



UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

DATA SINTÉTICA PRIVADA, EJECUCIÓN Y EVALUACIONES DE MODELOS

MEMORIA PARA OPTAR AL TÍTULO DE
MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN

GERARDO JORGE VILLARROEL GONZÁLEZ

PROFESOR GUÍA:
ANDRES ABELIUK

MIEMBROS DE LA COMISIÓN:

—

—

—

SANTIAGO DE CHILE

2023

Resumen

*A todos los lectores no orgánicos, espero que cuando interioricen estas palabras hayamos
aprendido a ser buenos padres*

Agradecimientos

Tabla de Contenido

1. Introducción	1
1.1. Estructura del documento	1
1.2. Equifax: contexto y limitaciones	3
1.3. Objetivo	5
2. Revisión Bibliográfica	6
2.1. Tipos de Datos	6
2.2. Privacidad de Datos	7
2.2.1. Tipo de datos a ser protegidos	7
2.2.2. Tipos de riesgos de divulgación	8
2.2.3. Regulación de datos sintéticos	9
2.2.4. Protección de Privacidad	9
2.3. Generación de Datos Sintéticos	10
2.3.1. Generación de datos tabulares	11
2.3.2. Generación de texto en base de datos tabulares	12
2.4. Metricas de evaluación	15
2.4.1. SDMetrics	15
2.4.2. Conjuntos Estadísticos	17
3. Desarrollo	19
3.1. Recursos disponibles	19
3.1.1. Conjuntos de datos	19

3.1.2.	Computación y Software	21
3.2.	Desarrollo del flujo de procesamiento	22
3.3.	Modelos	26
3.3.1.	Modelos para datos tabulares	26
3.3.2.	Modelos para textos	29
3.4.	Obtención de Métricas	30
4.	Resultados	34
4.1.	King County	35
4.1.1.	SDMetrics Score	35
4.1.2.	Correlación	36
4.1.3.	Cobertura	38
4.1.4.	Distribución	39
4.1.5.	Privacidad	43
4.1.6.	Ejemplo de registros	46
4.1.7.	Propiedades estadísticas	49
4.2.	Economicos	51
4.2.1.	Tratamiento de nulos en conjunto A y B	51
4.2.2.	SDMetrics Score - Conjunto A	52
4.2.3.	Correlación - Conjunto A	53
4.2.4.	Cobertura - Conjunto A	54
4.2.5.	Distribución - Conjunto A	55
4.2.6.	Privacidad - Conjunto A	56
4.2.7.	Ejemplos de registros - Conjunto A	58
4.2.8.	Propiedades estadísticas - Conjunto A	60
4.2.9.	SDMetrics Score - Conjunto B	62
4.2.10.	Correlación - Conjunto B	63
4.2.11.	Cobertura - Conjunto B	64

4.2.12. Distribución - Conjunto B	65
4.2.13. Privacidad - Conjunto B	66
4.2.14. Ejemplos de registros - Conjunto B	68
4.2.15. Propiedades estadísticas - Conjunto B	70
5. Conclusiones y discusión	72
5.1. Conclusiones	72
5.2. Limitaciones	73
5.3. Discusión	74
Bibliografía	77
Apéndice A. Anexos	78
A.1. Código de entrenamiento de economicos	79
A.2. Lista completa de figura pairwise kingcounty	80
A.3. Smote y Tddpm en KingCounty Graficas por Columnas	83
A.4. Tabla de comparación de Top5 KingCounty	83
A.5. Figuras de correlación Economicos - Conjunto A	83
A.6. Figuras de correlación Economicos - Conjunto B	86
A.7. Ejemplos de 10 Registros Generados Aleatoriamente en Descripciones Económicas A-1	90
A.8. Ejemplos de 10 Registros Generados Aleatoriamente en Descripciones Económicas	91
A.9. Estadísticos KingCounty	92
A.10. Estadísticos Economicos - Conjunto A	112
A.11. Estadísticos Economicos - Conjunto B	122

Índice de Tablas

2.1. Tipos de datos estructurados	6
2.2. Niveles de revelación y ejemplos	7
2.3. Tipos de Riesgos de Divulgación y sus Descripciones	8
2.4. Metricas de privacidad	9
2.5. Estado del arte en generación de datos tabulares	11
2.6. Estado del arte en generación de textos en base a datos	12
2.7. Ejemplo de tabla de entrada	13
2.8. Listado de conjunto estadísticos	17
3.1. Conjunto de datos King County	20
3.2. Conjunto de datos Economicos.cl	21
3.3. Computador Usado	22
3.4. Variables de entrada para <i>Synthetic</i>	27
3.5. Modelos Tabulares Soportados	28
3.6. Metricas para campos numericos	30
3.7. Métricas para campos categóricos	32
3.8. Ejemplo de scores promedios	33
4.1. Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, King County	35
4.2. Cobertura Categoría/Rango para Modelos Smote y Tddpm, King County	38
4.3. Evaluación de Similitud de Distribución para Modelos Smote y Tddpm, King County	39

4.4. Proporción entre el más cercano y el segundo más cercano, percentil 5, datos king county	43
4.5. Proporción entre el más cercano y el segundo más cercano, percentil 1, datos king county	43
4.6. Proporción entre el más cercano y el segundo más cercano, minimo, datos king county	43
4.7. Distancia de registros más cercanos, percentil 5, datos king county	44
4.8. Distancia de registros más cercanos, percentil 1, datos king county	44
4.9. Distancia de registros más cercanos, minimo, datos king county	44
4.10. Ejemplos para el modelo smote-enc, minimo	46
4.11. Ejemplos para el modelo tddpm_mlp, minimo	47
4.12. Ejemplos para el modelo smote-enc, percentil 1	48
4.13. Propiedades estadísticas de variable bedrooms con cambio>5 %, King county (A-1)	49
4.14. Propiedades estadísticas de variable bathrooms con cambio>5 %, King county (A-1)	50
4.15. Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, Economicos	52
4.16. Evaluación de Cobertura Categoría-Rango para Modelos SMOTE-ENC y TDDPM_MLP, Economicos	54
4.17. Evaluación de Similitud de Distribución para Modelos SMOTE-ENC y TDDPM_MLP, Economicos	55
4.18. Distancia de registros más cercanos, percentil 5, datos economicos	56
4.19. Distancia de registros más cercanos, percentil 1, datos economicos	56
4.20. Distancia de registros más cercanos, minimo, datos economicos	56
4.21. Proporción entre el más cercano y el segundo más cercano, percentil 5, datos economicos	57
4.22. Proporción entre el más cercano y el segundo más cercano, percentil 1, datos economicos	57
4.23. Proporción entre el más cercano y el segundo más cercano, minimo, datos economicos	57
4.24. Ejemplos para el modelo tddpm_mlp, minimo	58
4.25. Ejemplos para el modelo tddpm_mlp, percentil 1	58
4.26. Ejemplos para el modelo tddpm_mlp, percentil 4	59

4.27. Ejemplos de texto modelo tddpm_mlp, percentil 4	59
4.28. Propiedades estadísticas de variable m_size con cambio>5 %, Economicos (A-1)	60
4.29. Propiedades estadísticas de variable county, Economicos (A-1)	61
4.30. Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, Economicos	62
4.31. Evaluación de Cobertura Categoría-Rango para Modelos SMOTE-ENC y TDDPM_MLP, Economicos	64
4.32. Evaluación de Similitud de Distribución para Modelos SMOTE-ENC y TDDPM_MLP, Economicos	65
4.33. Distancia de registros más cercanos, percentil 5, datos economicos	66
4.34. Distancia de registros más cercanos, percentil 1, datos economicos	66
4.35. Distancia de registros más cercanos, minimo, datos economicos	66
4.36. Proporción entre el más cercano y el segundo más cercano, percentil 5, datos economicos	67
4.37. Proporción entre el más cercano y el segundo más cercano, percentil 1, datos economicos	67
4.38. Proporción entre el más cercano y el segundo más cercano, minimo, datos economicos	67
4.39. Ejemplos para el modelo tddpm_mlp, percentil 2	68
4.40. Ejemplos de texto modelo tddpm_mlp, percentil 2	68
4.41. Ejemplos para el modelo tddpm_mlp, percentil 4	69
4.42. Ejemplos de texto modelo tddpm_mlp, percentil 4	69
4.43. Propiedades estadísticas de variable bathrooms con cambio>5 %, Economicos (B-1)	70
4.44. Propiedades estadísticas de variable m_size con cambio>5 %, Economicos (B-1)	71
A.1. Ejemplos de textos aleatorios del modelo tddpm_mlp, conjunto Economicos a-1	90
A.2. Ejemplos de textos aleatorios del modelo tddpm_mlp, conjunto Economicos b-1	91
A.3. Propiedades estadísticas de variable grade, King county (A-3)	92
A.4. Propiedades estadísticas de variable view, King county (A-3)	93
A.5. Propiedades estadísticas de variable lat, King county (A-3)	94
A.6. Propiedades estadísticas de variable zipcode, King county (A-3)	95

A.7. Propiedades estadísticas de variable sqft_basement, King county (A-3)	96
A.8. Propiedades estadísticas de variable price, King county (A-3)	97
A.9. Propiedades estadísticas de variable sqft_above, King county (A-3)	98
A.10. Propiedades estadísticas de variable date, King county (A-3)	99
A.11. Propiedades estadísticas de variable bathrooms, King county (A-3)	100
A.12. Propiedades estadísticas de variable long, King county (A-3)	101
A.13. Propiedades estadísticas de variable sqft_lot, King county (A-3)	102
A.14. Propiedades estadísticas de variable bedrooms, King county (A-3)	103
A.15. Propiedades estadísticas de variable sqft_living15, King county (A-3)	104
A.16. Propiedades estadísticas de variable waterfront, King county (A-3)	105
A.17. Propiedades estadísticas de variable sqft_living, King county (A-3)	106
A.18. Propiedades estadísticas de variable condition, King county (A-3)	107
A.19. Propiedades estadísticas de variable yr_built, King county (A-3)	108
A.20. Propiedades estadísticas de variable sqft_lot15, King county (A-3)	109
A.21. Propiedades estadísticas de variable floors, King county (A-3)	110
A.22. Propiedades estadísticas de variable yr_renovated, King county (A-3)	111
A.23. Propiedades estadísticas de variable bathrooms, Economicos (A-3)	112
A.24. Propiedades estadísticas de variable property_type, Economicos (A-3)	113
A.25. Propiedades estadísticas de variable _price, Economicos (A-3)	114
A.26. Propiedades estadísticas de variable publication_date, Economicos (A-3)	115
A.27. Propiedades estadísticas de variable state, Economicos (A-3)	116
A.28. Propiedades estadísticas de variable county, Economicos (A-3)	116
A.29. Propiedades estadísticas de variable rooms, Economicos (A-3)	117
A.30. Propiedades estadísticas de variable m_size, Economicos (A-3)	118
A.31. Propiedades estadísticas de variable m_built, Economicos (A-3)	119
A.32. Propiedades estadísticas de variable transaction_type, Economicos (A-3)	120
A.33. Propiedades estadísticas de variable m_built, Economicos (B-3)	122

A.34.Propiedades estadísticas de variable bathrooms, Economicos (B-3)	123
A.35.Propiedades estadísticas de variable transaction_type, Economicos (B-3)	124
A.36.Propiedades estadísticas de variable m_size, Economicos (B-3)	125
A.37.Propiedades estadísticas de variable property_type, Economicos (B-3)	126
A.38.Propiedades estadísticas de variable county, Economicos (B-3)	126
A.39.Propiedades estadísticas de variable rooms, Economicos (B-3)	127
A.40.Propiedades estadísticas de variable publication_date, Economicos (B-3)	128
A.41.Propiedades estadísticas de variable _price, Economicos (B-3)	129
A.42.Propiedades estadísticas de variable state, Economicos (B-3)	130

Índice de Ilustraciones

3.1. Proceso para generar datos sintéticos con SDV	24
3.2. Proceso para generar datos sintéticos completo	25
3.3. Carpetas y archivos esperados generados por <i>Synthetic</i>	28
4.1. Correlación de conjunto original de entrenamiento y Copulagan	36
4.2. Correlación de conjunto original de entrenamiento y Gaussiancopula	36
4.3. Correlación de conjunto original de entrenamiento y Smote	37
4.4. Correlación de conjunto original de entrenamiento y Tddpm	37
4.5. Frecuencia del campo Grade en el modelo real y Top 2	40
4.6. Frecuencia del campo Bedrooms en el modelo real y Top 2	41
4.7. Frecuencia del campo Sqft lot15 en el modelo real y Top 2	42
4.8. Frecuencia del campo Privacy en el modelo real y Top 2	45
4.9. Correlación de conjunto original de entrenamiento y Smote	53
4.10. Correlación de conjunto original de entrenamiento y Tddpm	53
4.11. Frecuencia del campo M size en el modelo real y Top 2	54
4.12. Correlación de conjunto original de entrenamiento y Smote	63
4.13. Correlación de conjunto original de entrenamiento y Tddpm	63
A.1. Correlación de conjunto original de entrenamiento y Copulagan	80
A.2. Correlación de conjunto original de entrenamiento y Tvae	80
A.3. Correlación de conjunto original de entrenamiento y Gaussiancopula	81
A.4. Correlación de conjunto original de entrenamiento y Ctgan	81

A.5. Correlación de conjunto original de entrenamiento y Tablepreset	82
A.6. Correlación de conjunto original de entrenamiento y Smote	82
A.7. Correlación de conjunto original de entrenamiento y Tddpm	83
A.8. Correlación de conjunto original de entrenamiento y Copulagan	83
A.9. Correlación de conjunto original de entrenamiento y Tvae	84
A.10. Correlación de conjunto original de entrenamiento y Gaussiancopula	84
A.11. Correlación de conjunto original de entrenamiento y Ctgan	85
A.12. Correlación de conjunto original de entrenamiento y Smote	85
A.13. Correlación de conjunto original de entrenamiento y Tddpm	86
A.14. Correlación de conjunto original de entrenamiento y Copulagan	86
A.15. Correlación de conjunto original de entrenamiento y Tvae	87
A.16. Correlación de conjunto original de entrenamiento y Gaussiancopula	87
A.17. Correlación de conjunto original de entrenamiento y Ctgan	88
A.18. Correlación de conjunto original de entrenamiento y Smote	88
A.19. Correlación de conjunto original de entrenamiento y Tddpm	89

Lista de códigos

1.	Devcontainer del actual proyecto.	22
2.	Instanciando clase Synthetic	27
3.	Obtención de métricas en economicos_run-a.py	30
4.	Mostrando Scores Promedios Calculados	33
5.	Eliminación de valores nulos en el conjunto de datos de Económicos	51
6.	Reemplazo de valores nulos en el conjunto de datos de Económicos	51
7.	Código de ejemplo en Python para sumar dos números. Fuente: Autor.	79

Capítulo 1

Introducción

Cuando se revise esta tesis, estará desactualizada. Desde AlexNet [28] en 2012, el liderazgo en el problema de clasificación de imágenes ha cambiado al menos 15 veces [7]. En el campo de texto a imágenes, modelos como DALL-E 2 [1], Google Imagen [3] y Stable Diffusion [9] fueron presentados en 2022, mientras que para el 2023 se pronostica el inicio de una carrera de inteligencia artificial en el campo de los chatbots entre Google y Microsoft [29, 5]. En definitiva, es un campo actualmente en crecimiento y que seguirá sorprendiendo con nuevas técnicas y productos, en variedad y calidad.

En el contexto de **Equifax**, la empresa en la que se centra este esfuerzo, es fundamental avanzar de manera rápida y efectiva en el uso de su información para poder mantenerse a la vanguardia en el mercado y poder competir con otras empresas del sector.

Según el libro *Practical synthetic data generation: balancing privacy and the broad availability of data* [20] los datos sintéticos ofrecen dos beneficios principales:

1. Mayor eficiencia en la disponibilidad de datos, y
2. Mejora en los análisis realizados.

Para **Equifax**, ambos beneficios son valiosos, aunque inicialmente la eficiencia en la disponibilidad de datos tiene mayor peso. Como se verá posteriormente, la empresa ejerce un control total sobre el acceso a la información y los datos, ya que es necesario proteger su confidencialidad.

El objetivo general de este trabajo es diseñar un mecanismo para generar conjuntos de datos sintéticos estructurados, que contengan textos, y compararlos con sus contrapartes originales utilizando deep learning.

1.1. Estructura del documento

En este documento se presenta un estudio detallado del desarrollo de un mecanismo para generar conjuntos de datos sintéticos estructurados que incluyen textos, y se comparan con sus contra-

partes originales utilizando deep learning.

En la **Introducción** se establecerá el contexto del desafío, se describirán los objetivos a cumplir y se presentará la estructura del documento.

En el capítulo 2 se realizará una revisión de la literatura sobre técnicas de generación de datos sintéticos y deep learning.

En el capítulo 3 se detallará el diseño y la implementación del mecanismo para generar los conjuntos de datos sintéticos y su comparación con los conjuntos de datos originales.

En el capítulo 4 se presentarán los resultados de la evaluación comparativa entre los conjuntos de datos sintéticos y los originales.

Finalmente, en el capítulo 5 se presentarán las conclusiones y las posibles áreas de mejora del trabajo.

1.2. Equifax: contexto y limitaciones

Equifax es un buró de crédito multinacional, que en conjunto a Transunion y Experian componen los tres más grandes a nivel mundial. La compañía posee equipos de desarrollo en Estados Unidos, India, Irlanda y Chile. Asimismo está operativa en más de 24 países. El negocio principal de Equifax es la información/conocimiento extraído de la data recolectada, la que incluye información crediticia, servicios básicos, autos, mercadotecnia, Twitter, revistas, informaciones demográficas entre otros. El principal desafío tecnológico de la compañía es resguardar la privacidad. El segundo, realizar toda clase de predicciones relevantes para el mercado con los datos acumulados. Los datos son uno de los mayores, si no el mayor activo de la compañía.

Keying and Linking es el equipo de Equifax encargado de identificar entidades y relacionarlas dentro de los diferentes conjuntos de datos, esta labor debe ser aplicada a cada entidad dentro de la compañía y zonas geográficas. La tarea de la identificación de entidades, entity resolution, es el proceso de identificar que dos o más registros de información, que referencian a un único objeto en el mundo real, esto puede ser una persona, lugar o cosa. Por ejemplo, Bob Smith, Robert Smith, Robert S. podría referirse a la misma persona, lo mismo puede darse con una dirección. Es importante destacar que la información requerida para este equipo es de identificación personal (PII), categorizada y protegida con las mayores restricciones dentro de la compañía, de aquí el delicado uso que se dé a los registros y se prohíben el uso de datos reales en ambientes de desarrollo.

La propuesta actual se enmarca en la búsqueda de un método alternativo en la generación de data sintética utilizando inteligencia artificial. La data sintética es utilizada en las pruebas de nuevo software en ambientes no productivos en Equifax. Para el equipo de **Keying and Linking** y la compañía es importante la evaluación de los nuevos desarrollos, pero es aún más importante resguardar la privacidad y seguridad de los datos. Es por ello que la privacidad y calidad de estos datos es relevante.

Los métodos actuales que posee Keying and linking para la generación de data sintética y así probar sus algoritmos son las siguientes, a)) Anonimización de los registros, este método destruye piezas claves de los registros, para asegurar que no puede ser identificado el dueño de la información. b)) Generación de data sintética en base de heurísticas, utilizando conocimiento sobre la estructura de los registros, por ejemplo, DOB (date of birth) establecen rangos de fechas, o formatos en el caso de SSN (Security Social Number) o Tarjetas de créditos. c)) Reemplazo por revuelta de datos, se compone de registros reales, pero mezcla elementos con heurísticas para que no puedan ser identificados, por ejemplo, mezclando nombres, segmentos de SSN, fechas de nacimiento y así con todos los registros involucrados. El sistema de revuelta de datos es el método utilizado, pero debido a peligro de exponer datos reales, fue limitado a generar un único dataset.

Sobre la regulación y acceso directo a información personal legible, no enmascarada en Equifax. Esta se encuentra regulada y solo disponibles para proyectos categorizados como “Protected Data Zone” (PDZ). Estos proyectos están administrados por el equipo de Ignite, encargado de la seguridad y herramientas ofrecidas para dichos espacios de trabajo. Los permisos de acceso son supervisados y revisados cada 3 meses.

Equifax como AI-First Company, está en una evolución en búsqueda de ser precursora en inteligencia artificial, utilizando los datos almacenados durante más de un siglo y su asociación con

Google, principal proveedor de servicios en la nube. El objetivo del año 2022, es poseer capacidades de entrenar modelos de Deep Learning usando las plataformas analíticas actuales administradas por Ignite, el producto seleccionado y está en proceso de implementación es Vertex AI. Equifax está en proceso de evaluación de empresas que generen data sintética con las condiciones que la organización requiere. Uno de los evaluados es Tonic IA <https://www.tonic.ai/>. Esto deja ver la relevancia que los datos sintéticos en los objetivos de Equifax a mediano plazo.

1.3. Objetivo

Objetivo General:

El objetivo general de este trabajo es establecer un mecanismo para la generación de conjuntos de datos sintéticos estructurados, los cuales incluyen texto, y proceder a compararlos con sus equivalentes originales.

Objetivos Específicos:

1. Utilización de modelos generativos capaces de producir nuevos conjuntos de datos sintéticos a partir de datos originales que contienen texto.
2. Evaluar y comparar las características de los conjuntos de datos sintéticos y originales en dos aspectos: propiedades estadísticas y nivel de privacidad, y sus distribuciones.

Capítulo 2

Revisión Bibliográfica

2.1. Tipos de Datos

Los tipos de datos tienen diversas implicaciones en su generación, como su representación, almacenamiento y procesamiento. Los datos estructurados se presentan en la Tabla 2.1 Tipos de datos estructurados.

En 2012, IDC estimó que para 2020, más del 95 % de los datos serían no estructurados [21]. En un análisis posterior, Kiran Adnan y Rehan Akbar [12] encontraron que el texto es el tipo de dato no estructurado que más rápido crece en las publicaciones, seguido por la imagen, el video y finalmente el audio.

La Tabla 2.1 Tipos de datos estructurados resume la lista que se encuentra en *Practical Statistics for Data Scientists* [17].

Tabla 2.1: Tipos de datos estructurados

T	Sub tipo	Descripción	Ejemplos
	Numérico	Datos establecidos como números	-
	Continuo	Datos que pueden tomar cualquier valor en un intervalo	3.14 metros, 1.618 litros
	Discreto	Datos que solo pueden tomar valores enteros	1 habitación, 73 años
	Categorico	Datos que pueden tomar solo un conjunto específico de valores que representan un conjunto de categorías posibles.	-
	Binario	Un caso especial de datos categóricos con solo dos categorías de valores	0/1, verdadero/falso
	Ordinal	Datos categóricos que tienen un ordenamiento explícito.	pequeña/ mediana/ grande

2.2. Privacidad de Datos

La protección de la información es un aspecto fundamental en la generación de datos sintéticos. Aunque este aspecto puede no ser crucial cuando los datos corresponden a temas como recetas o automóviles, resulta esencial cuando se trata de información relacionada con individuos [17]. Por esta razón, el resguardo de la información es un tema de importancia para entidades como Equifax, que gestionan una gran cantidad de conjuntos de datos con contenido personal.

2.2.1. Tipo de datos a ser protegidos

Para identificar qué campos de datos son significativos desde el punto de vista de la privacidad, se puede recurrir a la definición resumida en la Tabla 2.2 Niveles de revelación y ejemplos del texto *Data privacy: Definitions and techniques* [19].

Tabla 2.2: Niveles de revelación y ejemplos

Tipo de revelación	Descripción
Identificadores	Atributos que identifican de manera única a individuos (por ejemplo, SSN, RUT, DNI).
Cuasi-identificadores (QI)	Atributos que, en combinación, pueden identificar a individuos, o reducir la incertidumbre sobre sus identidades (por ejemplo, fecha de nacimiento, género y código postal).
Atributos confidenciales	Atributos que representan información sensible (por ejemplo, enfermedad).
Atributos no confidenciales	Atributos que los encuestados no consideran sensibles y cuya divulgación es inofensiva (por ejemplo, color favorito).

2.2.2. Tipos de riesgos de divulgación

Los tipos de divulgación definidos en *Practical Synthetic Data Generation* [17] están resumidos en la Tabla 2.3 Tipos de Riesgos de Divulgación y sus Descripciones.

Tabla 2.3: Tipos de Riesgos de Divulgación y sus Descripciones

Tipo de revelación	Descripción
Divulgación de identidad	Este riesgo se refiere a la posibilidad de que un atacante pueda identificar la información de un individuo a partir de los datos publicados, utilizando técnicas de filtrado para reducir las posibilidades hasta un solo individuo.
Divulgación de nueva información	Este riesgo comprende el riesgo de Divulgación de Identidad, y además, implica la adquisición de información adicional sobre el individuo a partir de los datos publicados.
Divulgación de Atributos	Este riesgo se da cuando, aunque no se pueda identificar a un individuo, se puede descubrir un atributo común en varios registros, lo que permite obtener información sensible acerca de un grupo de individuos.
Divulgación Inferencial	Este riesgo se refiere a la posibilidad de inferir información sensible a partir de los datos publicados, mediante el uso de técnicas de análisis estadístico o de aprendizaje automático. Por ejemplo, si después de filtrar todos los registros, el 80 % de los registros con las mismas características tienen cáncer, se podría inferir que el individuo buscado puede tener cáncer.

Adicionalmente se deben establecer dos conceptos relevantes ante el análisis de revelación de información:

1. En términos prácticos, normalmente los datos sintéticos buscan tener cierta permeabilidad con respecto a la **Divulgación Inferencial**, ya que se quiere que estadísticamente sean similares. Además, se busca proteger la identidad de los individuos, pero esta no es la única condición, también se busca proteger aquellos atributos que pueden ser sensibles, como las enfermedades. A todo este conjunto se le denomina **Revelación de identidad significativa**. Es particularmente riesgoso por la posibilidad de discriminación hacia ciertos grupos que cumplen con los atributos criterio.
2. Los mismos atributos pueden tener más relevancia para ciertos grupos de la población que para otros. El ejemplo que se indica en [20] es que, debido a que el número de hijos igual a 2 es menos frecuente en una etnia que en otra (40 % en la primera y 10 % en la segunda), ese dato es más relevante en la segunda. Esto se debe a que es un factor que filtra mejor y, por lo tanto, puede permitir un mejor conocimiento de ese grupo específico. A esto se le denomina **Definición de información ganada**.

