Analítica de datos y visualización

Breve descripción:

El presente componente formativo aborda aspectos generales y claves sobre elementos, métodos y herramientas empleados para el desarrollo de reportes y tableros, a partir de los datos. Con su estudio responsable, el aprendiz se afianzará en fuentes de datos, transformación, machine *learning* y desarrollo de gráficos, usando datos nativos y cálculos con lenguajes de consulta.

Julio 2024

Tabla de contenido

Introducción 1

1. Modelos y metodologías de analítica 3

2. Aprendizaje automático *(machine learning)* 5

2.1. Aprendizaje supervisado 7

2.2. Aprendizaje no supervisado 10

2.3. Algoritmos y modelos matemáticos 11

2.4. Predictores 16

3. Herramientas de analítica de datos y visualización 19

4. Gestión de datos masivos 21

4.1. Lenguajes de consulta 23

4.2. Manipulación de datos 25

5. Tableros estadísticos, informes y resultados de visualización 26

Síntesis 30

Material complementario 31

Glosario 32

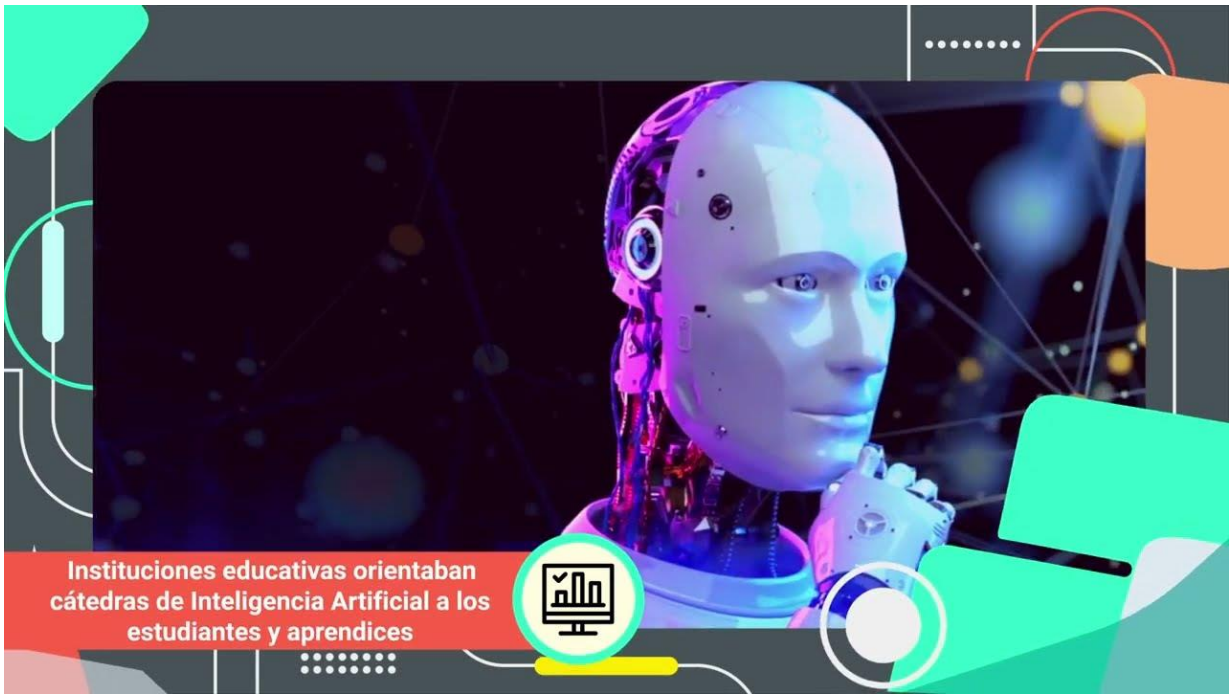
Referencias bibliográficas 34

Créditos 35

Introducción

Se da la bienvenida al estudio del componente formativo “Analítica de datos y visualización”. Para comenzar satisfactoriamente esta experiencia de aprendizaje, se le invita a explorar el video que se muestra a continuación. ¡Adelante!

Video 1. Introducción Analítica de datos y visualización



Enlace de reproducción del video

Síntesis del video: Introducción Analítica de datos y visualización
La Inteligencia Artificial (AI) y visualización de datos serán los principales temas del componente formativo. En relación con la Inteligencia Artificial, en los últimos años, se ha visto como un elemento lejano, lleno de complejidad, alta tecnología y difícil de implementar; si bien las instituciones educativas orientaban cátedras de

Inteligencia Artificial a los estudiantes y aprendices de formación tecnológica y profesional, el desarrollo y la aplicación eran relativamente escasos. Actualmente, estamos inmersos en estas tecnologías en todo momento, ha sido parte fundamental para el impulso de la cuarta revolución industrial y su acceso y desarrollo están al alcance de los profesionales de las industrias 4.0, es decir, programadores, analistas, científicos de datos, gestores de información, empresas, entre otros. En relación con la visualización de los datos, esta es quizás el principal objetivo de la analítica y todos los procesos técnicos; es el resultado final, donde se lleva a cabo la sinopsis de los procesos y resultados del negocio, para que el nivel de decisión corporativo conozca de manera acertada y actualizada, los procesos o estados de la organización para la toma de decisiones, fundamentadas en el conocimiento profundo del negocio. De esta manera, se generan acciones encaminadas a mejorar o corregir las estrategias u operaciones de la organización. La analítica de datos e inteligencia de negocios no es un asunto de moda o simplemente de entrar a una transformación digital corporativa, la inteligencia de negocio es una inversión importante, donde se conectan la tecnología, los procesos y los datos con quienes direccionan las empresas y quienes toman y asumen las decisiones del negocio. Siempre se debe tener claridad en el valor corporativo adicional que se obtendrá al implementar proyectos de esta naturaleza. Asimile este componente formativo con todos sus temas y afiance sus habilidades para apropiarse del proceso de gestión de datos, como las fases, arquitecturas, tecnologías, modelos y los resultados que las organizaciones esperan para generar valor empresarial, a partir del manejo y gestión de la información proveniente de diversas fuentes de la organización, y acercarse a cifras predictivas y conceptos de la analítica prescriptiva. ¡Éxitos en el proceso!

1. Modelos y metodologías de analítica

A partir del contexto de la inteligencia de negocio, se debe comenzar por entender bien los requerimientos específicos de las soluciones analíticas; si bien toma muchos elementos del desarrollo del *software* e incluso, se toma como una rama del desarrollo de aplicaciones informáticas, hay ciertas especificidades en los requisitos para la toma de decisiones que se deben asumir.

Los desarrollos de *software* se limitan a cumplir los requisitos de un área o un proceso determinado, como una isla en la empresa, donde no se integra información, ni formatos. Hoy en día, aún las empresas usan sistemas para cada área, por ejemplo, contabilidad, recursos humanos, área financiera, ventas, ofertas, *marketing*, producción, etc.

En los modelos y metodologías de analítica, se debe tener presente:

Manejo integral

Cada área, puede manejar su sistema de información por aparte, sin tener en cuenta integraciones o calidad de la información, generando muchísimos datos, pero sin explotar de manera óptima su riqueza escondida, la cual pudiera tener si se gestiona de manera integral y usando técnicas y tecnologías de la cuarta revolución industrial.

