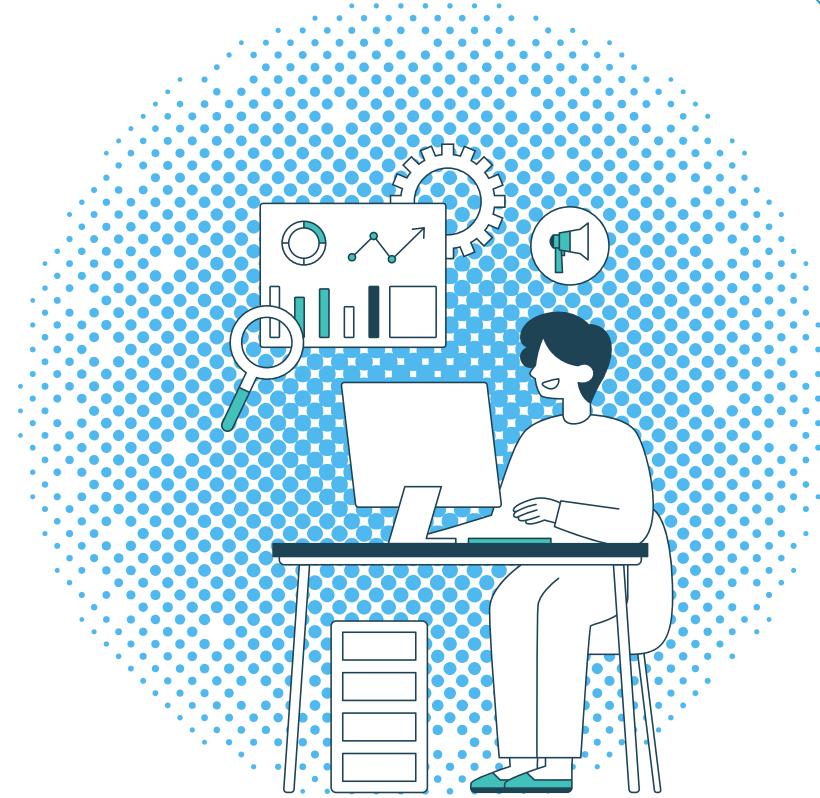


Análisis de datos

Carrera de Especialización en
Inteligencia Artificial



Programa de la materia

1 Introducción al análisis de datos

Flujo de trabajo típico. Aplicaciones. Configuración del entorno. Introducción a las principales bibliotecas.

2 Análisis exploratorio de datos

EDA, estadística descriptiva y visualización.

3 Preprocesamiento de datos y Feature Engineering

Tipos de variables. Codificación. Manejo de datos faltantes. Creación de nuevos features.

4 Pruebas estadísticas y validación de hipótesis

Pruebas de normalidad y de correlación. ANOVA y test de dependencia.

5 Reducción de dimensionalidad

Métodos para reducir la dimensionalidad y técnicas para selección de features.

6 Taller práctico

Análisis de datos completo de un dataset. Buenas prácticas. Discusión grupal.

7 Presentación de trabajos finales

Exposición por grupos y devolución de los docentes.

8 Automatización del análisis de datos

Herramientas para automatizar el EDA y otras tareas del flujo de trabajo.

Outliers

- Los **outliers** son puntos que difieren significativamente de otras observaciones en un dataset.
- Identificar outliers es crucial porque:
 - Indican errores o anomalías en los datos.
 - resaltar hallazgos inusuales pero significativos.
 - Afectar los modelos estadísticos o predictivos y llevar a interpretaciones o predicciones incorrectas.
- Si no se manejan adecuadamente, los outliers pueden sesgar el análisis y conducir a conclusiones erradas.

Técnicas para detectar outliers

- Análisis visual:
 - Boxplots
 - Scatterplots
 - Histogramas
- Métodos estadísticos:
 - datos menores a $(Q3 - 1.5 \cdot IQR)$ o mayores a $(Q3 + 1.5 \cdot IQR)$
 - frente a una distribución normal: datos menores a $(\text{media} + 3 \cdot STD)$ o mayores a $(\text{media} - 3 \cdot STD)$

Estrategias para manejar outliers

- **Eliminación.**
 - Si no son representativos o se los considera errores.
- **Imputación.**
 - Reemplazar por un valor acorde (media, mediana, percentil x)
 - Imputar a partir de un modelo. KNN por ejemplo.
- **Transformación.** Para minimizar el impacto.
 - Logarítmica (o raíz cuadrada)
- **Segmentación.**
 - Agrupar los outliers en una categoría separada (variables categóricas)
 - Analizar si se trata de errores o son valores legítimos!

