



Métricas de utilidad para datos sintéticos multidimensionales

Junio 2026



¿QUÉ ES LA CONFIDENCIALIDAD ESTADÍSTICA?

Definición Legal: Es la protección de los datos proporcionados por los Informantes del Sistema y de las personas físicas o morales objeto de la información. Su fin es evitar la identificación de los sujetos y prohibir el uso de los datos para fines no estadísticos, en cumplimiento con el artículo 47 de la Ley del Sistema Nacional de Información Estadística y Geográfica (LSNIEG)

Tipos de Identificación que se previenen:

Identificación Directa: Es el reconocimiento de la identidad utilizando únicamente los datos captados en la fuente, como el nombre, domicilio, números de identificación, registros, claves o cualquier dato análogo

Identificación Indirecta: Es el reconocimiento que se logra a través de la combinación de variables o datos presentes en diversas fuentes de información mediante el uso de técnicas o procedimientos especializados



\$582 MXN

Edad 42 años

Entidad Ver.

Modalidad
Trabajadores al servicio de
gobiernos estatales, municipales
y organismos
descentralizados



\$920 MXN

Edad 46 años

Entidad Jal.

Modalidad
Trabajadores permanentes y
eventuales de la ciudad



\$428 MXN

Edad 65 años

Entidad Ags.

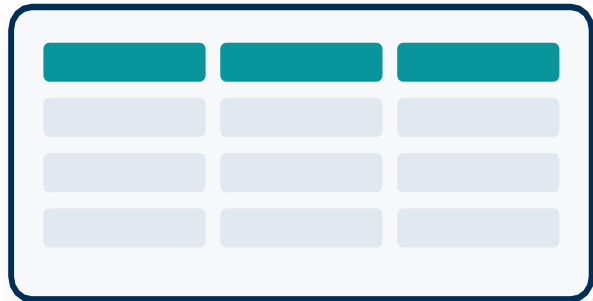
Modalidad
Trabajadores permanentes y
eventuales del campo

Uno fue generado sintéticamente.

¿Se puede distinguir?

¿Qué son los datos sintéticos?

Datos Reales



Microdatos sensibles



APRENDIZAJE



Datos Sintéticos



Microdatos estadísticamente similares

Caso práctico I

Ejemplo ENOE



Datos sintéticos a partir de la ENOE

hij5c	domestico	anios_esc	hrsocup	ingocup	ing_x_hrs	tpg_p8a	tcco	cp_anoc	imsssisste
2	3	11	50	10000	46.51163	0	3	0	1
2	6	12	0	0	0	0	0	0	0
0	8	10	0	0	0	0	0	0	0
2	8	9	0	0	0	0	0	0	0
3	3	6	36	0	0	0	2	0	1
1	2	11	17	0	0	0	0	0	4
0	9	9	0	0	0	0	0	0	0
0	3	12	30	12900	100	0	0	0	4
1	3	15	40	10000	58.13953	0	0	0	2
1	3	12	25	0	0	0	0	1	4
0	2	12	48	0	0	0	2	0	1
1	7	9	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Encuesta Nacional de Ocupación y Empleo (ENOE), población de 15 años y más de edad



La Encuesta Nacional de Ocupación y Empleo (ENOE) es la principal fuente de información sobre el mercado laboral mexicano al ofrecer datos mensuales y trimestrales de la fuerza de trabajo, la ocupación, la informalidad laboral, la subocupación y la desocupación. Constituye también el proyecto estadístico continuo más grande del país al proporcionar cifras nacionales y de cuatro tamaños de localidad, de cada una de las 32 entidades federativas y para un total de 39 ciudades.

En el momento del levantamiento en campo por la contingencia sanitaria, se implementó la Encuesta Telefónica de Ocupación y Empleo (ENOE^N) y sus resultados comprenden el periodo de abril a junio de 2020. Posteriormente, del tercer trimestre de 2020 al cuarto trimestre de 2020 la información proviene de la Encuesta Nacional de Ocupación y Empleo, Nueva Edición (ENOE^N). En tanto, la información del primer trimestre de 2005 al primer trimestre de 2020 y, a partir del primer trimestre de 2023, corresponde a la Encuesta Nacional de Ocupación y Empleo (ENOE). [Ver más.](#)

Desde el primer trimestre de 2005 a la fecha considera las estimaciones poblacionales trimestrales generadas por el Marco de Muestreo de la ENOE. [Ver más.](#)

