

Introducción a la Ciencia de Datos

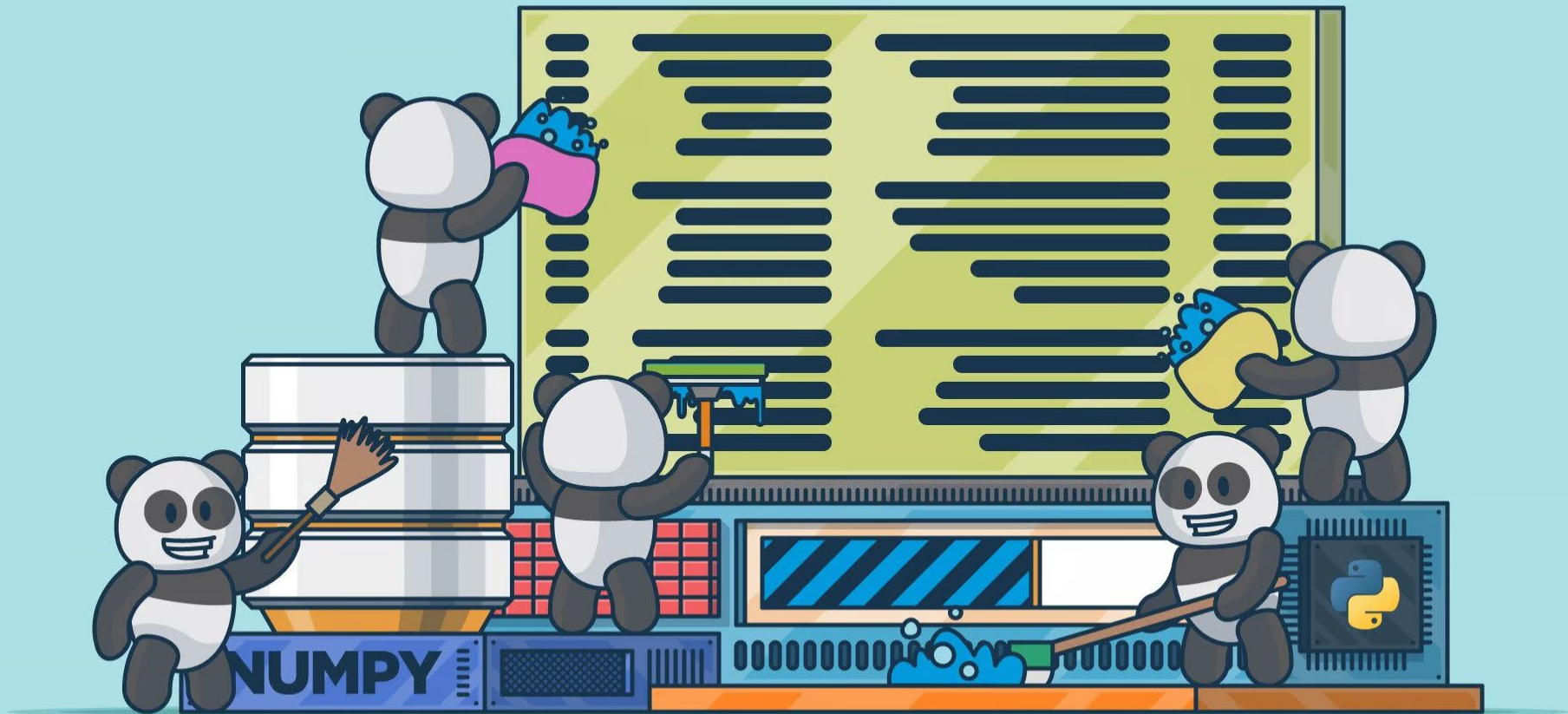
Maestría en Ciencias
de la Computación

Dr. Irvin Hussein López Nava

The slide features several decorative squares in various shades of blue and white. A large blue square with the number '02' is centered near the top. Other smaller squares are scattered around the slide, including a light blue square in the top left, a dark blue square in the top right, a 2x2 grid of squares (white, white, dark blue, white) to the right of the '02' square, a light blue square in the bottom left, a dark blue square in the bottom right, and a light blue square in the bottom right corner. A 2x2 grid of squares (white, white, white, white) is also visible in the top right area.

02

Procesamiento



Real Python

2.1 Limpieza de datos

Let's keep talking

(and listening)

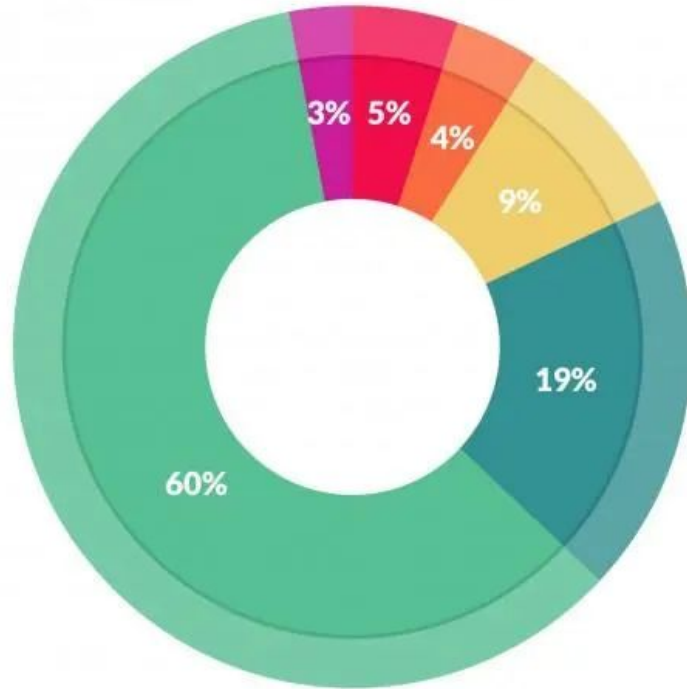
about

Data...

but cleaning



What data scientists spend the most time doing



- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

¿Qué es la limpieza de los Datos?



Definiciones

ChatGPT

Es el proceso de identificar y corregir errores y problemas en conjuntos de datos para garantizar su precisión y confiabilidad en el análisis y la toma de decisiones en ciencia de datos.

