Sociedad Mexicana de Materiales A.C.

TRONCO COMÚN

Módulo 1:

Introducción a la Minería de Datos

Dr. Irvin Hussein López Nava
Centro de Investigación Científica y de Educación Superior de Ensenada
Facultad de Ciencias, Universidad Autónoma de Baja California

DIPLOMADO EN INTELIGENCIA ARTIFICIAL

Sociedad Mexicana de Materiales A.C.

TEMARIO

Martes 10/feb

Introducción
(1 hora)
- ¿Qué es la minería de datos?
- Conceptos básicos
- Visualización

Ejercicios
(2 horas)
- Visualización de datos genéricos
- Exploración de datos reales

Limpieza
(1 hora)
- Preprocesamiento
- Reducción de dimensionalidad
- Selección de atributos
- Balanceo de clases

Ejercicios
(2 horas)
- Limpieza
- Reducción
- Selección
- Balanceo

Martes 17/feb

Evaluación
(1 hora)
- Partición de datos
- Métricas de rendimiento

Ejercicios
(1 hora)
- Validación de modelo simple
- Análisis de resultados

Sociedad Mexicana de Materiales A.C.

Limpieza y Preparación de Datos

Sociedad Mexicana de Materiales A.C.

Introducción

En **minería de datos**, los **conjuntos de datos** rara vez se encuentran en una forma directamente utilizable.

- Los datos suelen presentar inconsistencias, ausencias, escalas incompatibles o estructuras que no reflejan adecuadamente el fenómeno de interés.

La limpieza y preparación de datos constituye la fase en la que los datos se transforman en un objeto analítico, i.e., en una representación adecuada para el análisis y el modelado.

Sociedad Mexicana de Materiales A.C.

¿Por qué los datos reales no están “listos”?

Los **datos reales** reflejan procesos de medición imperfectos, decisiones humanas y limitaciones instrumentales.

- Valores faltantes, ruido y heterogeneidad no son excepciones, sino propiedades habituales de sistemas reales.

Asumir que los **datos** están listos para el análisis sin una preparación explícita suele conducir a modelos frágiles y conclusiones difíciles de justificar.

Sociedad Mexicana de Materiales A.C.

Preparación de datos

La **minería de datos** se concibe como un proceso iterativo que incluye comprensión del problema, preparación de datos, modelado y evaluación.

Errores o supuestos incorrectos en esta fase tienden a propagarse a lo largo del proceso, afectando resultados posteriores sin ser fácilmente detectables.

Sociedad Mexicana de Materiales A.C.

Decisiones clave

Durante esta etapa se toman decisiones como:

- cómo tratar valores faltantes,
- qué hacer con valores extremos,
- cómo escalar o transformar variables,
- qué atributos conservar o descartar.

Estas decisiones no son neutras: dependen del tipo de análisis, del modelo a emplear y del objetivo del problema.

Preparar datos implica diseñar las condiciones del experimento analítico.

Sociedad Mexicana de Materiales A.C.

Alcance de esta sección

En esta sección se abordarán técnicas y criterios para:

- evaluar la calidad de los datos,
- manejar valores faltantes y outliers,
- reducir dimensionalidad,
- seleccionar atributos relevantes,
- y preparar datos para análisis posteriores.

No se entrenarán modelos finales en esta etapa; el énfasis está en crear una **base sólida** y metodológicamente coherente para el modelado y la evaluación.

Sociedad Mexicana de Materiales A.C.

3.1 Calidad de datos y preprocesamiento

Sociedad Mexicana de Materiales A.C.

Datos tabulares

En la mayoría de los problemas, los conjuntos de datos se representan en forma **tabular**, donde cada fila corresponde a una observación y cada columna a un atributo.

Esta representación implica supuestos importantes:

- independencia entre observaciones,
- significado homogéneo de cada columna, y
- consistencia semántica a lo largo del conjunto de datos.

Sociedad Mexicana de Materiales A.C.

Tipos de datos

Los atributos pueden ser numéricos, categóricos, temporales o booleanos, y cada tipo impone restricciones sobre las operaciones que pueden realizarse.

- Tratar un atributo categórico como numérico, o viceversa, puede introducir relaciones artificiales o distorsionar medidas de similitud.

Identificar correctamente el **tipo** de cada atributo es un paso fundamental antes de aplicar técnicas estadísticas o de aprendizaje automático.

Sociedad Mexicana de Materiales A.C.

Tipos de datos en pandas

En **Pandas**, el tipo de dato de una columna determina cómo se almacenan los valores y qué operaciones están disponibles.

Tipos incorrectos suelen aparecer cuando:

- los datos se cargan desde archivos heterogéneos,
- existen valores faltantes o codificaciones mixtas,
- o se utilizan representaciones genéricas como texto.

Sociedad Mexicana de Materiales A.C.

Inspección temprana

Antes de aplicar técnicas de limpieza o transformación, es recomendable realizar una **inspección estructural básica** del conjunto de datos.

Esta inspección permite identificar:

- tipos incorrectos,
- columnas problemáticas,
- inconsistencias evidentes,
- y posibles errores de diseño del dataset.

Sociedad Mexicana de Materiales A.C.

¿Qué representan los valores faltantes?

Los **valores faltantes** son una característica frecuente en datos reales y no deben interpretarse automáticamente como errores.

Un valor faltante puede indicar:

- que una medición no se realizó,
- que la información no estaba disponible,
- o que el valor no es aplicable en cierto contexto.

Comprender qué representa la **ausencia de datos**
es tan importante como analizar los valores presentes.