Implementación con propósito

Al momento de pensar en una solución de analítica de negocio, es fundamental tener todas las preguntas que tanto el nivel de decisión, como el nivel operativo requieren.

Conciencia institucional

Cuando las organizaciones no tengan claridad sobre el uso del BI en el negocio, es necesario concientizar a los directivos que muchos problemas de la organización se deben a la falta de datos al instante, actualizados y que reflejen una realidad.

Oportunidad y practicidad

Si se toma como ejemplo, una feria de negocios importante, donde se presentan grandes oportunidades de conectar con clientes grandes, se espera que lleguen representantes de otras firmas; mientras, de manera tradicional, se trabaja confiando en la memoria y vagamente el ejecutivo de ventas recuerda que alguna vez su empresa ya había hecho negocios con este cliente que se acerca.

Actualización y equipamiento para la labor

Cuando las empresas tienen implementado un sistema de inteligencia de negocios, es de gran valor comercial, que el representante de ventas de la organización tenga en su dispositivo electrónico (celular, tableta, etc.), los datos de todos los clientes y sus historiales para que, con una búsqueda muy rápida, tenga de inmediato en pantalla, información que podría ser clave para conectar con nuevos clientes.

Una de las condiciones iniciales más importantes son las preguntas, es decir, qué se necesita saber del negocio, qué decisiones se planean tomar y qué insumos se requieren para la toma de estas decisiones.

En sí, se deben seleccionar varios elementos para el desarrollo de soluciones analíticas, tales como:

Metodología de desarrollo

Se recomienda elegir una de las metodologías ágiles y registrar todos los requerimientos y planeación, según el rol que cumpla en el proyecto.

Elegir herramienta de control y trazabilidad

Para llevar control y trazabilidad de todo el desarrollo del proyecto en relación de equipos de trabajo, funciones, actividades hechas, en proceso y pendientes. Trello y Jira son buena opción, existiendo otras más.

Herramientas de desarrollo

Si bien se ha hablado mucho de infraestructura y herramientas, en este caso habrá que concentrarse en la capa de visualización de datos, según la herramienta elegida, dando ciertas pautas para seguir el modelo de diseño.

2. Aprendizaje automático *(machine learning)*

En la vida cotidiana de un hogar promedio actual, podría ocurrir fácilmente que un niño de escuela, mientras hace sus tareas, pregunte en voz alta: "¿Cuántos departamentos tiene Colombia?". El padre, quien tradicionalmente debería saber la respuesta, recuerda que son 33 departamentos. Sin embargo, mientras se asegura de su respuesta, el dispositivo electrónico Alexa interrumpe primero, proporcionando la respuesta correcta de 32 departamentos.

Adicionalmente, amplía esta información con algunos datos complementarios. En este instante, la inteligencia artificial, es tomada como fuente de información rápida y confiable. Es así como los padres y profesores, que han sido tradicionalmente las

fuentes de conocimiento, van siendo desplazados en este sentido (para adquirir conocimientos), y empiezan otros retos y roles igual de importantes que deben asumir.

En relación con el aprendizaje automático, es importante tener en cuenta aspectos como:

Avance y actualización permanentes

La Inteligencia Artificial seguirá avanzando, las organizaciones y personas las seguirán asumiendo y consumiendo en su diario vivir, debido a que se encuentra en celulares, carros, compras, gestiones de gobierno, app, bancos, medios de comunicación, etc.

Apropiación y adaptación de las personas

La mayoría de las empresas y personas, en medio de la inmersión de diario vivir, ocasionan aumento de las habilidades digitales en los individuos, muchas veces de manera natural, sin enterarnos, esto nos hace más exigentes para la solución de problemas, gestión de procesos, así como conocedores de la tecnología.

Mejoras industriales

Los avances empresariales bajo el contexto de la cuarta revolución industrial, se van entendiendo mejor, aceptando y asumiéndolos en sus organizaciones. Algo similar debió pasar cuando llegó la electricidad al mundo, algunas personas y organizaciones no lo veían necesario sino complementario, como piensan actualmente, de la tecnología, algunas personas. No obstante, desde hace muchas décadas, la electricidad no es un elemento que se debata en un hogar o una organización.

La IA como representación de la inteligencia humana

La IA está fundamentada en una serie de algoritmos y métodos inspirados en procesos propios del cerebro, es decir, la programación se basa en entradas de datos. El *software* se programa para realizar procesos y asociar datos en tablas y organizar, sumar o dar reportes. Es un asunto más mecánico y de ejecutar pasos establecidos.

Inteligencia con poder de decisión, aprendizaje y predicción

En cierta manera, los cálculos y guardar datos también son características del cerebro, sin embargo, existen otras donde la programación ha avanzado ostensiblemente en los últimos años. Si el algoritmo tiene poder de decisión, aprendizaje, predicción, usa lenguaje natural humano y automatiza tareas, es un componente de IA.

Frecuencia y utilidad

La IA es una rama de las ciencias informáticas, si bien, la tecnología en analítica de datos no se enfoca al desarrollo de redes neuronales o aprendizajes de máquina, es interesante saber de qué se trata, porque, aunque no se programen estos sistemas, es una realidad el uso frecuente y útil de herramientas sofisticadas a las cuales conectamos nuestros datos, y nos pueden dar conocimientos importantes a partir de la IA.

2.1. Aprendizaje supervisado

Machine Learning (ML) es el área de las ciencias computacionales que hace parte de la IA, su enfoque es que las computadoras, en vez de ser programadas paso a paso, aprendan a partir de los datos. Cada solución de ML es específica para cada necesidad, tal y como se tiene el enfoque con la programación convencional. Los profesionales de

ML están dedicados al desarrollo de algoritmos genéricos que pueden extraer patrones de diferentes tipos de datos.

El ML enfocado a la ciencia de datos apunta a desarrollar procesos específicos como la identificación de la fuente de datos, desechar información inválida o no útil, limpiar, normalizar, relacionar, datos sesgados, etc.

Nota importante. Todas estas tareas podrían encontrar solución eficiente en la selección de soluciones de *machine learning,* cuya aplicación resulte apropiada, la elección del algoritmo más adecuado, el ajuste de los parámetros del método elegido, el análisis de los resultados, la identificación de comportamientos incorrectos, la vuelta a procesos anteriores con el fin de cambiar lo que resulte necesario para mejorar los resultados. (Bobadilla, 2020).

Si bien la aplicación de *Business Intelligence* (BI) es una manera interesante para el conocimiento de las organizaciones y otras bondades, la implementación de ML es ir más allá. Preste atención a los aspectos clave que se presentan a continuación:

Orientación del *Machine Learning*

Se orienta a mejorar predicciones cada vez más precisas, obtener información más profunda de los datos, reducir sobrecarga de tareas y mejorar las experiencias de clientes, por ejemplo, a través de *Chatbots* que vayan aprendiendo de un humano a cómo responder según las situaciones.

Diferentes tipos de aprendizaje

Varían en función de si se conoce o no la respuesta que se busca, del tipo de datos analizados, del entorno de los datos en cuestión y del tipo de análisis realizado (estadísticas, comparaciones, reconocimiento de imágenes, etc.).

