SPRINT 1: Introducción a la Inteligencia Artificial

Inteligencia artificial en la práctica

Videos

Fundamentos de los Datos (¡Consigue una credencial!)

**Introducción a los conceptos de datos**

**La duración es1 hora 15 minutos**

**MODULO 1: TIPOS DE DATOS**

[Nuestro mundo en datos](https://ourworldindata.org/)

<https://ourworldindata.org/>

[Crecimiento de datos en todo el mundo 2010-2028 | Statista](https://www.statista.com/statistics/871513/worldwide-data-created/)

<https://www.statista.com/statistics/871513/worldwide-data-created/>

**Módulo 2**

**Big data**

En este módulo, aprenderás sobre big data y sus características, llamadas las 5 V: Volumen, Variedad, Velocidad, Veracidad y Valor.

**Volumen**,

**Volumen** se refiere a las grandes cantidades de datos que se generan.

* ¡Hay exabytes, zettabytes y yottabytes de datos! ¿Qué es eso?  
  + Un *gigabyte*equivale aproximadamente a mil millones de bytes.
  + Un *exabyte*es igual a mil millones de gigabytes.
  + Un *zettabyte*es aproximadamente igual a mil exabytes.
  + Un *Yottabyte*son mil zettabyes.

**Variedad**,

**Variedad** se refiere a los diferentes tipos de datos a utilizar.

* Has aprendido sobre datos estructurados, como hojas de cálculo y bases de datos.
* Has aprendido sobre datos no estructurados, como informes de empresas, correos electrónicos, publicaciones de Facebook, archivos de audio, blogs, registros de atención médica e investigaciones de mercado.
* ¡Aún hay una gran variedad de datos disponibles!
* Las empresas necesitan utilizar diferentes tipos de herramientas y técnicas para gestionar la variedad de datos.

**Velocidad**,

**Velocidad**se refiere a la increíble rapidez con la que se generan nuevos datos y la rapidez con la que se mueven.

* Los datos se mueven constantemente y se generan extremadamente rápido. Es un proceso que nunca se detiene.
* Por ejemplo, cada minuto se publican 510 000 comentarios, se actualizan 293 000 estados y se suben 240 000 fotos a Facebook por parte de usuarios activos. Piensa en cuántos datos se acumulan a lo largo de horas, días y años.
* Existen tecnologías de transmisión en tiempo real o casi en tiempo real, locales y basadas en la nube que pueden procesar información muy rápidamente.
* Las empresas necesitan saber qué tan rápido se mueven los datos, desde el momento en que se toma una muestra de datos hasta el momento en que se utilizan. Pueden cambiar y cambian.

**Veracidad** y

**Veracidad** se refiere a la calidad y confiabilidad de los datos.

* Debes considerar el origen de los datos para asegurarte de que estén completos, tengan integridad y sean rastreables.
* Con las muchas formas de big data, la calidad y la precisión pueden ser menos controlables. Por ejemplo, un tweet puede contener hashtags, abreviaturas, errores tipográficos y lenguaje coloquial.
* Las empresas necesitan presupuesto y métodos para garantizar que los datos estén limpios y sean confiables. Esta es un área de enfoque para los científicos de datos.

**Valor**

**Valor**se refiere a la capacidad de convertir los datos en valor. La principal razón por la que las personas invierten tiempo en comprender el big data es para obtener valor de él.

* El valor no es solo ganancia. Pueden ser beneficios médicos o sociales, o puede ser la satisfacción del cliente, del empleado o personal.
* Las empresas deben presentar argumentos y tener una comprensión clara del valor que desean obtener de la recopilación y el uso de big data. Deben filtrar los datos "ruidosos" para encontrar lo que buscan.

**¡Sigue explorando!**

Consulta estos recursos:

* [Nuestro mundo digital y Big Data(se abre en una nueva pestaña)](https://www.futureoftech.org/big-data/1-our-digital-world-and-big-data/), un curso gratuito en línea ofrecido por Future of Tech de CompTIA
* [¿Qué tamaño tiene un petabyte, un exabyte o un yottabyte? ¿Cuál es el byte más grande?(se abre en una nueva pestaña)](https://www.zmescience.com/science/how-big-data-can-get/), un blog de ZME Science y escrito por Tibi Puiu para aprender todas las unidades de datos, desde el diminuto byte hasta el gigantesco yottabyte
* [Big Data in Practice(se abre en una nueva pestaña)](https://bernardmarr.com/big-data-in-practice/), un blog de Bernard Marr para aprender sobre las 10 áreas principales en las que se utiliza big data

**Módulo 3**

**Tipos de análisis de datos**

tipos de análisis de datos: análisis descriptivo, análisis de diagnóstico, análisis predictivo y análisis prescriptivo.

* **Análisis descriptivo: ¿Qué está pasando?**

El análisis descriptivo es el tipo de análisis de datos más simple y común.

El análisis descriptivo responde a la pregunta: "¿Qué está pasando?". Proporciona una instantánea de las tendencias y patrones comerciales y utiliza datos históricos y actuales.

El análisis descriptivo manipula datos sin procesar de múltiples fuentes para brindar al analista de datos información valiosa sobre el pasado y una visión de las métricas clave dentro de una empresa.

Estos hallazgos podrían indicar que algo está bien o mal, pero no explican por qué. Sin embargo, los hallazgos pueden ayudar a determinar cuáles son los mayores problemas y dónde empezar a investigar.

A continuación se muestran algunos ejemplos de datos utilizados para análisis descriptivos:

* + El número de suscriptores de un servicio
  + Datos de la encuesta de satisfacción del cliente
  + Informes de ingresos mensuales
  + Informes diarios de acciones
  + Datos demográficos sobre la población de clientes de una empresa, por ejemplo, datos que indican que el 30 % de los clientes son autónomos
* **Análisis diagnóstico: ¿Por qué se produce?**

Después de hacer la pregunta "¿Qué está pasando?", el siguiente paso es profundizar y preguntar "¿por qué?", como por ejemplo:  "¿Por qué se producen tendencias y patrones?". Aquí es donde entra en juego el análisis diagnóstico.

El análisis de diagnóstico toma la información obtenida del análisis descriptivo y la profundiza para encontrar las causas de problemas específicos.

Las empresas utilizan análisis de diagnóstico porque crea más conexiones entre datos e identifica patrones de comportamiento.

A continuación se muestran algunos ejemplos de análisis de diagnóstico:

* + Una empresa de transporte investiga la causa de la lentitud de los envíos en una región determinada.
  + Una empresa de atención médica examina diagnósticos y medicamentos recetados para identificar la influencia de los medicamentos.
  + Una empresa de TI analiza los datos de las solicitudes del servidor para identificar una pequeña cantidad de servidores que causan la mayor parte de las interrupciones del servicio de una organización.
* **Análisis predictivo: ¿Qué es probable que suceda en el futuro?**

El análisis predictivo tiene como objetivo realizar previsiones. Este tipo de análisis utiliza datos históricos para realizar predicciones sobre el futuro. Ya sea la probabilidad de un evento futuro, pronosticar una cantidad cuantificable o estimar un punto en el tiempo en el que algo podría suceder, todo esto se hace a través de modelos predictivos.