<https://chat.openai.com/>

Wikipedia

Es el proceso de detección y corrección de registros corruptos o inexactos de un conjunto de registros, y se refiere a la identificación de partes incompletas, incorrectas, inexactas o irrelevantes de los datos y, a continuación, la sustitución, modificación o eliminación de los datos sucios.

https://en.wikipedia.org/wiki/Data_cleansing

Gemini

Es un proceso fundamental en el manejo de información que consiste en identificar y corregir o eliminar registros de datos que son erróneos, incompletos, inconsistentes o duplicados. Es útil para asegurar que sean lo más precisos y confiables posible.

<https://gemini.google.com/>

La limpieza de datos

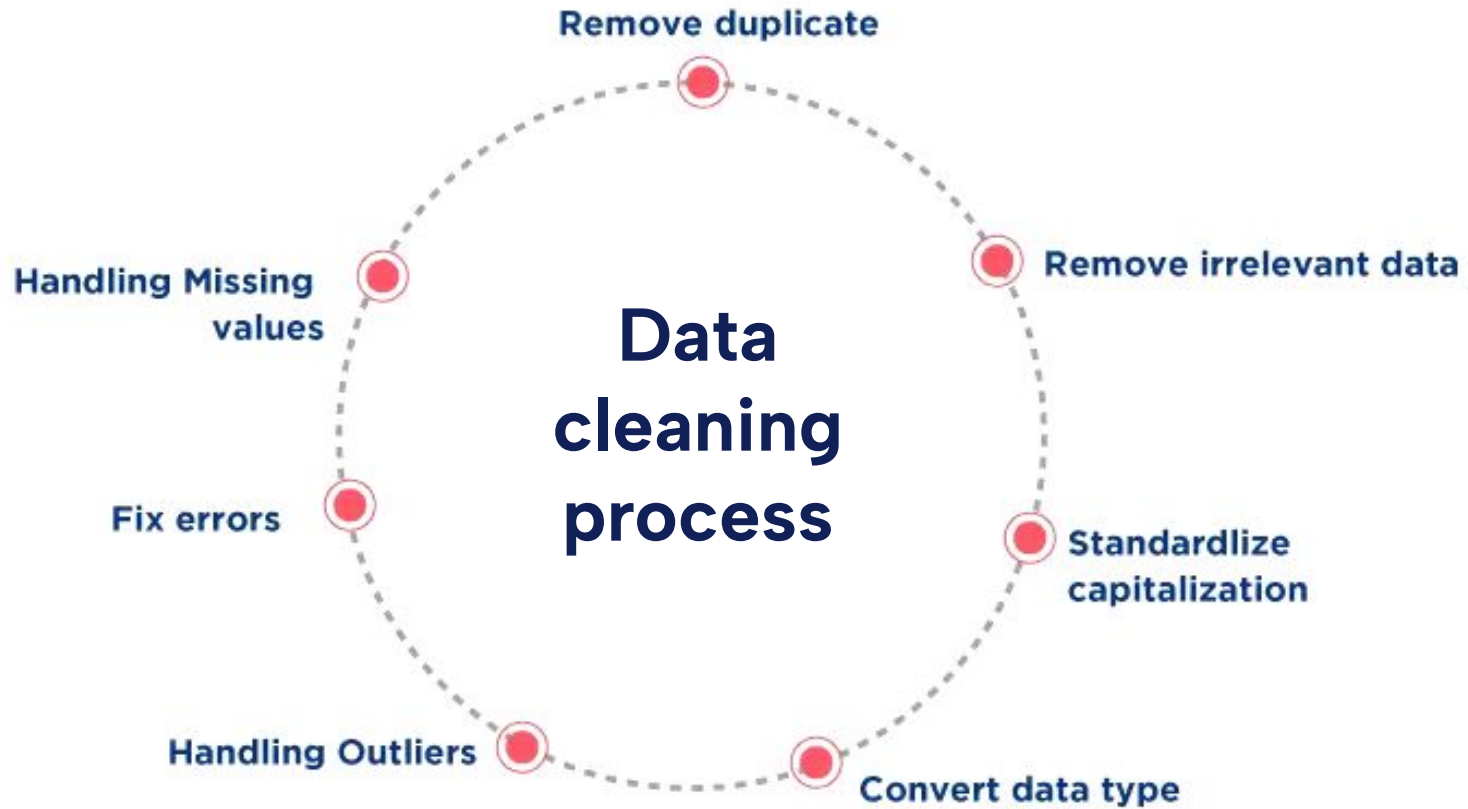
- Una vez completado el proceso de **ciencia de datos**, si los resultados no son satisfactorios, las dos cosas que pueden estar fallando son: (i) los **datos**, o (ii) los **modelos**.
- Elegir los **datos** adecuados es el primer paso, y después viene el **formato** y la representación de los datos.
 - Es importante estar seguros que los **datos** sobre los que se ha realizado el análisis están libres de cualquier tipo de incorrección.
- Como se ha mencionado, los **datos** del **mundo real** son desordenados.
 - Contienen faltas de ortografía, valores incorrectos, valores de datos irrelevantes, o valores faltantes.

¿Por qué es importante?

- **Calidad de Datos:** Datos limpios garantizan que los resultados del análisis sean precisos y confiables.
- **Toma de Decisiones:** Datos inexactos pueden llevar a decisiones erróneas, lo que puede ser costoso.
- **Eficiencia:** La limpieza de datos adecuada ahorra tiempo en las etapas posteriores del análisis.
- **Integración de Datos:** Es necesario tener datos limpios para combinar información de diferentes fuentes.

Un primer ejemplo

- Considere un conjunto de datos que incluye la columna de sexo.
 - Si los datos se rellenan manualmente, existe la posibilidad de que la columna de datos contenga registros de 'male' 'female', 'M', 'F', 'Male', 'Female', 'MALE', 'FEMALE', etc.
 - Al analizar esta columna, todos estos valores se considerarán distintos. Pero, en realidad, 'Male', 'M', 'male', and 'MALE' se refieren a lo mismo.
- Otros registros incluyen fechas, unidades de medida, y por supuesto, su respectiva codificación.
- En la etapa de **limpieza de datos** se requiere identificar los formatos incorrectos y corregirlos.



1. Eliminar duplicados

- Al trabajar con grandes conjuntos de datos, o incluso múltiples fuentes de datos, es probable que los **datos** incluyan valores duplicados.
- Estos valores duplicados añaden redundancia y pueden hacer que el análisis sea incorrecto. Los números duplicados en un conjunto de datos darán un recuento mayor que el número real.
- Por otro lado, en algunas circunstancias se requiere tal duplicidad, ya que el objetivo podría ser estudiar la ocurrencia de eventos.

