

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
Septiembre 2022

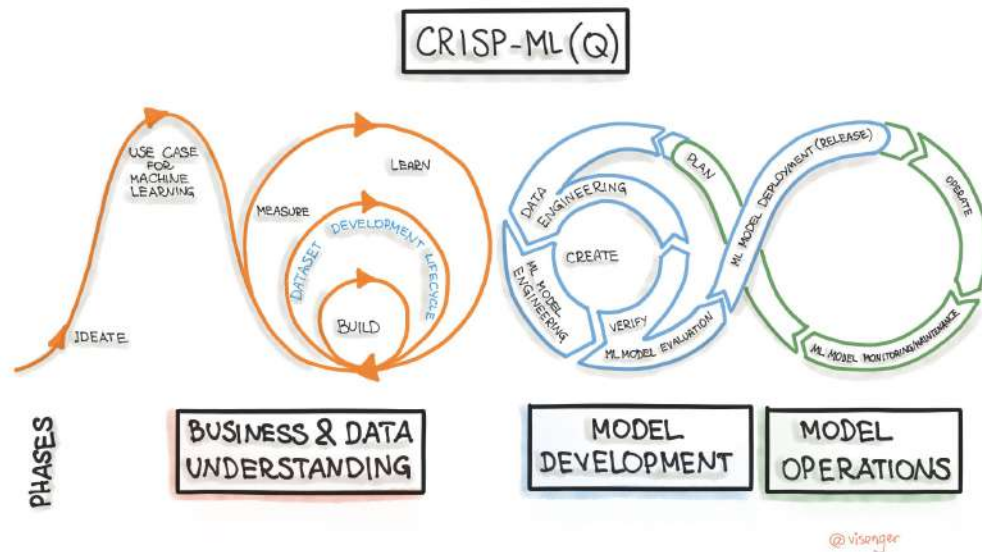


Agenda

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

Introducción

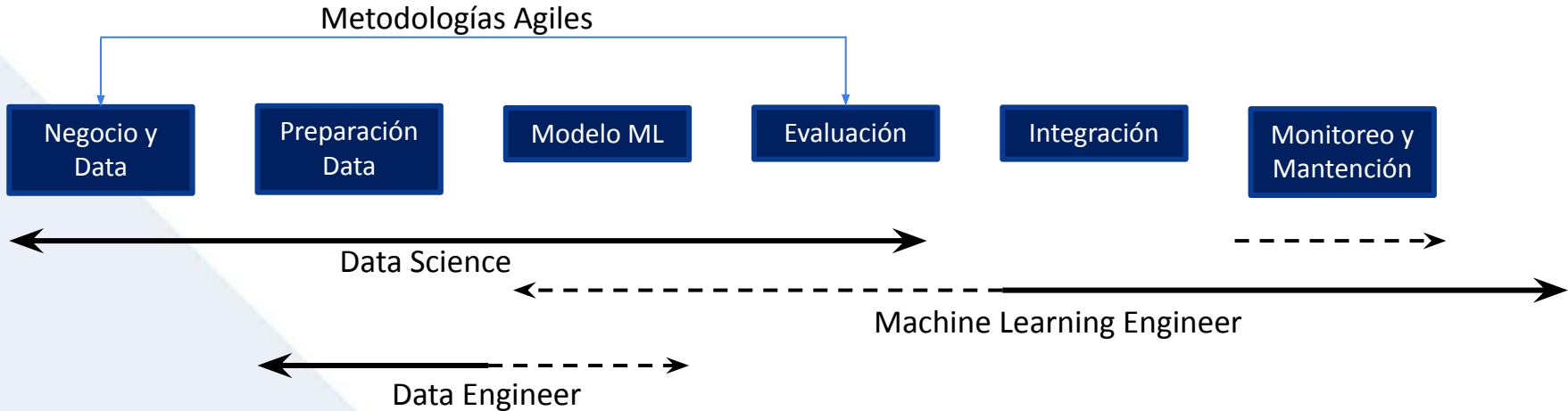
Figura 1: Ciclo de desarrollo de un proyecto de Machine Learning.



Fuente: [MLops](#)