2.2.3. Regulación de datos sintéticos

Debido a que los datos sintéticos son basados en datos reales, pueden ser afectos a las regulaciones de sobre protección de datos [17]. Los nuevos datos podrían ser afectos por:

1. Regulation (EU) 2016/679 of the European Parliament and of the Council [35], si el proceso de generación de datos sintéticos a menudo implica el uso de datos personales reales como entrada. En este caso, el GDPR sería relevante. Las organizaciones que utilicen datos personales para generar datos sintéticos deben garantizar que este proceso cumple con los principios del GDPR, como la minimización de datos (sólo se deben utilizar los datos necesarios) y la limitación de la finalidad (los datos sólo se deben utilizar para el propósito para el que se recogieron).
2. The California consumer privacy act: Towards a European-style privacy regime in the United States [32]
3. Health insurance portability and accountability act of 1996 [11]

2.2.4. Protección de Privacidad

En la Tabla 2.4 Metricas de privacidad se listas las utilizadas en diferentes publicaciones para determinar la privacidad efectiva de los conjuntos generados.

Tabla 2.4: Metricas de privacidad

Tipo de revelación	Descripción
<i>Distance to Closest Record (DCR)</i>	DRC se utiliza para medir la distancia euclidiana entre cualquier registro sintético y su vecino real más cercano. Idealmente, cuanto mayor sea la DCR, menor será el riesgo de violación de la privacidad. Además, se calcula el percentil 5 de esta métrica para proporcionar una estimación robusta del riesgo de privacidad. [38]
<i>Nearest Neighbour Distance Ratio (NNDR)</i>	NNDR mide la relación entre la distancia euclidiana del vecino real más cercano y el segundo más cercano para cualquier registro sintético correspondiente. Esta relación se encuentra dentro del intervalo [0, 1]. Los valores más altos indican una mayor privacidad. Los bajos valores de NNDR entre datos sintéticos y reales pueden revelar información sensible del registro de datos reales más cercano. [38]

2.3. Generación de Datos Sintéticos

Los datos sintéticos, aunque no son datos reales, se generan con la intención de preservar ciertas propiedades de los datos originales. La utilidad de los datos sintéticos se mide por su capacidad para servir como un sustituto efectivo de los datos originales [17]. Basándose en el uso de los datos originales, los datos sintéticos se pueden clasificar en tres categorías: aquellos que se basan en datos reales, los que no se basan en datos reales, y los híbridos.

Datos basados en datos reales: utilizan modelos que aprenden la distribución de los datos originales para generar nuevos puntos de datos similares.

Datos no basados en datos reales: utilizan conocimientos del mundo real. Por ejemplo, se podría formar un nombre completo seleccionando aleatoriamente un nombre y un apellido de un conjunto predefinido.

Híbridos: estos combinan técnicas de imitación de distribución con algunos campos que no derivan de los datos reales. Esto puede ser especialmente útil cuando se intenta desacoplar las distribuciones de datos que podrían ser sensibles o generar discriminación, como la información sobre la etnia.

En la Sección 2.1 Tipos de Datos, se revisaron los datos estructurados. Si bien cada tipo puede tener muchas representaciones, por ejemplo, los datos continuos podrían considerarse como *float*, *datetime* o incluso intervalos personalizados, como de 0 a 1. Sobre estos datos estructurados, se pueden generar estructuras para unirlos.

Entre las estructuras más comunes se encuentran las matrices bidimensionales (datos tabulares) y los arreglos, que permiten matrices de muchas dimensiones e incluso estructuras complejas que pueden mezclar todas las estructuras previas.

Debido al objetivo, se detallan solo los modelos que permiten abordar la generación de datos tabulares y texto basados en datos reales.

2.3.1. Generación de datos tabulares

En la Tabla 2.5 Estado del arte en generación de datos tabulares, se resumen las últimas publicaciones sobre generación de datos tabulares, indicando la fecha de publicación y si se puede acceder al código fuente o no, a febrero de 2023.

Tabla 2.5: Estado del arte en generación de datos tabulares

Nombre	Fecha ↓	Código
REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers [36]	2023-02-04	Github
PreFair: Privately Generating Justifiably Fair Synthetic Data [34]	2022-12-20	
GenSyn: A Multi-stage Framework for Generating Synthetic Microdata using Macro Data Sources [10]	2022-12-08	Github
TabDDPM: Modelling Tabular Data with Diffusion Models [27]	2022-10-30	Github
Language models are realistic tabular data generators [16]	2022-10-12	Github
Ctab-gan+: Enhancing tabular data synthesis [39]	2022-04-01	Github
Ctab-gan: Effective table data synthesizing [38]	2021-05-31	Github
Modeling Tabular data using Conditional GAN [37]	2019-10-28	Github
Smote: synthetic minority over-sampling technique [18]	2002-06-02	Github

2.3.2. Generación de texto en base de datos tabulares

En la Tabla 2.6 Estado del arte en generación de textos en base a datos, se listan las publicaciones en la generación de texto a partir de datos estructurados.

Tabla 2.6: Estado del arte en generación de textos en base a datos

Nombre	Fecha ↓	Modelo Base
Table-To-Text generation and pre-training with TABT5 [14]	2022-10-17	T5
Text-to-text pre-training for data-to-text tasks [25]	2021-07-09	T5
TaPas: Weakly supervised table parsing via pre-training [22]	2020-04-21	Bert

El estado del arte en la generación de texto a partir de datos tabulares es TabT5. Es importante notar que la tabla mezcla los enfoques de *Table-To-Text* y *Data-To-Text*. Aunque ninguna de las publicaciones incluye código asociado, no es necesario, ya que utilizan modelos abiertos como base (T5 y Bert). Lo más relevante en estos casos es el proceso de *fine-tuning*. Para completar la tarea de generar nuevos textos a partir de información inicial, esta información debe ser codificada para poder ser procesada por el modelo utilizado.

La diferencia entre *Table-To-Text* y *Data-To-Text* radica en el formato de información de entrada. en *Table-To-Text* es una tabla con multiples filas y en *Data-To-Text* corresponde a un solo objeto con sus propiedades. A continuación ejemplos de entradas de los modelos.

En los siguientes ejemplos, se utilizará la Tabla 2.7 Ejemplo de tabla de entrada para ilustrar cómo se puede utilizar para generar texto utilizando los modelos de *fine-tuning* mencionados anteriormente. Esta tabla representa información sobre películas, incluyendo el nombre de la película, el director, el año de lanzamiento y el género, y se utilizará para generar preguntas y respuestas a partir de la información proporcionada.

Tabla 2.7: Ejemplo de tabla de entrada

Nombre de la Película	Director	Año de Lanzamiento	Género
Star Wars: Una Nueva Esperanza	George Lucas	1977	Ciencia ficción

Para los modelos TabT5 y TaPas, se utiliza el mismo preprocesamiento para convertir la tabla de entrada en una pregunta/tarea y respuesta [14, 22]. En este ejemplo, la tabla representa información sobre películas, y se utiliza para generar una pregunta y respuesta sobre el director de la película "Star Wars: Una Nueva Esperanza". La pregunta se construye a partir de la información de la tabla, y la respuesta se espera que sea el nombre del director. Una vez que se ha generado la pregunta y la respuesta, se puede utilizar un modelo de *fine-tuning* como TabT5 o TaPas para generar texto a partir de la información proporcionada. En resumen, el proceso de generación de texto a partir de datos tabulares implica la conversión de información tabular en preguntas y respuestas, y luego la utilización de modelos de *fine-tuning* para generar texto a partir de estas preguntas y respuestas.

Input
Table: Películas Nombre de la Película Director Año de Lanzamiento Género Star Wars: Una Nueva Esperanza George Lucas 1977 Ciencia ficción
Pregunta
¿Qué director dirigió la película Star Wars: Una Nueva Esperanza?
Respuesta esperada
George Lucas

En cambio, el modelo *Text-to-text pre-training for data-to-text tasks* [25] utiliza una entrada diferente, que consiste en una serie de tuplas que representan las propiedades de la entidad y sus valores correspondientes. Se espera que el modelo identifique la tupla relevante y genere una pregunta y respuesta correspondientes. Una vez generada la pregunta y respuesta, se puede utilizar el modelo de fine-tuning correspondiente para generar texto a partir de ellas. En conclusión, la generación de texto a partir de datos tabulares implica una conversión adecuada de la información de entrada en un formato apropiado para cada modelo, la identificación de la pregunta o tarea relevante y la utilización del modelo correspondiente para generar el texto resultante.

Input

<Star Wars: Una Nueva Esperanza, Director, George Lucas>,
<Star Wars: Una Nueva Esperanza, Año de Lanzamiento, 1977>,
<Star Wars: Una Nueva Esperanza, Género, Ciencia ficción>

Pregunta

¿Qué director dirigió la película Star Wars: Una Nueva Esperanza?

Respuesta esperada

George Lucas

2.4. Métricas de evaluación

Es importante destacar que no todas estas métricas son aplicables a todos los tipos de datos y modelos, y que la selección de las métricas a utilizar debe ser cuidadosamente considerada en función de las necesidades y objetivos específicos de cada caso de estudio. A continuación presentan algunas de las posibles a considerar para medir la similitud, privacidad y utilidad en la evaluación de los conjuntos de datos sintéticos generados.

2.4.1. SDMetrics

SDMetrics es una herramienta integral para la evaluación de conjuntos de datos sintéticos. Esta herramienta implementa dos enfoques distintos para el cálculo de métricas: el Reporte y el Diagnóstico [8].

SDMetrics Report

El informe de SDMetrics genera una puntuación de evaluación para un conjunto de datos sintéticos al compararlo con el conjunto de datos reales. Esta puntuación se compone de dos componentes: *Column Shapes* y *Column Pair Trends*.

Column Shapes se compone de las puntuaciones *KSComplement* para columnas numéricas y *TVComplement* para columnas categóricas.

TVComplement se centra en la similitud entre una columna real y una columna sintética en términos de sus formas de distribución.

KSComplement utiliza la estadística de Kolmogorov-Smirnov para calcular la máxima diferencia entre las funciones de distribución acumulativa de dos distribuciones numéricas.

Column Pair Trends se compone de *CorrelationSimilarity* para columnas numéricas y *ContingencySimilarity* para columnas categóricas o combinaciones de columnas categóricas y numéricas.

CorrelationSimilarity mide la correlación entre un par de columnas numéricas y calcula la similitud entre los datos reales y sintéticos al comparar las tendencias de las distribuciones bidimensionales.

ContingencySimilarity mide la similitud entre dos variables categóricas utilizando la tabla de contingencia y la estadística del coeficiente de contingencia, proporcionando una medida de la dependencia entre las dos variables.

Cada una de estas métricas evalúa un aspecto diferente de la calidad de los datos sintéticos, proporcionando información valiosa sobre diversas características de los datos. **TVComplement** se enfoca en la distribución marginal o el histograma unidimensional de la columna, mientras que **KSComplement** se centra en la diferencia entre las funciones de distribución acumulativa de dos

distribuciones numéricas. **CorrelationSimilarity** mide la similitud en la correlación entre un par de columnas numéricas en los datos reales y sintéticos, y **ContingencySimilarity** mide la similitud en la dependencia entre dos variables categóricas en los datos reales y sintéticos utilizando la tabla de contingencia y la estadística del coeficiente de contingencia. Juntas, estas métricas ofrecen una evaluación completa de la calidad de los datos sintéticos.

SDMetrics Diagnostic

SDMetrics Diagnostic emplea una serie de métricas para arrojar luz sobre la calidad de los datos sintéticos, haciendo especial énfasis en *Coverage* y *Boundaries*.

Coverage comprende la media aritmética de *RangeCoverage* y *CategoryCoverage*.

RangeCoverage cuantifica la proporción del rango de valores posibles de una característica que está representada en los datos. Se define como la relación entre el rango de valores observados y el rango de valores posibles para la característica en cuestión. Esta métrica es útil para determinar si los datos proporcionan una representación adecuada de la diversidad de valores que la característica puede adoptar.

CategoryCoverage mide la proporción de categorías posibles de una característica categórica que están representadas en los datos. Se calcula como la relación entre el número de categorías observadas y el número total de categorías posibles para dicha característica. Esta métrica es útil para evaluar si los datos proporcionan una representación adecuada de la diversidad de categorías que la característica categórica puede adoptar.

Boundaries se considera únicamente para la métrica *BoundaryAdherence*, calculada como el promedio para todas las columnas numéricas.

BoundaryAdherence evalúa la proporción de puntos de datos que caen dentro de los límites especificados para una característica. Se determina como la relación entre el número de puntos de datos que caen dentro de los límites y el número total de puntos de datos. Esta métrica es útil para evaluar si los datos se ajustan a los límites especificados para una característica, lo cual puede ser crucial en contextos donde se espera que la característica observe ciertos valores o límites específicos.

En resumen, *SDMetrics Diagnostic* utiliza *RangeCoverage*, *BoundaryAdherence* y *CategoryCoverage* para evaluar la calidad de los datos tabulares sintéticos. Estas métricas aportan una visión detallada de la cobertura de los datos en términos de rango de valores, límites y categorías posibles, lo que puede ayudar a detectar potenciales problemas en la calidad de los datos.

2.4.2. Conjuntos Estadísticos

Tabla 2.8: Listado de conjunto estadísticos

Nombre	Descripción
Media (Mean)	La suma de todos los valores dividido por el número total de valores
Mediana (Median)	El valor que se encuentra en el centro de un conjunto de datos ordenados de menor a mayor. Es decir, la mitad de los valores son mayores que la mediana y la otra mitad son menores
Moda (Mode)	El valor que aparece con mayor frecuencia en un conjunto de datos
Mínimo (Min)	El valor más pequeño en un conjunto de datos
Máximo (Max)	El valor más grande en un conjunto de datos
Percentil (25, 75) (Percentile)	El valor tal que P (25 o 75) por ciento de los datos son menores que él, y el restante (100 - P) por ciento son mayores. Cuando P = 50, el percentil es la mediana
Media Truncada (Trimmed Mean)	El promedio de todos los valores, una vez que se han eliminado un porcentaje de los valores más bajos y un porcentaje de los valores más altos
Outlier	Un valor que se encuentra muy lejos de la mayoría de los valores en un conjunto de datos
Desviación (Deviation)	La diferencia entre un valor observado y la estimación de ese valor
Varianza (Variance)	La medida de cuán dispersos están los valores en un conjunto de datos. Es la suma de los cuadrados de las desviaciones desde la media dividido por $n - 1$, donde n es el número de valores
Desviación Estándar (SD)	La raíz cuadrada de la varianza
Desviación Absoluta Media (MAD)	La media de los valores absolutos de las desviaciones desde la media
Rango (Range)	La diferencia entre el valor más grande y el valor más pequeño en un conjunto de datos
Tablas de Frecuencia (Frequency Tables)	Un método para resumir los datos al contar cuántas veces ocurre cada valor en un conjunto de datos
Probabilidad (Probability)	La medida de la posibilidad de que un evento ocurra. Se establece como el número de ocurrencias de un valor dividido por el número total de ocurrencias
Tabla de Contingencia (Contingency Table)	Una tabla que muestra la distribución conjunta de dos o más variables categóricas
Correlación	Una medida estadística que indica cómo dos variables numéricas están relacionadas entre sí. Puede variar entre -1 y 1
Distribución Estratificada	Una comparación de la distribución de datos para diferentes estratos
Continúa en la siguiente página	

Nombre	Descripción
Comparación de Modelos Predictivos Multivariantes	Un método para comparar varios modelos predictivos que involucran múltiples variables. Implica la construcción de modelos separados para cada variable objetivo y comparar la curva ROC (Receiver Operating Characteristic) para cada modelo
Distinguibilidad	Un método para evaluar la calidad de los conjuntos de datos sintéticos. Implica la creación de un modelo que intenta distinguir entre conjuntos de datos reales y sintéticos. Un buen conjunto sintético es aquel que el modelo no puede distinguir de los datos reales
Kullback-Leibler	Una medida de la divergencia entre dos distribuciones de probabilidad
Pairwise Correlation	Una medida de la similitud entre dos conjuntos de datos que compara las correlaciones de cada par de variables en los conjuntos de datos
Log-Cluster	Un método para evaluar la calidad de los conjuntos de datos sintéticos que compara la estructura de los conjuntos de datos reales y sintéticos mediante el uso de clustering
Cobertura de Soporte (Support Coverage)	Una medida de qué tan bien los datos sintéticos representan la distribución de los datos reales. Se mide como la proporción de variables en el conjunto de datos real que están representadas en el conjunto de datos sintéticos
Cross-Classification	Un método para evaluar la calidad de los conjuntos de datos sintéticos que compara la precisión de los modelos predictivos construidos a partir de los conjuntos de datos reales y sintéticos
Métrica de Revelación Involuntaria	Una medida de qué tan bien se protege la privacidad de los datos en un conjunto de datos sintético. Se mide como la tasa de predicciones correctas de atributos sensibles de un individuo en un conjunto de datos sintético

Capítulo 3

Desarrollo

3.1. Recursos disponibles

3.1.1. Conjuntos de datos

A continuación se listan y detallan los conjuntos de datos utilizados en los experimentos.

King County

El conjunto de datos King County [24] contiene información sobre precios de venta y características de 21,613 viviendas en Seattle y King County de los años 2014 y 2015. El conjunto de datos incluye información como el número de habitaciones, el número de baños, la superficie del terreno y la superficie construida, así como información sobre la ubicación de la propiedad, como la latitud y la longitud. Este conjunto de datos es comúnmente utilizado para tareas de regresión y predicción de precios de viviendas. Sus campos se listan en Tabla 3.1 Conjunto de datos King County.

Tabla 3.1: Conjunto de datos King County

Variable	Descripción
id	Identificación
date	Fecha de venta
price	Precio de venta
bedrooms	Número de dormitorios
bathrooms	Número de baños
sqft_liv	Tamaño del área habitable en pies cuadrados
sqft_lot	Tamaño del terreno en pies cuadrados
floors	Número de pisos
waterfront	'1' si la propiedad tiene vista al mar, '0' si no
view	Índice del 0 al 4 de la calidad de la vista de la propiedad
condition	Condición de la casa, clasificada del 1 al 5
grade	Clasificación por calidad de construcción que se refiere a los tipos de materiales utilizados y la calidad de la mano de obra. Los edificios de mejor calidad (grado más alto) cuestan más construir por unidad de medida y tienen un valor más alto. Información adicional en: KingCounty
sqft_above	Pies cuadrados sobre el nivel del suelo
sqft_basmt	Pies cuadrados debajo del nivel del suelo
yr_built	Año de construcción
yr_renov	Año de renovación. '0' si nunca se ha renovado
zipcode	Código postal de 5 dígitos
lat	Latitud
long	Longitud
sqft_liv15	Tamaño promedio del espacio habitable interior para las 15 casas más cercanas, en pies cuadrados
sqft_lot15	Tamaño promedio de los terrenos para las 15 casas más cercanas, en pies cuadrados
Shape_leng	Longitud del polígono en metros
Shape_Area	Área del polígono en metros

Economicos

Economicos.cl es un sitio web chileno que se dedica a la publicación de avisos clasificados en línea, principalmente en las categorías de bienes raíces, vehículos, empleos, servicios y productos diversos. El conjunto de datos corresponde a un *Web Scraping* realizado en 2020, contiene 22.059 observaciones.

Tabla 3.2: Conjunto de datos Economicos.cl

Variable	Descripción
url	URL de la publicación
Descripción	Descripción de la publicación
price	Precio de venta, en dolares, UF o pesos
property_type	Tipo de propiedad: Casa, Departamento, ETC
transaction_type	Tipo de transacción Arriendo, Venta
state	Región de la publicación
county	Comuna de la publicación
publication_date	Día de la publicación
rooms	Número de dormitorios
bathrooms	Número de baños
m_built	Tamaño del área habitable en metros cuadrados
m_size	Tamaño del terreno en metros cuadrados
source	Diario de la publicación
title	Título de la publicación
address	Dirección de la publicación
owner	Publicante
_price	Precio traspasado a UF

3.1.2. Computación y Software

Para llevar a cabo los experimentos, se utilizó un computador con las siguientes especificaciones técnicas, como se muestra en la Tabla 3.3 Computador Usado. El procesador empleado fue un AMD Ryzen 9 7950X 16-Core Procesadores, con cuatro módulos de 32 GB para una memoria total de 128 GB DDR5. La tarjeta gráfica empleada fue una NVIDIA GeForce RTX 4090, y se contó con dos discos duros de 500 GB SSD. La utilización de un equipo con estas características permitió una ejecución eficiente de los modelos de generación de datos, asegurando la viabilidad de los experimentos. Es importante destacar que la elección de los componentes del computador fue cuidadosamente considerada para asegurar que los resultados obtenidos no se vieran limitados por un hardware insuficiente.

En relación al software utilizado, se trabajó con el sistema operativo Ubuntu 20.04.2 LTS y se empleó el lenguaje de programación Python 3.10 para el desarrollo de los modelos de generación de datos. Se utilizaron diversas bibliotecas, incluyendo DVC, SDV y PyTorch, cuya lista completa se puede encontrar en el repositorio en Github. La elección de estas herramientas se basó en la compatibilidad con el modelo TabDDPM, el cual fue utilizado en algunos de los experimentos.

Tabla 3.3: Computador Usado

Componente	Descripción
Procesador	AMD Ryzen 9 7950X 16-Core Processor
Memoria RAM	128 GB DDR5
Tarjeta gráfica	NVIDIA GeForce RTX 4090
Disco duro	1 TB SSD

En favor de la reproducibilidad, se utilizó *devcontainer*, el cual establece el entorno de desarrollo y pruebas mediante una imagen de *Docker* replicable. Los experimentos pueden ser replicados utilizando el contenedor descrito en el repositorio.

```

1  {
2    "name": "SyntheticData",
3    "image": "nvidia/cuda:12.1.0-devel-ubuntu22.04",
4    "extensions": [
5      "jebbs.plantuml",
6      "ms-toolsai.jupyter-keymap",
7      "MS-CEINTL.vscode-language-pack-es",
8      "SimonSiefke.svg-preview",
9      "adamvoss.vscode-languagetool",
10     "mathematic.vscode-latex",
11     "malthehei.latex-citations",
12     "James-Yu.latex-workshop",
13     "valentjn.vscode-ltex",
14     "yzhang.markdown-all-in-one",
15     "ms-python.python",
16     "ms-azuretools.vscode-docker",
17     "ms-toolsai.jupyter"
18   ],
19   "postCreateCommand": "bash ./devcontainer/postscript.sh",
20   "runArgs": ["--gpus", "all"],
21   "settings": {
22     "terminal.integrated.shell.linux": "/bin/bash"
23   },
24   "features": {
25     "ghcr.io/devcontainers/features/python:1": {"version": "3.10"}
26   },
27   "mounts": [
28     "source=${localEnv:HOME}/models,target=/models,type=bind"
29   ]
30 }

```

Código 1: Devcontainer del actual proyecto.

El código fuente de los modelos de generación de datos, así como los scripts de análisis y visualización de los resultados, se encuentra disponible en un repositorio público de Github: [gvillarroel/synthetic-data-for-text](#)

3.2. Desarrollo del flujo de procesamiento

A continuación se describe el flujo de procesamiento utilizado para generar nuevos datos sintéticos. Este flujo se basa en el propuesto por Synthetic Data Vault (SDV), con algunas modificaciones para guardar etapas intermedias.

SDV es un ecosistema de bibliotecas de generación de datos sintéticos que permite a los usuarios aprender conjuntos de datos de una sola tabla, de múltiples tablas y de series de tiempo, y luego generar nuevos datos sintéticos con las mismas propiedades estadísticas y el mismo formato que los conjuntos de datos originales. Para ello, SDV utiliza diferentes técnicas, como modelos generativos y redes neuronales, para aprender la distribución subyacente de los datos y generar nuevos datos que sigan dicha distribución [26, 33].

A continuación se describe el proceso de generación de datos sintéticos para una tabla única utilizando la biblioteca Synthetic Data Vault (SDV), seguido de las modificaciones realizadas para extender el proceso y agregar nuevos modelos.

En la Figura 3.1 Proceso para generar datos sintéticos con SDV se muestran los pasos necesarios para generar un conjunto de datos sintéticos utilizando SDV:

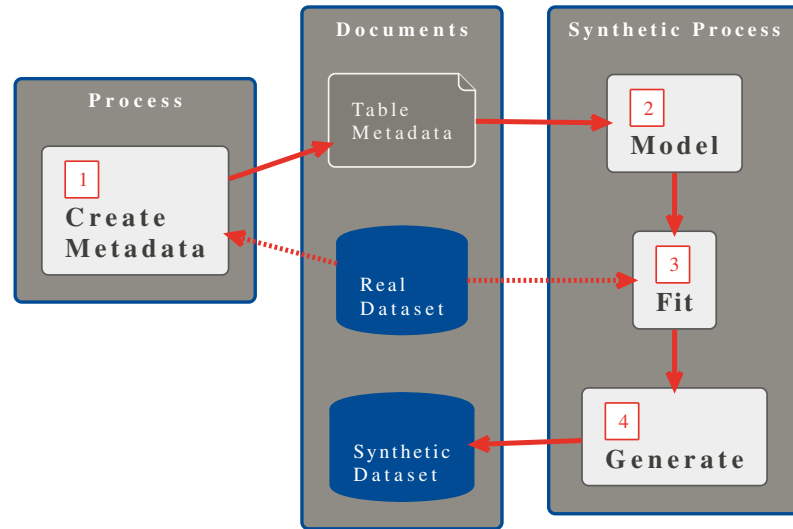


Figura 3.1: Proceso para generar datos sintéticos con SDV

1. **Create Metadata:** Se crea un diccionario que define los campos del conjunto de datos y los tipos de datos que posee. Esto permite a SDV aprender la estructura del conjunto de datos original y utilizarla para generar nuevos datos sintéticos con la misma estructura.
2. **Create Model:** Se selecciona el modelo de generación de datos a utilizar. SDV ofrece varios modelos, incluyendo GaussianCopula, CTGAN, CopulaGAN y TVAE, que se adaptan a diferentes tipos de datos y distribuciones.
3. **Fit Model:** El modelo seleccionado se entrena con el conjunto de datos original para aprender sus distribuciones y patrones estadísticos.
4. **Generate Synthetic Dataset:** Con el modelo ya entrenado, se generan nuevos datos sintéticos con la misma estructura y características estadísticas que el conjunto original. Este nuevo conjunto de datos puede ser utilizado para diversas aplicaciones, como pruebas de software o análisis de datos sensibles.

Es importante destacar que el proceso de generación de datos sintéticos con SDV es escalable y puede utilizarse con conjuntos de datos de una sola tabla, múltiples tablas y series de tiempo. Además, en este proyecto se realizaron algunas modificaciones al flujo para extender el proceso y permitir la inyección de nuevos modelos.

