Ciencia de Datos

José Alejandro Contreras Obregón

Colegio Panamericano

Senior Project

Prof. Gloria Gomez

29 de Mayo de 2024

Tabla de Contenidos

Introducción

“El conocimiento es poder.” Esta corta frase ha sido repetida durante siglos, y hoy es más cierta que nunca. La llegada de la era digital ha traído consigo una nueva revolución, como lo hicieron la fiebre del oro o el descubrimiento del petróleo en tiempos pasados, hoy son los datos y la información derivada de ellos los motores que mueven el mundo. Son una nueva divisa comercial para las empresas, son catalizadores para la investigación y el fundamento para una mejor toma de decisiones.

A raíz de su importancia, ha venido en auge un nuevo campo para suplir la necesidad de combinar una tecnología con capacidad analítica sin precedentes y unos conjuntos de información abrumadores para una simple mente humana. Esa disciplina tiene nombre, y se llama ciencia de datos.

Mediante este proyecto se busca se busca llevar a cabo un estudio que combine los aspectos generales de esta área del conocimiento y su caso de uso en el mercadeo de esta era digital. En los ocho capítulos que se presentan a continuación, se dará a conocer una pregunta problema que servirá como ancla a una investigación que recopile la normativa colombiana, la teoría de la problemática en cuestión y la experiencia del estudiante a la hora de llevar lo aprendido a la práctica.

Capítulo: 1 Descripción del Proyecto

La ciencia de datos como disciplina se enfoca en combinar la matemática, estadística y programación, para ejecutar métodos de análisis algorítmicos con el propósito de identificar patrones en conjuntos de datos, a fin de sacar conclusiones que permitan al profesional tomar decisiones de manera informada y desarrollar un planeamiento estratégico. En un entorno laboral, se espera de un científico de datos la capacidad de comunicar efectivamente sus hallazgos, de tal manera que sean fácilmente digeribles para otros miembros de su equipo de trabajo, generalmente mediante herramientas de software dedicadas a la visualización de datos sintetizados.

Teniendo esto en cuenta, el proyecto girará en torno la aplicación de esta área del conocimiento al mundo de los negocios y el emprendimiento, con base en la siguiente pregunta: ¿Qué técnicas de la ciencia de datos se emplean para llevar a cabo un estudio de mercado, a fin de incrementar la adquisición, retención y desarrollo de clientes actuales y potenciales en una empresa?

Capítulo: 2 Objetivos

Objetivo General:

Conocer las técnicas de la ciencia de datos que se emplean para llevar a cabo un estudio de mercado, a fin de incrementar la adquisición, retención y desarrollo de clientes actuales y potenciales en una empresa.

Objetivos Específicos:

- Investigar las limitaciones impuestas por la legislación colombiana con respecto a la recolección, el uso y la publicación de datos del público general.
- Reconocer el impacto de la gobernanza de datos en las jerarquías laborales, para asegurar su calidad y el subsecuente análisis.
- Investigar las regulaciones de la accesibilidad a bases de datos públicas y privadas y su debido proceso.
- Tener nociones elementales de conceptos y métodos estadísticos.
- Observar la implementación de algoritmos de aprendizaje de máquina e inteligencia artificial en el análisis de tendencias para desarrollar modelos predictivos.
- Estudiar los lenguajes de programación esenciales.
- Conocer las herramientas de software utilizadas por los profesionales.
- Explorar a profundidad las técnicas de limpieza para conjuntos de datos pequeños o de Big Data.
- Determinar los principios fundamentales para perfilar de manera óptima al consumidor, en concordancia con sus preferencias y necesidades.
- Analizar los conflictos ético-morales o de intereses que conlleva el manejo de datos y perfilamiento del consumidor.
- Estudiar la teoría e implementación detrás de las principales estrategias de mercadeo digital.

- Desarrollar habilidades de comunicación, ya sea para la socialización de resultados en el contexto laboral, o la implementación de estos mismos frente al público general mediante el mercadeo.
- Conocer el ambiente laboral de la Clínica Materno Infantil San Luis.
- Mejorar mis habilidades de comunicación.
- Determinar si hay vocación por la profesión de científico de datos.
- Obtener retroalimentación de mi asesor profesional.
- Explorar mis capacidades y aptitudes para la carrera en cuestión.

Capítulo 3: Marco Contextual

- **Razón Social:** Clínica Materno Infantil San Luis
- **Ubicación:** Cra. 26 #48-56, Bucaramanga, Santander
- **Sector Económico:** Salud privada.
- **Misión:** “Somos una organización que preserva la vida de la mujer y el niño, a través de la prestación de servicios de salud humanizados, de alta calidad y confiabilidad.”
- **Visión:** “Ser referente nacional en la atención integral y humanizada de la mujer y el niño al año 2029.”
- **Cobertura:** Regional: Nororiente colombiano.
- **Constitución:** Sociedad Anónima (S.A)
- **Usuarios:** Mujeres y niños necesitados de atención médica afiliados a las EPS Sanitas, SURA y Salud Total.
- **Servicios:** Atención especializada para la mujer y el niño en las áreas de urgencias, hospitalización, consulta externa, cirugías, cuidados intensivos, laboratorio clínico, radiología, ginecología y el concurso de diversas especialidades pediátricas que incluyen pero no se limitan a: hemato oncología, neumología, infectología, endocrinología, reumatología, nefrología.
- **Historia:** La Clínica Materno Infantil San Luis nace como iniciativa de los pediatras Fidel Rey, Álvaro Africano y Reynaldo Rey, quienes junto a varios de sus colegas del área metropolitana, consideraban imperativa la construcción de un edificio para hospitalizar a sus pacientes más delicados. La institución fue inaugurada el 2 de noviembre de 1983, y desde entonces ha experimentado dos ampliaciones en 1986 y 1997, además de la construcción de nuevos edificios en 2007, 2008 y 2015.

Capítulo 4: Marco Legal

Este proyecto busca estudiar el campo de las tecnologías de información mediante el área de la ciencia de datos. Su eje central, los usos de la computación en el análisis del consumidor dentro del mundo corporativo, de los negocios y el emprendimiento, con el propósito de mejorar su adquisición, desarrollo y retención. El propósito del capítulo es llevar a cabo una exploración de la legislación pertinente a un profesional de esta disciplina, para proporcionar al estudiante un entendimiento de las normativas que rigen su carrera en la República de Colombia.

Decreto 393 de 1991, por el cual se dictan normas sobre asociación para actividades científicas y tecnológicas, proyectos de investigación y creación de tecnologías.

Delimita las condiciones y el protocolo mediante el cual una entidad pública puede colaborar formalmente con particulares a través de diferentes sociedades o convenios especiales de cooperación, los cuales se destacan por la no creación de personas jurídicas. Se establecen los propósitos que pueden adoptar dichas asociaciones, en el caso de los convenios especiales, se fijan un conjunto de reglas para asegurar su funcionamiento. Se dispone del derecho privado como régimen legal, se permite la asociación con otras instituciones públicas y se habilita la posibilidad de la compra de acciones con respecto a los socios de la Nación en este tipo de acuerdos.

Un profesional debe ser conocedor de este decreto en dado caso de que tenga un cargo público o desee llevar a cabo algún tipo de colaboración con algún organismo gubernamental. Esto le permitirá a dicho individuo evaluar los parámetros y ajustar los detalles de dicha alianza para poder asegurar un cumplimiento de aquello dictaminado por el decreto y así efectuar una asociación satisfactoria.

Ley 1266 de 2008, Por La Cual Se Dictan Las Disposiciones Generales Del Hábeas Data y Se Regula El Manejo De La Información Contenida En Bases De Datos Personales, En

Especial La Financiera, Crediticia, Comercial, De Servicios y La Proveniente De Terceros Países y Se Dictan Otras Disposiciones.

Promulga un edicto que se centra en establecer las bases jurídicas para garantizar el Hábeas Data, el cual encapsula los derechos constitucionales de una persona natural sobre su información y su intimidad. Erige los cimientos legales que desarrolla la ley 1581 de 2012, con respecto a la administración de datos por parte del usuario y los operadores. Los contenidos de este documento se encargan de hacer la distinción entre los tipos de datos, públicos, personales y privados. Imparte las políticas de recopilación y presentación de estos registros en concordancia con el principio de interés público.

Un experto en esta profesión debe ser conocedor de los contenidos de este precepto para ejercer su cargo de tal manera que sea beneficioso para la sociedad. Al seguir los procedimientos delimitados y distinguir los tipos de datos definidos por este, podrá procurar la calidad de estos, respetar la privacidad de los ciudadanos y evitar ser sujeto a medidas punitivas por la rama judicial debido al incumplimiento de la ley 1266 de 2008.

Ley 1273 de 2009 por medio de la cual se modifica el código penal, se crea un nuevo bien jurídico tutelado - denominado "de la protección de la información y de los datos"- y se preservan integralmente los sistemas que utilicen las tecnologías de la información y las comunicaciones, entre otras disposiciones.

Reforma ciertos aspectos del Código Penal de Colombia a fin de establecer las medidas punitivas correspondientes a nuevos crímenes informáticos del mundo contemporáneo. Se centra principalmente en complementar el artículo 269, y explora problemáticas que incluyen pero no se limitan al uso de software malicioso, la violación de datos personales, el acceso abusivo a un sistema informático entre otros. Enfatiza en las condiciones de mayor punibilidad para los delitos en cuestión y empodera a los jueces municipales para procesar infracciones de esta índole.

Se debe tener un conocimiento de aquello englobado por la presente ley a modo de actuar en concordancia con aquello que delimita el Código Penal en el país. Siguiendo el principio de Ignorantia juris non excusat, es de suma importancia ser conocedor de las regulaciones y consecuencias establecidas por el gobierno para evitar una mala práctica a la hora de desempeñar un trabajo en el ámbito profesional.

Decreto 1389 De 2022 Por El Cual Se Adiciona El Título 24 A La Parte 2 Del Libro 2 Del Decreto Único 1078 De 2015, Reglamentario Del Sector De Tecnologías De La Información Y Las Comunicaciones, Con El Fin De Establecer Los Lineamientos Generales Para La Gobernanza En La Infraestructura De Datos Y Se Crea El Modelo De Gobernanza De La Infraestructura De Datos

Establece medidas a ser tomadas para garantizar el futuro de la gobernanza de datos en el país, de acuerdo con su influencia sobre el Plan Nacional de Infraestructura de los mismos. Segmenta su modelo en niveles estratégicos, tácticos y operativos que promuevan las políticas de innovación y digitalización de los mismos mediante lineamientos técnicos, la involucración de partes interesadas y mecanismos de participación para los usuarios. Delega la responsabilidad del cumplimiento de estas al Comité Nacional de Datos y demás entidades o servidores públicos adyacentes.

Es de beneficio reparar en los contenidos de este decreto, puesto que da un contexto del sistema que será necesario conocer para tramitar cierto tipo de procesos en un futuro próximo. Adicionalmente, la perspectiva que otorga sobre el panorama nacional de esta industria puede ser de una gran ayuda para abordar las problemáticas y tener un proceso efectivo de adaptación en los años venideros; algo que será de utilidad para el estudiante en dado caso que decida estudiar y ejercer esta carrera.

Ley 1480 De 2011 Por Medio De La Cual Se Expide El Estatuto Del Consumidor Y Se Dictan Otras Disposiciones

Tiene como objetivo proteger los derechos del consumidor mediante la imposición de condiciones que rigen el comercio de un producto o servicio. Fomenta la educación del comprador y otorga protecciones especiales a los menores de edad. Dispone de estándares mínimos en seguridad y calidad, así como define las reglas contractuales de garantía, devolución u otro reparo al cliente. Da estructura a los contenidos obligatorios de la información, publicidad y es altamente estricta con las prohibiciones de actos engañosos o hechos en mala fe.

Su importancia radica en que el área de investigación del proyecto gira en torno a la aplicación de la ciencia de datos frente al consumidor. Se debe tener en cuenta a la hora de perfilar y mercadear, para no vulnerar al usuario. Es esencial llegar a un compromiso óptimo entre la efectividad de la publicidad, y la ética detrás de la representación de esta misma o el uso de modelos predictivos y de inteligencia artificial enfocados en el mercado objetivo. Similarmente, un profesional debe ser muy cuidadoso con sus análisis del producto, para evitar tergiversar la representación de este, incluso si se hace de manera accidental.

Ley 1581 De 2012, Por La Cual Se Dictan Disposiciones Generales Para La Protección De Datos Personales.

Tiene como fin regular el acceso a información personal y bases de datos para su tratamiento al estatuir principios fundamentales que habilitan al Titular o algún tercero para ejecutar una consulta, ratificación o procesamiento de sus registros, de acuerdo con la confidencialidad, transparencia, edad legal entre otros. El acta se encarga de delimitar el debido proceso de obtención y manipulación de archivos por parte de entidades externas, con base en los deberes procesales del responsable frente al Titular. Esta establece a la Superintendencia de Industria y Comercio como entidad supervisora y le otorga el poder de imponer sanciones a todos aquellos que incumplan con las normativas impuestas por este documento.

La ley en cuestión es de gran relevancia para un científico de datos, puesto que estos profesionales enfocan la mayor parte de su tiempo en desglosar conjuntos de información pertenecientes al Big Data, que en algunos casos son recipientes de contenidos sensibles. Por consiguiente, para llevar a cabo sus proyectos dentro del margen de lo legal, es imperativo que los expertos en esta área sean conocedores de sus limitaciones jurídicas y sus obligaciones ante las corporaciones o ciudadanos que analizan.

Ley 1712 de 2014 Por Medio De La Cual Se Crea La Ley De Transparencia Y Del Derecho De Acceso A La Información Pública Nacional Y Se Dictan Otras Disposiciones.

Define qué tipo de datos son considerados públicos y aclara el deber de un sujeto obligado, el cual puede ser un organización o una persona, ya sea natural o jurídica, que la represente, de difundir contenidos de esta índole. Presenta las bases fundamentales, que guían el comportamiento de dichos sujetos obligados, de manera bastante favorable hacia quien realiza las pesquisas. Reglamenta la gestión documental y los métodos de accesibilidad a la información disponible. Aclara excepciones a la divulgación de esta con base en el interés general, los derechos civiles y la naturaleza no-divulgativa de algunos documentos.

Un profesional de esta área necesita de datos públicos para efectuar un análisis satisfactorio en una gran cantidad de sus proyectos. Consecuentemente, debe estar familiarizado con las restricciones, los mecanismos de obtención y las políticas de archivación electrónica para facilitar la implementación de los softwares empleados durante la descomposición de los mismos.

Ley 2162 de 2021 por medio de la cual se crea el ministerio de ciencia, tecnología e innovación y se dictan otras disposiciones.

Instaura el Ministerio de Ciencia, Tecnología e Innovación mediante la fusión de Colciencias en este nuevo organismo gubernamental. Se establecen una serie de objetivos que buscan fomentar el progreso de estas áreas en un futuro próximo a través de la política pública

y una planeamiento estratégico que cimiente el Sistema Nacional de Ciencia, Tecnología e Innovación (SNCTI). Adicionalmente se establece todo aquello que concierne la estructura administrativa y se delimitan las funciones a desempeñar por parte del ministerio. Se fijan la sede principal, los contratos, el patrimonio y demás aspectos claves para el funcionamiento adecuado de esta institución.

Se debe conocer esta ley para tener una contextualización sobre el ente regulador del campo al cual gira en torno esta profesión para poder entender el futuro desarrollo de la misma en el país. Es importante comprender su organización, protocolo y propósito para alinearse óptimamente con sus objetivos en pos de contribuir al mejoramiento del sector y en dado caso de que se necesita tramitar algo con esta entidad.

