

Metodología de datos sintéticos para modelos de *Machine Learning*

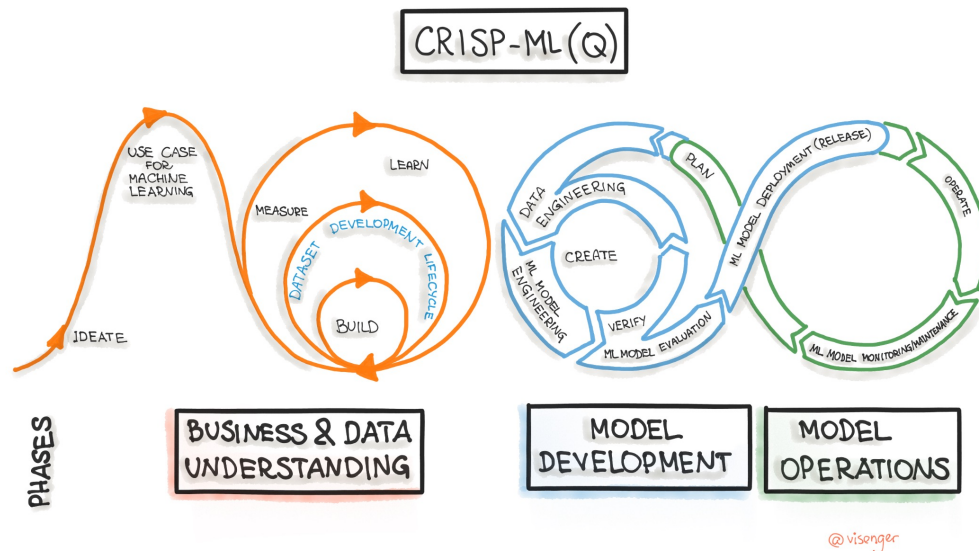
Franco A. Mansilla Ibáñez
www.francomansilla.com
Septiembre 2022

Agenda

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

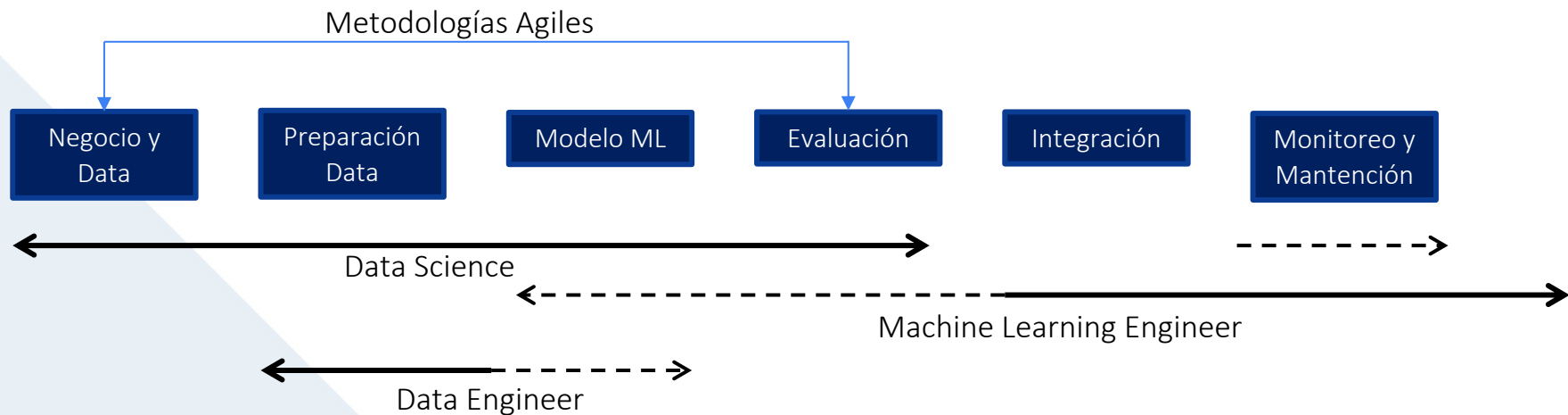
Introducción

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



Fuente: [MLops](#)