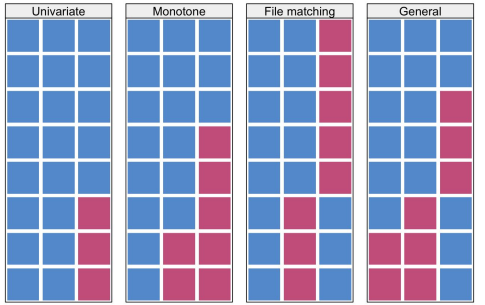
Sociedad Mexicana de Materiales A.C.

Patrones de valores faltantes

Los **valores faltantes** pueden aparecer de forma aislada o siguiendo patrones estructurados.

Por ejemplo:

- concentrados en ciertas variables,
- asociados a subconjuntos de observaciones,
- o correlacionados con otros atributos.



Sociedad Mexicana de Materiales A.C.

Valores faltantes por columna

Un primer paso consiste en cuantificar la presencia de valores faltantes en cada **atributo** o **variable**.

Columnas con altos niveles de ausencia pueden:

- perder poder explicativo,
- requerir imputación cuidadosa,
- o incluso ser descartadas.

La decisión depende tanto del **porcentaje de ausencia** como de la relevancia de la variable.

atributos

Edad	Género	Peso	Estatura	Servicio de salud	Enfermedades previas	Vacunas
5	F	38		1	1	2
6	M	39	113	2	2	1
12	M	38	121	2		
7	F	43	119	1	1	2
5	M	42	116	1	2	1
9	M					
6	M	49	140	1	2	1
	F	22	111	1	2	2
6	F	20	110	1	2	2
9	M	38	133		1	1
9	F		136	1	2	1
4	F	24	106	1	2	
6	M	48	101	1	1	2
5	M	28	122	1	1	1
8	M	21	111	2	2	1
4						
8	F	28	110	2	2	2
4	M		125	2	1	1
4	M	25	140	2	1	2
8	F	50	132	1		1
5				2	1	2
11		29	123	1	2	
12	F	45	116	2	2	1
4	F	28		2	2	1
8	M	43	127		1	2

Sociedad Mexicana de Materiales A.C.

Valores faltantes por fila

Además de analizar **columnas**, es útil examinar cuántos valores faltantes presenta **cada observación**.

Filas con múltiples ausencias pueden corresponder a **registros incompletos** o **problemáticos**.

En algunos casos, eliminar estas observaciones puede ser preferible a imputar múltiples valores sin información suficiente.

observaciones

Edad	Género	Peso	Estatura	Servicio de salud	Enfermedades previas	Vacunas
5	F	38		1	1	2
6	M	39	113	2	2	1
12	M	38	121	2		
7	F	43	119	1	1	2
5	M	42	116	1	2	1
9	M					
6	M	49	140	1	2	1
	F	22	111	1	2	2
6	F	20	110	1	2	2
9	M	38	133		1	1
9	F		136	1	2	1
4	F	24	106	1	2	
6	M	48	101	1	1	2
5	M	28	122	1	1	1
8	M	21	111	2	2	1
4						
8	F	28	110	2	2	2
4	M		125	2	1	1
4	M	25	140	2	1	2
8	F	50	132	1		1
5				2	1	2
11		29	123	1	2	
12	F	45	116	2	2	1
4	F	28		2	2	1
8	M	43	127		1	2

Sociedad Mexicana de Materiales A.C.

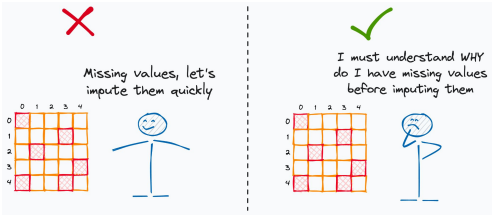
Eliminación ¿cuándo es razonable?

Eliminar filas o columnas con valores faltantes es una estrategia válida en situaciones controladas.

Esta opción suele ser razonable cuando:

- la proporción de datos eliminados es baja,
- la ausencia no está relacionada con la variable objetivo,
- y el tamaño del conjunto de datos lo permite.

Nota: **Eliminar** **sin analizar** el patrón de ausencia puede introducir sesgos inadvertidos.



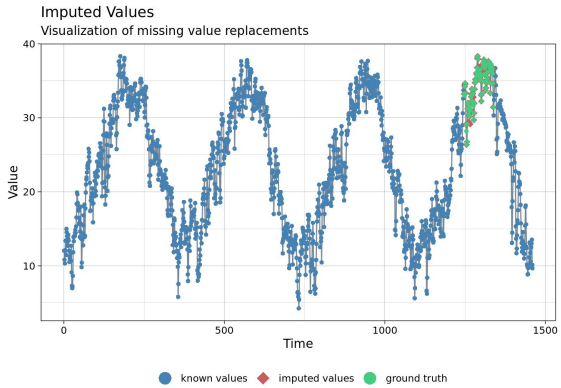
Sociedad Mexicana de Materiales A.C.

Imputación simple de valores faltantes

La **imputación** consiste en **reemplazar** valores faltantes por **estimaciones plausibles**.

Las estrategias simples utilizan **estadísticas básicas** como la media, la mediana o la moda.

Aunque son fáciles de aplicar, estas técnicas reducen la variabilidad y pueden ocultar la incertidumbre asociada a la ausencia de datos.



Sociedad Mexicana de Materiales A.C.