Algoritmos de aprendizaje

Los algoritmos de aprendizaje y la potencia de cálculo requerida también difieren en función de la tarea que se realiza.

Calidad de aprendizaje

Tal calidad depende del número de ejemplos relevantes que el *software* puede analizar (cuantos más ejemplos, más precisión se tendrá en el análisis de datos). También dependen de la cantidad de características que detallan los ejemplos (cuanto más sencillos y precisos, más rápido y acertado será el análisis: tamaño, peso, cantidad, velocidad, rangos, etc.).

Calidad de los datos

Si faltan muchos datos o se presentan falencias en las dimensiones de la calidad de datos, el análisis se verá afectado.

Cumplimiento máximo de criterios

El ML de predicción será más preciso y el análisis resultará más ajustado a la realidad. Así que, una vez que se hayan definido los objetivos y elementos de aprendizaje automático, y que las bases de datos estén en óptimas condiciones, podrá empezar a sacar el máximo partido al *machine learning.*

Principio del algoritmo

El tipo de aprendizaje está dado según el principio del algoritmo, sin decir que otro tipo de aprendizaje no es usado, este aprendizaje es el más implementado en la gestión de datos y otras aplicaciones y ha permitido gran ampliación en implementación de Inteligencia Artificial en las organizaciones y la vida cotidiana.

Fundamento del aprendizaje supervisado

Se fundamenta en el descubrimiento o el aprendizaje en la relación existente entre unas variables o datos de entrada y unas variables de salida, es decir, el aprendizaje surge de mostrarle a los algoritmos cuál es el resultado que se desea obtener para un determinado valor.

2.2. Aprendizaje no supervisado

Este paradigma de aprendizaje toma como base, únicamente, los datos de entrada, sin explicarle al sistema qué resultado es el que se espera obtener. Podría ser un poco difícil de concebir, porque si no hay una referencia previa, ¿de qué manera podrían los sistemas aprender?.

Este concepto es menos usado, pues sostiene mayores retos a la ciencia y a los algoritmos, donde a partir de un parámetro, el sistema deberá tratar de descubrir qué resultado o resultados posibles daría ese dato de entrada.

Estas son algunas generalidades que se deben tener en cuenta, respecto del aprendizaje no supervisado:

Tiene una ventaja, porque el entrenamiento en aprendizaje supervisado implicaría miles de horas, humanos enseñando y altos costos, debido a que para que un

sistema esté bien entrenado requiere al menos 100 mil ejemplos, esto es una tarea larga y costosa.

Por su parte, el aprendizaje no supervisado solo requiere de los datos de entrada, dar unos pocos parámetros de lo que se quiere y dejar todo a la máquina.

La desventaja es que requiere mucho procesamiento, puntos generales que pueda asociar a lo que se le parezca, y avance tecnológico con el fin que la máquina vislumbre y descubra el dato de salida.

Si bien, las técnicas ML más usadas están basadas en referencias de salidas, el aprendizaje no supervisado será el futuro, porque de cierta manera los sistemas usarán estas referencias y gran capacidad computacional para empezar a asociar.

Lograr que una máquina tenga sentido común es un objetivo muy difícil, sin embargo, los algoritmos actuales se van acercando un poco, por ejemplo, hay palabras que si bien es la misma para dos cosas o significados diferentes (palabras polisémicas), para los humanos es fácil según el contexto, pero para una máquina es difícil definir estas cosas.

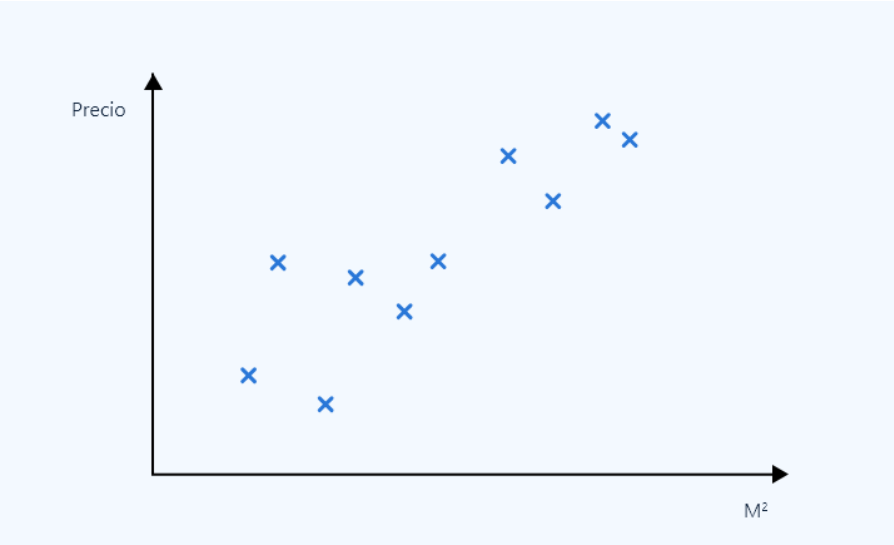
2.3. Algoritmos y modelos matemáticos

Los métodos matemáticos y algorítmicos empleados en la Inteligencia Artificial, pueden variar según los desarrollares. Aunque existen otros, los más comunes o típicos para **aprendizaje supervisado** son las Regresiones lineales y logísticas, Máquinas de vectores de soporte, árbol de decisiones y K-Media.

Regresiones lineales y logísticas

Estos algoritmos tienen un comportamiento basado en el historial de los datos, por ejemplo, se requiere predecir o asignar correctamente el valor de un inmueble; teniendo como base datos históricos, se tendría una gráfica como esta:

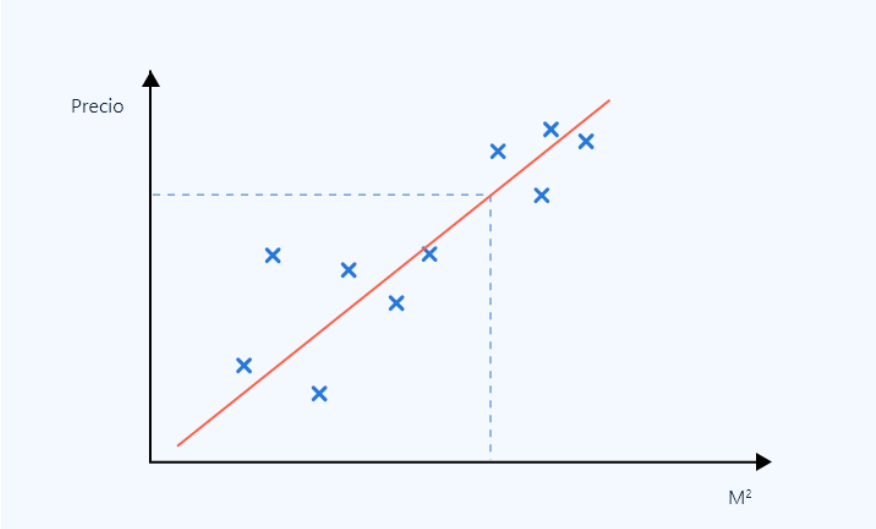
Figura 1. Datos base para los algoritmos regresiones lineales



Luego, a partir de una la regresión lineal, se calcula la línea que abarca los valores y se establece con mejor exactitud el valor del inmueble.