En un mundo de gran incertidumbre, poder predecir permite a las empresas tomar mejores decisiones.

Este tipo de análisis es más avanzado y a menudo puede depender del aprendizaje automático y del aprendizaje profundo.

A continuación se muestran algunos ejemplos de análisis predictivo:

* + Una empresa de software utiliza la segmentación de clientes para determinar oportunidades de ventas.
  + Un fabricante de automóviles pronostica la tasa de fallas de una pieza específica del vehículo para determinar las acciones de servicio recomendadas.
  + Un pronosticador meteorológico analiza las condiciones climáticas actuales en una parte del mundo para determinar las condiciones climáticas futuras en otras partes del mundo.
* **Análisis prescriptivo: ¿Qué debería suceder?**

El análisis prescriptivo combina la información de todos los análisis de datos anteriores para determinar el curso de acción a seguir para abordar un problema o tomar una decisión.

El propósito del análisis prescriptivo es prescribir qué acciones tomar para eliminar un problema futuro o aprovechar al máximo una tendencia prometedora.

El análisis prescriptivo se utiliza generalmente para un conjunto de acciones, en lugar de para una acción individual. Esto requiere un gran compromiso por parte de las empresas para poner en marcha la estrategia, el esfuerzo y los recursos. A medida que la tecnología continúa mejorando y más profesionales se capacitan en datos, más empresas ingresarán a este ámbito impulsado por los datos.

El análisis prescriptivo utiliza herramientas y tecnologías avanzadas, como aprendizaje automático, reglas comerciales y algoritmos. Esto hace que el análisis prescriptivo sea sofisticado de implementar y gestionar.

A continuación se muestran algunos ejemplos de análisis prescriptivo:

* + Una aplicación de tráfico que ayuda a las personas a elegir la mejor ruta para volver a casa y tiene en cuenta la distancia de cada ruta, la velocidad a la que se puede viajar en cada carretera y, fundamentalmente, las limitaciones de tráfico actuales
  + Un calendario de exámenes que verifica si los estudiantes tienen horarios conflictivos
  + Sistemas de inteligencia artificial (IA) de empresas basadas en datos como Facebook, TikTok y Netflix

**Fuentes**:

* [4 Types of Data Analytics to Improve Decision-Making(se abre en una nueva pestaña)](https://www.scnsoft.com/blog/4-types-of-data-analytics), ScienceSoft, de Alex Bekker, mayo de 2019
* [The 4 Types of Data Analytics(se abre en una nueva pestaña)](https://www.kdnuggets.com/2017/07/4-types-data-analytics.html), KDnuggets, de Thomas Maydon, julio de 2017
* [Types of Data Analysis(se abre en una nueva pestaña)](https://chartio.com/learn/data-analytics/types-of-data-analysis/), Chartio Data Tutorials, de Patrick Gibson, 2021

**¡Combina el proyecto!**

**Módulo 4**

**Pasos para analizar datos**

cuatro pasos a seguir en el proceso de análisis de datos: recopilar, limpiar, analizar y visualizar.

Antes de comenzar un proyecto de análisis de datos, un buen punto de partida es tener un objetivo de negocio en mente y un problema claramente definido. ¿Qué preguntas quieres responder? ¡Después de todo, el 90 % del análisis consiste en hacer las preguntas correctas! También debes determinar las métricas para medir el rendimiento. ¿Qué estás midiendo y cómo lo estás midiendo?

**Recopilar**

*"Este es el punto de partida del proceso. Este paso tiene que ver con recopilar los datos correctos y suficientes para las preguntas o problemas del proyecto que queremos investigar.*

*Primero determino los datos que puedo recopilar de las fuentes y bases de datos existentes que ya tenemos y que se relacionan con el problema que mi empresa quiere resolver. ¡Siempre recopilo estos datos primero!*

*Luego, averiguo si mi proyecto necesita nuevas fuentes de datos porque esto podría significar más tiempo para el proyecto y potencialmente una mayor inversión de mi grupo empresarial.*

*Mi equipo y yo usamos las herramientas de recopilación de datos de nuestra empresa y seguimos las pautas de recopilación de datos. También tenemos cuidado de almacenar de forma segura los datos en nuestros servidores en la nube.*

*Un punto crucial que me gustaría destacar es que debes recopilar suficientes datos para no distorsionar los resultados de tu análisis".*

**Limpiar**

*"A continuación, no todos los datos que recopilo serán útiles, así que ¡es hora de limpiarlos!*

*La limpieza de datos es el proceso de detectar y corregir registros faltantes o inexactos de un conjunto de datos.*

*Una gran parte de este paso es asegurarse de que los datos estén en un formato utilizable. Esto implica buscar lo que llamamos 'valores atípicos', tratar con valores nulos y buscar datos que puedan haberse ingresado incorrectamente. En pocas palabras, los datos sin procesar tendrán valores faltantes e inexactos que debo abordar.*

*Quizás hayas escuchado el término, 'organización de datos'. 'Ordeno' los datos para que estén en un formato utilizable para mi proyecto en nuestro sistema de base de datos. Por ejemplo, buscaré registros duplicados y los eliminaré.*

*No hay dos conjuntos de datos iguales, por lo que la forma en que limpio los datos puede variar. Limpio los datos en función del contexto. En un caso, ver una entrada en blanco podría equivaler a una entrada cero, por lo que son datos buenos y valiosos. Pero, en otro caso, ver una entrada en blanco podría significar que hay datos incompletos que necesito excluir. ¡Este es el arte de la ciencia de datos!*

*¡Guarda siempre tus datos ya que este es un proceso iterativo!*

*Ah, y un dato interesante. ¡Aquí es donde paso la mayor parte de mi tiempo, limpiando los datos! Calculo que los analistas de datos suelen dedicar entre el 70 y el 80 % de su tiempo a limpiar datos. Es mucho trabajo duro. Pero es imprescindible para poder pasar al análisis”.*

**Analizar**

*"Una vez que tengo los datos relevantes y están limpios, es hora de analizarlos. Este es el paso en el que los analistas de datos dedican entre el 20 y el 30 % de su tiempo. ¡Es la parte divertida y gratificante!*

*Puedo sentir curiosidad e investigar. Y mis habilidades para resolver problemas entran en juego. Aquí utilizo diferentes métodos estadísticos y analíticos y herramientas de software. Es importante que alinee mis métodos de análisis para que coincidan con la intención del problema.*

*Básicamente, identifico problemas y uso análisis para determinar las causas fundamentales de los problemas. Analizo tendencias, correlaciones, variaciones y valores atípicos para ayudarme a concentrarme en responder las preguntas (y cualquier pregunta u objeción que otros puedan tener).*

*A medida que manipulo datos, puedo descubrir que tengo exactamente los datos que necesito, pero lo más probable es que tenga que revisar mis preguntas originales o recopilar más datos. Esto puede generar análisis adicionales y es una de las razones por las que debes guardar siempre tus datos".*

**Visualizar**

***"Después de analizar los datos, interpreto los resultados. Pienso: '¿Qué aprendí de los resultados de mi análisis? ¿Los datos responden a mis preguntas originales? ¿Cómo?'.***

