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.

**Descripción general de herramientas y lenguajes de datos**

En este curso, obtendrás una introducción a las herramientas comunes de análisis y visualización de datos para que las conozcas y puedas reconocer cómo los proyectos de ciencia de datos las utilizan.

**Módulo 1**

**El valor del código abierto**

**El software de código abierto** es un software con código que se publica públicamente y que todo el mundo puede utilizar.

El software de código abierto es colaborativo, lo que significa que se apoya en una comunidad virtual de personas para revisar, modificar y compartir el código fuente entre sí. Los desarrolladores comparten conocimientos, ideas y código para crear soluciones de software más innovadoras.

Estos son cuatro aspectos clave del software de código abierto que conviene tener presente:

1. **Uso**: cualquiera puede utilizar y ejecutar el software para cualquier propósito, bajo licencia.
2. **Vista**: cualquiera puede ver el código fuente para comprender el funcionamiento del software.
3. **Modificación**: las mejoras, las correcciones de errores y soluciones pueden proceder de cualquier persona.
4. **Intercambio**: las contribuciones se basan en un propósito común y compartido.

Ahora, revisaremos alguna terminología y conceptos básicos que se utilizan en el mundo del código abierto.

* Una **comunidad** es cualquiera relacionado con un proyecto de código abierto. Es un grupo de personas que obtienen algún beneficio del proyecto.
* Un **colaborador** es alguien que contribuye a un proyecto.
* Un **committer** es alguien que revisa y aprueba los cambios realizados en el código fuente del proyecto. El committer tiene acceso de escritura (o permiso de escritura) a un repositorio de código fuente. Esto significa que están autorizados a actualizar los datos.
* El**código de conducta** protege a la comunidad y proporciona pautas sobre el comportamiento aceptable. Todos los proyectos deben tener un código de conducta.
* Las **pautas de contribución** describen cómo contribuir y colaborar. Proporcionan reglas sobre cómo la comunidad puede participar en el proyecto de código abierto.

El código abierto es una forma de trabajar en un proyecto que se basa en el acceso y la colaboración.

[Git(se abre en una nueva pestaña)](https://git-scm.com/) es un **sistema de control de versiones**para realizar el seguimiento de los cambios en el código fuente durante el desarrollo del software. Ayuda a coordinar el trabajo entre los programadores. Git es un software de código abierto con licencia. Se instala localmente en la computadora.

[GitHub(se abre en una nueva pestaña)](https://github.com/) es un **servicio en línea** que proporciona un lugar para alojar tanto código fuente como para contribuir y colaborar. Permite que las personas trabajen juntas en proyectos desde cualquier lugar. GitHub es un servicio, no un software. Ofrece cuentas gratuitas, profesionales y empresariales.

**Creación de una cuenta de GitHub**

La creación de una cuenta de GitHub es gratis. Es como crear una nueva cuenta en las redes sociales o suscribirse a un boletín informativo.

1. Ve a la página de afiliación de GitHub: [https://github.com/join(se abre en una nueva pestaña)](https://github.com/join).
2. Completa el formulario en línea.
   * Escribe un **nombre de usuario**. Esto mostrará tu nombre de usuario a todos en GitHub. Si un nombre de usuario no está disponible, sigue intentándolo.
   * Escribe tu **dirección de correo electrónico** y luego crea un **contraseña**.
   * Verifica tu cuenta.
   * Selecciona **Create account (Crear cuenta)**.

¡Hecho! Estás listo para [explorar GitHub(se abre en una nueva pestaña)](https://github.com/explore), revisar [temas(se abre en una nueva pestaña)](https://github.com/topics) o descubrir las nuevas [tendencias(se abre en una nueva pestaña)](https://github.com/trending).

En el siguiente paso, crearás un repositorio.

**Creación de un repositorio**

A continuación, crea tu primer repositorio para que puedas comprender la estructura del alojamiento de archivos en GitHub para proyectos de código abierto.

1. Ve a la página inicial de GitHub: [https://github.com/(se abre en una nueva pestaña)](https://github.com/). Asegúrate de iniciar sesión.
2. Busca "Recent Repositories" (Repositorios recientes) y selecciona el botón **New** (Nuevo).
3. Se muestra la página **Create a new repository** (Crear un nuevo repositorio). Ten presente que el nombre de usuario se muestra en el campo obligatorio **Owner**(Propietario).
4. Escribe un nombre en el campo **Repository name** (Nombre de repositorio).
   * El nombre del repositorio debe ser corto y fácil de recordar. Debe ser significativo para tu proyecto.
   * Puesto que es una actividad, puedes intentar utilizar un nombre de repositorio como "Proyecto 1" o el nombre que sugiera GitHub.
5. Escribe una descripción en el campo **Description** (Descripción).
   * Si bien esto es opcional, es muy importante describir el proyecto al público.
   * Como se trata de una actividad, puedes escribir lo que quieras o quizás algo como: "Aquí hay cosas buenas!"
6. En aras del código abierto, puedes seleccionar **Public** (Público) para que tu repositorio sea visible para todo el mundo en Internet. Todavía puedes elegir quién puede contribuir.
7. Debes marcar **Add a README file** (Agregar un archivo README). Es una práctica recomendada y un componente importante del repositorio.
   * [El archivo README(se abre en una nueva pestaña)](https://docs.github.com/en/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/about-readmes) indica a la gente la utilidad de tu proyecto, qué pueden hacer con él y cómo pueden utilizarlo.
   * Puesto que es una actividad, puedes inventarte un texto de momento y actualizarlo más adelante.
8. Puedes explorar, pero omite los campos **Add .gitignore** (Agregar .gitignore) y **Choose a license** (Elegir una licencia). Puedes dejar las selecciones predeterminadas.
9. Selecciona **Create repository** (Crear repositorio).

¡Felicidades! Acabas de crear un repositorio público en tu cuenta personal de GitHub. Tómate un momento para comprobarlo. Ten en cuenta el archivo README y el ícono para editarlo.

**Adición de un archivo de datos**

En este paso final, obtendrás información sobre cómo cargar un archivo con datos en tu repositorio.

1. Descarga y guarda este archivo **CSV (valores separados por comas)** denominado [“german\_credit\_data\_biased\_training.csv”(se abre en una nueva pestaña)](https://ibm.box.com/shared/static/e09ka2u5hrxww8ymcxmcwvtojdmvygk0.csv) en la computadora o dispositivo.
   * Un archivo CSV es un archivo de texto con un formato específico que permite guardar los datos en un formato estructurado de tabla.
   * **Nota:**Este archivo CSV es un archivo públicamente disponible que contiene datos ficticios de ejemplo para fines de capacitación. No está destinado a ningún otro fin. ¡De hecho, también está disponible en un repositorio de IBM en GitHub!
2. Ve al repositorio de GitHub que has creado.
3. En el elemento desplegable **Add file** (Agregar archivo), selecciona **Upload files** (Cargar archivos).
4. Puedes arrastrar el archivo CSV a tu repositorio o seleccionar **choose your files** (elegir archivos) para encontrarlo en la computadora o dispositivo.
5. En la sección **Commit changes** (Confirmar cambios), escribe un nombre en el campo **Add files via upload** (Agregar archivos mediante carga).
   * Por ejemplo, como se trata de una actividad, puedes escribir "Datos de ejemplo" ya que es el primer conjunto de datos en tu repositorio.
   * También puedes escribir una breve descripción en el campo **Add an optional extended description** (Agregar una descripción ampliada opcional). Es una buena práctica mantener los archivos organizados.
6. Selecciona **Commit changes** (Confirmar cambios).

¡Ya puedes ver el archivo CSV en tu repositorio! Puedes ver tu repositorio seleccionando el enlace. Tu repositorio está fechado, almacenado y contiene más de 5000 líneas, por lo que GitHub indica: “We can’t make this file beautiful and searchable because it’s too large" (No podemos poner bonito el archivo ni hacerlo consultable porque es demasiado grande). ¡Tienes datos!

**¡Sigue explorando!**

Consulta los siguientes recursos:

* [GitHub Docs(se abre en una nueva pestaña)](https://docs.github.com/en) - tutoriales en línea de GitHub para empezar y toda la documentación práctica en línea que puedas necesitar
* [Manual The Open Source Way 2.0(se abre en una nueva pestaña)](https://www.theopensourceway.org/the_open_source_way-guidebook-2.0.html) - un manual en línea de opensource.com y colaboradores sobre las prácticas recomendadas para crear y mantener comunidades de código abierto
* **Módulo 2**
* **Módulo 2**
* **Las herramientas para analizar y visualizar datos**
* Microsoft Excel

**¡Sigue explorando!**

Consulta los siguientes recursos:

* [Aprendizajes en video de Excel (se abre en una nueva pestaña)](https://support.microsoft.com/en-us/office/excel-video-training-9bc05390-e94c-46af-a5b3-d7c22f6990bb)– una serie de recursos introductorios en línea de Microsoft para comenzar a utilizar Excel
* [Excel Easy(se abre en una nueva pestaña)](https://www.excel-easy.com/) – tutoriales en línea de Excel Easy sobre cómo utilizar las funciones de MS Excel, con ejemplos
* Hojas de cálculo de Google

**Echa un vistazo...**

Puedes acceder al programa ahora. ¡Prueba [Hojas de cálculo de Google!(se abre en una nueva pestaña)](https://docs.google.com/spreadsheets/u/0/)

**¡Sigue explorando!**

Consulta el siguiente recurso:

* [Formación y ayuda sobre Hojas de cálculo de Google(se abre en una nueva pestaña)](https://support.google.com/a/users/answer/9282959?visit_id=637940335979772855-4286094511&hl=en&rd=1) - tutoriales en línea del Centro de Aprendizaje de Google Workspace para obtener información sobre los conceptos básicos de Hojas de cálculo
* Lenguaje de consulta estructurado (SQL)

El lenguaje de consulta estructurado (SQL) es un lenguaje estándar para comunicarte con bases de datos. Como el nombre sugiere, utilizas SQL cuando tienes datos estructurados en una base de datos relacional. SQL ha sido una referencia del Instituto Nacional Estadounidense de Estándares (ANSI) desde 1986 y de la Organización Internacional de Normalización (ISO) desde 1987.

SQL es un lenguaje de consulta, no un lenguaje de programación.

* El propósito es formular preguntas o "**consultar**" una base de datos relacional o modificar su contenido.
* Puedes utilizar las consultas SQL para realizar operaciones en una base de datos, como seleccionar, recuperar, actualizar y eliminar datos.
* Los comandos SQL, como "Select", "Insert", "Update", "Delete", "Create" y "Drop" pueden realizar la mayoría de cosas que necesitas hacer con una base de datos.

Con SQL, puedes:

|  |  |
| --- | --- |
| * Ejecutar consultas en una base de datos * Recuperar datos de una base de datos * Insertar registros en una base de datos * Actualizar registros en una base de datos * Eliminar registros de una base de datos | * Crear nuevas bases de datos * Crear nuevas tablas en una base de datos * Crear procedimientos almacenados en una base de datos * Crear vistas en una base de datos * Establecer permisos en tablas, procedimientos y vistas |

**¿Qué es NoSQL?**

NoSQL es una abreviatura de "no solo SQL". Si bien SQL te permite trabajar con datos estructurados en bases de datos relacionales, NoSQL te permite trabajar con datos no estructurados en bases de datos no relacionales. NoSQL es útil para Big Data y las aplicaciones web en tiempo real. Por ejemplo, una empresa como Twitter que recopila terabytes de datos de usuarios cada día puede utilizar NoSQL.

