«Talento Tech»

Data Analytics

con Python

Clase 01

Buenos Aires Ciudad

Clase N° 1 | Introducción a Data Analytics

Temario:

1. **Conceptos básicos en ciencias de datos**
    - **a.** Definiciones clave
    - **b.** Procesos en ciencias de datos
    - **c.** TIC
    - **d.** Roles profesionales
2. **Entornos de trabajo**
    - **a.** Jupyter Notebooks
3. **Python**
    - **a.** Estructuras de datos básicas
    - **b.** Librerías Python para análisis de datos

Presentación de la Empresa

Imaginá que recibís una invitación para participar en el **proceso de selección de SynthData**, una startup ubicada en Buenos Aires.
¿Cuál sería el reto? Completar una pasantía de aprendizaje que pondrá a prueba todas tus habilidades y aprendizaje.
A partir de este momento un equipo de expertos te acompañará en este emocionante viaje. Clase a clase encontrarás un texto que te dará la base teórica, una serie de ejercicios cortos que te ayudará a ganar habilidades y expertise en diferentes contextos en los que analizamos datos. A lo largo de la cursada, y con tus nuevas habilidades, el equipo de Data Analytics te guiará en el desarrollo de un **proyecto integral de análisis de datos** desde cero.

Acerca de SynthData

En **Synthdata,** transformamos datos en oportunidades estratégicas. Nuestro equipo de analistas de datos está comprometido a potenciar el éxito de las empresas a través de soluciones avanzadas de análisis y visualización. Mediante un enfoque personalizado, ayudamos a nuestros clientes a desentrañar insights clave que facilitan la toma de decisiones informadas y efectivas. Nuestra misión es simplificar la complejidad de los datos para convertirla en un motor de crecimiento claro y accesible para su negocio. ¡Descubra las posibilidades infinitas que podemos ofrecerle!

Equipo SynthData

Silvia

Project Manager y Data Scientist

Luis
BI Analyst

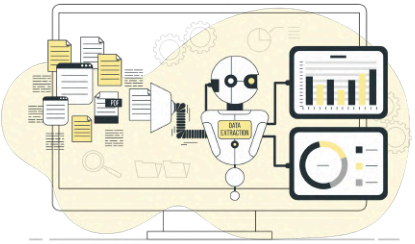
Matias
Data Analyst

Sabrina
Data Engineer

1. Conceptos básicos en ciencias de datos

Bienvenidos al curso de **Data Analytics**, una puerta de entrada al fascinante mundo de las ciencias de datos. En la era digital actual, la capacidad de analizar y extraer información valiosa de grandes volúmenes de datos se ha convertido en una habilidad esencial en diversas industrias. Este curso está diseñado para proporcionar una comprensión sólida de los conceptos básicos de la ciencia de datos, preparándote para enfrentar los desafíos y aprovechar las oportunidades que presentan los datos en tu entorno profesional.

a. Definiciones clave



La **ciencia de datos** es una disciplina que combina diversos campos como las matemáticas, la estadística, la informática y el conocimiento de dominio para extraer conocimiento e insights valiosos de los datos. A lo largo de los años, ha evolucionado con el desarrollo de nuevas tecnologías y métodos. A continuación, se definen algunos conceptos clave y tecnologías que son fundamentales en el ámbito de la ciencia de datos.

1. **Datos:** La materia prima de la ciencia de datos, que puede ser estructurada (como bases de datos) o no estructurada (como textos, imágenes o video). Los datos son la base para realizar análisis y obtener insights.
2. **Big Data:** Se refiere a conjuntos de datos tan grandes y complejos que las herramientas de procesamiento y análisis de datos tradicionales no son suficientes. Big Data se caracteriza por las "3 Vs": volumen, velocidad y variedad.
3. **Análisis de Datos:** Proceso de inspeccionar, limpiar y modelar datos con el objetivo de descubrir información útil, informando así la toma de decisiones. Incluye tanto análisis descriptivo como predictivo y prescriptivo.
4. **Modelado Predictivo:** Uso de modelos estadísticos y algoritmos de machine learning (aprendizaje automático) para predecir resultados futuros basados en datos históricos. Implica identificar patrones en los datos y aplicarlos a nuevos casos.
5. **Machine Learning (Aprendizaje Automático):** Parte de la inteligencia artificial que permite a las máquinas aprender de datos y mejorar su rendimiento en tareas específicas sin ser programadas explícitamente. Incluye algoritmos de clasificación, regresión y agrupamiento.
6. **Inteligencia Artificial (IA):** se refiere a la simulación de procesos de inteligencia humana por parte de sistemas computacionales. Esto incluye el

desarrollo de algoritmos y modelos que permiten a las máquinas realizar tareas que normalmente requieren inteligencia humana.

7. **Visualización de Datos:** Técnica que convierte datos en representaciones gráficas o visuales, facilitando la comprensión e identificación de patrones o tendencias en los datos.
8. **Minería de Datos:** Proceso de descubrir patrones significativos y relaciones en grandes conjuntos de datos utilizando técnicas de análisis estadístico, machine learning y visualización.

La **ciencia de datos** es un campo en constante evolución que integra una variedad de conceptos y tecnologías. Desde el manejo de grandes volúmenes de datos hasta la aplicación de técnicas avanzadas de análisis y visualización, las herramientas y metodologías de ciencia de datos son fundamentales para convertir datos en decisiones estratégicas. Comprender estos conceptos y tecnologías es esencial para cualquier profesional que busque adentrarse en la ciencia de datos y aprovechar el poder del análisis de datos en su organización.