Ejemplo práctico 1



Cardinalidad (v. categóricas)

- Es el número de valores únicos que puede tomar una variable en un conjunto de datos.

Baja

La mayoría de los
valores se repiten

↓
Estado civil

Día de la semana

Alta

Muchos valores
únicos

↓
Nombres de clientes

Direcciones de
correo electrónico

Extrema

Cada valor es único en la
mayoría de los casos.

↓
Números de serie

Códigos de transacciones

Codificación

- Consiste en transformar **variables categóricas** en **números** para que puedan ser utilizadas por algoritmos de ML.
- Estrategias de codificación:
 - One-hot encoding
 - Ordinal encoding
 - Target encoding
 - Frequency Encoding
 - Hashing Encoding
 - Cyclic Encoding
- La elección de la estrategia de codificación depende fuertemente de la cardinalidad de los datos.

Ejemplo práctico 2



Estrategias de codificación

Técnica	Consiste en ...	Ideal para...
One-Hot Encoding	crear una columna binaria (0/1) para cada categoría.	Baja cardinalidad
Ordinal Encoding	asignar un número entero secuencial a cada categoría.	Variables con un orden lógico
Target Encoding	reemplazar cada categoría con la media (u otra estadística) de la variable objetivo.	Alta cardinalidad
Frequency Encoding	sustituir cada categoría por su frecuencia de aparición.	Muchas categorías poco frecuentes
Hashing Encoding	aplicar una función hash para mapear categorías en un número fijo de dimensiones.	Cardinalidad extrema
Cyclic Encoding	transformar variables cíclicas en coordenadas en un círculo mediante funciones seno y coseno.	Variables con patrón cíclico

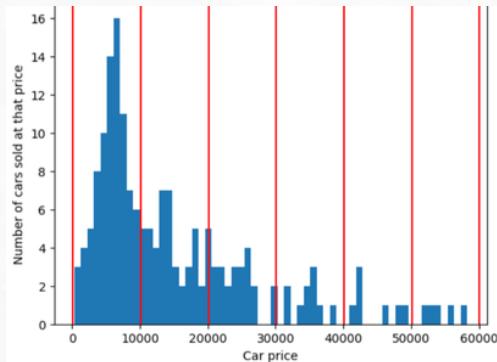
Estrategias de codificación (cont.)

Técnica	Ventajas	Desventajas
One-Hot Encoding	Simple y efectiva	Escala mal con alta cardinalidad
Ordinal Encoding	Conserva el orden natural	Puede inducir relaciones falsas
Target Encoding	Compacto y efectivo	Riesgo de sobreajuste
Por frecuencia	Útil para reducir categorías raras	Puede perder información
Hashing Encoding	Escalable y eficiente	Puede provocar colisiones
Codificación cíclica	Conserva continuidad temporal	No se recomienda en datos sin periodicidad clara

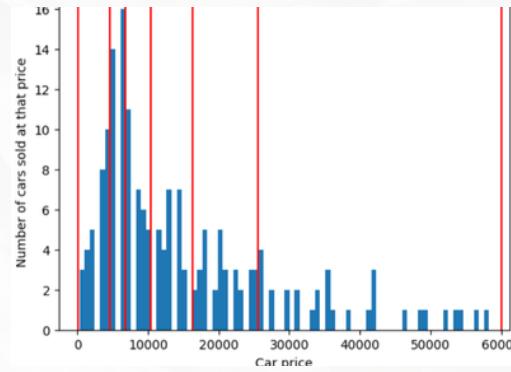
Discretización

- La discretización convierte **variables continuas** en variables discretas o **categóricas**.
- Se usa para mejorar la interpretación o el rendimiento de un modelo.

Métodos de discretización



- Por intervalos de igual rango de valores.



