CONFERENCIAS STATA LATAM 2022
Herramientas y aplicaciones estadísticas para Ciencia de Datos

Metodología de datos sintéticos para modelos de *Machine Learning*

Franco A. Mansilla Ibáñez
www.francomansilla.com
Septiembre 2022

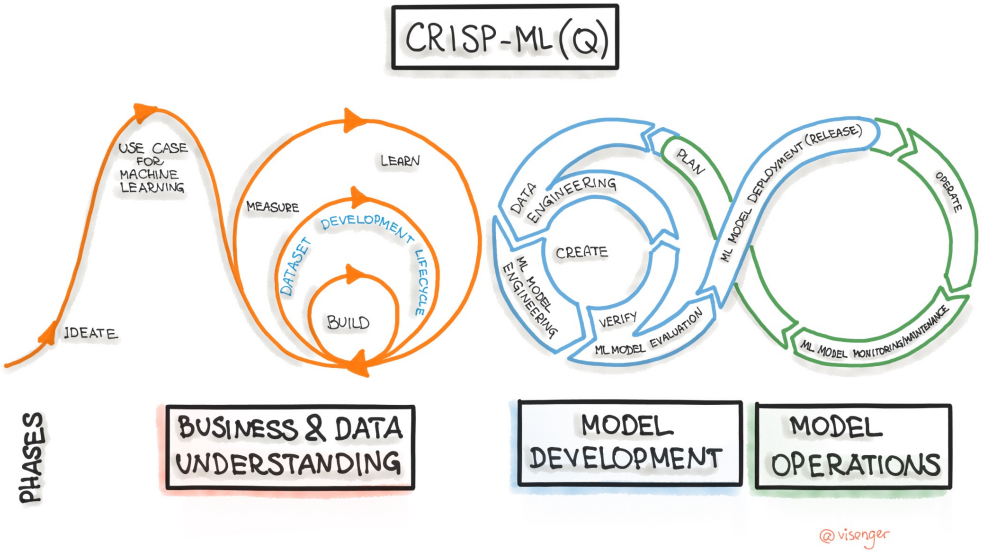
Software Shop | STATA

Agenda

1. Introducción
2. Pilares claves.
3. Algoritmos y Variables
4. Datos sintéticos.
5. Aplicación en Stata 17.

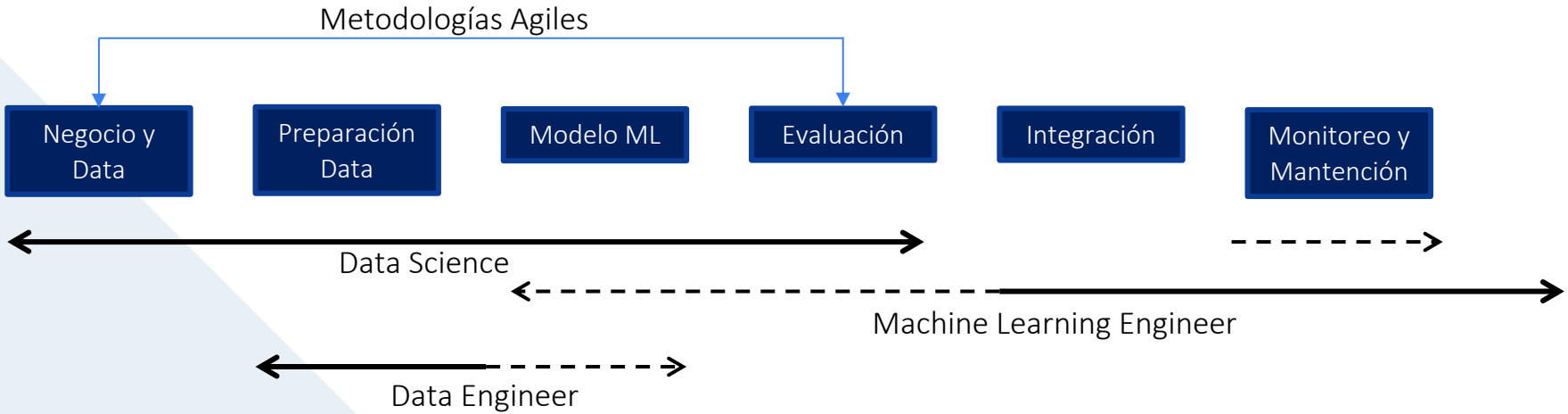
Introducción

Figura 1: Ciclo de desarrollo de un proyecto de Machine Learning.