En cuanto a nuestro objeto primario de estudio (los datos), es importante conocer la **pirámide de la Jerarquía del Conocimiento**, un modelo que organiza diferentes niveles de conocimiento y comprensión, desde los más básicos hasta los más complejos. Generalmente se representa en forma de pirámide, donde cada nivel superior se basa en el anterior. A continuación, se describen los niveles típicos de esta jerarquía:

Fuente: Hey, J.: The Data, Information, Knowledge, Wisdom Chaim: The Metaphorical Link

1. **Datos:** Recordemos que son hechos y cifras sin contexto. Por sí solos, no tienen significado. Ejemplo: "29ºC", "Rosario".

2. **Información:** Se refiere a datos que han sido organizados o procesados para darles significado. Ejemplo: "La temperatura en Rosario es de 29ºC".

3. **Conocimiento:** Es la comprensión e interpretación de la información. Implica la capacidad de aplicar la información en contextos específicos. Ejemplo: "Sabemos que 25ºC es una temperatura demasiado alta para Rosario en invierno".

4. **Sabiduría:** Es la capacidad de hacer juicios y tomar decisiones basadas en el conocimiento. Implica la experiencia y la reflexión. Ejemplo: "Los días de 29 grados en Rosario, es recomendable llevar agua y protector solar".

A veces se considera un nivel superior, **la inteligencia**, que implica la capacidad de aprender, adaptarse y resolver problemas complejos utilizando todos los niveles anteriores.

Esta jerarquía es útil en diversos campos, como la **educación**, la **gestión del conocimiento** y la **toma de decisiones**, ya que ayuda a entender cómo se construye el conocimiento y cómo se puede aplicar de manera efectiva.

b. Procesos en ciencias de datos

El **análisis de datos** es un proceso sistemático que permite extraer información valiosa de conjuntos de datos, a fin de tomar decisiones informadas y resolver problemas específicos. Este proceso puede variar en función del contexto y los objetivos, pero generalmente se puede dividir en varias etapas clave, cada una de las cuales desempeña un papel fundamental en la obtención de resultados significativos. A continuación, se describen los **principales procesos del análisis de datos:**

1. Definición del Problema

El primer paso en cualquier análisis de datos es **definir claramente el problema o la pregunta que se intenta responder**. Esto implica identificar los objetivos del análisis y establecer un marco específico que guiará todo el proceso. Una definición precisa permitirá focalizar los esfuerzos

de recopilación y análisis de datos.
Ejemplo: Una empresa puede querer analizar por qué ha disminuido la satisfacción del cliente en sus servicios. La pregunta clave sería: "¿Cuáles son los factores que afectan la satisfacción del cliente en forma negativa?"

2. Recolección de Datos

Una vez que se ha definido el problema, es necesario **recopilar los datos relevantes**. Esto puede incluir la recolección de datos primarios, mediante encuestas o entrevistas, así como la obtención de datos secundarios de fuentes ya existentes como bases de datos, informes u otras publicaciones.
Ejemplo: Para abordar el problema de la satisfacción del cliente, la empresa podría recopilar datos a través de encuestas a clientes y análisis de quejas registradas.

3. Preparación de Datos

La preparación de datos, también conocida como **limpieza de datos**, es una etapa crucial que implica la transformación y organización de los datos recolectados para asegurar que estén en condiciones óptimas para el análisis. Esto incluye la eliminación de duplicados, el manejo de valores faltantes y la conversión de datos en un formato adecuado.
Ejemplo: Puede que se encuentren respuestas faltantes en las encuestas. Se debe decidir si estas se imputan, se eliminan o se manejan de otra manera.

4. Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos es una fase en la que **se examinan los datos limpios para comprender su estructura, identificar patrones y buscar información relevante**. Esto incluye el uso de estadísticas descriptivas y visualizaciones para obtener una primera impresión de los datos.
Ejemplo: Utilizando gráficos de dispersión e histogramas, la empresa puede observar tendencias en los niveles de satisfacción del cliente en relación con variables como el tiempo de espera y la calidad del servicio.

5. Modelado

El modelado implica **aplicar técnicas estadísticas o algoritmos de machine learning para construir modelos que puedan predecir o explicar patrones en los datos**. Esta etapa puede incluir la selección del modelo adecuado, el ajuste de parámetros y la validación del modelo.
Ejemplo: La empresa podría construir un modelo de regresión que relacione la satisfacción del cliente con diferentes variables, tratando de predecir qué factores afectan más la satisfacción.

6. Evaluación

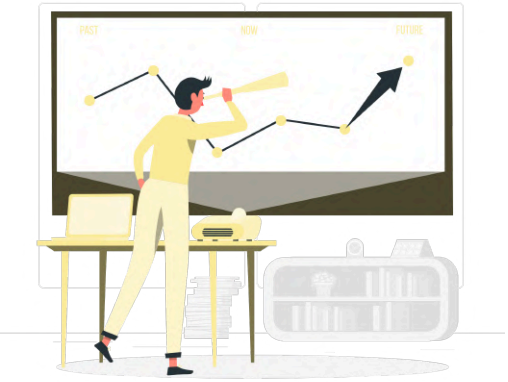
Una vez que se ha desarrollado un modelo, es esencial **evaluarlo para determinar su efectividad y precisión**. Esto incluye la revisión de métricas de rendimiento de un modelo de clasificación, que busca un equilibrio entre la precisión y la cantidad de aciertos, dependiendo de la naturaleza del análisis.
Ejemplo: La empresa evaluaría el modelo para asegurarse de que puede predecir la satisfacción del cliente con un nivel de precisión aceptable antes de implementarlo.