Cadena de búsqueda: exergam* OR "rehabilitation gam*" OR "active video gam*" OR "physical gam*" OR "virtual reality exercis*" OR freegam*



20 first entries sort by: citations: highest first

Authors	Article Title	Source Title	Cited by
Biddiss E.; et al.	Active video games to promote physical activity in children and youth: A systematic review	Archives of Pediatrics and Adolescent Medicine	396
Althoff T.; et al.	Influence of pokémon go on physical activity: Study and implications	Journal of Medical Internet Research	334
Anderson-Hanley C.; et al.	Exergaming and older adult cognition: A cluster randomized clinical trial	American Journal of Preventive Medicine	305
Maillot P.; et al.	Effects of interactive physical-activity video-game training on physical and cognitive function in older adults	Psychology and Aging	277
Peng W.; et al.	Is playing exergames really exercising? A meta-analysis of energy expenditure in active video games	Cyberpsychology, Behavior, and Social Networking	275
Graf D.L.; et al.	Playing active video games increases energy expenditure in children	Pediatrics	272
Sinclair J.; et al.	Considerations for the design of exergames	Proceedings - 5th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia	257
Greenberg B.S.; et al.	Orientations to video games among gender and age groups	Simulation and Gaming	256
Staiano A.E.; et al.	Exergames for Physical Education Courses: Physical, Social, and Cognitive Benefits	Child Development Perspectives	255
Hamari J.; et al.	Social motivations to use gamification: An empirical study of gamifying exercise	ECIS 2013 - Proceedings of the 21st European Conference on Information Systems	247

20 first entries sort by: citations: highest first

Authors	Article Title	Source Title	Cited by
Ijaz, K; et al.	Player Experience of Needs Satisfaction (PENS) in an Immersive Virtual Reality Exercise Platform Describes Motivation and Enjoyment	International Journal Of Human-Computer Interaction	998
Biddiss, E; et al.	Active Video Games to Promote Physical Activity in Children and Youth A Systematic Review	Archives Of Pediatrics & Adolescent Medicine	355
Althoff, T; et al.	Influence of Pokemon Go on Physical Activity: Study and Implications	Journal Of Medical Internet Research	270
Anderson-Hanley, C; et al.	Exergaming and Older Adult Cognition A Cluster Randomized Clinical Trial	American Journal Of Preventive Medicine	261
Peng, W; et al.	Is Playing Exergames Really Exercising? A Meta-Analysis of Energy Expenditure in Active Video Games	Cyberpsychology Behavior And Social Networking	247
Maillot, P; et al.	Effects of Interactive Physical-Activity Video-Game Training on Physical and Cognitive Function in Older Adults	Psychology And Aging	241
Graf, DL; et al.	Playing Active Video Games Increases Energy Expenditure in Children	Pediatrics	237
Staiano, AE; et al.	Exergames for Physical Education Courses: Physical, Social, and Cognitive Benefits	Child Development Perspectives	233
Hammami, A; et al.	Physical activity and coronavirus disease 2019 (COVID-19): specific recommendations for home-based physical training	Managing Sport And Leisure	224
Rosenberg, D; et al.	Exergames for Subsyndromal Depression in Older Adults: A Pilot Study of a Novel Intervention	American Journal Of Geriatric Psychiatry	210

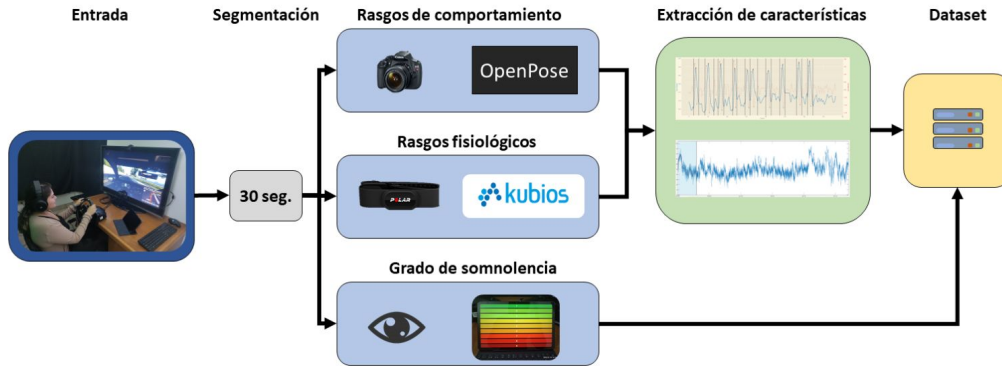
2. Eliminar datos irrelevantes

- Imagine analizar el servicio posventa de un producto. Los **datos** contienen varios campos como la fecha de solicitud de servicio, el número único de solicitud de servicio, el número de serie del producto, el tipo de producto, la fecha de compra del producto, etc.
- Aunque los campos parecen **relevantes**, los **datos** también pueden contener otros campos como: atendido por (nombre de la persona que inició la solicitud de servicio), ubicación del centro de servicio, detalles de contacto del cliente, etc., que podrían no ser útiles. Comúnmente, es la comprobación a nivel columna (atributos) que se realiza inicialmente.

2. Eliminar datos irrelevantes

- A continuación, se comprueban las filas (instancias). Supongamos que el cliente visitó el centro de servicio y se le pidió que volviera al cabo de 3 días para recoger el producto reparado. En este caso, habría dos registros diferentes que representan el mismo número de servicio.
 - Para el primer registro, el tipo de servicio es "primera visita" y el tipo de servicio es "recogida" para el segundo registro. Se puede agregar un campo nuevo para combinar la información, en este problema el estado de la solicitud.
- Para eliminar datos irrelevantes durante esta etapa, es importante comprender tales **datos** y el planteamiento del problema.