Los científicos de datos también utilizan NoSQL para comunicarse con las bases de datos NoSQL.

* El propósito es almacenar y trabajar con datos no estructurados, como muchas imágenes y texto.
* Los tipos de bases de datos NoSQL incluyen bases de datos de documentos, bases de datos de columnas anchas y bases de datos de gráficos.

En general, los profesionales de la ciencia de datos utilizan SQL para explorar, mantener y proteger datos para poder tomar mejores decisiones.

**¡Sigue explorando!**

Consulte los siguientes recursos:

* [SQL Tutorial(se abre en una nueva pestaña)](https://www.w3schools.com/sql/default.asp) – una serie de tutoriales en línea completos de w3schools para aprender SQL y talleres prácticos con ejercicios
* [Learn SQL(se abre en una nueva pestaña)](https://www.codecademy.com/learn/learn-sql) – un curso gratuito en línea para principiantes de Codecademy para aprender a comunicarte con bases de datos mediante SQL (Nota: Deberás registrarte para acceder a este curso gratuito).
* [SQL Commands Cheat Sheet: How to Learn SQL in 10 Minutes (se abre en una nueva pestaña)](https://www.freecodecamp.org/news/learn-sql-in-10-minutes/)– un blog de freeCodeCamp escrito por un investigador de IA que ofrece un inicio rápido en SQL para principiantes
* [SQL vs. NoSQL: What's the difference?(se abre en una nueva pestaña)](https://www.youtube.com/watch?v=Q5aTUc7c4jg) – un video de IBM Technology que analiza las cinco diferencias principales entre las bases de datos SQL y NoSQL
* Python

[Python(se abre en una nueva pestaña)](https://www.python.org/) es un lenguaje de programación gratuito, de código abierto y uso general que está disponible para que todos lo utilicen. Python fue creado por Guido van Rossum y lanzado en 1991. Fue diseñado con la intención de que fuera fácil y divertido de usar. El nombre, " Python", es un reconocimiento al grupo cómico británico, [Monty Python(se abre en una nueva pestaña)](https://en.wikipedia.org/wiki/Monty_Python).

Python se puede utilizar para crear aplicaciones web y admite proyectos de ciencia de datos. La filosofía de diseño de Python enfatiza un código que es fácil de leer y se destaca por el uso de espacios en blanco.

Python tiene una sintaxis como la del idioma inglés, con cierta influencia de las matemáticas. Por eso, a algunos programadores se les anima a programar sin código preparado. Otros programadores pueden descubrir que pueden escribir programas con menos líneas, en comparación a otros lenguajes de programación.

Además:

* puedes utilizar Python para conectarte a sistemas de bases de datos y leer y modificar archivos.
* Python puede manejar Big Data y realizar operaciones matemáticas complejas.
* puedes combinar Python con una biblioteca de software de análisis y manipulación de datos, como [pandas(se abre en una nueva pestaña)](https://pandas.pydata.org/). Python puede ayudarte a obtener información y crear visualizaciones de datos.

**Tómate un momento para revisar el ejemplo de Python**

Esta es una muestra de código Python. Es un programa Python de ejemplo que solicita al usuario una entrada numérica y luego comprueba si el número ingresado es par o impar.

Lee las líneas 1, 2 y 3. Luego, observa el código en el resto de las líneas.

**Fuente**: [Python Program to Check if a Number is Odd or Even by Programiz(se abre en una nueva pestaña)](https://www.programiz.com/python-programming/examples/odd-even)

Puedes probar tú mismo el programa Python anterior.

* Visita este [compilador Python en línea de Progamiz(se abre en una nueva pestaña)](https://www.programiz.com/python-programming/online-compiler/?ref=20631de9).
* Selecciona **Run** (Ejecutar).
* Junto a **Enter a number (Escribir un número)**: escribe un número de tu elección, como 8 o 303.
* Pulsa **Enter** (Intro) para ver si el código detecta si es un número par o impar.
* Puedes seleccionar **Run** (Ejecutar) para reiniciar y probar más números.

Python tiene una gran comunidad en crecimiento con una variedad de paquetes y bibliotecas de código abierto disponibles. Puedes aprender Python tanto si eres un programador principiante como si tienes experiencia con otros lenguajes de programación. La mayoría de los tutoriales de Python en línea asume que sabes cómo ejecutar programas en la computadora. Algunos sitios ofrecen codificación en el navegador para practicar mientras aprendes Python.

**Echa un vistazo...**

¿Quieres descargar e instalar Python en la computadora? Visita [python.org(se abre en una nueva pestaña)](https://www.python.org/).

**¡Sigue explorando!**

Consulta los siguientes recursos:

* [Beginner’s Guide to Python(se abre en una nueva pestaña)](https://wiki.python.org/moin/BeginnersGuide) - una guía en línea de Python que te ayuda a empezar rápidamente
* [Python Tutorial for Beginners: Learn Python Quickly(se abre en una nueva pestaña)](https://python.land/python-tutorial) – un tutorial en línea gratuito de PythonLand que enseña desde conceptos básicos a avanzados con muchos ejemplos que explican conceptos
* [Python Tutorial(se abre en una nueva pestaña)](https://www.w3schools.com/python/default.asp) – una serie completa de tutoriales en línea de w3schools para aprender Python y obtener práctica con ejercicios
* IBM Watson Studio

IBM tiene una solución llamada [IBM Watson Studio(se abre en una nueva pestaña)](https://www.ibm.com/cloud/watson-studio). IBM Watson Studio es un entorno de desarrollo integrado (IDE). IBM Watson Studio que lleva el nombre del fundador de IBM, reúne las herramientas de desarrollo y análisis más útiles, envolviéndolas en una plataforma de desarrollo que es lo suficientemente potente como para enfrentar desafíos a gran escala, y lo suficientemente simple para que los desarrolladores puedan dominarla rápidamente.

* Es un entorno colaborativo de ciencia de datos y aprendizaje automático.
* IBM Watson Studio trabaja con herramientas de código abierto.
* IBM Watson Studio ofrece una interfaz gráfica con operaciones integradas.
* No es necesario saber codificar para utilizar la herramienta.

Además, IBM Watson Studio tiene una [herramienta integrada de refinación de datos(se abre en una nueva pestaña)](https://www.ibm.com/cloud/data-refinery).

* La refinación de datos te permite preparar y transformar grandes cantidades de datos sin procesar en datos de alta calidad.
* Puedes visualizar los datos utilizando cuadros y gráficos integrados para comprender la distribución de tus datos.
* Puedes programar trabajos para que los datos produzcan resultados repetibles.

**Ejemplo de herramienta de refinación de datos en IBM Watson Studio: Tómate un momento para revisar**

Esta es una captura de pantalla de datos de reclamaciones de seguros de automóviles que se han cargado y se muestran en la interfaz de IBM Watson Studio. Los datos se pueden analizar con la**herramienta de refinación de datos**dentro de IBM Watson Studio. Puedes trabajar en un proyecto mediante las pestañas **Datos**, **Perfil** y **Visualizaciones**de la interfaz**.**

**Echa un vistazo...**

Si quieres profundizar más, mira este divertido **video de demostración**de 4 minutos llamado [Preparing data with Data Refinery(se abre en una nueva pestaña)](https://mediacenter.ibm.com/media/Lab%205%3A%20Preparing%20data%20with%20Data%20Refinery/1_zpxhku5t" \t "_blank) (Preparación de datos con Data Refinery). El video te ayudará a aprender cómo la herramienta de refinación de datos funciona en IBM Watson Studio. En el video, un equipo de ciencia de datos intenta predecir si es probable que un cliente incumpla el pago de su préstamo hipotecario. Inicia el video para obtener información sobre cómo un científico de datos recopila tres conjuntos de datos, los limpia y selecciona una visualización inicial.

**Fuente**: IBM Expert Labs presenta este video del curso de [Conceptos básicos de IBM Cloud Pak for Data(se abre en una nueva pestaña)](https://www.ibm.com/training/course/6XL536G).

**¡Sigue explorando!**

Consulta los siguientes recursos:

* [Prueba gratuita de IBM Watson Studio(se abre en una nueva pestaña)](https://www.ibm.com/cloud/watson-studio) - comienza la prueba gratuita de IBM Watson Studio en Cloud Pak for Data (**Nota**: Necesitarás registrarte para obtener una cuenta gratuita)
* [Data refinery(se abre en una nueva pestaña)](https://www.ibm.com/cloud/data-refinery) – página de un producto IBM para obtener más información sobre la herramienta de refinación de datos de IBM Watson Studio
* [Uso de Data Refinery para dar forma a los datos sin procesar (se abre en una nueva pestaña)](https://video.ibm.com/recorded/128592867)– video de IBM que demuestra cómo dar forma a datos de ejemplo sin procesar mediante la herramienta de refinación de datos
* Tableau

[Tableau(se abre en una nueva pestaña)](https://www.tableau.com/) es un software de visualización de datos e inteligencia empresarial popular para obtener información relevante de los datos. Muchas empresas utilizan Tableau para generar representaciones gráficas e ilustradas de los datos.

Con Tableau, puedes:

* analizar grandes volúmenes de datos.
* crear diferentes paneles, gráficos, tablas, mapas, historias, etc., para ayudar a tomar decisiones empresariales.
* realizar tareas sin experiencia en programación. Ofrece una interfaz intuitiva.
* diseñar visualizaciones interactivas.

**Tómate un momento para revisar el ejemplo de Tableau**

Este es un ejemplo de un panel de visualización de datos de Tableau del tráfico web entrante de un sitio web ficticio. El ejemplo muestra diferentes tipos de visualizaciones que una empresa puede utilizar para supervisar su tráfico web, incluye un gráfico de barras, un gráfico de dispersión y un mapa.