De esa manera, cuando una propiedad tiene ciertos metros cuadrados específicos, la salida que ofrece la máquina será un precio que coincida con la regresión lineal, como se expresa en la siguiente gráfica:

Figura 2. Línea a partir de valores conocidos



Con estas técnicas se pueden establecer grupos, por ejemplo, es posible definir si un inmueble es costoso o no lo es, con base en los datos y sus grupos.

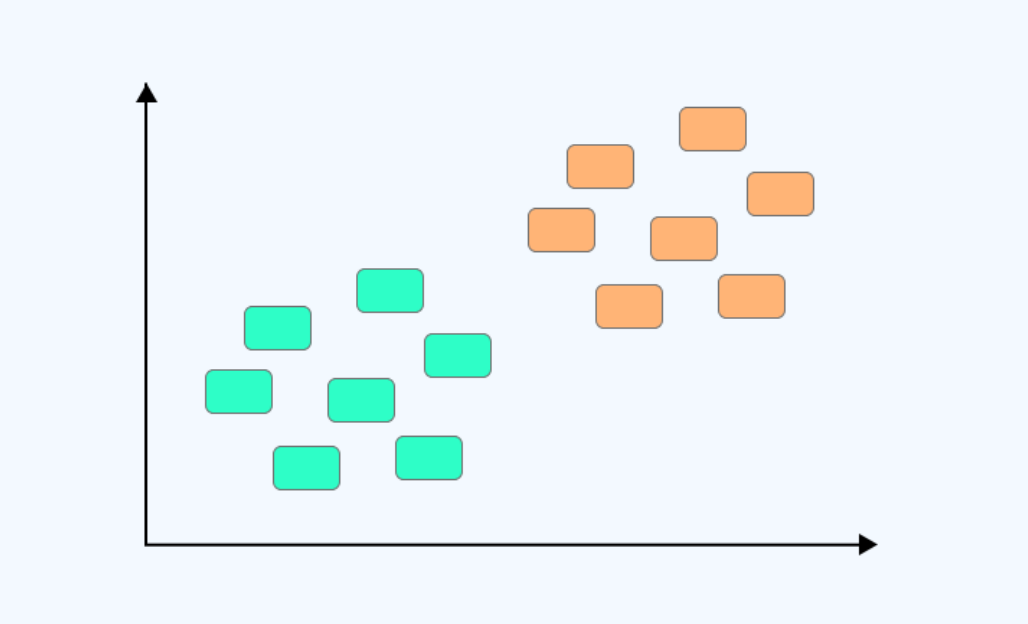
***K-MEANS cluestering o K-media* de agrupamiento**

En el caso anterior, se tenían valores históricos, sin embargo, ¿qué tal si se presenta el caso de que los datos no están categorizados y no hay un historial?.

Se deberá usar uno de los algoritmos o métodos no supervisados para que la máquina aprenda a identificar patrones y arroje respuestas.

El algoritmo por agrupación usa como base centroides o puntos de datos base, que procura detectar patrones similares y de esa manera identifica grupos, como se presenta a continuación.

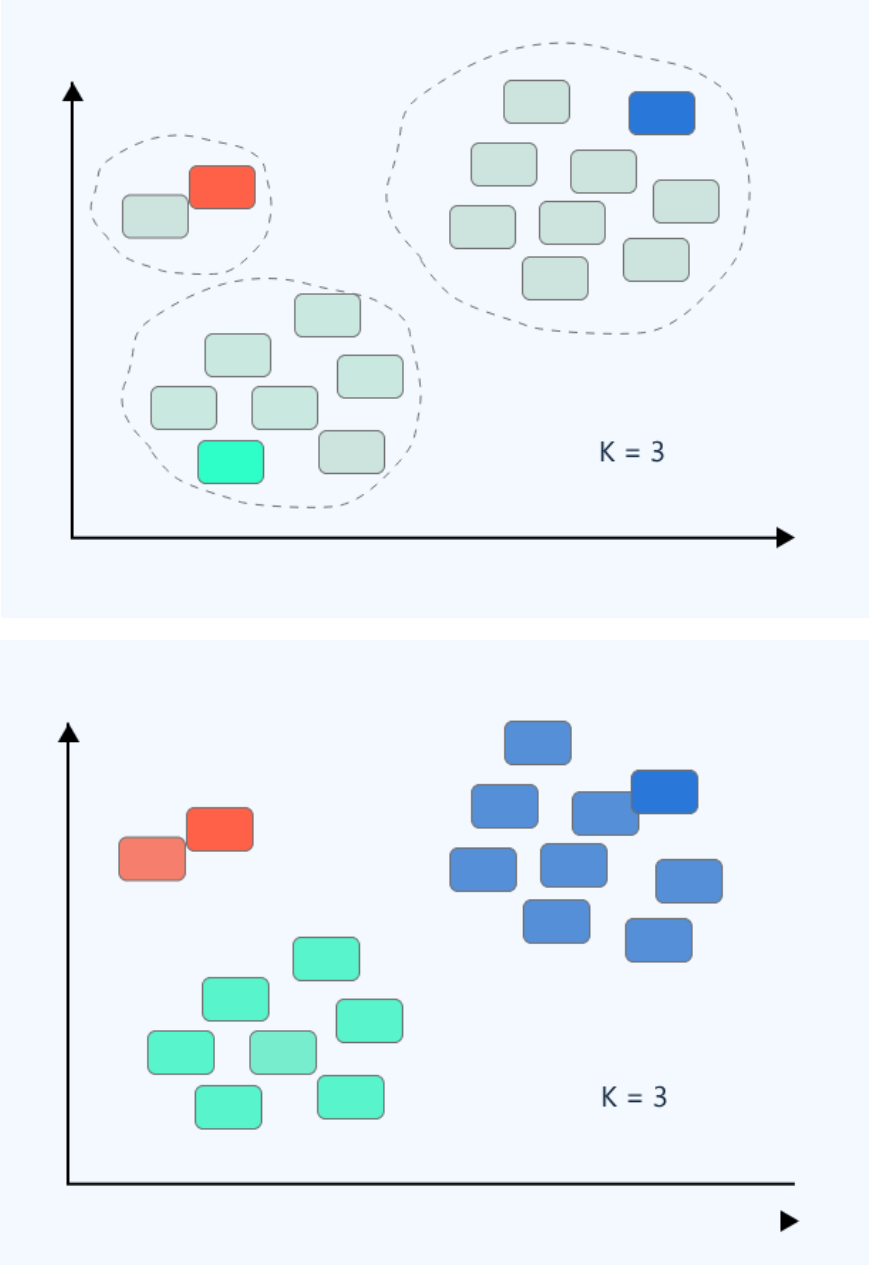
Figura 3. Agrupando (*clustering*)



En un primer momento, los datos no están clasificados, sin embargo, el algoritmo determina similitud según la distancia entre los mismos (en estos datos se nota que hay dos grupos por su cercanía).