Técnicas de imputación

- En un **primer nivel** se encuentran los métodos basados únicamente en la variable afectada, como la imputación por media, mediana o moda, que asumen una estructura simple de los datos.
- En un **segundo nivel** aparecen métodos que aprovechan relaciones entre variables, utilizando información de otros atributos para estimar valores plausibles.
- **Finalmente**, existen técnicas basadas en modelos estadísticos o de aprendizaje automático, que buscan capturar patrones más complejos, pero introducen supuestos adicionales y mayor complejidad computacional.

Sociedad Mexicana de Materiales A.C.

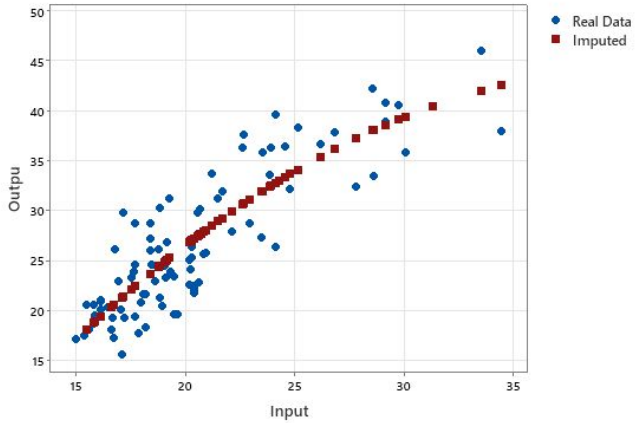
Riesgos asociados a la imputación

Toda **imputación** introduce supuestos adicionales en los datos.

Imputar sin considerar el contexto puede:

- distorsionar distribuciones,
- crear relaciones artificiales,
- o reforzar sesgos existentes.

Por ello, la **imputación** debe entenderse como una decisión analítica, no solo como una operación técnica.



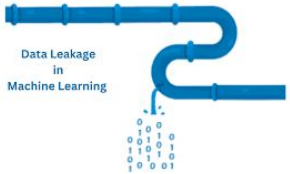
Sociedad Mexicana de Materiales A.C.

Imputación y *data leakage*

Un error común consiste en **imputar** valores utilizando información que no estaría disponible en un escenario real de predicción.

Por ejemplo, calcular estadísticas de imputación usando todo el conjunto de datos antes de separar **entrenamiento** y **prueba**.

Este tipo de filtración de información, conocido como ***data leakage***, conduce a evaluaciones optimistas y modelos poco confiables.



Sociedad Mexicana de Materiales A.C.

¿Qué es un *outlier*?

Un **valor atípico** (*outlier*) es una observación que se desvía significativamente del comportamiento general de los datos.

Esta desviación puede deberse a errores de medición, condiciones excepcionales o a la propia naturaleza del fenómeno estudiado.

Identificar un *outlier* no implica automáticamente que deba eliminarse; su interpretación depende del contexto del problema.

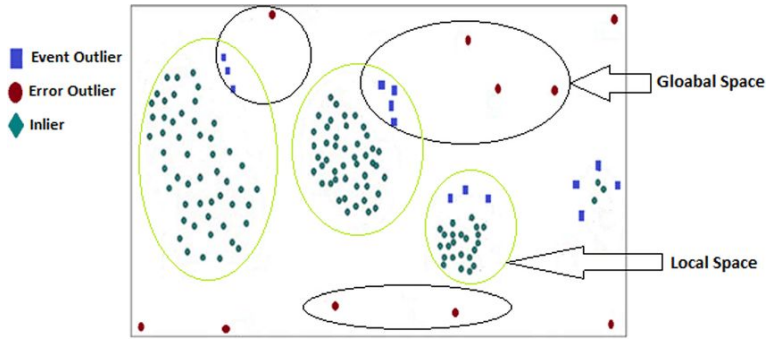
Sociedad Mexicana de Materiales A.C.

Outliers y errores: no son equivalentes

Aunque algunos ***outliers*** corresponden a errores, otros representan eventos reales pero poco frecuentes.

Eliminar sistemáticamente valores extremos puede suprimir información relevante, especialmente en dominios donde los eventos raros son de interés.

Por esta razón, es necesario distinguir entre errores de registro y observaciones válidas pero inusuales.



Sociedad Mexicana de Materiales A.C.

Detección visual de outliers

Las **herramientas de visualización** permiten identificar rápidamente valores atípicos.

Histogramas, diagramas de caja y escalas logarítmicas ayudan a revelar colas largas, asimetrías y valores extremos.

La detección visual no sustituye al análisis formal, pero es un primer filtro efectivo para orientar decisiones posteriores.

Sociedad Mexicana de Materiales A.C.

Estrategias para el manejo de *outliers*

- En un **primer nivel** se encuentran enfoques basados en reglas simples, como la eliminación directa de observaciones claramente erróneas o fuera de rango.
- Un **segundo nivel** incluye técnicas que modifican la representación de los datos, como transformaciones o recortes controlados, con el objetivo de reducir la influencia de valores extremos sin descartarlos por completo.
- **Finalmente**, existen enfoques que incorporan explícitamente los *outliers* en el análisis, ya sea mediante métodos estadísticos robustos o modelos que distinguen entre comportamiento típico y atípico.

Sociedad Mexicana de Materiales A.C.

Robustez frente a sobrelimpieza

Una **limpieza excesiva** puede eliminar patrones reales junto con el ruido.

Métodos robustos buscan reducir la sensibilidad a valores extremos sin descartarlos explícitamente.