***Para interpretar los resultados, creo visualizaciones de datos. Me ayudan a comparar conjuntos de datos y observar relaciones. Tenemos un par de herramientas que me encanta usar para esto.***

***Y, una vez que puedo sacar mis conclusiones, las visualizaciones me ayudan a mostrar gráficamente mis resultados de una manera que mi equipo y mi liderazgo pueden entenderlos fácil y concisamente. A menudo encuentro que elijo gráficos de barras, gráficos de líneas, gráficos de dispersión y mapas. Me gusta descubrir la mejor manera de representar los datos.***

***Normalmente tenemos una presentación y una reunión en equipo para revisar los resultados y las visualizaciones de datos. Nuestro objetivo es asegurarnos de que los conocimientos que he obtenido se puedan aplicar al problema empresarial y podamos tomar medidas, hacer cambios o reorientar los esfuerzos según sea necesario.***

***Este paso es un buen final para el proceso. Se trata de obtener información de los datos y poder aplicarla al negocio”***

**¡Es bueno saberlo! Extraer, transformar y cargar (ETL)**

Es posible que escuches el término "ETL" utilizado en entornos de trabajo basados en computadoras, en relación con datos, almacenamiento de datos y análisis. ETL es un acrónimo de **extraer, transformar y cargar (ETL**).

ETL es un proceso de integración de datos que combina datos de múltiples fuentes de datos en un único almacén de datos consistente que se carga en un almacén de datos u otro sistema de destino.

A medida que las bases de datos crecieron en popularidad en la década de 1970, se introdujo ETL como un proceso para integrar y cargar datos para el cálculo y el análisis, y eventualmente se convirtió en el método principal para procesar datos para proyectos de almacenamiento de datos.

ETL proporciona la base para el análisis de datos y los flujos de trabajo de aprendizaje automático. Las organizaciones a menudo utilizan ETL para:

* extraer datos de sistemas heredados
* limpiar los datos para mejorar la calidad de los datos y establecer la coherencia
* cargar datos en una base de datos de destino

**Fuente**: [What is ETL (Extract, Transform, Load)?(se abre en una nueva pestaña)](https://www.ibm.com/cloud/learn/etl" \t "_blank), IBM Cloud Learn Hub, abril de 2020

**Módulo 5**

**Visualización de datos**

Los analistas de datos utilizan visualizaciones como gráficos y mapas por dos razones:

1. Para explorar e interpretar datos durante el análisis para identificar patrones o tendencias
2. Para comunicar resultados y ayudar a las personas a comprender los conocimientos para tomar decisiones

Este es el [gráfico de Charles Minard de 1869(se abre en una nueva pestaña)](https://commons.wikimedia.org/wiki/File:Minard.png) que muestra la campaña rusa de 1812 del ejército de Napoleón Bonaparte y sus movimientos.

**La narración de datos** es el proceso de convertir los análisis de datos en una historia simple y comprensible para influir en una decisión comercial. Con el auge de los negocios digitales y la toma de decisiones basada en datos, la narración de datos es una habilidad importante. La idea es “conectar los puntos” entre los resultados y los encargados de tomar decisiones, quienes deben ser capaces de interpretar los datos.

La narración de datos implica una combinación de **datos**, **visualizaciones** y **narrativa**.

* Cuando la narrativa se combina con los datos, explica al público qué está sucediendo con los datos y por qué es importante obtener información.
* Cuando se aplican visualizaciones a los datos, estas brindan al público información que tal vez no se podría obtener sin gráficos o cuadros. Los patrones y tendencias surgen de todas las filas y columnas de una base de datos, con la ayuda de visualizaciones de datos.
* Cuando la narrativa y las visualizaciones se unen, pueden crear una historia de datos que puede influir, impulsar el cambio e involucrar al público.

La **proporción relativa** es el número o la cantidad de un subconjunto presente en la población de todos los puntos de datos. Por ejemplo, si hay 25 estudiantes en una clase, de los cuales 15 son niñas y 10 son niños, entonces la proporción de niñas es 15 de 25 (3 de 5) y la proporción de niños es 10 de 25 (2 de 5). GRAFICO CIRCULAR Y COLUMNA

La **categoría** es la relación entre un conjunto de elementos, en la que un elemento tiene una clasificación más alta, más baja o la misma, en comparación con un segundo elemento. Por ejemplo, los jugadores de videojuegos pueden clasificarse según su puntuación más alta en un torneo.GRAFICO BARRA

**Hora**

La **hora**es una serie de puntos de datos que se enumeran o secuencian en orden de tiempo, como, por ejemplo, la hora diaria de marea alta y marea baja en una playa. GRAFICO COLUMNA LINEAS

La **frecuencia** es el número de veces que ocurre un determinado evento. Por ejemplo, si hoy nieva dos veces, entonces la frecuencia de nieve en ese día en particular es 2. GRAFICO COLUMNA LINEAS

Una **correlación** es la relación entre dos variables aleatorias, que normalmente están relacionadas de forma lineal. Por ejemplo, existe una correlación entre la altura de los padres y la de sus hijos. GRAFICO BARRAS DISPERSIÓN

Un **gráfico circular**es un gráfico estadístico dividido en secciones para ilustrar proporciones numéricas. La suma de todas las "secciones" debe ser igual al 100 %. Los gráficos circulares son útiles para mostrar la proporción relativa de un número pequeño de elementos. Pueden mostrar muy fácilmente qué categoría es la más grande o la que tiene mayor impacto.

**gráficos de barras**son útiles para clasificar una gran cantidad de categorías, mostrar correlación y utilizar para los análisis de antes y después. Utiliza los gráficos de barras para comparar y clasificar. Los gráficos de barras ayudan a ilustrar el cambio a lo largo del tiempo.

**gráficos de columnas pueden**mostrar una proporción relativa entre elementos y datos con el tiempo, así como la frecuencia. Los gráficos de columnas muestran la información verticalmente. También son útiles para mostrar datos negativos.

Los **gráficos de líneas**pueden mostrar datos a lo largo del tiempo y la frecuencia; sin embargo, normalmente muestran estos datos en un continuo. Los gráficos de líneas pueden hacer un seguimiento de los cambios durante períodos de tiempo cortos y largos. Por ello, los gráficos de líneas son útiles para indicar pequeños cambios. Son eficaces para mostrar tendencias.

**gráficos de dispersión** muestran puntos graficados para mostrar la relación entre dos conjuntos de datos. Muestran una gran cantidad de datos y resaltan los valores atípicos. Los gráficos de dispersión también te pueden ayudar a identificar patrones.

cumple con el objetivo de ser **eficaz**, **atractivo** e **impactante**.

**¡Sigue explorando!**

Consulta estos recursos:

* [From Data to Viz(se abre en una nueva pestaña)](https://www.data-to-viz.com/) para una herramienta en línea que te permita aprender más sobre gráficos y ayudar a determinar el gráfico apropiado para los datos
* [El blog How to Choose the Right Data Visualization(se abre en una nueva pestaña)](https://chartio.com/learn/charts/how-to-choose-data-visualization/) ofrecido por Chartio - Data Tutorial y creado por Mike Yi y Mel Restori
* [Viz of the Day(se abre en una nueva pestaña)](https://public.tableau.com/app/discover/viz-of-the-day) en Tableau Public, donde personas de todo el mundo comparten sus visualizaciones

**Puntos clave por recordar**

1. 1

Es importante que las empresas de todas las industrias hallen conocimiento para mejorar potencialmente las operaciones, comprender mejor a los clientes o usuarios finales, aumentar las ganancias y mucho más.