ENOE_HOGT126

ENOE_COE2T126

ENOE_COE1T126

ENOE_SDEMT126

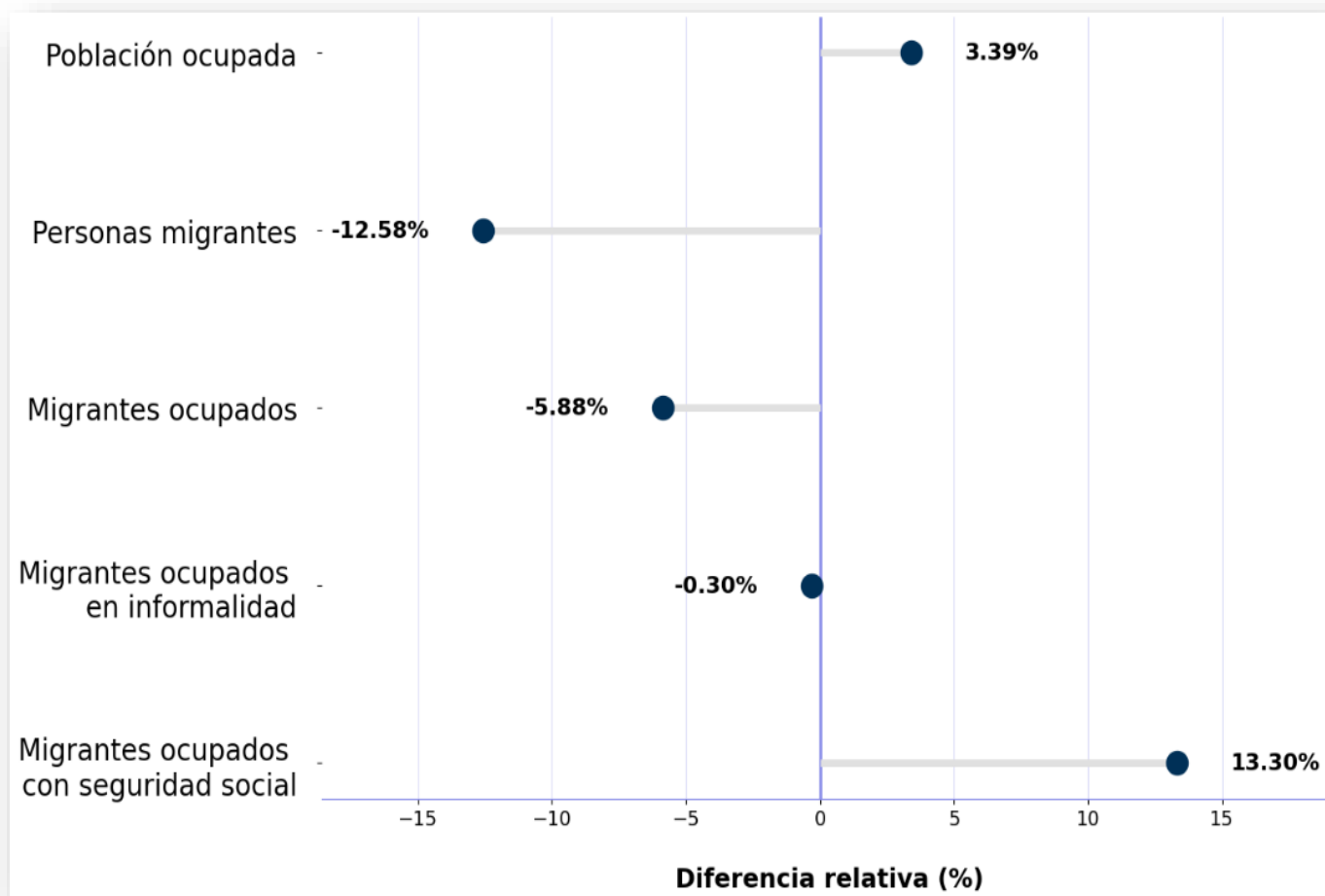
ENOE_VIVT126



hij5c	domestico	anios_esc	hrsocup	ingocup	ing_x_hrs	tpg_p8a	tcco	cp_anoc	imsssisste
2	3	11	50	10000	46.51163	0	3	0	1
2	6	12	0	0	0	0	0	0	0
0	8	10	0	0	0	0	0	0	0
2	8	9	0	0	0	0	0	0	0
3	3	6	36	0	0	0	2	0	1
1	2	11	17	0	0	0	0	0	4
0	9	9	0	0	0	0	0	0	0
0	3	12	30	12900	100	0	0	0	4
1	3	15	40	10000	58.13953	0	0	0	2
1	3	12	25	0	0	0	0	1	4
0	2	12	48	0	0	0	2	0	1
1	7	9	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
2	3	10	10	0	0	0	0	1	4
0	3	9	42	0	0	0	2	0	4
3	8	12	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	4	12	84	0	0	0	2	0	4
0	3	9	70	0	0	0	0	0	4
3	8	9	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
1	3	12	40	12900	75	0	0	1	4
0	3	9	15	9000	139.53488	0	0	0	4
2	8	9	0	0	0	0	0	0	0

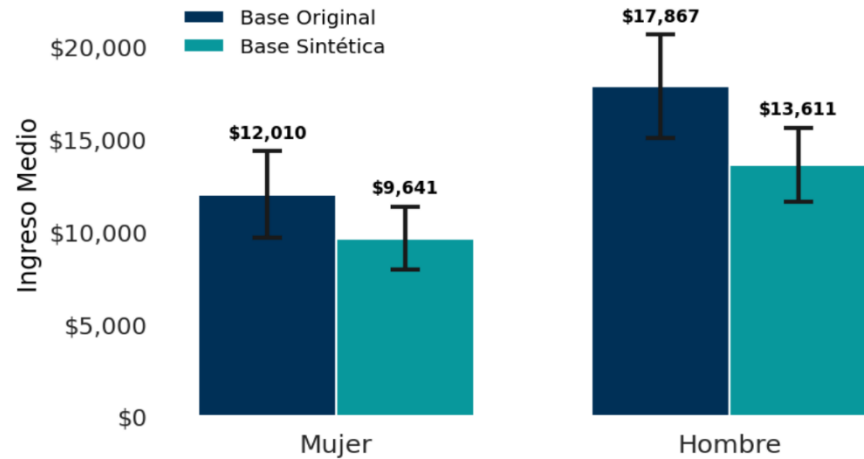
Reproducir totales

Indicador	real	sintético	CV	CV
			real	sintético
Población ocupada	59,552,660	61,571,312	0.5177	0.4896
Personas migrantes	708,573	619,407	4.7871	3.6312
Migrantes ocupados	364,617	343,177	6.0187	4.9167
% Migrantes ocupados en informalidad	65.72	65.52	3.7470	3.5966
% Migrantes ocupados con seguridad social	22.50	25.49	9.5164	8.7612