El **objetivo no es forzar** los datos a cumplir supuestos ideales, sino adaptar el análisis a su estructura real.

Sociedad Mexicana de Materiales A.C.

Documentar cada decisión

Cada decisión sobre el tratamiento de **valores faltantes**, o manejo de **valores atípicos**, debe quedar explícitamente **documentada**.

Esto incluye:

- la estrategia utilizada,
- las variables afectadas,
- y la justificación de la elección.

La documentación es esencial para la reproducibilidad,
la interpretación de resultados y la comunicación científica.

Sociedad Mexicana de Materiales A.C.

3.2 Reducción de dimensionalidad

Sociedad Mexicana de Materiales A.C.

La dimensionalidad

En muchos **problemas reales**, los conjuntos de datos contienen un gran número de atributos, no todos igualmente informativos.

A medida que **aumenta la dimensionalidad**, los datos se vuelven más dispersos y las relaciones relevantes pueden quedar ocultas por ruido o redundancia.

La **alta dimensionalidad** no solo afecta la interpretación, sino también la eficiencia y estabilidad de los métodos analíticos.

Sociedad Mexicana de Materiales A.C.

Impacto

La **alta dimensionalidad** afecta tanto la visualización como el modelado: las proyecciones directas resultan poco informativas y las medidas de distancia o similitud pierden capacidad discriminativa.

Como consecuencia, los modelos pueden volverse más inestables, menos interpretables y computacionalmente más costosos, sin un beneficio claro en desempeño.

La **reducción de dimensionalidad** busca mitigar estos efectos al concentrar la información relevante en un espacio más manejable.

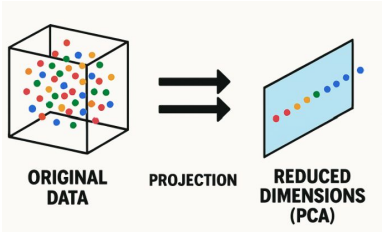
Sociedad Mexicana de Materiales A.C.

Análisis de Componentes Principales

Cuando los datos presentan **alta dimensionalidad** y **correlaciones** entre variables, muchas dimensiones aportan información redundante.

PCA permite **transformar** el conjunto original de variables en un nuevo sistema de ejes que captura la mayor variabilidad posible con menos dimensiones.

Esta transformación facilita la exploración visual, reduce el ruido y mejora el comportamiento de muchos algoritmos posteriores.



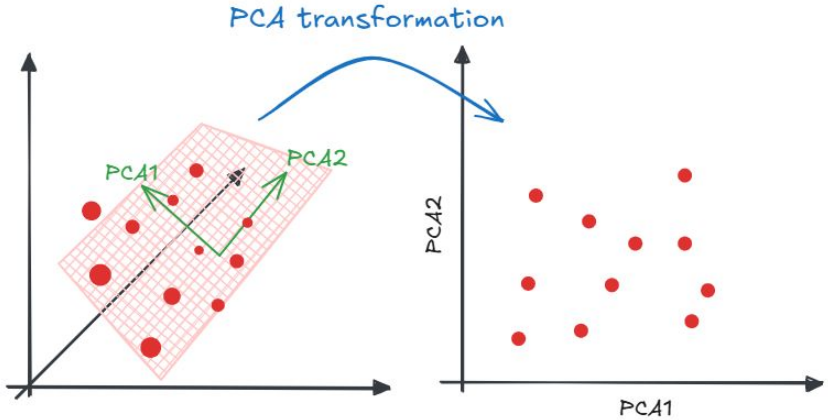
Sociedad Mexicana de Materiales A.C.

Intuición geométrica

PCA identifica **direcciones en el espacio de datos** donde la variabilidad es máxima.

Estas direcciones, llamadas **componentes principales**, son combinaciones lineales de las variables originales y están ordenadas por la cantidad de varianza que explican.

Geométricamente, el método busca rotar el sistema de coordenadas para alinear los ejes con la estructura dominante de los datos.



Sociedad Mexicana de Materiales A.C.

Componentes Principales

- Cada **componente principal** explica una fracción de la varianza total del conjunto de datos, y las primeras componentes suelen concentrar la mayor parte de la información relevante.
- La selección del **número de componentes** se basa en el análisis de la varianza acumulada, buscando un equilibrio entre reducción de dimensionalidad y preservación de la estructura original.
- Dado que PCA es sensible a la escala de las variables, el **escalamiento previo** resulta fundamental para evitar que atributos con mayor magnitud dominen la transformación.

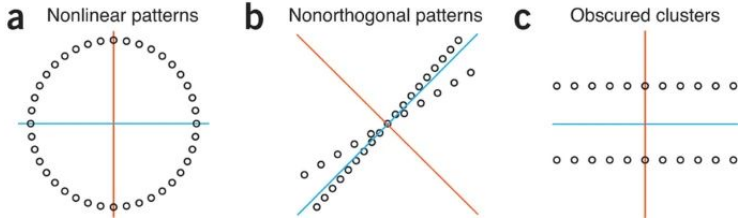
Sociedad Mexicana de Materiales A.C.

Limitaciones de PCA

PCA **asume relaciones lineales** entre variables y no captura estructuras no lineales complejas.

Además, los **componentes principales** suelen ser difíciles de interpretar directamente en términos de las variables originales.

Por estas razones, PCA es una herramienta exploratoria poderosa, pero no universal.



Sociedad Mexicana de Materiales A.C.

Proyecciones no lineales

En muchos conjuntos de datos reales, las relaciones entre variables **no siguen** estructuras lineales simples, lo que limita la capacidad de métodos como PCA para revelar patrones relevantes.

