



Preprocesamiento de Datos

Tutoría

¿Qué es una serie de tiempo? ¿Cuáles son sus componentes?



¿En qué consiste un
modelo ARIMA? ¿Cómo
se determinan los
parámetros?



¿Cómo evaluamos un
modelo de serie de
tiempo?



Autoaprendizaje

Recursos asincrónicos

- ¿Revisaste los recursos de la semana 7 (Guía y desafío)?
- ¿Tienes dudas sobre alguno de ellos?



Ideas fuerza



Las **metodologías** son útiles para **estandarizar y utilizar las mejores** prácticas al momento de trabajar.



Nos centramos en la metodología **CRISP-DM** principalmente en los primeros pasos de entendimiento del negocio, de la data y preparación de los datos .



Para cada etapa hay varios **estadísticos y métodos** que son útiles, pero lo importante es analizar caso a caso que es lo **mejor para mi problema.**

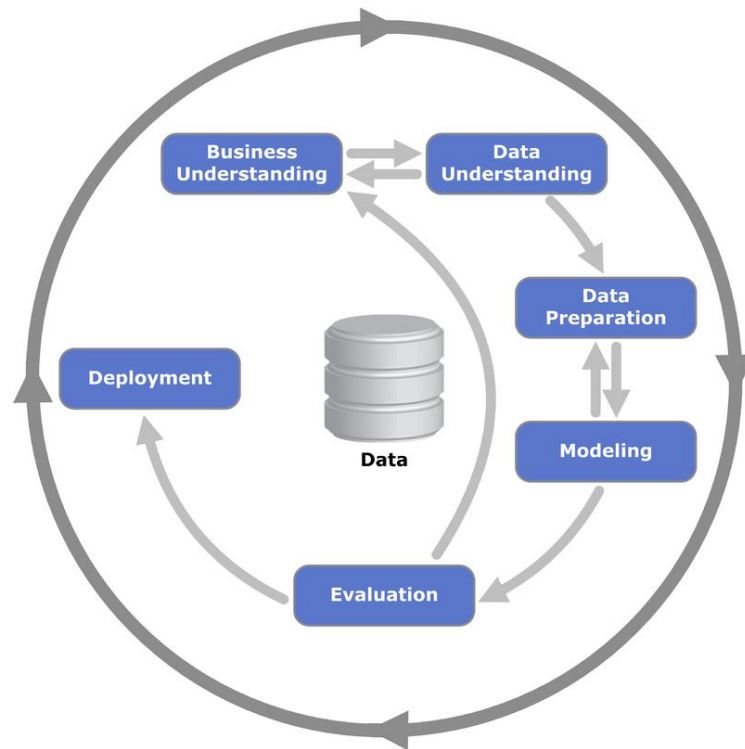
/* Metodología de Proyectos de Data Science */

Metodología de proyectos Data Science

CRISP DM

¿Por qué utilizar un marco de trabajo como CRISP DM?

1. Estructura Clara
2. Orientación al Negocio
3. Flexibilidad



/*Entendimiento y preparación de datos*/

Entendimiento de los datos

CRISP DM

Recopilación de datos

Identificar las diferentes fuentes de datos.

Adquisición de datos

Calidad de datos

Identificar valores faltantes, valores duplicados u otros problemas de datos.

Identificar outliers.

Concluir si la calidad de los datos permite modelar y primer filtro de variables.

Exploración de datos

Análisis univariado y multivariado.

Análisis entre variables y correlaciones.

Definir puntos importantes para la siguiente etapa y se itera.

Muestreo de datos

En caso de ser necesario se puede muestrear los datos para trabajarlos de mejor forma.

Asegurar que la muestra sea significativa.