Capítulo 5: Marco Teórico

El proyecto se centra en la ciencia de datos como área de investigación, un campo interdisciplinario relativamente nuevo, que tiene como objetivo aplicar la programación y la estadística para otorgar un enfoque moderno a la resolución de problemas mediante la toma de decisiones basada en datos. Esta investigación hace énfasis en sus usos dentro de un contexto empresarial, y puntualmente busca estudiar su importancia a la hora de examinar a profundidad los consumidores de una corporación.

En este capítulo, el estudiante tiene como propósito abarcar una amplia gama de temas que incluyen, pero no se limitan a las nociones elementales para un profesional de esta disciplina, las herramientas y algoritmos específicos de computación, y los conceptos foráneos a esta, como la mercadotecnia. Concretamente, busca responder la pregunta: ¿Qué técnicas de la ciencia de datos se emplean para llevar a cabo un estudio de mercado, a fin de incrementar la adquisición, retención y desarrollo de clientes actuales y potenciales en una empresa?

Ciencia de Datos

Especialidad de la computación que trabaja en conjunto con la estadística para manipular grandes conjuntos de datos a fin de revelar nuevos conocimientos y optimizar el rendimiento empresarial. Se destaca por su versatilidad de uso en diversos ámbitos laborales y suele requerir de colaboración con los distintos departamentos de una compañía para derivar conclusiones específicas (*¿Qué es la ciencia de datos?*, s.f.)

Historia de la Ciencia de Datos

Se origina como concepto en 1962, y busca transformar el análisis de los mismos mediante el establecimiento de una nueva visión con respecto al porvenir del campo. El término es usado por primera vez en 1974, y sólo tres años después se instaura la Asociación

Internacional de Computación Estadística. No es hasta 2001, que se le reconoce formalmente con la esperanza de suplir la necesidad de exploración en cuanto a sus aplicaciones en un mundo tecnológico (Muñoz y Ramírez, 2023).

Desde entonces, esta carrera se ha popularizado en ambientes académicos y laborales, y ahora se le denomina como el empleo más sexy del siglo XXI. Actualmente se destaca por su alta demanda profesional, por lo que las opciones de estudio trascienden las maestrías y ahora incluyen la opción de pregrado. A nivel nacional existen cinco programas de esta índole en los siguientes institutos de educación superior: Pontificia Universidad Javeriana, Universidad Externado de Colombia, Universidad de la Sabana, Universidad del Norte y Fundación Universidad Compensar.

Herramientas de Software

Un científico de datos debe ser diestro en el manejo de sus instrumentos para llevar a cabo su labor eficientemente. Es de suma importancia adquirir un entendimiento de las funcionalidades de los programas más usados en el mercado, para poder discernir entre ellos e identificar aquel que se adecue óptimamente a la resolución del problema en cuestión.

Entornos de Desarrollo Integrados (IDE)

Son aplicaciones de software que posibilitan el desarrollo de proyectos al reunir las principales herramientas de un programador en un solo lugar. Destacan por su versatilidad ya que permiten editar el código, compilarlo y correrlo para detectar errores en su funcionalidad (¿Qué es un IDE?..., s.f.).

Programas de datos

Permiten al profesional sacar el máximo provecho a su trabajo, pues tienen las facultades necesarias para mejorar la organización, descomposición, presentación y modelamiento de su material para extraer conclusiones poco aparentes.

Excel.

Lanzado en 1985 por Microsoft, es uno de los principales programas de hojas de cálculo a nivel mundial. Se ha popularizado en ámbitos empresariales por su facilidad a la hora de organizar datos y con ellos realizar diversas operaciones matemáticas y gráficos (Los Editores de Encyclopaedia Britannica, 2024).

Cuenta con una función integrada expresamente diseñada para el análisis de datos que debe ser habilitada por los usuarios. Es de suma utilidad para la analítica de datos a baja escala y provee las básicas de la estadística descriptiva así como la implementación de escenarios, hipótesis, regresión entre otros. Herramientas como las tablas dinámicas otorgan un gran valor añadido a Excel, pero son incapaces de competir contra softwares más robustos como R, Python y Tableau (Marin, 2023).

SPSS.

Es un servicio de suscripción pago para el análisis de información cualitativa, numérica y ordinal. Se divide en el IBM SPSS Modeler y SPSS Statistics los cuales funcionan en conjunción para crear una de las herramientas más robustas del mercado mediante la generación y resolución de hipótesis (SPSS Software | IBM, s. f.).

Tableau.

Fundado en 2003 como un proyecto de Stanford por Stolten, Hanrahan y Chabot, Tableau es una compañía que busca promover el entendimiento y uso de datos en la toma de decisiones mediante su software. Actualmente, su plataforma destaca por su producción de gráficos, interactividad, comunidad y su portal educativo (What is tableau?, 2024).

Lenguajes de Programación

Es la herramienta esencial para escribir un código, mediante la cual un desarrollador transmite una serie de directrices al ordenador. Actualmente, se cuenta con una gran variedad de lenguajes, de bajo y alto nivel, los cuales se diferencian por su sintaxis, semántica, propósito y funcionalidad. A continuación se explorarán algunos de los programas más utilizados por los científicos de datos (Chavez, 2023).

Python.

Creado en 1989 por Guido Van Rossum, se ha posicionado como uno de los lenguajes más prominentes de alto nivel en los últimos años. Su éxito se debe en parte a su facilidad de uso, la cual proviene de una sintaxis poco compleja y un gran número de librerías aplicables a diversos proyectos. Destaca por su versatilidad y sus facultades para desarrollar algoritmos de aprendizaje automático (*¿Qué es Python?*..., s.f.).

SQL.

Es un lenguaje de programación que tiene su nicho en las bases de datos relacionales por su capacidad de definirlas, consultarlas, manipularlas y optimizarlas. Su popularidad ha llevado a que sea estandarizado por la ISO, así promoviendo un panorama más consistente y fácil de trabajar. Hoy en día cuenta con diversas variaciones como lo son MySQL y NoSQL que tienen como cimiento SQL, pero cumplen funciones ligeramente diferenciadas (*¿Qué es SQL?*..., s.f.).

Lenguaje R.

Es un software expresamente desarrollado para facilitar el análisis estadístico de datos. Similarmente a los otros lenguajes, es óptimo para desglosar conjuntos de cifras u otra información, no obstante, sobresale por su potencial gráfico integrado. Adicionalmente, es una herramienta cuyo valor añadido se encuentra en su código abierto, comunidad y su habilidad para combinarse satisfactoriamente con alternativas como C + +, Python, entre otros (*Lenguaje R…,* 2019).

Java.

Está dividida en el lenguaje, API y la Java Virtual Machine lo cual la convierte en una de las plataformas informáticas más portables entre los distintos sistemas operativos. Se centra en la codificación en torno a objetos y es reconocida por su seguridad y la amplitud de bibliotecas y recursos en general. Es utilizado para trabajar con Big Data e inteligencia artificial *(¿Qué es Java?...,* s.f.).

Scala.

Desarrollado en 2004, es un lenguaje orientado a objetos que comparte muchas similitudes con Java, entre ellas el uso de la Java Virtual Machine. Como indica su nombre, este software se distingue por sus facultades para acoplarse a estructuras de datos de distintas escalas. Presentemente, Scala predomina en el mundo de los negocios y el emprendimiento (Ortega, 2023).

Algoritmos

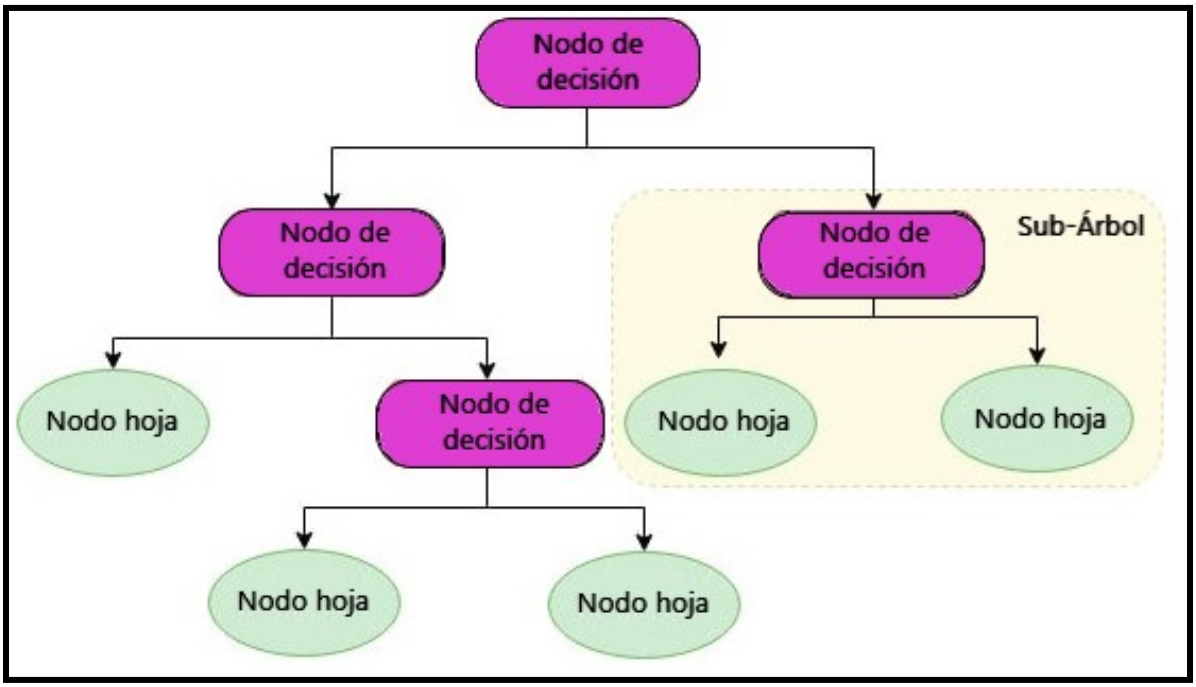
Es cualquier procedimiento en el mundo informático o la vida cotidiana que contenga secuenciación, selección e iteración. Es decir, tiene un conjunto de instrucciones que sigue con un orden definido, utiliza condicionales para determinar una respuesta o línea de código que ejecutar, y se puede repetir indefinidamente o de acuerdo a un condicional (E. J. Sepulveda, comunicación personal, 2021).

Clasificación

Se manifiestan principalmente como árboles de decisión, cuyo propósito es ramificar las diferentes posibilidades de acuerdo a unas características específicas. Son muy utilizados para predecir una clase de instancia mediante la evaluación de sus atributos en los nodos de decisión del árbol, los cuales dependiendo del valor de esta misma, conducen a la instancia por una ramificación hasta su clasificación final. Suelen ser usados para clasificar a los consumidores (*Los algoritmos más usados…*, 2023).

Figura 1

Árbol de decisión



Nota: Los nodos hoja representan la predicción final a la que se llega según las distintas ramificaciones del árbol. Adaptado de *Árboles de decisión en…, 2024,* sitiobigdata.com, https://sitiobigdata.com/2019/12/14/arbol-de-decision-en-machine-learning-parte-1/.

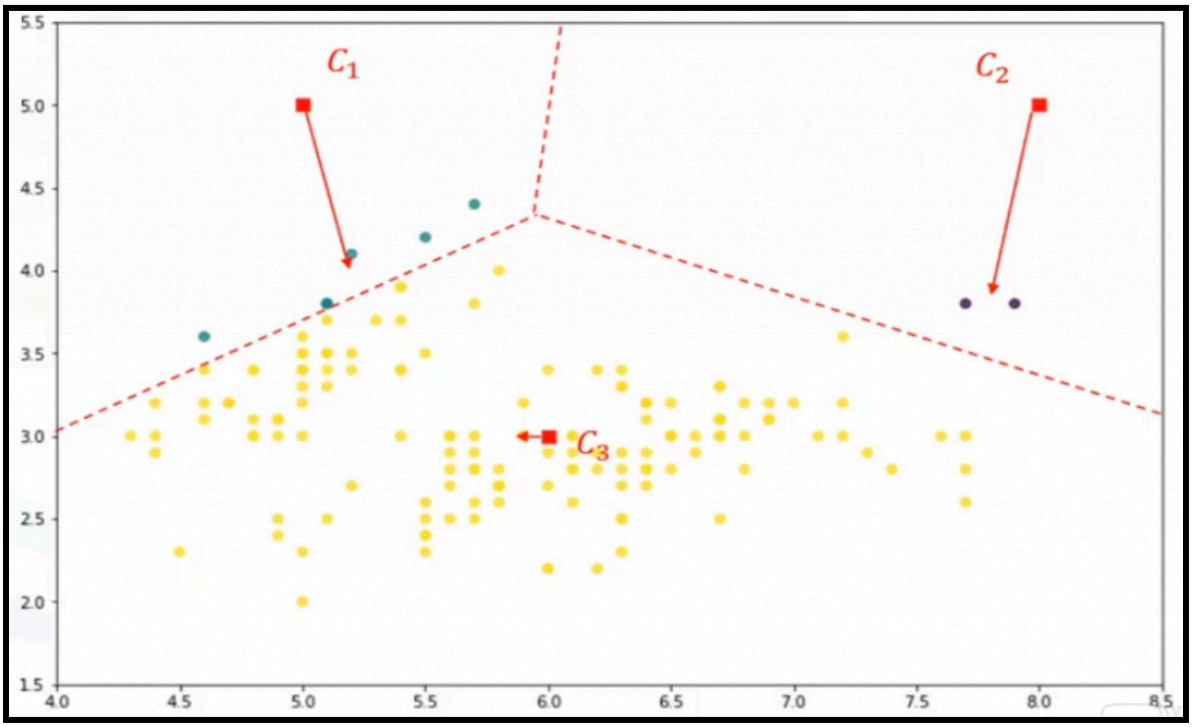
Por ejemplo, una empresa quiere predecir qué clientes son más propensos a comprar en temporadas de descuento, entonces analiza sus rasgos para categorizarlos, y con base en esto enfoca sus anuncios en estos individuos.

Agrupación

Tienen como objetivo la asociación de elementos similares a través de un análisis de sus tendencias generales. Buscan emplear este resultado para asignar cada valor a un grupo k de acuerdo a una distancia, la cual se maximiza con aquellos componentes distintos. Se diferencian de los algoritmos de clasificación pues no requieren una etiqueta previa y son usados para la segmentación del mercado (*Algoritmos de agrupación*…, 2022).

Figura 2

Agrupación por K-medias



Nota: La figura corresponde al producto de un algoritmo de K-medias de 3 conjuntos. C1, C2 y C3 son los centroides de cada grupo y de ellos se derivan las particiones, es decir las líneas punteadas de color rojo. Se organizan los puntos de acuerdo a la distancia de cada centroide. Recuperado de *Agrupación por K-medias…, 2021,* StatDeveloper, https://www.statdeveloper.com/agrupacion-por-k-medias/

A modo de muestra, un profesional de esta disciplina puede agrupar individuos de acuerdo a su rango de edad para proporcionarles anuncios más personalizados y afines a los gustos comunes de este conjunto.