En estos casos, **proyecciones no lineales** permiten capturar estructuras locales complejas, como agrupamientos curvos o *manifolds* de baja dimensión inmersos en espacios de alta dimensión.

Estas técnicas están orientadas principalmente a exploración visual, no a transformaciones estables para el modelado predictivo.

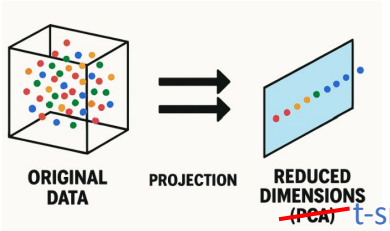
Sociedad Mexicana de Materiales A.C.

t-SNE

t-SNE construye una **representación de baja dimensión** preservando la similitud local entre observaciones, definida a partir de probabilidades de vecindad en el espacio original.

El método busca que puntos cercanos permanezcan cercanos tras la proyección, aun cuando se distorsione la estructura global del conjunto de datos.

Como resultado, t-SNE es especialmente útil para identificar clusters locales, subgrupos y separaciones no evidentes en proyecciones lineales.



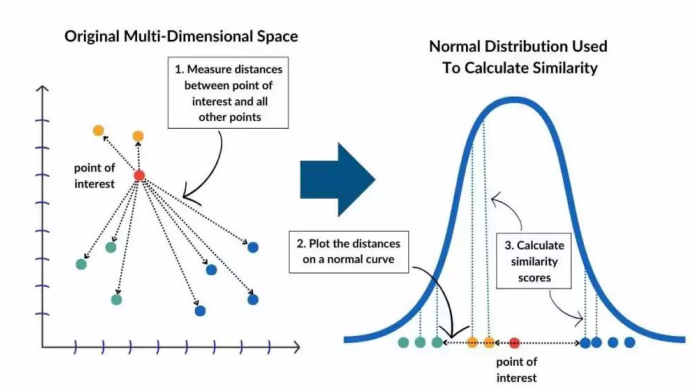
Sociedad Mexicana de Materiales A.C.

Fundamento de t-SNE

t-SNE asume que la estructura relevante de los datos está determinada por relaciones locales de vecindad, más que por la preservación de distancias globales.

La técnica modela explícitamente la **similitud** entre pares de puntos en el espacio original mediante distribuciones de probabilidad que indican qué tan probable es que dos observaciones sean vecinas.

La proyección de baja dimensión se obtiene **ajustando iterativamente** la posición de los puntos para minimizar la discrepancia entre las similitudes originales y las inducidas en el espacio reducido.

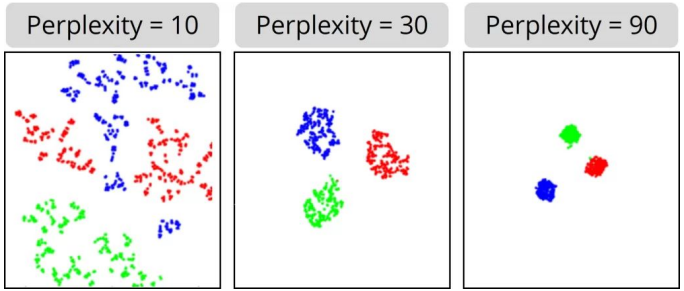


Sociedad Mexicana de Materiales A.C.

Parámetros clave y sensibilidad

El **comportamiento** de t-SNE depende fuertemente de parámetros como la perplexidad, que controla el tamaño efectivo del vecindario considerado para cada punto.

Cambios pequeños en estos parámetros pueden producir visualizaciones significativamente distintas,
por lo que la interpretación
debe hacerse con cautela.



Sociedad Mexicana de Materiales A.C.

Limitaciones de t-SNE

Las distancias globales en una proyección t-SNE no son interpretables: la separación entre clusters no refleja necesariamente relaciones reales en el espacio original.

Asimismo, la forma, tamaño o densidad visual de los grupos no debe interpretarse como evidencia cuantitativa sin análisis adicional.

Por estas razones, t-SNE debe utilizarse como apoyo a otras técnicas analíticas, no como evidencia concluyente por sí sola.

Sociedad Mexicana de Materiales A.C.

Usos prácticos de PCA y t-SNE

- En la práctica, PCA se utiliza principalmente como una herramienta de **preprocesamiento**: reducción de dimensionalidad antes de modelado, eliminación de redundancia y estabilización de métodos sensibles a la correlación entre variables.
- En contraste, t-SNE se emplea sobre todo con fines **exploratorios y diagnósticos**, para revelar estructuras locales, agrupamientos y separaciones no lineales que no son evidentes en el espacio original.

Sociedad Mexicana de Materiales A.C.

3.3 Selección de atributos

Sociedad Mexicana de Materiales A.C.

Motivación

En muchos conjuntos de datos, no todos los atributos **contribuyen de forma significativa** a la tarea analítica, y algunos pueden incluso introducir ruido o redundancia.

La **selección de atributos** busca identificar un subconjunto informativo que preserve la capacidad descriptiva o predictiva de los datos, reduciendo al mismo tiempo la complejidad del problema.