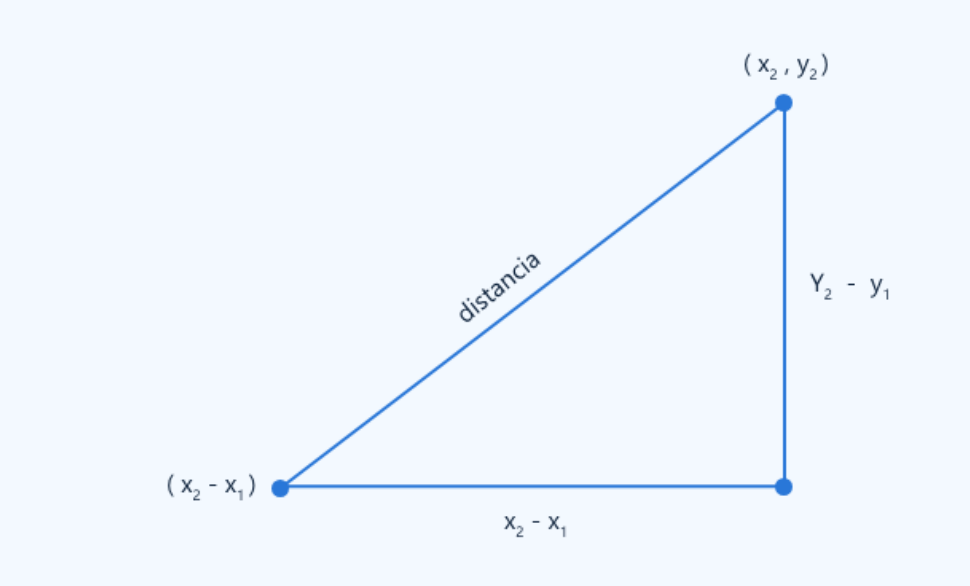
Para empezar a usar este tipo de algoritmos, se debe elegir el número de *clusters*, representado por K, Luego, aleatoriamente, se asignan centroides y se calcula, uno a uno, la cercanía de los datos al centroide, como se presenta en la siguiente figura.

Figura 4. Clusters y centroides



El algoritmo calcula estas distancias usando la **Distancia Euclidiana.** Se trata de una variación del teorema de Pitágoras y es una de las maneras más fáciles para calcular distancia, a partir de las posiciones de dos objetos en un plano cartesiano, como se presenta a continuación.

Figura 5. Distancia euclidiana



Esta función se repite las veces que sea necesario o se hayan parametrizado, para que el algoritmo autónomamente determine qué dato estaría en cuál *cluster* o qué clasificación.

Esto es solo un ejemplo de métodos o algoritmos que emplea el *machine learning,* pero hay muchos más tanto para aprendizaje supervisado como no supervisado.

2.4. Predictores

Adivinar el futuro ha sido uno de los factores más fascinantes que, a lo largo de toda la historia humana, ha ocupado a las civilizaciones y líderes, hasta el punto de desarrollar ansiedades por tal incertidumbre. Es por ello que es tan frecuente que el futuro sea un insumo para historias de cine, y atrapen y fascinen a todo el público, aparecen objetos como oráculos, bolas de cristal y profecías alrededor de múltiples *films* e historias.

Si bien la predicción, culturalmente, se contempla como saber el futuro, no necesariamente es así, la predicción enfatiza la capacidad de ver información oculta, ya sea en el pasado, el presente o el futuro, tal como se concibe la bola de cristal de algunas historias mágicas, donde este objeto no solo permite ver el futuro, sino también ver qué sucede en el presente en otros lugares; esta utilidad ya dejó de ser fascinante porque se convirtió en algo normal después del avance de los medios de comunicación, pero igual se sigue llamando predicción.

Predictores como funciones de herramientas

Un ejemplo de predictores usados por todos, ha sido la función autocompletar de Google en sus búsquedas y los teclados de celular donde, al escribir una palabra o parte de ella, el sistema procura autocompletar lo que se desea escribir, ahorrando tiempo y disminuyendo errores de digitación de manera ostensible. Esto se hace a partir de unas métricas y del aprendizaje que la máquina detectó por las tendencias propias del usuario o tendencias de búsqueda masiva localizada o mundial.

En las aplicaciones, también la predicción avanza según esta va aprendiendo; es el caso del traductor de Google: cuando esta herramienta era nueva, su traducción no tenía buena calidad, pero, actualmente, con el aporte de los mismos usuarios, las traducciones tienen en cuenta incluso ciertos contextos que la máquina ha aprendido a diferenciar.

En relación con los predictores, tenga en cuenta algunas generalidades como las que enuncian a continuación:

Resultado de una acción

La predicción no es un asunto nuevo; en su mínima expresión se podría resumir en el resultado de una acción, es decir, no se necesita ser gurú para saber que, si no se realizan ventas, el negocio se viene a pique, sin embargo, la IA es mucho más que esto y se responsabiliza de elementos más complejos.

Análisis predictivo

El objetivo de este análisis es la estimación de eventos futuros a partir de datos históricos o, incluso, descubrir fenómenos presentes que podrían desencadenar consecuencias más adelante, como, por ejemplo, la no satisfacción de clientes.

Construcción de algoritmos de predicción

Construir algoritmos o funciones de programación básicos no suele ser difícil, por ejemplo, en una ferretería se podría implementar un campo calculado que sugiera pedidos para comprar a proveedores, basado en la cantidad de productos vendidos en los últimos tres meses. Si bien, este es un componente que podría funcionar, no se considera exacto.

Datos entregados a modelos ML

Si estos datos se entregan a modelos ML de un proveedor de este servicio y se define el objetivo con claridad (número de unidades a comprar a proveedores), a medida que pasa el tiempo las predicciones de ventas y las decisiones de compra para el *stock* de inventarios serán cada vez más exactas, porque la IA tendrá muchísimas más variables para definir la decisión de compra. Además de las ventas, podría tener en

cuenta el mes, según comportamiento del mismo mes de los años anteriores, el precio de divisas, incluso si hay épocas de lluvia o no, etc.

Datos históricos

Para la construcción de estos modelos, una vez desarrollado el algoritmo predictivo (y reglas de negocio que apliquen), y configurado con claridad las metas y objetivos en el entorno del proveedor de IA, es necesario disponer de un conjunto de datos históricos. Por lo general, se tienen dos conjuntos: uno de datos de entrenamiento y otro de prueba.

Comprobación de exactitud

Al modelo se le pasan como entrada, los datos de entrenamiento para calibrar la predicción y, posteriormente, los datos de prueba. Después, se compara el resultado de la predicción con los valores reales (históricos) para comprobar su exactitud.

3. Herramientas de analítica de datos y visualización

Por practicidad y funcionalidad, las organizaciones deciden usar un ecosistema integrado que abarque todas las funcionalidades y extras de la inteligencia de negocio, desde el contexto de gestión de datos hasta el modelamiento y visualización de reportes. En este caso, la orientación estará enfocada a las herramientas propias para la visualización de datos.

Existen muchas opciones para las soluciones BI, las más populares son Tableau de la marca Salesforce, Power BI de Microsoft y otras como Qlik Sense de la empresa Qlik.

Sobre las herramientas de analítica de datos y visualización, tenga en cuenta los siguientes ítems:

Opciones en el mercado

En el mercado existe más de un centenar de posibilidades, es un sector de alto crecimiento y sus herramientas se van haciendo más especializadas, integrándose con IA, grandes capacidades de infraestructura y facilidad de manejo.