Regresión

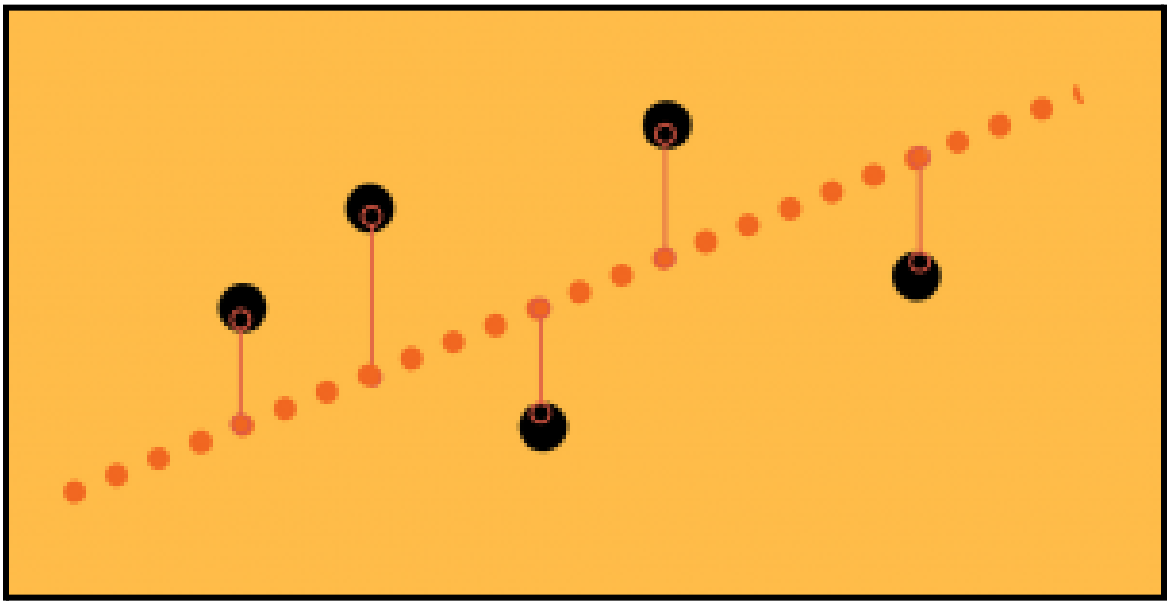
Es de carácter predictivo numérico, y funciona al relacionar los datos con una variable para aplicar modelos matemáticos que se ajusten a los valores para llevar a cabo ciertas estimaciones. Hay una variedad de tipos de regresión, ya sean lineales, no lineales o logísticos

(*Principales algoritmos utilizados…*, 2022). Pueden ser aplicados para el mercadeo de diversas maneras.

Para ilustrar, una empresa quiere determinar la cantidad de dinero a asignar al presupuesto de su próxima campaña publicitaria, para hacer esto puede usar la regresión lineal

Figura 3

Representación de regresión lineal



Nota: Se representa una regresión por el método de mínimos cuadrados, donde los puntos negros representan datos reales, y la línea ideal se establece al minimizar los residuos, es decir la distancia de los puntos a la misma. El valor de la y para la función en los distintos valores de x, es el valor que predice este algoritmo y que se usa para llevar a cabo los análisis subsecuentes. Tomado de *Principales algoritmos utilizados…, 2022,* Aprende Machine Learning, https://www.aprendemachinelearning.com/principales-algoritmos-usados-en-machine-learning/.

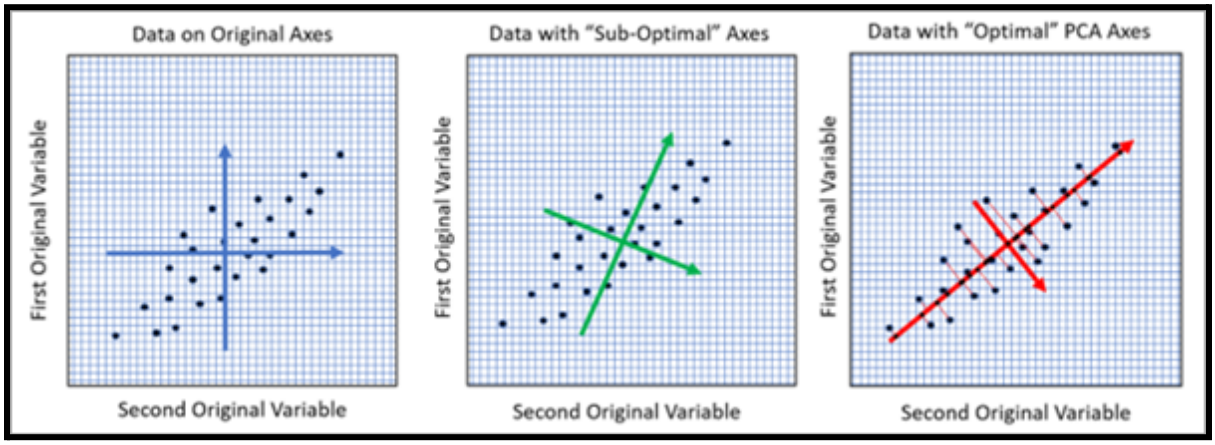
a fin de detallar la relación entre el monto invertido y las ventas generadas, y así encontrar la cantidad de inversión más rentable.

Reducción Dimensional

Encuentran un punto óptimo que tenga un equilibrio entre el tamaño muestral y la preservación de contenidos destacados. Suelen combinar algunos atributos o características de tal manera que se encapsule solo lo esencial y se elimine lo redundante. Facilitan mucho la examinación de los datos Estos algoritmos se pueden aplicar a prácticamente cualquier problema computacional, en los cuales cualquier compañía quiera usar de manera más eficiente sus recursos informáticos, por ejemplo en casos de mercadeo donde se busque recomendar un producto a un cliente al analizar solamente las características elementales que tengan alguna incidencia en esto. (*Los algoritmos más usados…*, 2023).

Figura 4

Transformación de datos con PCA



Nota: Se observan los efectos de ejecutar un Análisis de Componentes Principales (PCA) para revelar solo los puntos más relevantes. Los ejes de cada gráfico son indicativos de la varianza de las variables y son modificados por la rotación de los mismos, la cual busca ser más representativa de la relación entre ambos componentes. Adaptado de *Cómo funciona Reducción…,* (s. f.),

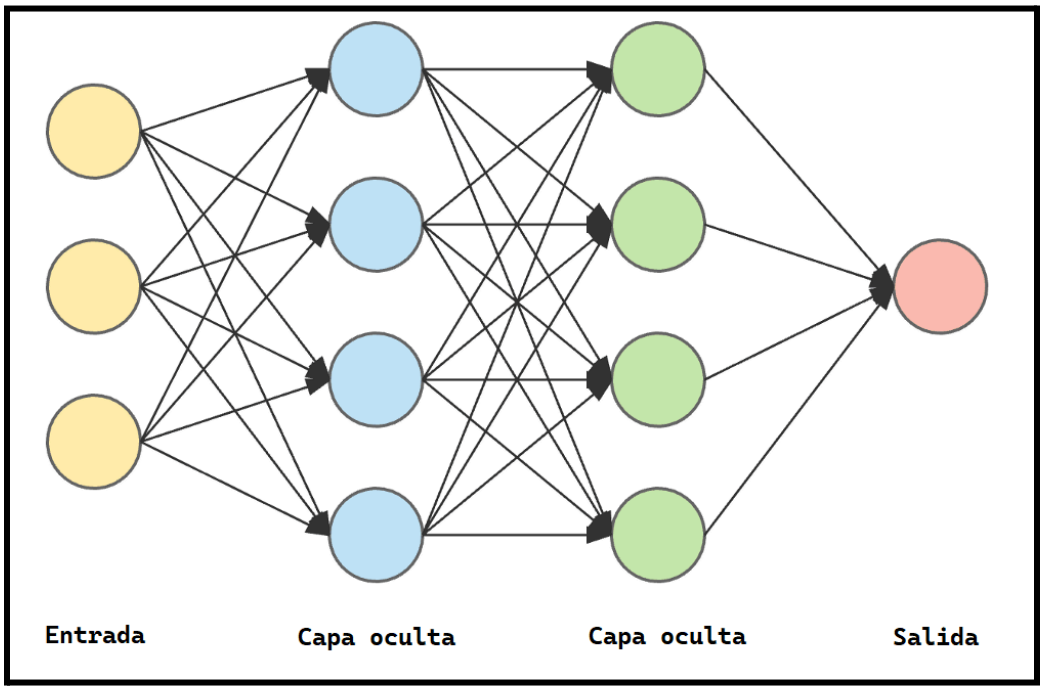
https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-statistics/how-dimension-reduction-works.htm

Redes Neuronales

Es un algoritmo de machine learning relacionado a la inteligencia artificial. Tiene las aptitudes necesarias para detectar todo tipo de relaciones y llevar a cabo los mismos procesos que los algoritmos anteriormente mencionados. Se divide en muchas partes pequeñas que conforman una red la cual emula el cerebro humano y puede aprender de manera no supervisada (*Los algoritmos más usados…*, 2023). Su carácter humano le permite llevar a cabo tareas específicas en el mercadeo, como un análisis de la percepción sobre cierto producto de una empresa, de acuerdo a un estudio de las opiniones positivas o negativas que circulan en internet.

Figura 5

Estructura general de redes neuronales



Nota: Se muestran las tres capas de las redes neuronales, la entrada alberga los datos principales, la sección oculta se encarga de desglosarlos y manipularlos, mientras que la salida retorna el resultado final. Es enfatizar que todos los nodos se encuentran conectados entre sí. Recuperado de *Qué son las redes neuronales y sus aplicaciones*, 2023, OpenWebinars.net, https://openwebinars.net/blog/que-son-las-redes-neuronales-y-sus-aplicaciones/

Conceptos Estadísticos

Son la base para examinar cualquier muestra de datos. Es necesario conocerlos para poder tener un planteamiento apropiado de una problemática, y llevar a cabo su subsecuente modelamiento para revelar tendencias claves.

Tipos de Variables

Se clasifican de acuerdo a propiedades o características comunes en concordancia con el fin del análisis. Aquellas consideradas independientes toman los valores de x en las gráficas y dictan el comportamiento de las independientes, que son modeladas mediante los puntos de la coordenada y. Se separan en cualitativas, es decir que no se operan de forma aritmética, como un color favorito, y en cuantitativas, las cuales representan magnitudes numéricas, como el tiempo. Estas se dividen en discretas y continuas, las primeras sólo contienen números enteros, y las segundas contienen decimales.

Las variables aleatorias asignan cifras a eventos reales cuyo comportamiento es susceptible a un grado azar que puede modelarse con funciones. Sus distribuciones de probabilidad surgen al juntar todas las probabilidades individuales y su sumatoria siempre debe ser igual a uno. Finalmente, están aquellas denominadas confusas, estas tienen su lugar en la recolección de datos ya que suelen influenciar el resultado, pero no se toman en cuenta; el objetivo de un experimento es su neutralización a través de la imposición de condiciones específicas. (L. Santamaría, comunicación personal, 22 de enero de 2024).

Estas son el pilar fundamental de cualquier exploración de datos ya que se encargan de delimitar el enfoque y el alcance de la misma. Es el trabajo de un profesional encontrar el equilibrio entre la precisión y la eficiencia a la hora de llevar a cabo su examinación. Como ejemplo, una segmentación de mercado emplea cantidades cuantitativas como la edad y métricas cualitativas como el sexo para analizar la demografía de los consumidores. La inclusión de otros criterios como la ubicación geográfica o el estado socioeconómico puede reducir significativamente al grupo poblacional que reúne todas las características, pero requerirá de mayor, esfuerzo, tiempo y capacidad computacional.

Estadísticas de Resumen

Es un grupo de informaciones clave que se derivan del grueso de los datos para dar a conocer concisamente las tendencias del conjunto en su totalidad. Se incluye la media o promedio, que es el resultado de sumar todos los valores y dividir por la cantidad de elementos que conforman la sumatoria. Se usa el mediano o cuartil dos (Q2), que representa el punto central de la población en un orden creciente si su número de componentes es impar, o con el valor medio de los dos valores centrales si es par. El Q1 representa el mediano de la primera mitad de todas las cifras, y el Q3 cumple la misma función para la segunda mitad.

Se encuentran las estadísticas de resumen que buscan explicar la variabilidad de los componentes de la muestra. Se destacan principalmente el rango, el rango intercuartil, y la desviación estándar. El primero es la diferencia entre la máxima y mínima medida, el segundo es lo mismo entre el Q1 y el Q3, y el tercero busca denotar la resta esperada entre el valor de cada punto y su valor proyectado. Finalmente es importante recalcar en los datos atípicos, los cuales sobresalen del resto por ser exorbitantemente pequeños o desmesurados. Pueden impedir que se lleve a cabo un análisis óptimo del conjunto por su influencia en la media y el rango, por lo que se eliminan en este proceso (S. A. Meza, comunicación personal, 18 de agosto de 2023).

Su importancia radica en que en un contexto empresarial no todos los ejecutivos o trabajadores involucrados en los proyectos están especializados en estadística. Consecuentemente, es clave para un profesional sintetizar los contenidos y la complejidad de sus métodos de tal manera que sean fácilmente digeribles para sus colaboradores. Esto se mantiene para los estudios de mercado, donde el analista debe contextualizar a los directivos de efectivamente con respecto a sus hallazgos,en función de que estos pueden proceder con el respaldo estadístico.

Correlación

Mide la fuerza con la que se corresponden dos variables, es decir como una influencia a la otra. Se dice que si es positiva el incremento de una variable produce un incremento en la otra, y viceversa si es negativa. Su coeficiente sólo produce valores correspondientes al intervalo [-1, 1]. Los estadísticos clasifican su intensidad de la siguiente manera (S. A. Meza, comunicación personal, 21 de septiembre de 2023):

Figura 6

Intensidad de la correlación según su coeficiente

Coeficiente de correlación (r)	Correlación
1 | -1	Perfecta
0.7 < r < 1 | -0.7 < r < -1	Fuerte
0.3 < r < 0.7 | -0.3 < r < -0.7	Moderada
0 < r < 0.3 | 0 < r < -0.3	Débil
0	No correlación

Nota: Valores del coeficiente de correlación mayores a 0 tendrán una asociación positiva, y mientras más cercanos sea este a 1 o -1 tendrá una correlación de mayor intensidad, es decir

el cambio en una variable produce mayor efecto en la otra. La figura 6 fue realizada por el autor utilizando información de su comunicación personal con S. A. Meza.

Es esencial para cualquier estudio de mercado y va de la mano con los algoritmos de regresión, puesto que mientras más fuerte sea esta, mejor será la precisión del análisis. Verbi gratia, si la percepción de los zapatos de una marca se correlacionan positivamente con el desempeño de sus atletas, la empresa podrá inferir que el mejor momento para lanzar una campaña de marketing será cuando sus deportistas estén demostrando un buen rendimiento.

Gráficas

Sintetizan un conjunto de información de tal manera que buscan exaltar tendencias para facilitar el proceso analítico al profesional. Son pieza clave para transmitir cualquier tipo de contenido relacionado a datos.

Distribución

Se describe a partir de la forma, variabilidad y tiene como propósito facilitar la identificación del comportamiento general de las estadísticas de resumen. Se describen como simétricas, sesgadas a izquierda o derecha y unimodales o bimodales. El sesgo está ligado principalmente a la concentración o agrupamiento de los datos, mientras que los nodos se relacionan principalmente con los picos en la frecuencia de un valor o grupo de valores específicos. A pesar de ser poco frecuente, se pueden observar huecos en algunos casos, los cuales se aparecen cuando un grupo de entradas, punto en coordenada x, no tiene ninguna salida, punto de la coordenada y.