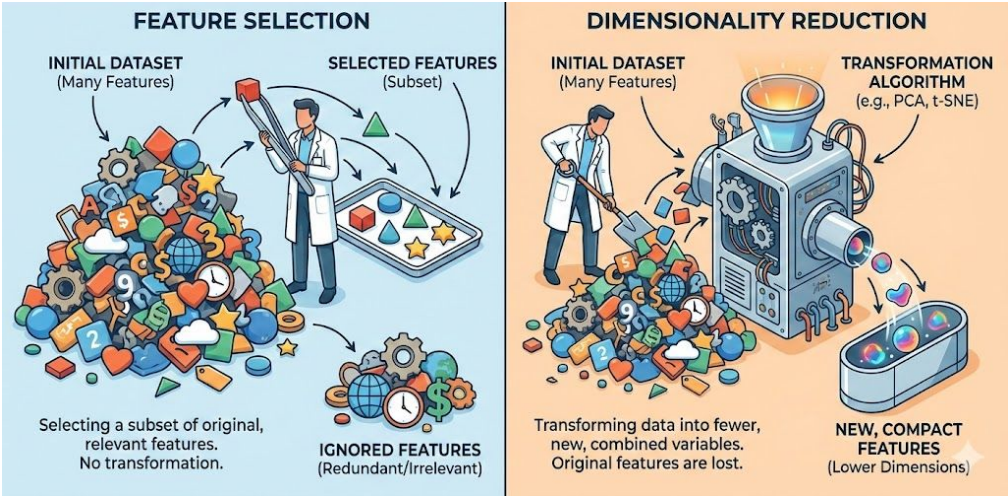
A diferencia de la **reducción de dimensionalidad**, este proceso mantiene el significado original de las variables, lo que favorece la interpretabilidad.

Sociedad Mexicana de Materiales A.C.

Selección vs. reducción

La **selección de atributos** elimina variables completas, mientras que la reducción de dimensionalidad transforma el espacio de representación mediante combinaciones de atributos.

Ambos enfoques persiguen objetivos similares: **simplificación y robustez**, pero difieren en su impacto sobre la interpretación y el uso posterior de los datos.

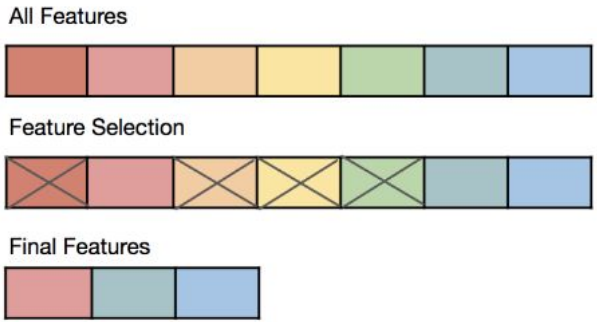


Sociedad Mexicana de Materiales A.C.

Criterios generales de selección

Un **atributo informativo** es aquel que aporta variabilidad relevante, está relacionado con la variable objetivo o contribuye a separar observaciones de interés.

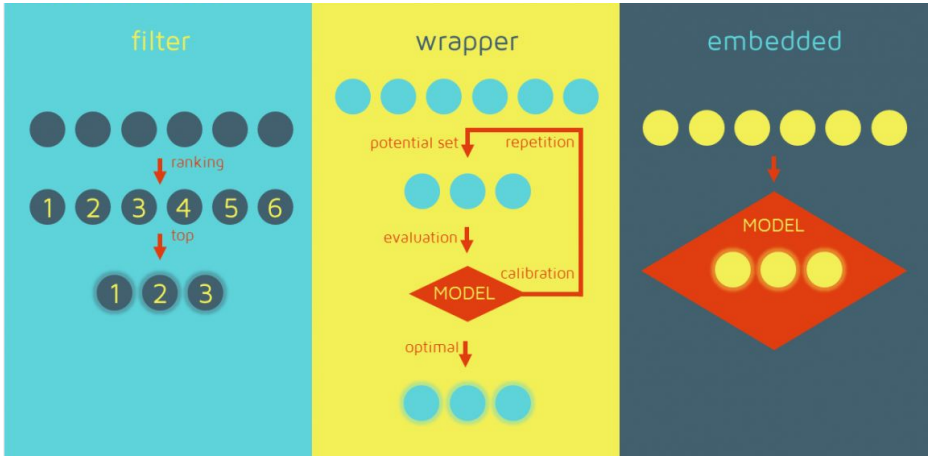
Por el contrario, atributos altamente correlacionados entre sí, con baja variación o dominados por ruido suelen aportar poco valor adicional.



Sociedad Mexicana de Materiales A.C.

Principales enfoques

Los métodos de selección de atributos suelen agruparse en tres grandes familias: métodos basados en **filtros**, métodos basados en ***wrappers*** y métodos **embebidos**, cada una con distintos compromisos entre costo computacional, interacción con el modelo y estabilidad.



Sociedad Mexicana de Materiales A.C.

Métodos Filter

Evalúan los atributos de forma independiente del modelo de aprendizaje, utilizando **criterios estadísticos** o medidas de dependencia con la variable objetivo.

Su principal ventaja es la **eficiencia computacional** y la **generalidad**, lo que los hace adecuados como etapa temprana de preprocesamiento.

Entre las técnicas más utilizadas se encuentran la ganancia de información, la información mutua, la correlación y pruebas estadísticas, que permiten rankear atributos según su relevancia individual.

Sociedad Mexicana de Materiales A.C.

Métodos *Wrapper*

Utilizan el desempeño de un modelo específico como criterio para evaluar subconjuntos de atributos.

Este enfoque permite capturar **interacciones complejas** entre variables, pero implica un costo computacional elevado debido a la exploración de múltiples combinaciones.