- Por puntos de corte (cuantiles)

Ejemplo práctico 3

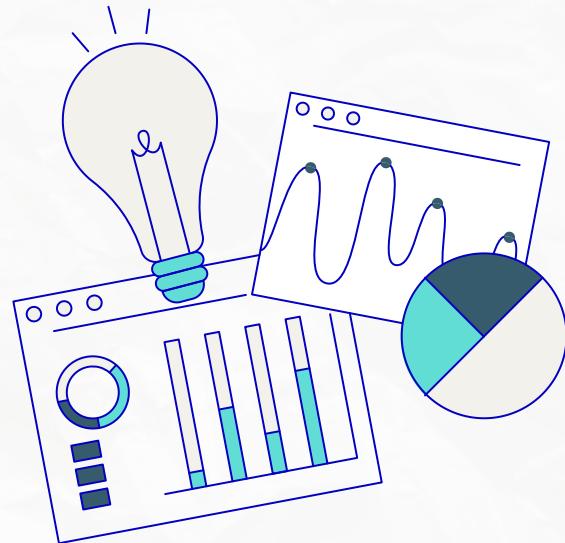


Métodos de discretización

Método	Descripción	Pros y contras
Intervalos iguales	Rangos de igual amplitud.	Simple. No sirve si los datos son asimétricos.
por puntos de corte	Basados en la distribución de los datos. Cada intervalo tiene el mismo número de observaciones.	Funciona bien para datos asimétricos.
K-means	Agrupa los datos en clusters con el algoritmo k-means	Captura agrupamientos naturales. Para datasets grandes requiere mucho cómputo.
Ancho fijo	Definidos por el usuario basados en el conocimiento del dominio.	Control total, pero requiere conocimiento de los datos y su distribución.

Desbalance

- Se produce cuando hay categorías que no están representadas de manera proporcional en el dataset.
- Es muy común en datos reales. Normalmente, hay una clase más frecuente que las otras (clase mayoritaria).
- Problema: los modelos de ML tienden a aprender mejor de las clases más fuertes y esto conduce a un mal desempeño para las clases menos frecuentes.
- Causas: bias al momento de recolectar los datos o, más frecuentemente, debido a la naturaleza del dominio en cuestión.



Técnicas para detectar desbalance

Técnica	Explicación
Análisis de la distribución	Análisis visual (gráfico de barras, pie chart) o Cálculo de las proporciones de cada clase
Matriz de confusión	Según cómo se clasificaron las instancias puede revelarse un desbalanceo de clases.
Índice de Gini $G = 1 - \sum p_i^2$	Mide la pureza de los datos. Indirectamente puede reflejar si hay un desbalance entre clases. ~0.5 indica un dataset balanceado.
Entropía de Shannon $H = - \sum_{i=1}^n p_i \log_2(p_i)$	Mide la incertidumbre de los datos. Mientras más alta la entropía, mayor la incertidumbre, más balanceados los datos.

Técnicas para mitigar el desbalance

- **Oversampling** (sobremuestreo). Implica duplicar o recrear en forma sintética las muestras de la clase minoritaria: Ej. SMOTE (*Synthetic Minority Over-sampling Technique*) o IA generativa.
- **Undersampling** (submuestreo). Implica remover muestras de la clase mayoritaria.
- **Modificación de la función de pérdida**. Implica ponderar o ajustar los pesos de las clases (*Weighted Loss Function*).
- **Técnicas de ensamble**. Algoritmos como Random Forest y XGBoost tienen mecanismos internos para manejar el desbalance.



Ejemplo práctico 4



Normalización y estandarización

- La normalización y la estandarización son dos técnicas para preparar las variables con el objetivo de mejorar el rendimiento de los modelos.
- **Normalización.** Se escalan las variables numéricas para que sus valores estén dentro de un rango específico, generalmente entre 0 y 1.
 - Útil cuando se usa un algoritmo que depende de la distancia (ej. KNN)
- **Estandarización (o Z-score).** Se transforman los datos de modo que tengan una media igual a 0 y una desviación estándar igual a 1. Adecuada cuando los datos ya tienen una distribución normal.
 - Útil para algoritmos que asumen distribución normal (Regresión lineal, SVM)

$$X_{\text{normalizado}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

$$X_{\text{estandarizado}} = \frac{X - \mu}{\sigma}$$

Ejemplo práctico 5



Programa de la materia

1 Introducción al análisis de datos

Flujo de trabajo típico. Aplicaciones. Configuración del entorno. Introducción a las principales bibliotecas.

