciudades, para eso hay que crear un grafico de quejas vs compras.

Imagen que contiene Diagrama

Descripción generada automáticamente

La línea de 45 grados indica un comportamiento normal, y las ciudades que estén lejos de esa línea son aquellas que muchos top offender o muy pocos. En el grafico no aparecen algunas ciudades con demasiada población porque hacen muchas compras y serien un punto muy alejado de los demás, así que simplemente se valido si en el caso de estas ciudades el comportamiento era normal, para saber si se pueden quitar del gráfico.

Las ciudades con más top offender están mas cerca al eje x como Puebla, y el caso contrario sería Mexicali.

**Acciones, algoritmos y toma de decisiones según los resultados del análisis:**

Ahora hay que convertir toda la información obtenida en una estrategia para solucionar el problema.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media

La minería de texto es el análisis de las palabras, después es aprender quienes son nuestros usuarios, los árboles de decisión ayudan a tomar decisión partiendo de una hipótesis a o b para predecir la mejor estrategia, le estadística bayesiana es muy útil para identificar patrones que se comportan de manera conjunta, las cadenas de montearlos son probabilidades conjuntas (estos son estudios que se hacen después del análisis, son métodos de validación).

Lo primera acción tomada es identificar quienes eran los top offenders, y añadirlos a la base de datos, la segunda opción es advertir a las personas para evitar que se comporten de esa manera pero en el caso de los oros las dudas eran diferentes, así que el servicio puede ser mas personalizado, y en los casos más extremos la opción es bloquearlos, finalmente todos los modelos necesitan una validación con A/B tests, esto quiere decir que tenemos la estrategia a (una llamada) y la estrategia b (un mensaje), después de usar la estrategia a con la mitad de los usuarios y la estrategia b con la otra mitad, se evaluara como se comportan los usuarios y cual es la estrategia más efectiva.