Universidad Popular del Cesar

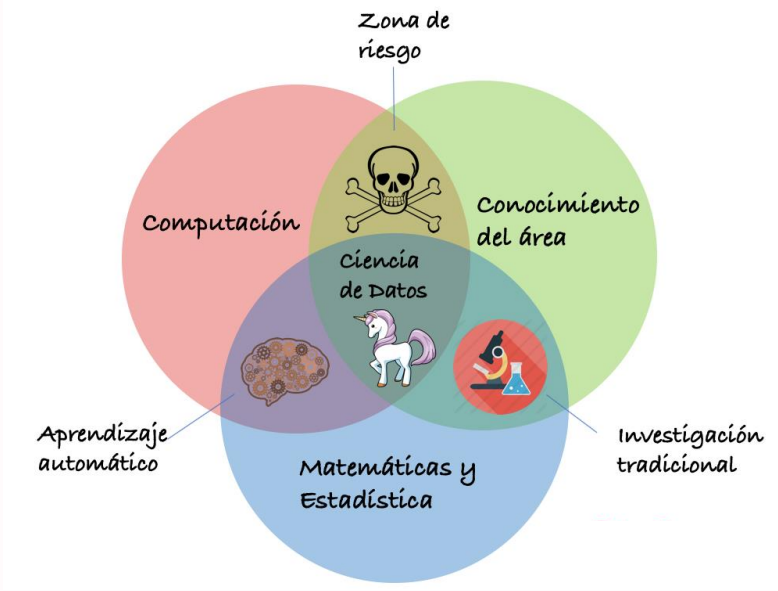
Procesamiento de Señales II
Ciencia de Datos II

Análisis y Visualización de Datos

Dr. José Ramón Iglesias
DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar

¿Qué es la Ciencia de Datos?

Conjunto de disciplinas que utiliza métodos científicos para obtener conocimientos a partir de los datos dándoles un mayor valor.



Statistical analysis and data reconfiguration

Data Analysis	Data Science	Machine Learning
Parte de preguntas concretas	Parte de una situación problemática	Parte de una tarea y un conjunto de datos
Busca explicar los datos para tomar decisiones	Busca un producto de datos	Busca optimizar una métrica de desempeño
Guiado por la intuición del analista	Guiado por interpretación de resultados	Guiado por la teoría de los modelos
Detecta patrones superficiales	Hace emerger patrones profundos	Detecta patrones profundos

Data Analysis

Explicar por qué se observa que los usuarios dejan de utilizar la plataforma pasados 6 meses