Complejidad de las organizaciones

La inteligencia de negocio no es una tecnología o una serie de herramientas ya establecidas, es decir, todo depende de qué tan grande y compleja sea la organización y qué cantidad de datos fluyen en los sistemas. Por lo que la inteligencia de negocio podría gestionarse desde la aplicación de Excel o Sheets de Google, hasta usar herramientas especializadas con proveedores como Amazon Web Services o Microsoft, entre otros.

Descarga e Instalación

Microsoft distribuye de manera gratuita la aplicación para conectar datos y generar reportes; esta aplicación se encuentra bajo el ecosistema de productividad Microsoft 365, la cual sí se cobra para algunas características extras, pero para efectos del aprendizaje, la versión descargable es suficiente y muy completa.

Fuente de datos

Este tipo de aplicaciones pueden ser muy sencillas y fáciles, pero también se podrán desarrollar *dashboards* muy complejos y robustos, según el nivel de conocimiento de la herramienta.

Diversidad de las fuentes de datos

Las fuentes de datos de Power BI son diversas; de manera nativa, el programa tiene decenas de opciones, que son las más comunes. Incluso, es tan compatible que, si existiera una fuente a partir de un desarrollo no comercial o poco común, da la posibilidad de programarlas y crearlas.

Desarrollo y representación gráfica de datos

Con la aplicación instalada y con la fuente de datos clara, la aplicación está lista para el desarrollo de informes. Es fundamental conocer la data y entender exactamente lo que se desea mostrar y qué interacciones con los datos son posibles de realizar.

Comportamiento de los tableros de mando

Los tableros de mando o *dashboards*, se comportan bajo un mismo contexto, es decir, cada elemento funciona como filtro o segmentador de datos. De esa manera, los datos tendrán mucha interacción y el usuario podrá filtrar y segmentar como a bien considere.

4. Gestión de datos masivos

La gestión del flujo de datos de extremo a extremo *(end to end),* es un término usado por algunas marcas y consiste en establecer la tubería por donde fluyen los datos; de esta manera, existe una toma de datos (extracción de datos origen), sigue su flujo haciendo filtraciones, depuraciones, limpieza y mejora en la calidad de los datos, para que lleguen a un gran estanque (bodega de datos), y luego puedan ser consumidos según los requerimientos del negocio.

Herramientas de desarrollo

Las soluciones *Open Source* es una sugerencia, al menos, para el proceso de aprendizaje, porque muchas organizaciones se alinean bajo ecosistemas de pago tales como Microsoft con *SQL Server Integration Services (SSIS), Qlik analityc* u otras herramientas comerciales.

Las herramientas *Open Source* cumplen con el principio de ser abiertas, es decir, que pertenecen o se matriculan a una comunidad que tiene acceso libre para su uso y participación del código fuente. Si bien son escasos los soportes técnicos, existen foros donde la comunidad misma ofrece su ayuda en temas o situaciones específicas.

¡Importante!

Muchas empresas usan este tipo de tendencias en las herramientas informáticas no por el ahorro del costo, sino también por su funcionalidad e impacto en el funcionamiento y utilidad.

Pentaho

Nacida en el año 2006, es una plataforma BI y se trata de una multiplataforma. Al ser un proyecto *Open Source,* hay una gran comunidad que brinda apoyo y ayuda con los errores que se puedan detectar en los proyectos desarrollados.

Es una plataforma integral con una suite de herramientas para completar las tareas y flujo de datos por toda la tubería de datos, transformación y carga.

Los principales componentes de Pentaho son:

- **Pentaho Server**

  Núcleo de la plataforma.

- **Pentaho Report Designer**

  Para desarrollar reportes a través de consultas de diversos orígenes de datos.

- **Pentaho Schema Workbench**

  Para administrar y crear cubos OLAP. Usa motor de datos llamado Mondrian, también de código abierto.

- **Pentaho Data Integration Kettle**

  Es una de las herramientas más usadas de Pentaho, permite crear procesos ETL diseñando y alimentando los *Data warehouse*.

Pentaho puede usarse como suite completa, es decir, usando todo el ecosistema de extremo a extremo o como componente individual de toda la solución BI, integrándose con otras herramientas.

Spoon (Pentaho Data Integration - Kettle)

Para poder realizar los procesos ETL, es necesario un entorno gráfico, eso es Spoon, una interfaz para realizar todos los procesos y tareas de extracción, transformación y carga de datos hacia las bodegas de datos.

4.1. Lenguajes de consulta

Los lenguajes de consulta son lenguajes de programación orientados al manejo y gestión de datos; cuando se desarrollan aplicaciones bajo lenguajes de programación como C++, Java, y otros, los lenguajes de consulta son incluidos en algunas funciones de las aplicaciones, sin embargo, dejar todo el trabajo de reportes y consultas en los desarrollos de *software*, podría resultar limitado y con problemas de desempeño, cuando se procesan grandes cantidades de datos.

A continuación, se describen los tres principales lenguajes de consulta que todo profesional de datos deberá tener en cuenta para desarrollar proyectos analíticos y transformación digital:

Es el principal lenguaje de consultas, su curva de aprendizaje es muy ágil, es decir, es un lenguaje de consultas y gestión de datos relativamente fácil de aprender. Todas las bases de datos relacionales están basadas en SQL, como principio de consulta.

Ha sido muy usado para aplicaciones de grandes volúmenes de datos, es usado para computación estadística y gráfica, porque cuenta con funcionalidades matemáticas y estadísticas muy importantes. Es el lenguaje usado de manera nativa en herramientas como Microsoft Query, usado para la conexión, consultas y transformación de fuentes externas de datos para las aplicaciones como Excel y Power BI.

Sin duda, es la tendencia para la aplicación de ciencia de datos y funcionalidades de Inteligencia Artificial. Cuenta con múltiples librerías que lo vuelven en una herramienta relativamente simple de aplicar y aprender, pero con potenciales de

procesamiento de datos, casi obligada para quienes desean incursionar en la ciencia de datos.

Lenguaje nativo de la herramienta Power BI de Microsoft. Cuenta con múltiples funcionalidades para no poner límites a la herramienta de inteligencia de negocios Power BI. Si se desea ser experto en reportes y *dashboard,* el manejo de este lenguaje es, sin duda, un elemento para dominar.

A diferencia de los lenguajes utilizados para el desarrollo de aplicaciones, la mayoría de estos lenguajes son interpretados. Esto significa que ejecutan las instrucciones directamente, sin necesidad de compilar previamente las líneas de código.

4.2. Manipulación de datos

En sí, la manipulación de datos es un elemento que permite varios elementos. La gestión de extremo a extremo comienza con los datos de origen de las áreas o procesos del negocio. El proceso realiza ingesta de datos de las fuentes, esto da como salida datos integrados en una herramienta específica para, luego, a través de la administración de los datos, se ordenen y se sincronicen.