desde otra perspectiva, considerar solo datos relevantes



Característica (No. de apariciones)		
Fisiológicos	Comportamiento	Múltiple
SD HR (6)	Izq. promedio Eye Closure (6)	Izq. promedio Eye Closure (6)
PNS Index (6)	Der. promedio Eye Closure (6)	Der. promedio Eye Closure (6)
RMSSD (6)	Izq. SD Eye Closure (6)	Der. PERCLOS (5)
Max. RR (5)	Der. PERCLOS (5)	Izq. PERCLOS (5)
AR poder absoluto AF ms2 (5)	Izq. PERCLOS (5)	Max. RR (4)
Promedio RR (5)	Der. min. Eye Closure (5)	Der. min. Eye Closure (4)
Poincaré SD1 (5)	Der. SD Eye Closure (4)	PNS Index (4)
Min. RR (5)	Der. parpadeos (4)	Promedio HR (4)
AR poder relativo LF pct (5)	Izq. duración min. de parpadeo (3)	Promedio RR (3)
AR LF frq (5)	Izq. tiempo promedio al cerrado del ojo (3)	Der. parpadeos (3)
NN50 (5)	Izq. tiempo mínimo a la apertura del ojo (3)	Izq. SD Eye Closure (2)
AR LF HF ratio (4)	Izq. velocidad promedio de la apertura del ojo (3)	Izq. mín. Eye Closure (2)
FFT poder absoluto MBF ms2 (4)	Der. tiempo mínimo a la apertura del ojo (2)	Der. SD Eye Closure (2)
SNS Index (4)	Izq. tiempo promedio a la apertura del ojo (2)	NN50 (2)
Promedio HR (4)	Izq. mín. Eye Closure (2)	AR AF frq (2)

3. Estandarización

- Asegurarse de que el **texto** de los **datos** es coherente. Lo mismo aplica para **datos numéricos**.
- Por ejemplo: tener como nombre de columna "Total-ventas" y "total ventas" es diferente (pese a que la mayoría de los lenguajes de programación ya distinguen entre mayúsculas y minúsculas).
- Se puede optar por el uso de estándares definidos.
 - Por ejemplo, **cobra case** es un estilo de escritura en el que la primera letra de cada palabra se escribe en mayúscula, y cada espacio se sustituye por el caracter de subrayado (_).

codecase my great variable name



myGreatVariableName

Camel Case

⌘1



my/great/variable/name

Slash Case

⌘2



my_great_variable_name

Snake Case

⌘3



MyGreatVariableName

Pascal Case

⌘4



My_Great_Variable_Name

Cobra Case

⌘5



my.great.variable.name

Dot Case

⌘6



my-great-variable-name

Dash Case

⌘7



my great variable name

Separate Words

↩

4. Conversión de tipo de datos

- Algunos lenguajes intentan adivinar los **tipos de datos**; en su mayor parte, tiene éxito, pero ocasionalmente hay que brindar un poco de ayuda.
- Los **tipos de datos** más comunes en **bases de datos** son los tipos de texto, numéricos y de fecha.
 - Como texto se pueden aceptar valores mixtos, incluidos alfabetos, 'dígitos' o incluso caracteres especiales.
 - Los tipos numéricos contienen valores enteros o números de punto flotante, y las operaciones matemáticas dependen de su consistencia.
 - En algunos casos, datos numéricos deben expresarse como textos.
 - Las fechas se pueden representar en diferentes formatos: 2023/09/12, 12 de septiembre de 2023, 12-09-2023, 1694494930000.

equivalencia entre tipos de datos en python

Pandas dtype	Python type	NumPy type	Usage
object	str or mixed	string_, unicode_, mixed types	Text or mixed numeric and non-numeric values
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values




esto también incluye los tipos de archivos

Data Handling

 python



What is an SQL query ? How to handle data stored in a database ?

	How to load data as pandas dataframe from your excel file stored locally ?
	How to handle data stored in CSV file?
	How to load data from a JSON file ?

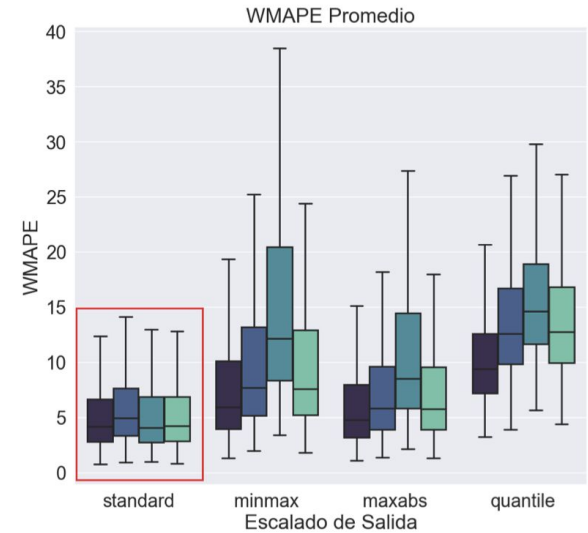
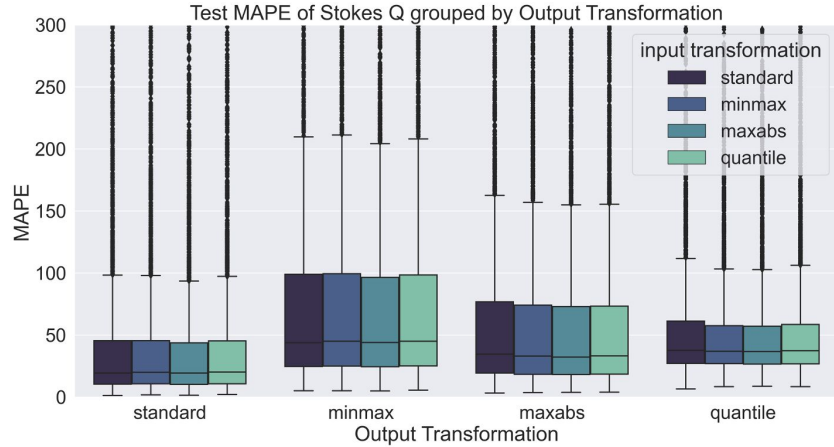
5. Manejo de valores atípicos

- Un **valor atípico** (*outlier*) es un **dato** que, estadísticamente, se desvía considerablemente de otras observaciones.
 - Ahora bien, podría reflejar variabilidad de la medición, pero también un error experimental.
- Los **valores atípicos** pueden identificarse con un análisis exploratorio, mediante un diagrama de caja o un diagrama de dispersión.
- Los **valores atípicos** dan como resultado **datos sesgados**.
 - Hay modelos que suponen que los **datos** siguen una distribución normal y los *outliers* pueden afectar el rendimiento de los modelos si los datos están sesgados.

5. Manejo de valores atípicos

- Hay dos formas comunes para manejar los **valores atípicos**.
 1. Eliminar las observaciones.
 2. Aplicar transformaciones como logaritmo, raíz cuadrada, box-cox, etc., para que los valores sigan una distribución normal, o casi normal.
- **Nota:** se debe considerar que al aplicar cualquier técnica, se puede derivar de igual forma en resultados sesgados.