7. Interpretación y Comunicación de Resultados

Después de evaluar el modelo, **los hallazgos obtenidos mediante el análisis deben interpretarse de manera clara y comprensible**. La comunicación efectiva de los resultados es clave, utilizando visualizaciones, informes y presentaciones para asegurarse de que los resultados lleguen a los interesados y se tomen en cuenta en la toma de decisiones.
Ejemplo: La empresa presentaría un informe a la dirección con visualizaciones que muestren cómo el tiempo de espera afecta la satisfacción del cliente y recomendaría estrategias para mejorar la experiencia del cliente.

8. Implementación y Seguimiento

Finalmente, **los resultados del análisis deben integrarse en las decisiones y estrategias comerciales**. Además, es importante realizar un seguimiento continuo para evaluar el impacto de las decisiones tomadas a partir del análisis de datos y ajustar estrategias según sea necesario.
Ejemplo: Si se implementan cambios para reducir el tiempo de espera, la empresa deberá realizar un seguimiento de la satisfacción del cliente y repetir el análisis para medir el impacto de esas decisiones.



El **análisis de datos** es un proceso cíclico que va más allá de la simple recolección de información. Cada etapa es crucial para garantizar que los resultados sean significativos y puedan utilizarse para guiar decisiones estratégicas. Al seguir este enfoque estructurado, las organizaciones pueden maximizar el valor que obtienen de sus datos y, en última instancia, mejorar su rendimiento y competitividad en el mercado.

c. Tecnologías de la Información y la Comunicación (TIC)

Las **Tecnologías de la Información y la Comunicación (TIC)** son herramientas y sistemas que facilitan la captura, el almacenamiento, el procesamiento, la transmisión y el análisis de información. En el contexto del análisis de datos, las TIC juegan un papel crucial, ya que **permiten a las organizaciones recolectar grandes volúmenes de datos, transformarlos en información significativa y tomar decisiones informadas**. A continuación, se definen, enumeran y ejemplifican algunas de las principales TIC orientadas al análisis de datos.

1. Sistemas de Gestión de Bases de Datos (DBMS)
Los sistemas de gestión de bases de datos permiten almacenar y organizar grandes volúmenes de datos de forma estructurada. Facilitan el acceso, la recuperación y la manipulación de datos.
Ejemplo: MySQL y PostgreSQL son DBMS de código abierto ampliamente utilizados que permiten gestionar bases de datos relacionales.

2. Herramientas de Visualización de Datos
Estas herramientas permiten representar datos gráficos de manera comprensible y visual, favoreciendo la interpretación rápida de información compleja.
Ejemplo: Tableau, Looker Studio y Power BI permiten crear dashboards interactivos y visualizaciones que ayudan a los usuarios a explorar y analizar datos de manera intuitiva.

3. Lenguajes de Programación para Análisis de Datos
Existen lenguajes de programación que están especialmente diseñados o son ampliamente utilizados para realizar análisis de datos, permitiendo manipular, procesar y modelar información.
Ejemplo: Python y R son lenguajes populares para el análisis de datos. Python, con bibliotecas como Pandas y NumPy, y R, que ofrece potentes paquetes estadísticos, son herramientas altamente eficientes para científicos de datos.

4. Plataformas de Big Data
Las plataformas de Big Data permiten el almacenamiento y procesamiento de conjuntos de datos masivos que no pueden ser gestionados por sistemas tradicionales. Estas herramientas suelen utilizar arquitecturas distribuidas.
Ejemplo: Apache Hadoop y Apache Spark son frameworks que permiten procesar grandes

volúmenes de datos de manera paralela y distribuida, optimizando el análisis en tiempo real.

5. Servicios en la Nube
Las soluciones en la nube proporcionan la infraestructura y las herramientas necesarias para el análisis de datos y su almacenamiento, eliminando la necesidad de mantener servidores físicos.
Ejemplo: Amazon Web Services (AWS), Google Cloud Platform (GCP) y Microsoft Azure ofrecen una variedad de servicios que incluyen almacenamiento de datos, análisis y machine learning, accesibles desde cualquier lugar con conexión a internet.

6. Herramientas de Machine Learning y AI
Las herramientas que permiten aplicar técnicas de machine learning e inteligencia artificial para extraer patrones y realizar predicciones a partir de datos.
Ejemplo: TensorFlow y Scikit-learn son bibliotecas de Python que facilitan la implementación de algoritmos de aprendizaje automático y permiten a los analistas construir modelos predictivos.

7. Sistemas de Business Intelligence (BI)
Las herramientas de BI ayudan a las organizaciones a analizar datos históricos y actuales para la toma de decisiones empresariales. Proporcionan análisis de rendimiento, informes y métricas clave.
Ejemplo: IBM Cognos y SAP BusinessObjects permiten generar informes y análisis detallados sobre el rendimiento empresarial a partir de datos integrados de diversas fuentes.