**Fuente**: [Fuentes de tráfico de Ellie Fields(se abre en una nueva pestaña)](https://public.tableau.com/app/profile/efields/viz/Traffic-sources/TrafficDashboard)

La visualización del ejemplo anterior es interactiva en Tableau. Puedes probarlo Visita [Fuentes de tráfico de Ellie Fields(se abre en una nueva pestaña)](https://public.tableau.com/app/profile/efields/viz/Traffic-sources/TrafficDashboard). Puedes desplazarte y seleccionar cualquier parte de la visualización para ver más información e interactuar con el gráfico.[(se abre en una nueva pestaña)](https://public.tableau.com/app/profile/efields/viz/Traffic-sources/TrafficDashboard)

**¡Sigue explorando!**

Consulta los siguientes recursos:

* [Prueba gratuita de Tableau(se abre en una nueva pestaña)](https://www.tableau.com/products/trial) – regístrate para obtener una prueba gratuita de Tableau
* [Videos gratuitos de formación en Tableau(se abre en una nueva pestaña)](https://www.tableau.com/en-gb/learn/training/20222) – serie de videos de Tableau para empezar a preparar datos y crear visualizaciones en la herramienta (Nota: necesitarás registrate para obtener una cuenta gratuita).
* [10 ejemplos de mapas interactivos en la visualización de datos(se abre en una nueva pestaña)](https://www.tableau.com/learn/articles/interactive-map-and-data-visualization-examples) – artículo de Tableau para ver e interactuar con 10 visualizaciones interactivas de ejemplo
* [Consejos para crear visualizaciones de datos efectivas e interesantes (se abre en una nueva pestaña)](https://www.tableau.com/learn/articles/data-visualization-tips)– artículo de Tableau que explica cómo hacer que las visualizaciones de datos sean más interesantes, dinámicas y relevantes
* [Visualización del día(se abre en una nueva pestaña)](https://public.tableau.com/app/discover/viz-of-the-day) - foro público de Tableau Public en el que puedes ver una nueva visualización todos los días de un usuario
* Matplotlib

Existen numerosas bibliotecas en Python y [Matplotlib(se abre en una nueva pestaña)](https://matplotlib.org/" \t "_blank) es una de las bibliotecas más utilizadas. Matplotlib es una biblioteca multiplataforma que proporciona varias herramientas para crear gráficos bidimensionales a partir de datos, en listas o matrices, en Python. John Hunter introdujo Matplotlib en 2002. Matplotlib es un proyecto comunitario de código abierto mantenido por y para sus usuarios.

* El script Matplotlib Python está estructurado de modo que, en la mayoría de instancias, unas pocas líneas de código pueden generar un gráfico de datos visual.
* Puedes crear [diferentes tipos de gráficos(se abre en una nueva pestaña)](https://matplotlib.org/stable/gallery/index.html), como gráficos de dispersión, histogramas, gráficos de barras, etc.
* Las visualizaciones pueden ser estáticas, animadas e interactivas.
* Se pueden exportar a muchos tipos de formatos de archivo diferentes.

**Tómate un momento para revisar el ejemplo de Matplotlib**

Este es un ejemplo de un script para crear un gráfico de barras agrupadas. El script incluye el etiquetado de las barras.

***Estudia el siguiente script de Matplotlib y luego revela el gráfico.***

import matplotlib.pyplot as plt

import numpy as np

labels = ['G1', 'G2', 'G3', 'G4', 'G5']

men\_means = [20, 34, 30, 35, 27]

women\_means = [25, 32, 34, 20, 25]

x = np.arange(len(labels)) # the label locations

width = 0.35 # the width of the bars

fig, ax = plt.subplots()

rects1 = ax.bar(x - width/2, men\_means, width, label='Men')

rects2 = ax.bar(x + width/2, women\_means, width, label='Women')

# Agrega algo de texto para las etiquetas, título y etiquetas de verificación personalizadas del eje X, etc.

ax.set\_ylabel('Scores')

ax.set\_title('Scores by group and gender')

ax.set\_xticks(x, labels)

ax.legend()

ax.bar\_label(rects1, padding=3)

ax.bar\_label(rects2, padding=3)

fig.tight\_layout()

plt.show()

**Expandir para mostrar el gráfico**

¡El script de ejemplo de Matplotlib crea un gráfico de barras agrupadas!

**Fuente**: [Grouped bar chart with labels de Matplotlib(se abre en una nueva pestaña)](https://matplotlib.org/stable/gallery/lines_bars_and_markers/barchart.html" \t "_blank)

**¡Sigue explorando!**

Consulta los siguientes recursos:

* [Tutoriales de introducción de Matplotlib(se abre en una nueva pestaña)](https://matplotlib.org/stable/tutorials/introductory/index.html) - tutoriales de Matplotlib que abarca los conceptos básicos de crear visualizaciones, así como algunas de las prácticas recomendadas
* [The Lifecycle of a Plot(se abre en una nueva pestaña)](https://matplotlib.org/stable/tutorials/introductory/lifecycle.html) – tutorial de Matplotlib que muestra el principio, mitad y final de una sola visualización mediante Matplotlib

**Módulo 3**

**¿Existe la herramienta perfecta?**

Puedes *analizar*datos con “**esta**”herramienta o “**ese**”lenguaje deprogramación.

Puedes *visualizar*datos con “**esa**”herramienta o “**este**”lenguaje deprogramación.

Todas las herramienta de transformación de datos están diseñadas con un uso específico en mente. A veces, las herramientas se solapan y tienen el mismo propósito. Pero cada herramienta tiene sus ventajas y desventajas. Las experiencias pueden variar y cada persona tendrá sus propias preferencias.

Hay muchos factores que una empresa puede tener en cuenta a la hora de elegir las herramientas de análisis y visualización de datos. Estas son varias preguntas que una empresa puede tener en cuenta:

* ¿Qué tipos de datos se analizarán?
* ¿Es necesario realizar procesos específicos, como el modelado de datos?
* ¿Cuál es el volumen de datos? ¿Es Big Data?
* ¿Qué tipos de visualizaciones se necesitan? Por ejemplo, ¿necesitan ser interactivas?
* ¿Qué nivel de tecnicismo se necesita y qué funciones laborales utilizarán las herramientas? Por ejemplo, ¿utilizarán la herramienta científicos de datos o profesionales del marketing?
* ¿Cual es el presupuesto?

Es importante que las empresas seleccionen las herramientas correctas para sus necesidades, en el momento adecuado.

**Determinación del problema**

**Preguntas para formular:**

* ¿Cuál es el problema empresarial?

**Técnicas a tener en cuenta:**

* Formulación de preguntas
* Uso de los 5 Por qué

**Herramientas a tener en cuenta**:

* Pensamiento de diseño

**Recopilación y limpieza de datos**

**Preguntas para formular:**

* ¿Dónde están los datos?
* ¿Son los datos correctos?
* ¿Cómo se pueden recopilar los datos?
* ¿Qué se debe cambiar?

**Técnicas a tener en cuenta**:

* Recopilación y limpieza de datos estructurados y no estructurados

**Herramientas a tener en cuenta**:

* Lenguaje de consulta estructurado (SQL)
* NoSQL
* Microsoft Excel
* Hojas de cálculo de Google

**Análisis de datos**

**Preguntas para formular:**

* ¿Qué destaca?

**Técnicas a tener en cuenta:**

* Regresión, clasificación y agrupamiento
* Visualizaciones iniciales para buscar tendencias

**Herramientas a tener en cuenta**:

* Python
* IBM Watson Studio
* IBM SPSS
* RStudio

**Validación del análisis y del modelo**

**Preguntas para formular:**

* ¿Esto resuelve el problema empresarial?
* ¿Tienen sentido los resultados?

**Técnicas a tener en cuenta**:

* Aprendizaje automático
* Modelos de negocio

**Herramientas a tener en cuenta**:

* Python
* IBM Watson Studio
* IBM SPSS
* RStudio

**Visualización de datos**

**Preguntas para formular:**

* ¿Cuál es la mejor manera de representar los datos?
* ¿La visualización conecta los conocimientos con el problema?

**Técnicas a tener en cuenta**:

* Narración de historias
* Comunicación de información
* Presentación de los hallazgos

**Herramientas a tener en cuenta**:

* Tableau
* Matplotlib
* Microsoft PowerPoint
* **Resumen**
* Se pueden utilizar diferentes técnicas y herramientas durante un proyecto de ciencia de datos. ¡Los proyectos generalmente implican tomar decisiones y emprender acciones! ¿Qué hará una empresa y cuándo lo hará?
* **Nota**: Estos ejemplos representan una muestra de cómo se alinean algunas herramientas en un proyecto de ciencia de datos. Hay muchos más.

**Puntos clave por recordar**

1. 1

El software de código abierto es un software con código que se publica públicamente para que cualquiera pueda utilizarlo y colaborar entre una comunidad virtual de personas. El código abierto permite a quienes trabajan en proyectos de ciencia de datos aprender nuevo software y desarrollar nuevas habilidades. Proporciona muchos beneficios potenciales a las empresas, como innovación acelerada, ahorro de costos y competitividad.

1. 2

**GitHub**es un servicio gratuito en línea que se utiliza para alojar, colaborar y contribuir en el código fuente mediante repositorios.

1. 3

**Microsoft Excel** es un software que se utiliza para introducir, manipular, analizar y visualizar datos. **Hojas de cálculo de Google** es una alternativa a Microsoft Excel.

1. 4

**Lenguaje de consulta estructurado (SQL)** es un lenguaje estándar para "consultar" o comunicarse con bases de datos para modificarlas.

1. 5

**IBM Watson Studio** tiene una herramienta de refinación de datos con cuadros y gráficos integrados para comprender la distribución de los datos.

1. 6

**Tableau**es un software para analizar grandes volúmenes de datos y crear diferentes tipos de visualizaciones, incluyendo paneles interactivos.

1. 7

**Matplotlib**es una biblioteca en Python que se utiliza para crear diferentes gráficos y diagramas.

1. 8

Hay muchas herramientas de ciencia de datos y lenguajes de programación que se pueden utilizar durante un proyecto de ciencia de datos. ¡Es importante seleccionar la herramienta correcta para cada trabajo, en el momento adecuado!

1. 9

Muchas herramientas no requieren experiencia previa en codificación. Algunas herramientas te permiten probarlas y utilizarlas de forma gratuita o consultar tutoriales en línea.

**Limpia, refina y visualiza datos con IBM Watson Studio**

En este curso, practicarás la limpieza, la refinación y la visualización de datos en una serie de simulaciones utilizando IBM Watson Studio con la herramienta de refinación de datos.

**Módulo 1**

**Introducción a IBM Watson Studio**

IBM Watson Studio ofrece muchas herramientas, incluida una herramienta de refinería de datos. Es beneficioso porque ahorra el tiempo de preparación de datos. Puede transformar rápidamente grandes cantidades de datos sin procesar en información consumible de gran calidad lista para el análisis.

Con él, los científicos de datos pueden:

* crear un flujo de trabajo para limpiar y dar forma a los datos
* comprender la calidad y distribución de los datos mediante docenas de gráficos, tablas y estadísticas integrados
* programar trabajos de datos para resultados repetibles
* visualizar datos para descubrir información

La herramienta de refinería de datos es interactiva y fácil de utilizar. No necesitas habilidades de codificación. ¡Hay más de 100 operaciones integradas disponibles para ayudarte a transformar datos!