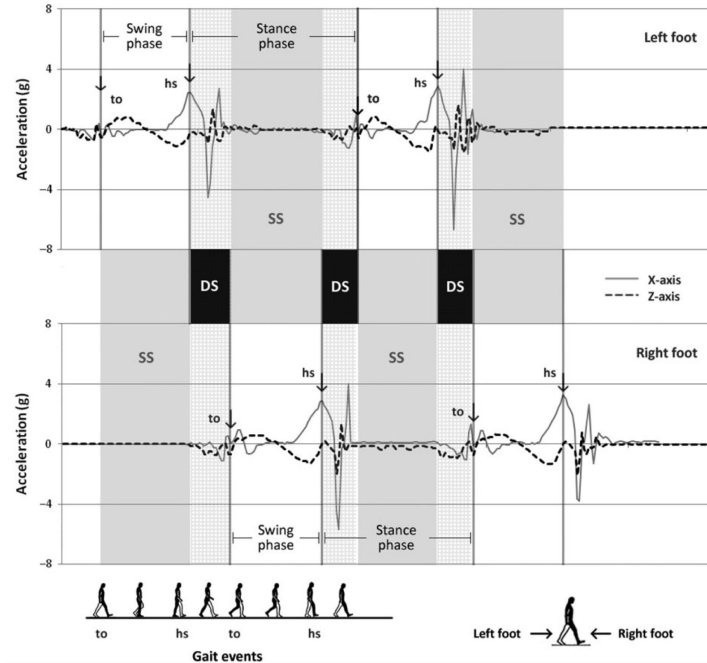
en ocasiones puede aplicarse con fines de visualización



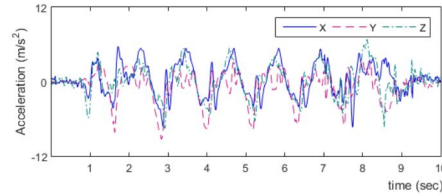
6. Corrección de errores

- Errores en los **datos** pueden hacer que se pierdan los hallazgos claves en el estudio.
 - Los sistemas que ingresan datos manualmente sin ningún tipo de verificación de datos casi siempre contendrán errores.
- Algunas validaciones de entrada para asegurar los datos:
 - Restringir (*a priori*) los valores esperados de entrada, e.g., 10 dígitos para el teléfono, y 3 más para identificar el país.
 - Realizar comprobaciones de validación (*a posteriori*) como que la fecha de compra debe ser mayor a la fecha de fabricación, el importe total debe ser igual a la suma de los demás importes, cualquier caracter especial que se encuentre en un campo que no lo permita, etc.

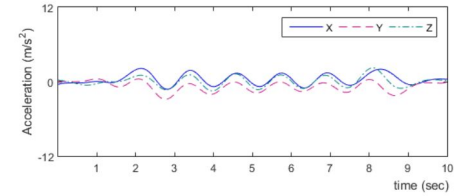
pero, tener cuidado!!



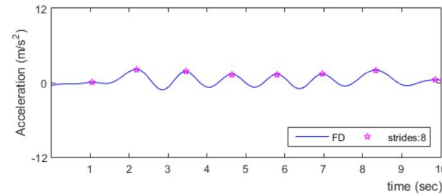
Eliminar patrones de interés



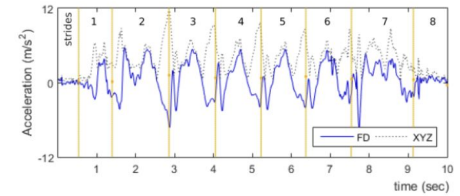
(a) Raw acceleration signals.



(b) Filtered and smoothed signals.



(c) Detection of strides.



(d) Segmentation of signals.

Uso en tareas específicas, e.g., segmentación

7. Traducción

- Algunos **conjuntos de datos** combinan información de varias fuentes, incluso en diferentes idiomas, lo que puede dar lugar a discrepancias lingüísticas.
- Los modelos de procesamiento del lenguaje natural (PLN), no pueden procesar más de un idioma, i.e., son monolingües.
- Por lo tanto, se estudia un solo idioma a la vez, y se requiere traducir todos los **datos** a un solo idioma.
- Actualmente, existen modelos robustos para la traducción automática.

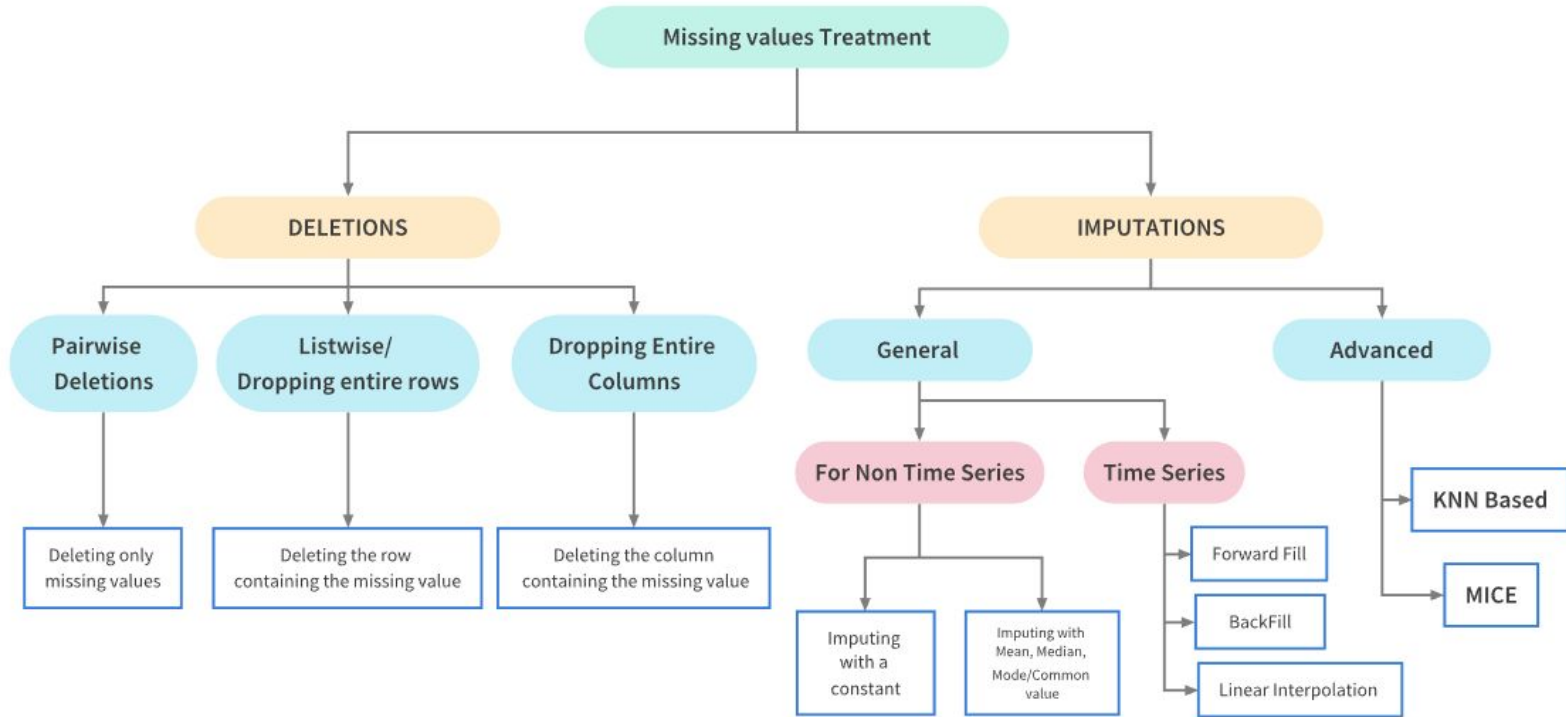
la nueva guerra silenciosa, por los datos