Identificando Outliers

Métodos

1. **Método IQR (Inter Quartil Range):** consideramos solo los datos entre $Q1 - 1.5 \cdot IQR$ y $Q3 + 1.5 \cdot IQR$, donde $IQR = Q3 - Q1$
2. **Z - Score:** calcula la puntuación Z para cada punto de datos, que determina a cuántas desviaciones estándar se encuentra un valor del promedio (se suele considerar outlier un valor por sobre 2 o 3 desviaciones estándar)
3. **Métodos basados en ML:** hay algoritmos como **dbscan** que detectan outliers, o KNN, que buscan valores con una distancia significativa que se consideran outliers.
4. **Otras:** algoritmos de detección de anomalías como **svm one class** o **isolation forest**.



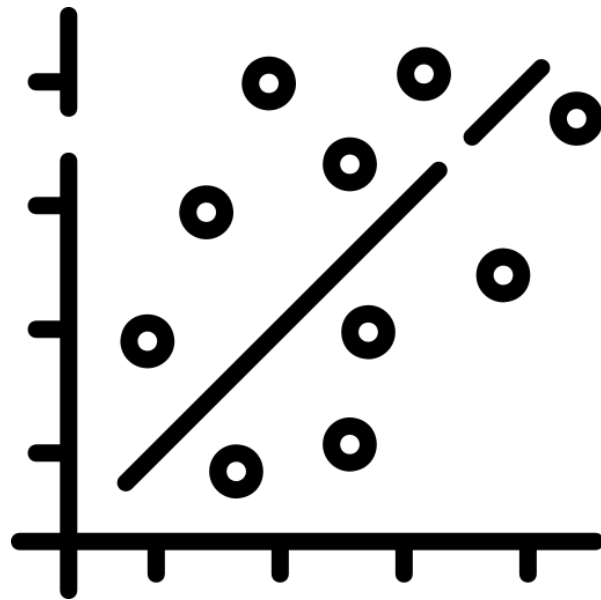
Correlaciones

Métodos de cálculo y detección

1. **Pearson:** mide la relación lineal entre variables numéricas

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

2. **Chi-Cuadrado:** mide la relación entre variables categóricas
3. **Test de Anova:** calcula si hay una diferencia en la variable numérica entre grupos de una categórica
4. **Otras:** kendall, spearman, v de cramer, test kolgomorov-smirnov, etc.



Coeficiente de Chi cuadrado

Categorica vs categorica

Frecuencias observadas		
	Acción	Romance
F	5	15
M	12	8

Frecuencias esperadas		
	Acción	Romance
F	8.5	11.5
M	8.5	11.5

Tabla chi-cuadrado		
	Acción	Romance
F	$(5 - 8.5)^2 / 8.5 = 1.441176$	$(8 - 11.5)^2 / 11.5 = 1,065217391$
M	$(12 - 8.5)^2 / 8.5 = 1.441176$	$(8 - 11.5)^2 / 11.5 = 1,065217391$

Test de ANOVA

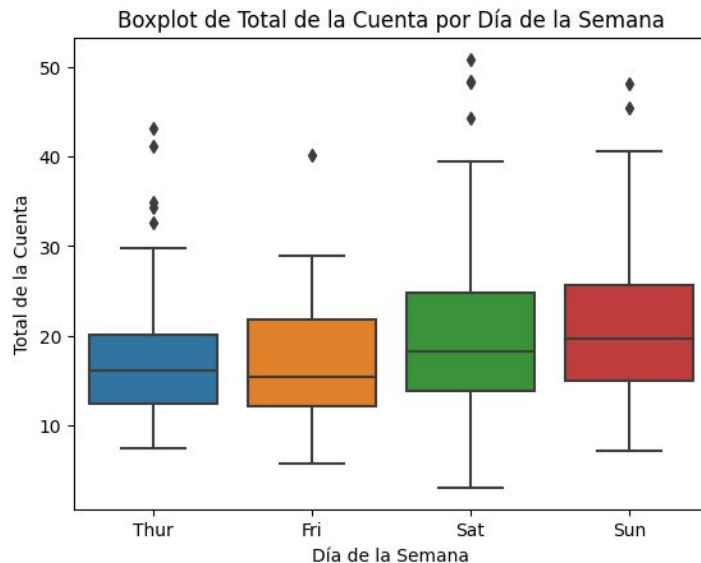
Numérica vs categórica

Es un test estadístico para ver si hay diferencia significativa entre diferentes grupos:

H0: No hay diferencia significativa

H1: Existe al menos un grupo diferente a los demás

No dice que grupo es diferente, pero con ayuda de un análisis posterior se puede observar.