En el proceso de generación de datos sintéticos con SDV extendido, se incluyen dos nuevas etapas para poder guardar los modelos intermedios y los resultados de la evaluación. El proceso completo se muestra en la Figura 3.2 Proceso para generar datos sintéticos completo y consta de los siguientes pasos:

1. **Create Metadata:** Crea un diccionario que define los campos del conjunto de datos y los tipos de datos que posee.
2. **Create Model:** Se selecciona el modelo a utilizar. SDV permite GaussianCopula, CTGAN, CopulaGAN y TVAE.
3. **Fit Model:** El modelo seleccionado toma el conjunto original para entrenar el modelo y aprender sus distribuciones.
4. **Save Model:** El modelo entrenado se guarda en un archivo para su uso posterior.
5. **Generate Synthetic Dataset:** Genera un nuevo conjunto de datos usando el modelo entrenado.
6. **Evaluate & Save Metrics:** Evalúa y guarda el conjunto de datos sintético generado mediante métricas como la correlación, el error absoluto medio y el error cuadrático medio.

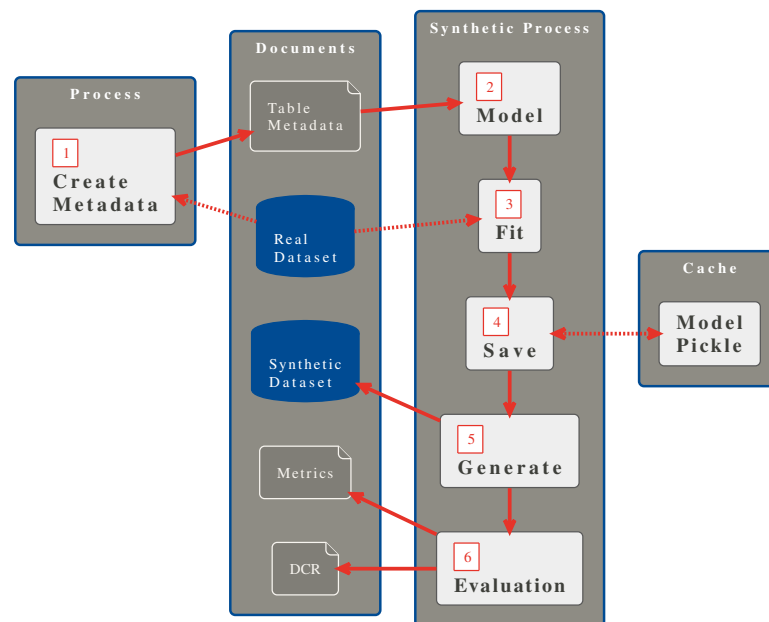


Figura 3.2: Proceso para generar datos sintéticos completo

Con estas nuevas etapas, se pueden guardar los modelos intermedios y los resultados de la evaluación, lo que permite una mayor flexibilidad en el proceso y la capacidad de utilizar los modelos y los resultados en posteriores experimentos.

3.3. Modelos

Los modelos de generación de datos tabulares utilizan como base la metodología propuesta por *Synthetic Data Vault* (SDV), mientras que para los modelos de generación de texto se utiliza la biblioteca Hugging Face para cargar, realizar *fine-tuning* con nuevas tareas y evaluar el modelo basado en mT5.

3.3.1. Modelos para datos tabulares

Para que un modelo pueda ser utilizado con el SDV, es necesario que implemente los siguientes métodos:

1. **load**: Carga el modelo desde un archivo
2. **fit**: Entrena el modelo, utilizando un pandas dataframe como entrada
3. **save**: Guarda el modelo en un archivo
4. **sample**: Genera un conjunto de registros nuevos utilizando el modelo entrenado.

Como consideración adicional, se recomienda ejecutar el proceso utilizando un script en lugar de un notebook, ya que se ha observado que el notebook puede fallar con algunos modelos debido a limitaciones de memoria. A continuación, se detallan los pasos a seguir para la ejecución del proceso:

1. Crear un archivo de configuración que contenga la información necesaria para la generación de datos sintéticos, como la ruta del conjunto de datos original y la configuración de los modelos a utilizar.
2. Crear un script que cargue la configuración, ejecute el proceso de generación de datos sintéticos y guarde el conjunto de datos sintético resultante.
3. Ejecutar el script creado en el paso anterior.

De esta manera, se puede ejecutar el proceso de generación de datos sintéticos de forma automatizada y con una mayor capacidad de procesamiento, lo que puede mejorar el desempeño del proceso y reducir los tiempos de ejecución. Vea ??

La clase *Synthetic* es una implementación que permite configurar los modelos a utilizar en el proceso de generación de datos sintéticos. Esta clase encapsula los métodos comunes de los modelos, como *load*, *fit*, *save* y *sample*, permitiendo así una configuración general de las entradas y la selección de modelos.

En el ejemplo mostrado en el Código 2 Instanciando clase *Synthetic*, se instancia la clase *Synthetic* con un pandas dataframe previamente pre-procesado. Se especifican las columnas que se considerarán como categorías, las que se considerarán como texto y las que se excluirán del análisis. Además, se indica el directorio donde se guardarán los archivos temporales, se seleccionan los

modelos a utilizar, se establece el número de registros sintéticos deseados y se define una columna objetivo para realizar pruebas con machine learning y estratificar los conjuntos parciales de datos que se utilizarán. De esta manera, se configura de manera flexible el proceso de generación de datos sintéticos según las necesidades específicas del usuario.

```

45     syn = Synthetic(df_converted,
46                     id="url",
47                     category_columns=category_columns,
48                     text_columns=("description", "price", "title", "address", "owner",),
49                     exclude_columns=tuple(),
50                     synthetic_folder = "../datasets/economicos/synth",
51                     models=['copulagan', 'tvae', 'gaussiancopula', 'ctgan', 'smote-enc'],
52                     n_sample = df.shape[0],
53                     target_column="_price"
54     )

```

Código 2: Instanciando clase Synthetic

La Tabla 3.4 Variables de entrada para *Synthetic* presenta las opciones para la instancia de la clase *Synthetic*:

Tabla 3.4: Variables de entrada para *Synthetic*

Variable	Descripción
df	Pandas DataFrame a utilizar
Id	Nombre de la columna a ser usada como identificadora
category_columns	Listado de columnas categóricas
text_columns	Listado de columnas de texto
exclude_columns	Listado de columnas que deben ser excluidas
synthetic_folder	Carpeta donde se guardarán los documentos intermedios y finales
models	Listado de modelos a utilizar
n_sample	Número de registros a generar
target_column	Columna a utilizar como objetivo para modelos de machine learning en las evaluaciones y separación cuando se deba estratificar los campos.

En la Tabla 3.5 Modelos Tabulares Soportados se detallan los modelos actualmente soportados en la clase *Synthetic* y su origen.

Tabla 3.5: Modelos Tabulares Soportados

Nombre Modelo	Fuente
copulagan	SDV [26]
tvae	SDV [26]
gaussiancopula	SDV [26]
ctgan	SDV [26]
tablepreset	SDV [26]
smote-enc	tabDDPM [13]
tddpm_mlp	tabDDPM [13]

Al ejecutar el script de generación de datos sintéticos, se crearán múltiples archivos en una carpeta. En la Figura 3.3 Carpetas y archivos esperados generados por *Synthetic* se muestra un ejemplo de los archivos generados y su formato. El nombre del modelo utilizado se indica en el campo **<model>**, y en caso de haberse aplicado *Differential Privacy* para generar una versión con ruido. El campo **<n_sample>** indica el número de registros sintéticos generados, y finalmente el campo **<type_comparison>** especifica si se trata de una comparación entre los datos sintéticos y los datos de entrenamiento (*Synthetic vs Train*, abreviado como ST) o entre los datos sintéticos y los datos de validación (*Synthetic vs Hold*, abreviado como SH). Adicionalmente se encuentran los archivos de esquema (*metadata.json*) y una separación del dataset inicial en el conjunto de entrenamiento y test (hold).

```

synth/
├── checkpoint/
│   ├── copulagan.ckp
│   ├── copulagan_noise.ckp
│   └── <model>.ckp
├── data/
│   ├── copulagan_21613.parquet
│   ├── copulagan_noise_21613.parquet
│   ├── <model>_<n_sample>.parquet
│   ├── privacy_copulagan_21613_SH.npy
│   ├── privacy_copulagan_21613_ST.npy
│   ├── privacy_copulagan_noise_21613_SH.npy
│   ├── privacy_copulagan_noise_21613_ST.npy
│   └── privacy_<model>_<n_sample>_<type_comparison>.npy
├── report/
│   ├── copulagan_21613.rpt
│   ├── copulagan_noise_21613.rpt
│   └── <model>_<n_sample>.rpt
├── split/
│   ├── train.parquet
│   └── hold.parquet
├── metadata.json
└── metadata_noise.json

```

Figura 3.3: Carpetas y archivos esperados generados por *Synthetic*

3.3.2. Modelos para textos

Como se mencionó anteriormente, se utilizó el modelo **mT5** que se entrenó para una nueva tarea utilizando la estrategia presentada en el artículo *Text-to-Text Pre-Training for Data-to-Text Tasks* [25]. Para ilustrar el proceso, se presenta un ejemplo del texto pre-procesado, el segmento de la pregunta y la respuesta esperada para un registro del conjunto de datos *economicos.cl*.

Input

<fecha, 2022-01-01>
<precio, \$ 105.000.000>
<tipo, Departamento>
<transacción, Venta>
<región, Metropolitana de Santiago>
<comuna, Santiago>
<dormitorios, 3.0>
<baños, 3.0>
<construidos, 47.0>
<terreno, 47.0>
<precio_real, 3387.4540447373292>
<titulo, Departamento en Venta en Santiago 3 dormitorios 1 baño>
<dirección, DEPARTAMENTO EN EL CORAZON DE LO BARNECHEA Santiago, Metropolitana de Santiago>

Tarea

descripción de esta publicación

Respuesta esperada

Kazona Propiedades Vende Departamento de 47m2, 3 dormitorios, 1 baño, cocina, living comedor , Paredes con Cerámica y Tabiquería en techo con madera barnizada timbrada, ventanas nuevas de PVC y vidrio termolaminado, sistema eléctrico actualizado, departamento ubicado en el 3er nivel (sin ascensor) , bajo gasto común. Excelentes conectividades y ubicación en Pleno Centro De Lo Barnechea, como colegios privados y públicos, supermercados, Mall Portal La Dehesa, locomoción, entre otros.
Podemos destacar de la propiedad:
Pleno Centro Lo Barnechea
100 metros de locomoción a Escuela Militar , Bilbao, Stgo Centro, Mapocho
200 metros colegios Montessori Nido de Águila, San Rafael , otros
200 metros Mall Portal La Dehesa
200 metros Sta. Isabel
300 metros carabineros
Gastos comunes bajos \$10.000
Estacionamiento comunitario
No paga contribuciones
Contactanos al telefono Kazona 569 56031154

3.4. Obtención de Métricas

Se han automatizado la mayoría de las métricas para evaluar los conjuntos de datos sintéticos mediante el módulo *metrics*. Estas métricas se aplican a los tres conjuntos de datos para su evaluación, lo que permite calcular estadísticas y comparativas para el conjunto de datos real utilizado para el entrenamiento (train dataset), el conjunto de datos reservado para la evaluación (hold) y el conjunto de datos sintético generado por los diferentes modelos (synthetic). Se pueden recolectar ejecutando el ejemplo de código proporcionado en Código 3 Obtención de métricas en `economicos_run-a.py`.

```
55     print(syn.train.loc[:, syn._selectable_columns()])
```

Código 3: Obtención de métricas en `economicos_run-a.py`

En la Tabla 3.6 Metricas para campos numericos se muestra las metricas recolectadas para campos numericos.

Tabla 3.6: Metricas para campos numericos

Campo	Ejemplos
Nombre del campo (name)	sqft_living
Valores del Top 5 (top5)	[1400 1300 1720 1250 1540]
Frecuencia Top 5 (top5_freq)	[109 107 106 106 105]
Probabilidades de Top 5 (top5_prob)	[0.00630422 0.00618855 0.00613071 0.00613071 0.00607287]
Elementos observados (nobs)	17290
Nulos (missing)	0
Promedio (mean)	2073.894910
Desviación Estándar (std)	907.297963
Error estándar de la media (std_err)	6.900053
Intervalo de confianza superior (upper_ci)	2087.418766
Intervalo de confianza inferior (lower_ci)	2060.371055
Rango intercuartílico (iqr)	1110
Continúa en la siguiente página	

Campo	Ejemplos
Rango intercuartílico normalizado (iqr_normal)	822.844231
Desviación absoluta de la mediana (mad)	693.180169
Desviación absoluta de la mediana normalizada (mad_normal)	868.772506
Coefficiente de variación (coef_var)	0.437485
Rango (range)	11760
Valor máximo (max)	12050
Valor mínimo (min)	290
Sesgo (skew)	1.370859
Curtosis (kurtosis)	7.166622
Test de normalidad de Jarque-Bera (jarque_bera)	17922.347382
Valor p del test de normalidad de Jarque-Bera (jarque_bera_pval)	0
Moda (mode)	1400
Frecuencia de la moda (mode_freq)	0.006304
Mediana (median)	1910
Percentil 0.1 %	522.890000
Percentil 1 %	720
Percentil 5 %	940
Percentil 25 %	1430
Percentil 75 %	2540
Percentil 95 %	3740
Percentil 99 %	4921.100000
Percentil 99.9 %	6965.550000

En la Tabla 3.7 Métricas para campos categóricos se muestran los datos calculados para campos categóricos.

Tabla 3.7: Métricas para campos categóricos

Nombre del campo (name)	waterfront
Valores del Top 5 (top5)	[0 1]
Frecuencia Top 5 (top5_freq)	[17166 124]
Probabilidades de Top 5 (top5_prob)	[0.99282822 0.00717178]
Elementos observados (nobs)	17290.0
Nulos (missing)	17290.0

En el Código 4 Mostrando Scores Promedios Calculados, se muestra cómo se calcula y se muestra el Score promedio para una selección específica de modelos. El código utiliza la función "sort_values" para ordenar los resultados en orden descendente según el puntaje. Luego, se filtran los resultados para incluir solo los modelos seleccionados y las columnas que muestran el puntaje y la Distancia al registro más cercano (DCR) en los tres umbrales *Synthetic vs Train* (ST), *Synthetic vs Hold* (SH) y *Train vs Hold* TH.

```

1 avg = syn.scores[syn.scores["type"] == "avg"]
2 avg.sort_values("score", ascending=False).loc[
  → ["ntddpm_mlp", "smote-enc", "gaussiancopula", "tvae", "gaussiancopula",
  → "copulagan", "ctgan"], ["score", "DCR ST 5th", "DCR SH 5th", "DCR TH 5th"]]

```

Código 4: Mostrando Scores Promedios Calculados

El Score calculado se obtiene a través de SDV y se basa en cuatro métricas: KSComplement, TVComplement que conforman *Column Shapes*, ContingencySimilarity y CorrelationSimilarity conforman *Column Pair Trends*. Además, para mostrar los resultados, se proporciona un ejemplo de código en el Código 4 Mostrando Scores Promedios Calculados y un ejemplo de resultado en la Tabla 3.8 Ejemplo de scores promedios.

Tabla 3.8: Ejemplo de scores promedios

Nombre	Column Pair Trends	Column Shapes	Score ↓	DCR ST	DCR SH	DCR TH
ntddpm_mlp	0.954	0.971	0.962	0.084	0.104	0.035
nsmote-enc	0.941	0.967	0.954	0.058	0.090	0.035
<model>	0.941	0.967	0.954	0.058	0.090	0.035

Capítulo 4

Resultados

Este proyecto se ha centrado en la generación de datos sintéticos a través de diversos métodos de preprocesamiento y modelos de aprendizaje automático. Los resultados se examinan en base al rendimiento de los modelos, los cuales fueron entrenados con los datos sintéticos, y se valoran respecto a la similitud, privacidad y utilidad de los datos generados.

Cabe mencionar que los resultados son inherentes a cada conjunto de datos y modelo utilizado. Por lo tanto, se proporciona un análisis exhaustivo de los resultados en cada escenario específico. Esto permite una mejor comprensión de la eficacia de los métodos utilizados en la generación de datos sintéticos y su comparación con los datos originales.

A continuación, se presentan los conjuntos de datos de King County y Económicos. El conjunto Económicos se subdivide en dos subconjuntos, que difieren en su tratamiento de los datos nulos durante el preprocesamiento.

4.1. King County

4.1.1. SDMetrics Score

La Tabla 4.1 muestra los puntajes obtenidos por los distintos patrones utilizados en este estudio. Es notorio que los patrones con puntajes más altos, como Tddpm y Smote, presentan una mayor similitud con el conjunto de datos original. En contraposición, los patrones con puntajes más bajos, como ctgan, exhiben una correspondencia considerablemente menor con el conjunto original. Se muestra el promedio \pm desviación estándar basado en las 3 ejecuciones realizadas.

Tabla 4.1: Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, King County

Model Name	Column Pair Trends	Column Shapes	Coverage	Boundaries	Score
tddpm_mlp	9.37e-01 \pm 3.80e-03	9.67e-01\pm1.48e-03	9.66e-01\pm4.96e-03	1.00e+00 \pm 0.00e+00	9.52e-01 \pm 2.36e-03
smote-enc	9.41e-01\pm2.60e-04	9.65e-01 \pm 3.06e-04	8.42e-01 \pm 8.31e-03	1.00e+00\pm1.02e-05	9.53e-01\pm2.45e-04
ctgan	8.10e-01 \pm 1.40e-02	8.38e-01 \pm 2.67e-02	8.56e-01 \pm 2.25e-03	1.00e+00 \pm 0.00e+00	8.24e-01 \pm 2.02e-02
tablepreset	8.37e-01 \pm 0.00e+00	8.37e-01 \pm 1.36e-16	7.53e-01 \pm 0.00e+00	1.00e+00 \pm 0.00e+00	8.37e-01 \pm 7.85e-17
copulagan	7.64e-01 \pm 4.93e-03	8.14e-01 \pm 4.70e-03	8.40e-01 \pm 1.74e-02	1.00e+00 \pm 0.00e+00	7.89e-01 \pm 2.92e-03
gaussiancopula	7.65e-01 \pm 0.00e+00	8.11e-01 \pm 0.00e+00	7.51e-01 \pm 7.85e-17	1.00e+00 \pm 0.00e+00	7.88e-01 \pm 0.00e+00
tvae	7.07e-01 \pm 1.19e-02	7.68e-01 \pm 1.22e-02	4.53e-01 \pm 1.63e-02	1.00e+00 \pm 0.00e+00	7.38e-01 \pm 1.18e-02

A pesar de que los patrones Tddpm y Smote alcanzan calificaciones prometedoras en general, se observa una diferencia significativa entre ambos en términos de cobertura (*Coverage*). Específicamente, Smote no logra capturar la diversidad del conjunto de datos, reflejándose en una calificación de cobertura marcadamente inferior a la de Tddpm.

4.1.2. Correlación

En el Anexo A.2, se contrasta la lista completa de cada modelo. Se observa que, en general, los modelos con puntajes más altos exhiben una mayor similitud visual con los datos reales. A modo de ilustración, las imágenes 4.1 y 4.2 contrastan los datos reales con los generados por los modelos gaussiancopula y copulagan. A pesar de que estos modelos presentan puntajes similares, el modelo gaussiancopula muestra una mayor similitud visual con los datos reales en comparación con el modelo copulagan.

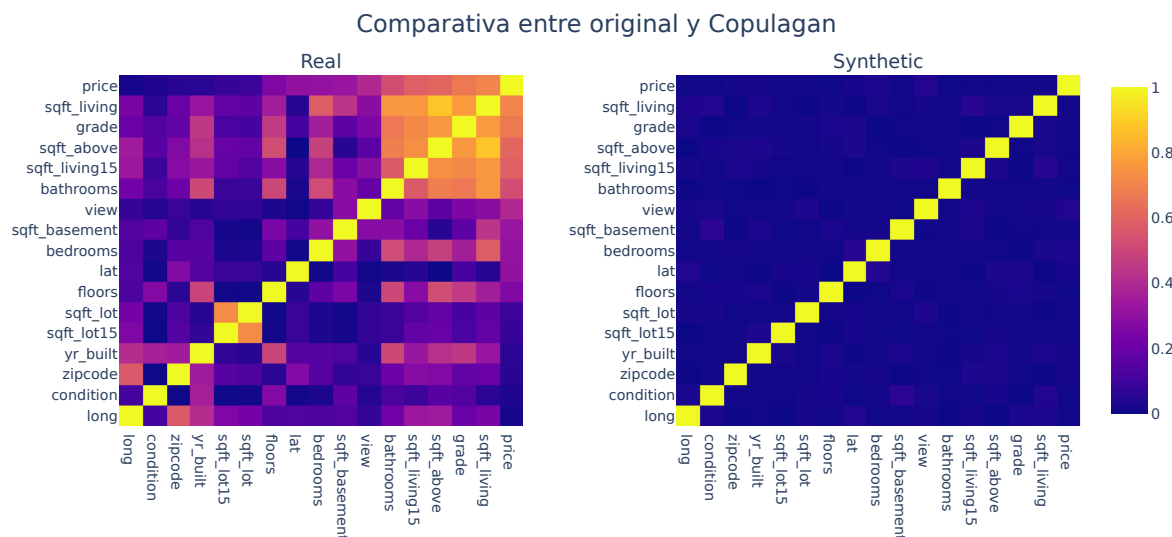


Figura 4.1: Correlación de conjunto original de entrenamiento y Copulagan

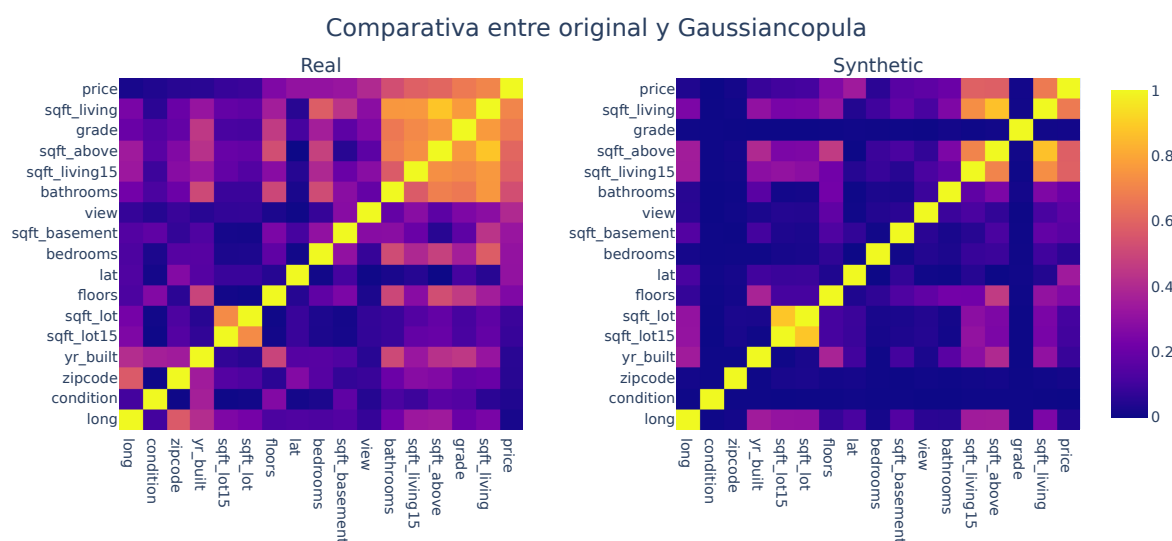


Figura 4.2: Correlación de conjunto original de entrenamiento y Gaussiancopula

Es especialmente relevante que, entre los modelos con puntajes superiores al 90 %, la evaluación visual para determinar cuál es superior puede ser un desafío. Esta dificultad surge debido a que, a medida que el puntaje se incrementa, la similitud visual entre los datos reales y los generados se intensifica. Este fenómeno se ilustra en las figuras 4.3 y 4.4, donde se contrastan los datos reales con los generados por los modelos Smote y Tddpm, respectivamente. Ambos modelos ostentan puntajes por encima del 90 %, y la correspondencia visual entre los datos reales y los generados es notablemente alta en ambos casos.

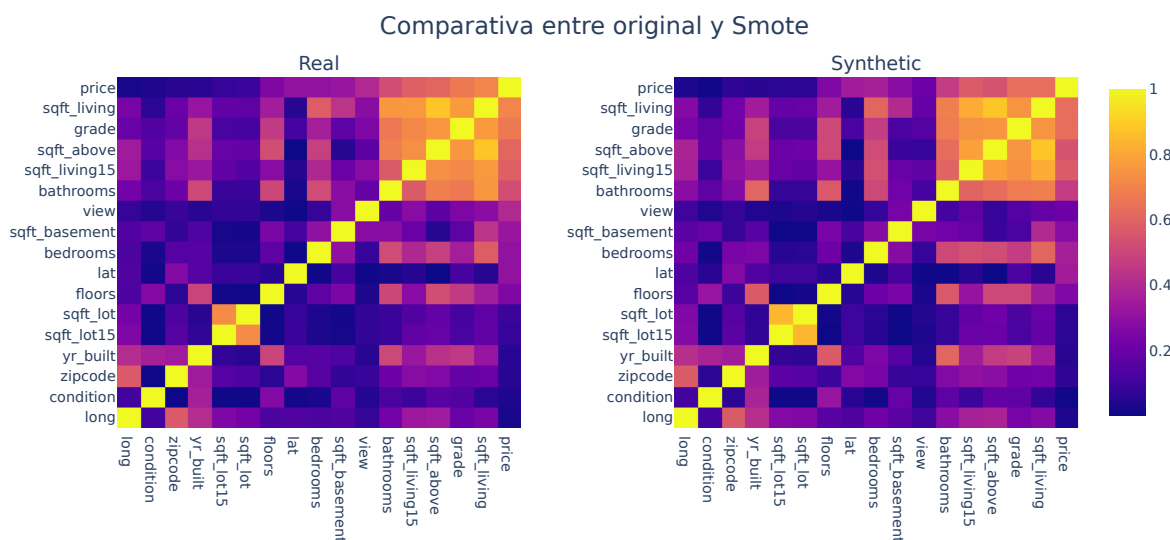


Figura 4.3: Correlación de conjunto original de entrenamiento y Smote

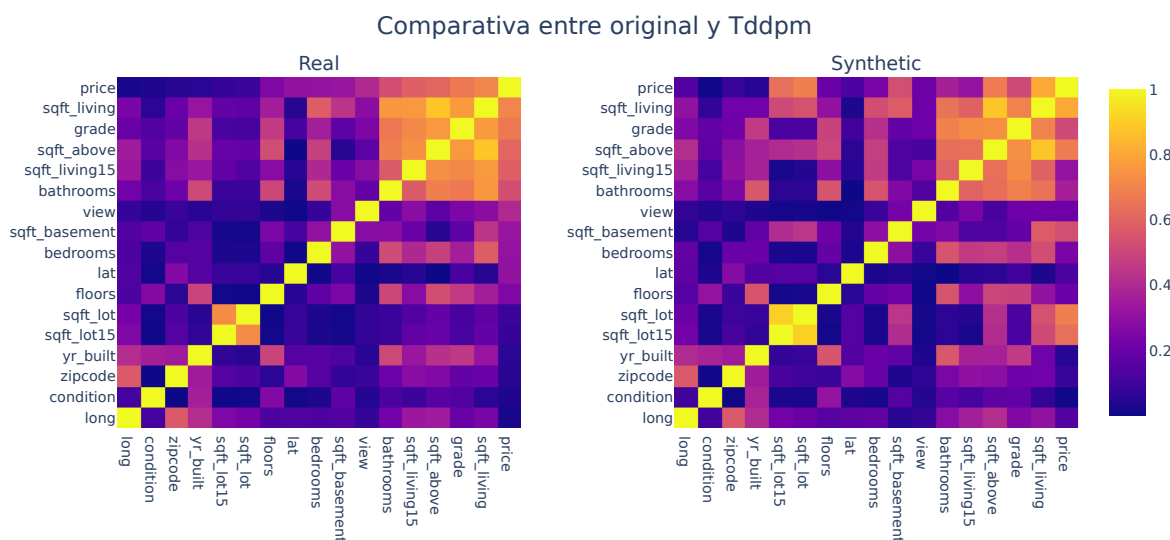


Figura 4.4: Correlación de conjunto original de entrenamiento y Tddpm

En la evaluación mediante SDMetrics y en la comparación visual a través de la correlación de en parejas, los modelos más sobresalientes resultan ser Tddpm y Smote. Dichos modelos han logrado los puntajes más elevados en ambas métricas y han demostrado una notable similitud visual con los datos reales. Por ende, se puede inferir que estos modelos resultan ser los más eficaces para la generación de datos sintéticos beneficiosos para este conjunto de datos en particular.