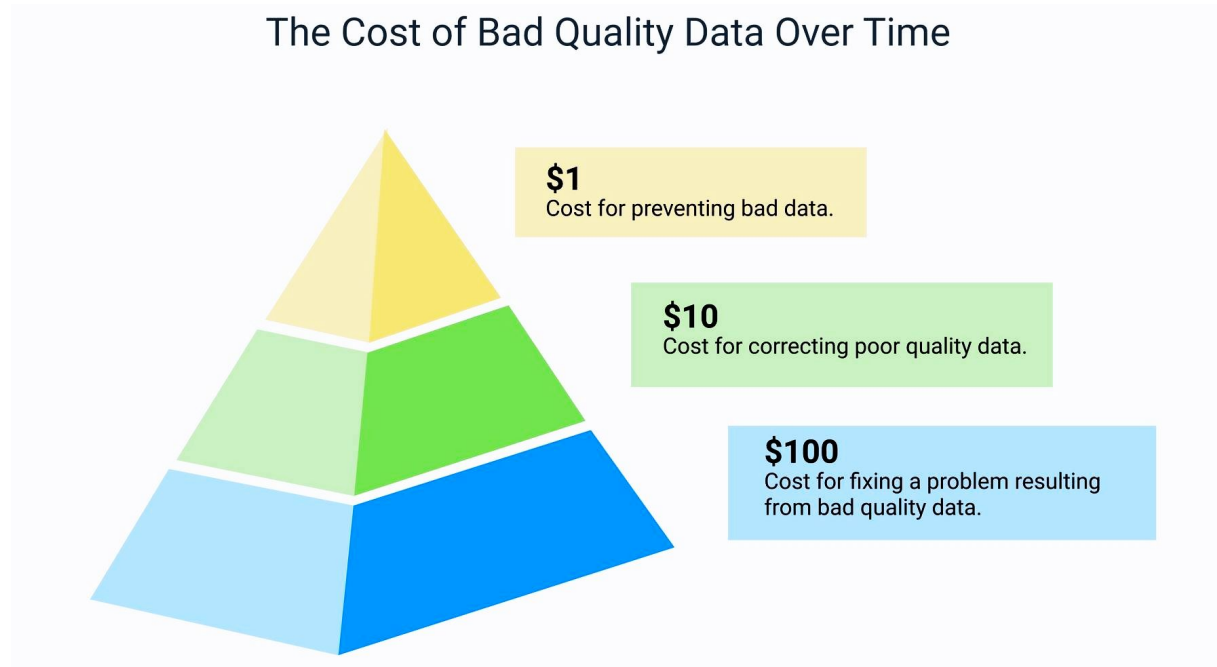
8. Datos incompletos

- Los **datos** de la **vida real** pueden contener **valores faltantes** que requieren un procesamiento particular. Dos enfoques distintos:
 - Eliminar los registros completos a los que les faltan valores, o
 - Completar los valores faltantes utilizando alguna técnica estadística o recopilando datos.
- Una regla general es eliminar los registros incompletos si representan menos del 5% del total, pero depende del tipo de análisis la importancia de los valores faltantes, e.g., en datos desbalanceados.

ideas para hacerlo con Python



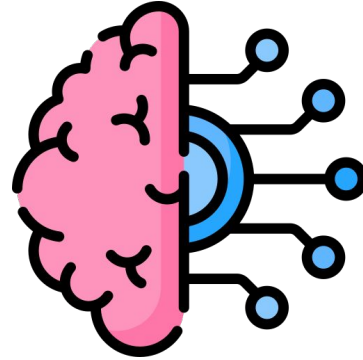
Costo de la baja calidad de los datos



Caso de estudio

¿es un mensaje con contenido sexista?

nadie ir tratar tanto bien
hombre querer meter primero
vez



EXIST dataset: sEXism Identification in Social neTworks

- El sexismo se puede definir como “prejuicio, estereotipo o discriminación, típicamente contra las mujeres, por motivos de sexo”.
- Detectar el sexismo en línea puede resultar difícil, ya que puede expresarse de formas muy diferentes.
 - El sexismo puede parecer “amistoso”: la afirmación “Las mujeres deben ser amadas y tratadas siempre como un cristal frágil” puede parecer positiva, pero en realidad es considerar que son más débiles que los hombres.
 - El sexismo puede sonar “gracioso”, como es el caso de los chistes o el humor sexistas (“Hay que amar a las mujeres... sólo eso... Nunca las entenderás”).
 - El sexismo puede sonar “ofensivo” y “odioso”.

Task 1: Identification de sexismo

- La primera subtarea es una clasificación **binaria**: decidir si un texto determinado (tuit) es sexista (i.e., es sexista si describe una situación sexista o critica un comportamiento sexista). Ejemplos:
 - SEXIST:
 - "Mujer al volante, tenga cuidado!"
 - "People really try to convince women with little to no ass that they should go out and buy a body. Like bih, I don't need a fat ass to get a man. Never have."
 - NOT SEXIST:
 - "Alguien me explica que zorra hace la gente en el cajero que se demora tanto."
 - "@messyworldorder it's honestly so embarrassing to watch and they'll be like 'not all white women are like that'"

Diferentes niveles de limpieza

Crudo

Nadie te va a tratar tan bien como un hombre que te lo quiere meter por primera vez.