* **Tu lista de tareas por hacer**

Mediante IBM Watson Studio y la herramienta de refinería de datos, podrás:

* + aprovisionar el servicio Watson Studio
  + configurar un nuevo proyecto
  + importar los datos de las reclamaciones de seguros de automóviles
  + limpiar el conjunto de datos
  + refinar el conjunto de datos
  + crear y estudiar una visualización de datos para obtener información
* **Tu conjunto de datos**
  + Estás utilizando un conjunto de datos ficticio, que está basado en el registro de eventos de una compañía de seguros de automóviles real.
  + Es un archivo de valores separados por comas (CSV), cada fila es un registro de datos.
  + Hay 975 filas y 38 columnas de datos.
  + Cada fila contiene los datos de la reclamación de una persona. La persona presentó una reclamación de reparación de automóvil después de un accidente. Tu compañía de seguros de automóviles ya ha aprobado todas estas reclamaciones.
* **Ten presente esta información sobre el seguro de automóvil**
  + Por lo general, los conductores contratan una póliza de seis meses con una compañía de seguros de automóviles.
  + El conductor paga todos los meses, o de una sola vez, una cuota o prima a la compañía.
  + Algunos factores determinan el costo de la póliza, como:
    - el tipo de vehículo asegurado, especialmente su registro de seguridad y qué tan caro es repararlo
    - el historial del conductor, por ejemplo, cuantos más multas por exceso de velocidad tenga el conductos, más arriesgado es el conductor
    - la edad del conductor, por ejemplo, los adolescentes son más caros de asegurar porque tienen menos experiencia y, por lo tanto, tienen un mayor riesgo
  + Los conductores con menos accidentes y multas en su historial, los conductores a tiempo parcial que no conducen mucho, las personas que realizan cursos de educación vial y las familias con muchos automóviles disfrutan de primas más bajas.

**Módulo 2**

**Comienza tu proyecto**

**Transcripción de simulación Iniciar un proyecto y cargar datos**

1. Después de iniciar sesión en IBM Cloud, encontrarás el Panel de control. Este Panel de control te permite acceder a las herramientas, servicios, recursos y más en tu cuenta de IBM Cloud. Selecciona **Catalogar** en la barra de herramientas superior.

2. Este es el catálogo de productos y servicios de IBM. Observa que en el lado izquierdo abajo de **Categoría** hay una gran colección de herramientas disponibles. Selecciona la categoría **IA/Aprendizaje automático**.

3. Verás que se enumeran varios productos de IA y aprendizaje automático. Busca y selecciona la opción **Watson Studio**.

4. Ahora es el momento de disponer IBM Watson Studio como tu servicio. Es buena idea elegir una ubicación que te quede cerca y seguir usando la misma ubicación en todos tus proyectos. En **Seleccionar una ubicación**, selecciona **Dallas (sur de EE. UU.)** de la lista desplegable.

5. Para esta simulación, estás utilizando una cuenta gratuita, por lo que no necesitas revisar un plan de precios.

**Desplázate hacia abajo** y selecciona la **flecha Siguiente** para continuar.

6. En la sección **Configura tu recurso**, hay un **Nombre del servicio** generado automáticamente. Sería bueno mantener "Watson Studio" y luego agregar un nombre que sea relevante para tu proyecto. Escribe " Watson Studio-AutoIns" en el campo de **Nombre del servicio**.

7. No es necesario que completes los demás campos. En la parte inferior derecha, haz de cuenta que has marcado que has leído y aceptado los acuerdos de licencia, y luego selecciona **Crear**.

8. ¡Has dispuesto el servicio IBM Watson Studio! Observa que el nombre del servicio que seleccionaste se muestra en la parte superior izquierda. Ahora necesitas crear un nuevo proyecto. Selecciona **Iniciar en IBM Cloud Pak for Data**.

9. Estás trabajando en un nuevo proyecto, así que selecciona **+** junto a **Proyectos** y luego selecciona **Crear un proyecto vacío**.

10. En la sección **Definir detalles**, puedes nombrar tu proyecto. Escribe "Seguro de automóviles" en el **campo Nombre** y presiona **Intro**.

11. Observa que en **Almacenamiento** IBM Watson Studio ya ha dispuesto Cloud Object Storage (COS), por lo que tienes una base de datos. Este COS puede almacenar datos no estructurados como imágenes y texto. Pronto trabajarás con datos estructurados en un archivo de valores separados por comas (CSV). Selecciona **Crear**.

12. ¡Has creado el proyecto! Este es el panel de control de los Proyectos de IBM Watson Studio. Puedes ver el nombre de tu proyecto en la parte superior. Esta es la pestaña **Descripción general**. Selecciona la pestaña **Recursos**.

13. Es hora de cargar tu conjunto de datos. Selecciona **Nuevo recurso** para abrir tu unidad local y encontrar el archivo AutoInsClaims.csv con el que trabajarás para el proyecto.

14. Arrastra y suelta el archivo AutoInsclaims.csv en **Soltar archivos de datos o busca archivos para cargar** en Datos de este proyecto.

15. Después de unos momentos, los datos se cargarán. A la izquierda, selecciona **Todos los recursos**.

16. Observa que tu archivo CSV está allí mismo. Pasa el cursor por encima y selecciona **AutoInsClaims.csv**.

17. Se abrirá una página de **Vista previa del recurso** de modo que puedas obtener una vista previa de tus datos. Observa que estás trabajando con 38 columnas y muchas filas de datos. Selecciona la **flecha Siguiente** para continuar.

Has dispuesto correctamente el servicio IBM Watson Studio, has configurado un nuevo proyecto y has cargado el conjunto de datos en tu proyecto, de modo que estás listo para comenzar a trabajar con él.

**Módulo 3  
Comprensión y limpieza de datos**

En este módulo, obtendrás experiencia práctica en una simulación para comprender y limpiar los datos en IBM Watson Studio con la herramienta de refinería de datos.

**Transcripción de la simulación Comprender y limpiar los datos**

1. Has cargado los datos de tu archivo CSV a tu proyecto y esta es una vista previa. Puedes comenzar a explorar tu conjunto de datos aquí en la vista previa o puedes comenzar a utilizar la herramienta de refinación de datos. Vamos a empezar. Selecciona **Preparar datos** para comenzar el refinamiento de datos.

2. La herramienta de refinación de datos puede tardar varios minutos en leer y refinar los datos. Antes, el conjunto de datos era de solo lectura y ahora es editable para que puedas trabajar con él. Observa que hay tres pestañas en la parte superior: **Datos**, **Perfil** y **Visualizaciones**. Selecciona la **flecha Siguiente** para continuar.

**Estás listo para comprender y limpiar los datos**

Recuerda que estos datos provienen de reclamaciones de seguros de automóviles que la compañía aprobó. El objetivo de tu proyecto es predecir reclamaciones fraudulentas. Por lo tanto, debes examinar la **previsibilidad** de los datos en **cada columna**. ¿Los datos de la columna te ayudarán a predecir un fraude? Algunas columnas lo harán y otras no. Tú decides. Vamos a ver.

Selecciona **X** para cerrar esta ventana y continuar.

3. Las primeras cuatro columnas son números de identificación, como la identificación del conductor y la identificación de la póliza. Estos datos no ayudan a predecir quién podría cometer fraude. Son simplemente identificadores. Vamos a asegurarnos. Selecciona la pestaña **Perfil**.

4. La pestaña **Perfil** de la herramienta de refinación de datos proporciona estadísticas descriptivas que la herramienta calculó a partir de los datos. Te ayuda a analizar y decidir qué conservar y qué eliminar. Selecciona la **flecha Siguiente** para continuar.

5. Al observar las estadísticas de las primeras cuatro columnas, hay 975 eventos únicos. Si ves estadísticas que simplemente cuentan las filas en un archivo CSV, entonces debes eliminar esa columna. No hay nada interesante en los datos y no son predecibles. Selecciona la pestaña **Datos**.

6. Vamos a borrarlos. Selecciona los **tres puntos** al lado de la columna **HOUSEHOLD\_ID**. Esto te brinda varias opciones de lo que puedes hacer con la columna. Selecciona **Eliminar columna**.

7. Haz lo mismo con los demás datos de identificación que no ayudarán a tu proyecto a predecir comportamientos fraudulentos. Selecciona los **tres puntos** al lado de la columna **DRIVER\_ID**. Selecciona **Eliminar columna**.

8. Selecciona los **tres puntos** al lado de la columna **POLICY\_ID**. Selecciona **Eliminar columna**.

9. Selecciona los **tres puntos** al lado de la columna **CLAIM\_ID**. Selecciona **Eliminar columna**.

10. ¡Ahora estás comprendiendo y limpiando los datos! Has pasado de 38 columnas a 34 columnas. Sigamos entendiendo y limpiando. Selecciona la **flecha Siguiente** para continuar.

11. La columna **INCIDENT\_CAUSE** es interesante. Se trata de la causa del accidente automovilístico en el que 1 es para error del conductor, 2 es para causas naturales, 3 es para error de otro conductor, 4 es para un delito y 5 es para otras causas. Podría ayudarte a predecir un fraude. Selecciona la pestaña **Perfil** para investigar más a fondo.

12. La herramienta de refinación de datos muestra estadísticas interesantes para el campo **INCIDENT\_CAUSE**. Observa que hay un valor para **Desviación estándar**. ¿Qué significa eso? Selecciona la **flecha Siguiente** para continuar.

**¿Qué es la desviación estándar?**

La desviación estándar es un cálculo estadístico que indica qué tan dispersos están los datos en relación con la media.

• La media es el promedio de un conjunto de datos.

• Una desviación estándar baja significa que los datos están agrupados alrededor de la media.

• Una desviación estándar alta indica que los datos están más dispersos.

Para ti, cualquier columna en la pestaña **Perfil** que tiene un valor en el campo de **Desviación estándar** podría contener datos interesantes para realizar predicciones sobre tu proyecto. Conserva estos datos.

Selecciona **X** para cerrar esta ventana y continuar.

13. Continuemos. Selecciona la pestaña **Datos** para investigar más columnas de datos.

14. Observa que la columna **DESCRIPCIÓN** está en blanco. Es posible que te topes con datos faltantes mientras trabajas en un proyecto. No es de ayuda. Selecciona los **tres puntos** y selecciona **Eliminar columna**.

15. La siguiente columna para **CLAIM\_STATUS** es interesante. Tiene datos sobre el estado de la reclamación en la cual 1 significa abierto, 2 significa aprobado, 3 significa pagado y 4 está marcado como fraude. Esta columna seguramente podría ayudarte a predecir un fraude, así que consérvala. Selecciona la **flecha Siguiente** para continuar.

16. Puedes conservar la columna **ODOMETER\_AT\_LOSS** porque podría ser interesante considerar el kilometraje del vehículo en el momento del accidente. Selecciona la **flecha Siguiente** para continuar.

17. A continuación, revisa las columnas **LOSS\_EVENT\_TIME** y **CLAIM\_INIT\_TIME**. Selecciona la **flecha Siguiente**

para continuar.

**¡Recuerda ser curioso!**

Para este proyecto de ciencia de datos, es una buena idea mantener columnas que tengan información de fecha y hora. Nunca se sabe lo que podría revelar en términos de predicciones. Por ejemplo, si alguien esperó mucho tiempo para presentar una reclamación en lugar de presentarla el día del accidente automovilístico, ¿qué podría significar eso?

Observa cómo estás siendo curioso. ¡Eres como un detective siguiendo pistas! Los científicos de datos necesitan pensar de esta manera.

Puedes conservar las columnas **LOSS\_EVENT\_TIME** y **CLAIM\_INIT\_TIME**. Selecciona **X** para cerrar esta ventana y continuar.

18. Es importante guardar mientras trabajas. ¡Y guarda a menudo! En la parte superior derecha, selecciona el ícono **Guardar**.

