#3 Data preprocessing

by Saturdays.AI

Saturdays.AI LATAM

Saturdays.AI
LATAM

Week 3

Saturdays.AI
LATAM

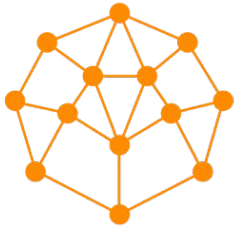
Objetivo

Aprender diferentes técnicas para el preprocesado de los datos

Metodología de Ciencia de Datos.

Week 3

Agenda



- Introducción
- Data cleaning
  - Valores faltantes
  - Valores duplicados
- Data transformation
  - Codificación de los datos
  - Escalamiento de los datos
  - Normalización de los datos
- Data reduction
  - Selección de características
  - Extracción de características

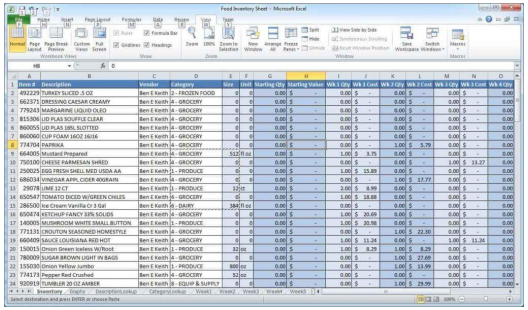
Datooos!!! Muchos datos!!

- Recurso **más valioso** hoy en el mundo actual.
- Según el Foro Económico Mundial, para el 2025 estaremos generando alrededor de 463 exabytes de datos a nivel mundial por día.
- Pero,

¿Todos estos **datos son lo suficientemente adecuados** para ser utilizados por algoritmos de aprendizaje automático?

Saturdays.AI
LATAM

Hablando de datos pensamos ...



- Grandes conjuntos de datos con una gran cantidad de filas y columnas.

- No siempre es el caso: los **datos pueden estar en muchas formas diferentes**: tablas estructuradas, imágenes, archivos de audio, videos, etc.

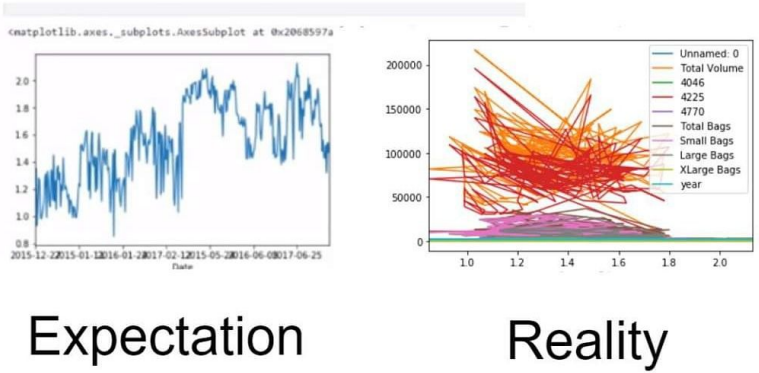
- Las **máquinas no entienden** el texto libre, las imágenes o los datos de video tal como están, **entienden los 1 y 0**.

Saturdays.AI
LATAM

Tenemos un problema

- Los datos del mundo real a menudo son
  - incompletos,
  - inconsistentes,
  - contienen muchos errores.

me everytime:



La calidad de los datos afecta directamente la capacidad de nuestro modelo para aprender.

Saturdays.AI
LATAM

Data preprocessing

- Paso muuuy importante en todo proceso de Machine Learning
    - Aplicar transformaciones a los datos para llevarlos a un estado que la máquina ahora puede analizar fácilmente.