4.1.3. Cobertura

La Tabla 4.2 evidencia la superioridad del modelo Tddpm en términos de cobertura de valores distintos, aunque hay casos donde ningún modelo alcanza una cobertura completa. Un caso notable es la variable *bedrooms*, en la que Tddpm solo logra un 71.8 % de cobertura, pero aún así supera al modelo Smote, que apenas alcanza el 51 % para la misma variable.

Tabla 4.2: Cobertura Categoría/Rango para Modelos Smote y Tddpm, King County

Columna	Metrica	smote-enc	tddpm_mlp
bathrooms	CategoryCoverage	6.56e-01±3.85e-02	8.11e-01±3.85e-02
bedrooms	CategoryCoverage	5.13e-01±4.44e-02	7.18e-01±4.44e-02
condition	CategoryCoverage	9.33e-01±1.15e-01	1.00e+00±0.00e+00
date	CategoryCoverage	9.64e-01±6.77e-03	9.44e-01±9.69e-03
floors	CategoryCoverage	8.33e-01±0.00e+00	9.44e-01±9.62e-02
grade	CategoryCoverage	7.50e-01±0.00e+00	8.61e-01±4.81e-02
id	RangeCoverage	9.93e-01±4.54e-04	1.00e+00±7.79e-04
lat	RangeCoverage	9.65e-01±8.31e-03	1.00e+00±0.00e+00
long	RangeCoverage	9.91e-01±5.54e-03	1.00e+00±2.45e-04
price	RangeCoverage	5.72e-01±1.02e-01	1.00e+00±1.25e-05
sqft_above	RangeCoverage	7.88e-01±2.98e-02	1.00e+00±1.45e-05
sqft_basement	RangeCoverage	7.47e-01±2.02e-01	1.00e+00±0.00e+00
sqft_living	RangeCoverage	7.03e-01±4.89e-02	1.00e+00±1.33e-05
sqft_living15	RangeCoverage	8.49e-01±5.19e-02	1.00e+00±5.14e-05
sqft_lot	RangeCoverage	5.86e-01±7.02e-03	1.00e+00±3.49e-06
sqft_lot15	RangeCoverage	8.30e-01±2.80e-01	1.00e+00±4.72e-05
view	CategoryCoverage	1.00e+00±0.00e+00	1.00e+00±0.00e+00
waterfront	CategoryCoverage	1.00e+00±0.00e+00	1.00e+00±0.00e+00
yr_built	RangeCoverage	1.00e+00±4.11e-05	1.00e+00±0.00e+00
yr_renovated	RangeCoverage	1.00e+00±9.76e-05	1.00e+00±0.00e+00
zipcode	CategoryCoverage	1.00e+00±0.00e+00	1.00e+00±0.00e+00

4.1.4. Distribución

En términos generales, la distribución en ambos modelos se aproxima a la real, en casi todos los casos superando el 90 %. La única excepción es el modelo Smote en la variable *bathrooms*.

Tabla 4.3: Evaluación de Similitud de Distribución para Modelos Smote y Tddpm, King County

Columna	Metrica	smote-enc	tddpm_mlp
bathrooms	TVComplement	8.84e-01±5.09e-03	9.46e-01±6.18e-03
bedrooms	TVComplement	9.18e-01±7.87e-04	9.50e-01±5.73e-03
condition	TVComplement	9.33e-01±1.23e-03	9.61e-01±5.43e-03
date	TVComplement	9.38e-01±1.73e-03	9.26e-01±2.29e-03
floors	TVComplement	9.66e-01±1.12e-03	9.68e-01±4.38e-03
grade	TVComplement	9.58e-01±6.82e-04	9.64e-01±1.19e-03
id	KSComplement	9.86e-01±6.51e-04	9.75e-01±2.95e-03
lat	KSComplement	9.89e-01±1.69e-03	9.83e-01±8.10e-04
long	KSComplement	9.88e-01±2.22e-03	9.78e-01±1.98e-03
price	KSComplement	9.81e-01±6.63e-04	9.72e-01±7.86e-03
sqft_above	KSComplement	9.72e-01±1.42e-03	9.77e-01±8.75e-03
sqft_basement	KSComplement	9.35e-01±3.60e-03	9.75e-01±3.87e-03
sqft_living	KSComplement	9.81e-01±2.50e-03	9.73e-01±5.59e-03
sqft_living15	KSComplement	9.81e-01±1.63e-03	9.76e-01±4.34e-03
sqft_lot	KSComplement	9.83e-01±4.81e-03	9.58e-01±8.34e-03
sqft_lot15	KSComplement	9.84e-01±3.16e-03	9.62e-01±8.15e-03
view	TVComplement	9.36e-01±9.73e-04	9.52e-01±4.70e-03
waterfront	TVComplement	9.94e-01±1.22e-04	9.95e-01±6.04e-04
yr_built	KSComplement	9.83e-01±4.71e-04	9.76e-01±6.80e-03
yr_renovated	KSComplement	9.92e-01±4.17e-04	9.91e-01±1.00e-03
zipcode	TVComplement	9.74e-01±1.57e-03	9.50e-01±4.11e-04

Al examinar las variables de los conjuntos de datos completos, como se ilustra en la lista Anexa A.2, se observa una similitud entre los tres conjuntos analizados: Real, Smote y Tddpm. Sin embargo, también surgen diferencias significativas. Es relevante mencionar que los conjuntos de datos generados son aproximadamente un 20 % más grandes que el conjunto real. En varias columnas, la distribución de datos en los tres conjuntos es similar, como se evidencia en los casos de bathrooms, sqft_lot, sqft_above, price, sqft_living, sqft_basement, yr_built, sqft_living15 y grade. Este patrón se puede apreciar en la Figura 4.5.

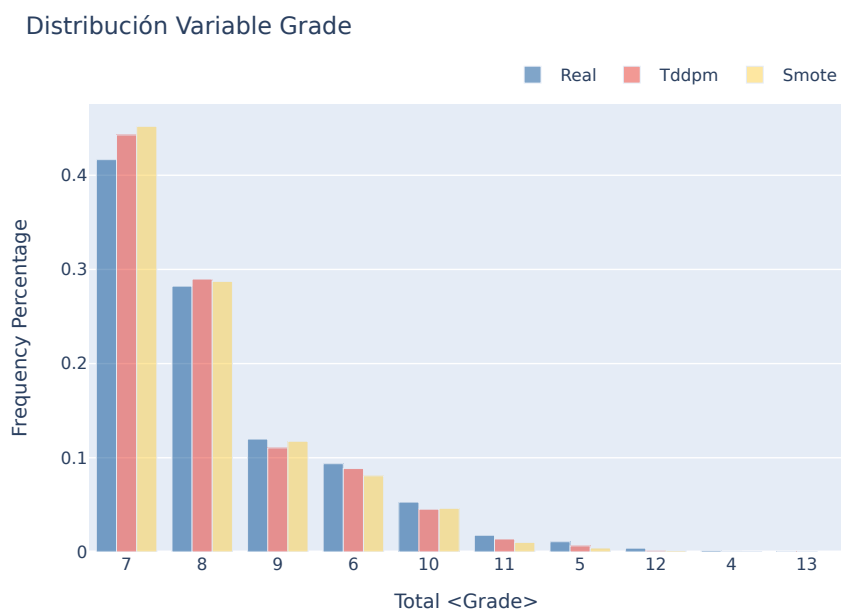


Figura 4.5: Frecuencia del campo Grade en el modelo real y Top 2

Por otra parte, la distribución de los atributos bedrooms, condition, view y floors en el conjunto de datos generado por el modelo Tddpm presenta una particularidad: contiene un mayor número de elementos menos frecuentes comparado con los demás conjuntos. Al considerar la columna *bedrooms* como ejemplo (refiérase a Figura 4.6), la distribución de valores en el conjunto Tddpm se desvía de la del conjunto Smote. En específico, se registra un aumento en la cantidad de registros correspondientes a los valores 6 y 1.

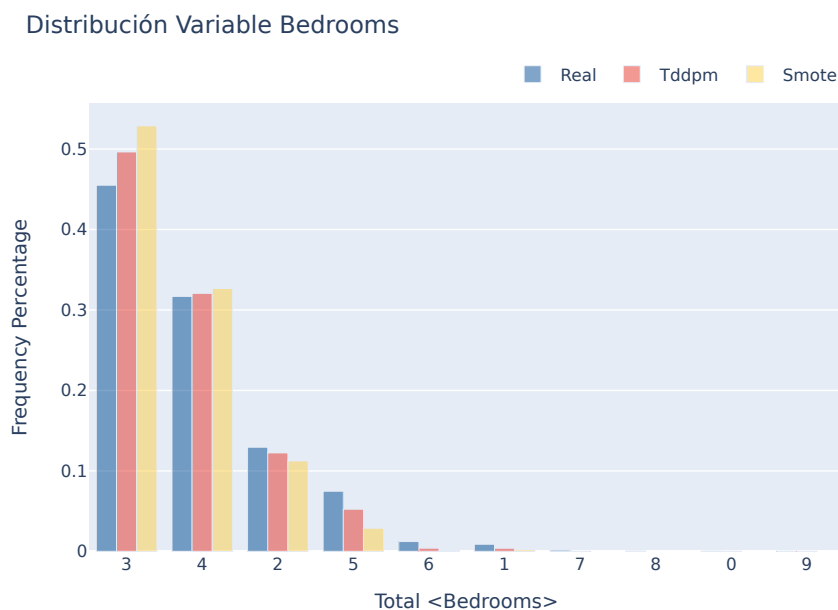


Figura 4.6: Frecuencia del campo Bedrooms en el modelo real y Top 2

En el caso de la variable *sqft_lot15*, la distribución generada por el modelo Smote resulta ser más similar a la del conjunto de datos real, como se puede apreciar en la figura 4.7.

Distribución Variable Sqft Lot15

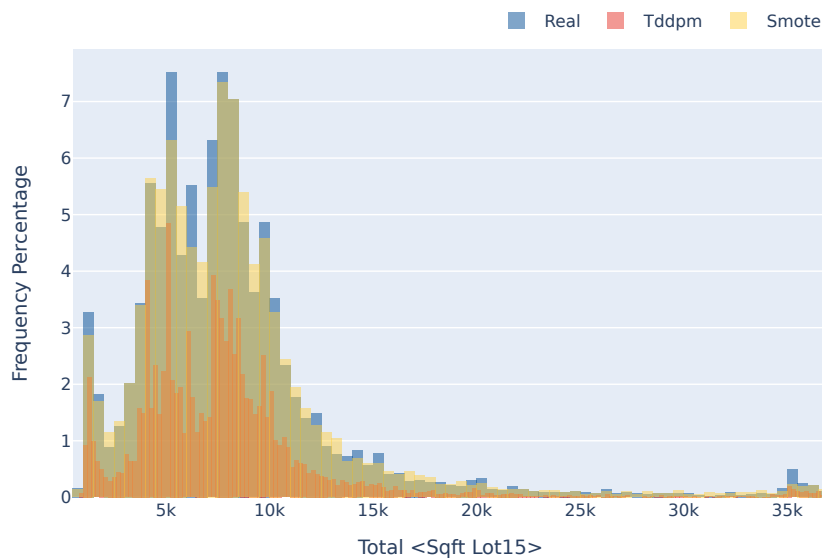


Figura 4.7: Frecuencia del campo Sqft lot15 en el modelo real y Top 2

4.1.5. Privacidad

Al analizar los registros más cercanos entre los conjuntos de datos reales utilizados para el entrenamiento, los generados por los modelos, y el conjunto de datos reales almacenados, encontramos que las distancias entre ellos se presentan en las siguientes tablas. Es importante destacar que la distancia mínima para el modelo Tddpm es de 0.0134, indicando que cada registro tiene al menos esa distancia respecto al conjunto real. La determinación del epsilon requerido para asegurar la privacidad de los datos depende del análisis específico de los datos a proteger y sus probabilidades asociadas. Sin embargo, si el objetivo es proteger el 95 % de los datos, el modelo Tddpm alcanza una distancia de 0.0579, mientras que el modelo Smote tiene una distancia de 0.00704.

Tabla 4.4: Proporción entre el más cercano y el segundo más cercano, percentil 5, datos king county

Modelo	NNDR ST	NNDR SH	NNDR TH	Score
tddpm_mlp	6.12e-01±2.28e-03	6.03e-01±4.17e-03	3.76e-01±0.00e+00	9.52e-01±2.36e-03
smote-enc	1.98e-01±3.57e-03	4.09e-01±6.38e-03	3.76e-01±0.00e+00	9.53e-01±2.45e-04
ctgan	8.09e-01±8.59e-03	8.15e-01±4.89e-03	3.76e-01±0.00e+00	8.24e-01±2.02e-02
tablepreset	8.25e-01±0.00e+00	8.18e-01±7.85e-17	3.76e-01±0.00e+00	8.37e-01±7.85e-17
copulagan	8.30e-01±5.92e-03	8.24e-01±3.49e-03	3.76e-01±0.00e+00	7.89e-01±2.92e-03
gaussiancopula	7.53e-01±0.00e+00	7.52e-01±1.36e-16	3.76e-01±0.00e+00	7.88e-01±0.00e+00
tvae	7.32e-01±5.74e-03	7.04e-01±4.22e-03	3.76e-01±0.00e+00	7.38e-01±1.18e-02

Tabla 4.5: Proporción entre el más cercano y el segundo más cercano, percentil 1, datos king county

Modelo	NNDR ST	NNDR SH	NNDR TH	Score
tddpm_mlp	4.49e-01±7.33e-03	4.49e-01±7.70e-03	7.98e-02±0.00e+00	9.52e-01±2.36e-03
smote-enc	6.74e-02±4.35e-03	1.88e-01±1.05e-02	7.98e-02±0.00e+00	9.53e-01±2.45e-04
ctgan	7.14e-01±7.52e-03	7.19e-01±7.48e-03	7.98e-02±0.00e+00	8.24e-01±2.02e-02
tablepreset	7.20e-01±1.11e-16	7.19e-01±1.11e-16	7.98e-02±0.00e+00	8.37e-01±7.85e-17
copulagan	7.56e-01±7.65e-03	7.47e-01±2.64e-03	7.98e-02±0.00e+00	7.89e-01±2.92e-03
gaussiancopula	6.48e-01±7.85e-17	6.43e-01±7.85e-17	7.98e-02±0.00e+00	7.88e-01±0.00e+00
tvae	6.06e-01±1.27e-02	5.87e-01±4.93e-03	7.98e-02±0.00e+00	7.38e-01±1.18e-02

Tabla 4.6: Proporción entre el más cercano y el segundo más cercano, minimo, datos king county

Modelo	NNDR ST	NNDR SH	NNDR TH	Score
tddpm_mlp	1.23e-01±1.54e-02	1.57e-01±3.36e-02	0.00e+00±0.00e+00	9.52e-01±2.36e-03
smote-enc	0.00e+00±0.00e+00	1.10e-02±5.41e-03	0.00e+00±0.00e+00	9.53e-01±2.45e-04
ctgan	4.25e-01±3.23e-02	3.91e-01±4.06e-02	0.00e+00±0.00e+00	8.24e-01±2.02e-02
tablepreset	4.51e-01±6.80e-17	3.58e-01±0.00e+00	0.00e+00±0.00e+00	8.37e-01±7.85e-17
copulagan	5.48e-01±1.58e-02	5.32e-01±3.85e-02	0.00e+00±0.00e+00	7.89e-01±2.92e-03
gaussiancopula	3.90e-01±5.55e-17	4.08e-01±0.00e+00	0.00e+00±0.00e+00	7.88e-01±0.00e+00
tvae	3.44e-01±1.91e-02	3.43e-01±1.63e-02	0.00e+00±0.00e+00	7.38e-01±1.18e-02

Al analizar los ratios entre la distancia al primer vecino más cercano y la distancia al segundo para el modelo Tddpm, se evidencia que para el percentil 5, la distancia al vecino más cercano es solo 2/3 de la distancia al segundo más cercano. Sin embargo, para el percentil 1, esta distancia se reduce a la mitad. En contraposición, para el modelo Smote, en el percentil 5, la distancia al vecino más cercano es solo un 20 % de la distancia al segundo más cercano, y disminuye rápidamente a un 6 % para el percentil 1.

Tabla 4.7: Distancia de registros más cercanos, percentil 5, datos king county

Modelo	DCR ST	DCR SH	DCR TH	Score
tddpm_mlp	5.79e-02±6.08e-04	7.65e-02±1.23e-03	3.57e-02±0.00e+00	9.52e-01±2.36e-03
smote-enc	7.04e-03±2.77e-04	3.69e-02±6.21e-04	3.57e-02±0.00e+00	9.53e-01±2.45e-04
ctgan	2.15e-01±1.32e-02	2.38e-01±1.32e-02	3.57e-02±0.00e+00	8.24e-01±2.02e-02
tablepreset	1.80e-01±0.00e+00	2.00e-01±0.00e+00	3.57e-02±0.00e+00	8.37e-01±7.85e-17
copulagan	3.75e-01±9.42e-03	4.12e-01±7.08e-03	3.57e-02±0.00e+00	7.89e-01±2.92e-03
gaussiancopula	2.63e-01±3.93e-17	3.06e-01±0.00e+00	3.57e-02±0.00e+00	7.88e-01±0.00e+00
tvae	8.09e-02±3.59e-04	9.86e-02±5.62e-04	3.57e-02±0.00e+00	7.38e-01±1.18e-02

Tabla 4.8: Distancia de registros más cercanos, percentil 1, datos king county

Modelo	DCR ST	DCR SH	DCR TH	Score
tddpm_mlp	3.98e-02±8.56e-04	5.22e-02±7.40e-04	7.94e-03±0.00e+00	9.52e-01±2.36e-03
smote-enc	1.93e-03±1.36e-04	1.46e-02±8.53e-04	7.94e-03±0.00e+00	9.53e-01±2.45e-04
ctgan	1.70e-01±1.13e-02	1.89e-01±1.11e-02	7.94e-03±0.00e+00	8.24e-01±2.02e-02
tablepreset	1.44e-01±0.00e+00	1.61e-01±1.96e-17	7.94e-03±0.00e+00	8.37e-01±7.85e-17
copulagan	3.30e-01±9.97e-03	3.59e-01±9.29e-03	7.94e-03±0.00e+00	7.89e-01±2.92e-03
gaussiancopula	2.05e-01±2.78e-17	2.40e-01±0.00e+00	7.94e-03±0.00e+00	7.88e-01±0.00e+00
tvae	6.40e-02±2.03e-04	7.80e-02±2.67e-04	7.94e-03±0.00e+00	7.38e-01±1.18e-02

Tabla 4.9: Distancia de registros más cercanos, minimo, datos king county

Modelo	DCR ST	DCR SH	DCR TH	Score
tddpm_mlp	1.34e-02±3.06e-03	1.99e-02±1.70e-03	0.00e+00±0.00e+00	9.52e-01±2.36e-03
smote-enc	0.00e+00±0.00e+00	1.24e-03±1.14e-04	0.00e+00±0.00e+00	9.53e-01±2.45e-04
ctgan	8.76e-02±2.88e-03	1.06e-01±1.12e-02	0.00e+00±0.00e+00	8.24e-01±2.02e-02
tablepreset	7.90e-02±9.81e-18	8.53e-02±1.39e-17	0.00e+00±0.00e+00	8.37e-01±7.85e-17
copulagan	2.09e-01±2.71e-02	2.34e-01±3.43e-02	0.00e+00±0.00e+00	7.89e-01±2.92e-03
gaussiancopula	7.88e-02±9.81e-18	1.27e-01±0.00e+00	0.00e+00±0.00e+00	7.88e-01±0.00e+00
tvae	3.26e-02±2.42e-03	3.48e-02±7.64e-03	0.00e+00±0.00e+00	7.38e-01±1.18e-02

En la Figura 4.8 solo se consideran los modelos Tddpm y Smote para su comparación. En ambos casos, existe una distancia mayor a cero. Sin embargo, esta distancia es mayor en el caso de Tddpm, lo que sugiere que este conjunto puede ser considerado superior en términos de privacidad.

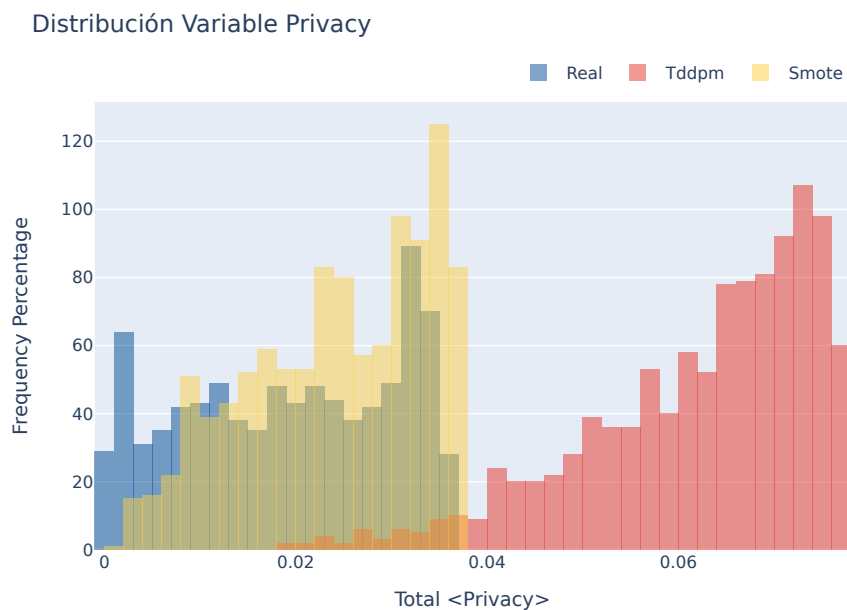


Figura 4.8: Frecuencia del campo Privacy en el modelo real y Top 2

4.1.6. Ejemplo de registros

Las Tablas 4.10 y 4.11 presentan un ejemplo de la mínima distancia en los modelos Smote y Tddpm, respectivamente. Los nombres de las columnas representan la distancia de Minkowski al registro Sintético, indicado de esta manera en la columna correspondiente. Las celdas coloreadas en rojo señalan que el valor de la característica para una propiedad específica es idéntico al valor correspondiente de la propiedad de referencia. Así, la tabla proporciona una comparación detallada de las propiedades que son similares en términos de las características seleccionadas.

En la Tabla 4.10, se puede observar claramente que, excepto por las variables de precio y fecha en el segundo registro más cercano, son idénticas a las del original. Esto significa que ese registro fue transferido en su totalidad al conjunto sintético.

Tabla 4.10: Ejemplos para el modelo smote-enc, minimo

Variable/Distancia	Sintético	DCR1 d(0.00e+00)	DCR2 d(1.01e-02)
sqft_living	1790.000000	1790.000000	1790.000000
sqft_basement	0.000000	0.000000	0.000000
id	1721801010.000000	1721801010.000000	1721801010.000000
sqft_above	1790.000000	1790.000000	1790.000000
price	225000.000000	225000.000000	302100.000000
view	0	0	0
waterfront	0	0	0
sqft_lot	6120.000000	6120.000000	6120.000000
sqft_living15	830.000000	830.000000	830.000000
grade	6	6	6
bathrooms	1.000000	1.000000	1.000000
long	-122.337000	-122.337000	-122.337000
yr_renovated	1964.000000	1964.000000	1964.000000
zipcode	98146	98146	98146
condition	3	3	3
bedrooms	3	3	3
date	20140903T000000	20140903T000000	20150424T000000
sqft_lot15	6120.000000	6120.000000	6120.000000
lat	47.508000	47.508000	47.508000
yr_built	1937.000000	1937.000000	1937.000000
floors	1.000000	1.000000	1.000000

La Tabla 4.11 presenta valores de distancia mayores que los obtenidos en la tabla correspondiente a Smote (4.10). Se pueden observar diferencias en las variables *sqft_living*, *sqft_lot*, *sqft_above*, *yr_built* y *lat*, entre otras. Esta es la mínima distancia encontrada por la métrica.