Técnicas comunes incluyen *forward selection*, *backward elimination* y búsquedas heurísticas, donde el subconjunto de atributos se construye o refina iterativamente en función del desempeño del modelo.

Sociedad Mexicana de Materiales A.C.

Métodos *Embedded*

Integran la selección de atributos **directamente en el proceso de entrenamiento** del modelo.

Este enfoque ofrece un balance entre eficiencia y capacidad de capturar relaciones relevantes, aunque los resultados dependen del modelo específico utilizado.

Ejemplos representativos incluyen regularización L1, árboles de decisión y métodos basados en importancia de variables, donde la selección emerge como parte natural del ajuste del modelo.

Sociedad Mexicana de Materiales A.C.

Consideraciones prácticas

La **selección de atributos** no es un paso puramente técnico, sino una decisión analítica que impacta directamente la interpretabilidad, estabilidad y desempeño de los modelos.

Cada criterio de selección —filtros, wrappers o embebidos— implica supuestos distintos sobre los datos y el objetivo del análisis.

Por ello, debe entenderse como un **proceso iterativo**, informado por el conocimiento del dominio, la exploración de los datos y la validación del modelo, más que como una optimización automática aislada.

Sociedad Mexicana de Materiales A.C.

3.4 Balanceo de clases

Sociedad Mexicana de Materiales A.C.

El desbalance

En muchos problemas de clasificación, las clases **no están representadas de manera uniforme**, lo que introduce sesgos durante el entrenamiento de los modelos.

Este **desbalance** puede llevar a modelos que optimizan métricas globales mientras fallan sistemáticamente en clases minoritarias, que suelen ser las de mayor interés analítico.

El **balanceo de clases** busca mitigar este efecto, mejorando la capacidad del modelo para aprender patrones relevantes en todos los grupos.

Sociedad Mexicana de Materiales A.C.

Balanceo como decisión analítica

El balanceo de clases **no es un paso obligatorio**, sino una intervención condicionada por el objetivo del análisis y las métricas de evaluación.

En algunos contextos, preservar la distribución original **es preferible**; en otros, es necesario modificarla para evitar modelos trivialmente sesgados.

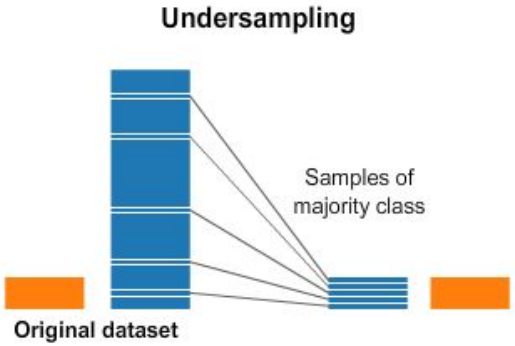
Por ello, **el balanceo** debe entenderse como una decisión analítica informada, no como una corrección automática.

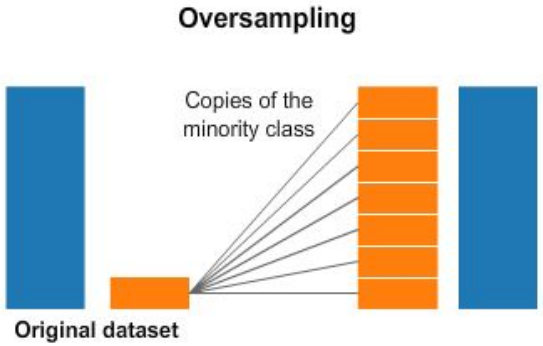
Sociedad Mexicana de Materiales A.C.

Técnicas para balanceo

Pueden agruparse en dos enfoques principales: **submuestreo** y **sobremuestreo**.

Cada familia introduce distintos compromisos entre preservación de información, riesgo de sobreajuste y complejidad computacional.





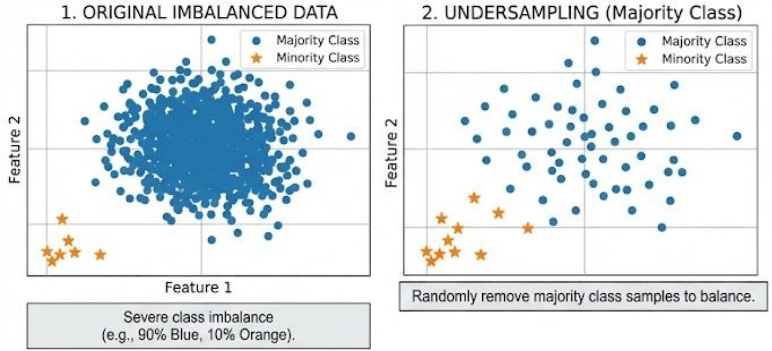
Sociedad Mexicana de Materiales A.C.

Submuestreo

El **submuestreo** **reduce** el número de observaciones de la clase mayoritaria para equilibrar la distribución.

Técnicas comunes incluyen submuestreo aleatorio y métodos guiados por proximidad o densidad, que buscan conservar observaciones representativas.

Este enfoque es eficiente, pero puede descartar información relevante si se aplica de forma agresiva.



