

# Guía de estudio — Introducción al análisis y estructura de datos



¡Hola! Te damos la bienvenida a esta guía de estudio.

## ¿En qué consiste esta guía?

La siguiente guía de estudio tiene como objetivo profundizar los contenidos que hemos visto en clase, además de abordar temas adicionales que complementan aquellos vistos en la clase.

**¡Vamos con todo!**



## Tabla de contenidos

<b>Tipos de analítica</b>	<b>2</b>
Analítica descriptiva	2
Analítica diagnóstica	2
Analítica predictiva	3
Analítica prescriptiva	3
<b>Problemas y analíticas</b>	<b>4</b>
<b>Más metodologías</b>	<b>6</b>
KDD	6
SEMMA	7



**¡Comencemos!**

## Tipos de analítica

Los diferentes tipos de análisis de datos se relacionan con diferentes problemas de negocios de la siguiente manera:

### Analítica descriptiva

Es una forma de análisis de datos que se enfoca en describir y resumir los datos existentes. Este tipo de análisis se emplea para conseguir una comprensión general de los datos, identificar patrones y tendencias, y proporcionar información sobre la distribución de los datos.

Se basa en técnicas estadísticas simples, como **medias, desviaciones estándar, histogramas y gráficos de dispersión**. También se pueden usar técnicas más avanzadas, como la **minería de datos** y la **inteligencia artificial**, para explorar y visualizar los datos de manera más detallada.

**El análisis descriptivo es importante porque proporciona una base para la toma de decisiones** y ayuda a entender el contexto en el que se encuentran los datos, siendo un punto de partida para la investigación y el análisis posterior, y puede ser el primer paso en la resolución de problemas complejos. A fin de cuentas, toda estrategia de mejora, todo plan comienza por comprender en forma acabada el estado actual de las cosas.

### Analítica diagnóstica

Es una forma de análisis de datos que se centra en determinar las causas raíz de un problema o un comportamiento anómalo, utilizando técnicas de **minería de datos, estadísticas y aprendizaje automático** para explorar los datos y encontrar patrones o relaciones que puedan explicar el comportamiento observado.

El objetivo de la analítica diagnóstica es responder preguntas específicas, como: ¿Por qué un producto no se está vendiendo tan bien como se esperaba? ¿Por qué una aplicación web está experimentando una baja tasa de retención de usuarios? ¿Qué está causando un aumento en los costos de producción?

Se centra en identificar la causa subyacente de un fenómeno, mientras que la descriptiva se centra en describir el fenómeno en sí mismo. Naturalmente, la descripción detallada del estado actual de las cosas es la que **nos permite detectar comportamientos especiales o anómalos, de los cuales nos interesará conocer la causa**. La analítica diagnóstica es, a su vez, un paso previo necesario antes de aplicar la analítica predictiva o la prescriptiva, ya que

es importante comprender la raíz de un problema antes de tratar de predecirlo o solucionarlo.

## Analítica predictiva

Es una rama de la analítica de datos que se centra en utilizar datos y **algoritmos** para hacer predicciones sobre eventos futuros, empleando modelos matemáticos y técnicas de aprendizaje automático para analizar datos históricos y hacer predicciones sobre resultados futuros. Estas predicciones pueden ser usadas para tomar decisiones informadas en una amplia variedad de contextos, desde la gestión del inventario en una empresa hasta la identificación de tendencias en el mercado. Por tanto, es una herramienta valiosa para cualquier organización que desee mejorar su capacidad para tomar decisiones basadas en datos.

## Analítica prescriptiva

La analítica prescriptiva es una rama avanzada de la analítica de datos que se enfoca en identificar acciones futuras óptimas a partir de la información y los datos disponibles. Es un enfoque predictivo que no solo identifica lo que podría suceder en el futuro, sino que también recomienda acciones específicas que se pueden tomar para influir en los resultados futuros.

La analítica prescriptiva se basa en técnicas avanzadas de aprendizaje automático, inteligencia artificial y **modelos de simulación** para predecir la probabilidad de distintos resultados futuros y recomendar acciones para optimizar los resultados. Esto permite a las organizaciones tomar decisiones más informadas y estratégicas en tiempo real.

Además, se utiliza en una amplia variedad de industrias, incluyendo finanzas, marketing, atención médica y logística, entre otras. Por ejemplo, en la industria financiera, se pueden emplear modelos de análisis prescriptivo para identificar oportunidades de inversión y recomendar acciones específicas para maximizar los retornos. En la atención médica, se pueden usar modelos prescriptivos para predecir y prevenir las complicaciones de salud y recomendar acciones para mejorar la atención del paciente.



En resumen, los diferentes tipos de analítica se emplean para abordar diferentes problemas de negocios, desde entender la situación actual hasta tomar decisiones estratégicas para mejorar el desempeño y los resultados:

1. **Analítica descriptiva:** se utiliza para describir los patrones y tendencias existentes en los datos, lo que ayuda a entender mejor la situación actual del negocio.