Las TIC orientadas al análisis de datos proporcionan a las empresas las herramientas necesarias para transformar datos en información útil y estratégica. La combinación de estas tecnologías permite la creación de soluciones innovadoras que optimizan la toma de decisiones y ofrecen ventaja competitiva en un entorno empresarial cada vez más orientado a datos. Al comprender y utilizar estas herramientas, los profesionales del análisis de datos pueden maximizar el valor que los datos aportan a sus organizaciones.

d. Ramas de las ciencias de datos

Las ciencias de datos son un **campo interdisciplinario** que se centra en la extracción de conocimientos y patrones a partir de datos, integrando diversas disciplinas y técnicas. Esta combinación de conocimientos permite abordar problemas complejos y ofrecer soluciones basadas en el análisis de datos. A continuación, se describen las **principales ramas y disciplinas que forman el ámbito de las ciencias de datos:**

1. Estadística

La estadística es fundamental en las ciencias de datos. Se encarga de la recolección, análisis e interpretación de datos, proporcionando las herramientas necesarias para resumir información y extraer conclusiones. La estadística incluye métodos descriptivos, inferenciales y probabilísticos que permiten estimar parámetros, realizar pruebas de hipótesis y construir modelos predictivos.

2. Matemáticas

Las matemáticas son la base teórica que sustenta muchas técnicas utilizadas en la ciencia de datos. Las ramas más relevantes incluyen:

- **Álgebra lineal:** Fundamental para la comprensión de algoritmos de aprendizaje automático, especialmente aquellos que involucran matrices y vectores.
- **Cálculo:** Utilizado en la optimización de funciones y en algoritmos de aprendizaje automático que se basan en el ajuste de modelos.

3. Informática y Programación

La informática proporciona el marco tecnológico necesario para el manejo y procesamiento de datos. Las ciencias de datos utilizan diversos lenguajes de programación, siendo Python y R los más populares. La programación no solo permite el análisis de datos, sino también la automatización de procesos y la implementación de algoritmos de machine learning.

4. Machine Learning (Aprendizaje Automático)

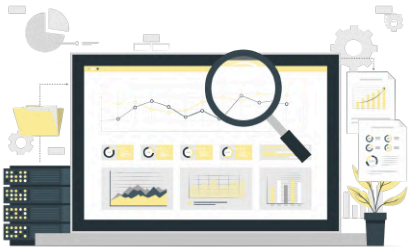
5. Big Data

6. Visualización de Datos

La visualización de datos es la disciplina que se encarga de representar gráficamente información compleja para facilitar su comprensión e interpretación. Utiliza herramientas y técnicas que permiten crear gráficos, dashboards y otros tipos de representaciones visuales, ayudando a comunicar insights de manera efectiva.

7. Minería de Datos (Data Mining)

La minería de datos se refiere al proceso de descubrir patrones ocultos y relaciones en grandes conjuntos de datos mediante técnicas estadísticas y de machine learning. Este proceso incluye la limpieza, preparación y análisis de datos para extraer información relevante que puede ser utilizada para la toma de decisiones.



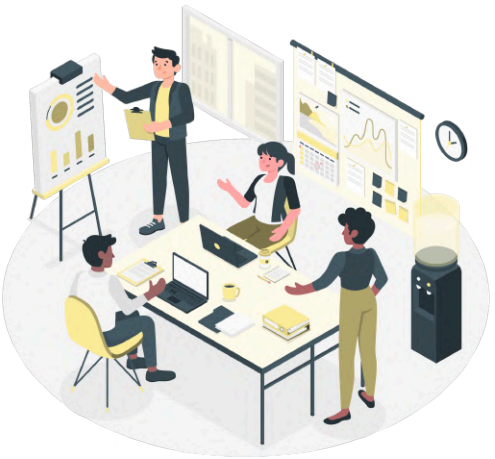
8. Ciencias Sociales y Humanas

Las ciencias sociales juegan un papel crucial en el contexto de las ciencias de datos, especialmente en la comprensión de comportamientos y tendencias humanas. La interacción y colaboración de los científicos de datos con expertos en sociología, psicología y otras disciplinas humanas es esencial para contextualizar los hallazgos y garantizar que las soluciones propuestas sean relevantes y aplicables a situaciones de la vida real.

9. Ética y Privacidad de los Datos

La ética en el manejo de datos y la privacidad es una consideración fundamental en las ciencias de datos. Esta disciplina se enfoca en el uso responsable de los datos, asegurando que se respeten los derechos de los individuos y que se utilicen prácticas justas y equitativas en el análisis y la explotación de información.

Las **ciencias de datos** integran una variedad de ramas y disciplinas, lo que las convierte en un **campo multidimensional y en contínuo crecimiento**. La colaboración entre expertos de diferentes áreas permite abordar problemáticas complejas y aprovechar al máximo el valor de los datos. Con el avance de la tecnología y la creciente disponibilidad de datos, el campo de las ciencias de datos seguirá evolucionando, abriendo nuevas oportunidades y desafíos en la búsqueda de conocimiento y soluciones basadas en datos.



e. Roles profesionales

- **Científico de Datos**: Se encarga de analizar y modelar datos para extraer información valiosa.
- **Ingeniero de Datos**: Se enfoca en la arquitectura y el flujo de datos, asegurando que los datos estén disponibles y sean accesibles.
- **Analista de Datos**: Realiza análisis descriptivos y proporciona informes sobre los datos.
- **Especialista en Machine Learning**: Desarrolla modelos predictivos y algoritmos de aprendizaje automático

Reflexión final

En un mundo impulsado por datos, la **capacidad de transformar información en conocimiento** se ha convertido en un recurso invaluable para las empresas y organizaciones. A medida que nos adentramos en la era digital, la práctica de **Data Analytics** no solo nos permite comprender el pasado, sino también facilita anticipar el futuro y mejorar nuestras decisiones estratégicas.