- **Requiere muuuucho tiempo.**

Saturdays.AI
LATAM

Definiciones

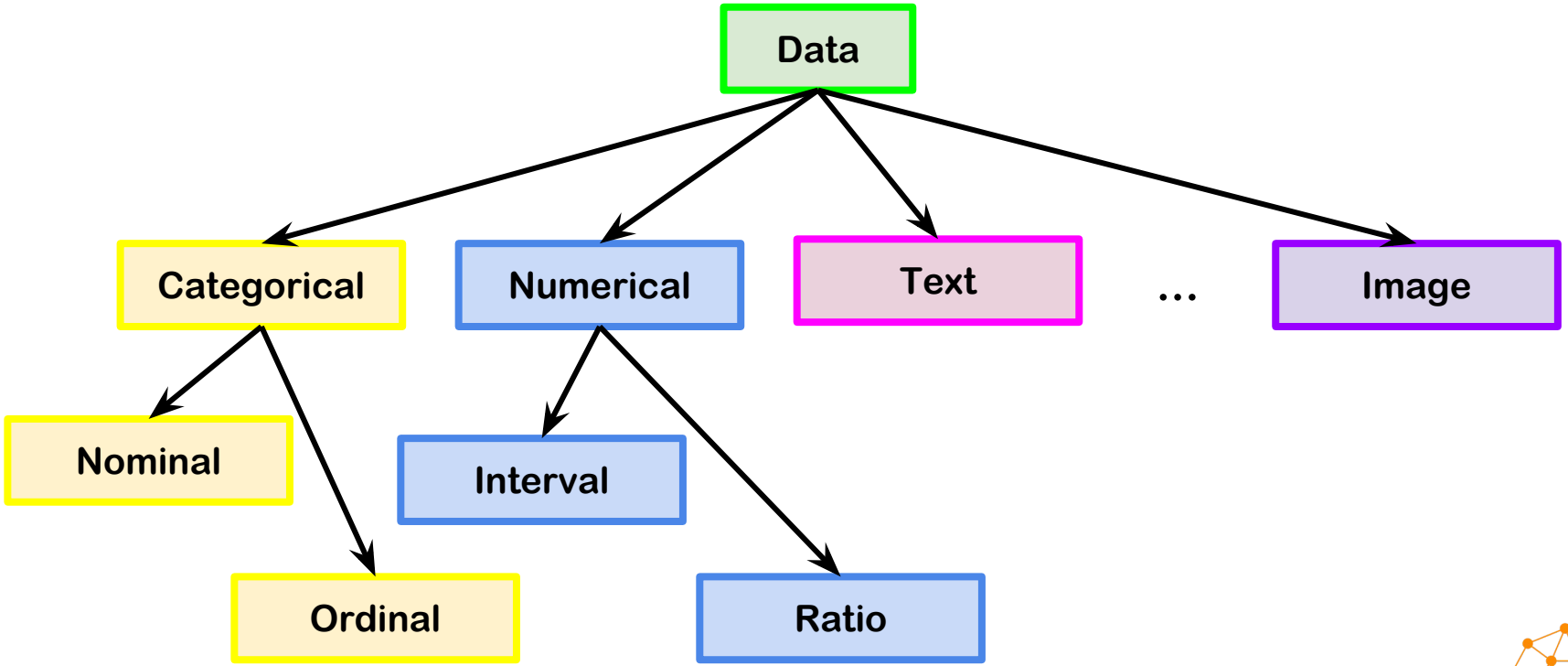
Dataset:

- **Colección de datos**
  - Ejemplos: registros de interacciones, eventos, observaciones.
- Descritos mediante una serie de características o **features**.
  - Ejemplos: la masa de un objeto físico o el momento en que ocurrió un evento, etc..

Feature vector

Features					Label
Position	Experience	Skill	Country	City	Salary ($)
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Definiciones: Tipos de features



Saturdays.AI
LATAM

Tipos de features : Categorical

- Características cuyos valores se toman de un **conjunto definido de valores**.
  - Ejemplos?.... En el chat :D

Nominal

- Variables categóricas **sin un orden implícito**
- **Ejemplo:** Los colores de un carro: negro, morado, rosa

Ordinal

- Variables categóricas **con un orden natural implícito**
- **Ejemplo:** Los tamaños de la ropa: chico, mediano, grande

Saturdays.AI
LATAM

Tipos de features : Numerical

- Características representadas por **números cuyos valores son continuos o discretos.**
  - Ejemplos?.... En el chat :D