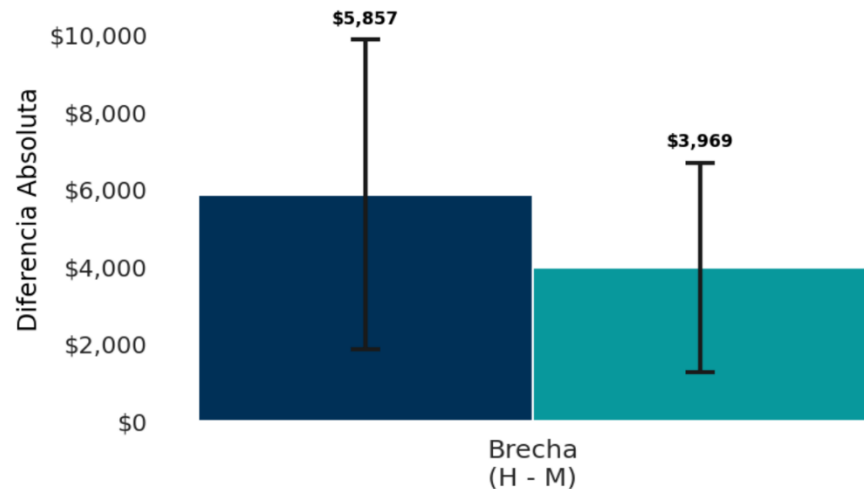


Brecha de género

Ingreso Medio Estimado por Género (IC al 95%)



Brecha de Ingreso Comparativa (IC al 95%)

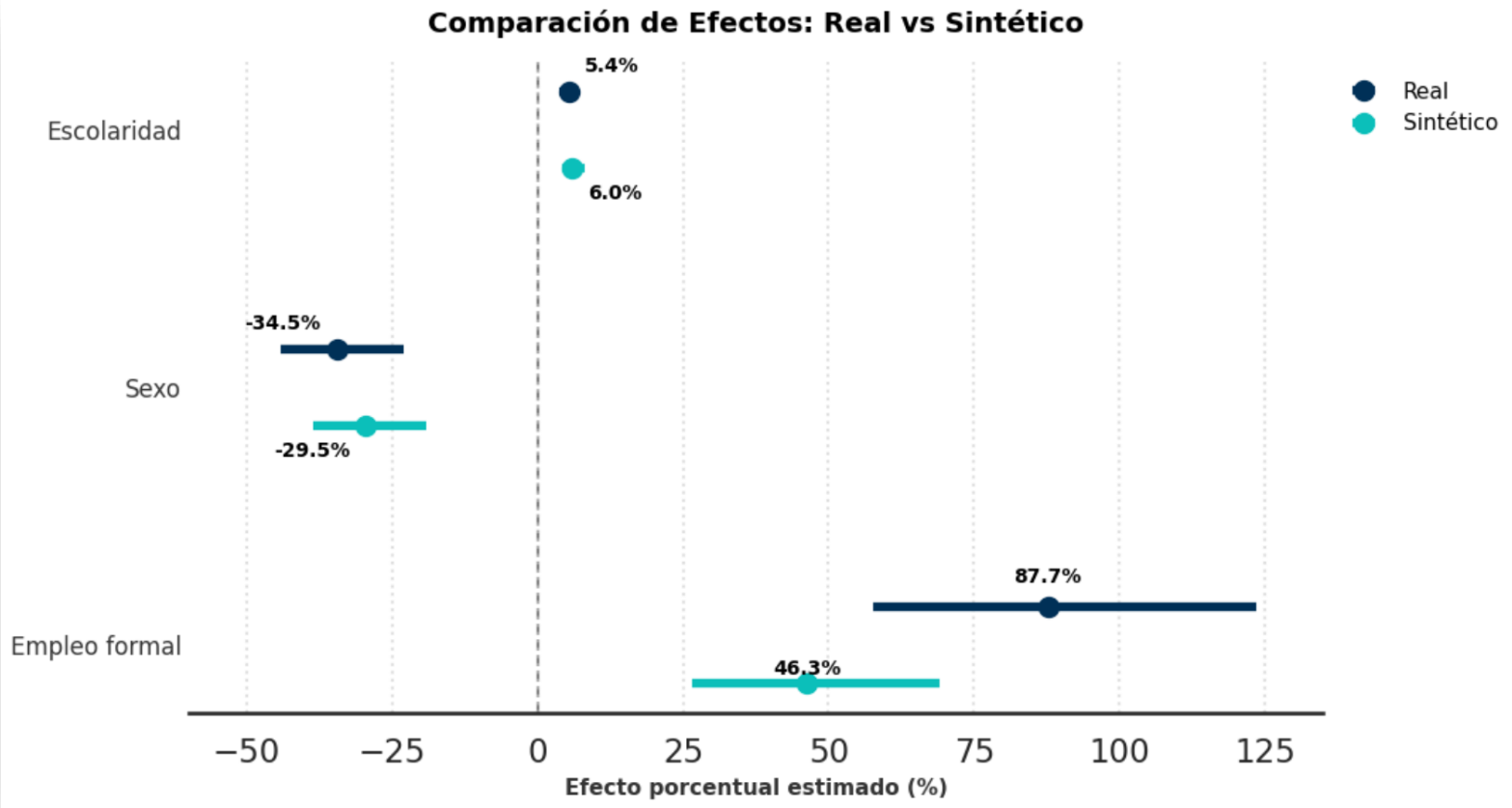


Datos reales **Datos sintéticos**

Brecha Estimada	\$5,857	\$3,969
P Valor	0.0044	0.0044
CV	34.90%	34.92%

Análisis de regresión

$$\ln(\text{Ingreso}) = \beta_0 + \beta_1(\text{Escolaridad}) + \beta_2(\text{Sexo}) + \beta_3(\text{Formalidad}) + \varepsilon$$