Es importantel recordar que el verdadero valor de los datos no reside únicamente en su cantidad, sino en el significado que extraemos de ellos. La colaboración entre diversos roles, desde analistas y científicos de datos hasta ingenieros y expertos en inteligencia artificial, es esencial para maximizar el potencial de los datos. Cada miembro desempeña un papel crítico en un proceso que va desde la recolección y preparación de datos hasta la modelización y comunicación de resultados.
Además, debemos considerar los retos éticos que acompañan el análisis de datos, como la privacidad y la seguridad. La confianza y la transparencia son claves para construir relaciones sólidas con clientes y usuarios en la gestión de su información.

Al final, **Data Analytics** es más que una disciplina técnica; es una herramienta que puede empoderar a las empresas para innovar, crecer y hacer una diferencia positiva en la sociedad. Al aprovechar el poder de los datos de manera responsable y efectiva, tenemos la oportunidad de transformar no solo negocios, sino también comunidades enteras.

2. Entornos de trabajo

a. Jupyter Notebooks

Los **Jupyter Notebooks** son aplicaciones web interactivas que permiten crear y compartir documentos que contienen código, ecuaciones, visualizaciones y texto narrativo. Son ampliamente utilizados en la comunidad de ciencia de datos, investigación y enseñanza, brindando un entorno versátil para análisis de datos, aprendizaje automático, y exploración de datos, entre otras tareas. Jupyter permite a los usuarios combinar elementos de programación con documentación de manera fluida, lo que facilita el entendimiento y la presentación de resultados.

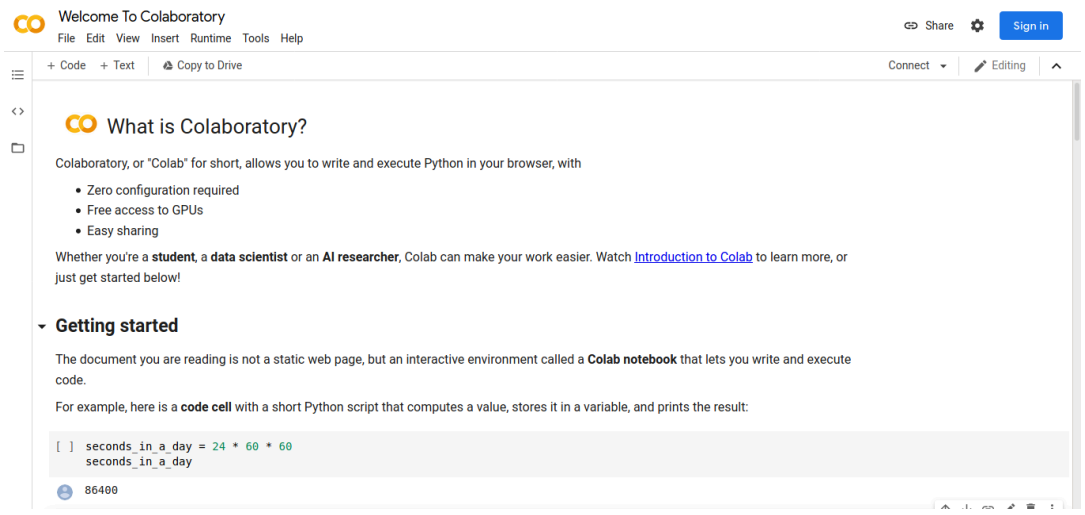
Una de las características más destacadas de Jupyter Notebooks es su **capacidad para ejecutar código en vivo**. Los usuarios pueden escribir código en múltiples lenguajes de programación, siendo Python el más popular, pero también soportando R, Julia y otros. Esto permite realizar experimentos de manera inmediata, visualizar resultados y ajustar el código en tiempo real, lo que es extremadamente útil para la prototipación y el desarrollo iterativo de proyectos. Cada bloque de código se puede ejecutar de forma independiente, lo que permite mantener un flujo de trabajo flexible y eficiente.

Además, **Jupyter Notebooks** admite la inclusión de texto formateado utilizando Markdown, lo que significa que los usuarios pueden agregar explicaciones, anotaciones y ecuaciones matemáticas de manera intuitiva. Esto ayuda a documentar el proceso de análisis y facilita la comprensión de los resultados para otros usuarios o para futuras referencias. La capacidad de combinar código y documentación en un solo archivo hace que los notebooks sean herramientas ideales para la educación y la divulgación científica, ya que pueden servir como material de aprendizaje y recursos colaborativos.

Los Jupyter Notebooks son altamente integrables y permiten la incorporación de gráficos y visualizaciones generadas por librerías populares de Python como Matplotlib, Seaborn y Plotly, enriqueciendo aún más la presentación de datos. Además, su naturaleza open-source fomenta una comunidad activa que contribuye a su desarrollo y mejora continua. En resumen, Jupyter Notebooks ofrecen un entorno interactivo y accesible para el análisis de datos y la enseñanza, simplificando el proceso de desarrollo y comunicación de proyectos de ciencia de datos y programación.