1. 2

El análisis de datos es el proceso de recopilar, limpiar y transformar datos para obtener información que ayude a tomar decisiones mejores e informadas.

1. 3

Los datos estructurados se pueden organizar en bases de datos con muchas columnas y filas. Los datos no estructurados carecen de organización y pueden ser datos como tweets, comentarios de clientes, registros médicos, etc.

1. 4

Los datos cuantitativos se pueden contar y son numéricos. Los datos cualitativos no se pueden contar y no son numéricos. Cuando pienses en datos cualitativos, piensa en "calidad": algo que pueda categorizar sobre los datos.

1. 5

Big data se caracteriza por las 5 V: Volumen, Variedad, Velocidad, Veracidad y Valor.

1. 6

Hay cuatro tipos de análisis de datos que responden a preguntas clave, se complementan entre sí y aumentan en complejidad:

* + Descriptivo
  + Diagnóstico
  + Predictivo
  + Prescriptivo

1. 7

Si bien no existe un proceso de análisis de datos único y definido, estos son los pasos típicos a seguir:

* + Recopilar
  + Limpiar
  + Analizar
  + Visualizar

1. 8

Una visualización de datos, como un gráfico, un mapa o un cuadro, te ayuda a explorar e interpretar los datos durante el análisis para identificar patrones o tendencias, comunicar resultados y ayudar a las personas a comprender los conocimientos para tomar decisiones.

1. 9

Un tipo de visualización puede mostrar los datos de una manera mejor que otro tipo de visualización. El objetivo es tener una visualización que sea eficaz, atractiva e impactante.

**La ciencia de datos en nuestro mundo**

**La duración es2 horas**

**Módulo 1**

**¿Qué es la ciencia de datos?**

**¿Qué es la ciencia de datos?**  
La ciencia de datos es la comprensión del mundo a través del análisis científico de los datos digitales.

**La ciencia de datos** combina el método científico, las matemáticas y la estadística, la programación especializada, el análisis avanzado, la inteligencia artificial (IA) e, incluso, la narración para descubrir y explicar la información empresarial oculta en los datos.

La ciencia de datos es un **enfoque multidisciplinario** para extraer **información procesable** a partir del volumen masivo y creciente de datos que recopilan las empresas actuales.

El objetivo de la ciencia de datos es extraer valor de los datos en todas sus formas.

**Los 5 Por qué**

Al analizar datos, encontrarás un problema y necesitarás comprender el por qué. Los [5 Por qué(se abre en una nueva pestaña)](https://en.wikipedia.org/wiki/Five_whys) es una técnica valiosa en la resolución de problemas que es fácil de recordar. Puedes determinar la causa raíz de un problema **haciendo la pregunta "¿Por qué?" cinco veces**.

Primero pregunta "¿Por qué?", obtén una respuesta y vuelve a preguntar "¿Por qué?". ¡Y así sucesivamente! Cada respuesta guiará la siguiente pregunta "¿Por qué?". La respuesta al quinto "¿Por qué?" deberá revelar la causa raíz del problema.

Algunos problemas tienen más que una sola causa raíz. Si deseas descubrir varias causas raíz, debes repetir este método planteando una secuencia diferente de preguntas cada vez.

**Nota**: El número 5 es la norma general y el número de preguntas "¿Por qué?" que hay que plantear puede variar.

La ciencia, la tecnología y los datos son áreas que están vinculadas y han dado origen a la ciencia de datos. Ten en cuenta este concepto clave mientras aprendes sobre el campo de la ciencia de datos.

* La formación científica permite al científico de datos formular una hipótesis y seguir el método científico.
* Trabajar con tecnología y herramientas permite al científico de datos empezar a clasificar los datos y formular predicciones.
* El científico de datos permite obtener información de los datos estructurados en bases de datos relacionales y datos sin estructurar, como material de video, blogs y tuits.

La ciencia, la tecnología y los datos han estado vinculados desde los primeros días de la ciencia en los siglos XIV y XV. Cada área continúa evolucionando.

**Ciencia:** proporciona conocimientos matemáticos, como el álgebra lineal, la estadística, la teoría de la probabilidad y más, para ayudar a resolver problemas científicos.

**Tecnología:** proporciona inteligencia empresarial, minería de datos, big data, análisis predictivo, aprendizaje automático, etc.

**Datos:** proporciona el estado actual del mundo en el que hay datos estructurados (como conjuntos de datos y bases de datos) y datos no estructurados (como videos, blogs, imágenes y música).

En este momento, la ciencia de datos (**la ciencia más la tecnología más el mundo de datos**) es uno de los campos más importantes.

**El análisis de datos frente a la ciencia de datos**

El análisis de datos y la ciencia de datos son dos términos que se suelen usar en el mismo contexto. Sin embargo, es importante saber que son definiciones diferentes.

Ambos campos trabajan con datos y comparten un mismo objetivo: traducir el análisis de datos en inteligencia empresarial.

La principal diferencia entre el análisis de datos y la ciencia de datos es lo que los analistas de datos y los científicos de datos **hacen**con los datos, es decir las **tácticas** usadas. Así es cómo se diferencian:

* **Los analistas de datos** recopilan y examinan grandes conjuntos de datos para identificar tendencias, previsiones y visualizaciones de datos para narrar una historia convincente mediante información procesable. Esta información ayuda a las empresas a tomar decisiones informadas sobre sus necesidades empresariales.
* **Los científicos de datos**diseñan y crean nuevos procesos para el modelado de datos. Emplean algoritmos, análisis predictivo y estadístico. Los científicos de datos tienen habilidades técnicas para organizar los datos no estructurados y crear sus propias metodologías para realizar predicciones basadas en tendencias de datos.

**¡Sigue explorando!**

Consulta los siguientes recursos:

* [What is data science? (se abre en una nueva pestaña)](https://www.ibm.com/think/topics/data-science)– un artículo de IBM que ofrece un tutorial rápido sobre la ciencia de datos y las herramientas
* [Data and Analytics(se abre en una nueva pestaña)](https://www.comptia.org/resources/data-and-analytics) - ofrece recursos y artículos en línea de CompTIA sobre el campo, las habilidades necesarias y las industrias que usan datos

**Módulo 2**

**Las metodologías de ciencia de datos**

CRISP-DM significa**Cross-Industry Standard Process for Data Mining o Proceso estándar intersectorial de minería de datos).**

Fundado por la iniciativa del [Programa estratégico europeo de investigación y desarrollo en tecnologías de la información(se abre en una nueva pestaña)](https://en.wikipedia.org/wiki/European_Strategic_Programme_on_Research_in_Information_Technology_(ESPRIT)), CRISP-DM es una forma comprobada de orientar la minería de datos. Cualquier industria puede usar este metodología para ayudar a estructurar un proyecto de ciencia de datos. CRISP-DM es un enfoque de ciencia de datos flexible e integral.