Las gráficas simétricas permiten deducir que la media será igual al mediano, aquellas que estén sesgadas a la derecha tendrán un mediano menor al promedio, y aquellas que tengan un sesgo hacia la izquierda presentaran una tendencia opuesta a la anteriormente mencionada. Es común que los conjuntos bimodales tengan una desviación típica considerable

y un rango intercuartil más elevado de lo usual. Si una distribución se considera normal, se puede utilizar la desviación estándar y la media para encontrar la probabilidad de que un cierto valor sea inferior o superior al resto del conjunto con una tabla Z (S. A. Meza, comunicación personal, 27 de agosto de 2023).

Tipos de Gráficas

Diferentes tipos de datos y propósitos requieren de herramientas distintas para enfatizar en aspectos particulares de cada conjunto. Elementos como los gráficos de tarta, de barra y las tablas, trabajan con cantidades discretas y cualitativas para resaltar la frecuencia de sus componentes, ya sea de manera absoluta o relativa; un histograma cumple la misma función pero se adecúa a cifras continuas.

Otras representaciones como el box plot, o diagrama de caja, se prestan para recalcar en la distribución de las cifras, puesto que condensa las principales estadísticas de resumen (mediano, cuartiles, rango, rango intercuartil, datos atípicos) en un solo lugar. Los gráficos de dispersión son idóneos para acentuar las relaciones entre variables y se usan para modelar la correlación. Trabajan en conjunción con métodos matemáticos como la regresión lineal de mínimos cuadrados y se analiza su precisión mediante los residuos de la función (S. A. Meza, comunicación personal, 27 de agosto de 2023).

Técnicas de Recolección de Datos

Aunque la labor de un profesional en este campo implica enfocar la mayoría de su tiempo en desglosar ciertas informaciones, ocasionalmente debe ser él mismo quien se ocupe de reunir su material de trabajo. Esto puede conllevar la manipulación de contenidos ya existentes o el proceso de obtención de los mismos, mediante ciertos estudios. Una buena adquisición de datos es la base para un buen análisis de mercado.

Estudios Observacionales

Recopilan información sobre un grupo poblacional para determinar correlación. Nunca determinan causalidad ya que no existe un dominio sobre las variables independientes y confusas. Se usan los censos, métodos de muestreo, aleatorios, estratificados, de agrupamiento entre otros, dependiendo del tamaño de la muestra. La recolección de datos es susceptible a los sesgos, que en algunos casos se aplican de forma intencionada para producir resultados engañosos (L. Santamaría, 2023).

Se emplea frecuentemente para obtener información de productos y estudios de mercado en general. A modo de muestra, una empresa envía una encuesta a sus clientes en la cual pregunta por su edad y les pide una calificación de uno a diez del producto X. Los datos recogidos son usados para examinar cómo se relacionan este factor y la afinidad por dicho servicio para encontrar una relación entre las dos variables.

Experimentos

Buscan demostrar causalidad mediante una serie de condiciones rigurosas que disminuyen la influencia de las variables confusas. Para ser considerado un experimento, es imperativo establecer un grupo de control con placebos para tener un punto de comparación, el tratamiento se aplica intencionalmente, de forma aleatoria, y se debe procurar la replicación en entornos uniformes. La aleatoriedad permite generalizar los resultados a ciertas poblaciones, y la replicación busca dar credibilidad a las conclusiones.

Entre los diseños más comunes se encuentran los pares aleatorios y los bloques, en el primero se agrupan un par de sujetos con rasgos similares y solo uno recibe es intervenido, en el segundo se forman dos grupos, el primero posee cierta característica y al segundo le falta; se subdividen en referencial y práctico. En un experimento ciego el paciente desconoce si es sometido al tratamiento, en el doble ciego, quien lo suministra no conoce el contenido de este y en un triple ciego ocurre lo mismo con el responsable del análisis (L. Santamaría, comunicación personal, 20 de noviembre de 2023). Se pueden emplear para el análisis de mercado de

manera similar a los estudios observacionales, pero su rigurosidad hace que el proceso resulte más complejo y costoso, por lo que no es recomendable en todos los casos.

Probabilidad

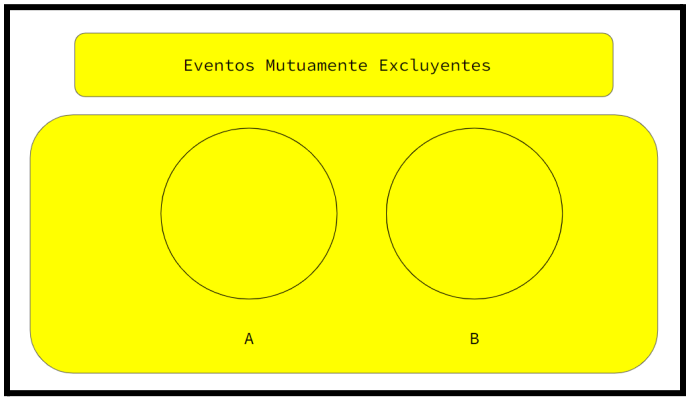
Es un número en el intervalo [0, 1] que se obtiene a partir de la división entre los casos favorables de un evento, es decir aquellos en los que se da dicha situación, y el grupo de escenarios totales, el cual representa todos los escenarios posibles. Una probabilidad de cero permite concluir que la ocurrencia en cuestión es imposible, mientras que una valor de uno expresa que es una certeza (L. Santamaría, 2024).

Eventos Mutuamente Excluyentes

Son sucesos que bajo ninguna circunstancia pueden ocurrir simultáneamente, ya que el uno imposibilita al otro. Los estadísticos lo manifiestan de la siguiente manera: $A \cap B = 0$, donde $A \cap B$ son el número de casos en los que se cumplen ambos acontecimientos (L. Santamaría, comunicación personal, 8 de enero de 2024). Una manifestación real de esto podría ser la siguiente: si dos marcas hacen el lanzamiento de su producto al mismo tiempo, se entiende que si una persona ha asistido al evento de la compañía A, este no pudo haber estado presente en la reunión de la empresa B.

Figura 7

Representación de eventos mutuamente excluyentes



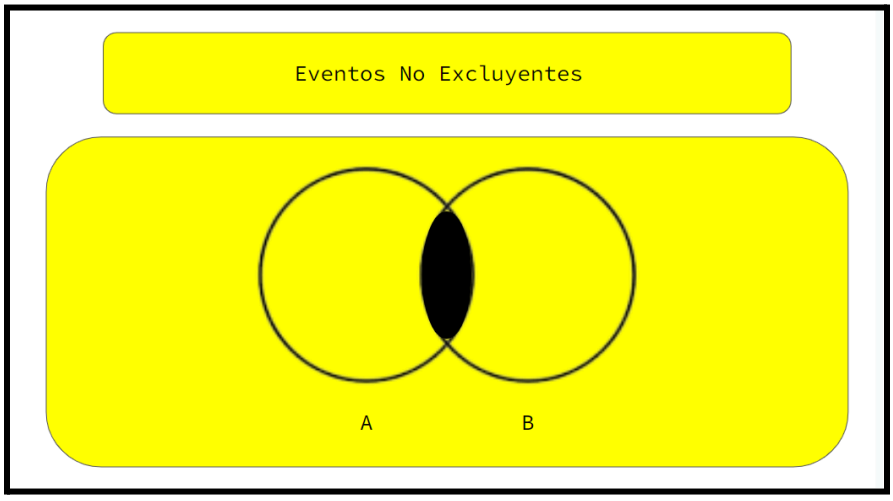
Nota: La esfera izquierda representa un evento A y la derecha un suceso B. Se entiende que todo aquello encapsulado por la esfera es una probabilidad favorable de que se de dicho caso. Al no estar superpuestos los diagramas, se da por hecho que la P(A ∩ B) es = 0. La figura 7 fue realizada por el autor utilizando información de su comunicación personal con L. Santamaría.

Eventos no Excluyentes

Son lo contrario de los eventos mutuamente excluyentes, por lo que la probabilidad de la intersección A ∩ B ≠ 0. Aunque es posible que las dos posibilidades se den al mismo tiempo, es importante recalcar que este no siempre es el caso. (L. Santamaría, 2024). Para ilustrar, si un cliente potencial compra los productos de la marca A, no implica que no pueda comprarle a una empresa B, por lo que un científico de datos debe discriminar este tipo de evento para incluirlo en su análisis de mercado

Figura 8

Representación de eventos no excluyentes



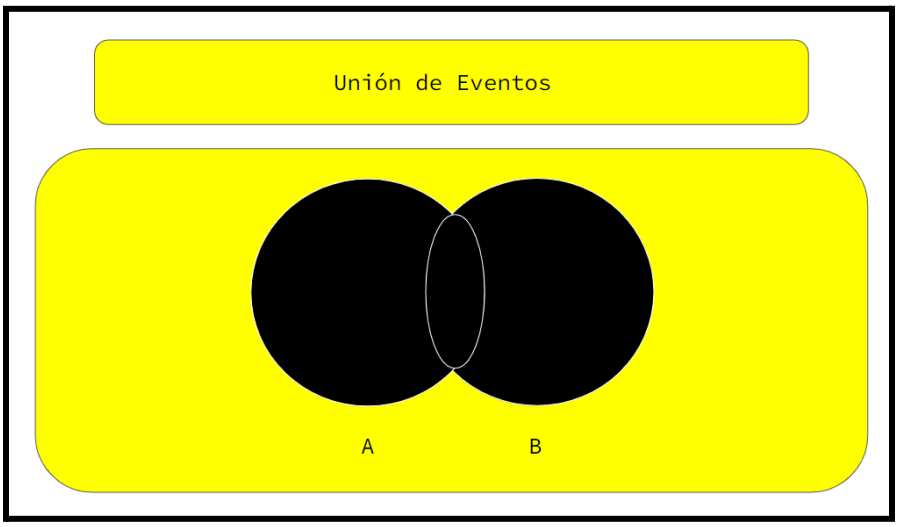
Nota: El área sombreada de negro representa la superposición de las esferas, que se entiende como su intersección (A ∩ B), es decir el caso de que se den los eventos A y B. La figura 8 fue realizada por el autor utilizando información de su comunicación personal con L. Santamaría.

Unión de Eventos

Se usa para expresar la posibilidad de los casos A, B, o su intersección. Matemáticamente se denota de la siguiente manera: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (L. Santamaría, comunicación personal, 11 de enero de 2024). La unión de eventos puede ser común en el mundo del mercadeo y las ventas, donde se frecuentan campañas promocionales que fomentan el comercio. Por ejemplo, una reducción porcentual por la compra de varias prendas, situación A, otro descuento por compras mayores a un monto específico, opción B, o la aplicación de ambos beneficios dado que se cumplan ambas condiciones, $A \cap B$, serían contemplados simultáneamente.

Figura 9

Representación de la unión de eventos



Nota: Se sombrea la totalidad de las dos esferas para plasmar la unión, ya que esta representa la probabilidad de que ocurran ambos eventos, es decir se contempla la posibilidad de solamente A, solo B, y A y B. La figura 9 fue realizada por el autor utilizando información de su comunicación personal con L. Santamaría.

Eventos Independientes

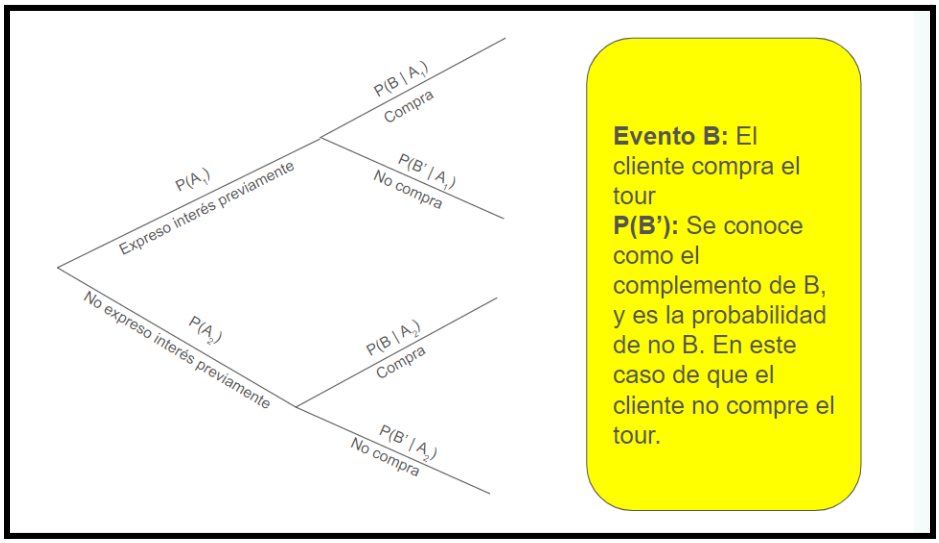
Son aquellos en los que un suceso o caso favorable no tiene incidencia alguna en una ocurrencia alternativa. En este tipo de eventos, P(A ∩ B), P mayúscula entendida como probabilidad de, se puede escribir como la P(A) multiplicada por la P(B) (L. Santamaría, 2024). Como muestra, si se hacen dos campañas publicitarias, una en medios tradicionales y otra en redes sociales, la respuesta del consumidor en cualquiera de ellas no afecta el éxito de la otra.

Condicional

Son el opuesto de los eventos independientes, por lo tanto el resultado del primero influye en las probabilidades del segundo. Se busca calcular la P(A) dado que se cumple B, P(A | B), mediante el teorema de Bayes, P(A | B) = (P(B | A) * P(A)) / P(B), o una versión simplificada donde P(A | B) = P(A ∩ B) / P(B). Suelen ser representados mediante diagramas de árbol (L. Santamaría, comunicación personal, 14 de enero de 2024). Una situación del mundo real donde se puede observar el uso de probabilidad condicional es la siguiente: Una agencia de viajes quiere conocer la posibilidad que un cliente compre un tour a Estados Unidos, dado que previamente haya expresado interés en visitar este destino turístico.

Figura 10

Diagrama de árbol condicional



Nota: La figura 10 es una representación gráfica del ejemplo del autor en el párrafo superior y fue realizada por el mismo utilizando información de su comunicación personal con L. Santamaría.

Binomial

Encapsula aquellas situaciones donde todos los escenarios posibles resultan en solo dos casos, P(A) o P(B). Se puede aplicar cuando los eventos no son independientes, tienen una misma posibilidad de manifestarse, son representables por un número natural y tienen una cantidad n de ensayos definidos. Se calcula mediante la siguiente fórmula, nCx * p^x * (1 - p)^n - x (L. Santamaría, 2024). Para ejemplificar, si una empresa manda 10,000 comunicaciones electrónicas dando a conocer su producto, donde el porcentaje de que el usuario lo abra es del 20%, y quiere saber la probabilidad de que por lo menos 2,500 personas hayan abierto el mail, pueden recurrir a la distribución binomial. En este caso hay dos alternativas, el individuo abre el correo o no abre el correo, P(A) y P(B), los eventos serían independientes ya que el hecho de que un usuario lea el mensaje no afecta los demás eventos, la expectativa de que se abra es de 0.2 para cada individuo y hay n ensayos de 10,000 emails.

Comunicación

En un entorno laboral, se espera de un científico de datos la capacidad de comunicar efectivamente sus hallazgos, de tal manera que sean fácilmente digeribles para otros miembros de su equipo de trabajo. Consecuentemente, un profesional debe poseer nociones expositivas para complementar sus habilidades blandas con una propuesta de valor acorde a su especialización.