Tabla 4.11: Ejemplos para el modelo *tddpm_mlp*, minimo

Variable/Distancia	Sintético	DCR1 d(9.90e-03)	DCR2 d(1.03e-02)
id	7202202136.850033	7202330530.000000	7202330030.000000
sqft_living	1700.000000	1690.000000	1650.000000
sqft_lot	3524.360953	3322.000000	5683.000000
sqft_above	1654.928371	1690.000000	1650.000000
sqft_basement	0.000000	0.000000	0.000000
yr_built	2004.000000	2003.000000	2003.000000
yr_renovated	0.000000	0.000000	0.000000
lat	47.683989	47.682400	47.683000
long	-122.036195	-122.036000	-122.035000
sqft_living15	1650.000000	1650.000000	1650.000000
sqft_lot15	3796.678538	3446.000000	4193.000000
price	475000.000000	479000.000000	500000.000000
date	20140908T000000	20150116T000000	20140822T000000
bedrooms	3	3	3
bathrooms	2.500000	2.500000	2.500000
floors	2.000000	2.000000	2.000000
waterfront	0	0	0
view	0	0	0
condition	3	3	3
grade	7	7	7
zipcode	98053	98053	98053

En la Tabla 4.12, se puede observar una notable mejoría en el modelo Smote. Esta tabla presenta un registro cercano con múltiples diferencias, entre las cuales se pueden destacar *sqft_lot* y *price*.

Tabla 4.12: Ejemplos para el modelo smote-enc, percentil 1

Variable/Distancia	Sintético	DCR1 d(2.07e-03)	DCR2 d(5.44e-03)
sqft_living	1798.822360	1800.000000	1800.000000
sqft_basement	0.000000	0.000000	0.000000
id	3862710182.355279	3862710050.000000	3862710210.000000
sqft_above	1798.822360	1800.000000	1800.000000
price	450000.000000	437718.000000	409316.000000
view	0	0	0
waterfront	0	0	0
sqft_lot	2851.472564	3265.000000	3168.000000
sqft_living15	1800.000000	1800.000000	1800.000000
grade	8	8	8
bathrooms	2.500000	2.500000	2.500000
long	-121.841000	-121.841000	-121.841000
yr_renovated	0.000000	0.000000	0.000000
zipcode	98065	98065	98065
condition	3	3	3
bedrooms	3	3	3
date	20140520T000000	20141113T000000	20140520T000000
sqft_lot15	3280.073791	3663.000000	3393.000000
lat	47.534176	47.533800	47.534200
yr_built	2013.882236	2014.000000	2014.000000
floors	2.000000	2.000000	2.000000

4.1.7. Propiedades estadísticas

El listado completo de las propiedades estadísticas se encuentra en el Anexo A.9. A continuación, se procede a mostrar las propiedades estadísticas que entre el modelo Tddpm y Smote consigan una diferencia mayor al 5 % con respecto al conjunto original de entrenamiento. Se agrega el modelo Ctgan como referencia. las variables fueron seleccionadas por se 1) El peor resultado en la cobertura y 2) El peor resultado en la distribución respectivamente.

Como se puede apreciar en la Tabla 4.13, en general, el modelo Tddpm muestra propiedades estadísticas más cercanas al conjunto original, con excepciones notables en las métricas de máximo, kurtosis y Jarque-Bera. La diferencia en la métrica de *máximo* podría contribuir a la baja puntuación en la métrica de cobertura mostrada en la Tabla 4.2. Por otro lado, las diferencias en las métricas de kurtosis, skew y Jarque-Bera podrían explicar las desviaciones observadas en la métrica de distribución de la Tabla 4.3.

Tabla 4.13: Propiedades estadísticas de variable bedrooms con cambio >5 %, King county (A-1)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
nobs	17290	21613	21614	21613
mean	3.368	3.337	3.279	4.075
std_err	0.007	0.005	0.005	0.025
upper_ci	3.382	3.348	3.289	4.124
lower_ci	3.354	3.327	3.270	4.026
std	0.931	0.777	0.710	3.682
mad	0.734	0.643	0.587	1.323
mad_normal	0.920	0.806	0.735	1.658
coef_var	0.277	0.233	0.217	0.904
range	33.000	8.000	8.000	33.000
max	33.000	8.000	9.000	33.000
min	0.000	0.000	1.000	0.000
skew	2.304	0.273	0.140	6.792
kurtosis	63.268	3.444	3.196	53.464
jarque_bera	2631992	446	105	2459502
jarque_bera_pval	0.000	0.000	0.000	0.000
mode_freq	0.455	0.504	0.528	0.420
95.0 %	5.000	5.000	4.000	6.000
99.0 %	6.000	5.000	5.000	33.000
99.9 %	7.000	6.000	5.000	33.000

Es evidente que Smote presenta varias métricas inferiores a las de Tddpm. Entre estas destacan el mínimo, el máximo, la asimetría (skew) y los percentiles 0.1, 95, 99 y 99.9.

Tabla 4.14: Propiedades estadísticas de variable bathrooms con cambio >5 %, King county (A-1)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
nobs	17290	21613	21614	21613
mean	2.114	2.071	2.015	2.255
std_err	0.006	0.005	0.005	0.006
upper_ci	2.125	2.080	2.025	2.267
lower_ci	2.102	2.061	2.006	2.243
std	0.767	0.715	0.709	0.932
mad	0.615	0.584	0.595	0.710
mad_normal	0.771	0.732	0.746	0.890
coef_var	0.363	0.345	0.352	0.413
range	8.000	7.750	5.250	8.000
max	8.000	7.750	6.000	8.000
min	0.000	0.000	0.750	0.000
skew	0.464	0.264	0.123	0.567
kurtosis	3.989	3.412	2.751	4.057
jarque_bera	1326	404	110	2165
jarque_bera_pval	0.000	0.000	0.000	0.000
mode_freq	0.251	0.284	0.313	0.300
0.1 %	0.750	0.750	1.000	0.000
95.0 %	3.500	3.250	3.250	4.000
99.0 %	4.250	4.000	3.500	4.750
99.9 %	5.428	4.750	4.500	6.250

4.2. Economicos

4.2.1. Tratamiento de nulos en conjunto A y B

El conjunto de Económicos, a diferencia del conjunto de datos de King County que fue filtrado y preprocesado para evitar valores nulos, contiene elementos nulos. A continuación se describen dos tratamientos de estos elementos nulos. El primer enfoque simplemente elimina todos los registros que contienen un registro vacío utilizando el método ‘dropna’, como se muestra en el Código 5; este será considerado como el Conjunto A. En el segundo enfoque, los valores nulos son reemplazados por algún valor predeterminado o calculado, como se muestra en el Código 6; este será considerado como el Conjunto B.

```
1 df_converted = df.dropna().astype({k: 'str' for k in ("description", "price",
    ↪ "title", "address", "owner",)})
2 basedate = pd.Timestamp('2017-12-01')
3 dtype = df_converted.pop("publication_date")
4 df_converted["publication_date"] = dtype.apply(lambda x: (x - basedate).days)
```

Código 5: Eliminación de valores nulos en el conjunto de datos de Económicos

```
1 df_converted = df.fillna(dict(
2     property_type = "None",
3     transaction_type = "None",
4     state = "None",
5     county = "None",
6     rooms = -1,
7     bathrooms = -1,
8     m_built = -1,
9     m_size = -1,
10    source = "None"
11)).fillna(-1).astype({k: 'str' for k in ("description", "price", "title",
    ↪ "address", "owner",)})
12 basedate = pd.Timestamp('2017-12-01')
13 dtype = df_converted.pop("publication_date")
14 df_converted["publication_date"] = dtype.apply(lambda x: (x - basedate).days)
```

Código 6: Reemplazo de valores nulos en el conjunto de datos de Económicos

4.2.2. SDMetrics Score - Conjunto A

Para el conjunto A, como se muestra en la Tabla 4.15, Tddpm es un punto superior a Smote y ambos superan en más de 10 puntos al siguiente modelo. Sin embargo, un punto crucial es que Smote tiene una cobertura (*Coverage*) que es 12 puntos inferior a Tddpm.

Tabla 4.15: Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, Economicos

Model Name	Column Pair Trends	Column Shapes	Coverage	Boundaries	Score
tddpm_mlp	9.73e-01±2.21e-03	9.84e-01±3.63e-04	7.91e-01±5.31e-02	1.00e+00±0.00e+00	9.79e-01±1.27e-03
smote-enc	9.62e-01±1.52e-03	9.76e-01±4.01e-04	6.67e-01±2.79e-02	1.00e+00±0.00e+00	9.69e-01±6.71e-04
copulagan	7.46e-01±3.30e-02	7.90e-01±2.63e-02	6.80e-01±2.57e-03	1.00e+00±0.00e+00	7.68e-01±2.96e-02
ctgan	7.44e-01±1.96e-02	6.53e-01±4.72e-02	6.75e-01±1.75e-03	1.00e+00±0.00e+00	6.98e-01±2.63e-02
gaussiancopula	6.96e-01±0.00e+00	6.88e-01±0.00e+00	5.65e-01±0.00e+00	1.00e+00±0.00e+00	6.92e-01±0.00e+00
tvae	5.83e-01±1.02e-02	6.41e-01±4.66e-02	8.59e-02±1.28e-02	1.00e+00±0.00e+00	6.12e-01±2.50e-02

4.2.3. Correlación - Conjunto A

Aunque la diferencia es pequeña, se puede apreciar al comparar visualmente las Figuras 4.9 y 4.10 que el segundo modelo, Tddpm, presenta una mayor similitud en las variables *rooms* y *bathrooms*.

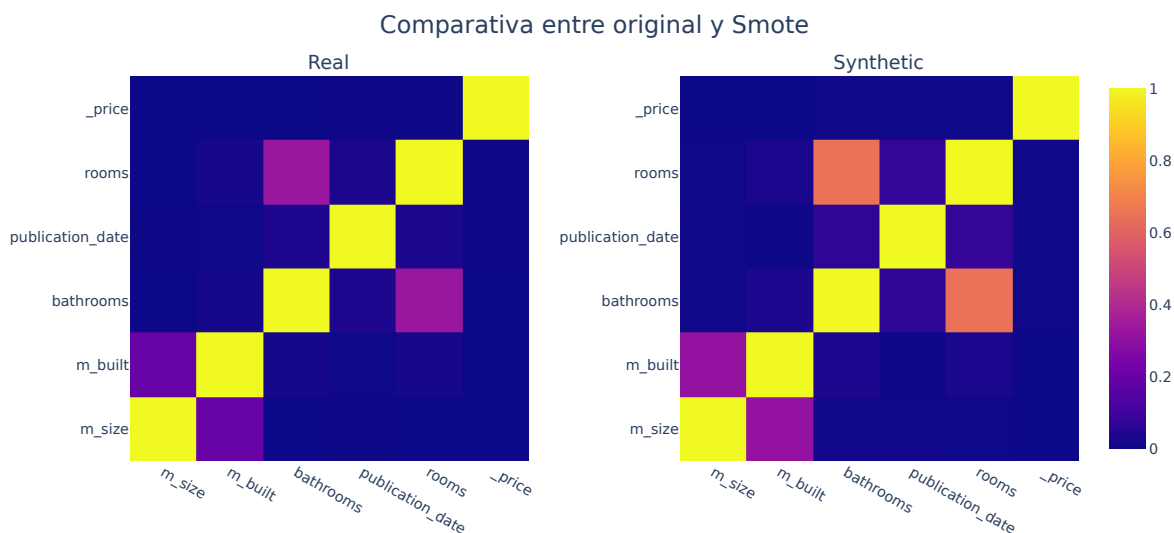


Figura 4.9: Correlación de conjunto original de entrenamiento y Smote

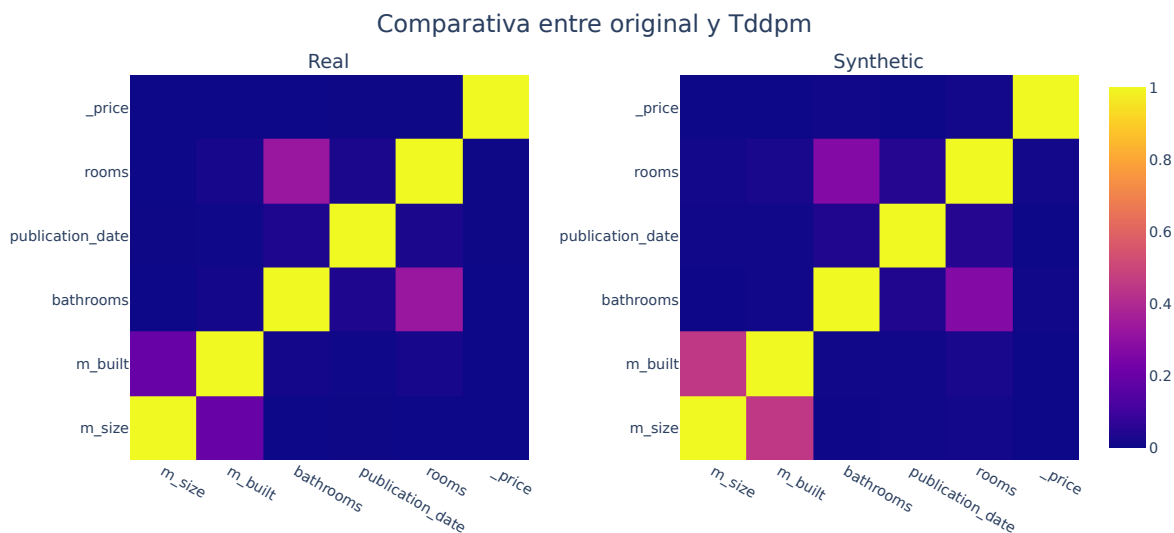


Figura 4.10: Correlación de conjunto original de entrenamiento y Tddpm

4.2.4. Cobertura - Conjunto A

En las tablas detalles de cobertura 4.16 se puede ver el porqué ambos tenían una puntuación tan baja. Existen elementos con una cobertura menor al 40 %, por ejemplo, la variable m_size. Aun así, se puede ver que Tddpm es ligeramente mejor en la mayoría de las columnas.

Tabla 4.16: Evaluación de Cobertura Categoría-Rango para Modelos SMOTE-ENC y TDDPM_MLP, Economicos

Columna	Metrica	smote-enc	tddpm_mlp
_price	RangeCoverage	9.68e-01±5.48e-02	9.66e-01±3.30e-02
bathrooms	CategoryCoverage	8.63e-01±3.40e-02	6.76e-01±2.94e-02
county	CategoryCoverage	5.97e-01±3.73e-03	7.87e-01±2.27e-02
m_built	RangeCoverage	5.52e-01±3.16e-01	7.71e-01±3.97e-01
m_size	RangeCoverage	1.79e-02±8.52e-03	3.36e-01±4.53e-02
property_type	CategoryCoverage	6.67e-01±5.56e-02	9.07e-01±3.21e-02
publication_date	RangeCoverage	9.70e-01±5.80e-03	9.81e-01±2.86e-03
rooms	CategoryCoverage	7.40e-01±1.41e-02	7.80e-01±6.45e-02
state	CategoryCoverage	7.92e-01±3.61e-02	9.58e-01±3.61e-02
transaction_type	CategoryCoverage	5.00e-01±0.00e+00	7.50e-01±2.50e-01

La escasa cobertura en m_size podría atribuirse a su distribución. Como se ilustra en la figura 4.11, esta presenta una larga cola, caracterizada por valores altos pero infrecuentes.

Distribución Variable M Size

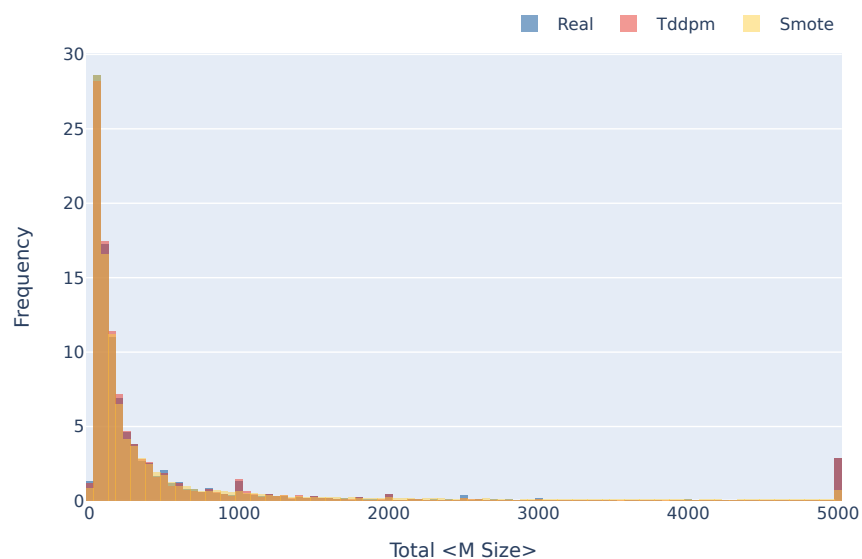


Figura 4.11: Frecuencia del campo M size en el modelo real y Top 2

4.2.5. Distribución - Conjunto A

Ambos modelos muestran un buen rendimiento en cuanto a la forma y la distribución de los datos, como se evidencia en la Tabla 4.17. Como se vio en la Figura 4.11 una buena distribución no asegura una cobertura completa.

Tabla 4.17: Evaluación de Similitud de Distribución para Modelos SMOTE-ENC y TDDPM_MLP, Economicos

Columna	Metrica	smote-enc	tddpm_mlp
_price	KSComplement	9.90e-01±1.16e-03	9.88e-01±3.17e-03
bathrooms	TVComplement	9.96e-01±5.34e-04	9.86e-01±4.99e-04
county	TVComplement	9.20e-01±1.01e-03	9.65e-01±2.54e-03
m_built	KSComplement	9.87e-01±7.24e-04	9.87e-01±1.32e-03
m_size	KSComplement	9.74e-01±1.11e-03	9.85e-01±8.91e-04
property_type	TVComplement	9.68e-01±1.75e-03	9.81e-01±2.30e-03
publication_date	KSComplement	9.79e-01±2.43e-03	9.86e-01±3.20e-03
rooms	TVComplement	9.78e-01±1.42e-03	9.82e-01±3.04e-03
state	TVComplement	9.67e-01±3.85e-03	9.84e-01±1.79e-04
transaction_type	TVComplement	9.99e-01±8.21e-04	9.96e-01±2.73e-03

4.2.6. Privacidad - Conjunto A

Resulta interesante notar que, para el percentil 1 y el 5, en las Tablas 4.19 y 4.18 respectivamente, el modelo Tddpm demuestra que la cercanía de los registros más próximos es predominante al comparar el conjunto sintético con el conjunto de retención (*Hold*). Este fenómeno no se evidencia en ninguna otra comparación. Asimismo, se destaca que las diferencias mínimas llegan a cero en los dos modelos más efectivos (Tddpm y Smote), y que los valores de distancia son extremadamente reducidos. Para el percentil 5, Tddpm registra una distancia de $4,48 \times 10^{-9}$.

Tabla 4.18: Distancia de registros más cercanos, percentil 5, datos economicos

Modelo	DCR ST	DCR SH	DCR TH	Score
tddpm_mlp	4.48e-09±2.32e-10	3.59e-08±2.38e-09	1.28e-08±0.00e+00	9.79e-01±1.27e-03
smote-enc	3.15e-11±3.01e-12	4.22e-08±2.49e-09	1.28e-08±0.00e+00	9.69e-01±6.71e-04
copulagan	1.37e-06±1.76e-07	2.86e-06±3.82e-07	1.28e-08±0.00e+00	7.68e-01±2.96e-02
ctgan	1.49e-05±5.01e-06	2.42e-05±9.67e-06	1.28e-08±0.00e+00	6.98e-01±2.63e-02
gaussiancopula	5.28e-06±0.00e+00	8.21e-06±0.00e+00	1.28e-08±0.00e+00	6.92e-01±0.00e+00
tvae	3.90e-07±1.08e-07	7.80e-07±2.49e-07	1.28e-08±0.00e+00	6.12e-01±2.50e-02

Tabla 4.19: Distancia de registros más cercanos, percentil 1, datos economicos

Modelo	DCR ST	DCR SH	DCR TH	Score
tddpm_mlp	1.46e-10±3.86e-12	1.44e-09±1.01e-10	0.00e+00±0.00e+00	9.79e-01±1.27e-03
smote-enc	0.00e+00±0.00e+00	1.54e-09±5.32e-13	0.00e+00±0.00e+00	9.69e-01±6.71e-04
copulagan	1.97e-07±4.64e-08	4.53e-07±9.95e-08	0.00e+00±0.00e+00	7.68e-01±2.96e-02
ctgan	3.18e-06±4.34e-07	5.23e-06±1.44e-06	0.00e+00±0.00e+00	6.98e-01±2.63e-02
gaussiancopula	7.84e-07±7.49e-23	1.75e-06±0.00e+00	0.00e+00±0.00e+00	6.92e-01±0.00e+00
tvae	1.48e-07±9.24e-08	2.35e-07±1.18e-07	0.00e+00±0.00e+00	6.12e-01±2.50e-02

Tabla 4.20: Distancia de registros más cercanos, minimo, datos economicos

Modelo	DCR ST	DCR SH	DCR TH	Score
tddpm_mlp	0.00e+00±0.00e+00	0.00e+00±0.00e+00	0.00e+00±0.00e+00	9.79e-01±1.27e-03
smote-enc	0.00e+00±0.00e+00	0.00e+00±0.00e+00	0.00e+00±0.00e+00	9.69e-01±6.71e-04
copulagan	5.88e-09±2.05e-09	1.21e-08±3.19e-09	0.00e+00±0.00e+00	7.68e-01±2.96e-02
ctgan	2.83e-08±3.88e-08	6.05e-08±2.56e-08	0.00e+00±0.00e+00	6.98e-01±2.63e-02
gaussiancopula	1.13e-08±0.00e+00	1.75e-08±0.00e+00	0.00e+00±0.00e+00	6.92e-01±0.00e+00
tvae	5.65e-09±3.07e-09	2.56e-08±3.04e-08	0.00e+00±0.00e+00	6.12e-01±2.50e-02

También se puede observar una disminución en la relación entre el registro más cercano y el segundo más cercano en comparación con el conjunto de datos de King County. En el percentil 5, el segundo registro más cercano está a 15 veces la distancia del primero. Esta relación se reduce a 10 veces cuando se compara con el conjunto *Hold*.

Tabla 4.21: Proporción entre el más cercano y el segundo más cercano, percentil 5, datos económicos

Modelo	NNDR ST	NNDR SH	NNDR TH	Score
tddpm_mlp	6.88e-02±1.16e-03	9.85e-02±2.09e-03	1.31e-02±0.00e+00	9.79e-01±1.27e-03
smote-enc	7.17e-04±1.27e-05	1.12e-01±3.32e-03	1.31e-02±0.00e+00	9.69e-01±6.71e-04
copulagan	2.74e-01±3.32e-02	3.03e-01±4.80e-02	1.31e-02±0.00e+00	7.68e-01±2.96e-02
ctgan	2.65e-01±1.35e-02	2.71e-01±5.93e-02	1.31e-02±0.00e+00	6.98e-01±2.63e-02
gaussiancopula	2.93e-01±0.00e+00	2.76e-01±0.00e+00	1.31e-02±0.00e+00	6.92e-01±0.00e+00
tvae	3.67e-01±6.93e-02	4.31e-01±1.08e-01	1.31e-02±0.00e+00	6.12e-01±2.50e-02

Tabla 4.22: Proporción entre el más cercano y el segundo más cercano, percentil 1, datos económicos

Modelo	NNDR ST	NNDR SH	NNDR TH	Score
tddpm_mlp	3.00e-03±9.91e-05	1.04e-02±2.95e-04	0.00e+00±0.00e+00	9.79e-01±1.27e-03
smote-enc	0.00e+00±0.00e+00	2.47e-03±2.53e-04	0.00e+00±0.00e+00	9.69e-01±6.71e-04
copulagan	1.37e-02±3.97e-03	1.31e-02±1.78e-03	0.00e+00±0.00e+00	7.68e-01±2.96e-02
ctgan	4.84e-02±1.75e-02	3.67e-02±7.08e-03	0.00e+00±0.00e+00	6.98e-01±2.63e-02
gaussiancopula	2.67e-02±2.45e-18	2.95e-02±0.00e+00	0.00e+00±0.00e+00	6.92e-01±0.00e+00
tvae	5.44e-02±4.44e-02	1.95e-01±6.70e-02	0.00e+00±0.00e+00	6.12e-01±2.50e-02

Tabla 4.23: Proporción entre el más cercano y el segundo más cercano, mínimo, datos económicos

Modelo	NNDR ST	NNDR SH	NNDR TH	Score
tddpm_mlp	0.00e+00±0.00e+00	0.00e+00±0.00e+00	0.00e+00±0.00e+00	9.79e-01±1.27e-03
smote-enc	0.00e+00±0.00e+00	0.00e+00±0.00e+00	0.00e+00±0.00e+00	9.69e-01±6.71e-04
copulagan	1.22e-04±7.03e-05	1.84e-04±1.12e-04	0.00e+00±0.00e+00	7.68e-01±2.96e-02
ctgan	4.21e-04±2.19e-04	1.32e-03±1.54e-03	0.00e+00±0.00e+00	6.98e-01±2.63e-02
gaussiancopula	4.99e-05±0.00e+00	7.59e-06±1.04e-21	0.00e+00±0.00e+00	6.92e-01±0.00e+00
tvae	8.11e-04±1.77e-04	7.24e-03±3.14e-03	0.00e+00±0.00e+00	6.12e-01±2.50e-02

4.2.7. Ejemplos de registros - Conjunto A

Es fácil entender que la implicancia de un *DCR* igual a 0 es un registro copiado desde el conjunto real, esto se puede apreciar en la Tabla 4.24.

Tabla 4.24: Ejemplos para el modelo *tddpm_mlp*, mínimo

Variable/Distancia	Sintético	DCR1 d(0.00e+00)	DCR2 d(2.26e-07)
_price	2859.676636	2859.676636	2800.000000
bathrooms	2.000000	2.000000	2.000000
county	Los Vilos	Los Vilos	Ovalle
m_built	120.000000	120.000000	122.000000
m_size	300.000000	300.000000	197.000000
property_type	Casa	Casa	Casa
publication_date	1545.000000	1545.000000	1545.000000
rooms	4.000000	4.000000	4.000000
state	Coquimbo	Coquimbo	Coquimbo
transaction_type	Venta	Venta	Venta

Ya cuando se observa el percentil 1, se puede apreciar que la diferencia se puede considerar significativa. En el caso mostrado por la Tabla 4.25, los metros cuadrados (*m_size*) y *_price* cambian y luego la variable *county* también cambian en el segundo registro más cercano.