Interval

- Con una unidad de medida definida
- Representa valores como 0 y menores que 0

- **Ejemplo:** temperatura en Celsious

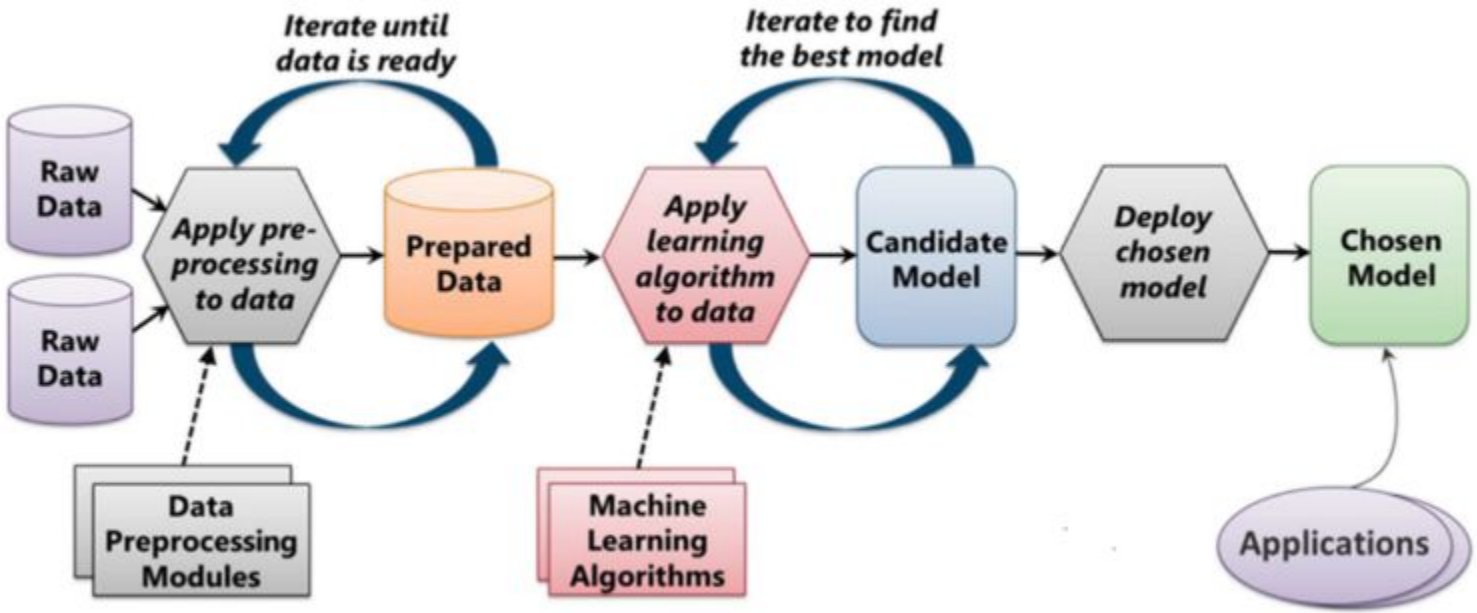
Ratio

- Con una unidad de medida definida
- Representa valores de 0 y mayores a 0

- **Ejemplo:** estatura y peso

Saturdays.AI
LATAM

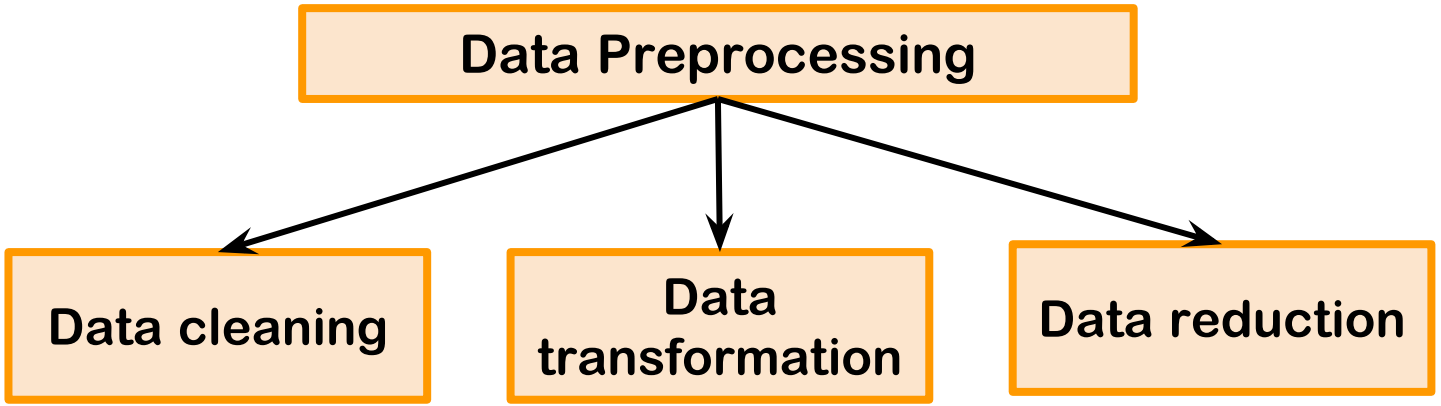
Metodología de Ciencia de Datos.



From "Introduction to Microsoft Azure" by David Chappell

Saturdays.AI
LATAM

Data Preprocesing ....