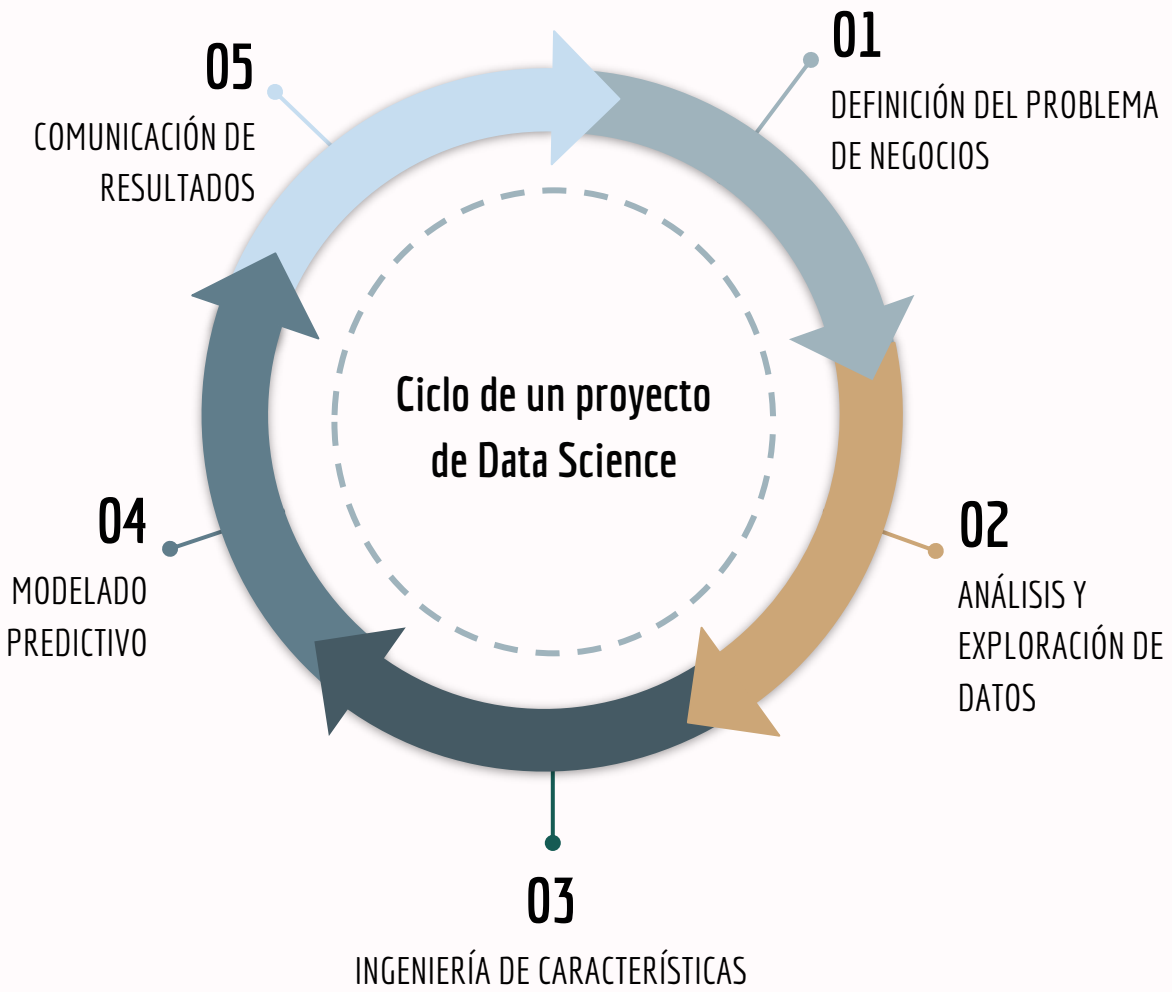
Data Science

Encontrar grupos de usuarios (segmentos) que tienen patrones de compras similares, para elegir distintas estrategias de marketing

Machine Learning

Clasificar usuarios en “high value” vs “low value” de acuerdo a cuánto se predice que gastarán en la plataforma