	Datos reales	Datos sintéticos
Pseudo R ²	0.2489	0.2729

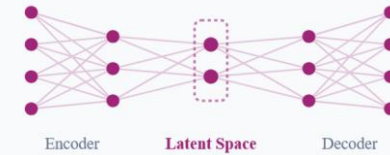
Estado del Arte

GANs



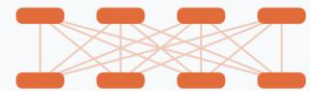
CTGAN

Autoencoders Variacionales



TVAE

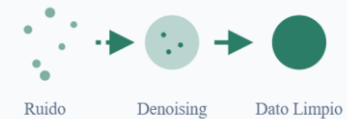
Transformers



Mecanismo de Auto-Atención (Self-Attention)

RealtabFormer

Difusion Models



TabDDPM

Necesidad de comparar

PRIVACIDAD

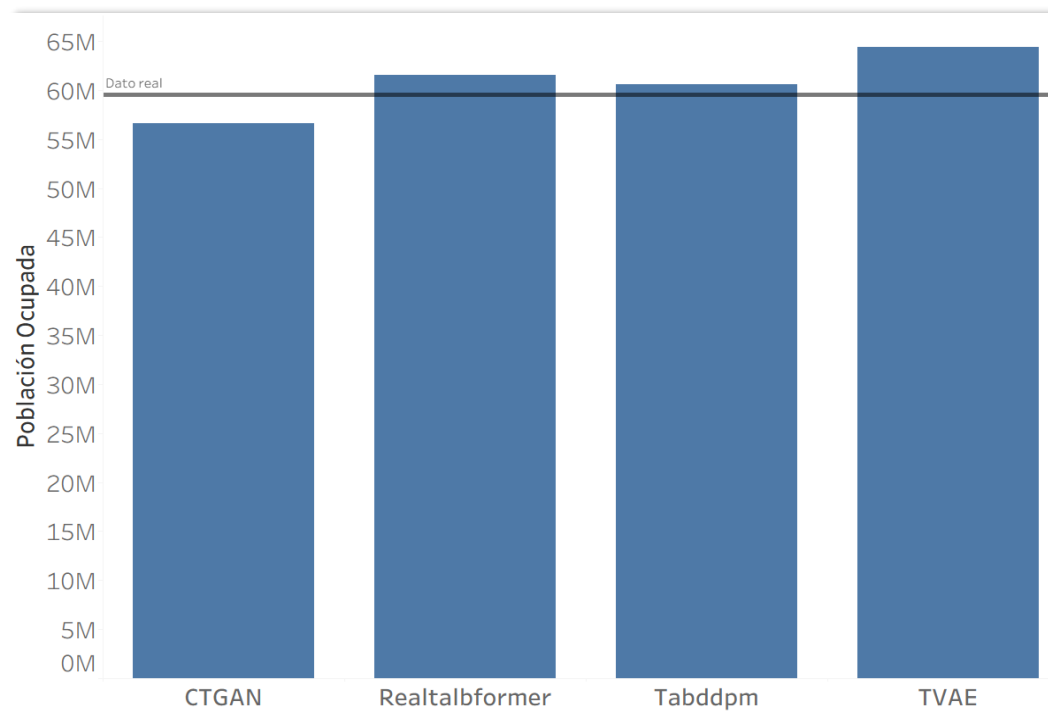
Copia exacta

Índice original: 68546

Índice sintético: 7

tipo_fila	indice	edad	sexo	migrante	ocupado	informal	seguridad_social	ingocup	anios_esc
ORIGINAL	68546	32	Hombre	No migrante	Ocupado	Informal	Sin seguridad social	10750	14.0
SINETICA	7	32	Hombre	No migrante	Ocupado	Informal	Sin seguridad social	10750	14.0

UTILIDAD



¿Cómo evaluamos los datos sintéticos generados?