2. **Analítica diagnóstica:** se emplea para identificar las causas subyacentes de los problemas de negocios, lo que ayuda a tomar decisiones informadas para resolver los problemas.
3. **Analítica predictiva:** se usa para predecir el comportamiento futuro de los datos, lo que ayuda a tomar decisiones estratégicas para mejorar el desempeño del negocio.
4. **Analítica prescriptiva:** se utiliza para identificar las acciones más efectivas a tomar en una situación dada, lo que ayuda a optimizar los resultados del negocio.

## Problemas y analíticas

A continuación, veremos que estos diferentes enfoques se aplican en problemas de negocios, según lo que queramos obtener y las preguntas que se nos planteen.



**Ejemplo:** Una empresa de comercio electrónico quiere analizar los ingresos generados por sus ventas en los últimos 12 meses. Para ello, contrata los servicios de un data Analyst que aplica los siguientes pasos.

1. **Selección de datos:** la empresa recopila los datos de ingresos generados por las ventas de los últimos 12 meses. En esta recopilación se busca que tengan el mayor nivel de detalle (caracterización de los compradores, montos máximos y mínimos de compra, tiempo que pasan los compradores en la tienda, etc.)
2. **Análisis estadístico:** la empresa emplea una hoja de cálculo, como Microsoft Excel o Google Sheets, para calcular la media de ingresos generados por las ventas, de los tiempos, de los montos diarios, etc. Del mismo modo, se calculan otros indicadores de tendencia central como la mediana o la moda, indicadores de posición como los quintiles o cuartiles y medidas de dispersión como la desviación estándar. Existen otras herramientas de análisis estadístico que pueden aplicarse y que verán en el desarrollo de esta carrera.
3. **Presentación de resultados:** El analista puede presentar un primer informe o **reporte** de la situación actual.
4. **Interpretación de resultados:** el reporte presentado permite relacionar los datos y extraer conclusiones. Por ejemplo, se puede observar que un producto es comprado cada mes por un número similar de clientes, en cantidades que resultan convenientes para la empresa. Sin embargo, se pueden observar problemas: algunos productos se venden poco, existe mucho abandono de carritos de compra antes de pagar, etc. Naturalmente, queremos averiguar por qué se produce cada situación y para ello aplicamos un enfoque **diagnóstico**, que nos permite diseñar soluciones de manera más adecuada. Así podemos, por ejemplo:

- Identificar **correlaciones** entre datos, que nos permitan inferir el origen de los problemas presentados, así como las relaciones que sería conveniente mantener.
  - Analizar los datos con mayor profundidad, cruzando una mayor cantidad de ellos. Así podremos realizar **segmentaciones** específicas para diseñar soluciones más adecuadas en cada caso. Quizás el comportamiento de los clientes es sustancialmente diferente según el precio del artículo, la posibilidad de pagarlo o no en cuotas, la edad del cliente, etc., y todos estos datos en conjunto generan separaciones que sea conveniente atender para diseñar estrategias específicas.
5. **Análisis predictivo:** una vez identificadas las variables más importantes, y los posibles problemas y correlaciones, es posible iniciar el análisis predictivo que nos permita anticiparnos a los escenarios futuros, para preverlos y también poder mejorar los estados actuales. Para esto se suelen aplicar los siguientes pasos:
- **Selección de variables:** se seleccionan las variables más relevantes que se utilizarán para hacer predicciones. Por ejemplo, se podrían seleccionar las variables que hayan tenido un impacto más significativo en las ventas en el pasado.
  - **Creación del modelo:** en esta etapa, se usan algoritmos de aprendizaje automático para crear un modelo predictivo basado en los datos históricos. Por ejemplo, se podría usar un modelo de regresión lineal para predecir las ventas futuras. Existen otras diferentes estrategias que aprenderás a lo largo de esta carrera.
  - **Validación del modelo:** es importante validar el modelo para asegurarse de que sea preciso, por medio de pruebas de simulación, validación cruzada, etc.

Para realizar este análisis, suele utilizarse el **A/B Test**, en el que se separan dos grupos para probar el efecto de un cambio en una variable. Así, por ejemplo, se prueba en un segmento de clientes expresar un descuento en monto en dinero ("antes, \$2000, ahora, \$1000") como se ha hecho hasta ahora, o cambiar la forma para otro grupo ("%50 de descuento"). Es imprescindible mantener un grupo de control en el que no se realicen cambios, a fin de poder aislar de manera efectiva la variable estudiada.

6. **Analítica prescriptiva:** Es el momento de hacer las recomendaciones a la empresa: ¿Qué estrategias pueden ser más efectivas? ¿Cómo mantener un esquema que ha dado resultado? Dependiendo de la profundidad del análisis, es posible incluso anticipar situaciones que actualmente no sean problemáticas, pero que podrían serlo en el futuro si cambian algunas variables.