19. A continuación, la columna **POLICE\_REPORT** es importante y podría ser útil. Ten en cuenta que el tipo de datos se llama Entero. Esto significa que es un 0 (que significa no) o un 1 (que significa sí). O bien se presentó una denuncia ante la policía o no. Puedes conservar esta

columna.Seleccionala**flechaSiguiente**para continuar.

20.Pasemos alacolumna**CLAIM\_AMOUNT**.¿Cuántocostóarreglar elautomóvildespués delaccidente? Estoes muyinteresante para tuproyecto.Tuempresa pagaestas reclamaciones.Conserva esta columna.Seleccionala**flechaSiguiente**para continuar.

21.Acontinuación,revisala columna**FLAG\_FOR\_FRAUD\_INV**.Seleccionala**flecha Siguiente**paracontinuar.

**Sehan marcadoreclamacionespor fraude**

La columna **FLAG\_FOR\_FRAUD\_INV**es única. Los datosprovienen del equipodeinvestigadores de fraude que revisaronyconcluyeron sila reclamaciónde la personaenesafilano erafraude(0)osí era fraude (1).Se ingresómás tarde, alfinalde año.Estas sonrespuestas conocidaspara ti.Estosdatos son un conjuntodeentrenamiento para quelaherramientaconstruyaun modelo deentrenamientosupervisado.IBMWatson Studioanalizaráestacolumna,seleccionarádatos al azar ycrearápatrones parati.

Puedes conservar la columna**FLAG\_FOR\_FRAUD\_INV**.Además, recuérdalapara más tardecuandovisualices los datos.

Selecciona**X** paracerrar estaventana y continuar.

22.Acontinuación,lacolumna**PRIMARY\_DRIVER\_ID**tiene datosde identificador.Recuerda que losdatosde tipode identificaciónnoseránpredecibles paraeste proyecto.Seleccionalos**trespuntos** y selecciona**Eliminarcolumna**.

23.Esimportante guardarmientrastrabajas.¡Yguardaamenudo!En la partesuperiorderecha,seleccionaelícono**Guardar**.

24.Pasemos alas columnas **FIRST\_NAME**y **LAST\_NAME**.¿Podrían estos datos ayudarteapredecir unfraude? Seleccionala**flechaSiguiente**paracontinuar.

**Ten cuidadocon la información personal (IP)**

Algunascolumnasdetu conjuntodedatos representan información personal(IP)sobre lapersonaficticiaque realizóla reclamación delautomóvil que pagótuempresa.

Lainformación personal escualquierinformación relacionadaconun individuovivoeidentificable.Cuando se recopila informaciónpersonal,puede llevar a identificar aunapersonaenparticular.Por ejemplo,nombres y apellidos,números deteléfono,direcciones decorreoelectrónico,números de matrículayfechasde nacimiento son informaciónpersonal.Debesestar atentoadichos datos enlos proyectos y siestápermitidosuuso.

Para este proyectode cienciade datos,si lopiensas,la informaciónpersonalnoespredecible en cuantoa fraude. Elnombre o ladirección de correo electrónico dealguien noteayudarán apredecir unfraude.Eliminaremos estos datos por ti.

Selecciona**X** paracerrar esta ventana y continuar.

25.Hastaahora, has eliminadocolumnas con datos faltantes onopredecibles.Hay otropasoparalimpiar los datos.Debes examinar el**tipo de datos** decada columna para estar segurode que seacorrecto.La herramientade refinaciónde datosya harealizadoalgunascorrecciones.Seleccionala**flecha Siguiente**para continuar.

26.Ten en cuentaqueel tipode datospara**ODOMETER\_AT\_LOSS**se muestra correctamente como

decimal. Selecciona la**flecha Siguiente** paracontinuar.

27.Pero,eltipodedatos para**LOSS\_EVENT\_TIME**esunacadena.Lascadenasson unasecuenciade letras,dígitos,signos de puntuación, etc.,comounadirecciónpostal.Estacolumnadebeser correctaparaser unafecha para quelaherramientapuedarealizar cálculos.Seleccionalos **trespuntos**,selecciona**Convertirtipode columna**.

28.La ventanaizquierdate permite personalizar eltipode datosparatucolumna. Desde elmenú desplegablede**Tipo**,selecciona**Fecha**.

29.Enla listade**Seleccionaelpedido…**ahora verás que**mda**estáseleccionado.Esto significaquelos datossemostraráncomomes (m),día(d)y año (a).Selecciona**Aplicar**.

30.Observa quetu actualizaciónseguardóyla columna **LOSS\_EVENT\_TIME**ahora muestra correctamenteeltipodedatos comoFecha.Elrestodelos tipos de datos seactualizaránpor ti,paraquepuedas seguir adelante.Seleccionala**flechaSiguiente**para continuar.

31.Ahoraestás listoparaguardar tu trabajo y crear untrabajoderefineríaenlaherramienta.Estocreaunarchivo CSVseparadoque contienetus cambios.En labarra deherramientas superior,a la derecha dedondeguarda, selecciona**Guardary crearun trabajo**.

32.Para crear untrabajoparatuproyecto,comienza escribiendo"Predicción de fraude ensegurosde automóviles"enelcampo**Nombre**y luego presiona**Intro**.

33.Puedesdejaren blancoelcampo**Descripción**.Selecciona**Siguiente**.

34.Enla sección**Configurar**,puedes aceptar los valores predeterminados.Selecciona**Siguiente**.

35.Enla sección**DocumentoTransaccional**,puedes mantenerlaprogramación desactivada. Ten encuentaque siestabastrabajando alargoplazo en este proyecto,talvez quierasconfigurar eltrabajoparaque se ejecute amedianoche sillevamuchotiempo,de modo que esté listocuandocomiences atrabajar.Selecciona**Siguiente**.

36.Enla sección**Notificar**,puedesmantener lasnotificaciones desactivadas.Selecciona**Siguiente**.

37.Observaque el archivoCSV original quecargaste se encuentra en**Fuente**.En**Objetivo**seencuentraelnuevoarchivoCSV que la herramientade refinaciónde datosestácreandoparati. Tienela palabra"\_shaped"enelnombredel archivo.Estees el archivoal que deberás acceder en el futuroparatuproyecto.Selecciona**Crear yejecutar**.

38.¡Prestaatención ala notificación!El trabajofue creado exitosamente.Seleccionala**flechaSiguiente**para continuar.

Limpiaste exitosamente elconjuntode datosaleliminar lascolumnasfaltantesy lascolumnascondatos que noson predecibles.También corregisteeltipodedatos quesemuestran en unacolumnapararealizar cálculos adecuados y creasteeltrabajoen laherramienta.

**¡Buen trabajo hasta ahora!**

Los pasos que acabas de completar para comprender y limpiar el conjunto de datos llevó su tiempo. Es un paso importante y largo en un proyecto de ciencia de datos. Los analistas y científicos de datos dedican mucho tiempo a este tipo de trabajo.

¡Has obtenido una gran experiencia práctica!

* Utilizaste la pestaña **Datos** para revisar muchas columnas para comprender mejor los datos con los que trabajas para investigar el problema empresarial de tu empresa sobre las reclamaciones fraudulentas.
* Determinaste qué columnas debes **mantener** ya que **contienen datos predecibles** que pueden ayudarte a predecir los fraudes, como la causa del incidente y las columnas con información sobre las horas y las fechas.
* Determinaste qué columnas debes **eliminar** ya que **no contienen datos predecibles**, como aquellas en las que faltan datos, contienen números de identificación e información personal.
* Utilizaste la pestaña**Perfil**para consultar las estadísticas descriptivas, como la desviación estándar, que proporciona la herramienta de refinería de datos. Estas estadísticas te ayudaron a decidir si una columna es interesante y puede contener datos predictivos que debes mantener.
* Además, corregiste el **tipo de datos** en una columna para asegurarte de que la herramienta pueda realizar cálculos y garantizar que es un buen conjunto de datos para tu proyecto.

A continuación, estarás listo para continuar trabajando y refinando los datos. Pero primero, haz un breve cuestionario.

**Módulo 4**

**Refinería de datos**

**Transcripción de simulación Refinar los datos**

1. Aquí está el archivo CSV "\_shaped" que creó la herramienta de refinación de datos. Ten en cuenta que ahora estás trabajando con 19 columnas de datos tras tus esfuerzos de limpieza. Es hora de refinarlo en función de la hipótesis de tu patrocinador comercial de que las reclamaciones superiores a $10 000 pueden ser fraudulentas. Selecciona la **flecha Siguiente** para continuar.

2. Estás creando una nueva columna que muestra qué valores de la columna **CLAIM\_AMOUNT** son mayores a $10 000. Selecciona los **tres puntos** de la columna de **CLAIM\_AMOUNT** y selecciona **Nuevo paso…**

3. Aparece un menú **LIMPIAR** para la columna de **CLAIM\_AMOUNT**. Selecciona **Calcular**.

4. Estás buscando los datos de las reclamaciones de automóviles que pagó tu compañía que son superiores a $10 000. Selecciona el menú desplegable debajo de **Operador**.

5. Ahora, selecciona **Es mayor que**.

6. Escribe "10000" en el campo **Valor**. Luego, marca la casilla para **Crear una nueva columna para los resultados**.

7. Es hora de seleccionar un nombre para tu nueva columna. Escribe "EXCESSIVE\_CLAIM\_AMOUNT" en el campo **Nombre de nueva columna** y presiona **Intro**.

8. Querrás que esta nueva columna sea la que esté más a la derecha en el conjunto de datos, de modo que puedas mantener el valor predeterminado en **Posición de la nueva columna**. Selecciona **Aplicar**.

9. ¡Tu nueva columna llamada **EXCESSIVE\_CLAIM\_AMOUNT** se ha agregado al final del conjunto de datos! Selecciona la **flecha Siguiente** para continuar.

10. Vamos a examinarlo. Ten en cuenta que los valores son **verdaderos** o **falsos**. Este es el tipo de datos booleanos. Será más fácil ver estos datos si los cambia de booleanos a enteros. Selecciona los **tres puntos** al lado de la columna **EXCESSIVE\_CLAIM\_AMOUNT**.

11. Selecciona **Convertir tipo de columna…**

12. Este menú te brinda opciones para convertir el tipo de datos de tu columna. En el menú desplegable de **Tipo**, selecciona **Entero**.

13. Deja todo como está aquí y selecciona **Aplicar**.

14. Ahora los datos en tu última columna se muestran como **0** o **1**. **0** significa que la reclamación en esa fila no es superior a $10 000. **1** significa que la reclamación en esa fila es superior a $10 000. Estos serán datos muy útiles para determinar si es cierta la hipótesis de que las reclamaciones superiores a $10 000 pueden ser fraudulentas. Selecciona la **flecha Siguiente** para continuar.

Agrega con éxito una nueva columna para refinar el conjunto de datos de modo que puedas investigar la hipótesis de que las reclamaciones superiores a $10 000 pueden ayudar a predecir un fraude.