Visualización de Datos

Se entiende como el uso de componentes de apoyo como gráficas, mapas y tablas entre otros para facilitar el entendimiento y el análisis de datos. Su eficiencia radica en su capacidad de resaltar ciertas relaciones no aparentes en grupos, debido a que permite organizar los elementos de tal manera que pueden surgir patrones específicos para esclarecer las conexiones implícitas entre variables. Además cabe recalcar que la implementación de dichos materiales visuales aporta una estimulación sensorial que contribuye a la transmisión de los contenidos en cuestión, especialmente a las masas que buscan maneras llamativas de recibir la información.

Los componentes visuales tienen roles específicos de acuerdo a sus características. Los esquemas favorecen estructuras jerárquicas, mientras que los mapas presentan conceptos claves o de carácter espacial. Las gráficas son la representación por excelencia de los conjuntos numéricos por sus aptitudes para acentuar proporción, distribución y demás. Finalmente, los diagramas destacan por su manejo de información multivariada, hecho por el cual se emplean en el campo de la estadística (H. Parraga & J. Yntusca, 2023).

Narrativa de Datos

La narración de datos es una de las principales formas de potencializar el alcance de los mismos frente a una audiencia, especialmente en campañas de mercadeo. Su éxito recae en una gran variedad de factores, entre ellos su intrínseca conexión con la naturaleza del ser humano, la cual ha interiorizado el uso de la narrativa como método para transmitir el conocimiento a través de los siglos. Sin embargo, su mayor fortaleza se encuentra en su habilidad para convertir un “aburrido” conjunto numérico, en un mensaje capaz de romper la apatía y sembrar el interés en un grupo para moldear su percepción y llevar a cabo una persuasión exitosa.

Algunos académicos sugieren que las buenas historias comparten elementos universales entre sí. Afirman que la característica principal de estas es el establecimiento de

una relación de causa y consecuencia, en la cual la primera actúa como anzuelo para mejorar la recepción del mensaje materializado como el efecto. Aseguran que un origen inesperado es el cimiento de la curiosidad, y por consiguiente de la sorpresa y de un impacto prolongado. Matei y Hunter (2021) lo resumen de la siguiente manera: Una historia convincente despierta la imaginación del público y le hace pensar. Una historia hábil obliga a una persona a adoptar una actitud ... las historias de datos, en su mejor momento, nos empujan del creer, al saber.

Mercadeo

El mercadeo es un término que encapsula todas las acciones de una empresa con el propósito de crear o fortalecer una relación existente con sus clientes. Busca generar interés y promover la lealtad hacia la marca a través del estudio del consumidor y la exposición del producto o servicio en cuestión (Patrutiu-Baltes, 2016).

Inbound Marketing

Se diferencia de otros tipos de marketing por su enfoque en el consumidor y carácter no intrusivo. Emplea métodos de creación de contenido de alta calidad en redes sociales y otras plataformas para darse a conocer y construir su reputación. Está estrechamente relacionado con el perfilamiento del consumidor dado que potencia su alcance mediante contenidos personalizados que integran al usuario. Las estadísticas revelan que con la llegada de la era digital, esta forma de mercadeo se ha posicionado como una de las más eficientes gracias a un cambio cultural en el mercado y el surgimiento de herramientas innovadoras (PATRUTIU-BALTES, 2016).

SEO

Acrónimo para Search Engine Optimization, es una práctica que busca sacar el mayor provecho de una página web, para incrementar el volumen del tráfico que recibe en internet. Se

basa en métricas que utilizan los motores de búsqueda como el contenido, las URL, palabras claves, etc.. Recientemente, se comenzó a explorar el uso del big data para alcanzar una mayor optimización (*What is SEO?*..., 2021).

Anuncios

Forma de publicidad que tiene como fin el posicionamiento de una empresa y la divulgación de un producto a una población específica de acuerdo a los estudios de mercado. Se propagan mediante medios tradicionales como los periódicos y la televisión, medios digitales como las redes sociales entre otros. El auge del big data en los últimos años ha revolucionado esta industria, la cual ha gravitado lejos de productos generalizados para reemplazarlos con anuncios dirigidos (*Publicidad: Qué es, elementos e importancia,* 2022).

Segmentación

Clasifica los consumidores de acuerdo a rasgos comunes para determinar cómo emplear la mercadotecnia de manera más eficiente en cada caso. Frecuentemente, se hace en torno a la demografía, psicografía, conducta y estado socioeconómico. Permite construir un perfil del consumidor, entendido como el modelo del cliente idóneo, para focalizar los esfuerzos apropiadamente (S. Orjuela & M. Chaparro, 2008).

Conflictos Éticos y Morales

A luz de los recientes escándalos mediáticos se han dado a conocer una gran variedad de riesgos asociados al uso de big data analytics. Expertos y miembros del público general se cuestionan cómo estos comportamientos vulneran sus derechos a la privacidad, especialmente con relación a la manipulación de información sensible por parte de entes privados. Sus preocupaciones son multifacéticas, por un lado, está todo lo que conlleva una práctica maliciosa, por el otro se encuentran las repercusiones de un uso pobre y poco profesional. Un análisis errado o con sesgos importantes, puede ser el causante de la implementación de

políticas discriminatorias que vayan en detrimento del desarrollo de la organización o ciertos grupos poblacionales. Asimismo, la digitalización de algunos contenidos es motivo de angustia y abre la posibilidad de que se exploten brechas de seguridad para ejecutar ciberataques. Es de suma importancia, que un científico de datos sea consciente de su responsabilidad y las ramificaciones de su trabajo para procurar su calidad y fomentar una ética profesional, especialmente en un contexto de mercadeo, donde se ha vuelto común que los gigantes tecnológicos vendan la información de sus usuarios a anunciantes u otros terceros (*The ethical implications of big data analytics*, 2023).

Capítulo 6:Marco Conceptual

- **Análisis de Tendencias:** Es una técnica estadística que permite estudiar una o más variables en un período de tiempo aportando información para la toma de decisiones en la empresa. [1]
- **Aprendizaje Automático:** Es el proceso mediante el cual se usan modelos matemáticos de datos para ayudar a un equipo a aprender sin instrucciones directas. Se considera un subconjunto de la inteligencia artificial (IA). El aprendizaje automático usa algoritmos para identificar patrones en los datos, y esos patrones luego se usan para crear un modelo de datos que puede hacer predicciones. [2]
- **Atributos:** Son las características individuales que diferencian un objeto de otro y determinan su apariencia, estado u otras cualidades. Los atributos se guardan en variables denominadas de instancia, y cada objeto particular puede tener valores distintos para estas variables. [3]
- **Bases de Datos:** Es una recopilación de datos sistemática y almacenada electrónicamente. Puede contener cualquier tipo de datos, incluidos palabras, números, imágenes, vídeos y archivos. Puede usar un software denominado sistema de administración de bases de datos (DBMS) para almacenar, recuperar y editar datos. [4]
- **Big Data:** El término “big data” abarca datos que contienen una mayor variedad y que se presentan en volúmenes crecientes y a una velocidad superior. Esto también se conoce

---

[1] Arias, E. R. (2022, 24 noviembre). *Análisis de tendencia*. Economipedia.tomado de https://economipedia.com/definiciones/analisis-de-tendencia.html

[2] *¿Qué es el aprendizaje automático?* | *Microsoft Azure*. (s. f.).Tomado de https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-machine-learning-platform

[3] *Conceptos básicos de la Programación Orientada a Objetos*. (s. f.). Tomado de http://www.sc.ehu.es/sbweb/fisica/cursoJava/fundamentos/clases1/clases.htm

[4] *¿Qué es una base de datos? - Explicación de las bases de datos en la nube - AWS*. (s. f.). Amazon Web Services, Inc.Tomado de https://aws.amazon.com/es/what-is/database/#:~:text=Una%20base%20de%20datos%20es,almacenar%2C%20recuperar%20y%20editar%20datos.

como “las tres V”. Dicho de otro modo, el big data está formado por conjuntos de datos de mayor tamaño y más complejos, especialmente procedentes de nuevas fuentes de datos. Estos conjuntos de datos son tan voluminosos que el software de procesamiento de datos convencional sencillamente no puede gestionarlos. Sin embargo, estos volúmenes masivos de datos pueden utilizarse para abordar problemas empresariales que antes no hubiera sido posible solucionar. [5]

- **Compilación del Código:** Es el proceso de transformar un programa informático escrito en un lenguaje en un conjunto de instrucciones en otro formato o lenguaje. Un compilador es un programa de computadora que realiza dicha tarea. [6]
- **Conjunto:** Es la agrupación de diferentes elementos que comparten entre sí características y propiedades semejantes. [7]
- **Consumidor:** Es una persona u organización que consume bienes o servicios, que los productores o proveedores ponen a su disposición en el mercado y que sirven para satisfacer algún tipo de necesidad. [8]
- **Demografía:** Es la ciencia que estudia a las poblaciones humanas de manera estadística, es decir, en base a datos numéricos y cálculos que permiten analizar diversos aspectos como el tamaño, la densidad, la distribución y las tasas de vitalidad de una población. Los estadísticos que utiliza se obtienen mediante instrumentos de evidencia científica (bases de datos, encuestas, censos y otros). [9]

---

[5] *¿Qué es el big data?* (s. f.).Tomado de https://www.oracle.com/co/big-data/what-is-big-data/
[6] *Compilar - Glosario de MDN Web Docs: Definiciones de términos relacionados con la Web | MDN*. (2023, 13 noviembre). MDN Web Docs.Tomado de https://developer.mozilla.org/es/docs/Glossary/Compile
[7] Equipo editorial, Etecé. (2021, 5 agosto). *Conjunto - Concepto, tipos, ejemplos y otras acepciones*. Concepto.Tomado de https://concepto.de/que-es-un-conjunto/
[8] Galán, J. S. (2022, 24 noviembre). *Consumidor*. Economipedia.Tomado de https://economipedia.com/definiciones/consumidor.html
[9] Equipo editorial, Etecé. (2023, 20 noviembre). *Demografía - Concepto, tipos, importancia y características*. Concepto.de https://concepto.de/demografia/

- **Distribución Normal:** Es un modelo teórico que aproxima el comportamiento de una variable aleatoria a una situación ideal, utilizando la media y la desviación típica como parámetros clave. [10]
- **Estudio de Mercado:** Es el proceso mediante el cual realizamos la recolección y análisis de información que sirve para identificar las características de un mercado y comprender cómo funciona. Esta investigación es utilizada por diversos ramos de la industria para garantizar la toma de decisiones y entender mejor el panorama comercial al que se enfrentan al momento de realizar sus operaciones. [11]
- **Grupo de Control:** Es el grupo que no recibe el nuevo tratamiento que está en estudio. Se compara este grupo con el grupo que recibe el nuevo tratamiento para determinar si el nuevo tratamiento es eficaz. [12]
- **Instancia:** Se refiere a un objeto específico que se crea a partir de una clase. Para entenderlo mejor, primero debemos aclarar qué es una clase. En la programación orientada a objetos, una clase es como un plano o una plantilla que define las propiedades y comportamientos de un objeto. [13]
- **Lenguajes de Alto Nivel:** Son aquellos que se encuentran más cercanos al lenguaje natural de las personas, que al lenguaje máquina. Están dirigidos a solucionar problemas mediante el uso de EDD 's. Se tratan de lenguajes independientes de la arquitectura del ordenador y de su hardware … Estos lenguajes permiten al programador olvidarse por completo del funcionamiento interno de la máquina/s para la que están diseñando el

---

[10] Rodó, P. (2024, 24 enero). *Distribución normal: Qué es, cómo se calcula y ejemplos*. Economipedia.Tomado de https://economipedia.com/definiciones/distribucion-normal.html
[11] QuestionPro. (s. f.). *¿Qué es un estudio de mercado?* | *QuestionPro*.Tomado de https://www.questionpro.com/es/estudio-de-mercado.html
[12] *Diccionario de cáncer del NCI*. (s. f.). Instituto Nacional del Cáncer.Tomado de https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/grupo-de-control
[13] Admin. (s. f.). *Que es una instancia en programación*. Programación Desde Cero.Tomado de https://programacion.top/conceptos/instancia/

programa. Tan solo necesitan un traductor que consiga transformar el código fuente del lenguaje de alto nivel a un código cercano a las características de la máquina. [14]

- **Lenguajes de Bajo Nivel:**Son totalmente dependientes de la máquina, es decir, dependen directamente del hardware donde van a ejecutarse. Por ello, los programas que se realizan con este tipo de lenguajes no se pueden migrar o utilizar en otras máquinas, con otros tipos de procesadores. Al estar prácticamente diseñados a medida del hardware, aprovechan al máximo las características del mismo. Son extremadamente rápidos, aunque las operaciones que les podemos pedir también son extremadamente simples. [15]
- **Librerías:** Es una colección de código desarrollado previamente que los programadores pueden utilizar para desarrollar *software* de manera más ágil. Estas colecciones de código reutilizable suelen resolver problemas o necesidades comunes de desarrollo. [16]
- **Muestra:** Es un subconjunto de datos perteneciente a una población de datos. Estadísticamente hablando, debe estar constituido por un cierto número de observaciones que representen adecuadamente el total de los datos. [17]
- **Muestreo Aleatorio:** Cada uno de los elementos a analizar que componen el universo tienen la misma posibilidad de ser escogidos como parte de la muestra … se estudian mediante la teoría de la probabilidad. Esta es una disciplina que define una serie de reglas para determinar la ocurrencia de ciertos fenómenos o procesos. [18]
- **Placebo:** Sustancia inactiva u otra intervención que tienen la misma apariencia y se administran de la misma forma que el medicamento o tratamiento activo que se está

---

[14] *Tipos de lenguajes de programación*. (s. f.). DesarrolloWeb.com.tomado de https://desarrolloweb.com/articulos/2358.php
[15] Ibid..
[16] Unir, V. (2024, 29 enero). ¿Qué son las librerías en programación y para qué sirven? *UNIR*.de https://www.unir.net/ingenieria/revista/librerias-programacion/#:~:text=Una%20librer%C3%ADa%20de%20programaci%C3%B3n%20es,o%20necesidades%20comunes%20de%20desarrollo.
[17] López, J. F. (2022, 24 noviembre). *Muestra estadística*. Economipedia.tomado de https://economipedia.com/definiciones/muestra-estadistica.html
[18] Westreicher, G. (2022, 24 noviembre). *Aleatorio*. EconomipediaTomado de. https://economipedia.com/definiciones/aleatorio.html

probando. Los efectos del medicamento activo u otra intervención se comparan con los efectos del placebo. [19]

- **Posicionamiento Empresarial:** Es una estrategia de marketing que las marcas crean para definir su identidad de marca mientras comunican su propuesta de valor de la marca, que es el motivo por el que un cliente prefiera su marca sobre otras. Además, el posicionamiento de marca se utiliza cuando una empresa quiere posicionarse de cierta manera ante sus audiencias para que los clientes generen asociaciones entre la marca y su propuesta de valor. [20]
- **Regresión Lineal de Mínimos Cuadrados:**Es un método común para estimar los coeficientes de las ecuaciones de regresión lineal que describen la relación entre una (o varias) variables independientes cuantitativas y una variable dependiente. Según el número de variables, podemos hacer una regresión simple o múltiple. [21]
- **Replicación:** El término, también conocido como reproducibilidad o replicación, se refiere a la capacidad de repetir un experimento en diferentes situaciones, con diferentes sujetos e investigadores. Esto con el fin de comprobar la seguridad de los hallazgos del primer experimento y comprobar su viabilidad. [22]
- **Residuos:** La diferencia entre la predicción y el valor observado es el residual. Si graficamos los valores observados y superponemos la línea de regresión ajustada, los

---

[19] *Diccionario de cáncer del NCI*. (s. f.-b). Instituto Nacional del Cáncer.Tomado de https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/placebo

[20] *¿Qué es el posicionamiento de marca y por qué es tan importante?* (2024, 12 febrero). Amazon Ads.de https://advertising.amazon.com/es-mx/library/guides/brand-positioning#:~:text=El%20posicionamiento%20de%20marca%20es%20el%20valor%20%C3%BAnico%20que%20una,prefiera%20su%20marca%20sobre%20otras.