#3.1 Data cleaning

Saturdays.AI
LATAM

Data cleaning

- Los datos pueden faltar durante la extracción o recolección de datos.
- Reducen la calidad de cualquiera de nuestras métricas de rendimiento.
- **“Data cleaning”: Técnica que implica el manejo de datos faltantes, datos ruidosos, etc.**

EDA

Los valores faltantes pueden aparecer como un signo de interrogación (?) O un cero (0) o menos uno (-1) o un espacio en blanco.

Data cleaning - Missing values

Ignorarlos: Solo cuando el conjunto de datos que tenemos es bastante grande.

Eliminar filas con datos faltantes:

- Estrategia simple y a veces efectiva.
- Falla si muchos objetos tienen valores faltantes.
- Si una *característica tiene valores perdidos en su mayoría, entonces esa característica en sí también se puede eliminar (+ 75%).*

Estimar valores perdidos:

- Si solo falta un porcentaje razonable de valores (20%)
- Métodos más comunes:
  - Manualmente.
  - Promedio.
  - El valor más probable de la característica respectiva.
  - ML algoritmo.

Saturdays.AI
LATAM

¿Que usarias?

Ejemplo 1

ID	Children	Age of youngest child	Did you drink Coca-Cola in the last 24 hours?	How many colas did you drink in the past 24 hours?
1	No		No	
2	Yes	18	Yes	2
3	No		No	
4	Yes	13	No	
5	Yes	8	Yes	1

Ejemplo 2

ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High

Saturdays.AI
LATAM

Data cleaning - Duplicate values

- Puede suceder cuando la misma persona envía un formulario más de una vez.

- En la mayoría de los casos, los duplicados **se eliminan** para no dar a ese objeto de datos en particular una ventaja o sesgo.

Saturdays.AI
LATAM


Time to Code!!! :D
Saturdays.AI
LATAM

#3.3 Data transformation

Feature Enconding

Saturdays.AI
LATAM

Data transformation

- Proceso en el que toma datos de su estado de origen sin procesar, y los transforma en datos listos para el análisis.

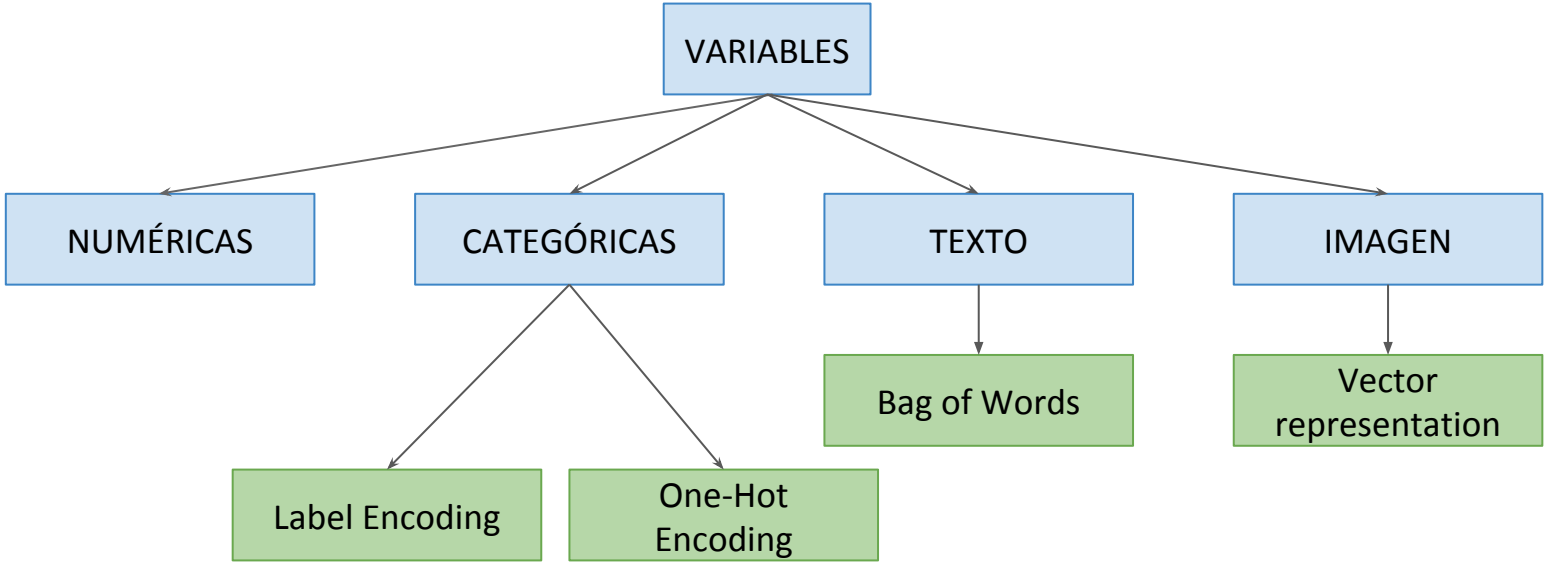
Saturdays.AI
LATAM

Data transformation: Feature encoding

- Consiste en realizar transformaciones en los datos para aceptarse como entrada para algoritmos de aprendizaje automático, manteniendo su significado original.
- Los algoritmos de aprendizaje automático sólo puede aceptar valores numéricos.