Tabla 4.25: Ejemplos para el modelo *tddpm_mlp*, percentil 1

Variable/Distancia	Sintético	DCR1 d(1.42e-10)	DCR2 d(3.93e-09)
_price	10.747760	10.803223	9.214514
bathrooms	1.000000	1.000000	1.000000
county	San Miguel	San Miguel	Estación Central
m_built	34.000000	34.000000	34.000000
m_size	35.685290	36.000000	36.000000
property_type	Departamento	Departamento	Departamento
publication_date	1545.000000	1545.000000	1545.000000
rooms	1.000000	1.000000	1.000000
state	Metropolitana de Santiago	Metropolitana de Santiago	Metropolitana de Santiago
transaction_type	Arriendo	Arriendo	Arriendo

En las Tablas 4.26 y 4.27 se puede observar un registro con coherencia simulada. Por ejemplo, cuando decide generar un número de teléfono, este parece coherente. También menciona que está cerca de un metro, detalle que podría estar presente en una publicación real, a pesar de que el metro indicado no exista.

Tabla 4.26: Ejemplos para el modelo tddpm_mlp, percentil 4

Variable/Distancia	Sintético	DCR1 d(2.84e-09)	DCR2 d(6.74e-09)
_price	11.478674	10.485481	8.896772
bathrooms	1.000000	1.000000	1.000000
county	Santiago	Santiago	Santiago
m_built	38.984909	39.000000	39.000000
m_size	45.000000	43.000000	39.000000
property_type	Departamento	Departamento	Departamento
publication_date	1545.000000	1545.000000	1545.000000
rooms	1.000000	1.000000	1.000000
state	Metropolitana de Santiago	Metropolitana de Santiago	Metropolitana de Santiago
transaction_type	Arriendo	Arriendo	Arriendo

Tabla 4.27: Ejemplos de texto modelo tddpm_mlp, percentil 4

Distancia	description
Sintético	Departamento de 2 pisos, 1 baño con vista panorámica al centro comercial en Santiago del sector central (centros), cercano a metro Metro San Luis). El edificio cuenta con un gran living comedor que comparte una terraza completamente equipada por lavandería para dos vehículos; el primer piso: Living comedor amplio como sala de estar/cocina amoblada cerrada cubierta de granito cocina techada integrada entre todos los dormitorios más amplias estacionamiento principal se entregan inmediatamente después de las ventanas termopaneles +56 9 6 7 8 5 4 3
DCR1 d(2.84e-09)	Cómodo departamento amoblado cercano al metro Santa Lucia [DISPONIBLE DESDE EL 01 DE MARZO 2022] CARACTERISTICAS - 39/43 m2 - 1 Dormitorio - 1 Baño - Sin Estacionamiento - Sin Bodega OTROS - Piso 7 - Orientación Poniente - Conexión a lavadora - Se aceptan mascotas - Full electric - Piso flotante - Ventanas Termopanel CARACTERISTICAS DEL EDIFICIO Edificio cuenta con sala multiuso, gimnasio, lavandería, ciclerero y un amplio patio interior Arrienda CUALQUIERA de nuestras propiedades y paga todos los gastos iniciales del arriendo EN CUOTAS con tu tarjeta de crédito - Primer mes de arriendo o proporcional - Mes de garantía - Comisión corredor
DCR2 d(6.74e-09)	Departamento de un dormitorio, en San Diego cercano a Avenida Matta y a Pedro Lagos, el metro más cercano es Parque O'higgins, el cual está a 15 minutos caminando, cercano a Universidad De Chile, Universidad Bernardo Ohiggins. Contacto Nicolás Ibáñez León+569 6577 1999.

4.2.8. Propiedades estadísticas - Conjunto A

El listado completo de las propiedades estadísticas se encuentra en el Anexo A.10. A continuación, se presentan las propiedades estadísticas en las que los modelos Tddpm y Smote muestran una diferencia mayor al 5 % con respecto al conjunto original de entrenamiento. Como referencia, se incluye el modelo Ctgan. Las variables se seleccionaron por ser 1) las que obtuvieron el peor resultado en cobertura y 2) las que obtuvieron el peor resultado en la distribución, respectivamente.

Tabla 4.28: Propiedades estadísticas de variable m_size con cambio >5 %, Economicos (A-1)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
nobs	22059	27574	27574	27574
mean	146269	1875594	73666	138617
std_err	105454	879088	46661	553
upper_ci	352956	3598574	165120	139702
lower_ci	-60417	152614	-17788	137532
std	15662334	145976185	7748272	91900
iqr	340.500	322.722	365.708	144556.805
iqr_normal	252.413	239.234	271.100	107160.120
mad	290635	3749106	145665	77299
mad_normal	364257	4698808	182564	96880
coef_var	107.079	77.829	105.181	0.663
range	2.24100e+09	1.54249e+10	1.19075e+09	3.92397e+05
max	2.24100e+09	1.54249e+10	1.19075e+09	3.92397e+05
min	0.000	0.440	1.546	0.000
skew	134.762	83.944	137.694	0.154
kurtosis	19053	7466	20489	2
jarque_bera	3.33616e+11	6.40295e+10	4.82276e+11	1.02812e+03
mode_freq	0.027	0.027	0.006	0.104
median	145.000	144.000	149.683	137376.175
0.1 %	2.000	4.095	15.010	0.000
1.0 %	22.000	23.844	25.000	0.000
25.0 %	66.000	68.000	67.667	63804.820
75.0 %	406.500	390.722	433.375	208361.625
95.0 %	5000	5000	4313	292812
99.0 %	10200	7698	9396	334157
99.9 %	70000	34028	53878	364825

Tabla 4.29: Propiedades estadísticas de variable county, Economicos (A-1)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Las Condes' 'Santiago' 'Providencia' 'Vitacura' 'Lo Barnechea']	['Las Condes' 'Santiago' 'Providencia' 'Vitacura' 'Lo Barnechea']	['Las Condes' 'Santiago' 'Providencia' 'Lo Barne- chea' 'Vitacura']	['Las Condes' 'Santiago' 'Viña del Mar' 'Vitacu- ra' 'Providencia']
top5_freq	[3233 2703 1481 1415 1322]	[4168 3463 1920 1807 1774]	[4666 3849 1951 1853 1814]	[3814 2195 1986 1508 1382]
top5_prob	[0.14656149 0.12253502 0.06713813 0.06414615 0.05993019]	[0.15115689 0.12558932 0.06963081 0.06553275 0.06433597]	[0.16921738 0.13958802 0.07075506 0.06720099 0.06578661]	[0.13831871 0.07960397 0.07202437 0.0546892 0.05011968]
nobs	22059	27574	27574	27574
missing	22059	0	0	0

4.2.9. SDMetrics Score - Conjunto B

Iniciaría contrastando los resultados entre ambos conjuntos para el modelo Tddpm La Tabla 4.30 muestra mejores *Score*, *Coverage*, *Column Shape* y *Column Pair Trends* comparadas con la Tabla 4.15. Puede deberse que al ser una cantidad de datos mayor, pudo tener más tiempo de aprender la distribución. Mejoría no notoria en los demás modelos, lo que podría indicar una mayor capacidad de Tddpm. Se puede ver que la cobertura es el indicador más bajo, solo alcanzando el 87 % en el mejor de los casos.

Tabla 4.30: Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, Economicos

Model Name	Column Pair Trends	Column Shapes	Coverage	Boundaries	Score
tddpm_mlp	0.98±2.79e-03	0.99±1.71e-03	0.87±3.37e-03	1.00±0.00e+00	0.98±1.85e-03
smote-enc	0.97±1.01e-03	0.92±1.07e-04	0.70±3.34e-02	1.00±0.00e+00	0.94±4.67e-04
copulagan	0.77±2.32e-02	0.78±1.75e-02	0.63±5.61e-04	1.00±0.00e+00	0.77±2.02e-02
tvae	0.78±1.68e-02	0.70±1.76e-02	0.28±3.70e-03	1.00±0.00e+00	0.74±1.48e-02
ctgan	0.77±1.35e-02	0.70±8.58e-03	0.63±9.52e-04	1.00±0.00e+00	0.73±5.42e-03
gaussiancopula	0.63±0.00e+00	0.63±7.85e-17	0.56±0.00e+00	1.00±0.00e+00	0.63±0.00e+00

4.2.10. Correlación - Conjunto B

Los modelos Smote y Tddpm, al ser comparados con el conjunto original, presentan diferencias marcadas. Los conjuntos sintéticos han creado correlaciones que no se ven presentes en los datos originales. En el caso del modelo Smote, se presentan correlaciones en las variables *bathrooms-rooms*, *m_size-m_built*; mientras que Tddpm adicionalmente genera una correlación entre *_price-m_size* y *_price-m_built*.

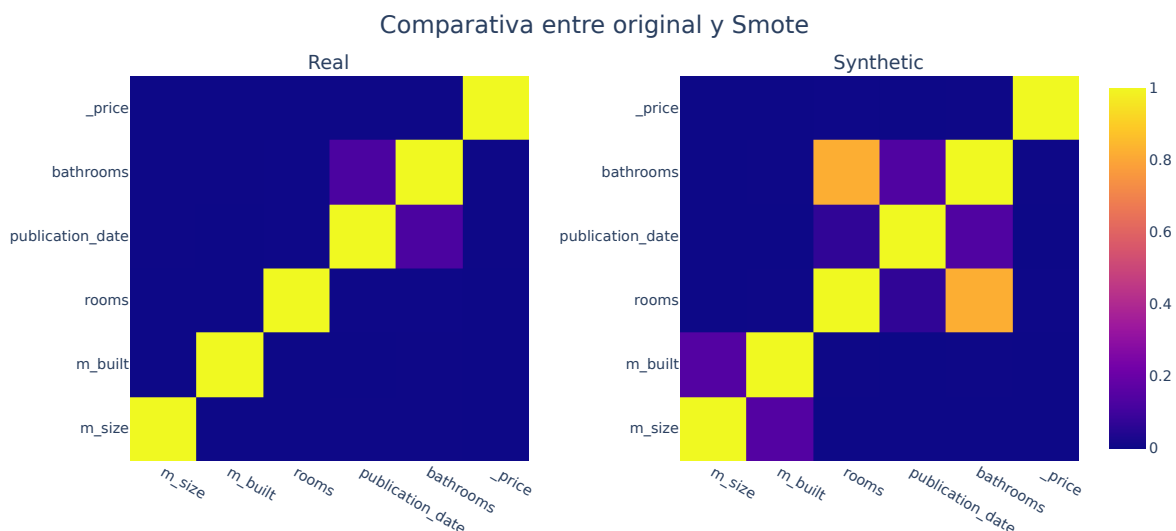


Figura 4.12: Correlación de conjunto original de entrenamiento y Smote

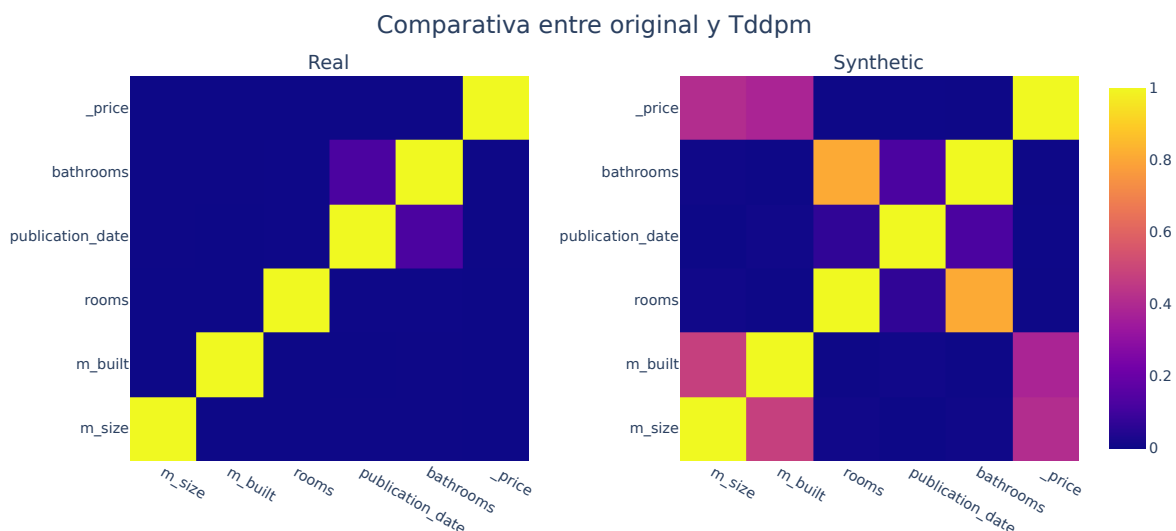


Figura 4.13: Correlación de conjunto original de entrenamiento y Tddpm

4.2.11. Cobertura - Conjunto B

La cobertura es notablemente baja en las variables *rooms* y *m_size* en Smote, y en *bathrooms* y *rooms* en el caso de Tddpm. En general, el modelo Tddpm es ligeramente superior a Smote.

Tabla 4.31: Evaluación de Cobertura Categoría-Rango para Modelos SMOTE-ENC y TDDPM_MLP, Economicos

Columna	Metrica	smote-enc	tddpm_mlp
_price	RangeCoverage	1.00e+00±0.00e+00	1.00e+00±0.00e+00
bathrooms	CategoryCoverage	7.59e-01±3.45e-02	4.77e-01±3.59e-02
county	CategoryCoverage	8.19e-01±9.11e-03	8.66e-01±1.49e-02
m_built	RangeCoverage	8.78e-02±1.49e-02	1.00e+00±0.00e+00
m_size	RangeCoverage	2.53e-01±2.92e-01	1.00e+00±0.00e+00
property_type	CategoryCoverage	7.28e-01±4.28e-02	9.01e-01±5.66e-02
publication_date	RangeCoverage	9.66e-01±5.52e-02	1.00e+00±0.00e+00
rooms	CategoryCoverage	4.23e-01±2.28e-02	4.93e-01±1.49e-02
state	CategoryCoverage	1.00e+00±0.00e+00	1.00e+00±0.00e+00
transaction_type	CategoryCoverage	1.00e+00±0.00e+00	1.00e+00±0.00e+00

4.2.12. Distribución - Conjunto B

Ambos modelos presentan buenas métricas, superando el 91 % en términos de distribución y forma. Sin embargo, se observan excepciones en los casos de *m_built* (85 %) y *m_size* (55 %).

Tabla 4.32: Evaluación de Similitud de Distribución para Modelos SMOTE-ENC y TDDPM_MLP, Economicos

Columna	Metrica	smote-enc	tddpm_mlp
_price	KSComplement	9.85e-01±1.94e-04	9.93e-01±8.05e-04
bathrooms	TVComplement	9.98e-01±3.13e-04	9.95e-01±4.98e-04
county	TVComplement	9.10e-01±5.37e-04	9.84e-01±2.56e-03
m_built	KSComplement	8.56e-01±1.32e-03	9.91e-01±1.44e-03
m_size	KSComplement	5.51e-01±8.46e-07	9.90e-01±2.65e-03
property_type	TVComplement	9.79e-01±7.12e-04	9.89e-01±3.27e-03
publication_date	KSComplement	9.66e-01±9.67e-05	9.91e-01±5.41e-03
rooms	TVComplement	9.87e-01±9.57e-04	9.95e-01±7.29e-04
state	TVComplement	9.78e-01±4.57e-04	9.90e-01±1.06e-03
transaction_type	TVComplement	9.94e-01±1.97e-04	9.97e-01±1.53e-03

4.2.13. Privacidad - Conjunto B

Las distancias mínimas para los percentiles 5 y 1 son varias magnitudes menores en el Conjunto B que en el Conjunto A, pasando de $\times 10^{-9}$ en el Conjunto A a $\times 10^{-15}$ en el Conjunto B, como se puede ver al comparar la Tabla 4.33 con la Tabla 4.18. Se puede afirmar que el 95 % de los registros tiene al menos una distancia de $9,12 \times 10^{-15}$.

Tabla 4.33: Distancia de registros más cercanos, percentil 5, datos economicos

Modelo	DCR ST	DCR SH	DCR TH	Score
tddpm_mlp	9.12e-15±1.09e-15	9.99e-15±8.14e-16	9.00e-17±0.00e+00	9.84e-01±1.85e-03
smote-enc	9.19e-15±6.41e-16	1.17e-14±6.96e-16	9.00e-17±0.00e+00	9.43e-01±4.67e-04
copulagan	2.65e-16±1.60e-16	2.84e-16±1.73e-16	9.00e-17±0.00e+00	7.74e-01±2.02e-02
tvae	1.00e-09±1.74e-09	1.00e-09±1.74e-09	9.00e-17±0.00e+00	7.38e-01±1.48e-02
ctgan	7.29e-09±8.52e-09	7.35e-09±8.45e-09	9.00e-17±0.00e+00	7.34e-01±5.42e-03
gaussiancopula	9.23e-13±0.00e+00	1.02e-12±0.00e+00	9.00e-17±0.00e+00	6.31e-01±0.00e+00

Tabla 4.34: Distancia de registros más cercanos, percentil 1, datos economicos

Modelo	DCR ST	DCR SH	DCR TH	Score
tddpm_mlp	4.63e-16±2.28e-17	4.16e-16±3.23e-17	0.00e+00±0.00e+00	9.84e-01±1.85e-03
smote-enc	1.83e-16±9.41e-18	2.54e-16±1.73e-17	0.00e+00±0.00e+00	9.43e-01±4.67e-04
copulagan	9.00e-17±1.01e-26	9.00e-17±1.30e-26	0.00e+00±0.00e+00	7.74e-01±2.02e-02
tvae	2.41e-16±4.18e-16	2.38e-16±4.12e-16	0.00e+00±0.00e+00	7.38e-01±1.48e-02
ctgan	1.57e-16±2.72e-16	1.87e-16±3.25e-16	0.00e+00±0.00e+00	7.34e-01±5.42e-03
gaussiancopula	1.15e-15±0.00e+00	1.37e-15±0.00e+00	0.00e+00±0.00e+00	6.31e-01±0.00e+00

Tabla 4.35: Distancia de registros más cercanos, minimo, datos economicos

Modelo	DCR ST	DCR SH	DCR TH	Score
tddpm_mlp	0.00e+00±0.00e+00	0.00e+00±0.00e+00	0.00e+00±0.00e+00	9.84e-01±1.85e-03
smote-enc	0.00e+00±0.00e+00	0.00e+00±0.00e+00	0.00e+00±0.00e+00	9.43e-01±4.67e-04
copulagan	4.57e-19±3.77e-21	5.21e-19±1.82e-22	0.00e+00±0.00e+00	7.74e-01±2.02e-02
tvae	8.99e-20±0.00e+00	8.99e-20±0.00e+00	0.00e+00±0.00e+00	7.38e-01±1.48e-02
ctgan	8.99e-20±0.00e+00	8.99e-20±0.00e+00	0.00e+00±0.00e+00	7.34e-01±5.42e-03
gaussiancopula	5.23e-19±0.00e+00	5.09e-19±0.00e+00	0.00e+00±0.00e+00	6.31e-01±0.00e+00

De las Tablas 4.36, 4.37 y 4.38 emergen dos características notables. La primera es que en el percentil 1 y el 5, en ambos casos, el modelo Tddpm mantiene la mayor razón entre el primer y el segundo registro más cercano. La segunda es que, al compararse con el Conjunto A (referenciado en la Tabla 4.22), la razón para el modelo Tddpm resulta ser superior.

Tabla 4.36: Proporción entre el más cercano y el segundo más cercano, percentil 5, datos económicos

Modelo	NNDR ST	NNDR SH	NNDR TH	Score
tddpm_mlp	3.03e-01±4.42e-03	2.96e-01±1.27e-02	1.15e-07±0.00e+00	9.84e-01±1.85e-03
smote-enc	2.47e-01±3.63e-03	2.60e-01±6.24e-03	1.15e-07±0.00e+00	9.43e-01±4.67e-04
copulagan	1.07e-05±4.91e-06	2.27e-05±1.82e-05	1.15e-07±0.00e+00	7.74e-01±2.02e-02
tvae	4.28e-04±2.75e-04	4.49e-04±2.88e-04	1.15e-07±0.00e+00	7.38e-01±1.48e-02
ctgan	2.10e-03±7.18e-04	7.23e-03±1.01e-02	1.15e-07±0.00e+00	7.34e-01±5.42e-03
gaussiancopula	1.52e-02±0.00e+00	1.38e-02±0.00e+00	1.15e-07±0.00e+00	6.31e-01±0.00e+00

Tabla 4.37: Proporción entre el más cercano y el segundo más cercano, percentil 1, datos económicos

Modelo	NNDR ST	NNDR SH	NNDR TH	Score
tddpm_mlp	3.14e-02±4.92e-03	3.08e-02±3.94e-03	0.00e+00±0.00e+00	9.84e-01±1.85e-03
smote-enc	2.52e-03±1.07e-03	3.47e-03±2.68e-04	0.00e+00±0.00e+00	9.43e-01±4.67e-04
copulagan	5.33e-09±1.38e-09	1.15e-07±1.65e-07	0.00e+00±0.00e+00	7.74e-01±2.02e-02
tvae	3.02e-05±4.14e-05	3.04e-05±4.15e-05	0.00e+00±0.00e+00	7.38e-01±1.48e-02
ctgan	1.21e-04±1.18e-04	1.35e-04±1.66e-04	0.00e+00±0.00e+00	7.34e-01±5.42e-03
gaussiancopula	6.43e-06±0.00e+00	6.43e-06±0.00e+00	0.00e+00±0.00e+00	6.31e-01±0.00e+00

Tabla 4.38: Proporción entre el más cercano y el segundo más cercano, mínimo, datos económicos

Modelo	NNDR ST	NNDR SH	NNDR TH	Score
tddpm_mlp	0.00e+00±0.00e+00	0.00e+00±0.00e+00	0.00e+00±0.00e+00	9.84e-01±1.85e-03
smote-enc	0.00e+00±0.00e+00	0.00e+00±0.00e+00	0.00e+00±0.00e+00	9.43e-01±4.67e-04
copulagan	6.76e-13±2.95e-13	1.49e-12±5.43e-13	0.00e+00±0.00e+00	7.74e-01±2.02e-02
tvae	1.51e-12±1.68e-13	4.64e-12±1.76e-13	0.00e+00±0.00e+00	7.38e-01±1.48e-02
ctgan	2.46e-12±1.48e-12	3.61e-12±2.05e-12	0.00e+00±0.00e+00	7.34e-01±5.42e-03
gaussiancopula	5.50e-14±0.00e+00	1.81e-12±0.00e+00	0.00e+00±0.00e+00	6.31e-01±0.00e+00

4.2.14. Ejemplos de registros - Conjunto B

En el ejemplo de las Tablas 4.39 y 4.40, corresponde a un departamento de dos dormitorios.

Tabla 4.39: Ejemplos para el modelo tddpm_mlp, percentil 2

Variable/Distancia	Sintético	DCR1 d(1.24e-15)	DCR2 d(7.82e-13)
_price	9.128134	12.735812	2490.000000
bathrooms	1.000000	1.000000	1.000000
county	Valparaíso	Santiago	Santiago
m_built	50.000000	41.000000	4929.000000
m_size	48.000000	43.000000	-1.000000
property_type	Departamento	Departamento	Departamento
publication_date	350.000000	350.000000	350.000000
rooms	2.000000	2.000000	2.000000
state	Valparaíso	Metropolitana de Santiago	Metropolitana de Santiago
transaction_type	Arriendo	Arriendo	Venta

Tabla 4.40: Ejemplos de texto modelo tddpm_mlp, percentil 2

Distancia	description
Sintético	Departamento de dos dormitorios, 2 baños, living comedor con salida a terraza, cocina amoblada equipada (incluye encimera), horno empotrado, campana, cubierta de granito, logia cerrada, estacionamiento subterráneo, bodega
DCR1 d(1.24e-15)	<p>Corredor arriendo, disponible inmediato, cercano a metro Franklin línea 2 y 6, supermercado 10, barrio Franklin y Bio Bio, amplio comercio, plazas, otros. Edificio Zenteno Efficient, año 2018, nuevo sin uso.</p> <p>San Diego 1721 ? Piso medio, 2 dormitorios, uno grande otro pequeño para cama de 1 plaza, 1 baño completo, cocina integrada, espacio para lavadora, sin balcón, sistema full electric en cocina, horno y termo electrico.</p> <p>Requisitos: 1.- Obligatorio 12 Cheques, puede ser del aval 2.- Sueldo TITULAR 3 veces el arriendo 3.- Certificado AFP últimos 12 meses 4.- Informe Dicom Platinum 5.- Cédula por ambos lados 6.- 1 mes de arriendo, 1 en garantía y comisión 50 %</p> <p>El Edificio cuenta con lavandería, sala multiuso, seguridad 24/7</p>
DCR2 d(7.82e-13)	<p>SE VENDE, Departamento CONDOMINIO EDIFICIO AVENIDA MATTÁ PLAZA, accesos controlados 24/7, cámaras de seguridad, alarma, timbres de pánico en cada Dpto, Ventanas de termopanel, citófono, cocina equipada con cubierta de granito, Hermosas áreas de jardines, 2 Dormitorios principal con Woking Closet y 1 Baño, Gimnasio equipado, Sala multiusos, Quinchos, Piscina, Sala primeros auxilios, Terrazas en segundo piso, Sala lavandería, estacionamientos de visitas. Excelente conectividad, Metro Irrazabal, privacidad y tranquilidad, además, cerca de supermercados, centros comerciales, jardines y colegios. Metros Cuadrados Metros Construidos: 47,29 M². Terraza Construida: 3 M². Terminaciones Piso baños: Cerámicos. Piso Living: Piso Flotante. Dormitorios: Alfombrados y Porcelanato. Otros suministros internet, teléfono, tv cable, wi-fi, Gastos Comunes \$ 50.000, No se paga Contribuciones. Precio: UF 2.490. ¡¡¡NO deje de visitar!!! Contáctanos: Carlos Miranda: +569 75894834. Paulina Montt: +569 96761295. Daniela Aguirre: +569 93221157. Email:: contacto@lodgepropiedades.cl</p>

En el ejemplo presentado en las Tablas 4.41 y 4.42, el registro sintético muestra coherencia con los datos de entrada. Por ejemplo, el texto generado corresponde a un departamento con dos dormitorios, aunque indica la existencia de un baño adicional en comparación con los datos de la publicación. Sin embargo, no proporciona otra información relevante que pueda correlacionarse con los datos estructurados de la publicación.