[21] *Regresión lineal de los mínimos cuadrados (OLS)*. (s. f.). XLSTAT, Your Data Analysis Solution.de https://www.xlstat.com/es/soluciones/funciones/regresion-lineal-de-los-minimos-cuadrados-ols#:~:text=La%20regresi%C3%B3n%20lineal%20de%20los%20m%C3%ADnimos%20cuadrados%20(en%20ingl%C3%A9s%20OLS,cuantitativas%20y%20una%20variable%20dependiente.

[22] García-Bullé, S. (2022, 2 noviembre). *¿Qué es la replicabilidad y por qué está en crisis?* Observatorio / Instituto Para el Futuro de la Educación.Tomado de https://observatorio.tec.mx/edu-news/replicabilidad-ciencia/#:~:text=El%20t%C3%A9rmino%2C%20tambi%C3%A9n%20conocido%20como,experimento%20y%20comprobar%20su%20viabilidad.

residuos para cada observación serían la distancia vertical entre la observación y la línea de regresión. [23]

- **Sintaxis:** La sintaxis de un lenguaje de programación se refiere a las reglas y estructuras que se utilizan para escribir el código, siendo equivalente a las reglas gramaticales y de redacción que rigen a cada idioma. La sintaxis define cómo deben escribirse las instrucciones, variables y estructuras de control para que el programa funcione correctamente. [24]
- **Software:** Designa a todo componente intangible (y no físico) que forma parte de dispositivos como computadoras, teléfonos móviles o tabletas y que permite su funcionamiento. El software está compuesto por un conjunto de aplicaciones y programas diseñados para cumplir diversas funciones dentro de un sistema. Además, está formado por la información del usuario y los datos procesados. [25]
- **Tabla Z:** Es una herramienta utilizada en estadística para encontrar el área bajo la curva de la distribución normal estándar, la cual tiene una media de 0 y una desviación estándar de 1. La tabla Z presenta una serie de valores que corresponden al área bajo la curva de la distribución normal estándar para una puntuación Z dada. La puntuación Z es una medida que indica cuántas desviaciones estándar se encuentra un valor respecto a la media de la distribución. [26]
- **Tratamiento:** Son aquellas fuentes cuyo efecto sobre la respuesta es de particular interés para el experimentador. [27]

---

[23] Statologos. (2021, 7 mayo). *¿Qué son los residuos en las estadísticas?* Tomado de https://statologos.com/residuos/
[24] Biraki, M. D. (2023, 29 agosto). *¿Qué es un lenguaje de programación? HACK A BOSS*.Tomado de. https://www.hackaboss.com/blog/que-es-lenguaje-de-programacion
[25] Equipo editorial, Etecé. (2023a, noviembre 19). *Software - Qué es, concepto, tipos, ejemplos, hardware*. Concepto. https://concepto.de/software/
[26] es.z-table.com. (2023, 7 abril). *Tabla z*.Tomado de https://es.z-table.com/
[27] Del Estado de Hidalgo, U. A. (s. f.). *Boletín científico - Ingenio y Conciencia No. 1*.Tomado de https://www.uaeh.edu.mx/scige/boletin/sahagun/n1/e1.html

Capítulo 7: Resultados del Proceso

La práctica como culminación del Senior Project conlleva una oportunidad de condensar todas las etapas del proceso en un ejercicio que permite al estudiante poner a prueba lo aprendido durante cada capítulo. Consecuentemente, para extraer el máximo provecho de esta oportunidad, es de suma importancia comparar y contrastar la teoría y su aplicación en el mundo real, a fin de obtener una perspectiva más holística de la disciplina en cuestión y el proyecto en general.

Para llevar a cabo este análisis de manera óptima es imperativa una recapitulación de los principales temas abordados en Capítulo 5: Marco Teórico. Estos incluyen pero no se limitan a la historia de la carrera, las herramientas de software, los algoritmos computacionales, los principios de la estadística descriptiva e inferencial, la comunicación de resultados, el mercadeo y la ética profesional. Igualmente, es importante recalcar que dichas temáticas giran en torno a un eje central cuyo propósito era dar a conocer las técnicas de la ciencia de datos empleadas para ejecutar un estudio de mercado en el mundo empresarial.

La jornada laboral por su parte fue más representativa del trabajo de un analista tradicional, por lo que ciertos aspectos de la investigación como los algoritmos computacionales son poco relevantes para esta ponderación. Esto se debe a que la ciencia de datos como ámbito de estudio definido fue instaurada hace poco tiempo a comparación de otros campos, hecho que ha llevado a que no esté satisfactoriamente desarrollada en el área metropolitana de Bucaramanga.

La práctica como tal fue constituida por cuatro fases principales de acuerdo al Ciclo Deming (PHVA). La primera de ellas, planear, se dividió en un marco teórico sobre el caso de estudio del dengue y la obtención de contenidos pertinentes a la exploración. En este paso se evidenciaron dos aspectos claves del trabajo previo que guiaron el curso del análisis. Por un lado, la indagación con respecto a los tipos de variables fue el cimiento para determinar cómo

se generaría la base de datos a observar en los días laborales. Lo encontrado sobre las variables independientes, dependientes, cualitativas y cuantitativas que fueran relevantes tuvo un rol clave para establecer un balance a la hora de extraer la información de la Clínica San Materno Infantil San Luis en la Búsqueda Activa. Por el otro, lo averiguado sobre el lenguaje de programación SQL fue utilizado para llevar a cabo esta tarea en conjunto con el asesor. Lo aprendido sobre las facultades de esta herramienta para consultar y manipular una fuente de reportes tan extensa fue concordante con lo evidenciado en el ejercicio real y sirvió como base para comprender los aspectos del trabajo que excedían lo investigado.

En relación al hacer, la segunda etapa del ciclo, se empleó el software de Microsoft Excel para segmentar adecuadamente las entradas anteriormente definidas. Dicho programa fue parte del marco teórico, no obstante, se le dio un enfoque más superficial a sus funcionalidades, por lo que lo reflejado en la práctica excedió en complejidad las búsquedas previas. Mientras que lo indagado sólo hacía mención de las tablas dinámicas, la práctica reveló nuevos aspectos sobre la configuración de sus filtros, valores y capacidad multivariada de análisis que trabajaba en conjunto con las fórmulas matemáticas ya integradas en la aplicación. En lo respectivo a desglosar los datos, los conocimientos de estadística descriptiva y componentes visuales demostraron ser útiles para completar las labores de manera satisfactoria. Recurrir a las estadísticas de resumen resultó ser un gran facilitador para entender el comportamiento de ciertas variables a la vez que se hacían más evidentes las medidas de tendencia central y de variabilidad en la muestra. Lo estudiado con respecto a la distribución permitió utilizar la simetría a fin de generar gráficos más representativos mediante transformaciones matemáticas.

El tercer aspecto, verificar, resaltó una gran falencia en el proceso de exploración ya que se omitió una sección sobre las investigaciones contextuales de cualquier problema en cuestión. Aunque estar familiarizado con los métodos de análisis estadísticos es esencial, las tendencias encontradas no tienen una verdadera utilidad si no pueden ser adaptadas al entorno

general. Fueron la pesquisa del caso de estudio, dengue, y el diagnóstico socioeconómico de Santander los que otorgaron un verdadero entendimiento del panorama completo. Por ejemplo, determinar que hay más pacientes procedentes de Bucaramanga a comparación de Floridablanca puede llevar a la conclusión de que hay mayor incidencia de la patología en un municipio, lo cual no puede ser corroborado si no se tiene en cuenta la proporción en relación a la población total de cada uno.

Finalmente, la cuarta fase, actuar, puso en práctica la narrativa de datos y el apoyo visual. Como producto final se construyó un informe con dos elementos esenciales, un reporte completo que documentara todos los aspectos del proceso y una infografía en canva que hiciera más digerible la información y se centrara sólo en las conclusiones esenciales. Para esto la idea de convertir un “aburrido” conjunto numérico en un mensaje capaz de sembrar el interés y evocar el pensamiento crítico fue el principio clave a la hora de sintetizar los hallazgos. Del mismo modo, los roles encontrados para los distintos tipos de gráficos coincidieron con su uso práctico, lo que confirmó sus facilidades y el propósito averiguado.

Del Capítulo 7 se puede concluir que el enfoque del marco teórico estuvo algo desatinado con respecto al ejercicio laboral debido a la cercanía con el análisis de datos tradicional. Su extensión temática fue superior a la requerida para este trabajo y consecuentemente la profundidad de ciertos temas más aterrizados a las prácticas no fue suficiente. La indagación de mercadeo, probabilidad, algoritmos y diseño experimental pudo haber sido redirigida hacia el software de Microsoft Excel, cuyo rol fue primordial para llevar a cabo el proceso exitosamente. A raíz de esto se resalta la importancia de procurar la mayor compatibilidad posible entre el plan de operaciones y la pregunta de investigación.

Capítulo 8: Conclusiones

Conclusiones del Proyecto

Habiendo concluído el Senior Project puedo decir que considero se ha dado respuesta a la pregunta de investigación a rasgos generales y a un grado con el cual estoy satisfecho desde lo teórico, y a un nivel parcial desde un punto de vista práctico. Pienso que me enfoqué más en el conocimiento del área que en la pregunta como tal, por lo que pude abarcar todos los propósitos establecidos y hacerme una idea que me permitió responder el interrogante más indirectamente.

Los algoritmos son la técnica principal de un científico de datos para llevar a cabo un estudio de mercado. Son la manifestación del trabajo en conjunto de un propósito claro, el conocimiento estadístico y líneas de código que permiten derivar conclusiones apoyadas por datos para generar una experiencia más personalizada y que se ajuste al SEO, el consumidor y la estrategia de mercadeo de la empresa. Son una herramienta altamente precisa y cuyas variantes como la clasificación, agrupación, regresión entre otras se adaptan a todo tipo de problemas para brindar soluciones en este espacio digital. Son habilitados por herramientas como SQL y potenciados por programas altamente especializados como SPSS y altamente propagados como lo es Excel que expanden la esencia de la disciplina a diversos públicos.

Con respecto a los objetivos que conciernen la legislación, la gobernanza de datos y la accesibilidad a bases de datos, fueron cubiertos a profundidad por el marco legal. Las lecturas del Decreto 1389 de 2022, la ley 1581 de 2012 y demás contribuyeron ampliamente a esto y resolvieron mis dudas sobre el debido proceso, la naturaleza de distintos tipos de información y de cómo la gobernanza se aplica en la nación para establecer mejores lineamientos estratégicos.

De igual forma, el marco teórico abarcó la gran mayoría de las metas que establecí para mí. En general la naturaleza de los objetivos giraba en torno a observar, estudiar o

conocer, por lo que esta fase se prestó para llevar a cabo esta finalidad. Mediante este capítulo pude abarcar todos los temas desde las herramientas de software, hasta los algoritmos esenciales y los principios de mercadeo, por lo que creo se cumplió adecuadamente.

Finalmente, está la práctica. Creo que este ejercicio me ayudó a complementar lo teórico y sobrepasar lo inicialmente planteado en algunos casos. No solo conocí programas con Excel, pero pude trabajarlo y volverme más o menos diestro con la aplicación en el transcurso de esos días. Igualmente creo que mi propósito de mejorar las habilidades de comunicación se cumplió en estos días y a través del proyecto en general. La redacción de cada capítulo impulsaba a generar un producto que transmitiera mis hallazgos o ideas de manera clara y concisa. La práctica me obligó a interactuar en un nuevo ambiente y adaptar mi comunicación al mismo, y estoy seguro de que las presentaciones de la próxima semana contribuirán de igual manera conforme me acoplo a un público adulto y estudiantil.

Con respecto al desarrollo del proyecto creo que fue el esperado. Los capítulos teóricos estuvieron muy bien organizados, de tal modo que no hubo sorpresas ni ambigüedad. Igualmente la práctica, hubo un buen entendimiento con el asesor, quien fue muy transparente con las actividades y diseñó un plan fácil de entender en el acuerdo. Desde el año pasado algunos seniors me dijeron, el senior le va a dar lo que usted le de, y creo que eso se evidenció a través del año. Trabajé duro, cumplí con mi deber y al final considero que se han dado buenos resultados, no solo en lo académico pero también en el proceso de aprendizaje y en lo personal.

Conclusiones Personales

El Senior Project es sin duda alguna una experiencia única y de gran valor para un estudiante de doce a mi parecer, sea de una u otra manera. Como proyecto le pude encontrar lo bueno y lo malo, y al sopesar estos dos factores puedo decir que es una experiencia que pienso debería continuar aplicándose en el colegio y la cual repetiría.

Primero lo bueno, aprendí bastante y de bastantes cosas a través de todos los marcos y la práctica. Creo que el conocer la legislación y la teoría que rodean la disciplina de mi elección es algo que puede otorgar perspectiva, y cuanto menos algún grado de cultura sobre el tema. Sin embargo, debo decir que esto fue lo menos importante para mí. Al fin y al cabo creo que en esta etapa final de formación lo esencial no es llenarme de conocimientos puntuales como los algoritmos de redes neuronales o la probabilidad binomial, sino el desarrollar habilidades que me permitan ser un mejor estudiante en la universidad, un mejor profesional en el futuro y ante todo un individuo más completo.

Estoy bastante seguro de que parte de lo investigado se verá condenado al olvido, en mi caso y en el de mis compañeros, pero creo que la naturaleza de un proyecto como este trasciende esa faceta. Las miles de palabras que redacté, la planeación que se debe tener para un proyecto tan extenso, incluso si fue impuesta, los viajes a la empresa y demás componentes dejaron su impronta así fuera de manera sútil. El escribir cada capítulo me parece algo muy valioso, el saber escribir es en sí el saber pensar, y el saber pensar es lo más importante para mí. El saber planear da un timón a cada uno para que escoja su rumbo, sea para algo tan simple como separar un proyecto en capítulos más digeribles, o para trabajar en el proyecto más importante en el que jamás se trabajará, la vida de cada uno. Los días de práctica me otorgaron perspectiva sobre un nuevo mundo de oficina que no me gustó demasiado, pero ante todo me dejaron una experiencia.

Ahora lo malo, realmente no disfruté de muchos aspectos del proyecto y en muchas ocasiones lo sentí algo tedioso y aburrido. Hacer un marco teórico no tiene nada de divertido para mí, ni leer extensos documentos legales, eso lo tengo por seguro. Creo que el enfocarme en conocer el qué y el cómo de mi pregunta no fue demasiado valioso. Hubiera preferido mil veces usar las varias horas que me tomó desarrollar algo como el marco teórico en aprender a hacer lo que estaba investigando, así fuera a un nivel elemental. Lo aprendido de los algoritmos que investigué o los tipos de mercadeo sobre los que indagué no me entusiasman demasiado

con respecto a la carrera, contrariamente las experiencias prácticas como el AP Computer Science me otorgaron mucha más claridad con respecto a mi futuro que este proyecto, el cual en mi caso no tuvo un efecto significativo en mi decisión.