En esta parte se preparan y optimizan los datos para aplicarles técnicas de analítica, visualizando datos y generando modelos de inteligencia artificial para predecir comportamientos y, en casos avanzados, sugerir acciones para que, en la siguiente etapa, ya con ***insigth*** y datos claros, se generen alertas programadas y en algunos

casos, que el sistema tome algunas decisiones automatizadas (como subir o bajar precios a productos, publicar o no productos, entre otras acciones que pueden automatizarse). Al final, se toman acciones ya sean humanas o por los sistemas de información, de manera inteligente.

5. Tableros estadísticos, informes y resultados de visualización

Los tableros estadísticos, también llamados cuadros de mandos o *dashboard*, permiten mostrar información consolidad a alto nivel. Los *dashboard* son las maneras más populares en BI de presentar datos, debido a que permite la compresión de manera sencilla los hechos del negocio y las situaciones presentadas.

Los tableros estadísticos se centran en diferentes aspectos, como:

- **Emplear recursos gráficos**

  Emplear recursos gráficos como elemento principal para representar datos y cifras.
- **Presentar aspectos del negocio**

  Presentar aspectos del negocio de manera general, pero que con la interactividad pueda llegar al detalle hasta que los datos lo permitan.
- **Crear diseños amigables**

  Por lo general todas las herramientas permiten crear diseños muy atractivos visualmente y amigables para la navegación e interacción con los datos.

- **Generar actualizaciones**

  Por lo general estos recursos permiten actualizaciones de manera fácil incluyendo o quitando datos o elementos visuales según las necesidades y sugerencias de los clientes o líderes.

Tableros estadísticos como herramienta para monitorizar el negocio

Un ***dashboard***, permite visualizar de manera ágil y actualizada los procesos de negocio, pues muestra información clave de fácil entendimiento,
los *dashboard* modernos tienen un "*refresh* de datos" casi en tiempo real, lo que les permite a las personas del nivel de decisión o coordinación tener el negocio y sus procesos a la mano con el detalle que requiera, permitiendo tomar con más velocidad las decisiones diarias.

Los tableros de mando contienen los siguientes elementos:

- Permiten empelar variedad de elementos (gráficos, tablas, alertas, mapas, etiquetas, etc.).
- La interacción se aplica de manera integral a todos los elementos del tablero.
- La información que se presenta se basa en indicadores claves del negocio, dando idea clara de cómo van las áreas o aspectos del negocio o proceso.
- Exhibe las tendencias del negocio para tomar decisiones inteligentes, con base en datos y cifras.
- Con la analítica predictiva, los tableros mostrarán hechos futuros con cierta precisión que ayudará también a la toma de decisiones más acertadas.

Usuarios de los tableros de mando

El perfil de los usuarios que usan estos *dashboard* son, por lo general:

- Alta dirección, con el objetivo de comprender lo que sucede en el negocio.
- Gerentes, que deben monitorizar procesos de negocio.
- Usuarios de negocio: estos necesitan realizar análisis exploratorio de datos.

En este sentido, los tableros o cuadros de mando aportan valor al nivel estratégico, táctico y operativo.

Informes y resultados de visualización

Los *dashboard* no son la única manera de ver resultados del negocio o visualizar los datos. Cada área del negocio podría generar sus propios reportes a partir del desarrollo de BI, incluso desde la propia data *warehouse* se puede suministrar a usuarios, datos filtrados según sus interesas particulares del negocio, para que ellos mismos puedan elaborar sus reportes con los datos del despliegue analítico, garantizando que las cifras, independiente del área coincidan entre ellas.

Lo anterior, dando respuesta a que las organizaciones y los usuarios, cada vez más, tienen más habilidades digitales y dominio de herramientas analíticas así no sean usuarios informáticos propiamente dicho.

En relación con los informes y resultados de visualización, tenga presente:

Excel

La herramienta Excel, puede integrarse con bases de datos y otras herramientas de bases de datos que traigan las dimensiones para que los usuarios de Excel puedan

realizar sus propios informes, imprimirlos o presentarlos, o simplemente para tomar decisiones con contexto. Estos OLAP se emplean con el mismo principio de tablas dinámicas o pivotes de datos.

Los informes

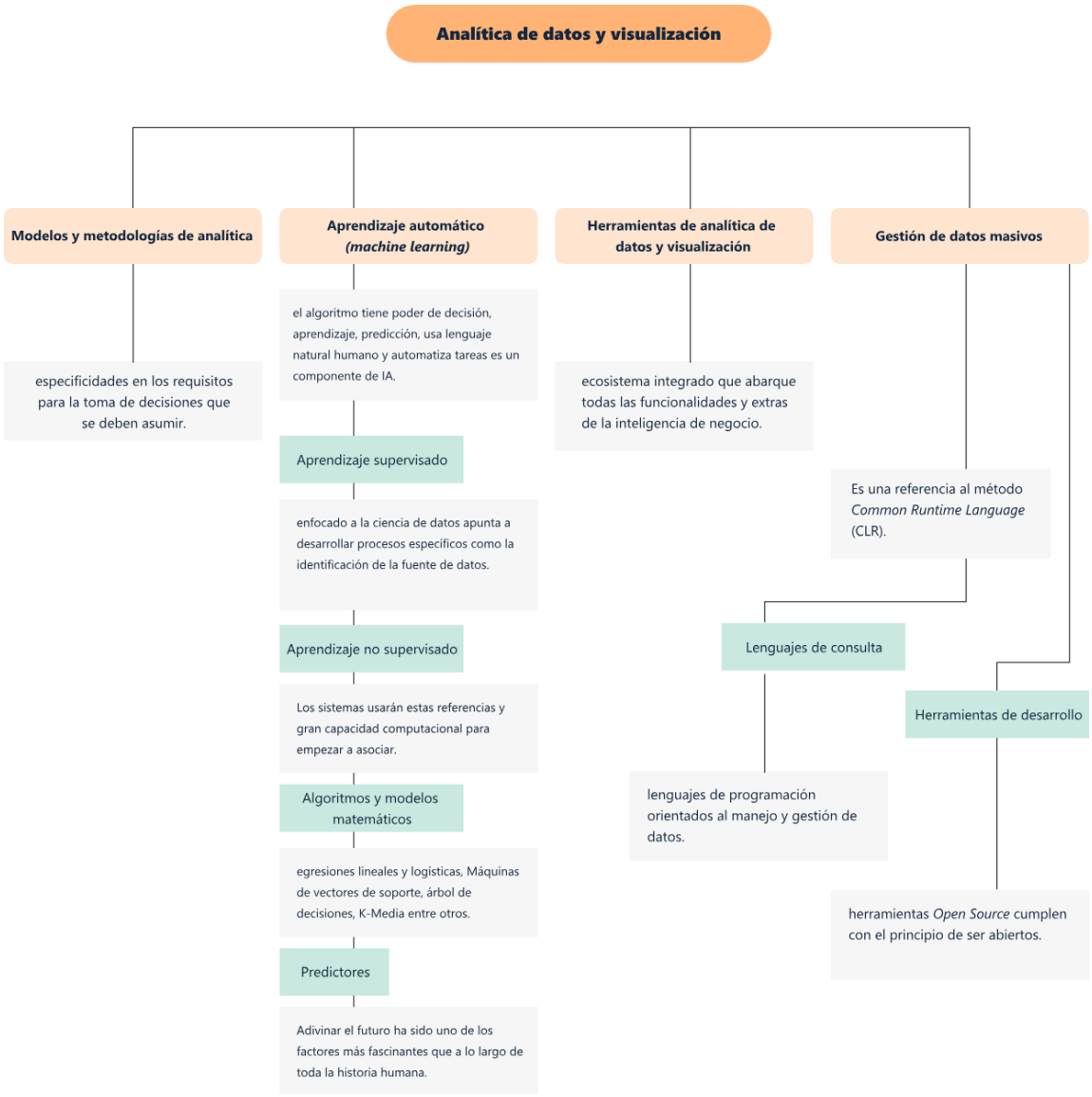
Hablando propiamente de informes, es un paso más allá de lo que son los reportes analíticos. Las tendencias de estos informes son, incorporar mayores capacidades visuales, proporcionando mayor libertad a los responsables de área o procesos para la creación y consumo para incluir en los informes ampliados estos recursos en archivos PDF, o presentaciones Power Point o similares.