CRISP-DM consta de seis fases con flechas que indican las dependencias más importantes y frecuentes entre las fases:

1. Comprensión del negocio
2. Comprensión de los datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Implementación

La secuencia de las fases no es estricta. CRISP-DM es **iterativo**, lo que significa que las fases se pueden repetir para mejorar gradualmente el resultado. Los resultados de algunas fases pueden requerir que el ciclo del proyecto vuelta a las fases anteriores.

Lo que hace única a la metodología CRISP-DM es que empieza con la **comprensión del negocio**. Esta fase se centra en comprender los objetivos y requisitos del proyecto desde la perspectiva empresarial y definir el problema de datos a resolver.

CRISP-DM es uno de los enfoques más populares para los proyectos de ciencia de datos.

KDD significa **Extracción de conocimientos en bases de datos.**

KDD representa el proceso general de recopilación de datos y su refinamiento metódico. KDD suele constar de cinco pasos:

1. Selección
2. Preprocesamiento previo
3. Transformación
4. Minería de datos
5. Interpretación y evaluación

La metodología KDD puede ayudar a las empresas a mantenerse informadas de las necesidades y comportamientos de los clientes y predecir las futuras tendencias de compra para seguir siendo competitivas. Sin embargo, el proceso no aborda la realidad moderna de los proyectos de ciencia de datos, como la implementación de una arquitectura de big data, las consideraciones éticas o los diferentes roles en un equipo de ciencia de datos.

KDD es **iterativo**, lo que significa se se pueden integrar y transformar nuevos datos para obtener resultados diferentes y más apropiados. Los conocimientos adquiridos se pueden reciclar en el proceso, lo que mejora su eficacia.

**Fuente:** [KDD and Data Mining(se abre en una nueva pestaña)](https://www.datascience-pm.com/kdd-and-data-mining/), Data Science Process Alliance, de Nick Hotz, julio de 2021

SEMMA representa sus cinco pasos:

1. **S**ample o Muestrear
2. **E**xplore o Explorar
3. **M**odify o Modificar
4. **M**odel o Modelar
5. **A**ssess o Evaluar

SEMMA es una metodología de ciencia de datos que ayuda a convertir los datos en conocimientos. SEMMA puede ayudar a resolver un abanico de problemas empresariales, como la identificación de fraudes, la retención y la rotación de clientes, el marketing de bases de datos, la fidelización de clientes, la segmentación del mercado y el análisis de riesgos.

El [Instituto SAS(se abre en una nueva pestaña)](https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm) desarrolló SEMMA como un proceso de minería de datos. SEMMA está enfocado principalmente en las tareas de modelado de proyectos de minería de datos.

SEMMA también es un proceso **iterativo,**en la que la respuesta a un conjunto de preguntas suele llevar a preguntas mucho más interesantes y específicas.

**Fuente:** [What is SEMMA?(se abre en una nueva pestaña)](https://www.datascience-pm.com/semma/" \t "_blank), Data Science Process Alliance, de Nick Hotz, mayo de 2021

**Módulo 3**

**Seguimiento de una metodología de ciencia de datos**

Estos son los pasos de la metodología de ciencia de datos que aprenderás:

1. Comprensión del negocio
2. Exploración y preparación de datos
3. Representación y transformación de datos
4. Visualización y presentación de datos
5. Modelos de datos de trenes
6. Implementar modelos de datos

**Paso 1: Comprensión del negocio**

Todos los proyectos, independientemente de lo grande que sean, empiezan con la **comprensión del negocio**. Antes de realizar cualquier exploración de datos, el equipo deberá comprender el problema que se debe resolver.

El **patrocinador empresarial** desempeña un papel crítico. Los patrocinadores empresariales ocupan un puesto de liderazgo. Inician el proyecto porque tiene un "punto débil", por lo que plantean el problema empresarial al equipo de proyectos de ciencia de datos  y luego apoya el proyecto.

El equipo de proyectos de ciencia de datos examina el problema empresarial. Para ello utilizan el **pensamiento de diseño**. El pensamiento de diseño es una metodología de resolución de problemas que se enfoca en el usuario, sintiendo empatía por el usuario y determinando la mejor experiencia de usuario.

Como parte de este método, el equipo de proyectos de ciencia de datos puede organizar un taller de pensamiento de diseño y aplicar técnicas para:

* definir el problema
* determinar los objetivos del proyecto
* desarrollar personajes o personajes ficticios que representen a los usuarios finales típicos
* documentar los requisitos de la solución desde una perspectiva empresarial

La definición del problema ayuda a sentar las bases para la posible resolución del problema empresarial. El patrocinador empresarial también debe participar durante todo el proyecto para brindar conocimientos, revisar los resultados de los datos y garantizar que el proyecto mantenga el rumbo.

Una vez que se haya expuesto el problema empresarial claramente, el **científico de datos** del equipo definirá el enfoque analítico para resolver el problema. Esto implica expresar el problema en el marco de las técnicas estadísticas y de aprendizaje automático. Por ejemplo, si el objetivo es predecir la respuesta de un cliente a los términos "sí" o "no", el enfoque analítico se podría definir como crear, probar e implementar algo denominado el modelo de regresión logística. ¡Los científicos de datos son expertos que cuentan con muchas tecnologías y métodos en su caja de herramientas!

**Paso 2: Exploración y preparación de datos**

Los científicos de datos identifican y recopilan datos de fuentes existentes y, a menudo nuevas, de una empresa. Pueden ser datos estructurados y no estructurados que sean relevantes para el problema. Pueden recuperar datos de fuentes como:

* Archivos estáticos, como hojas de cálculo
* Bases de datos
* Internet

Si el científico de datos encuentra problemas o brechas en la recopilación de datos, éste puede necesitar revisar los requisitos de los datos y recopilar más datos.

Una vez que los datos estén en un formato, el científico de datos necesita trabajar en ellos, puede empezar la **exploración de datos**. La exploración **inicial**de un conjunto de datos es importante porque ayuda a los científicos de datos a encontrar patrones y relaciones y descubrir información inicial de los datos. Estas son algunas preguntas que el científico de datos puede pensar durante la exploración inicial de los datos:

* ¿Qué características de los datos parecen prometedoras para un análisis más profundo?
* ¿La exploración ha revelado nuevas características sobre los datos?
* ¿La exploración ha cambiado la hipótesis inicial?

A continuación, el científico de datos **prepara los datos**. La preparación de datos es muy importante y el paso que consume más tiempo en un proyecto de ciencia de datos. Implica la construcción del conjunto de datos que se utilizará en el paso de modelado. Además, la preparación de datos incluye limpiar los datos, combinarlos de varias fuentes y asegurarse de que no haya brechas en los datos. La preparación de datos también incluye limpiar o "pulir" los datos para que estén listos para su transformación.