Para concluir, el Senior Project me pareció muy bien organizado. La carga académica es muy razonable y se ven evidenciados los muchos años de desarrollo y mejoría. Todo fue muy bien explicado desde el principio del año y hubo transparencia con respecto a factores específicos como fechas de entrega, al igual que con lo que se esperaba de cada estudiante. Fue una experiencia provechosa y puedo decir con confianza que cada persona se llevó algo del proceso, por muy pequeño que ese algo pueda ser.

Lista de Referencias

Algoritmos de agrupación: qué son y cuándo se utilizan - The Black Box Lab. (2022, 30 junio). The Black Box Lab. Recuperado de: https://theblackboxlab.com/2022/06/17/algoritmos-de-agrupacion-machine-learning/

Árbol de decisión en Machine Learning - sitiobigdata.com. (2023, 14 septiembre). sitiobigdata.com. Recuperado de: https://sitiobigdata.com/2019/12/14/arbol-de-decision-en-machine-learning-parte-1/

Chavez, J. (2023, 20 marzo). Lenguaje de programación: Qué es, tipos y características. *Ceupe*. Recuperado de: https://www.ceupe.com/blog/lenguaje-de-programacion.html

Congreso de Colombia. (2009, 5 enero). *Ley 1273 De 2009 por medio de la cual se modifica el Código Penal, se crea un nuevo bien jurídico tutelado - denominado "de la protección de la información y de los datos"- y se preservan integralmente los sistemas que utilicen las tecnologías de la información y las comunicaciones, entre otras disposiciones.* Sistema Único de Información Normativa. Recuperado 31 de marzo de 2024, de https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Leyes/1676699

Congreso de Colombia. (2011, 12 octubre). *Ley 1480 de 2011 por medio de la cual se expide el Estatuto del Consumidor y se dictan otras disposiciones.* Sistema Único de Información Normativa. Recuperado 27 de noviembre de 2023, de https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Leyes/1681955

Congreso de Colombia. (2012, 18 octubre). *Ley 1581 De 2012 por la cual se dictan disposiciones generales para la protección de datos personales*. Sistema Único de Información Normativa. Recuperado 27 de noviembre de 2023, de https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Leyes/1684507

Congreso de la República. (2008, 31 diciembre). *Ley 1266 de 2008 por la cual se dictan las disposiciones generales del hábeas data y se regula el manejo de la información contenida en bases de datos personales, en especial la financiera, crediticia, comercial,*

de servicios y la proveniente de terceros países y se dictan otras disposiciones. Sistema Único de Información Normativa. Recuperado 27 de noviembre de 2023, de https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Leyes/1676616

Congreso de la República. (2014, 6 marzo). *Ley 1712 de 2014 por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones.* Sistema Único de Información Normativa. Recuperado 27 de noviembre de 2023, de https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Leyes/1687091

Congreso de la República de Colombia. (2021, 6 diciembre). *Ley 2162 De 2021 por medio de la cual se crea el Ministerio de Ciencia, Tecnología e Innovación y se dictan otras disposiciones.* Sistema Único de Información Normativa. Recuperado 31 de marzo de 2024, de https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Leyes/30043653

Cómo funciona Reducción de dimensión—ArcGIS Pro | Documentación. (s. f.). Recuperado de: https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-statistics/how-dimension-reduction-works.htm

Data storytelling is not storytelling with data: A framework for storytelling. . .: EBSCOhost. (s. f.). Recuperado de: https://web.p.ebscohost.com/ehost/detail/detail?vid=4&sid=745dd98f-7278-4e7e-b505-e010537f935e%40redis&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#AN=152511150&db=aqh

Duque, I., Lombana, M., Valderrama, C., Mayolo, Á., Cressien, T., Muñoz, V., Botero, A., Alvis, N., & Oviedo, J. (2022, 28 julio). *Decreto 1389 de 2022 por el cual se adiciona el Título 24 a la Parte 2 del Libro 2 del Decreto Único 1078 de 2015, Reglamentario del Sector de Tecnologías de la Información y las Comunicaciones, con el fin de establecer los lineamientos generales para la gobernanza en la infraestructura de datos, y se crea el Modelo de gobernanza de la infraestructura de datos.* Sistema Único de Información

Normativa. Recuperado 27 de noviembre de 2023, de https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Decretos/30044499

Gaviria, C., & Montenegro, A. (1991, 8 febrero). *Decreto 393 De 1991 por el cual se dictan normas sobre asociación para actividades científicas y tecnológicas, proyectos de investigación y creación de tecnología*. Sistema Único de Información Normativa. Recuperado 2 de abril de 2024, de https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Decretos/1088238

H. Parraga, V., & J. Yntusca, R. (2023, 13 abril). *Visualización de Datos Trabajo De Suficiencia Profesional*. Repositorio Académico UPC. Recuperado de: https://repositorioacademico.upc.edu.pe/handle/10757/668390

Hojas, I. M. (2021, 17 enero). *Agrupación por K-medias*. StatDeveloper. Recuperado de: https://www.statdeveloper.com/agrupacion-por-k-medias/

Huet, P. (2023, 24 octubre). Qué son las redes neuronales y sus aplicaciones. *OpenWebinars.net*. Recuperado de: https://openwebinars.net/blog/que-son-las-redes-neuronales-y-sus-aplicaciones/

Lenguaje R, ¿Qué es y por qué es tan usado en big data? (2019, 29 noviembre). *UNIR*. Recuperado de: https://www.unir.net/ingenieria/revista/lenguaje-r-big-data/

Los algoritmos más usados en machine learning. (2023, 24 julio). Nodd3r. Recuperado de: https://nodd3r.com/blog/los-algoritmos-mas-usados-en-machine-learning

Los Editores de Encyclopaedia Britannica. (2024, 1 febrero). *Microsoft Excel | Description & History*. Encyclopedia Britannica. Recuperado de: https://www.britannica.com/technology/Microsoft-Excel

Marín, R. (2023, 14 abril). *Aprovecha al máximo las herramientas de análisis de datos en Excel*. Canal Informática y TICS. Recuperado de: https://www.inesem.es/revistadigital/informatica-y-tics/analisis-de-datos-en-excel/

Muñoz, R. S., & García, M. (2023, 23 mayo). *El origen y evolución de la ciencia de datos (Data science)*. Fundación iS+D. Recuperado de: https://isdfundacion.org/2021/07/02/el-origen-y-evolucion-de-la-ciencia-de-datos-data-science/

Ortega, K. (2023, 8 agosto). *¿Qué es el lenguaje de programación Scala? - Saint Leo University*. Saint Leo University. Recuperado de: https://worldcampus.saintleo.edu/noticias/sistemas-computacionales-desarrollo-de-software-que-es-el-lenguaje-de-programacion-scala

Patrutiu-Baltes, L. (2016). *View of Inbound marketing - the most important digital marketing strategy*. Bulletin of theTransilvania University of Braşov. Recuperado de: https://webbut.unitbv.ro/index.php/Series_V/article/view/3946/3116

Principales algoritmos utilizados | Aprende Machine Learning. (2022, 15 diciembre). Aprende Machine Learning. Recuperado de: https://www.aprendemachinelearning.com/principales-algoritmos-usados-en-machine-learning/

Publicidad: Qué es, elementos e importancia. (2022, 12 mayo). *Ceupe*. Recuperado de: https://www.ceupe.com/blog/publicidad.html

¿Qué es Java? - Explicación del lenguaje de programación Java - AWS. (s. f.). Amazon Web Services, Inc. Recuperado de: https://aws.amazon.com/es/what-is/java/

¿Qué es la ciencia de datos? | IBM. (s. f.). Recuperado de: https://www.ibm.com/es-es/topics/data-science

¿Qué es Python? - Explicación del lenguaje Python - AWS. (s. f.). Amazon Web Services, Inc. Recuperado de: https://aws.amazon.com/es/what-is/python/

¿Qué es SQL? - Explicación de Lenguaje de consulta estructurado (SQL) - AWS. (s. f.). Amazon Web Services, Inc.Recuperado de: https://aws.amazon.com/es/what-is/sql/

¿Qué es un IDE? - Explicación de los entornos de Desarrollo Integrado - AWS. (s. f.). Amazon Web Services, Inc.Recuperado de: https://aws.amazon.com/es/what-is/ide/

S. Orjuela, L., & M. Chaparro, A. (2008). *Perfil Del Consumidor Y Comportamiento De Compra En La Tienda La Riviera Del Centro Comercial "El Retiro"*. Repositorio Institucional Javeriano. Recuperado de: https://repository.javeriana.edu.co/bitstream/handle/10554/9229/tesis317.pdf

SPSS Software | IBM. (s. f.). IBM. Recuperado de: https://www.ibm.com/spss

The ethical implications of big data analytics. (2023, 30 octubre). *IABAC®*. Recuperado de: https://iabac.org/blog/the-ethical-implications-of-big-data-analytics#:~:text=The%20Ethical%20Dilemmas%20of%20Big,privacy%2C%20security%2C%20and%20bias

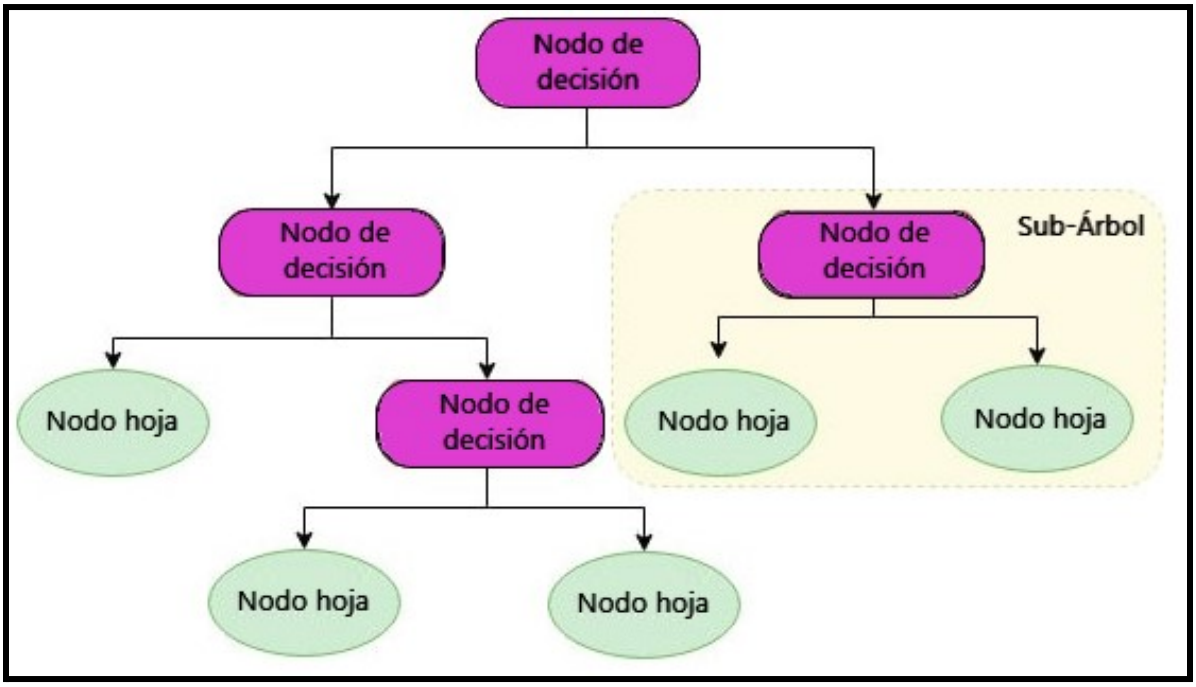
What is SEO? (Search Engine Optimization). (2021, 4 octubre). Michigan Technological University. Recuperado de: https://www.mtu.edu/umc/services/websites/seo/what-is/

What is tableau? (2024). Tableau. Recuperado de: https://www.tableau.com/why-tableau/what-is-tableau

Lista de Figuras

Figura 1

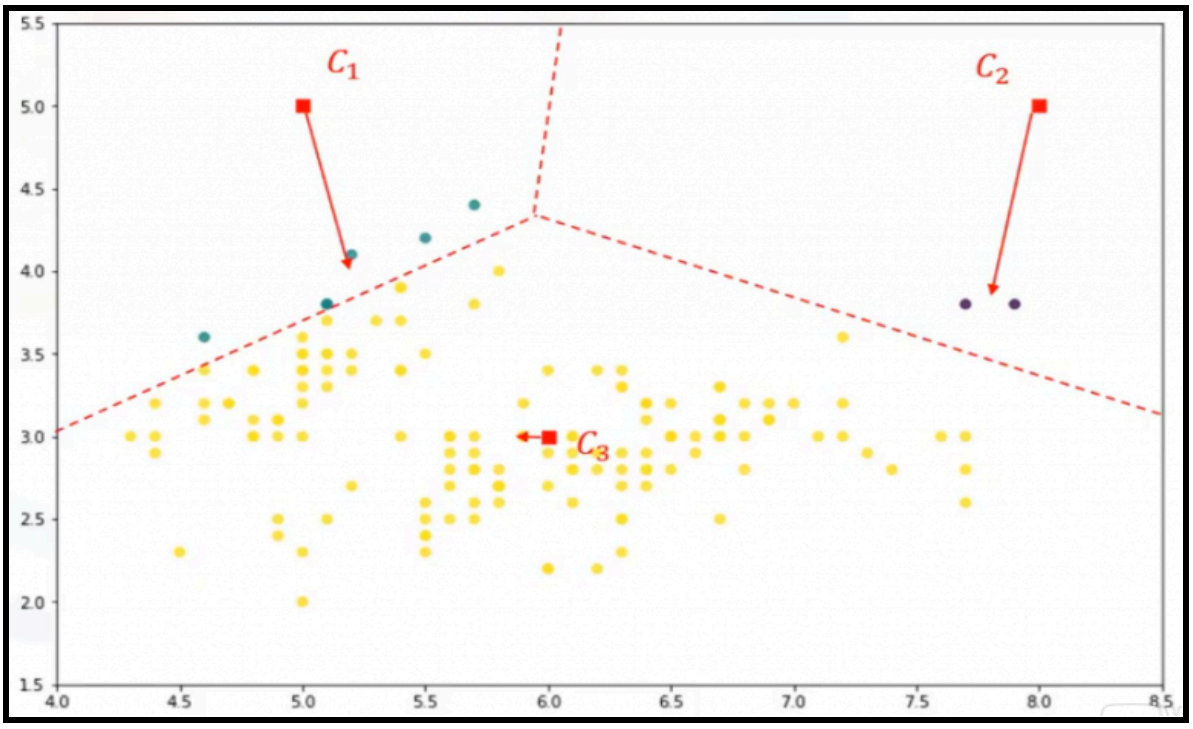
Árbol de decisión



Nota: Los nodos hoja representan la predicción final a la que se llega según las distintas ramificaciones del árbol. Adaptado de *Árboles de decisión en…, 2024,* sitiobigdata.com, https://sitiobigdata.com/2019/12/14/arbol-de-decision-en-machine-learning-parte-1/.