Utilidad



Univariado

JSD, Hellinger, KS/TVD



Bivariado

Correlación, información mutua



Multivariado

pMSE, NNA, PCA

Privacidad



Copias

DCR, Hitting Rate



Proximidad

NNDR, vecinos cercanos



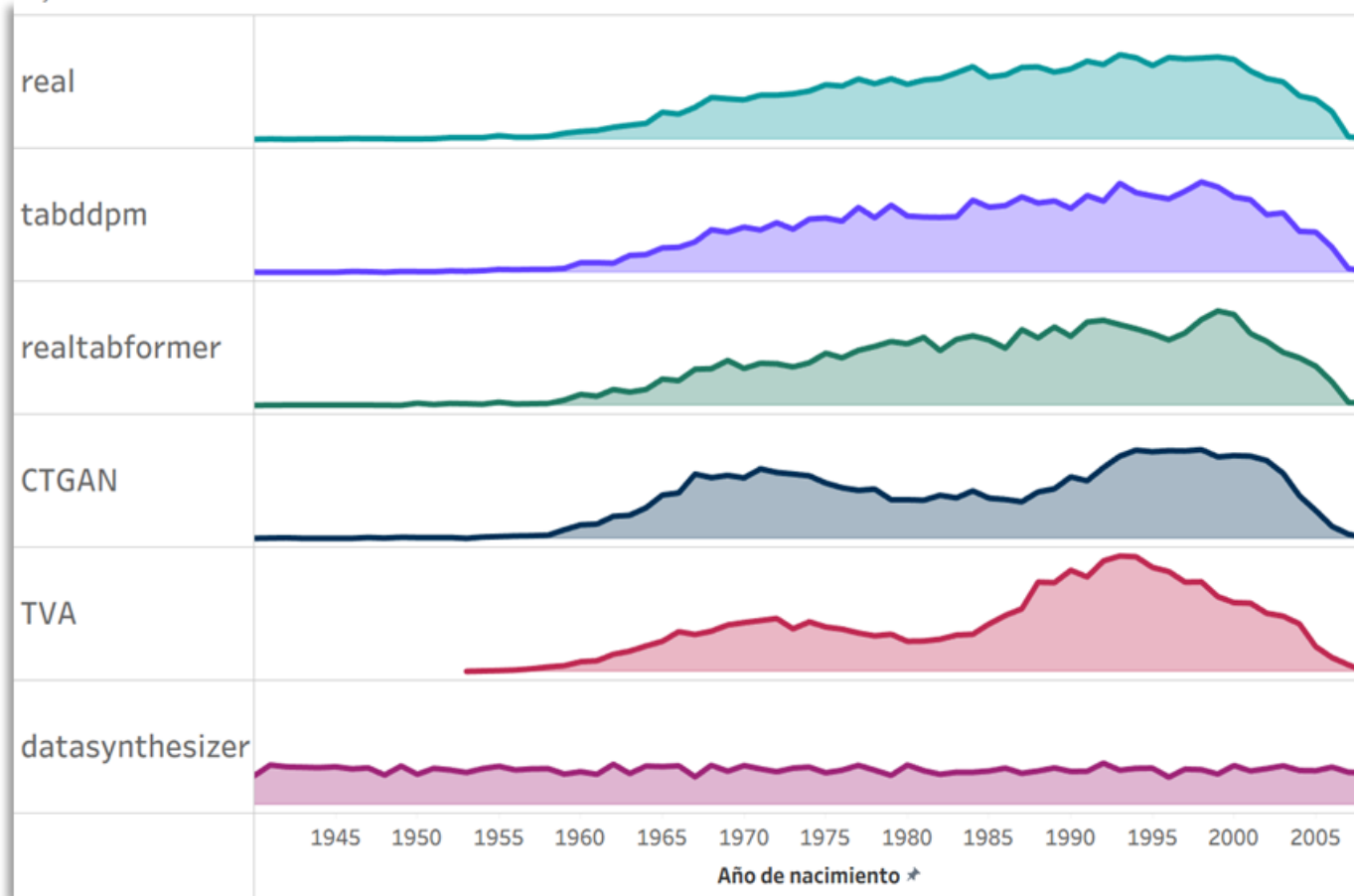
Inferencia

MIA, ADR, Caja negra /
caja blanca

Caso práctico II

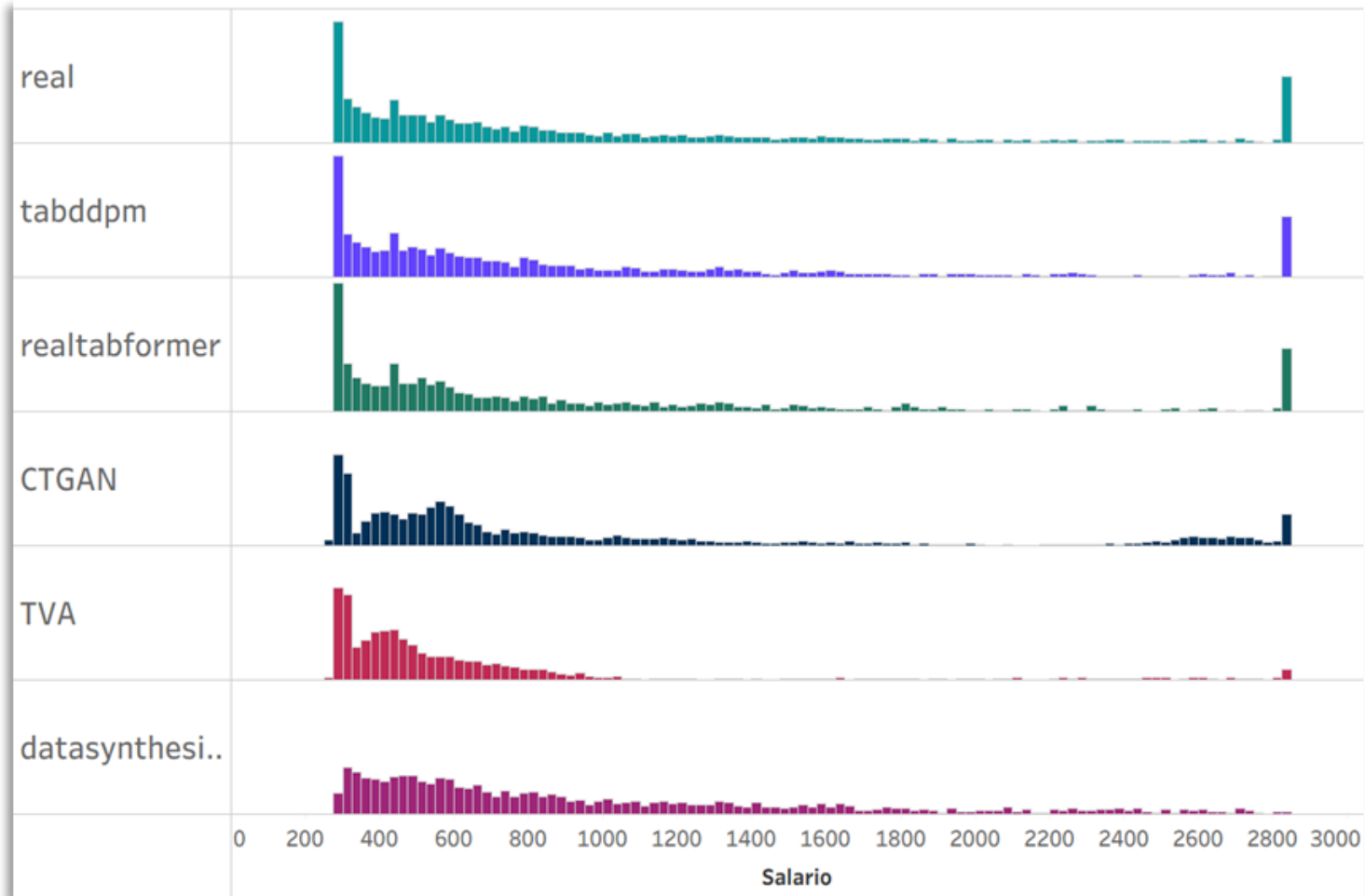
Registro Administrativos

Análisis exploratorio



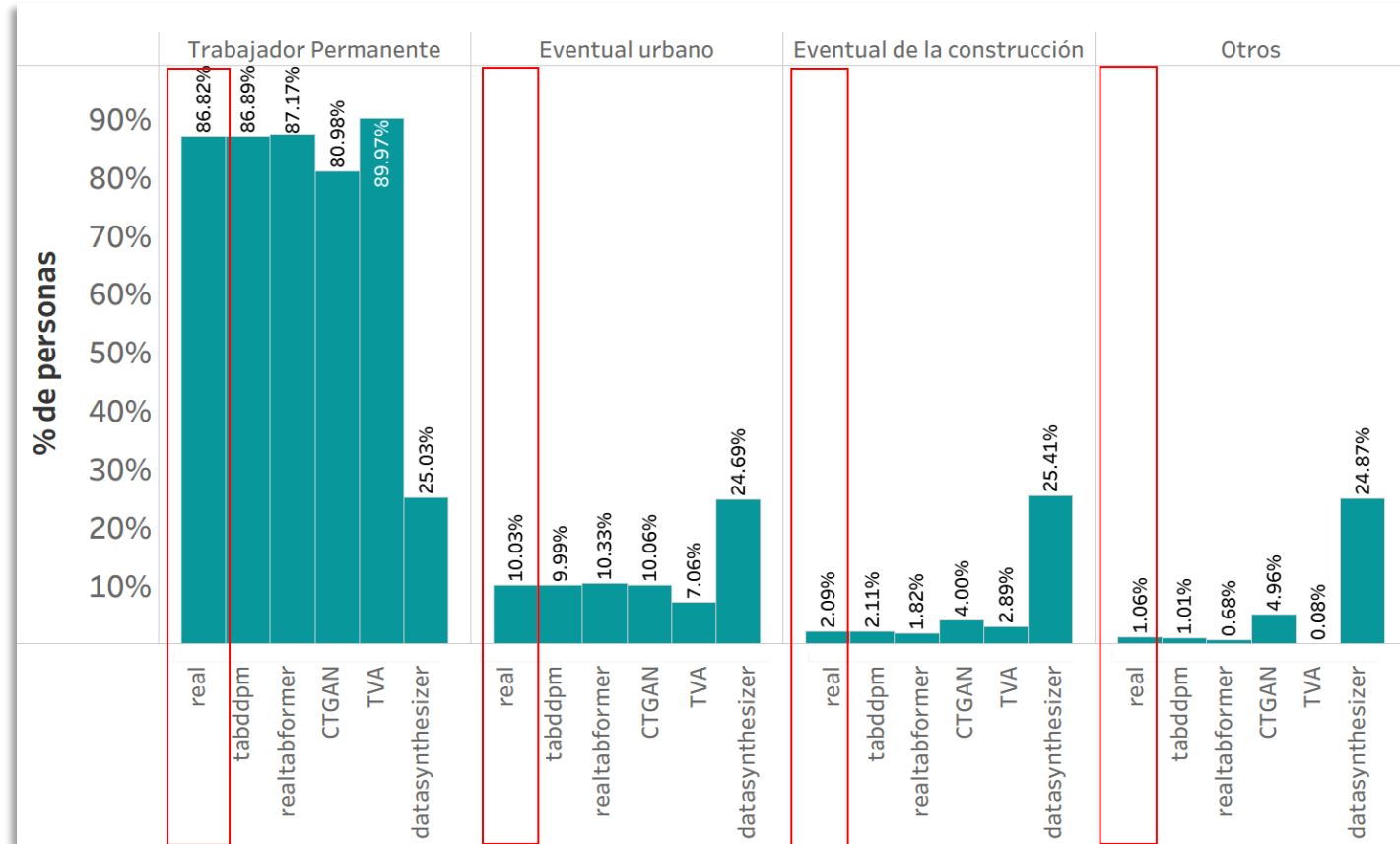