**Módulo 5**

**Visualización de datos**

**Transcripción de simulación Visualizar los datos**

1. Es hora de ver qué historia cuentan los datos. Selecciona la pestaña **Visualizaciones**.

2. La herramienta ofrece muchos tipos de gráficos. Selecciona el ícono de **flecha doble**.

3. ¡Mira todas las opciones de tipos de gráficos! Ten en cuenta que una visualización no comenzará siendo perfecta. Debes experimentar para probar diferentes gráficos. Para tu proyecto, una buena visualización inicial para explorar datos es un gráfico de dispersión. Selecciona **Gráfico de dispersión**.

4. Selecciona las columnas de datos que deseas visualizar. Observa tu eje x. Es tu variable independiente. Quieres que esta sea la nueva columna que agregaste porque deseas ver si las reclamaciones superiores a

$10 000 se clasifican como 0 para ausencia de fraude o 1 para fraude. En el **Eje x**, selecciona

**EXCESSIVE\_CLAIM\_AMOUNT**.

5. Observa tu eje y. Es tu variable dependiente. Para tu proyecto, debe ser un monto en dólares para que puedas ver cuántas reclamaciones superiores a $10 000 corresponden a 0 para ausencia de fraude o 1 para fraude. En el **Eje Y**, selecciona **CLAIM\_AMOUNT**.

6. ¡Observa la visualización inicial! Tienes una distribución de puntos en tu gráfico de dispersión. Los puntos arriba del **0** representan todas las reclamaciones por debajo de $10 000. Los puntos arriba del **1** representan todas las reclamaciones superiores a $10 000. Esto es interesante hasta ahora. Selecciona la **flecha Siguiente** para continuar.

7. Pero es necesario profundizar más. Recuerda que los investigadores de fraude concluyeron qué reclamaciones eran fraudes reales y agregaron una columna al conjunto de datos. Estas son las "respuestas" que necesitas. Agreguemos estos datos en código de colores. En **Mapa de colores**, selecciona **FLAG\_FOR\_FRAUD\_INV**.

8. ¡Esto revela algo interesante en el diagrama de dispersión! Los puntos azules son reclamaciones individuales que los investigadores concluyeron que *no son fraude*. Los investigadores concluyeron que los puntos rosados *eran fraude*. Selecciona la **flecha Siguiente** para continuar.

9. Esta es una visualización interactiva en IBM Watson Studio. ¡Vamos a explorar! Pasa el cursor sobre diferentes áreas del gráfico para ver algunos puntos de datos. Selecciona la **flecha Siguiente** para continuar.

10. Seleccione el **área azul** en el eje y justo arriba del 0.

11. En primer lugar, esto significa que hay algunas reclamaciones *inferiores a* $10 000 que *no son* fraude. Esta reclamación por $1720 está en azul. El color azul significa que los investigadores consideraron que no se trataba de fraude. Ahora selecciona el **área rosa** en el eje y justo encima de allí.

12. En segundo lugar, esto significa que hay algunas reclamaciones *inferiores a* $10 000 que *son* fraude. Esta reclamación por $2730 está en rosa. El color rosa significa que los investigadores consideraron que se trataba de un fraude. ¡Interesante! Ahora selecciona el **área rosa** arriba del 1 justo arriba de 40 000 000.

13. En tercer lugar, se puede ver que hay muchas reclamaciones *superiores a* $10 000 que *son fraude* porque el color rosa significa que los investigadores consideraron que eran fraude. Esto significa que la hipótesis del patrocinador comercial de que las reclamaciones superiores a $10 000 podrían indicar fraude es correcta. Selecciona la **flecha Siguiente** para continuar.

14. Pero, ¡mira! También hay reclamaciones *superiores a*$10 000 que *no son fraude* porque el color azul significa que los investigadores consideraron que no se trataba de fraude. Por lo tanto, no todas las reclamaciones superiores a $10 000 son fraudulentas. ¡Interesante!

Seleccionala**flechaSiguiente** para continuar.

15.Estaes una visualizacióninformativa.Puedes usarlaparaestimar estos conocimientos y compartirlos conelpatrocinador comercial,asíqueselecciona**Guardar visualizaciónenelproyecto** enlapartesuperior derechaparaguardarloy verloenel futuro.

Hascreadoconéxitounavisualizaciónde diagramade dispersiónenlaherramienta de refinaciónde datos paramostrar datosquepuedasutilizarparaestimar información.

**Puntos clave por recordar**

1. 1

IBM Watson Studio es un entorno de desarrollo integrado (IDE) que ofrece la herramienta de refinería de datos para limpiar, dar forma y visualizar datos.

1. 2

El objetivo de tu proyecto de ciencia de datos es predecir el fraude y la actividad sospechosa para ayudar a reducir considerablemente las pérdidas debidas a los fraudes. El fraude es un problema para la compañía de seguros de automóviles.

1. 3

Trabajaste con un conjunto de datos de reclamaciones de seguros de automóviles ficticias con 975 filas y 38 columnas.

1. 4

Tuviste la oportunidad de obtener experiencia práctica en simulaciones de IBM Watson Studio y la herramienta de refinería de datos para:

* + aprovisionar el servicio de Watson Studio
  + configurar un nuevo proyecto
  + importar datos de reclamaciones de seguros de automóviles
  + limpiar el conjunto de datos
  + refinar el conjunto de datos
  + crear y estudiar una visualización de datos para obtener información

1. 5

Has considerado la hipótesis de la patrocinadora empresarial y has sacado conclusiones de la visualización de datos que creaste.

Si deseas obtener más información sobre IBM Watson Studio y la herramienta de refinería de datos, consulta los siguientes recursos:

* [Versión de prueba de IBM Watson Studio(se abre en una nueva pestaña)](https://dataplatform.cloud.ibm.com/registration/stepone?context=cpdaas&apps=data_science_experience%2Cwatson_machine_learning%2Ccos) – para utilizar una versión de prueba de IBM Watson Studio en Cloud Pak Data (**Nota**: Necesitarás registrarte para obtener una cuenta gratuita).
* [Data Refinery(se abre en una nueva pestaña)](https://www.ibm.com/cloud/data-refinery) – la página de productos de IBM para obtener más información sobre la herramienta de refinería de datos de IBM Watson Studio
* [Preparing data with Data Refinery(se abre en una nueva pestaña)](https://mediacenter.ibm.com/media/Lab%205%3A%20Preparing%20data%20with%20Data%20Refinery/1_zpxhku5t) – video de IBM Experts Lab que demuestra cómo un científico de datos recopila tres conjuntos de datos, los limpia y selecciona una visualización inicial
* [Use the Data Refinery to shape raw data (se abre en una nueva pestaña)](https://video.ibm.com/recorded/128592867)– video de IBM que demuestra cómo dar forma a datos sin procesar de ejemplo mediante la herramienta de refinería de datos

**Tu Futuro en Datos: Panorama Laboral**

En este curso aprenderá sobre el mercado laboral y proyecciones de una carrera en datos. Esto incluye sus responsabilidades y conjunto de habilidades que una persona científica o analista de datos, así como los recursos y oportunidades de aprendizaje para que puedas explorar más.

Después de completar este curso, deberías poder:

* reconocer la demanda global de analistas de datos y científicos de datos en el mercado laboral
* reconocer el futuro del campo del análisis de datos
* identificar las industrias en las que trabajan los profesionales de datos
* identificar las principales responsabilidades de un analista de datos y un científico de datos
* identificar las habilidades que necesitan los profesionales de datos
* identificar las herramientas que debes conocer al iniciarte en el rubro
* identificar recursos para obtener más información y mantenerte actualizado en el campo de la ciencia de datos
* **Módulo 1**
* **El mercado laboral y el futuro**

### Profesión de mayor crecimiento

El Foro Económico Mundial (FEM) colocó los datos y la inteligencia artificial (IA) entre las siete profesiones con mayor crecimiento en 2020, con una tasa de crecimiento del 41 %. Esta es la tasa de crecimiento más alta de todas las profesiones mencionadas en el informe.  
– [Informe sobre los empleos del futuro del FEM(se abre en una nueva pestaña)](http://www3.weforum.org/docs/WEF_Jobs_of_Tomorrow_2020.pdf)

### 33 %

Se proyecta que los empleos de análisis de datos crecerán un 33 % entre 2020 y 2030 en Estados Unidos, lo que es mucho más rápido que el promedio de todas las ocupaciones.  
– [Oficina de Estadísticas Laborales de EE. UU.](https://www.bls.gov/ooh/math/mathematicians-and-statisticians.htm)

* **Más de 137 000**

Analytics Insight ha pronosticado que para 2025 habrá 137 630 vacantes de empleo en ciencia de datos en India, frente a los 62 793 empleos de 2020.  
– [Analytics Insight(se abre en una nueva pestaña)](https://www.analyticsinsight.net/analytics-insight-predicts-137630-new-jobs-in-data-science-in-india-by-2025/?utm_source=skillsbuild.org" \t "_blank)

* **Puestos 3 y 7**

Glassdoor compiló una lista de los *50 mejores trabajos en Estados Unidos* *para 2022* según el potencial de ingresos, la satisfacción laboral y las vacantes laborales. El rol de científico de datos ocupa el puesto número 3. El rol de ingeniero de datos ocupa el puesto número 7.  
– [Glassdoor](https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm" \t "_blank)

El **análisis predictivo** incluso podría usarse en tu hogar a través de **dispositivos de Internet de las cosas (IoT)**, como termostatos inteligentes, electrodomésticos y monitores de actividad física, por nombrar algunos. Estos tipos de dispositivos de IoT recopilan datos tuyos para predecir tu comportamiento y avanzar aún más en la automatización del hogar.

Además, muchas personas se han acostumbrado a los asistentes virtuales, como Google Search, Alexa y Siri, que utilizan **procesamiento del lenguaje natural (PLN)** comprender y procesar el lenguaje humano en tiempo real. El PNL tiene el potencial de hacer que tanto las aplicaciones empresariales como las de consumo sean más fáciles de usar. Cuando se utiliza junto con **inteligencia artificial (IA)**, el PNL podría ayudar a los profesionales a resolver desafíos globales, como la energía limpia.

**Fuente:** [¿Qué es el análisis de datos? de CompTIA(se abre en una nueva pestaña)](https://www.comptia.org/content/guides/what-is-data-analytics)

En general, en el futuro los negocios se realizarán de manera diferente debido al análisis de datos. Las empresas con una sólida base en el área de datos llevarán la delantera. Los profesionales de datos ayudan a las empresas a crecer, resolver problemas y sobresalir en este mundo digital de datos. Un trabajo en análisis de datos proporciona una carrera segura, con muchas oportunidades de crecimiento profesional. ¡El futuro es brillante!

**Módulo 2**

**Trabajar con datos**

En este módulo, aprenderás sobre las principales responsabilidades de un analista de datos y un científico de datos, y las numerosas industrias que dependen de los datos.