Fuente: MLops

Pilares claves



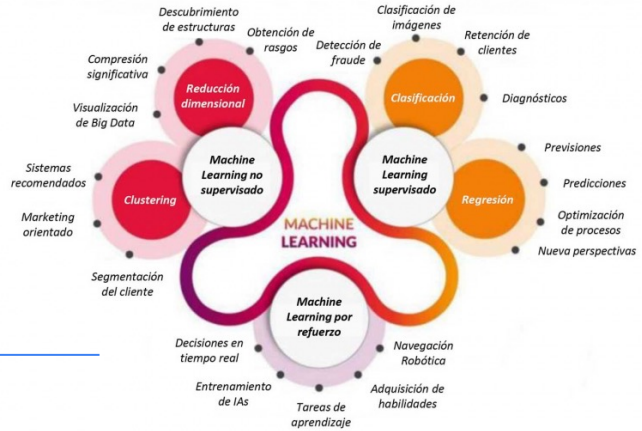
Algoritmos y Variables

Operacionales

Toma de Decisiones

Apoyo Gestión

Reducción de Tiempo

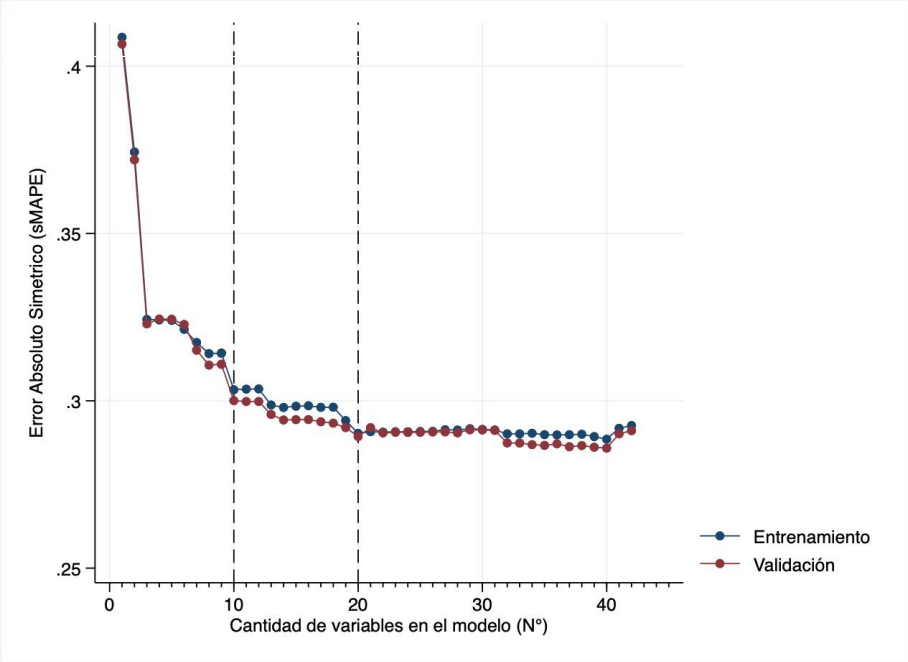


Algoritmos y Variables (cont.)

Modelos Clásicos v/s Machine Learning

- BigData.
- Sesgo y Varianza.
- Capacidad Tecnológica.

Fuente: Elaboración propia.



Datos Sintéticos

En la actualidad existe mucha investigación en metodologías para predecir datos en función a un contexto.

- Aleatoriedad.
- Aleatoriedad en función comportamiento (distribución de probabilidad).
- Anonimización y pseudoanonimización.
- Predicción por cluster.
- Predicción en imágenes

Datos Sintéticos

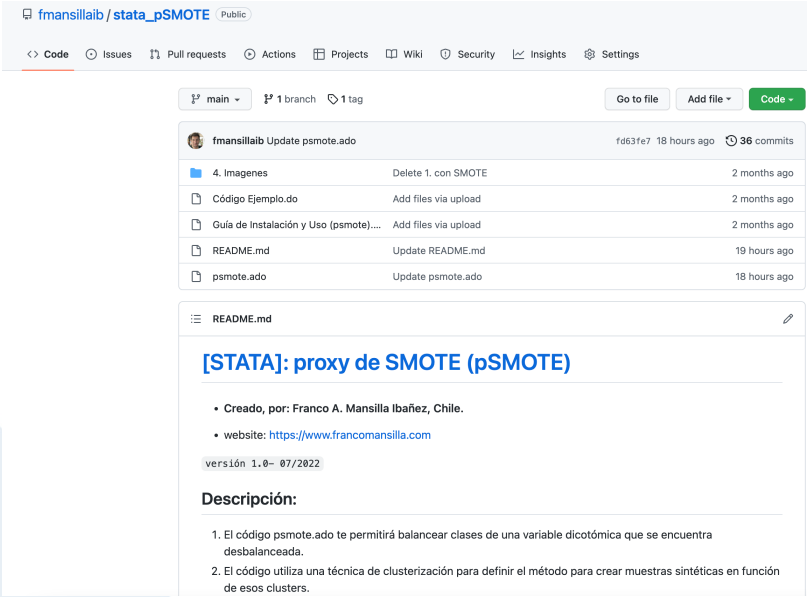
¿para que se usan?

- Nivelar Clases
- Completar imágenes
- Agregación de Datos
- Predecir futuros comportamientos
- Anonimizar datos

Técnica: Generative Adversarial Networks (GANs);
Técnica: Synthetic Minority Oversampling Technique (SMOTE)

Aplicación

Código pSMOTE: https://francomansilla.com/github



Do.file