2 Análisis exploratorio de datos

EDA, estadística descriptiva y visualización.

3 Preprocesamiento de datos y Feature Engineering

Tipos de variables. Codificación. Manejo de datos faltantes. Creación de nuevos features.

4 Pruebas estadísticas y validación de hipótesis

Pruebas de normalidad y de correlación. ANOVA y test de dependencia.

5 Reducción de dimensionalidad

Métodos para reducir la dimensionalidad y técnicas para selección de features.

6 Taller práctico

Análisis de datos completo de un dataset. Buenas prácticas. Discusión grupal.

7 Presentación de trabajos finales

Exposición por grupos y devolución de los docentes.

8 Automatización del análisis de datos

Herramientas para automatizar el EDA y otras tareas del flujo de trabajo.

Estadística inferencial

- Utiliza muestras de los datos para sacar conclusiones o "inferir" algo sobre la población entera.



- Estadística descriptiva vs inferencial:
 - Descriptiva: "¿Qué es?" (resumen de los datos que tengo).
 - Inferencial: "¿Qué podrá ser?" (adivinar algo sobre el panorama completo).
- ¿Cómo se relaciona con AI/ML?
 - Pre-training: ¿Qué cosas importan?
 - Post-training: ¿Se puede confiar en el modelo?

¿Por qué es importante?

- Asegura que los datos son representativos.
- Ayuda a detectar biases y anomalías.
- Apoya el proceso de selección de features y validación de hipótesis antes de entrenar modelos.



Antes de AI/ML: informar

- Validar supuestos.
- Identificar features o relaciones importantes.



Ejemplos

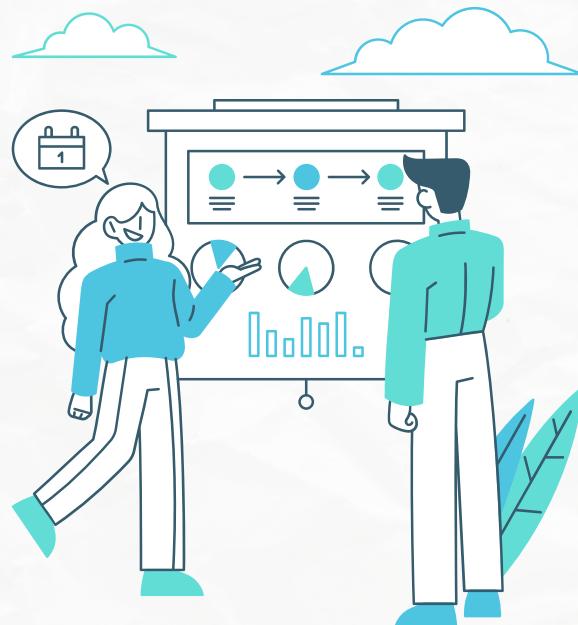
- Pruebas de linealidad: verificar si las relaciones entre variables son lineales (Por ej., para regresión).
- Estimar valores o predecir tendencias.

Después de AI/ML: validar

- Tests de significancia: verifica que las predicciones del modelo no son casualidad (Ej., p-valores para importancia de features).
- Intervalos de confianza: estimar la incertidumbre de las predicciones (Ej., "IC de 95% para la métrica accuracy").
- Tests de hipótesis: entender si un modelo funciona mejor que otro (Ej., aplicar un t-test sobre las tasas de error).
- Ejemplo: "La métrica accuracy de este modelo es suficientemente mejor que la baseline?"



Tests estadísticos inferenciales



Son tests de hipótesis, que prueban si un efecto observado es estadísticamente significativo.

Algunos de los más utilizados, son:

- Tests t
- ANOVA
- Chi-cuadrado
- Correlación

¿Cómo funcionan los tests de hipótesis?

- 1) Plantear la hipótesis: definir H_0 (la hipótesis nula o lo contrario a lo que sospecho) y la alternativa, H_1 (lo que sospecho).
- 2) Seleccionar un nivel de significancia (umbral): comúnmente se utiliza 0,05 (muy improbable, 5% o menos)
- 3) Calcular el test estadístico.
- 4) Determinar el p-valor (¿qué tan probable es ver otra cosa en la muestra si H_0 es cierta?): si $p < \alpha$, rechazo H_0 (mi sospecha se considera cierta)
- 5) Sacar una conclusión: interpretar los resultados en el contexto del problema y los datos.