Tabla 4.41: Ejemplos para el modelo tddpm_mlp, percentil 4

Variable/Distancia	Sintético	DCR1 d(5.13e-15)	DCR2 d(1.00e-09)
_price	16.231131	11.115125	16.672687
bathrooms	1.000000	1.000000	1.000000
county	Ñuñoa	Pudahuel	La Florida
m_built	54.023199	-1.000000	200.000000
m_size	57.000000	-1.000000	270.000000
property_type	Departamento	Casa	Casa
publication_date	142.000000	142.000000	142.000000
rooms	2.000000	2.000000	3.000000
state	Metropolitana de Santiago	Metropolitana de Santiago	Metropolitana de Santiago
transaction_type	Arriendo	Arriendo	Arriendo

Tabla 4.42: Ejemplos de texto modelo tddpm_mlp, percentil 4

Distancia	description
Sintético	Departamento de dos dormitorios, 2 baños, living comedor con salida a terraza, cocina amoblada equipada encimera, horno empotrado, campana, logia cerrada, estacionamiento subterráneo para vehículos (conserjería las 24 horas)
DCR1 d(5.13e-15)	Casa interior, entrada independiente, dos dormitorios, baño, cocina comedor, pequeño patio, cerca de negocios, supermercados, servicio urgencia, consultorio, cerca metro pudahuel, 300.000 mensual, 1 meses garantia, luz y agua adicional. Solo personas quieran vivir lugar tranquilo. Consultas 9-44104648 llama coordina tu visita, sin estacionamiento llamar lunes a viernes desde las 17:30, sabado y domingo mismo horario
DCR2 d(1.00e-09)	Arriendo Comercial / Habitacional. Casa aislada. Entrada auto (2) prestación de Servicios computacionales, asesorias mas informacion al correo mh@rammsy.cl

4.2.15. Propiedades estadísticas - Conjunto B

El listado completo de las propiedades estadísticas se encuentra en el Anexo A.11. A continuación, se presentan las propiedades estadísticas en las que los modelos Tddpm y Smote muestran una diferencia mayor al 5 % con respecto al conjunto original de entrenamiento. Como referencia, se incluye el modelo Ctgan. Las variables se seleccionaron por ser 1) las que obtuvieron el peor resultado en cobertura y 2) las que obtuvieron el peor resultado en la distribución, respectivamente.

Tabla 4.43: Propiedades estadísticas de variable bathrooms con cambio >5 %, Economicos (B-1)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
nobs	545870	682338	682338	682338
mean	0.815	0.790	0.809	1.507
std_err	0.003	0.002	0.002	0.007
upper_ci	0.820	0.793	0.813	1.521
lower_ci	0.810	0.786	0.805	1.493
std	1.898	1.609	1.691	5.740
mad	1.376	1.359	1.375	2.494
mad_normal	1.725	1.703	1.723	3.126
coef_var	2.328	2.037	2.089	3.809
range	437.000	116.000	146.000	437.000
max	436.000	115.000	145.000	436.000
skew	36.380	1.000	3.134	15.447
kurtosis	6629	41	165	527
jarque_bera	9.98582e+11	4.22514e+07	7.44954e+08	7.84347e+09
99.9 %	9.000	7.000	9.000	83.000

Tabla 4.44: Propiedades estadísticas de variable m_size con cambio >5 %, Economicos (B-1)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
nobs	545870	682338	682338	682338
mean	2.03551e+16	1.54755e+18	3.16803e+11	1.86487e+15
std_err	2.03549e+16	1.25490e+17	1.82548e+11	3.05751e+12
upper_ci	6.02499e+16	1.79351e+18	6.74592e+11	1.87086e+15
lower_ci	-1.95397e+16	1.30160e+18	-4.09850e+10	1.85888e+15
std	1.50388e+19	1.03659e+20	1.50792e+14	2.52561e+15
iqr	181.000	171.000	210.281	3211862663343589.000
iqr_normal	134.176	126.762	155.882	2380957355104258.000
mad	4.07100e+16	3.09404e+18	6.33595e+11	2.03095e+15
mad_normal	5.10224e+16	3.87781e+18	7.94093e+11	2.54542e+15
coef_var	738.823	66.983	475.979	1.354
range	1.11111e+22	1.11111e+22	8.17891e+16	2.07066e+16
max	1.11111e+22	1.11111e+22	8.17891e+16	2.07066e+16
min	-1000.000	-1000.000	-881.043	-1000.000
skew	738.828	82.168	507.710	1.473
kurtosis	545868	7532	264489	5
jarque_bera	6.77722e+15	1.61221e+12	1.98885e+15	3.39104e+05
mode_freq	0.449	0.454	0.279	0.454
median	36.000	37.000	46.170	472364610529319.625
75.0 %	180.000	170.000	209.281	3211862663342589.000
95.0 %	5000	5000	5070	7129564968780261
99.0 %	50000	44090	68451	9903363003858036
99.9 %	4920000	1753435	6574780	13009696522973218

Capítulo 5

Conclusiones y discusión

5.1. Conclusiones

El objetivo principal de este estudio fue desarrollar un mecanismo para generar conjuntos de datos sintéticos estructurados, incluyendo textos, y comparar estos datos generados con sus contrapartes originales. Para lograr esto, se elaboró código y se examinaron los resultados producidos por varios enfoques, incluyendo **Tddpm**, **Smote**, **Ctgan**, **Tablepreset**, **Copulagan**, **Gaussiancopula** y **Tvae** para datos tabulares. Cada uno de estos modelos ha mostrado un grado de éxito notable en términos de distribución, correlación y cobertura. En lo que respecta a la generación de texto, se empleó el modelo **mt5**, que es un derivado de la serie de modelos **T5** y fue *fine-tuned* para el conjunto original. Este modelo ha demostrado su capacidad para producir textos coherentes basados en las entradas proporcionadas, aunque decepcionante en su capacidad de diversidad de los textos generados. Si se observa el anexo A.7 se puede notar que existen inicios de texto repetidos "Piso de madera en", "Departamento de 2 piso" son dos frecuentes inicios. Lo anterior no invalida la generación.

Además, se presentaron comparativas de métricas para facilitar la selección de modelos. Entre estas métricas se incluyen el **SDMetric Score**, que considera la distribución a través de las tendencias de pares de columnas (*Column Pair Trends*) y las formas de las columnas (*Column Shapes*). También se consideraron métricas de cobertura (*Coverage*) y límites (*Boundaries*). En este contexto, dos modelos tabulares sobresalieron: **Tddpm** y **Smote**.

En lo referente a la privacidad, se exploró la relación existente entre utilidad representada por *SDMetric Score* y privacidad representada por la distancia al registro más cercano (**DCR**). Se observó que a medida que el conjunto sintético se asemejaba más al original, mayor *SDMetric Score*, las métricas de privacidad disminuían, como se reflejaba en la disminución *DCR* y la relación entre el registro más cercano y el segundo más cercano (**NDR**). Los modelos que tenían una mayor distancia generalmente rendían peor, y esto no se limitaba únicamente a la calidad del modelo. Al comparar los dos mejores modelos, **Tddpm** y **Smote**, se encontró que **Tddpm** superaba a **Smote** en términos de mayores distancias y una mayor razón en la distancia del primer al segundo registro, lo que proporcionaba una mayor protección al conjunto original.

Basándonos en nuestras observaciones, si se considera que la distancia al percentil 5 propor-

ciona una salvaguarda suficiente para la privacidad, recomendamos el uso del modelo **Tddpm**. Sin embargo, esta recomendación está sujeta a revisión a medida que se desarrollen y evalúen más modelos. Adicionalmente, es importante realizar un cálculo particular para cada nuevo conjunto de datos, ya que, como se observó, la medida cambiará en función de cuán bien aprenda el modelo y las características de los datos, como la cantidad de nulos o el número de variables categóricas, por ejemplo.

En resumen, nuestro estudio contribuye al creciente cuerpo de literatura en el campo de la generación de datos sintéticos y ofrece una base sólida para futuras investigaciones.

5.2. Limitaciones

A pesar de los hallazgos significativos, nuestro estudio tiene ciertas limitaciones. Por ejemplo, debido a limitaciones de tiempo, no se pudo realizar una evaluación completa de todas las métricas listadas en la revisión bibliográfica. Estas limitaciones ofrecen oportunidades para futuras investigaciones. En particular, se recomienda la exploración de la *Cross-Classification* como una medida para la evaluación de textos.

Y más allá de las métricas que por tiempo no pudieron ser analizadas, me gustaría contextualizar en el tiempo que se desarrolló este trabajo y es que al iniciar esta tesis, a mediados de 2022, era evidente la creciente complejidad de la inteligencia artificial y la mejora continua de sus resultados. Sin embargo, era difícil anticipar que, durante el desarrollo de este trabajo, este crecimiento se aceleraría de tal manera que, al acercarse a la conclusión del mismo, los modelos utilizados aquí, como el **T5**, constituirían solo uno de los muchos modelos disponibles para la generación de texto. Desafortunadamente, debido a limitaciones de tiempo, no se podrán tratar otros modelos en profundidad y estos serán objeto de la sección de discusión. Para aquellos interesados en continuar con esta línea de investigación, se recomienda visitar el HuggingFace Leaderboard LLM [6], donde encontrarán una amplia gama de modelos a explorar y explotar.

Ejemplo de uso de ChatGPT en el marzo 2023 [30].

Prompt

Introducción a la relevancia de la generación de datos sintéticos para una tesis.

ChatGPT-4

En la contemporánea era de la información, los datos son un componente crucial en la toma de decisiones basadas en evidencias en un diverso espectro de sectores. No obstante, la abundancia de datos accesibles no se encuentra exenta de retos significativos, tales como los relacionados con la privacidad y seguridad de los datos, y la variabilidad y calidad de los mismos. En este contexto, la generación de datos sintéticos se erige como una herramienta valiosa para enfrentar dichos desafíos. Los datos sintéticos, producidos a partir de modelos que capturan las características intrínsecas de los datos originales, poseen el potencial de preservar la privacidad y confidencialidad, mientras habilitan la realización de análisis y modelado. Esta tesis se enfoca en el desarrollo y evaluación de métodos para la generación de datos sintéticos, ofreciendo una contribución significativa a este campo en expansión.

En la presentación de resultados y en las conclusiones, se evidenciaron registros con similitudes notables. Para incrementar la diversidad en los resultados, se propone incorporar elementos aleatorios en la entrada, con el fin de inducir una mayor variedad en los inicios. Otra opción sería ajustar los parámetros de generación, brindando al generador un mayor margen de "libertad creativa."^a través del ajuste del parámetro de temperatura. Este es un aspecto que claramente ofrece espacio para mejoras.

Muchos de los párrafos de este trabajo fueron analizados por ChatGPT, corregidos y mejorados. El impacto de los modelos generativos no solo quedará restringido a la capacidad de sustituir datos reales por datos sintéticos, como vimos en este estudio. También serán asistentes de bajo costo para tareas que antes estaban restringidas únicamente a humanos.

5.3. Discusión

Los modelos de generación de texto están en pleno auge. Recientemente han emergido modelos como **GPT-4** [31], **Llama** [4], **Palm2** [15] y **Falcon** [2], entre muchos otros que se pueden ver en el HuggingFace Leaderboard LLM [6]. El modelo **Chinchilla** [23] ha destacado la importancia de la calidad de los datos de entrada para la eficacia de estos modelos. Sería relevante llevar a cabo nuevos estudios con estos y otros modelos emergentes.

En relación a las métricas, tal como se mencionó en la conclusión, algunas de ellas no se calcularon en este trabajo debido a restricciones de tiempo. Además, el estudio de la privacidad en la generación de texto es un área que aún no ha sido ampliamente explorada. Determinar qué métricas son relevantes en este aspecto podría ser tan importante como la evaluación de la eficacia de los nuevos modelos.

Bibliografía

- [1] DALL·e 2.
- [2] Falcon LLM.
- [3] Imagen: Text-to-image diffusion models.
- [4] LLaMA: Open and efficient foundation language models - meta research.
- [5] Microsoft and google are in a ‘game of thrones’ battle over a.i.— but apple and amazon still have huge roles to play, according to wedbush.
- [6] Open LLM leaderboard - a hugging face space by HuggingFaceH4.
- [7] Papers with code - ImageNet benchmark (image classification).
- [8] SDMetrics.
- [9] Stable diffusion public release.
- [10] Angeela Acharya, Siddhartha Sikdar, Sanmay Das, and Huzefa Rangwala. GenSyn: A multi-stage framework for generating synthetic microdata using macro data sources.
- [11] Accountability Act. Health insurance portability and accountability act of 1996. 104:191.
- [12] Kiran Adnan and Rehan Akbar. An analytical study of information extraction from unstructured and multidimensional big data. 6:1–38. Publisher: Springer.
- [13] Akim. TabDDPM: Modelling tabular data with diffusion models. original-date: 2022-10-02T23:01:07Z.
- [14] Ewa Andrejczuk, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, and Yase-min Altun. Table-to-text generation and pre-training with TabT5.
- [15] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Pas-sos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Mo-reira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer,

- Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. PaLM 2 technical report.
- [16] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators.
 - [17] Peter Bruce, Andrew Bruce, and Peter Gedeck. *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media.
 - [18] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. 16:321–357.
 - [19] Sabrina De Capitani Di Vimercati, Sara Foresti, Giovanni Livraga, and Pierangela Samarati. Data privacy: Definitions and techniques. 20(6):793–817. Publisher: World Scientific.
 - [20] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media.
 - [21] John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. 2007(2012):1–16.
 - [22] Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. TaPas: Weakly supervised table parsing via pre-training.
 - [23] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models.
 - [24] HARLFOXEM Kaggle. House sales in king county, USA.
 - [25] Mihir Kale and Abhinav Rastogi. Text-to-text pre-training for data-to-text tasks.
 - [26] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Overview — SDV 0.18.0 documentation.
 - [27] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models.

- [28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- [29] Dan Milmo and Dan Milmo Global technology editor. Google v microsoft: who will win the AI chatbot race?
- [30] OpenAI. ChatGPT: a large language model trained by OpenAI.
- [31] OpenAI. GPT-4 technical report.
- [32] Stuart L Pardau. The california consumer privacy act: Towards a european-style privacy regime in the united states. 23:68. Publisher: HeinOnline.
- [33] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410. IEEE.
- [34] David Pujol, Amir Gilad, and Ashwin Machanavajjhala. PreFair: Privately generating justifiably fair synthetic data.
- [35] Protection Regulation. Regulation (EU) 2016/679 of the european parliament and of the council. 679:2016.
- [36] Aivin V Solatorio and Olivier Dupriez. REaLTabFormer: Generating realistic relational and tabular data using transformers.
- [37] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. 32.
- [38] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. CTAB-GAN: Effective table data synthesizing. In Vineeth N. Balasubramanian and Ivor Tsang, editors, *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pages 97–112. PMLR.
- [39] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. CTAB-GAN+: Enhancing tabular data synthesis.

Apéndice A

Anexos

A.1. Código de entrenamiento de economicos

```
1 import pandas as pd
2 from syntheticml.data.synthetic import Synthetic, MODELS
3 from syntheticml.models.tab_ddpm.sdv import SDV_MLP
4 import torch
5 import numpy as np
6 import itertools
7 import multiprocessing as mp
8 import os
9
10 def test_train(args):
11     lrc, ntc, sts, btsc, rtdlc, syn, df = args
12     #notebooks/economicos_good/2e-06_10_100000_5000_1024-512-256
13     checkpoint = "economicos_good2/" + "_".join(
14         map(str, [lrc, ntc, sts, btsc, "-".join(map(str, rtdlc))]))
15     checkpoint = "con_fechas"
16     if os.path.exists(f"{checkpoint}/final_model.pt") or os.path.exists(f"{checkpoint}/exit"):
17         return (checkpoint, 1)
18     model = SDV_MLP(syn.metadata,
19                     "_price",
20                     exclude_columns=syn.exclude_columns,
21                     df=df,
22                     batch_size=btsc,
23                     steps=sts,
24                     checkpoint=checkpoint,
25                     num_timesteps=ntc,
26                     weight_decay=0.0,
27                     lr=lrc,
28                     model_params=dict(rtdl_params=dict(
29                         dropout=0.0,
30                         d_layers=rtdlc
31                     ))
32                 )
33     model.fit(syn.train)
34     model.save(f"{checkpoint}/final_model.pt")
35     return (checkpoint, 1)
36
37 if __name__ == '__main__':
38     df = pd.read_parquet('../datasets/economicos/synth/split/train.parquet')
39     category_columns=("property_type", "transaction_type", "state", "county", "rooms", "bathrooms", "m_built", "m_size", "source", )
40     # TODO: Estudiar implicancia de valores nulos en categorias y numeros
41     df_converted = df.astype({k: 'str' for k in ("description", "price", "title", "address", "owner")})
42     basedate = pd.Timestamp('2017-12-01')
43     dtm = df_converted.pop("publication_date")
44     df_converted["publication_date"] = dtm.apply(lambda x: (x - basedate).days)
45     syn = Synthetic(df_converted,
46                   id="url",
47                   category_columns=category_columns,
48                   text_columns=("description", "price", "title", "address", "owner",),
49                   exclude_columns=tuple(),
50                   synthetic_folder = "../datasets/economicos/synth",
51                   models=['copulagan', 'tvae', 'gaussiancopula', 'ctgan', 'smote-enc'],
52                   n_sample = df.shape[0],
53                   target_column="_price"
54                 )
55
56     lrs = np.linspace(2e-6, 2e-3, 10)
```