¡En general, los profesionales de datos trabajan en todas partes! Los encontrarás en grandes empresas globales y organizaciones medianas en todas las industrias, como:

* Finanzas
* Seguros
* Cuidado de la salud
* Servicios científicos
* Tecnología de la información (TI)
* Marketing
* Gestión de empresas
* Fabricación
* Administración pública
* Servicios educativos
* Artes y entretenimiento
* Venta minorista
* Haz clic para voltear
* **Finanzas**
* En **finanzas**, el análisis de datos puede ayudar a los bancos y a las instituciones financieras a detectar fraudes y clientes que podrían incumplir sus préstamos.
* **Cuidado de la salud**
* En el **cuidado de la salud**, el análisis de datos puede ayudar a proporcionar diagnósticos que salven vidas basándose en los datos del paciente. El análisis de datos también puede ayudar a descubrir nuevos métodos de desarrollo de fármacos.
* **Fabricación**
* En el área de **fabricación**, el análisis de datos puede ayudar a reducir costos al resolver problemas con la cadena de suministro y las fallas de los equipos.
* **Venta minorista**
* En la **venta minorista**, el análisis de datos puede ayudar a las empresas a comprender las necesidades de los clientes y los hábitos de compra para predecir tendencias. Las empresas minoristas pueden recomendar nuevos productos para impulsar su negocio.

**Responsabilidades del analista de datos**

a responsabilidad principal de un analista de datos es **recopilar, limpiar y analizar datos utilizando los métodos y herramientas estándar de una empresa para detectar tendencias, encontrar patrones y hacer pronósticos.**

Los analistas de datos se centran en un problema comercial específico identificado por el patrocinador comercial o el científico de datos, como, por ejemplo, "¿Por qué el equipo de TI recibe tantas solicitudes por defectos este año en comparación con el año pasado?". Los analistas de datos recopilan y analizan datos para encontrar información y luego resolver el problema comercial.

Los analistas de datos también deben presentar sus hallazgos al equipo del proyecto y al patrocinador comercial. Los analistas de datos presentan sus hallazgos utilizando visualizaciones de datos, como gráficos y cuadros, para que el equipo pueda tomar decisiones e impulsar la acción.

**En general, para realizar el trabajo de manera eficaz, los analistas de datos necesitan...**

* comprender la dirección y los objetivos del negocio
* explorar el significado detrás de los números
* analizar las causas de determinados eventos basándose en las conclusiones de los datos
* presentar conocimientos utilizando un lenguaje y gráficos fáciles de entender
* contribuir a la toma de decisiones empresariales ofreciendo información
* determinar la importancia de los valores atípicos y si deben considerarse o eliminarse

¡El trabajo de un científico de datos es más científico! La responsabilidad principal de un científico de datos es **utilizar estadísticas, algoritmos de aprendizaje automático y lenguajes de programación para extraer datos relevantes para formular una hipótesis y probarla para ayudar a una empresa a tomar decisiones acertadas y predecir resultados futuros.**

Los científicos de datos colaboran con los analistas de datos para recopilar y limpiar grandes cantidades de datos. Los datos incluyen "datos oscuros" que no están estructurados, como imágenes y texto.

Los científicos de datos diseñan procesos de modelado de datos, crean algoritmos y desarrollan modelos predictivos para extraer los datos que la empresa necesita. Luego, ayudan a analizar los datos para sacar conclusiones y compartir conocimientos con el equipo.

Los científicos de datos podrían trabajar en proyectos sobre problemas globales a gran escala con puntos críticos, como el cambio climático. Los científicos de datos que conocen de un negocio en particular pueden determinar un problema a investigar. De lo contrario, los científicos de datos trabajan en estrecha colaboración con los patrocinadores comerciales para comprender los problemas y objetivos del negocio, y determinar cómo se pueden utilizar los datos para resolver los problemas y lograr esos objetivos.

**En general, para realizar el trabajo de manera eficaz, los científicos de datos necesitan…**

* identificar los problemas que ofrecen mayores oportunidades para que una empresa los resuelva
* desarrollar una hipótesis para investigar y encontrar patrones ocultos
* extraer y analizar grandes cantidades de datos
* utilizar técnicas y algoritmos de aprendizaje automático
* crear, entrenar e implementar modelos de inteligencia artificial (IA) personalizados para predecir resultados futuros
* interpretar datos para obtener información útil
* proponer soluciones y estrategias para afrontar retos empresariales

**Conclusiones clave**

* ¡La ciencia de datos es una labor en equipo! Puedes colaborar dentro de tu equipo y en diferentes programas dentro de una empresa.
* Los proyectos de ciencia de datos fomentan la colaboración entre zonas horarias y culturas de distintas partes del mundo.
* Si estás empezando y te interesa la ciencia de datos, Armen te sugiere que:
  + seas curioso y preguntes: "¿Por qué?"
  + tengas una sólida formación en estadística

**Conclusiones clave**

* Harás tu mejor trabajo cuando hagas lo que amas y ames lo que haces.
* Un día en la vida de un científico de datos puede parecer diferente cada día, dependiendo del proyecto. Siempre mirarás los números para determinar cómo cuentan una historia para que puedas obtener información y tomar decisiones valiosas.
* Si estás empezando y te interesa la ciencia de datos, DeShaun te sugiere lo siguiente:
  + Ten curiosidad por las matemáticas
  + Ábrete a los desafíos y a aprender cosas nuevas de todos los recursos en línea disponibles sobre ciencia de datos y herramientas

**Módulo 3  
Qué buscan los empleadores**

**Habilidades a desarrollar y herramientas a conocer**

**Habilidades que necesitan los analistas de datos**

Habilidades básicas

* Análisis de datos
* Ciencia de datos
* Calidad de los datos

Habilidades especializadas

* Habilidades de comunicación
* Trabajo en equipo y colaboración
* Resolución de problemas
* Investigación
* Orientado a los detalles
* Redacción

Habilidades con las herramientas

* Microsoft Excel
* SQL
* Python
* Tableau
* Microsoft Power BI

**Habilidades que necesitan los científicos de datos**

Habilidades básicas

* Análisis de datos
* Ciencia de datos
* Aprendizaje automático
* Modelos predictivos
* Inteligencia artificial (IA)

Habilidades especializadas

* Habilidades de comunicación
* Trabajo en equipo y colaboración
* Resolución de problemas
* Investigación
* Creatividad
* Redacción
* Habilidades de presentación

Habilidades con las herramientas

* SQL
* Python
* Tableau
* Herramientas adicionales y experiencia en programación, como R, Apache Hadoop, TensorFlow, Scala, Scikit-learn y más

Ten en cuenta que hay muchas habilidades que los analistas de datos y los científicos de datos tienen en **común**, como el análisis de datos, la comunicación, la resolución de problemas e incluso las herramientas.

Cada conjunto de habilidades no es una lista acumulativa. Sin embargo, las capacidades presentadas representan una selección de algunas de las principales habilidades necesarias para realizar tareas diarias, gestionar responsabilidades y tener éxito en cada puesto de trabajo respectivo.

**Herramientas y lenguajes de análisis de datos populares**

Esta tabla destaca algunas herramientas de análisis de datos y lenguajes de programación populares. Debido a que estas herramientas y lenguajes de programación se utilizan ampliamente en el análisis de datos, es importante que estés familiarizado con al menos algunos de ellos si estás buscando una carrera en el campo de datos.

| **Herramienta** | **Objetivo** |
| --- | --- |
| **SQL** | SQL, o lenguaje de consulta estructurada, es un lenguaje de programación de propósito especial para administrar datos almacenados en sistemas de gestión de bases de datos relacionales |
| **Microsoft Excel** | Una herramienta de hoja de cálculo sencilla pero potente para la recopilación y el análisis de datos |
| **Python** | Inicialmente diseñado como un lenguaje de programación orientada a objetos (OOP) para el desarrollo de software y web, y luego mejorado para la ciencia de datos |
| **R** | Un lenguaje de programación para modelado estadístico, visualización y análisis de datos |
| **SAS** | Un paquete de software estadístico ampliamente utilizado para Business Intelligence (BI), gestión de datos y análisis predictivo |
| **Microsoft Power BI** | Una solución de análisis empresarial de Microsoft que está disponible en tres versiones: Desktop, Pro y Premium |
| **Tableau** | Una herramienta de BI desarrollada para que los analistas de datos visualicen, analicen y comprendan los datos |
| **Apache Spark** | Un motor de análisis integrado para el procesamiento de big data, diseñado para desarrolladores, investigadores y científicos de datos |
| **Google Analytics** | Un servicio de análisis web que rastrea e informa sobre el tráfico de sitios web |
| **IBM Watson Studio** | Un entorno de desarrollo interactivo (IDE) que proporciona acceso a herramientas para analizar y visualizar datos, limpiar y dar forma a datos, ingerir datos en tiempo real o crear y entrenar modelos de aprendizaje automático |

**Fuente:** [*What Is Data Analytics?*, CompTIA, Inc., https://www.comptia.org/content/guides/what-is-data-analytics(se abre en una nueva pestaña)](https://www.comptia.org/content/guides/what-is-data-analytics).

**Python y R son dos lenguajes de programación muy populares utilizados en la ciencia de datos, por lo que podría ser un excelente lugar para comenzar. Python y R son de código abierto y gratuitos, lo que significa que cualquiera puede aprender a programar en estos lenguajes**

Los proyectos de ciencia de datos implican trabajar en equipo para intentar resolver problemas del mundo real que tienen un impacto la vida de las personas. Por lo tanto, es importante tener **habilidades laborales** como la comunicación, el trabajo en equipo, el pensamiento crítico, la inteligencia emocional y otras. Estas habilidades también pueden denominarse habilidades "de empleabilidad", "sociales" o "profesionales". Independientemente de la terminología, los empleadores buscan candidatos potenciales con más que solo habilidades técnicas. Los analistas de datos y los científicos de datos necesitan una **fuerte combinación de habilidades técnicas y laborales**.

* Ser **curioso** en tu naturaleza. La curiosidad es un factor importante para comprender el problema empresarial, los datos y lo que se puede hacer con ellos.
* Estar dispuesto a **asumir una posición** y tener confianza en tu posición.
* Ser **flexible** y **adaptable**. Ten en cuenta que, mientras asumes una posición, aprenderás de los datos y podrías encontrarte en el lugar equivocado. Por lo tanto, es importante saber que es un proceso de aprendizaje y que debes adaptarte a lo que descubres en los datos. Por ejemplo, un científico de datos podría necesitar modificar la hipótesis.
* Siéntete cómodo usando la **narración**. Una vez que hayas analizado los datos y determinado algunas respuestas, debes contar una historia. De lo contrario, tus conclusiones podrían permanecer ocultas. Contar grandes historias sobre insights es clave en este campo.  
  En conclusión, ¡sé curioso, decidido, flexible y un buen narrador!

**Piensa en ti mismo**

¿Eres un pensador crítico y un buen comunicador que quiere resolver problemas utilizando datos? ¿Te gusta investigar en busca de datos y "aprender sobre la marcha"? Si respondiste "sí" a estas preguntas, entonces tal vez una carrera en el análisis de datos o la ciencia de datos sea para ti. Es útil tener en cuenta tus **habilidades actuales, intereses personales y las habilidades que puedes desarrollar**.

**Módulo 4**

**Recursos útiles y primeros pasos**

**Objetivos de aprendizaje**

Después de completar este módulo, deberías poder:

* identificar recursos para obtener más información y mantenerte actualizado en el campo de la ciencia de datos

**Recursos para explorar**

Esta lección tiene como objetivo brindarte inspiración para que busques más información en el futuro.