Tests-t (T-Student)

- Test T de 1 muestra: compara la media de un grupo contra un valor conocido o media de referencia de la población.
 - Ej., un fabricante de chocolates quiere saber el peso promedio de una muestra se aleja del valor publicado.
- Test T independiente: compara medias de dos grupos independientes.
 - Ej., queremos comparar la efectividad de dos medicamentos distintos para la presión y los probamos en dos grupos de personas diferentes.
- Test para muestras relacionadas: compara la media del mismo grupo antes y después de un evento.
 - Ej., queremos saber queremos saber si una dieta es efectiva. Para eso, pesamos al mismo grupo de personas antes y después de la dieta.

Test Chi-cuadrado

- Sirve para averiguar la dependencia o independencia entre dos variables categóricas.
- Para el cálculo del p-valor se utiliza una tabla de contingencia.

		Práctica deportiva		
Sensación de bienestar			Total	
	Sí	No		
Sí	20	25	45	
No	10	45	55	
Total	30	70	100	

Frecuencias esperadas

$(30 \times 45) / 100 = 13,5$	$(70 \times 45) / 100 = 31,5$
$(30 \times 55) / 100 = 16,5$	$(70 \times 55) / 100 = 38,5$

Test ANOVA (análisis de varianza)

- Es una extensión del test t pero para más de 2 grupos.
¿Cuánto de la dispersión en una variable puede ser explicada si la dividimos en grupos?
- Test de 1 vía: se utiliza para analizar los efectos de una variable categórica independiente (3+ categorías) sobre una variable continua dependiente.
- Test de 2 vías: para probar el efecto de dos variables categóricas independientes (factores) sobre una variable numérica continua.
 - Ej., analizar si el género y el nivel de educación tienen un efecto en el salario. Se analizan 3 pares de hipótesis!



Correlación

- Mide la “fuerza” y la “dirección” de una relación entre dos variables.
- El resultado está entre -1 y 1:
 - 1: están muy relacionadas.
 - -1: están muy relacionadas pero cuando una crece, la otra decrece.
- Pearson: asume que la distribución de los datos es **normal** y la relación entre las variables es **lineal**.
- Spearman y Kendall se basan en los rangos de los datos y solo piden que la relación entre las variables sea **monótona**. Funcionan mejor con más datos.



Casos de uso

Test	Tipos de variables	¿Qué testea?	Supuestos/ Consideraciones
Chi-cuadrado	Dos variables categóricas	Asociación entre variables	<ul style="list-style-type: none"> Cantidad de muestras >20 Cantidades >5 en las celdas de la tabla cont.
Tests t	Continua o categórica (2 categorías)	Media vs un valor de referencia, entre grupos o muestras relacionadas	<ul style="list-style-type: none"> Normalidad Observaciones independientes
ANOVA	<ul style="list-style-type: none"> 1 vía: categórica 3+ grupos y una continua 2 vías: 2 categóricas con 2+ grupos cada una y una continua 	<ul style="list-style-type: none"> 1 vía: diferencias entre las medias de 3+ grupos 2 vías: efecto de 2 factores y de su interacción 	<ul style="list-style-type: none"> Normalidad Homocedasticidad de la varianza (las varianzas de los grupos son iguales) Las mediciones de un grupo no dependen de las de otro grupo La variable dependiente es algo medible (continua)
Correlación de Pearson	2 variables continuas	Relación lineal	<ul style="list-style-type: none"> Ambas variables tienen distribución Normal Tienen una relación lineal No hay presencia de valores extremos
Correlación de Spearman	2 continuas u ordinales	Relación monótona	<ul style="list-style-type: none"> Monótona, no necesariamente lineal
Correlación de Kendall	2 continuas u ordinales	Relación monótona	<ul style="list-style-type: none"> Monótona, no necesariamente lineal Funciona mejor cuando hay "empates" (datos que se repiten) pero mayor costo computacional que Spearman

En resumen...

- El análisis inferencial contribuye a validar supuestos antes del entrenamiento de modelos de IA.
- Los resultados se pueden usar como guía para seleccionar features y diseñar los modelos.
- Recordar que es importante verificar los supuestos antes de ejecutar los tests.

Ejemplo práctico