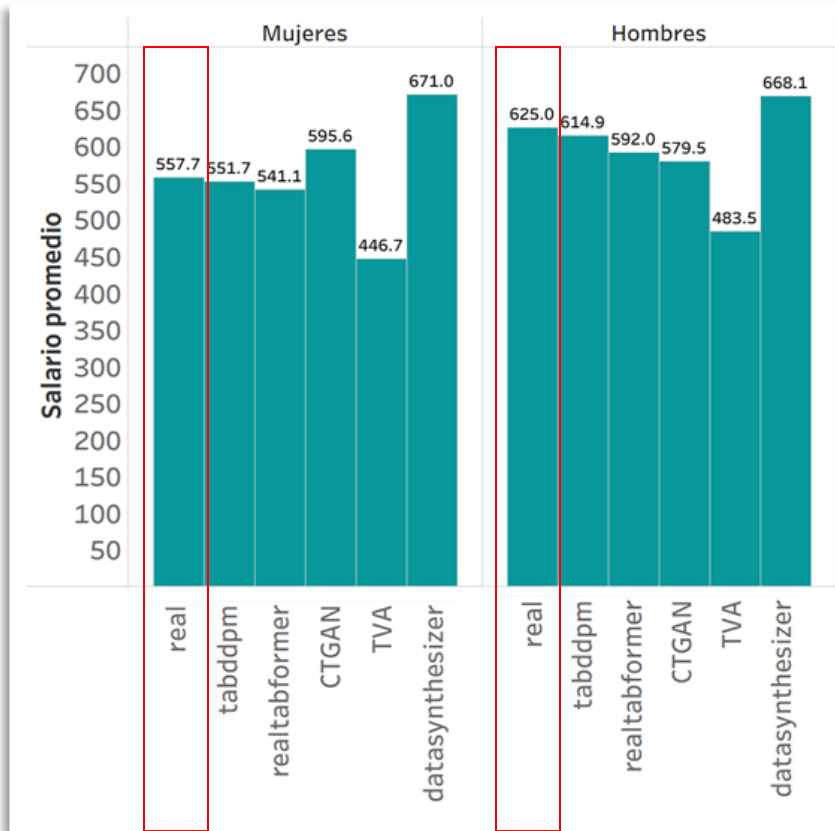
- tabddpm y realtabformer capturan de una buena forma la distribución del año de nacimiento
- Por el contrario, Datasynthesizer muestra dificultades para capturar esta misma distribución