Hay **varias plataformas y entornos que facilitan el uso de Jupyter Notebooks**, cada una con características particulares que pueden ser útiles dependiendo de las necesidades del usuario. A continuación, se presentan algunas de las plataformas más destacadas:

1. **Anaconda:** Anaconda es una distribución popular de Python que incluye Jupyter Notebook como parte de su suite de herramientas. Proporciona un entorno fácil de usar para instalar y gestionar paquetes de ciencia de datos y machine learning. Anaconda Navigator es una interfaz gráfica que permite lanzar Jupyter Notebooks y gestionar entornos y paquetes sin necesidad de utilizar la línea de comandos.
2. **Google Colab:** Google Colaboratory, o simplemente Google Colab, es una plataforma en línea gratuita que permite crear y ejecutar Jupyter Notebooks en la nube. Proporciona acceso a recursos computacionales, incluyendo GPU, lo que es especialmente útil para tareas de machine learning. Además, Colab facilita la colaboración, permitiendo a múltiples usuarios trabajar en el mismo notebook en tiempo real.



3. **JupyterHub:** Es una plataforma diseñada para facilitar el uso de Jupyter Notebooks en entornos educativos y de investigación. Permite a múltiples

usuarios acceder y utilizar notebooks en un servidor compartido, lo que es ideal para aulas y proyectos colaborativos. Los administradores pueden gestionar usuarios y recursos de manera eficiente.

4. **Microsoft Azure Notebooks:** Esta es una oferta de servicio en la nube de Microsoft que permite crear y ejecutar Jupyter Notebooks sin necesidad de configuración local. Azure Notebooks proporciona una manera sencilla de almacenar y compartir notebooks en línea, lo que es útil para la colaboración y el acceso a recursos computacionales en la nube de Azure.
5. **Kaggle Kernels:** Kaggle, una plataforma de competiciones de ciencia de datos, ofrece "Kernels", que son entornos basados en Jupyter Notebooks donde los usuarios pueden escribir y ejecutar código. Esta característica hace que sea fácil trabajar con conjuntos de datos disponibles en Kaggle y participar en competiciones, además de permitir la colaboración y compartir notebooks con la comunidad.

Estas plataformas, entre otras, brindan diversas opciones para trabajar con Jupyter Notebooks, ofreciendo características como colaboración en tiempo real, acceso a recursos de computación en la nube, y entornos preconfigurados para facilitar el trabajo en proyectos de ciencia de datos y análisis.

En este curso utilizaremos Google Colab, y al finalizar, te recomendamos que utilices las herramientas de Kaggle para ganar experiencia y crear tu propio portfolio.

3. Python

a. Estructuras de datos básicas

Python ofrece varias **estructuras de datos integradas**, cada una con características específicas que permiten **almacenar, organizar y manipular información de manera eficiente**.

A continuación, se enumeran y definen las principales estructuras de datos en Python:

1. **Listas:**
   - Las listas son colecciones ordenadas y mutables que permiten almacenar múltiples elementos en un solo objeto. Los elementos pueden ser de diferentes tipos de datos, incluyendo otros objetos de lista. Las listas permiten la modificación de sus contenidos mediante operaciones como agregar, eliminar o cambiar elementos. Se definen utilizando corchetes (`[]`), por ejemplo, `mi_lista = [1, 2, 3, 'hola']`.
2. **Tuplas:**
   - Las tuplas son similares a las listas en que también son colecciones de elementos, pero a diferencia de estas, son inmutables, lo que significa que no se pueden cambiar una vez creadas. Las tuplas son útiles para almacenar datos que no deben ser alterados a lo largo de la ejecución del programa. Se definen utilizando paréntesis (`()`), por ejemplo, `mi_tupla = (1, 2, 3, 'hola')`.
3. **Diccionarios:**
   - Los diccionarios son estructuras de datos que almacenan pares de clave-valor de manera desordenada y mutable. Cada valor se puede acceder a través de su clave única, lo que permite realizar búsquedas rápidas. Son ideales para representar relaciones entre datos, como un directorio de contactos. Se definen utilizando llaves (`{}`), por ejemplo, `mi_diccionario = {'nombre': 'Juan', 'edad': 30}`.

4. **Conjuntos (Sets):**
    - Los conjuntos son colecciones no ordenadas de elementos únicos y mutables. Son útiles para realizar operaciones matemáticas como uniones, intersecciones y diferencias. Al igual que los diccionarios, los conjuntos se definen utilizando llaves (`{}`), pero sin pares de clave-valor, por ejemplo, `mi_conjunto = {1, 2, 3, 4}`.Los conjuntos mutables son los sets, y los inmutables son los frozensets.
5. **Cadenas de texto (Strings):**
    - Si bien no son una estructura de datos compuesta en el sentido tradicional, las cadenas de texto se consideran una estructura de datos fundamental en Python. Son secuencias inmutables de caracteres que se pueden utilizar para almacenar y manipular texto. Se definen utilizando comillas simples (`'`) o dobles (`"`), por ejemplo, `mi_cadena = "Hola, mundo"`.

Estas son algunas de las estructuras de datos más comunes en Python. Cada una tiene sus propias características y ventajas, lo que permite a los desarrolladores elegir la más adecuada según las necesidades específicas de su programa, optimizando así el rendimiento y la eficiencia en la gestión de datos.

Material Adicional:

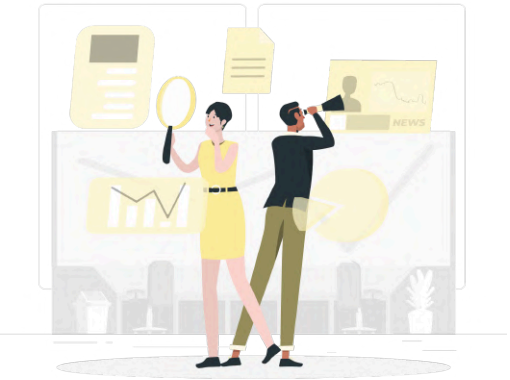
Fuente: Python.org

- Estructuras de datos en Python

Fuente: W3Schools.com

- Python Dictionaries
- Python Tuples
- Python Sets
- Python frozenset() Function

b. Librerías Python para análisis de datos



Python ofrece una amplia **variedad de bibliotecas diseñadas específicamente para el análisis de datos**, cada una con sus propias características y funcionalidades.