Limpieza

nadie te va a tratar tan bien como un hombre que te lo quiere meter por primera vez

Tokenización

['nadie', 'te', 'va', 'a', 'tratar', 'tan', 'bien', 'como', 'un', 'hombre', 'que', 'te', 'lo', 'quiere', 'meter', 'por', 'primera', 'vez']

StopWords

['nadie', 'va', 'tratar', 'tan', 'bien', 'hombre', 'quiere', 'meter', 'primera', 'vez']

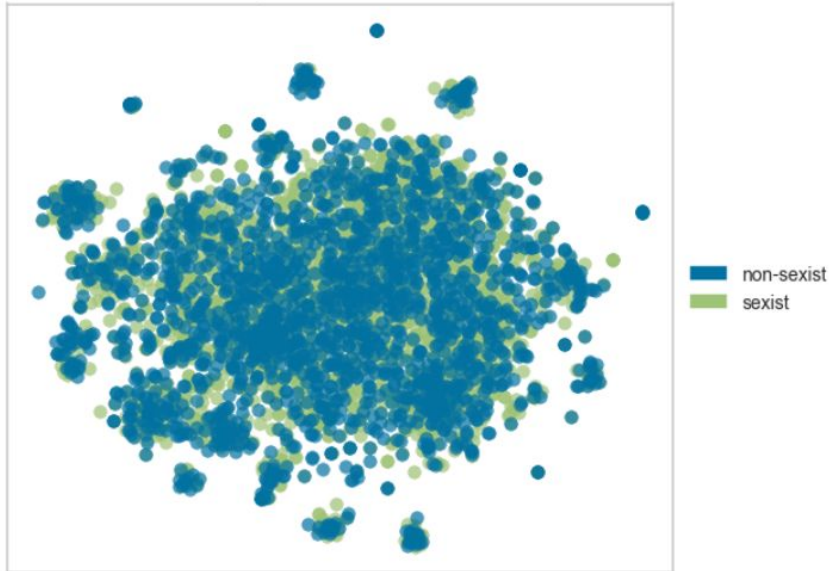
Lematización

nadie ir tratar tanto bien hombre querer meter primero vez

Exploración de las representaciones

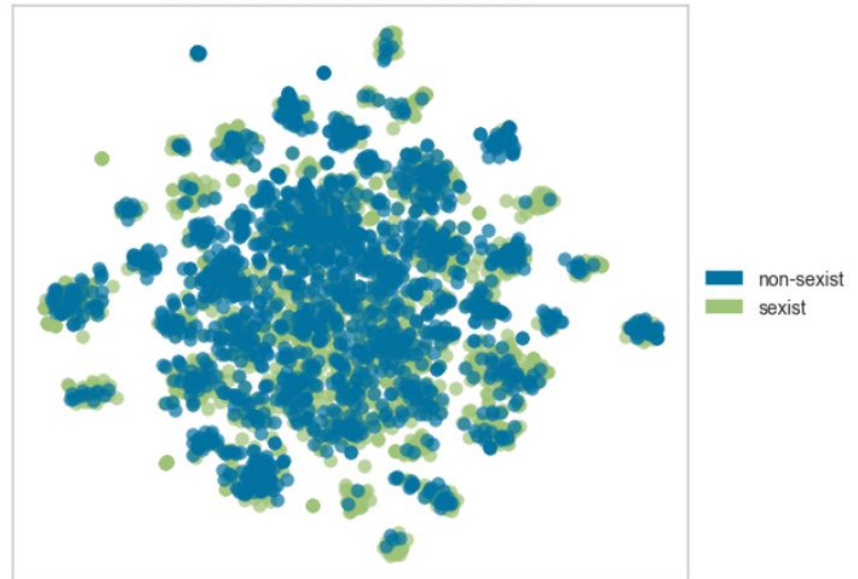
T-SNE datos crudos

TSNE Projection of 3541 Documents



T-SNE datos lematizados

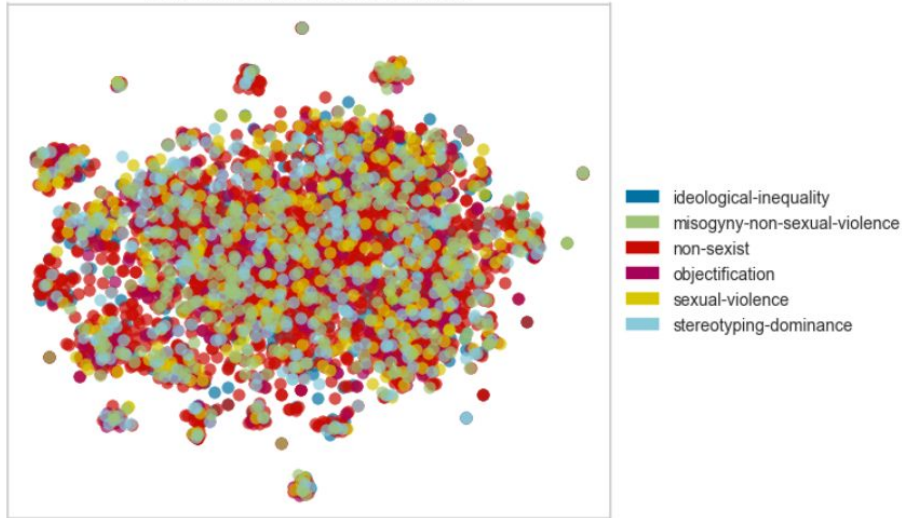