Saturdays.AI
LATAM

Tipos de características



Variables categóricas

- Una variable **categórica** es una variable que puede tomar uno de un número **limitado** de posibles **valores o categorías**.

ID	TIPO
1	COMPRA
2	VENTA
3	VENTA

Variables categóricas - Label Encoding

- Asigna a cada categoría un valor numérico distinto.
- Útil cuando las categorías **tienen relación**.
    - Ejemplo nivel socioeconómico: bajo = 1, medio = 2, alto = 3.
- **Ligero computacionalmente**, genera solamente una columna.
- Si las categorías **no tiene relación** puede generar sesgos a los algoritmos de ML.

ID	TIPO	TIPO
1	COMPRA	1
2	VENTA	2
3	VENTA	2

Saturdays.AI LATAM

Variables categóricas - One-Hot Encoding

- Asigna a cada categoría un vector que contiene 1 y 0 que denota la presencia o ausencia de la característica.
- Útil cuando las categorías **no tienen relación** mejorando la precisión de de los algoritmos.
- Puede generar **demasiadas columnas**.

ID	TIPO	COMPRA	VENTA
1	COMPRA	1	0
2	VENTA	0	1
3	VENTA	0	1

Saturdays.AI LATAM

Variables de texto

PROD_DESC
Unspecified vomiting of pregnancy, unspecified as to episode of care or not applicable
Retained (old) foreign body following penetrating wound of orbit.
Subjective visual disturbance, unspecified.

Variables de texto - Bag of Words

- **Bag-of-words** es una representación numérica de variables de texto que describe la aparición de palabras dentro de un documento.
- Este utiliza:
  - **Vocabulario** de palabras conocidas.
  - Una **métrica** que evalúa la **presencia** de palabras del vocabulario.

Saturdays.AI
LATAM

Variables de texto - Bag of Words

Dataset

ID	PROD_DESC
1	Algodón, poliéster, sin cordones. Preferible lavar a máquina.
2	Solo algodón, con cordones. No lavar a máquina.
3	Playera azul. No lavar a máquina.

Vocabulario

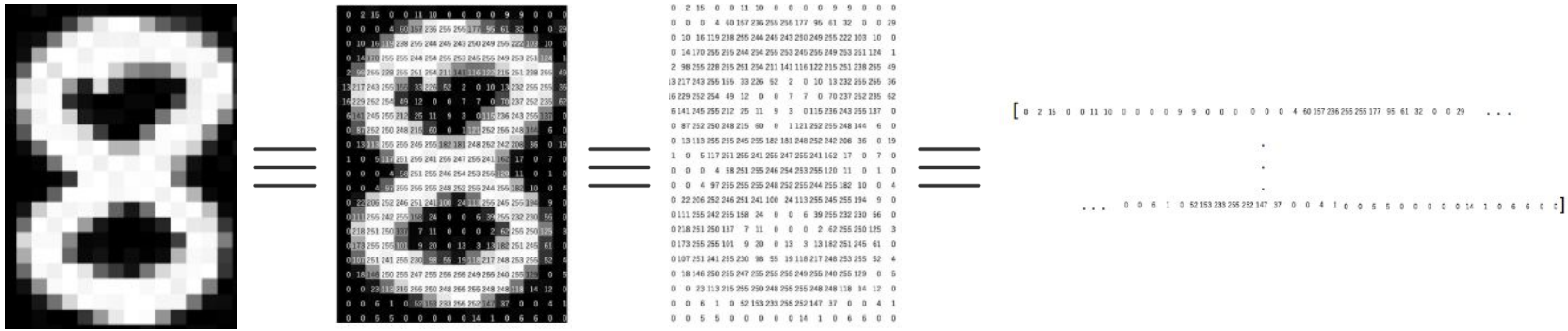
1. Algodón
2. Poliéster
3. Cordones
4. Preferible
5. Lavar
6. Máquina.
7. Playera
8. Azul