**Reflexiona:**

- ¿Qué otros ejemplos de situaciones se te ocurren, en los que podrían aplicarse enfoques similares para el análisis?
- ¿Qué herramientas de las mencionadas te llamó más la atención, o te interesaría aprender más adelante?



## Más metodologías

### KDD

La metodología KDD (Knowledge Discovery in Databases) es un proceso sistemático de extracción de información útil y valiosa a partir de grandes volúmenes de datos, que se compone de varios pasos que van desde la limpieza y preparación de los mismos hasta la identificación de patrones y tendencias útiles.

La metodología KDD es ampliamente utilizada en la minería de datos y en la inteligencia artificial y se aplica a una amplia variedad de contextos como el análisis de mercado, la identificación de tendencias en el comportamiento de los clientes, la detección de fraude y la toma de decisiones en tiempo real.

También, es un proceso sistemático para extraer conocimiento o información útil de grandes cantidades de datos. Aquí hay un ejemplo de cómo podría ser aplicada la metodología KDD en un contexto de negocios:

1. **Selección de datos:** se recopilan y seleccionan los datos relevantes para el problema de negocio específico que se está tratando de resolver.
2. **Preprocesamiento de datos:** se lleva a cabo una limpieza, integración y normalización de los datos recopilados para prepararlos para su uso.
3. **Selección de características:** Se identifican las características fundamentales de los datos, para luego ser empleadas como entrada para los algoritmos de minería de datos.
4. **Aplicación de técnicas de minería de datos:** se utilizan técnicas de minería de datos como clustering, regresión, redes neuronales, etc. para analizar los datos y extraer patrones y relaciones interesantes (muchas de estas técnicas las abordaremos más adelante en esta carrera).

5. **Interpretación y evaluación:** se interpretan los resultados y se evalúan para determinar su relevancia y utilidad en el contexto del problema de negocio.
6. **Presentación de resultados:** se presentan los resultados a los tomadores de decisiones relevantes en un formato claro y accesible.

En este ejemplo, una empresa podría aplicar la metodología KDD para analizar grandes cantidades de datos de ventas para identificar patrones y tendencias en los comportamientos de compra de los clientes. Los resultados podrían ser usados para mejorar la estrategia de marketing y aumentar las ventas.

## SEMMA

SEMMA es un acrónimo de Sample, Explore, Modify, Model y Assess (Muestreo, Exploración, Modificación, Modelado y Evaluación), que hace referencia a una metodología para el análisis de datos que se utiliza en el campo de la minería de datos. SEMMA se compone de las siguientes etapas:

- a. **Sample:** muestreo de los datos.
- b. **Explore:** exploración de los datos para obtener una comprensión más profunda de ellos.
- c. **Modify:** modificación de los datos para hacerlos más útiles para el análisis.
- d. **Model:** construcción de modelos de datos que describan los patrones y tendencias en los datos.
- e. **Assess:** evaluación de los modelos para determinar su validez y utilidad.

La metodología SEMMA es un enfoque iterativo y flexible para el análisis de datos que permite a los analistas experimentar y explorar los datos antes de comenzar a construir modelos. Esto ayuda a asegurar que los modelos sean precisos y significativos.

Aquí hay un ejemplo de cómo se puede aplicar la metodología SEMMA a un problema real:

- a. **Muestreo:** se seleccionan una porción representativa de los datos, teniendo en cuenta las variables y las relaciones que se esperan analizar.
- b. **Exploración:** se utilizan gráficos y estadísticas para identificar patrones y tendencias en los datos.
- c. **Modificación:** se realiza una limpieza y transformación de los datos para prepararlos para el análisis.
- d. **Modelado:** se construyen modelos predictivos que describan los patrones y tendencias identificadas en los datos.
- e. **Evaluación:** se evalúan los modelos para determinar su precisión y eficacia. Si es necesario, se hacen ajustes y se repiten los pasos de modelado y evaluación.



**Ejemplo:** Una compañía de servicios financieros quiere utilizar la metodología SEMMA para predecir la probabilidad de que los clientes cancelen su servicio en el futuro.

- Primero, recopilan una muestra de datos de los clientes, incluyendo información sobre su historial de pago, la duración del servicio, la ubicación geográfica, etc.
- Luego, exploran los datos para identificar tendencias y patrones.
- A continuación, limpian y transforman los datos para prepararlos para el análisis.
- Construyen un modelo predictivo empleando técnicas de aprendizaje automático, y evalúan el modelo para determinar su precisión y eficacia.
- Finalmente, pueden usar el modelo para predecir la probabilidad de que los clientes cancelen su servicio y tomar medidas preventivas para minimizar la cancelación.