@visonger

Pilares claves



Algoritmos y Variables

Operacionales



Toma de Decisiones



Apoyo Gestión



Reducción de Tiempo

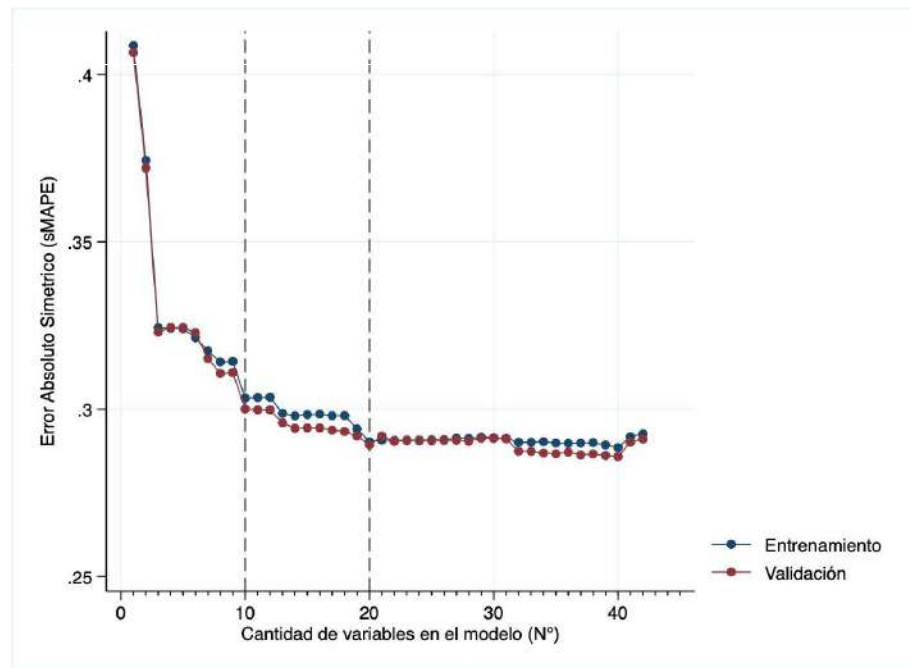


Algoritmos y Variables (cont.)

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 clúster.
- Predicción en imágenes

Datos Sintéticos

¿para que se usan?

Nivelar Clases

Completar
imágenes

Agregación de
Datos

Predecir futuros
comportamientos

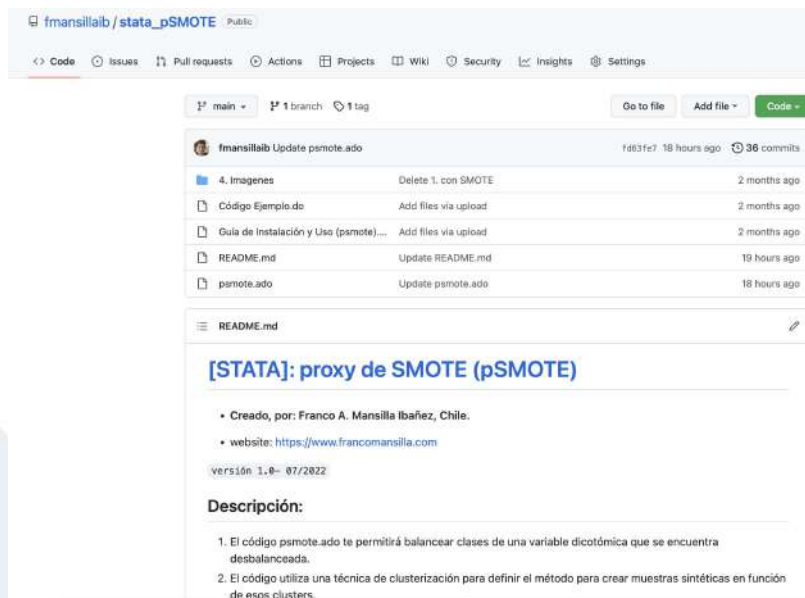
Anonimizar
datos

Técnica: Generative Adversarial Networks (GANs);

Técnica: Synthetic Minority Oversampling Technique (SMOTE)

Aplicación

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



The screenshot shows the GitHub repository page for 'francomansilla/stata_pSMOTE'. The repository is public and has 36 commits. The file list includes: 4. Imágenes, Código Ejemplo.do, Guía de Instalación y Uso (psmote)..., README.md, and psmote ado. The README.md file is selected, showing the title '[STATA]: proxy de SMOTE (pSMOTE)'. It was created by Franco A. Mansilla Ibañez, Chile, and the website is https://www.francomansilla.com. The version is 1.0-07/2022. The description states: 1. El código psmote.ado te permitirá desbalancear 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 Ibañez *
6  * www.francomansilla.com *
7  * www.software-shop.com *
8  * Conferencia STATA 09/2022 *
9  *-----*
10
11 * Definición pre-eliminar *
12 *-----*
13
14 clear all
15 set more off, permanently
16
17 * Cargar BD
18 import delimited "/Volumes/GoogleDrive-11186884732940162537/M1
19
20 * Renombrar variables
21 drop v1
22 ds *, varwidth(32)
23
24 global var_all = r(varlist)
25
26 local number=1
27 foreach i in $var_all {
28     rename `i' x`number'
29     local ++number
30 }
31
32 rename x3 fraude
33 drop x1 x2 x49 x50
34
35 *-----*
36 * Análisis de la Data *
37 *-----*
38 * 1. Tabulación de Fraude
```

Conferencias Stata LATAM 2022

Herramientas y aplicaciones estadísticas para Ciencia de Datos

Organiza:



Conozca más sobre STATA
escaneando el código QR.