Representación Bag of Words

ID	Algodón	Poliéster	Cordones	Preferible	Lavar	Máquina	Playera	Azul
1	1	1	1	1	1	1	0	0
2	1	0	1	0	1	1	0	0
3	0	0	0	0	1	1	1	1

Saturdays.AI
LATAM

Imágenes - Vector representation



Fuente:Link


Time to Code!!! :D
Saturdays.AI
LATAM

#3.2 Transformación de datos

Escalamiento / Feature scaling
Normalización de datos / and Feature normalization

Saturdays.AI
LATAM

Feature scaling vs Feature normalization

- Cambiar los valores de las columnas numéricas en el conjunto de datos a una escala común

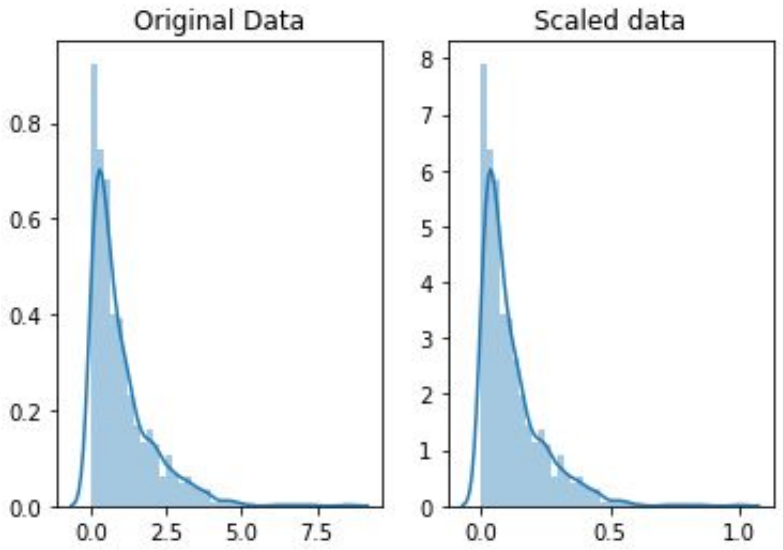
Feature scaling, cambia el **rango** de sus datos

Feature normalization cambiando la forma de la **distribución** de sus datos.

Saturdays.AI
LATAM

Feature scaling

- Transformación para que se **ajusten a un rango específico**.
  - Ejemplo: 0–100 o 0–1.
- Útil cuando se usa en métodos basados en medidas de qué tan separados están los puntos de datos



Observe que la forma de los datos no cambia, pero que en lugar de variar de 0 a 8, ahora varía de 0 a 1.

Saturdays.AI LATAM

Feature scaling

Con Feature Scaling generalmente nos referimos a generar un cambio en el rango de los valores sin cambiar la forma de la distribución de los mismos, de esta forma es posible mejorar el performance de un algoritmo de machine learning o su tiempo de ejecución cuando sus features se encuentran en una escala similar o cercana a ser normalmente distribuida.

Algunos algoritmos que pueden mejorar su performance mediante estas técnicas son:

* Regresión logística y lineal
* Nearest Neighbors
* Redes Neurales
* Análisis de componentes principales

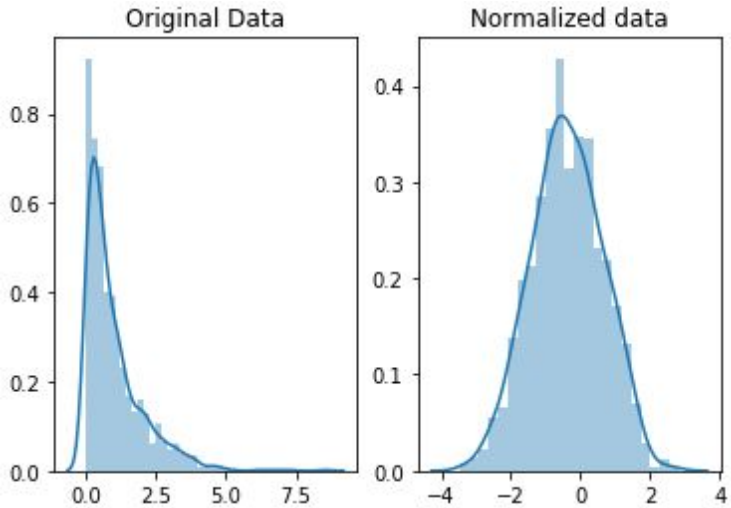
Saturdays.AI
LATAM

Feature normalization

- Pretendemos cambiar los valores de columnas numéricas en el dataset a una escala en común, sin distorsionar las diferencias en los rangos de valores. En machine learning no siempre los dataset requieren normalización, sólo es requerido cuando los features tienen rangos diferentes.
- La normalización es útil cuando se desconoce la distribución de los datos o cuando la distribución no es Gaussiana (en campana).