Los científicos de datos no pueden asumir que los datos estén listos para utilizarse, incluso si son datos estructurados. Por lo general, los datos reales necesitan un poco de trabajo porque pueden:

* estar incompletos o tener valores incorrectos
* estar dañados con líneas rotas o tener campos en lugares equivocados
* ser demasiado aleatorios
* ser irrelevantes
* ser un valor atípico, es un valor que se encuentra alejado de otros valores y que sesgará los datos
* ser un valor omitido en algunos campos

Para los científicos de datos puede ser un desafío verificar manualmente grandes volúmenes de datos almacenados. Por lo tanto, los científicos de datos utilizan procesos automatizados y herramientas para preparar los datos de forma rápida y precisa.

**Paso 3: Representación y transformación de datos**

El paso de representación y transformación de datos de la metodología de ciencia de datos trata de:

* comprender los datos
* evaluar la calidad de los datos
* descubrir la información inicial sobre los datos

Piensa en los científicos de datos como detectives que investigan un caso. Los científicos de datos utilizan muchas técnicas, herramientas y tipos de análisis para **representar mejor y transformar los datos** para detectar información. Esta lección presenta algunas de estas técnicas, pero recuerda que hay muchas más.

**Estadísticas descriptivas**

Para comprender los datos, el científico de datos puede utilizar un enfoque matemático, como la estadística descriptiva. **La estadística descriptiva** resume cuantitativamente un conjunto de datos. Puede responder a la pregunta: "¿Qué ocurre?". Los científicos de datos pueden crear una tabla para describir un gran conjunto de datos complejos y realizar observaciones sobre:

* Número (N): ¿Cuál es el número total de observaciones?
* Significado: ¿Cuál es el promedio de un conjunto de dos o más números?
* Media: ¿Cuál es el número medio o "centro" en una lista ordenada de números?
* Modo: ¿Cuál es el valor más observado en un conjunto de datos?
* Mínimo: ¿Cuál es el extremo mínimo de un conjunto de datos?
* Máximo: ¿Cuál es el extremo máximo de un conjunto de datos?
* Desviación estándar: ¿Cómo se distribuyen los datos en relación con la media?

A veces, la estadística puede ser engañosa por sí sola, por lo que es importante contar con otras técnicas. Los científicos de datos también utilizan **visualizaciones exploratorias**. Las visualizaciones exploratorias ayudan a que los datos complejos sean más accesibles y reveladores. Los científicos de datos utilizan visualizaciones iniciales, como diagramas, gráficos y mapas para descubrir distribuciones, encontrar patrones y comprender tendencias.

La estadística descriptiva, las técnicas de visualización y muchas otras técnicas ayudan a los científicos de datos a comprender los datos y **evaluar su calidad**. Los equipos de ciencia de datos deben validar la calidad de los datos que utilizan como entrada para el modelado predictivo, ya que unos datos de mala calidad conducirán a un rendimiento deficiente del modelo más adelante en el proceso.

**Para facilitar el análisis, los científicos de datos suelen necesitar normalizar los datos en formatos aptos para modelos para que sean "datos ordenados" en lugar de "datos desordenados".**

**Paso 4: Visualización y presentación de datos**

**La visualización de datos** es la culminación de los esfuerzos del equipo de ciencia de datos para ver la información que han producido sus actividades de transformación de datos.

Las visualizaciones ayudan a los científicos de datos a poner a prueba sus hipótesis y a comprobar sus suposiciones.

Los datos deben **narrar una historia** y abordar el problema empresarial o pregunta que el proyecto intenta responder.

Un**gráfico circular** es un gráfico circular estadístico dividido en secciones para ilustrar proporciones numéricas. La suma de todas las "secciones" debe ser igual al 100 %. Los gráficos circulares son útiles para mostrar la proporción relativa de un pequeño número de elementos y qué categoría es la más grande o tiene el mayor impacto.

**Ejemplo**: una cadena de supermercados utiliza un gráfico circular para mostrar el porcentaje de beneficios totales de diferentes ubicaciones de supermercados. El supermercado con los mayores beneficios aparece resaltado en el gráfico circular.

**Los gráficos de barras** son útiles para clasificar un gran número de categorías. También son útiles para mostrar la correlación o el análisis de antes y después. Utiliza los gráficos de barras para comparar y clasificar. Los gráficos de barras pueden mostrar cambios a lo largo del tiempo, comparando diferentes categorías o partes de un todo.

**Ejemplo**: una farmacia en línea observa un aumento del tráfico web durante la pandemia de COVID-19. La farmacia utiliza un gráfico de barras para capturar el tráfico mensual durante un año. El gráfico de barras ayuda a la farmacia a identificar el aumento en el tráfico.

**Los gráficos de columnas** pueden mostrar una proporción relativa entre elementos y datos con el tiempo, así como la frecuencia. Los gráficos de columnas muestran la información verticalmente. También son útiles para mostrar datos negativos.

**Ejemplo**: una gran empresa realiza un seguimiento del número de accidentes que ocurren en los almacenes. Si el número de accidentes cae por debajo del promedio mensual, la empresa utiliza un gráfico de columnas para diferenciarlo en su informe mensual.

**Los gráficos de líneas** pueden mostrar los datos con el tiempo, además de la frecuencia. Sin embargo, suelen mostrar estos datos a lo largo de un continuo. Los gráficos de líneas pueden hacer un seguimiento de los cambios durante períodos de tiempo cortos y largos. Por ello, los gráficos de líneas son útiles para indicar tendencias y pequeños cambios.

**Ejemplo**: un servicio de asistencia técnica de TI debe medir el rendimiento del servicio para cerrar los problemas de los clientes. Por lo que el servicio de asistencia técnica utiliza un gráfico de líneas para realizar el seguimiento del número de chats y correos electrónicos que responde el equipo al mes.

**Los gráficos de dispersión** muestran puntos graficados para mostrar la relación entre dos conjuntos de datos. Muestran una gran cantidad de datos y resaltan los valores atípicos. Los gráficos de dispersión también ayudan a identificar patrones.

**Ejemplo**: un banco utiliza un gráfico de dispersión para explicar a los ejecutivos que un conjunto de datos de miles de llamadas de clientes revela que los niveles de satisfacción del clientes son más altos cuando el tiempo de respuesta de llamada es más bajo.

**¡Sigue explorando!**

Consulta los siguientes recursos:

* [From Data to Viz(se abre en una nueva pestaña)](https://www.data-to-viz.com/): una herramienta en línea para obtener información sobre los gráficos y ayudar a determinar el gráfico más apropiado para los datos
* [How to Choose the Right Data Visualization(se abre en una nueva pestaña)](https://chartio.com/learn/charts/how-to-choose-data-visualization/): un tutorial de datos de Chartio (Atlassian) de Mike Yi y Mel Restori
* [Viz of the Day(se abre en una nueva pestaña)](https://public.tableau.com/app/discover/viz-of-the-day): un sitio de Tableau Public que cuenta con visualizaciones que gente de todo el mundo ha compartido

**Paso 5: Entrenamiento de modelos de datos**

¿Qué se entiende por un **modelo**?