TSNE Projection of 3538 Documents



Exploración de las representaciones

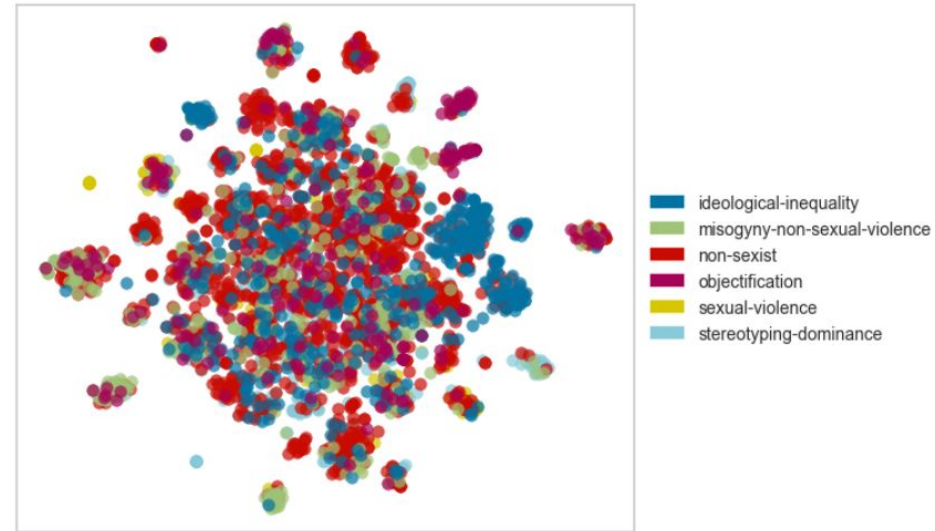
T-SNE datos crudos

TSNE Projection of 3541 Documents

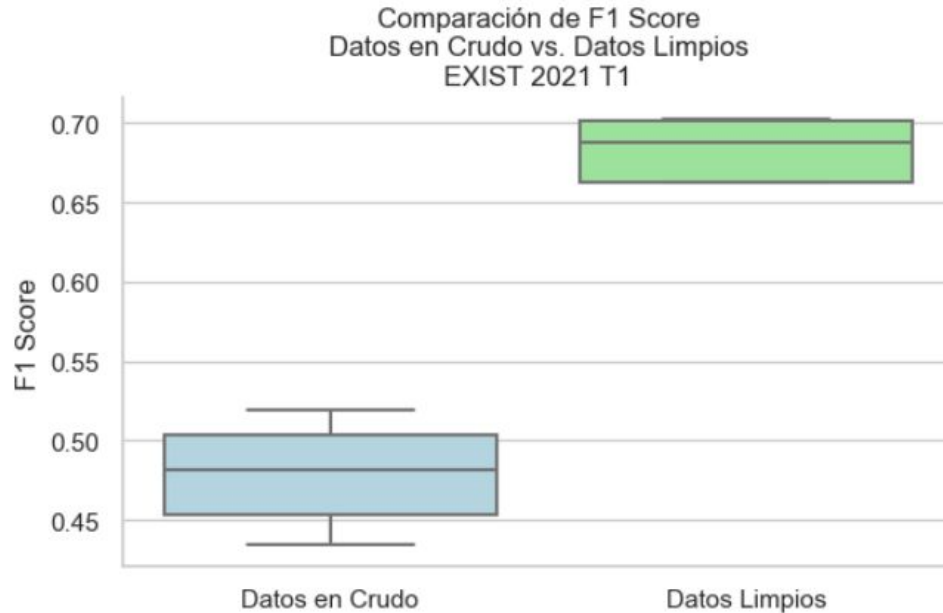


T-SNE datos lematizados

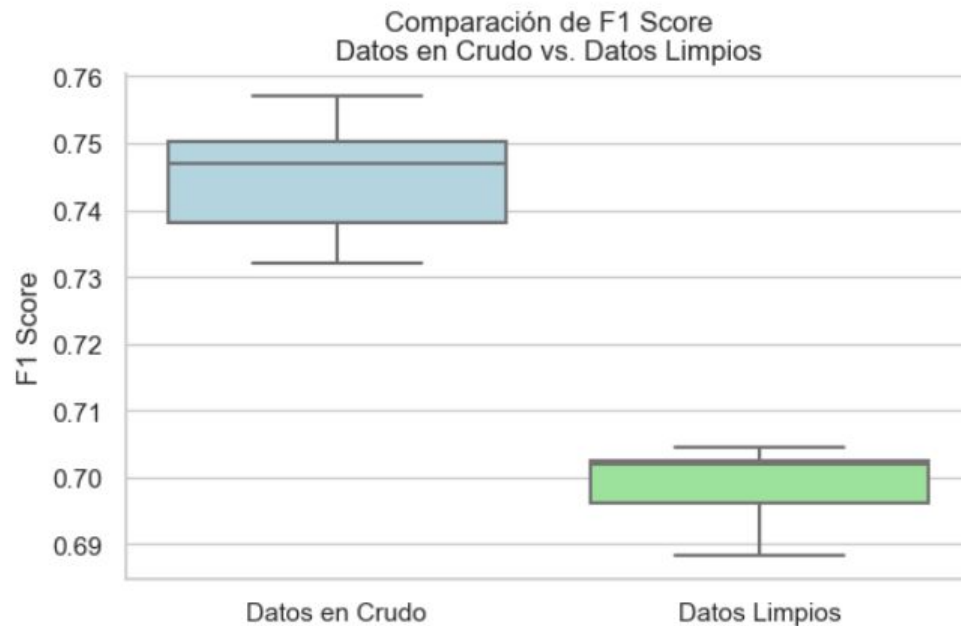
TSNE Projection of 3538 Documents



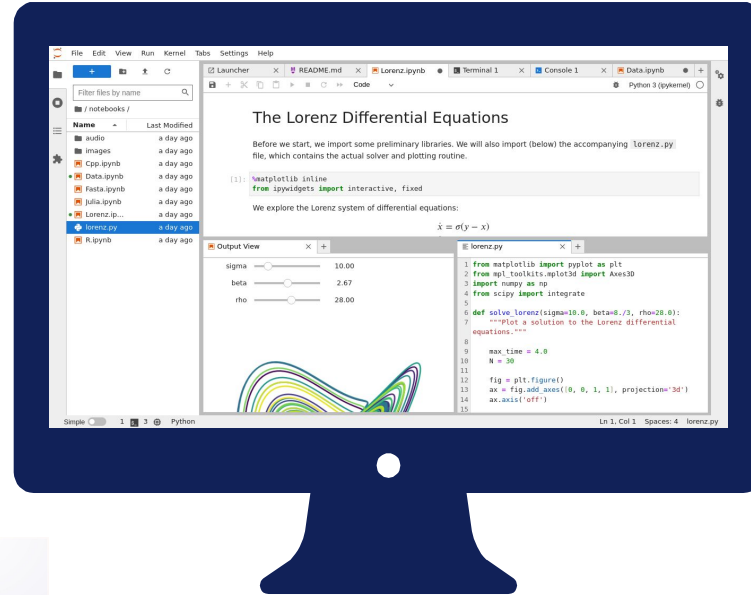
Bag of words + Bayesian Network



BETO transformer



(Go to live notebook)



Gracias!

¿Alguna pregunta?

hussein@cicese.mx

<https://sites.google.com/view/husseinlopeznava>



CREDITS: This presentation was based on a template by Slidesgo, and includes icons by Flaticon.