Feature normalization - Standardization

- **Z-Normalization (Standardization):**
  - Transformación para que se **ajusten a una distribución normal.**



Saturdays.AI
LATAM


Time to Code!!! :D
Saturdays.AI
LATAM

#3.3 Data reduction

Saturdays.AI
LATAM

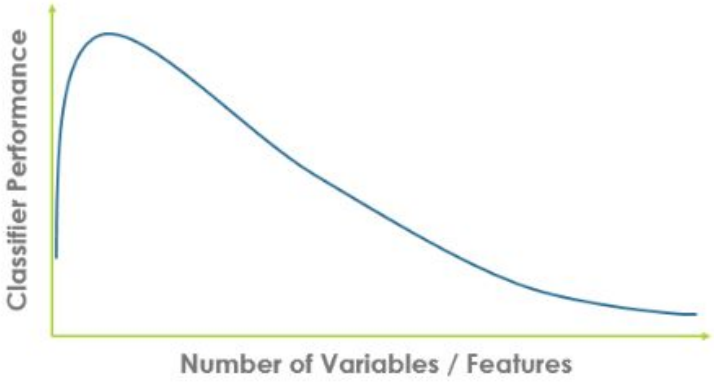
Data reduction

- Manejamos una gran cantidad de datos.
  - Problema:
    - **Gran volumen de datos = Análisis más difícil**.
  - Solución: Técnicas de reducción de datos.

- Objetivo:
  - Reducir el almacenamiento de datos.
  - Eliminar características innecesarias.
  - Reducir los costos computacionales durante análisis.

The curse of dimensionality

- "**Dimensionalidad**" simplemente se refiere al **número de características** en su conjunto de datos.



- Cuando el **número de características es muy grande en relación con el número de observaciones** en su conjunto de datos, ciertos algoritmos luchan por entrenar modelos efectivos.

Saturdays.AI
LATAM

Data reduction - Feature selection

- Filtrar características irrelevantes o redundantes de su conjunto de datos.

Feature selection != Feature extraction

1. Feature selection **mantiene un subconjunto** de las características originales
2. Feature extraction **crea nuevas**.

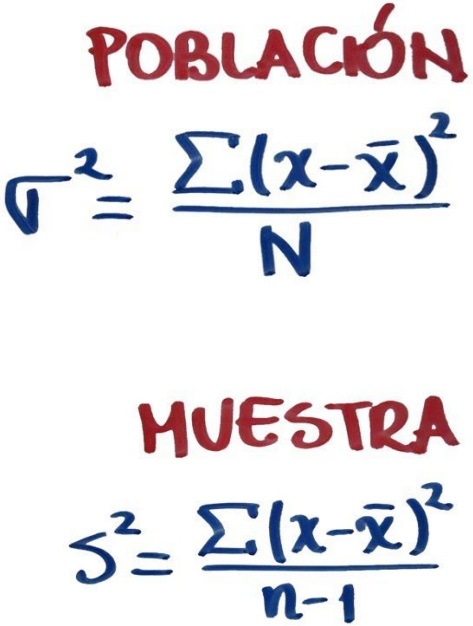
Saturdays.AI
LATAM

Feature selection - Variance thresholds

- Eliminan características cuyos **valores no cambian mucho de una observación a otra** (es decir, su varianza cae por debajo de un umbral).

- Estas características proporcionan poco valor.

- Debido a que la **variación depende de la escala, siempre debes normalizar** sus características primero.



Saturdays.AI
LATAM

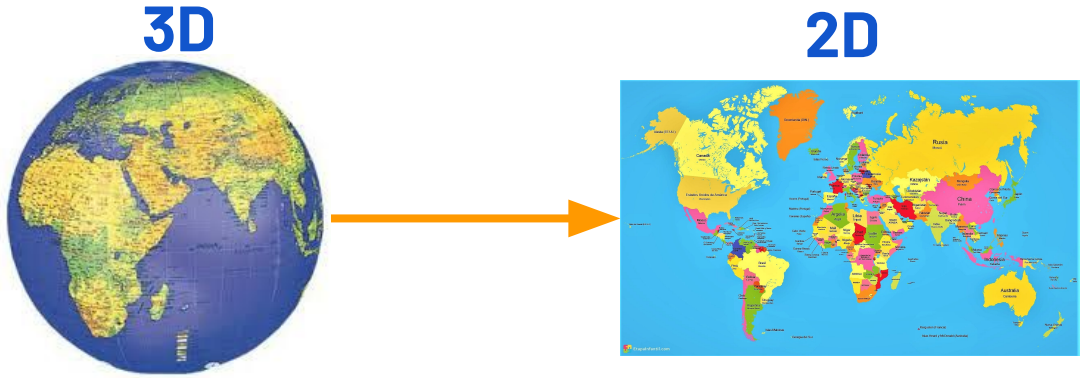
Feature selection - Correlation thresholds