¡Manos a la obra! Entendimiento y preparación de datos



Manos a la obra

Entendimiento y preparación de datos

Veremos ahora cómo aplicar estos procedimientos con la ayuda de Python. Para ello, sigue los pasos que te presentará tu profesor en el archivo Jupyter Notebook.

En esta ocasión veremos:

- Detección de outliers
- Correlaciones



/* Dimensionalidad y características*/

Dimensionalidad

La maldición de la dimensionalidad

La **maldición de la dimensionalidad** es un término utilizado en estadísticas y aprendizaje automático para describir los desafíos y problemas que surgen cuando trabajamos con conjuntos de datos de alta dimensionalidad, es decir, conjuntos de datos que tienen un gran número de características o variables en comparación con el número de observaciones.

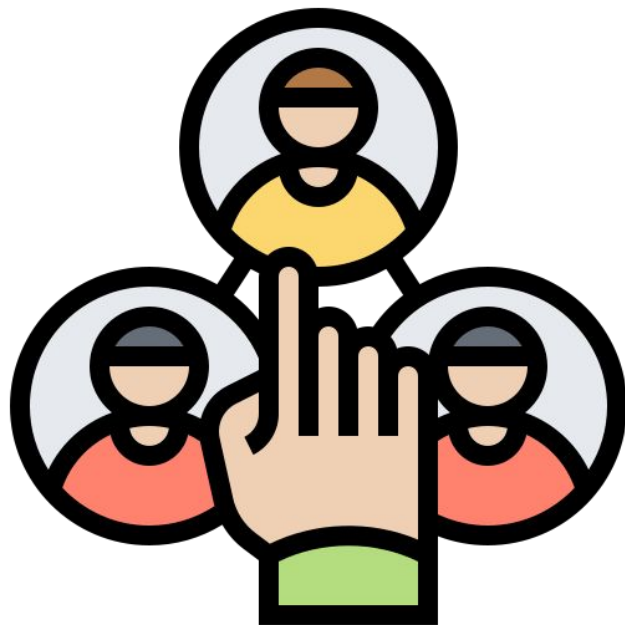
Algunos de los efectos de la maldición de la dimensionalidad son los siguientes:

- **Espacio de características disperso**
- **Requerimientos computacionales**
- **Sobreajuste**

Selección de características

Métodos de selección

- **Backward Selection:** se comienza con todas las características y se eliminan iterativamente con algún criterio de evaluación.
- **Forward Selection:** se comienza con un conjunto vacío y se agregan características de forma iterativa.
- **Métodos de filtro:** se evalúan características filtrando por algún criterio de correlación.
- **Otras:** Lasso, Random Forest



¡Manos a la obra! Entendimiento y preparación de datos



Manos a la obra

Entendimiento y preparación de datos

Veremos ahora cómo aplicar estos procedimientos con la ayuda de Python. Para ello, sigue los pasos que te presentará tu profesor en el archivo Jupyter Notebook.

En esta ocasión veremos:

- Forward selection
- Filtro por correlaciones
- Selección por Lasso



Desafío

"Preprocesamiento de datos"

- ¿Leíste el desafío de esta semana? ¿Comprendes bien lo que se solicita en cada caso?
- ¿Hay contenidos que necesitas repasar antes de comenzar este desafío?
- ¿Necesitas algún ejemplo o indicación para alguna pregunta o requerimiento específico?





Próxima sesión...

Continuaremos poniendo en práctica los conocimientos aprendidos y reforzando la etapa de modelamiento

{desafío}
latam_

*Academia de
talentos digitales*