A continuación, se enumeran y describen algunas de las bibliotecas más populares y útiles para este propósito:

1. **Pandas:** es una de las bibliotecas más utilizadas para la manipulación y el análisis de datos en Python. Proporciona estructuras de datos flexibles como DataFrame y Series, que permiten almacenar y manipular datos en forma tabular, similar a una hoja de cálculo. Pandas ofrece herramientas eficientes para la lectura y escritura de datos, filtrado, agrupamiento, y tratamiento de datos faltantes, facilitando el análisis exploratorio de datos.
2. **NumPy:** es fundamental para el cálculo numérico en Python. Proporciona un objeto de matriz multidimensional, llamado ndarray, que permite realizar operaciones matemáticas y estadísticas de manera eficiente. NumPy es esencial para el manejo de datos numéricos, y muchas otras bibliotecas de ciencia de datos, incluida Pandas, se construyen sobre esta base. Su uso es crucial para el procesamiento de grandes volúmenes de datos y cálculos matemáticos avanzados.
3. **Matplotlib:** es una biblioteca de visualización de datos en 2D que permite crear gráficos estáticos, animados e interactivos. Es extremadamente versátil y se puede utilizar para generar una amplia gama de visualizaciones, incluyendo gráficos de líneas, histogramas, diagramas de dispersión, y mucho más. Matplotlib es frecuentemente utilizado junto con Pandas y NumPy para presentar visualmente los resultados del análisis de datos.
4. **Seaborn:** es una biblioteca de visualización de datos basada en Matplotlib que proporciona una interfaz más sencilla y atractiva para crear gráficos estadísticos. Incorpora temas estéticamente agradables y colores que mejoran la presentación de los gráficos. Seaborn permite realizar análisis de datos más complejos, facilitando la creación de gráficos de distribuciones, relaciones y categorizaciones, lo que es especialmente útil para el análisis exploratorio de datos.
5. **SciPy:** se utiliza junto con NumPy para realizar cálculos científicos y técnicos. Ofrece funcionalidades avanzadas para optimización, integración, interpolación, álgebra lineal, estadísticas y procesamiento de señales. Es

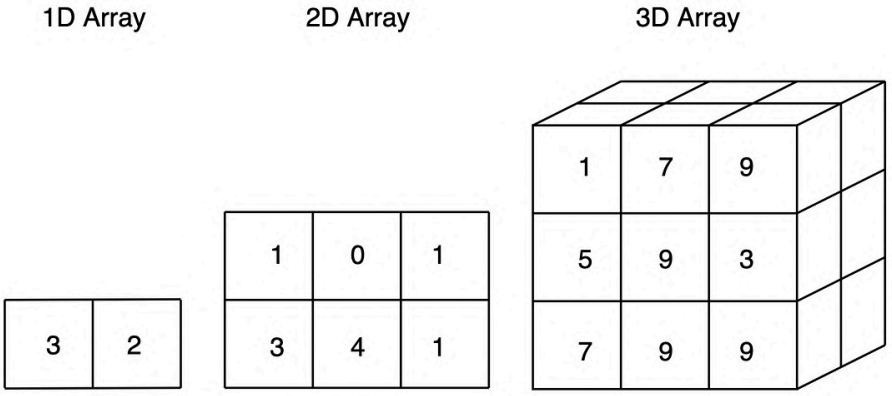
ideal para quienes necesitan realizar cálculos complejos en el contexto del análisis de datos.

6. **Scikit-learn:** es una de las bibliotecas más populares para el aprendizaje automático en Python. Proporciona herramientas para realizar tareas de clasificación, regresión, agrupamiento y reducción de dimensionalidad. Su API simple y consistente facilita la implementación de algoritmos de machine learning y es muy utilizada en proyectos de análisis de datos que requieren modelos predictivos.
7. **Statsmodels:** proporciona clases y funciones para el análisis estadístico. Ofrece herramientas para elaborar modelos estadísticos, incluyendo regresiones lineales y no lineales, autocorrelación, y pruebas de hipótesis. Statsmodels es particularmente útil para los analistas que necesitan realizar análisis estadísticos más profundos en sus datos.

Estas bibliotecas, entre muchas otras, conforman el **ecosistema de análisis de datos en Python**, ofreciendo herramientas potentes y flexibles para la **manipulación, análisis, visualización y modelado de datos**. Esta diversidad permite a los analistas y científicos de datos seleccionar las herramientas que mejor se adapten a sus necesidades, optimizando el rendimiento y la eficiencia en sus proyectos.

Además de las estructuras de datos básicas de Python, las librerías nos habilitan otras estructuras. NumPy proporciona la potente estructura de datos ndarray para trabajar con arreglos multidimensionales, mientras que Pandas ofrece las estructuras Series y DataFrame para el manejo y análisis de datos estructurados. Estas herramientas son cruciales para realizar análisis de datos y cálculos numéricos de manera eficiente en Python.