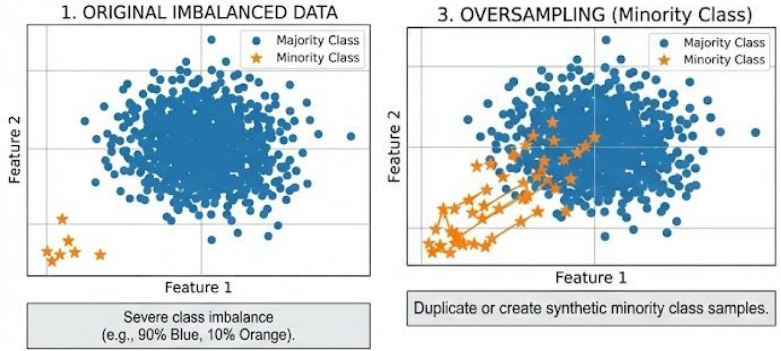
Sociedad Mexicana de Materiales A.C.

Sobremuestreo

El **sobremuestreo** **incrementa** artificialmente la representación de las clases minoritarias, generalmente mediante duplicación de observaciones existentes.

Métodos como *random oversampling* son simples de aplicar, pero pueden inducir sobreajuste al repetir exactamente los mismos ejemplos.

Por esta razón, suelen utilizarse como línea base o en combinación con otras técnicas.

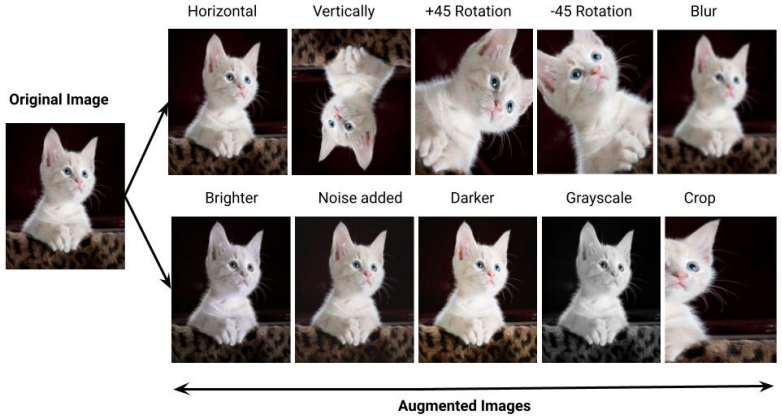


Sociedad Mexicana de Materiales A.C.

Aumento de datos

Las técnicas de *data augmentation* buscan generar **nuevas** observaciones plausibles para la clase minoritaria, en lugar de replicar datos existentes.

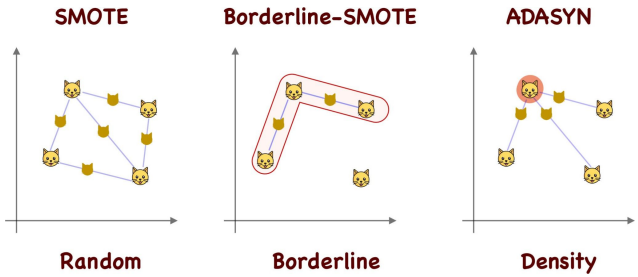
Estos métodos amplían el espacio de representación de la clase minoritaria, reduciendo el riesgo de sobreajuste y mejorando la generalización.



Sociedad Mexicana de Materiales A.C.

SMOTE

- *Synthetic Minority Over-sampling Technique* genera **nuevas observaciones** interpolando entre vecinos cercanos de la clase minoritaria, ampliando el espacio de representación sin recurrir a duplicaciones exactas.
- Este enfoque preserva la estructura local de los datos y reduce el riesgo de sobreajuste asociado al sobremuestreo simple.
- Extensiones de SMOTE: Borderline-SMOTE, que enfatiza regiones cercanas a la frontera entre clases, y ADASYN, que adapta la generación sintética según la dificultad local del problema.



Sociedad Mexicana de Materiales A.C.

Riesgos del balanceo

Las técnicas de **balanceo** influyen directamente en cómo el modelo representa la estructura del problema y en los resultados obtenidos.

Aplicarlas **sin una validación** adecuada puede introducir artefactos, especialmente si el balanceo se realiza antes de separar los conjuntos de entrenamiento y prueba, generando fuga de información.

Por ello, el **balanceo** debe integrarse explícitamente en el diseño del experimento, documentando las técnicas empleadas, el momento de su aplicación y su impacto en las métricas de evaluación, como parte de un flujo analítico riguroso.

Sociedad Mexicana de Materiales A.C.

¿Qué se construyó en esta sección?

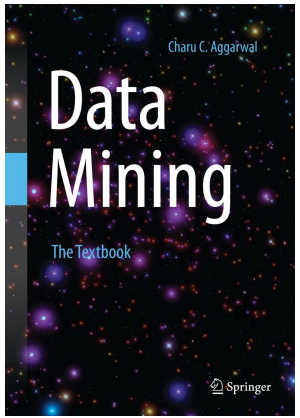
Manejo de valores faltantes y su impacto → Identificación y tratamiento de valores atípicos → Reducción de dimensionalidad para visualización

Técnicas para la Selección de atributos → Balanceo de clases y generación sintética → 

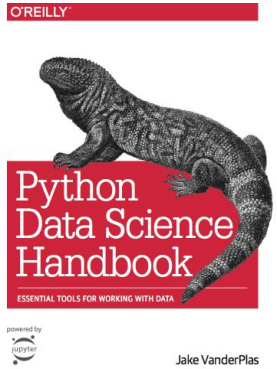
17/feb

Sociedad Mexicana de Materiales A.C.

Bibliografía



Aggarwal, C. C. (2015). *Data mining: the textbook* (Vol. 1, No. 3). New York: springer.



VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. O'Reilly Media, Inc.

https://jakevdp.github.io/PythonDataScienceHandbook/