- Eliminan características que están **altamente correlacionadas con otras** (es decir, sus valores cambian de manera muy similar a la de los demás).

- Estas características proporcionan información redundante.

Saturdays.AI
LATAM

Data reduction - Feature extraction



- Creación de un **nuevo conjunto de características más pequeñas** que aún captura la mayor parte de la información útil.

Saturdays.AI
LATAM

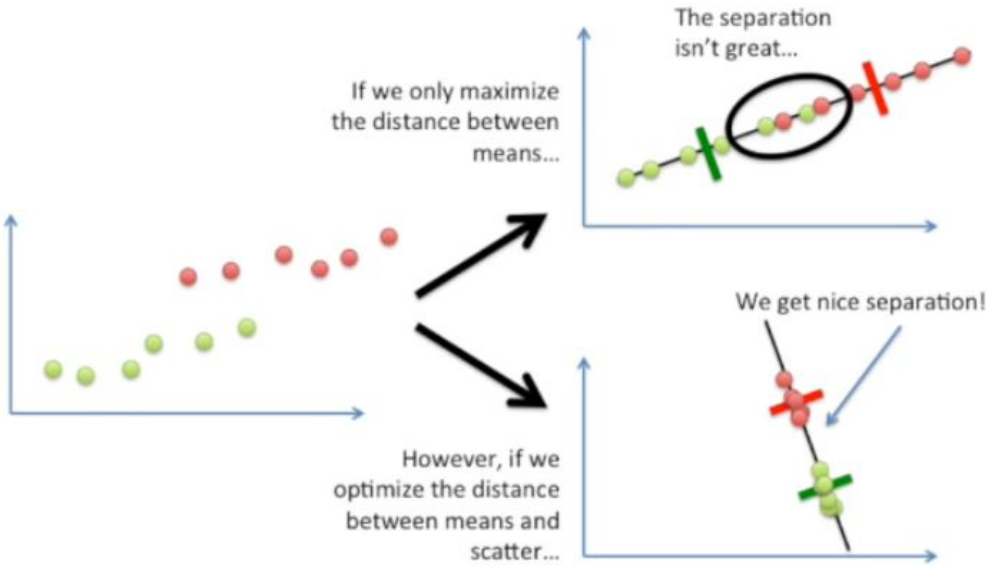
Feature extraction - Principal Component Analysis (PCA)

- Crea combinaciones lineales de las características originales.
- Las nuevas características son ortogonales = que no están correlacionadas.
- No supervisado.
- Los nuevos componentes se ordenan según la máxima varianza: PC1 explica la mayor variación en su conjunto de datos, PC2 explica la segunda variación más, ...

Saturdays.AI
LATAM

Feature extraction - Linear Discriminant Analysis (LDA)

- Crea combinaciones lineales de sus características originales.
- PCA vs LDA
  - LDA es supervisado
  - LDA maximiza la separabilidad entre clases.



Fuente: https://www.youtube.com/watch?v=azXCzI57Yfc

Saturdays.AI LATAM


Time to Code!!! :D
Saturdays.AI
LATAM

#3.4 Conclusión

Saturdays.AI
LATAM

¿Que técnica usar?
¿ Cuál es mejor?

"No Free Lunch" theorem

En resumen …. Ningún algoritmo funciona mejor para cada problema.

Solución: Prueba muchos algoritmos diferentes para tu problema, usa un "test set" de datos para evaluar el rendimiento y seleccionar al ganador.

Saturdays.AI
LATAM

Contacto

María Inés Calderón Zetter

ines@saturdays.ai

https://www.linkedin.com/in/ineszetter/

Favio Vázquez

favio@closter.net

https://www.linkedin.com/in/faviovazquez/

Lea Vega Romero

lea@saturdays.ai

https://www.linkedin.com/in/lea-vega-66a18011b/

Saturdays.AI
LATAM