Estos son algunos recursos que puedes explorar, marcar como favoritos y tener en cuenta si deseas explorar más sobre datos y mantenerte informado de los últimos avances en este campo. Esta es una lista seleccionada. Hay muchas organizaciones y sitios web para explorar en función de tus intereses.

* **Organizaciones profesionales a tener en cuenta**
  + [CODATA(se abre en una nueva pestaña)](https://codata.org/) es el Comité de Datos del Consejo Científico Internacional (ISC). La misión de CODATA es conectar datos y personas para avanzar en la ciencia y mejorar nuestro mundo.
  + La [Asociación de Científicos de Datos (ADaSci)(se abre en una nueva pestaña)](https://www.adasci.org/) es un organismo profesional global de profesionales en ciencia de datos y aprendizaje automático.
  + [Data Science Africa (DSA)(se abre en una nueva pestaña)](http://www.datascienceafrica.org/) es un grupo profesional de intercambio de conocimientos sin fines de lucro cuyo objetivo es reunir a investigadores líderes y profesionales de la ciencia de datos en África, y brindar capacitación sobre métodos de ciencia de datos de última generación.
  + [INFORMS(se abre en una nueva pestaña)](https://www.informs.org/) es el Instituto de Investigación de Operaciones y Ciencias de la Gestión, cuya misión es promover y fomentar la ciencia y la tecnología de la toma de decisiones para salvar vidas, ahorrar dinero y resolver problemas.
* **Publicaciones que puedes consultar**
  + La [revista de ciencia de datos CODATA(se abre en una nueva pestaña)](https://datascience.codata.org/) es una revista electrónica de acceso abierto y revisada por pares que publica artículos sobre la gestión, difusión, uso y reutilización de datos de investigación y bases de datos en todos los dominios de investigación, incluida la ciencia, la tecnología, las humanidades y las artes.
  + La [Revista Internacional de Ciencia y Análisis de Datos (JDSA)(se abre en una nueva pestaña)](https://www.springer.com/journal/41060/) reúne a líderes de opinión, investigadores, profesionales de la industria y usuarios potenciales del análisis y la ciencia de datos para desarrollar el campo, discutir nuevas tendencias y oportunidades, intercambiar ideas y prácticas, y promover colaboraciones transdisciplinarias y entre dominios.
  + [Scientific Data(se abre en una nueva pestaña)](https://www.nature.com/sdata/) es una revista de acceso abierto y revisada por pares para descripciones de conjuntos de datos e investigaciones que promueven el intercambio y la reutilización de datos científicos.
  + La [Harvard Data Science Review(se abre en una nueva pestaña)](https://hdsr.mitpress.mit.edu/) es una iniciativa de una plataforma de acceso abierto que presenta pensamiento fundamental, hitos de investigación, innovaciones educativas y aplicaciones importantes, con un énfasis principal en la reproducibilidad, replicabilidad y legibilidad.
* **Noticias y blogs para mantenerte informado**
  + [Analytics Insight(se abre en una nueva pestaña)](https://www.analyticsinsight.net/) ofrece información, últimas noticias y una revista que presenta opiniones y puntos de vista de los principales líderes y ejecutivos de la industria que comparten su trayectoria, experiencias, historias de éxito y conocimientos para hacer crecer negocios rentables.
  + [Towards Data Science(se abre en una nueva pestaña)](https://towardsdatascience.com/) es una publicación en línea en la que los autores independientes que siguen sus reglas y pautas pueden publicar su trabajo, compartir sus conocimientos y experiencia e involucrar a una amplia audiencia en [Medium(se abre en una nueva pestaña)](https://medium.com/).[(se abre en una nueva pestaña)](https://www.kdnuggets.com/)
  + [Data Science Central(se abre en una nueva pestaña)](https://www.datasciencecentral.com/) es un recurso en línea líder para profesionales de datos. Desde estadísticas y análisis hasta aprendizaje automático e IA, Data Science Central ofrece una experiencia comunitaria que incluye una rica plataforma editorial, interacción social, soporte basado en foros y la información más reciente sobre tecnología, herramientas, tendencias y carreras.
  + El [Data Science 101 Blog(se abre en una nueva pestaña)](https://ryanswanstrom.com/datascience101/) contiene una gran cantidad de información útil para personas que recién comienzan su camino para convertirse en científicos de datos.
  + [Datanami(se abre en una nueva pestaña)](https://www.datanami.com/) es un portal de noticias dedicado a proporcionar información, análisis y novedades sobre tendencias emergentes y soluciones en big data.
* **Oportunidades de aprendizaje gratuitas para desarrollar tus habilidades**

**Nota:** Necesitas registrarte para obtener una cuenta gratuita para aprovechar las siguientes oportunidades de aprendizaje en línea.

* + [Kaggle(se abre en una nueva pestaña)](https://www.kaggle.com/learn) ofrece microcursos gratuitos en línea para ayudarte a adquirir las habilidades que necesitas para realizar proyectos independientes de ciencia de datos. Kaggle también te permite desarrollar tus habilidades en ciencia de datos y aprendizaje automático compitiendo en competencias de Kaggle. Encuentra ayuda en la [documentación(se abre en una nueva pestaña)](https://www.kaggle.com/docs/competitions) o aprende sobre [Competiciones comunitarias(se abre en una nueva pestaña)](https://www.kaggle.com/c/about/community).
  + [Udemy(se abre en una nueva pestaña)](https://www.udemy.com/) ofrece un curso gratuito y breve basado en video llamado [Introducción a la ciencia de datos(se abre en una nueva pestaña)](https://www.udemy.com/course/an-introduction-to-data-science/) donde podrás aprender sobre conceptos básicos de los datos.
  + [freeCodeCamp.org(se abre en una nueva pestaña)](https://www.freecodecamp.org/) ofrece un curso basado en videos de YouTube llamado [Ciencia de datos: introducción(se abre en una nueva pestaña)](https://www.youtube.com/watch?v=ua-CiDNNj30) de Barton Poulson. El curso cubre los fundamentos de la ciencia de datos, obtención de datos, codificación, matemáticas y estadísticas. También encontrarás muchos otros tipos de oportunidades de aprendizaje gratuitas en [freeCodeCamp(se abre en una nueva pestaña)](https://www.freecodecamp.org/" \t "_blank).
  + [DataCamp(se abre en una nueva pestaña)](https://www.datacamp.com/) ofrece el curso gratuito de dos horas [Comprender la ciencia de datos(se abre en una nueva pestaña)](https://www.datacamp.com/courses/understanding-data-science) que introduce la ciencia de datos sin codificación involucrada.
  + [W3Schools(se abre en una nueva pestaña)](https://www.w3schools.com/) es el sitio para desarrolladores web más grande del mundo que ofrece una variedad de tutoriales en línea gratuitos con experiencia práctica. El sitio incluye tutoriales sobre algunos lenguajes populares de programación de ciencia de datos, como [Python(se abre en una nueva pestaña)](https://www.w3schools.com/python/), [R(se abre en una nueva pestaña)](https://www.w3schools.com/r/default.asp) y [SQL(se abre en una nueva pestaña)](https://www.w3schools.com/sql/default.asp).
  + [Codecademy(se abre en una nueva pestaña)](https://www.codecademy.com/) ofrece clases de codificación gratuitas en 12 lenguajes de programación diferentes, incluidos Python y SQL.
  + [Code Patterns(se abre en una nueva pestaña)](https://developer.ibm.com/patterns/) de IBM Developer ofrece soluciones de código gratuito a problemas que los desarrolladores enfrentan todos los días, en todas las tecnologías.

También puedes aprovechar [IBM SkillsBuild(se abre en una nueva pestaña)](https://skillsbuild.org/) para impulsar tu futuro en la tecnología con habilidades laborales, cursos, credenciales digitales, etc.

Si deseas planificar con anticipación, consulta **certificaciones**para demostrar tus competencias a potenciales empleadores. Por ejemplo, [CompTIA Data+(se abre en una nueva pestaña)](https://www.comptia.org/certifications/data" \t "_blank) es una certificación en análisis de datos para profesionales que inician su carrera profesional y tienen como objetivo desarrollar y promover la toma de decisiones comerciales basada en los datos.

* **¡Amplía tus conocimientos!**Cuanto más familiarizado estés con el análisis de datos, más oportunidades tendrás para explorar. Trata de investigar lo que te interesa, en términos de tipos de análisis, estadísticas, IA, diferentes industrias, los últimos proyectos de ciencia de datos que están teniendo impacto, etc.
* **¡Sigue aprendiendo!**Este es el comienzo de tu experiencia de aprendizaje. Puedes continuar buscando en línea oportunidades de aprendizaje adicionales y consulta las que se presentan en este curso. Investiga y determina los requisitos educativos para lo que quieres hacer.
* **¡Explora oportunidades!**Si estás buscando empleo, puedes comenzar a explorar el mercado laboral. Consulta las ofertas de empleo en línea para identificar requisitos y calificaciones comunes. Obtén una idea de qué trabajos pueden interesarte en el futuro y trabaja para cumplir con los requisitos.

**Ahora que has completado este curso, deberías poder:**

* reconocer la demanda global de analistas de datos y científicos de datos en el mercado laboral
* reconocer el futuro del campo del análisis de datos
* identificar las industrias en las que trabajan los profesionales de datos
* identificar las principales responsabilidades de un analista de datos y un científico de datos
* identificar las habilidades que necesitan los profesionales de datos
* identificar las herramientas que debes conocer al iniciarte en el rubro
* identificar recursos para obtener más información y mantenerte actualizado en el campo de la ciencia de datos
* 1

Existe una necesidad global de personas con habilidades especializadas en el campo de los datos para ayudar a las empresas a tomar decisiones basadas en datos que marquen la diferencia en este mundo digital.

* 2

Un trabajo en análisis de datos o ciencia de datos puede generar muchas oportunidades y posibilidades de crecimiento profesional.

* 3

El análisis de datos ayuda a muchos tipos de organizaciones y empresas a resolver problemas en todas las industrias, desde las finanzas y la atención médica hasta la fabricación y el comercio minorista.

* 4

La responsabilidad principal de un analista de datos es recopilar, limpiar y analizar datos utilizando los métodos y herramientas estándar de una empresa para identificar tendencias, encontrar patrones y hacer pronósticos.

* 5

La responsabilidad principal de un científico de datos es utilizar estadísticas, algoritmos de aprendizaje automático y lenguajes de programación para extraer datos relevantes para formular una hipótesis y probarla para ayudar a las empresas a tomar decisiones acertadas y predecir resultados futuros.

* 6

Es importante desarrollar habilidades básicas o técnicas, como en análisis de datos y ciencia de datos, y habilidades laborales, como comunicación y resolución de problemas.

* 7

Los analistas de datos y los científicos de datos deben conocer herramientas de análisis de datos y lenguajes de programación ampliamente utilizados, como SQL, Microsoft Excel, Python y R.

* 8

Si estás pensando en un nuevo camino profesional en el área de datos, ¡será útil que seas una persona curiosa, decidida y flexible que pueda contar una historia!