* **¿Qué es un modelo?** Un modelo de datos identifica los datos, los atributos de datos y las relaciones o asociaciones con otros datos. Un modelo de datos proporciona una vista generalizada de los datos que representa el escenario empresarial y los datos reales.
* **¿Por qué crear un modelo?** El científico de datos puede desarrollar un enfoque mas sistemático para abordar el problema empresarial identificado mediante la creación de un modelo. El objetivo principal de crear un modelo es realizar mejores predicciones para la empresa y obtener una mejor comprensión del sistema que se está modelando.
* **Aprendizaje supervisado**
* En el aprendizaje supervisado, la máquina ingiere muchas preguntas y sus respuestas, básicamente un conjunto de información estructurada previamente. La información puede, por ejemplo, ser dibujos e imágenes de animales, algunos de los cuales son perros y estar etiquetada como "perro". La máquina intenta  identificar patrones de modo que cuando vea una nueva foto de un perro y se le pregunte "¿Qué es?", pueda responder "perro" con gran precisión.
* El aprendizaje supervisado entrena a las máquinas con datos para crear reglas generales que pueden aplicarse a problemas futuros. Cuanto mejor sea el conjunto de datos de entrenamiento, mejor será el resultado.
* **Aprendizaje no supervisado**
* En el aprendizaje no supervisado, la máquina ingiere una enorme cantidad de información, se le formula una pregunta y se le permite determinar cómo responderla por sí sola. Por ejemplo, la máquina puede recibir muchas fotos y artículos sobre perros. La máquina ingiere y clasifica la información dentro de todas las fotos y artículos. Cuando a la máquina se le muestra una nueva foto de un perro, la máquina debe ser capaz de identificarla como un perro con una precisión razonable.
* El aprendizaje no supervisado entrena máquinas con un gran volumen de datos no etiquetados o no estructurados.
* **Aprendizaje reforzado**
* Los humanos y las máquinas pueden aprender mediante el aprendizaje reforzado. El aprendizaje reforzado es una técnica de aprendizaje automático basada en la retroalimentación. A través del aprendizaje reforzado, la máquina determina cómo comportarse en un entorno realizando y observando los resultados de sus acciones. Por cada acción "buena", la máquina recibe una retroalimentación positiva (una recompensa). Por cada acción "mala", la máquina recibe una retroalimentación negativa (una penalización). Como resultado, la máquina aprende automáticamente mediante su experiencia y retroalimentación.
* El aprendizaje reforzado no implica un objetivo específico. Más bien, implica aprender por ensayo y error o "aprender sobre la marcha". El aprendizaje reforzado se utiliza ampliamente en los vehículos autónomos, drones y otras aplicaciones robóticas.

**¡Sigue explorando!**

Consulta los siguientes recursos:

* [¿Qué es el machine learning?:(se abre en una nueva pestaña)](https://www.ibm.com/think/topics/machine-learning) una introducción al aprendizaje automático de IBM que proporciona una descripción de la historia del aprendizaje automático, las definiciones importantes, las aplicaciones y las preocupaciones en las empresas actuales
* [Machine Learning Tutorial: A Step-by-Step Guide for Beginners:(se abre en una nueva pestaña)](https://www.simplilearn.com/tutorials/machine-learning-tutorial) tutoriales en línea de Simplilearn sobre el aprendizaje automático, los tipos de aprendizaje, los pasos y más
* [Machine Learning Key Terms, Explained:(se abre en una nueva pestaña)](https://www.kdnuggets.com/2016/05/machine-learning-key-terms-explained.html) una descripción rápida de KDnuggets de 12 conceptos importantes del aprendizaje automático

|  | **Aprendizaje supervisado** | **Aprendizaje no supervisado** |
| --- | --- | --- |
| **Proceso** | Se proporcionan las variables de entrada y salida | Solo se proporcionan datos de entrada |
| **Datos de entrada** | Los algoritmos se entrenan utilizando datos etiquetados | Los algoritmos se utilizan con datos que no están etiquetados |
| **Complejidad** | Método más sencillo | Computacionalmente complejo |
| **Uso de datos** | Utiliza datos de entrenamiento para identificar un vínculo entre la entrada y las salidas | No utiliza datos de salida |
| **Precisión de los resultados** | Método altamente preciso y confiable | Método menos preciso y confiable |
| **Ejemplos de uso** | Detección de fraudes, clasificación de imágenes, previsión meteorológica, previsión de mercado y estimación de la esperanza de vida | Segmentación de clientes, marketing dirigido, comprensión significativa y sistemas que recomiendan música o películas en streaming |

**Paso 6: Implementación de modelos de datos**

La implementación de un modelo es el paso en el que el modelo de aprendizaje automático se integra en el entorno de producción de la empresa. Los científicos de datos realizan este paso utilizando el conjunto de herramientas y el software elegidos por la empresa. Una vez que se guarda e implemente el modelo, éste se puede utilizar para seguir creando**mejores predicciones para soluciones futuras**. El modelo opera según un plan y el científico de datos lo debe mantener.

**Módulo 4**

**Aplicación de la ciencia de datos en el mundo real**

La ciencia de datos puede:

* identificar y predecir enfermedades y personalizar recomendaciones en el **sector de los cuidados de la salud**
* optimizar las rutas de envío en tiempo real para el **transporte**
* evaluar con precisión el rendimiento de los atletas en los **deportes**
* prevenir la evasión fiscal y predecir las tasas de encarcelamiento para los **gobiernos**
* automatizar la colocación de anuncios digitales en el **comercio electrónico**
* mejorar las experiencias en línea para los **videojuegos**
* crear algoritmos para identificar socios compatibles para **redes sociales**

**Fuente:** [30 Data Science Applications and Examples(se abre en una nueva pestaña)](https://builtin.com/data-science/data-science-applications-examples), Built In, de Mae Rice y actualizaciones de Matthew Urwin, agosto de 2024

Explora ejemplos de cómo las empresas aplican la ciencia de datos para mejorar y resolver problemas globales.

En conclusión, la ciencia de datos se aplica en todos los sectores de muchas maneras:

* La ciencia de datos ha cambiado la forma en la que se descubren y desarrollan medicamentos farmacéuticos para uso humano.
* Los minoristas pueden desarrollar nuevos productos que respondan a las necesidades de los clientes.
* Las empresas pueden detectar y prevenir amenazas de ciberseguridad.
* Las organizaciones pueden trabajar para resolver desafíos mundiales como el cambio climático, la pobreza, la desigualdad y el terrorismo.

Los datos dicen mucho a las empresas y ofrecen un horizonte potencialmente ilimitado para la investigación y la tecnología.

**¡Sigue explorando!**

Consulta los siguientes recursos:

* [Applications of Data Science and Business Analytics(se abre en una nueva pestaña)](https://www.kdnuggets.com/2020/12/greatlearning-applications-data-science-business-analytics.html) - blog de KDnuggets que examina cómo las empresas alcanzan el potencial de la ciencia de datos
* [Data Science Use Cases Guide(se abre en una nueva pestaña)](https://www.datacamp.com/blog/data-science-use-cases-guide) – blog de DataCampo y escrito por Elena Kosourova para obtener información sobre casos de uso de la ciencia de datos y cómo implementar la ciencia de datos en diferentes sectores para impulsar el crecimiento y la toma de decisiones
* [Is Data at the Center of Healthcare for the Future? (se abre en una nueva pestaña)](https://www.news-medical.net/health/Is-Data-at-the-Center-of-Healthcare-for-the-Future.aspx)– blog de News-Medical.Net y escrito por la Dra. Liji Thomas, que describe el auge del Big Data en los cuidados de la salud y las aplicaciones de la ciencia de datos en este sector.