Figura 2

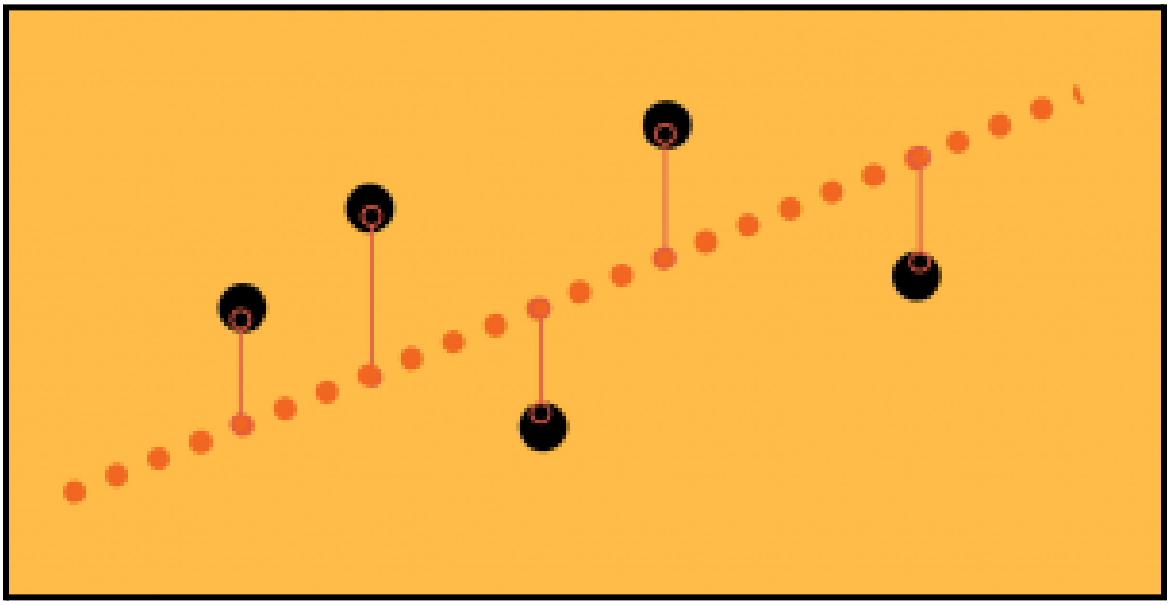
Agrupación por K-medias



Nota: La figura corresponde al producto de un algoritmo de K-medias de 3 conjuntos. C1, C2 y C3 son los centroides de cada grupo y de ellos se derivan las particiones, es decir las líneas punteadas de color rojo. Se organizan los puntos de acuerdo a la distancia de cada centroide. Recuperado de *Agrupación por K-medias…, 2021,* StatDeveloper, https://www.statdeveloper.com/agrupacion-por-k-medias/

Figura 3

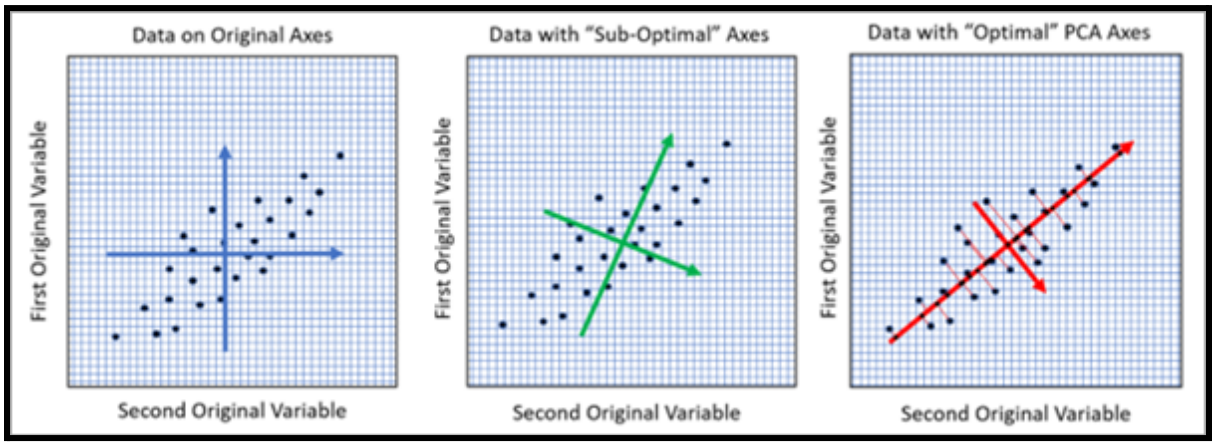
Representación de regresión lineal



Nota: Se representa una regresión por el método de mínimos cuadrados, donde los puntos negros representan datos reales, y la línea ideal se establece al minimizar los residuos, es decir la distancia de los puntos a la misma. El valor de la y para la función en los distintos valores de x, es el valor que predice este algoritmo y que se usa para llevar a cabo los análisis subsecuentes. Tomado de *Principales algoritmos utilizados…, 2022,* Aprende Machine Learning, https://www.aprendemachinelearning.com/principales-algoritmos-usados-en-machine-learning/.

Figura 4

Transformación de datos con PCA

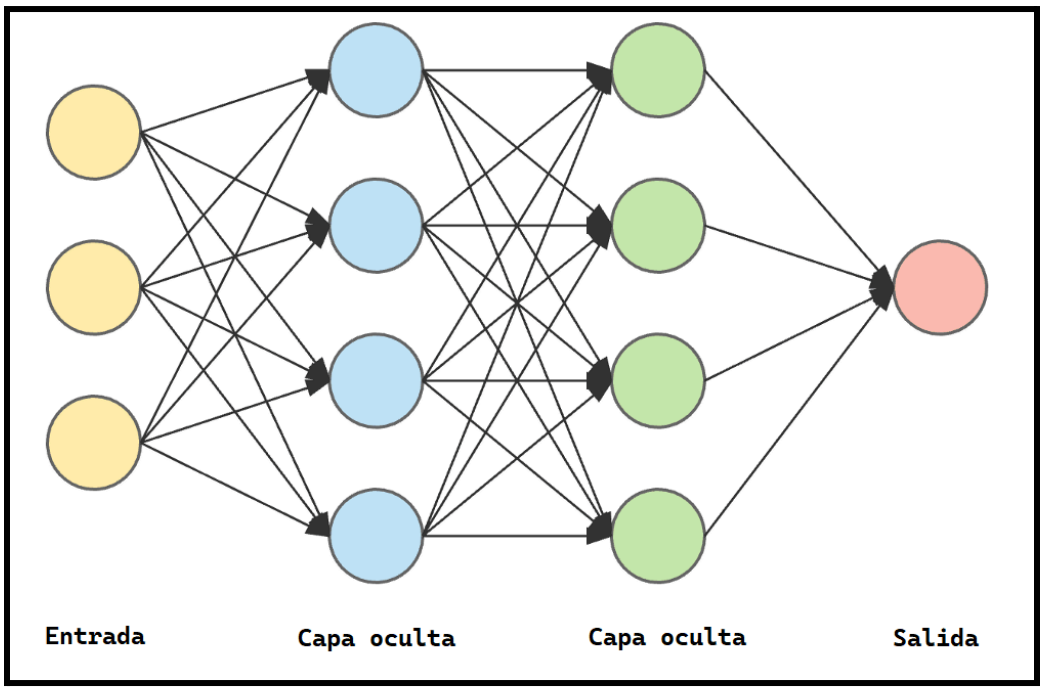


Nota: Se observan los efectos de ejecutar un Análisis de Componentes Principales (PCA) para revelar solo los puntos más relevantes. Los ejes de cada gráfico son indicativos de la varianza de las variables y son modificados por la rotación de los mismos, la cual busca ser más representativa de la relación entre ambos componentes. Adaptado de *Cómo funciona Reducción…,* (s. f.),

https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-statistics/how-dimension-reduction-works.htm

Figura 5

Estructura general de redes neuronales



Nota: Se muestran las tres capas de las redes neuronales, la entrada alberga los datos principales, la sección oculta se encarga de desglosarlos y manipularlos, mientras que la salida retorna el resultado final. Es enfatizar que todos los nodos se encuentran conectados entre sí. Recuperado de *Qué son las redes neuronales y sus aplicaciones*, 2023, OpenWebinars.net, https://openwebinars.net/blog/que-son-las-redes-neuronales-y-sus-aplicaciones/

Figura 6

Intensidad de la correlación según su coeficiente

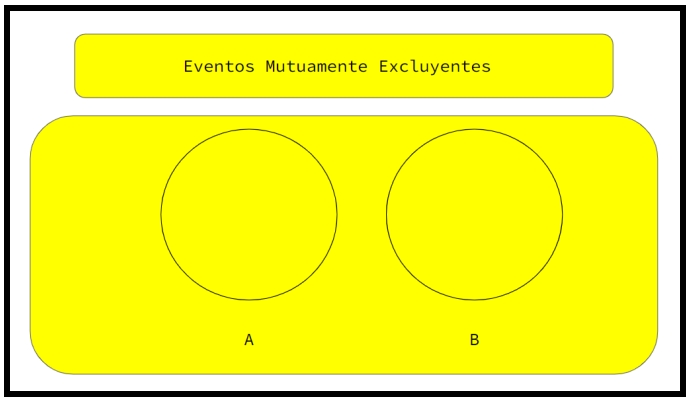
Coeficiente de correlación (r)	Correlación
1 | -1	Perfecta
0.7 < r < 1 | -0.7 < r < -1	Fuerte

0.3 < r < 0.7 | -0.3 < r < -0.7	Moderada
0 < r < 0.3 | 0 < r < -0.3	Débil
0	No correlación

Nota: Valores del coeficiente de correlación mayores a 0 tendrán una asociación positiva, y mientras más cercanos sea este a 1 o -1 tendrá una correlación de mayor intensidad, es decir el cambio en una variable produce mayor efecto en la otra. La figura 6 fue realizada por el autor utilizando información de su comunicación personal con S. A. Meza.

Figura 7

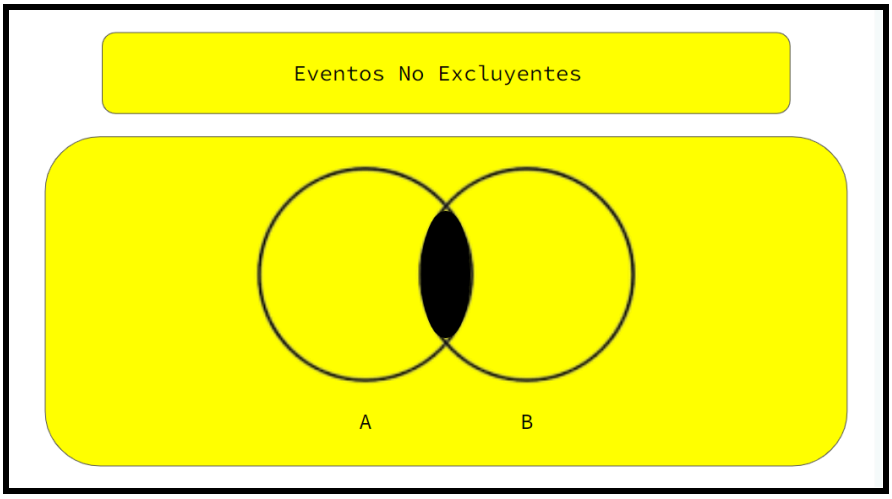
Representación de eventos mutuamente excluyentes



Nota: La esfera izquierda representa un evento A y la derecha un suceso B. Se entiende que todo aquello encapsulado por la esfera es una probabilidad favorable de que se de dicho caso. Al no estar superpuestos los diagramas, se da por hecho que la P(A ∩ B) es = 0. La figura 7 fue realizada por el autor utilizando información de su comunicación personal con L. Santamaría.

Figura 8

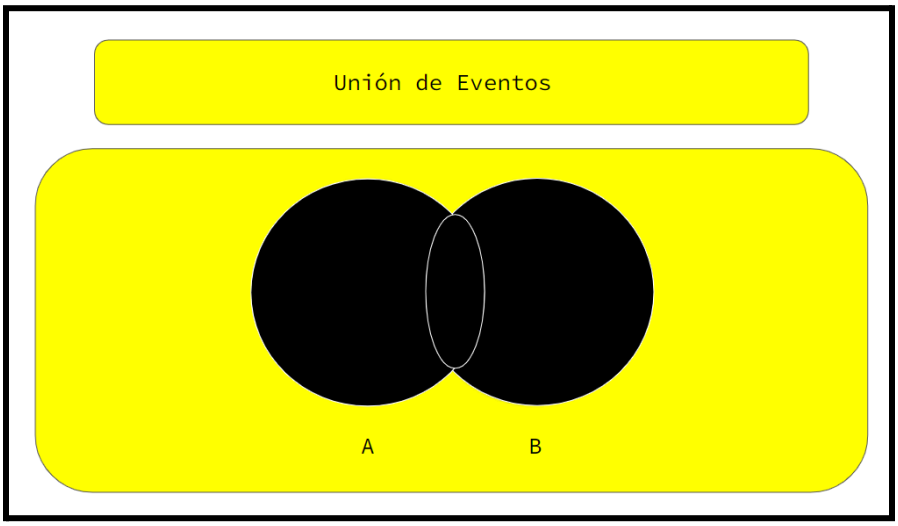
Representación de eventos no excluyentes



Nota: El área sombreada de negro representa la superposición de las esferas, que se entiende como su intersección (A ∩ B), es decir el caso de que se den los eventos A y B. La figura 8 fue realizada por el autor utilizando información de su comunicación personal con L. Santamaría.

Figura 9

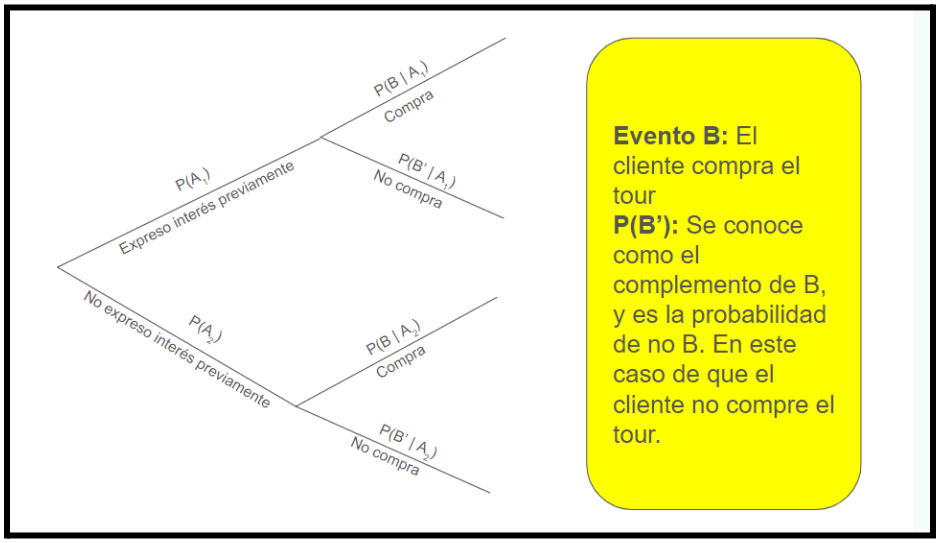
Representación de la unión de eventos



Nota: Se sombrea la totalidad de las dos esferas para plasmar la unión, ya que esta representa la probabilidad de que ocurran ambos eventos, es decir se contempla la posibilidad de solamente A, solo B, y A y B. La figura 9 fue realizada por el autor utilizando información de su comunicación personal con L. Santamaría.

Figura 10

Diagrama de árbol condicional



Nota: La figura 10 es una representación gráfica del ejemplo del autor en el párrafo superior y fue realizada por el mismo utilizando información de su comunicación personal con L. Santamaría.