Reporting

Las herramientas de informes o *reporting*, permiten responder preguntas como ¿qué pasó?, debido a que es una de las primeras fases de la analítica descriptiva y que, según las herramientas, se pueden generar o exportar cuando incluyen un motor de generación de informes.

Síntesis

Aquí finaliza el estudio de los temas de este componente formativo. En este punto, explore con atención el esquema que se muestra enseguida y realice su propia síntesis de los conceptos estudiados. **¡Adelante!**



Material complementario

Tema	Referencia APA del material	Tipo	Enlace
Algoritmos y modelos matemáticos	Data Silvestre. (2021). Distancia Euclidiana en Python | Métricas y Distancias #1 [video].	Video	https://www.youtube.com/watch?v=vklKtK5oCfg
Gestión de datos masivos	The QA Testing Channel. (2017). Comandos SQL Básicos en Base de Datos.	Video	https://www.youtube.com/watch?v=Yc1-Rg2whNU
Gestión de datos masivos	Learning BI. (2017). Introducción Pentaho Data Integration [video].	Video	https://www.youtube.com/watch?v=o7If1a-gkyI
Gestión de datos masivos	Be Intellligence. Business intelligence	video	https://www.youtube.com/watch?v=Pmdps2kK_5M

Glosario

A

Algoritmo: pasos programados para que las máquinas realicen una función o tarea. Los algoritmos pueden programarse o en caso de la IA se auto ajustan o calibran sin necesidad de intervención humana.

Aprendizaje Profundo: *Deep Learning,* se refiere a los programas que emplean redes neuronales programadas para tener procesos de machine *learning* más avanzados y complejos.

AWS: Amazon Web Service*,* plataforma de computación o servicios en la nube, cuenta con múltiples servicios entre los cuales muchas herramientas tienen grandes componentes de inteligencia artificial y gestión de datos.

C

***Chatbots*:** *chats* operados por robots o *chats* inteligentes que interactúan con personas o clientes sin necesidad de intervención humana.

D

***Datasets*:** conjunto de datos guardados en un sistema, ya sea en una o varias bases de datos. Por lo general son datos estructurados y están disponibles para gestión y uso que se desee dar según los objetivos del negocio.

I

***Insigth*:** en términos de informática y *marketing* se refiere a las verdades (a veces relativa según tiempo y condiciones), que generan los datos o los comportamientos de consumo.

M

Minería de datos: *Data mining,* es el uso de grandes volúmenes de datos para la obtención de situación, circunstancias o verdades a partir de patrones y características de los datos. Existen varias técnicas para la aplicación de minería de datos.

P

Palabras polisémicas: palabras que tienen varios significados, el significado lo da el contexto en el que se da la comunicación.

Referencias bibliográficas

Browner. M. (2020). Máquinas predictivas: la sencilla economía de la inteligencia artificial. https://dokumen.pub/maquinas-predictivas-prediction-machines-spanish-edition-la-sencilla-economia-de-la-inteligencia-artificial-1nbsped-8494949381-9788494949388.html

Curto Díaz, J. (2016). Introducción al business intelligence. https://cursos.yura.website/wp-content/uploads/2020/03/Introduccion_al_Business_Intelligence.pdf

Dot CSV. (2019). ¿Qué es el Aprendizaje Supervisado y No Supervisado? [video]. YouTube. https://youtu.be/oT3arRRB2Cw

Zambelli. R (2024). ¿Qué es el Machine Learning y cómo usarlo en la gestión industrial?. https://blog-es.checklistfacil.com/machine-learning/?utm_term=&utm_campaign=LATAM-ES-PAID-CF-GOOGLE-SEARCH-LM_NOVOS_LEADS-DSA-BLOG&utm_source=google&utm_medium=cpc&hsa_acc=6707140990&hsa_cam=21096577828&hsa_grp=160250856136&hsa_ad=693261475572&hsa_src=g&hsa_tgt=dsa-2284541207217&hsa_kw=&hsa_mt=&hsa_net=adwords&hsa_ver=3&gad_source=1&gclid=CjwKCAjwqMO0BhA8EiwAFTLgIAqb8LevzY8hmeCkm9H9GcC_TwoE49Fnr6dqoYkrvwRMnj4720jy7BoChYAQAvD_BwE

Créditos

ECOSISTEMA DE RECURSOS EDUCATIVOS DIGITALES

Milady Tatiana Villamil Castellanos	Responsable del Ecosistema	Dirección General
Claudia Johanna Gómez Pérez	Responsable de Línea de Producción	Regional Santander - Centro Agroturístico

CONTENIDO INSTRUCCIONAL

Jaime Hernán Tejada	Experto Temático	Centro de la Industria, la Empresa y los Servicios - CIES
Javier Ricardo Luna Pineda	Diseñador Instruccional	Centro de la Industria, la Empresa Y Los Servicios
Fabián Leonardo Correa Díaz	Diseñador Instruccional	Regional Santander – Centro Industrial del Diseño y la Manufactura
Ana Catalina Córdoba Sus	Metodólogo para Formación Virtual	Regional Santander – Centro Industrial del Diseño y la Manufactura
Rafael Neftalí Lizcano Reyes	Responsable Equipo de Desarrollo Curricular	Regional Santander – Centro Industrial del Diseño y la Manufactura
Sandra Paola Morales Páez	Evaluadora Instruccional	Regional Santander - Centro Agroturístico
Lucenith Pinilla Moreno	Actividad Didáctica	Regional Santander - Centro Agroturístico

DISEÑO Y DESARROLLO DE RECURSOS EDUCATIVOS DIGITALES

Edison Eduardo Mantilla Cuadros	Diseñador de Contenidos Digitales	Regional Santander - Centro Agroturístico
Pedro Alonso Bolívar González	Desarrollador *Fullstack*	Regional Santander - Centro Agroturístico
Maria Alejandra Vera Briceño	Animadora y Productora Multimedia	Regional Santander - Centro Agroturístico

VALIDACIÓN RECURSO EDUCATIVO DIGITAL

Laura Paola Gelvez Manosalva	Validadora de Recursos Educativos Digitales	Regional Santander - Centro Agroturístico
Erika Fernanda Mejía Pinzón	Evaluadora Para Contenidos Inclusivos y Accesibles	Regional Santander - Centro Agroturístico