Análisis exploratorio



- tabddpm y realtabformer modelan correctamente tanto el sesgo general como el pico secundario en la cola

Análisis exploratorio



Evaluación

Validación utilidad

	CTGAN	TVA	realtabformer	datasynthesizer	tabddpm
jsd_avg	0.0844	0.103	0.0234	0.2242	0.011
corr_diff_corr_m at_diff	0.9195	0.6271	0.1435	0.8663	0.2003

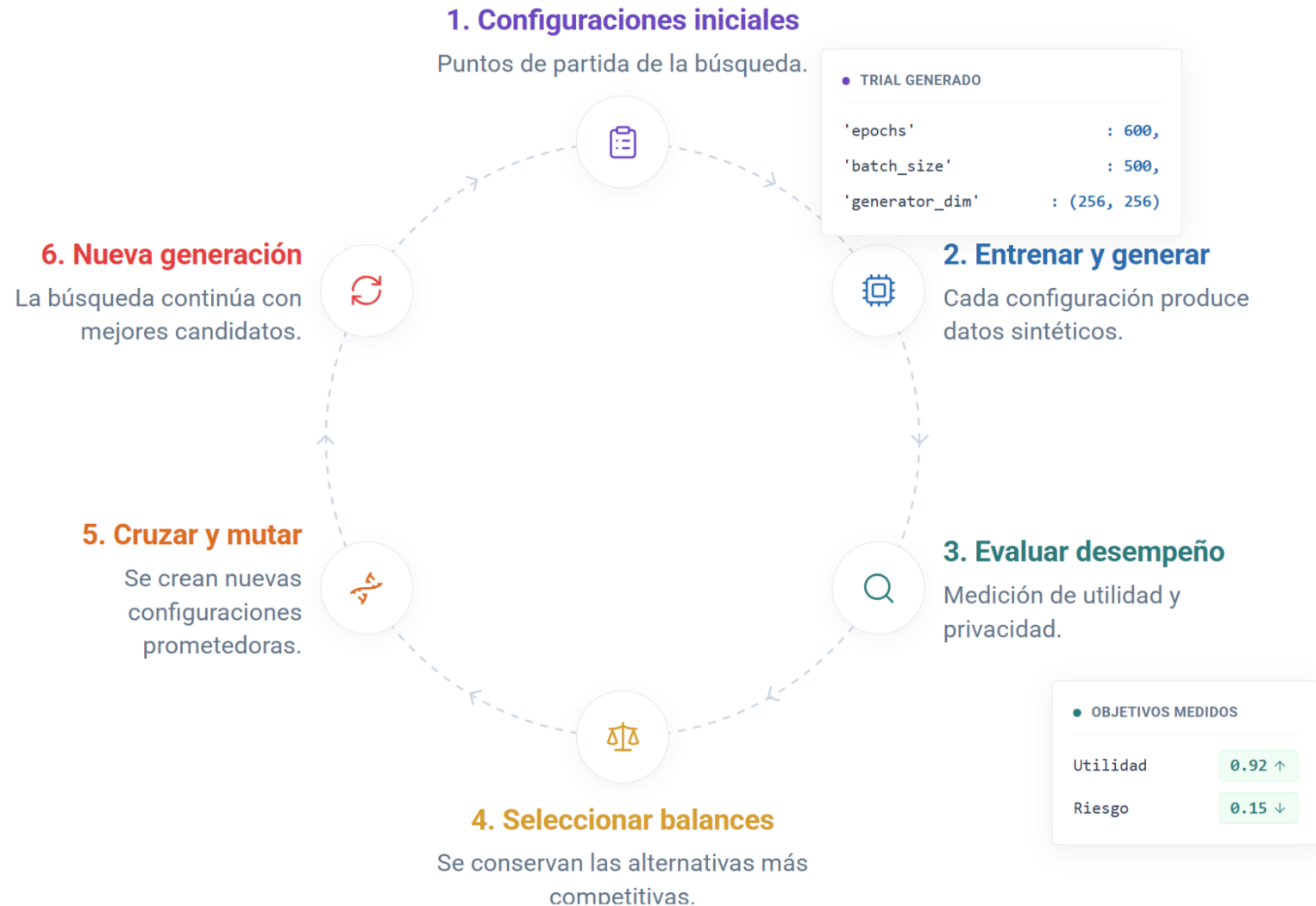
Validación Privacidad

mia_MIA recall	0.0961	0.1311	0.4208	0.015	0.451
dcr_mDCR	1.791	1.0644	0.9891	3.41	0.00137

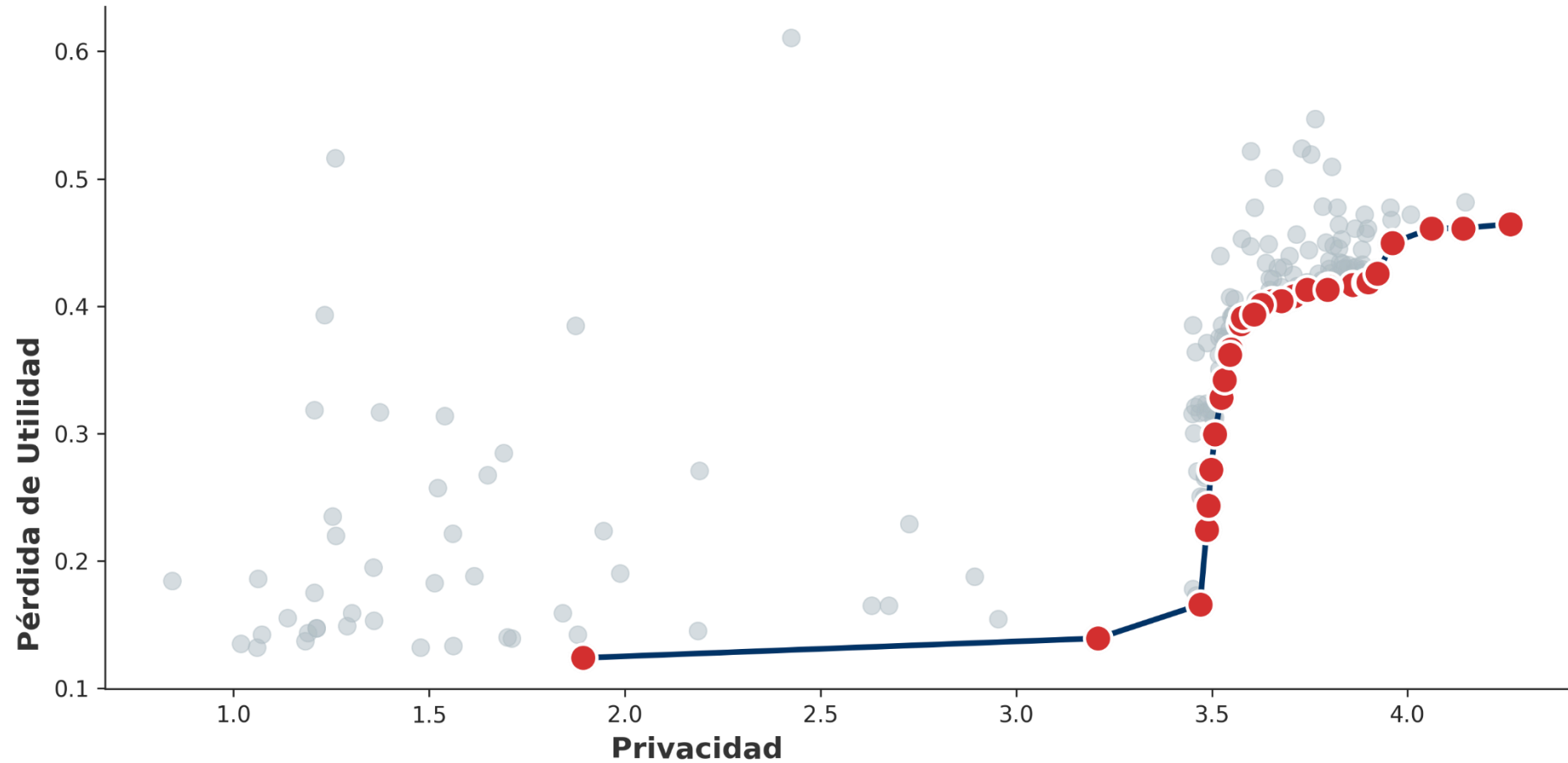
Hiperparámetros



Algoritmo NSGA-II



Capacidad de selección



Flujo de Propuesto



Optimización

Búsqueda multiobjetivo.

Max: DCR
Min: JSC, Corr



Generación

Entrenamiento final con hiperparámetros óptimos seleccionados.



Privacidad Final

Validación de seguridad.

MIA Recall
(Membership Inference Attack)



Utilidad Final

Reporte estadístico.

Univariada, Bivariada y Multivariada



¡GRACIAS!

CONOCIENDO
**MÉ
XI
CO**

800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