Introducción a estructuras de Datos en NumPy



1. **ndarray (N-dimensional array):** La estructura de datos principal de NumPy es el objeto ndarray, que permite almacenar y manipular datos en forma de arreglos multidimensionales. Estos arreglos pueden ser de cualquier número de dimensiones (1D, 2D, 3D, etc.) y son mucho más eficientes en términos de rendimiento y uso de memoria en comparación con listas de Python. Los elementos del ndarray son del mismo tipo de dato, lo que permite realizar operaciones matemáticas vectorizadas de manera eficiente.

Introducción a estructuras de Datos en Pandas

1. **Series:** es una colección unidimensional similar a una lista o un array de NumPy, pero con etiquetas en los índices. Cada elemento en una Series se puede acceder utilizando su índice, y permite almacenar y manipular datos de diferentes tipos (números, cadenas, objetos, etc.). Es útil para representar datos unidimensionales con etiquetas.
2. **DataFrame:** el DataFrame, que es una colección bidimensional de datos que se asemeja a una tabla o una hoja de cálculo. Un DataFrame está compuesto por varias Series, donde cada columna puede contener diferentes tipos de datos. Los DataFrames permiten realizar operaciones de filtrado, agrupamiento y agregación, además de facilitar la manipulación de datos y su análisis, gracias a su capacidad para manejar etiquetas en filas y columnas.

	year	month	day	dep_time	dep_delay	arr_time	arr_delay	carrier	tailnum	flight	origin	dest	air_time	distance	hour	minute
0	2014	1	1	1.0	96.0	235.0	70.0	AS	N508AS	145	PDX	ANC	194.0	1542	0.0	1.0
1	2014	1	1	4.0	-6.0	738.0	-23.0	US	N195UW	1830	SEA	CLT	252.0	2279	0.0	4.0
2	2014	1	1	8.0	13.0	548.0	-4.0	UA	N37422	1609	PDX	IAH	201.0	1825	0.0	8.0
3	2014	1	1	28.0	-2.0	800.0	-23.0	US	N547UW	466	PDX	CLT	251.0	2282	0.0	28.0
4	2014	1	1	34.0	44.0	325.0	43.0	AS	N762AS	121	SEA	ANC	201.0	1448	0.0	34.0

Ejercicios Prácticos

SYNTHDATA
SINTETIZA CONOCIMIENTO

Actividad 1: Primer bloque de texto en Colab

Contexto

Tu primera guía en SynthData será Silvia, la Project Manager. En su onboarding, suele pedir a las personas en pasantía que documenten sus descubrimientos de forma clara y visual. Aunque parezca simple, crear un buen bloque de texto en Colab es el primer paso para construir notebooks útiles, legibles y colaborativos.

Objetivos

Aprender a utilizar los bloques de texto de Google Colab para documentar código, explicar hallazgos y agregar referencias visuales o externas. Esta habilidad te permitirá entregar trabajos entendibles tanto para colegas técnicos como para usuarios no especializados.

Ejercicio práctico

1. Creá un nuevo bloque de texto en tu notebook de Colab.
2. Agregá un título de nivel 1 con tu nombre (usando #).
3. Escribí un breve párrafo de presentación, incluyendo formato en **negrita**, *cursiva*, y una lista con al menos 3 ítems.
4. Insertá un link externo (por ejemplo, a una página de documentación oficial de Python).
5. Desde tu Google Drive, insertá al menos una imagen relevante para tu presentación (podés ser vos, tu setup, un mapa, lo que quieras).
6. Cerrá con una cita breve o una frase que te represente usando el formato de bloque de cita (>).

¿Por qué importa esto en SynthData?

En SynthData, los notebooks no son solo espacios de código: son reportes vivos. Se comparten con clientes, con otros equipos, y se usan para dejar evidencia de decisiones técnicas. Saber cómo documentar con claridad, destacar lo importante y conectar visualmente los bloques de análisis es tan importante como saber programar. Silvia lo repite siempre: *"El mejor código del mundo no sirve si nadie puede leerlo después."*

Actividad 2: Pares, impares y primeros pasos en Python

Contexto

En esta segunda tarea, tu mentor será Luis, Analista de BI. Él te va a acompañar en los primeros pasos con código. El desafío es entender cómo interactuar con los usuarios a través de la consola, tomar datos de entrada y procesarlos usando lógica condicional.

Objetivos

Familiarizarte con los bloques de código en Colab y practicar la escritura de scripts básicos en Python. Aprender a usar `input()` y condicionales `if/else` para resolver un problema concreto y simple: detectar si un número es par o impar.

Ejercicio práctico

1. Insertá un nuevo bloque de código en tu notebook de Colab.
2. Escribí un programa que:
    a. Pida al usuario que ingrese un número entero usando `input()`
    b. Convierta ese valor a entero (`int()`)
    c. Use una estructura if/else para verificar si el número es divisible por 2
    d. Imprima un mensaje indicando si el número es par o impar
3. Ejecutá el bloque al menos dos veces, con un número par y uno impar.

No te preocupes por validar errores si se ingresa algo incorrecto. Asumimos que el usuario siempre colabora 🙂

¿Por qué importa esto en SynthData?

Aunque este ejercicio parece elemental, es la base de todas las decisiones condicionales que usamos más adelante para filtrar datos, construir dashboards, o automatizar reportes. Luis trabaja con estructuras como estas todos los días para calcular métricas condicionales y responder preguntas como: ¿cuántos usuarios activos hubo esta semana en comparación con la anterior? Dominar la lógica `if/else` es el primer paso hacia ese tipo de análisis.


Buenos Aires
aprende
Agencia de Habilidades para el Futuro
Buenos
Aires
Ciudad