Pilares claves



—————→ Habilidad Fuerte

- - - - -→ Habilidad Débil

Algoritmos y Variables

Operacionales



Toma de Decisiones



Apoyo Gestión



Reducción de Tiempo

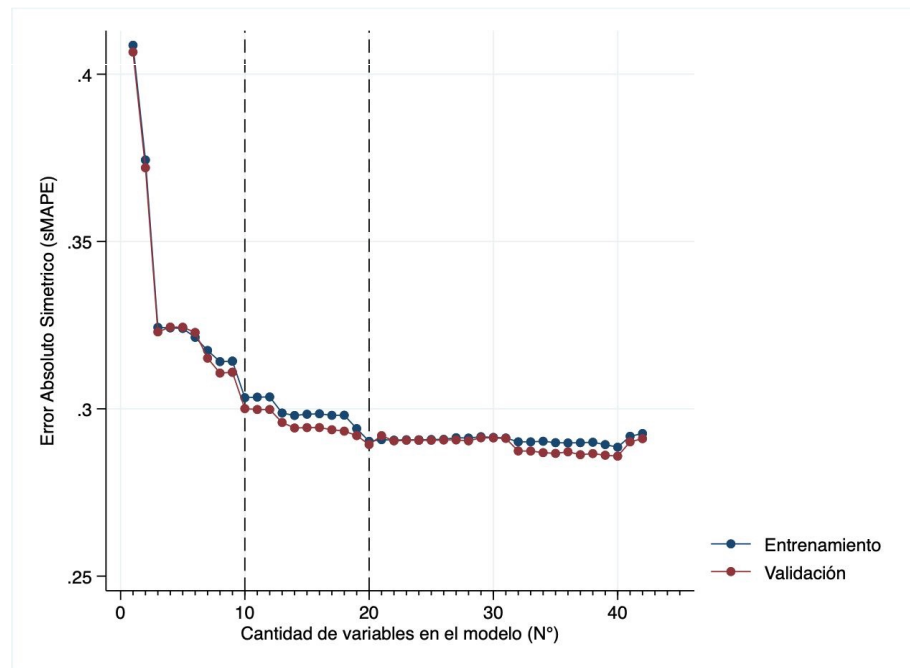


Algoritmos y Variables (cont.)

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>

fmansillaib / stata_pSMOTE Public

< Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 1 tag

Go to file Add file Code

File	Commit	Time
4. Imágenes	Delete 1. con SMOTE	2 months ago
Código Ejemplo.do	Add files via upload	2 months ago
Guía de Instalación y Uso (psmote)....	Add files via upload	2 months ago
README.md	Update README.md	19 hours ago
psmote.ado	Update psmote.ado	18 hours ago

README.md

[STATA]: proxy de SMOTE (pSMOTE)

- Creado, por: Franco A. Mansilla Ibáñez, Chile.
- website: <https://www.francomansilla.com>

versión 1.0- 07/2022

Descripción:

1. El código psmote.ado te permitirá balancear clases de una variable dicotómica que se encuentra desbalanceada.
2. El código utiliza una técnica de clusterización para definir el método para crear muestras sintéticas en función de esos clusters.

Do.file

```

1  * ~~~~~ *
2  * CONFERENCIA DE STATA - SEPT. 2022 *
3  * ~~~~~ *
4
5  * Franco A. Mansilla Ibáñez *
6  * ~~~~~ *
7  * www.francomansilla.com *
8  * www.software-shop.com *
9  * ~~~~~ *
10 * Conferencia STATA 09/2022 *
11 * ~~~~~ *
12
13 * ===== *
14
15 * ~~~~~ *
16 * Definición pre-eliminar *
17 * ~~~~~ *
18
19 clear all
20 set more off, permanently
21
22 * Cargar BD
23 import delimited "Volumes/GoogleDrive-111868847232940162537/Mi
24
25 * Renombrar variables
26 drop v1
27 ds *, varwidth(32)
28
29 global var_all = r(varlist)
30
31 local number=1
32 foreach i in $var_all {
33     rename `i' x`number'
34     local ++number
35 }
36
37
38 rename x3 fraude
39 drop x1 x2 x49 x50
40
41
42 * ~~~~~ *
43 * Análisis de la Data *
44 * ~~~~~ *
45
46 * 1. Tabulación de Fraude

```