**Módulo 5**

**La ciencia de datos es un deporte de equipo**

En esta lección, obtendrás información y compararás los roles únicos del:

* analista de datos

**¿Qué hace Joseph?**

Joseph está en la oficina todos los días, trabaja con datos estructurados en bases de datos y utiliza la estadística.

* La responsabilidad principal de Joseph es recopilar, organizar, limpiar y analizar grandes volúmenes de datos. Utiliza métodos y herramientas normalizados de su empresa para identificar tendencias, encontrar patrones y hacer predicciones.
* Joseph conoce el sector y utiliza sus conocimientos del negocio.
* Joseph comunica su hallazgos y muestra la información a los patrocinadores empresariales mediante visualizaciones y presentaciones de datos para ayudarles a tomar decisiones y adoptar acciones.

**¿Cuáles son las características y habilidades útiles que hay que tener?**

* Hay que ser metódico
* Pensador crítico
* Tener experiencia en el uso de herramientas de transformación de datos para limpiarlos y herramientas de visualización para mostrar la información
* Ser un gran comunicador con habilidades de presentación
* científico de datos

**¿Qué hace Zhanna?**

Zhanna está en la oficina y participa de principio a fin en los proyectos.

* Dado el problema empresarial, Zhanna desarrolla una hipótesis para investigar y encontrar patrones ocultos.
* Zhanna trabaja con datos de muchas fuentes. Dependiendo del proyecto, puede viajar o desplazarse al campo para recopilar datos y mediciones. Los datos pueden ser estructurados o no estructurados, si ya se han "pulido".
* Zhanna realiza experimentos para crear modelos personalizados mediante las herramientas y metodología de ciencia de datos de la empresa.
* Zhanna utiliza técnicas como el aprendizaje automático para crear y entrenar modelos que predigan resultados futuros.
* En general, Zhanna transforma los datos en conocimientos para generar información procesable que pueda utilizarse para mejorar los resultados futuros.

**¿Cuáles son las características y habilidades útiles que hay que tener?**

* Sentir siempre curiosidad y preguntarte "por qué".
* Tener una mentalidad científica y de investigación
* Ser solucionador de problemas
* Tener conocimientos de matemáticas, estadística y aprendizaje automático
* Tener experiencia en herramientas de análisis de datos
* ingeniero de datos

**¿Qué hace Ronnie?**

Ronnie gestiona la infraestructura de datos de la empresa. Su responsabilidad principal es establecer sistemas y procesos que los analistas y científicos de datos puedan utilizar y en los que puedan confiar al trabajar con datos.

* Ronnie entiende el flujo de datos y transforma grandes volúmenes de datos sin procesar en "canalizaciones" útiles para proyectos.
* Ronnie se enfoca en las herramientas y utiliza técnicas de programación avanzadas.
* Ronnie trabaja en los pasos para probar e implementar los modelos de aprendizaje automático en la producción para la empresa.

**¿Cuáles son las características y habilidades útiles que hay que tener?**

* Ser experto en tecnología
* Tener conocimientos de matemáticas, estadística y aprendizaje automático
* Ser experto en programación
* Estar familiarizado con la arquitectura de la infraestructura (IA

**Notas interesantes**

* Todos los miembros del equipo son responsables de **comprender el problema empresarial** para que puedan trabajar en la propuesta de una solución.
* Las organizaciones utilizan **diferentes puestos**para los roles que trabajan con datos. Dependiendo del tamaño de la organización, una persona puede tener un rol y responsabilidades combinados.
* Hay otros roles que no se cubren aquí. Por ejemplo, puedes encontrar a alguien que sea un **periodista de datos**. El periodista de datos mantiene una mentalidad centrada en el cliente y su responsabilidad principal es la comunicación de los resultados. Si bien no es un rol técnico, el periodista de datos es un comunicador altamente calificado que convierte los hechos en una historia convincente con información y visualizaciones de datos.

**¡Cuánto has progresado!**

¡Y has aprendido muchísimo! Has aprendido sobre el campo de la ciencia de datos, incluidas las metodologías ampliamente adoptadas. Has visto un proyecto de ejemplo mientras avanza por los pasos de una metodología de ciencia de datos, la aplicación de la ciencia de datos en el mundo real y el rol de los científicos de datos y sus colegas.

Ahora que has completado este curso, deberías poder:

* definir la ciencia de datos
* reconocer la importancia de tener curiosidad para resolver problemas con los datos
* diferenciar entre los campos de análisis de datos y la ciencia de datos
* identificar tres metodologías de ciencia de datos adoptadas ampliamente
* explorar un escenario de proyecto de datos e identificar las principales tareas mientras avanza por una metodología
* reconocer las industrias y aplicaciones de la ciencia de datos que ayudan a resolver problemas mundiales y descubrir innovaciones
* comparar los roles y características del analista de datos, el científico de datos y el ingeniero de datos

**Puntos clave por recordar**

1. 1

La ciencia de datos combina el método científico, las matemáticas y la estadística, la programación especializada, el análisis avanzado, la inteligencia artificial (IA) e, incluso, la narración para descubrir y explicar la información empresarial oculta en los datos.

1. 2

Los científicos de datos utilizan metodologías que incluyen procesos o actividades que se deben realizar para obtener resultados. Las metodologías también son científicas, por lo que la clave es que son repetibles.

1. 3

Las tres metodologías de ciencia de datos ampliamente utilizadas son: Cross-Industry Standard Process for Data Mining (CRISP-DM), Extracción de conocimientos en bases de datos (KDD) y Muestrear, Explorar, Modificar, Modelar y Evaluar (SEMMA).

1. 4

Las actividades clave que tienen lugar durante un proyecto de ciencia de datos son:

* 1. Comprensión del negocio
  2. Exploración y preparación de datos
  3. Representación y transformación de datos
  4. Visualización y presentación de datos
  5. Entrenamiento de modelos de datos de trenes
  6. Implementación de modelos de datos

1. 5

Quienes trabajan en la ciencia de datos utilizan tecnologías y herramientas para crear modelos. Los modelos se utilizan para predecir resultados o descubrir patrones subyacentes. La intención es obtener conocimientos que conduzcan a acciones que mejoren los resultados futuros.

1. 6

Los científicos de datos utilizan el aprendizaje automático para entrenar modelos. El aprendizaje automático consiste esencialmente en enseñar a una computadora a resolver problemas. Permite a una máquina aprender a partir de los datos sin programarla con reglas. La máquina puede aprender de los datos que se le proporcionan.

1. 7

Las visualizaciones de datos deben ser efectivas, atractivas e impactantes.

1. 8

¡Una de las características más importantes de un científico de datos es ser siempre curioso!

1. 9

Se están llevando a cabo proyectos de ciencia de datos en el mundo real, en las plataformas de redes sociales y sectores como el de los cuidados de la salud, el transporte, los deportes, el comercio electrónico, etc.

1. 10

¡La ciencia de datos es una labor en equipo! Los analistas de datos, los científicos de datos y los ingenieros de datos colaboran para resolver problemas empresariales.