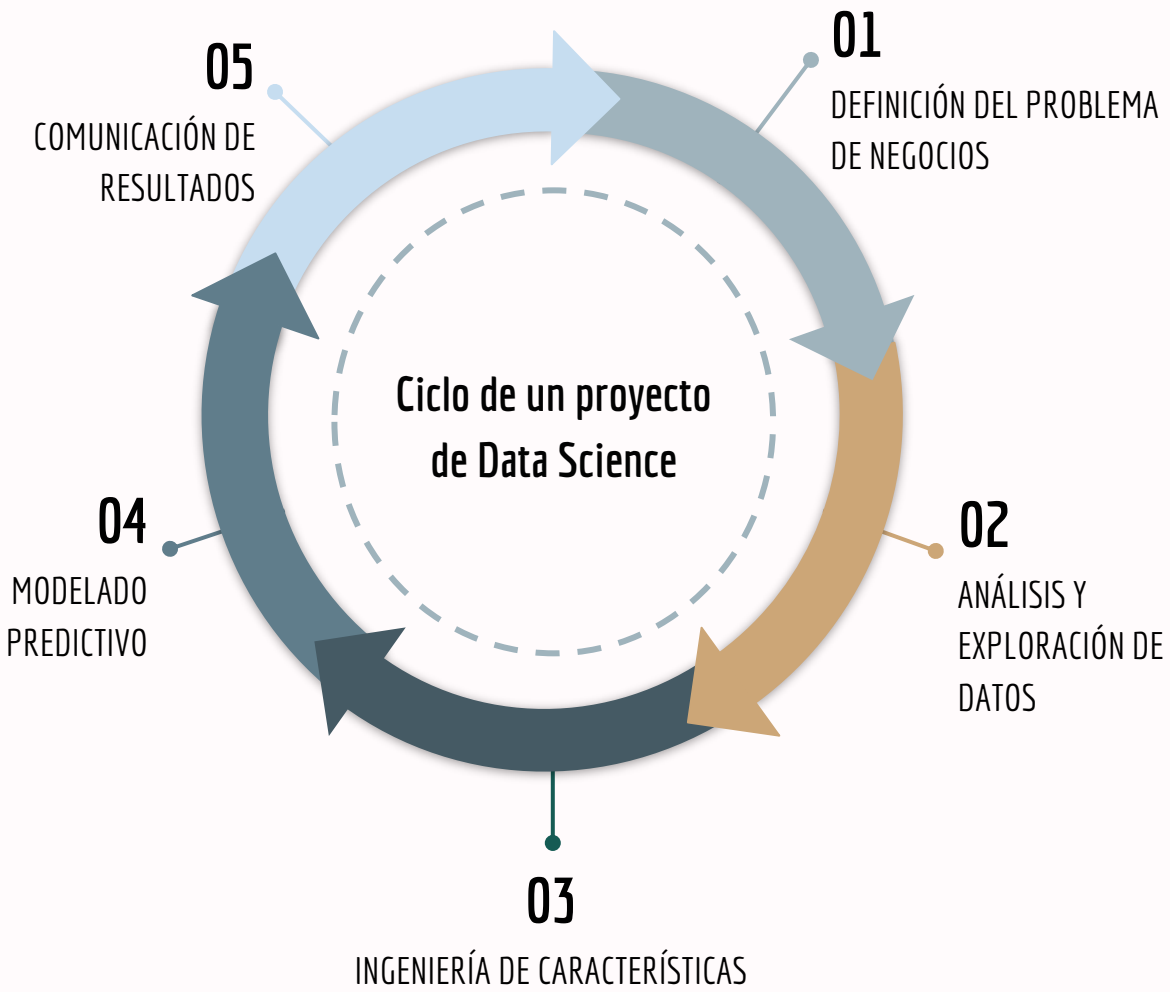
El ciclo sin fin



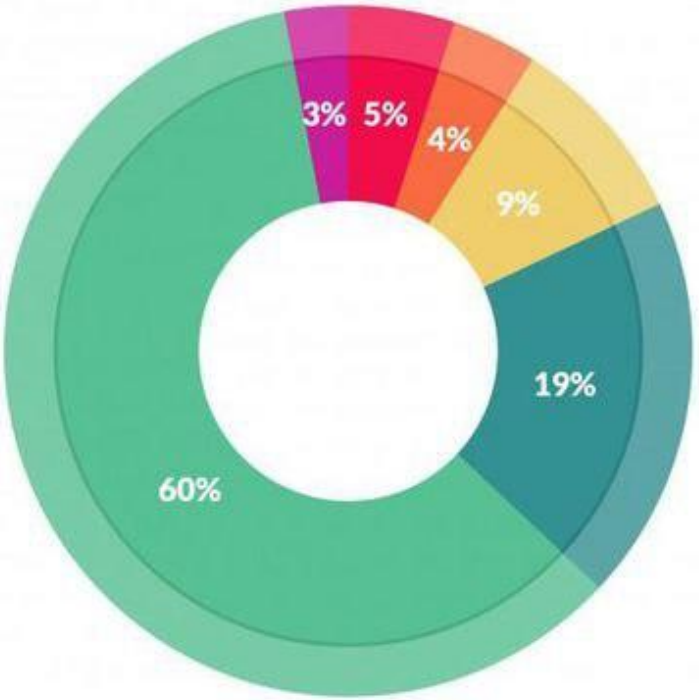
El ciclo sin fin

Durante esta materia, veremos conceptos involucrados en:

1. Herramientas estadísticas y visualizaciones para la etapa 02.
2. Herramientas estadísticas necesarias para interpretar los resultados de la etapa 04.
3. Visualización y comunicación efectiva para la etapa 05.



REALIDAD



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%*

Sesgos en los datos

¿Qué observa en esta imagen?

Sesgos en los datos

¿Qué observa en esta imagen?

- Bananas
- Etiquetas
- Bananas en estantes

Sesgos en los datos

El sesgo no proviene de algoritmos de IA, proviene de personas.

Habilidades a aprender durante esta materia

1. *Seleccionar* y *aplicar* herramientas estadísticas ***adecuadas***
2. *Diseñar* procesos de análisis de datos **sistemáticos**
3. *Obtener resultados* a partir de un conjunto de datos y *contextualizarlos*
4. *Explicar* resultados y conclusiones de forma **correcta** y **efectiva**
5. *Implementar* pipelines de análisis de datos en Python

¿Qué pasa si usamos Machine Learning sin saber análisis de datos?

- No sabemos **qué modelar**, a menos que alguien más se los diga.
- No podemos interpretar correctamente el **impacto** de sus resultados
  - Impacto a largo plazo, provocados por sesgos, unfairness, filtrado de información
  - Impacto en métricas de negocios, por ejemplo evaluado a través de test A/B
- Perdemos mucho tiempo en desarrollar modelos que no responden la pregunta correcta, y por lo tanto son menos accionables

¿Qué sucede si hacemos Data Science sin entender Machine Learning?

- Estamos limitados a análisis simples. O usamos modelos sin saber cómo funcionan...
- No entendemos qué tipo de **restricciones imponen los modelos** que elegimos, por ejemplo, modelos lineales, hiperplanos, etc.
- Aplicamos modelos a conjuntos de datos para los cuales no son adecuados, por ejemplo, redes neuronales sin normalizar las columnas
- Perdemos mucho tiempo en optimizar los modelos, porque no sabemos cómo

Material y Herramientas de trabajo

- Google colab → Para leer notebooks de python (00 Inicios en Python.ipynb)

- GitHub → Repositorio de los documentos