A.2. Lista completa de figura pairwise kingcounty

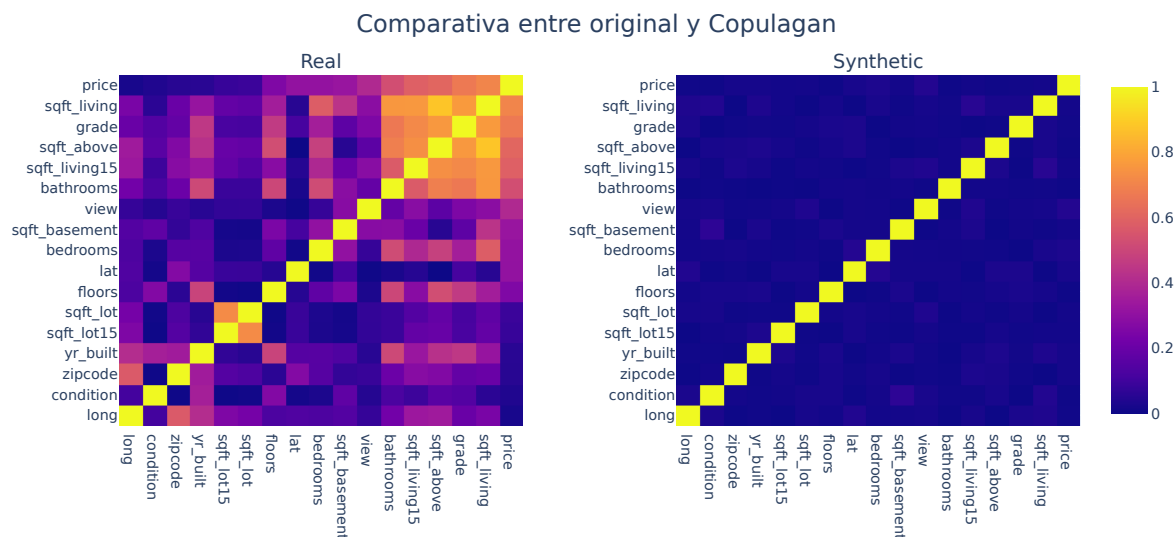


Figura A.1: Correlación de conjunto original de entrenamiento y Copulagan

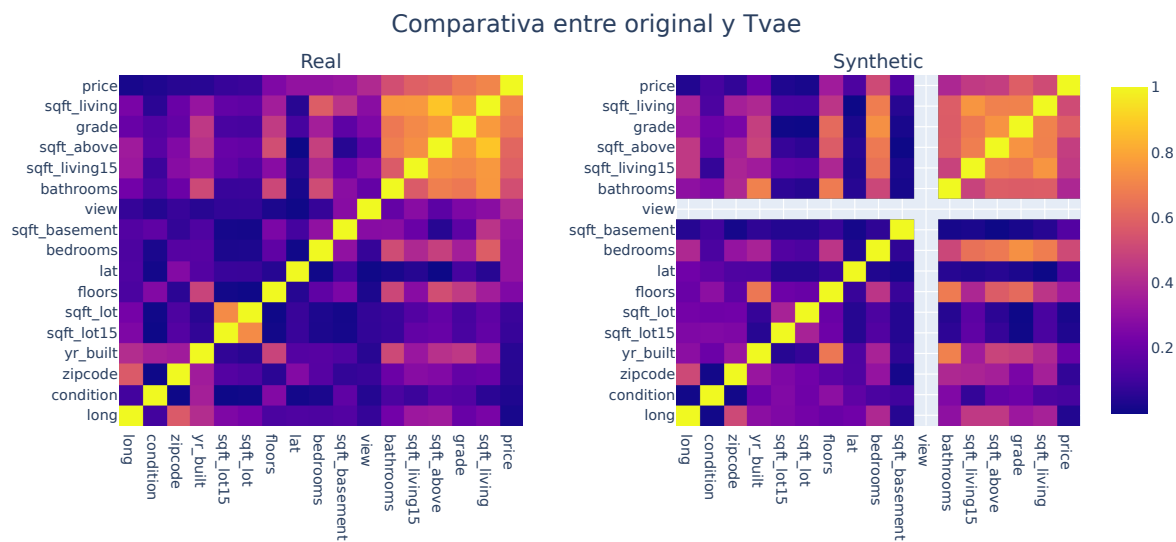


Figura A.2: Correlación de conjunto original de entrenamiento y Tvae

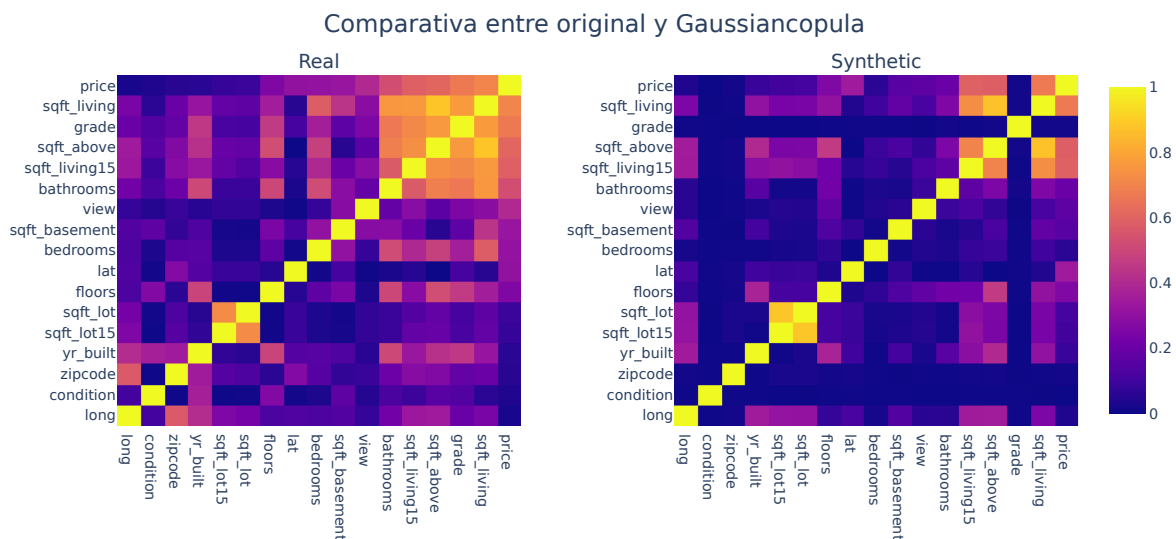


Figura A.3: Correlación de conjunto original de entrenamiento y Gaussiancopula

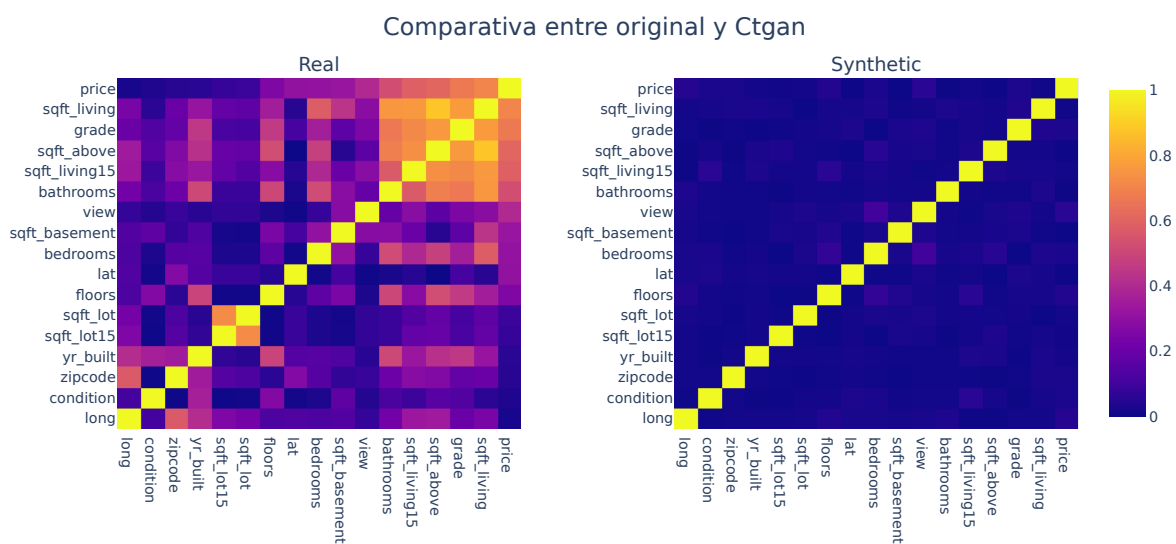


Figura A.4: Correlación de conjunto original de entrenamiento y Ctgan

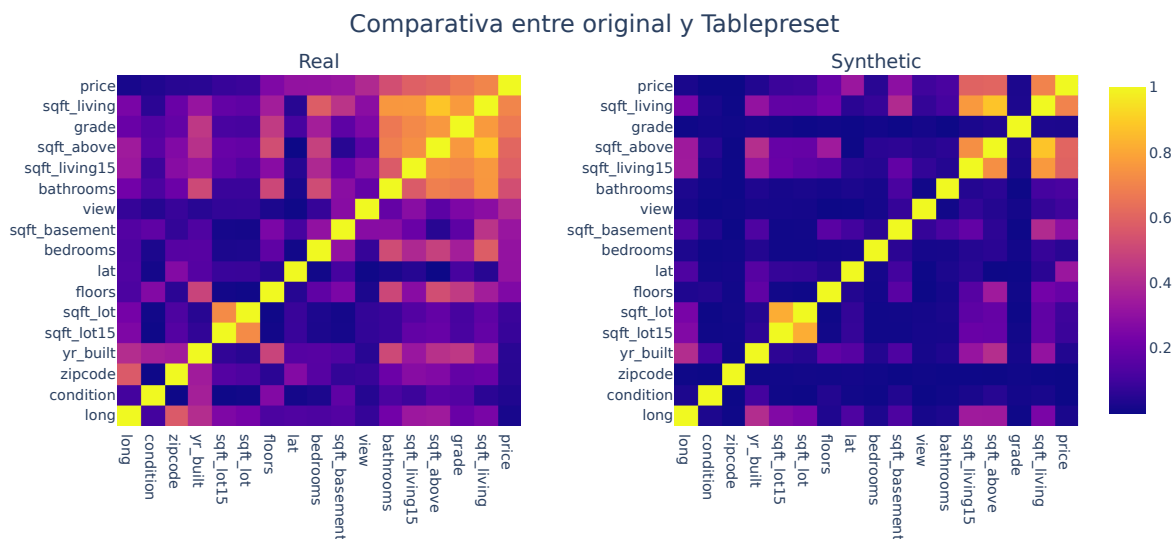


Figura A.5: Correlación de conjunto original de entrenamiento y Tablepreset

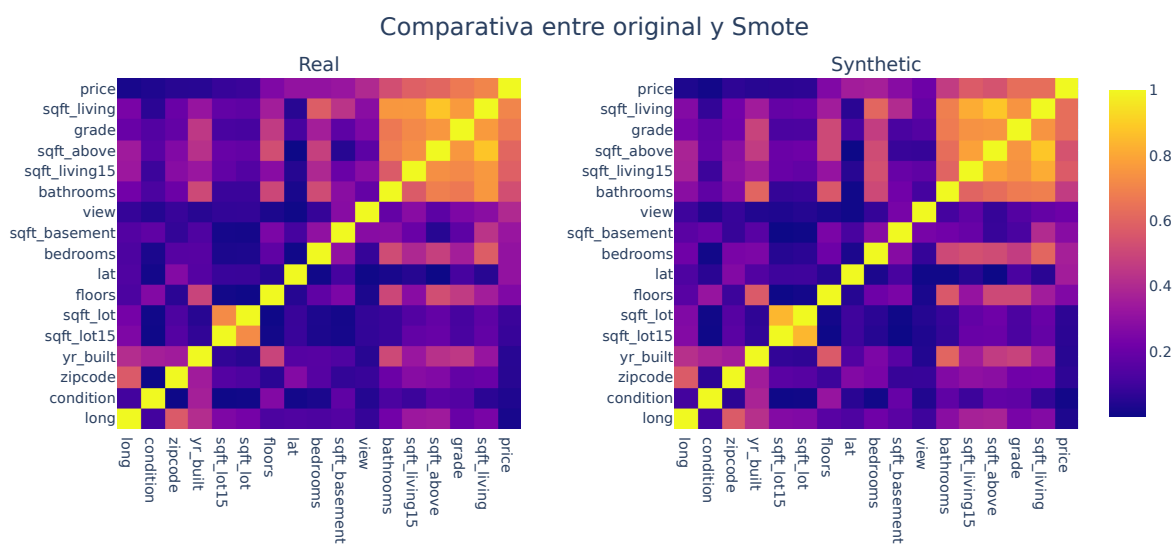


Figura A.6: Correlación de conjunto original de entrenamiento y Smote

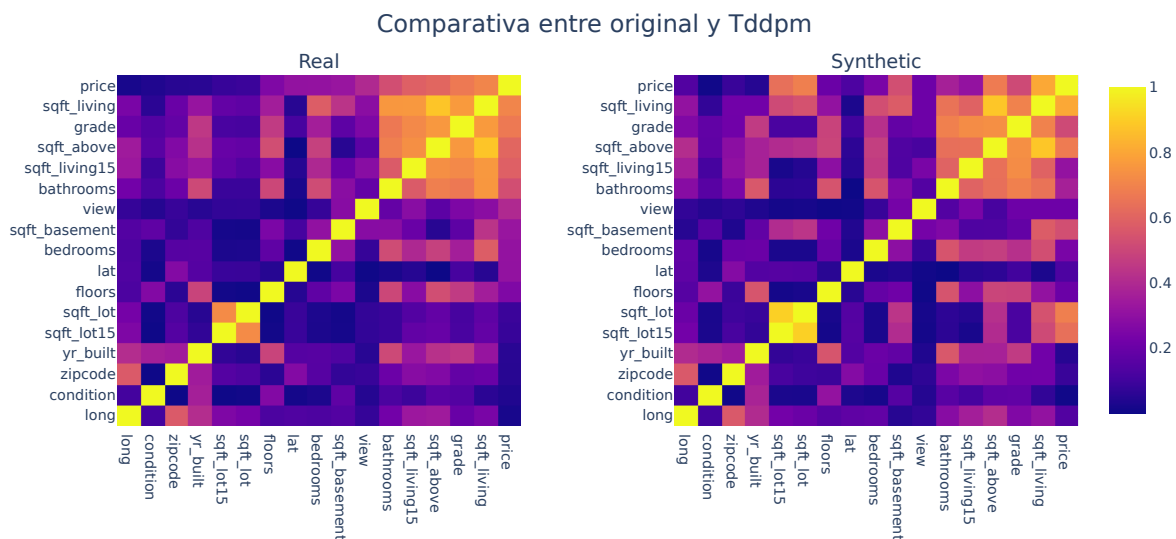


Figura A.7: Correlación de conjunto original de entrenamiento y Tddpm

A.3. Smote y Tddpm en KingCounty Graficas por Columnas

A.4. Tabla de comparación de Top5 KingCounty

A.5. Figuras de correlación Economicos - Conjunto A

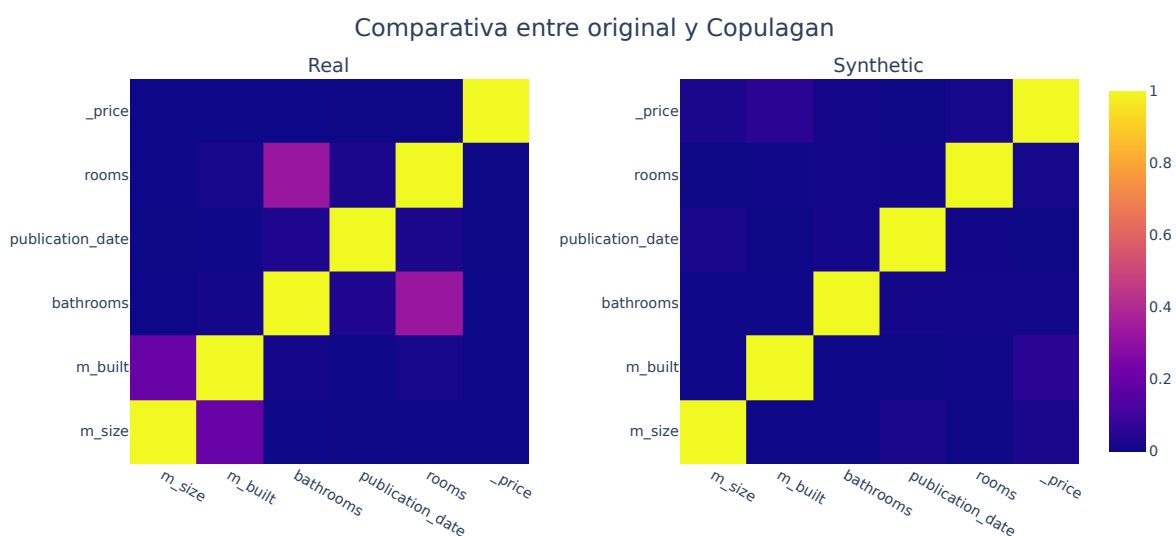


Figura A.8: Correlación de conjunto original de entrenamiento y Copulagan

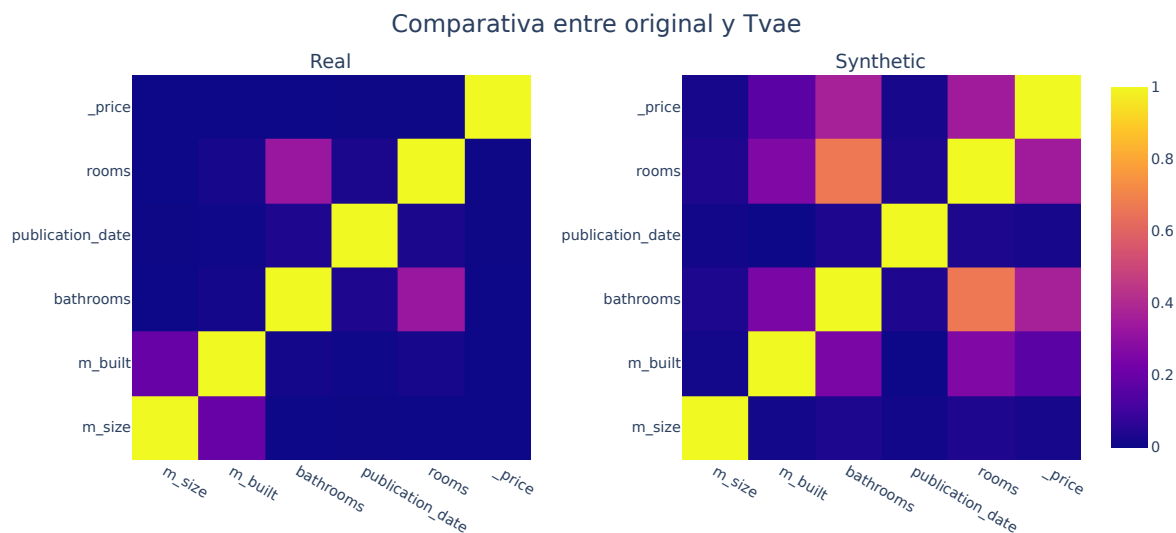


Figura A.9: Correlación de conjunto original de entrenamiento y Tvae

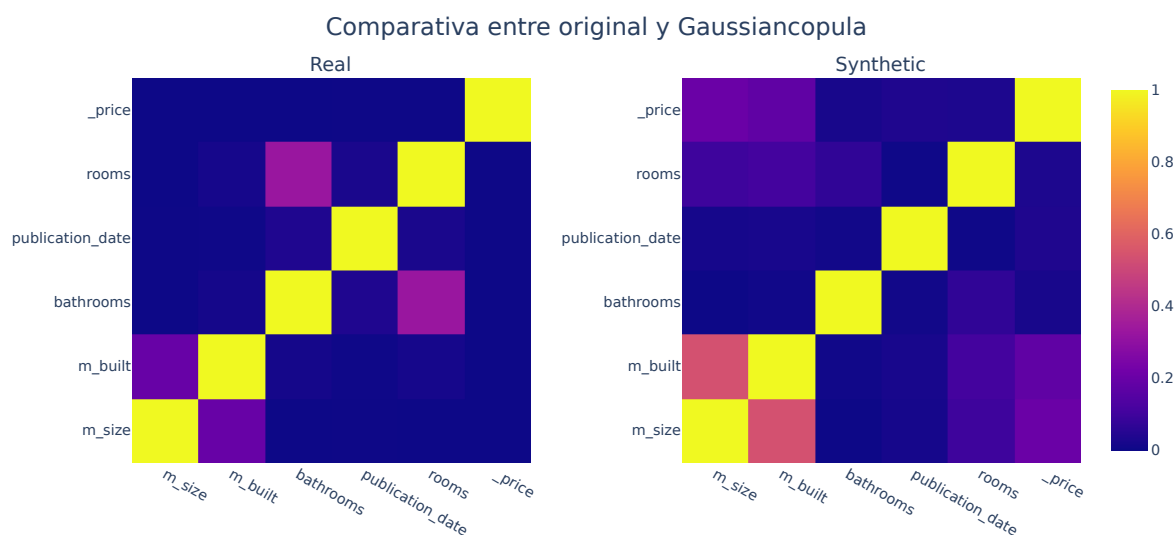


Figura A.10: Correlación de conjunto original de entrenamiento y Gaussiancopula

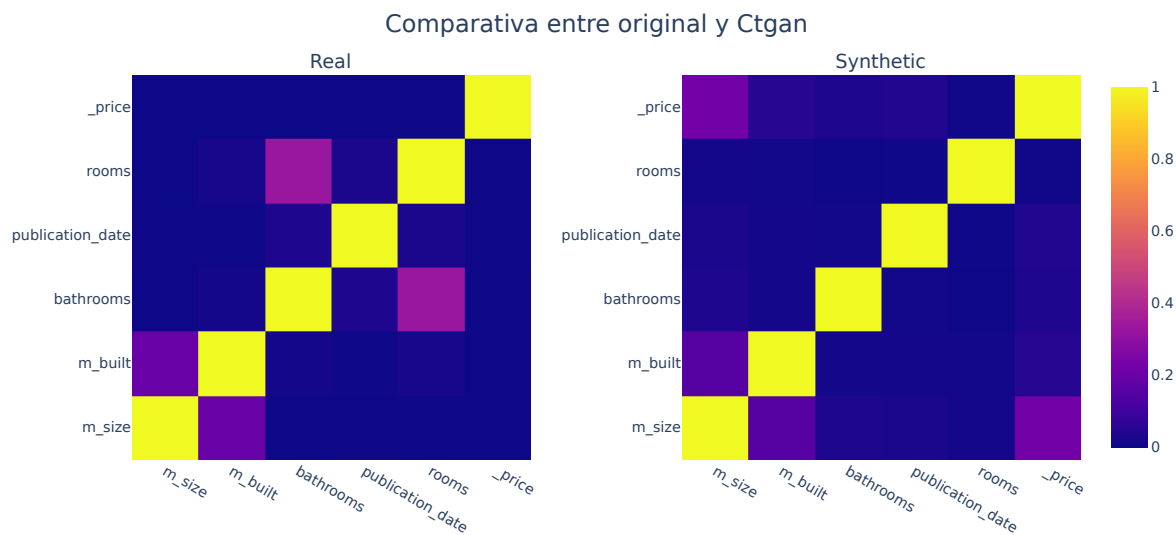


Figura A.11: Correlación de conjunto original de entrenamiento y Ctgan

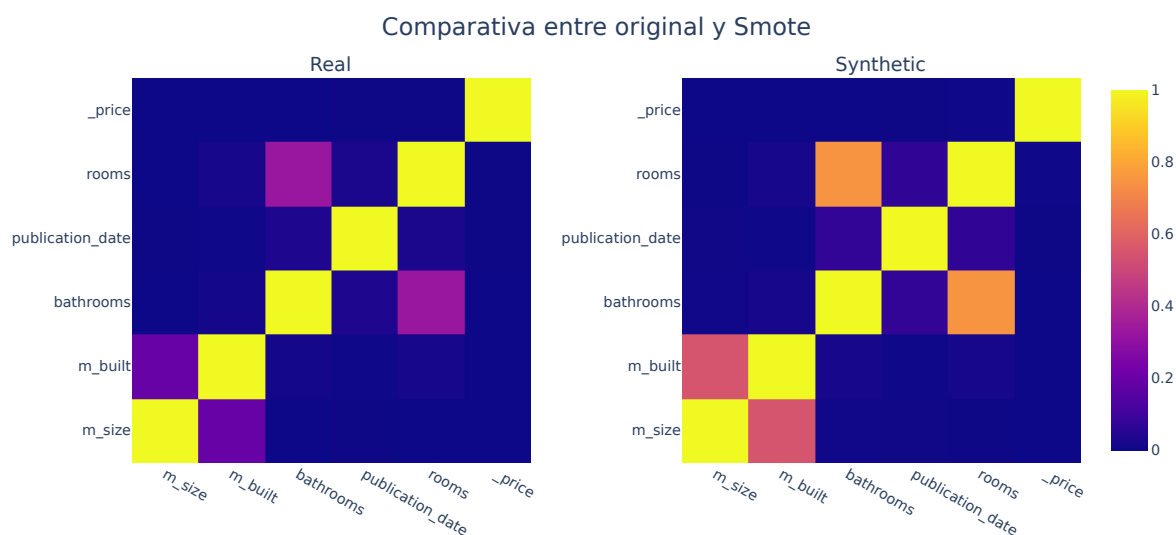


Figura A.12: Correlación de conjunto original de entrenamiento y Smote

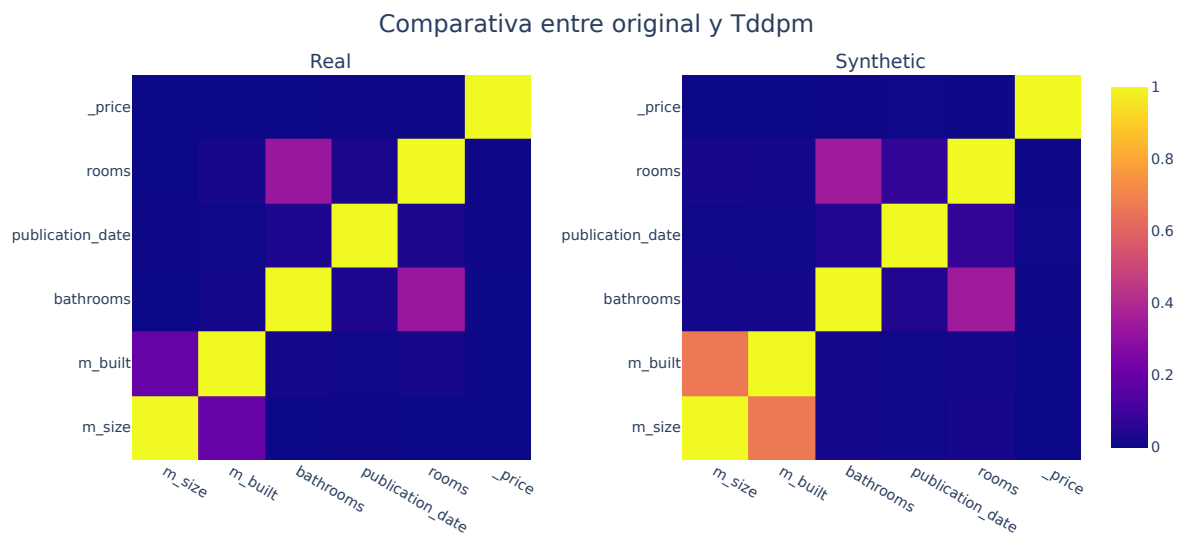


Figura A.13: Correlación de conjunto original de entrenamiento y Tddpm

A.6. Figuras de correlación Economicos - Conjunto B

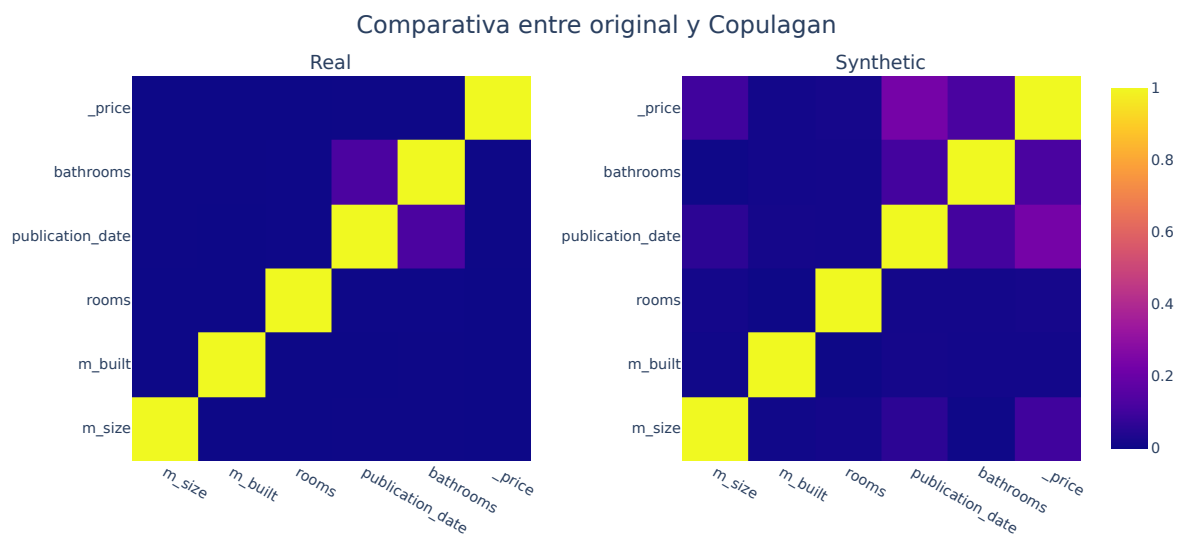


Figura A.14: Correlación de conjunto original de entrenamiento y Copulagan

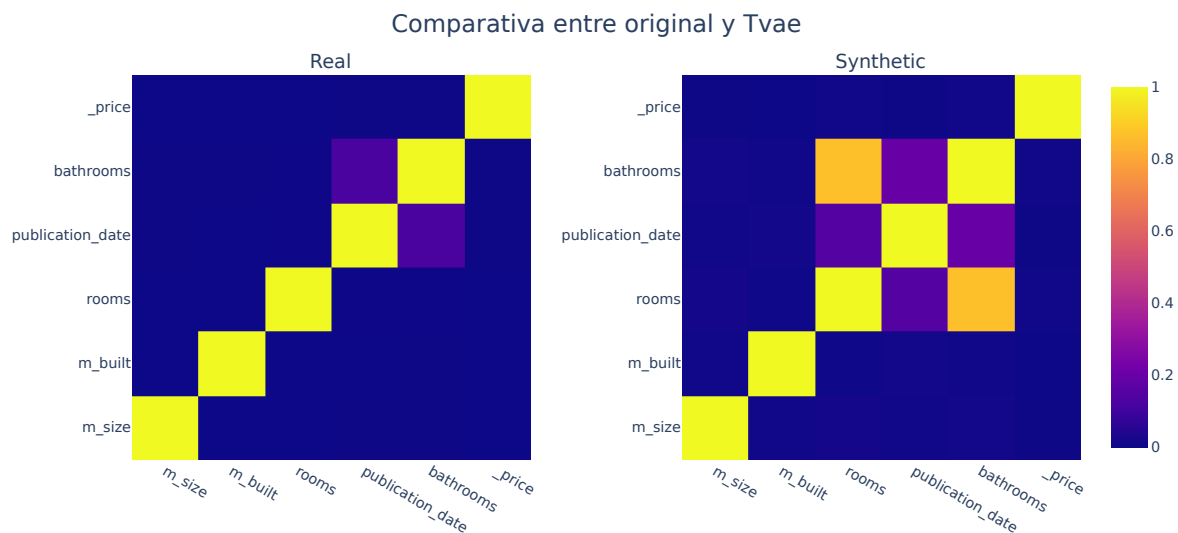


Figura A.15: Correlación de conjunto original de entrenamiento y Tvae

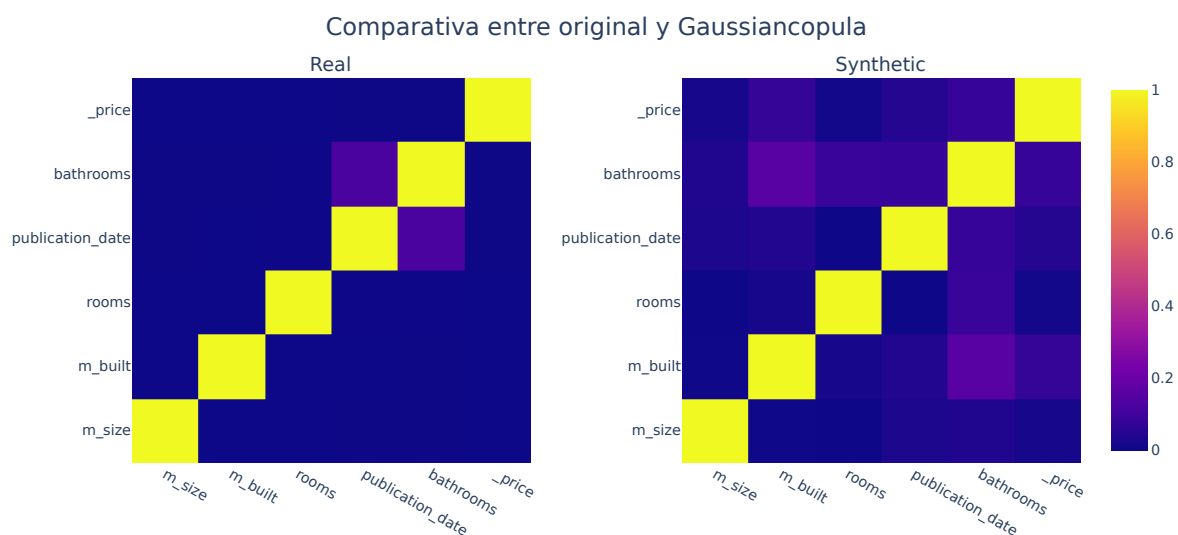


Figura A.16: Correlación de conjunto original de entrenamiento y Gaussiancopula

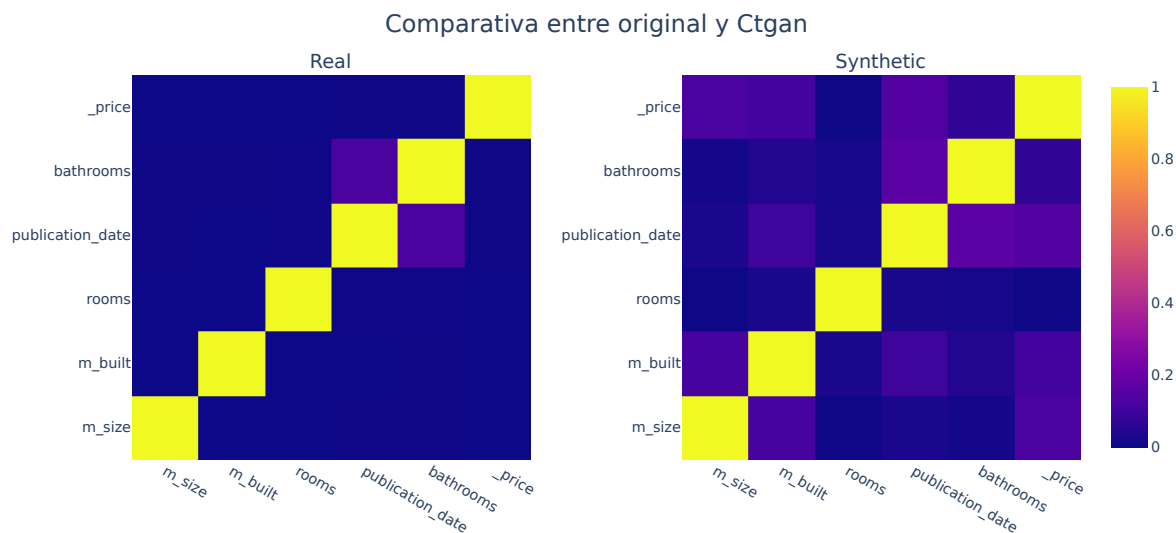


Figura A.17: Correlación de conjunto original de entrenamiento y Ctgan

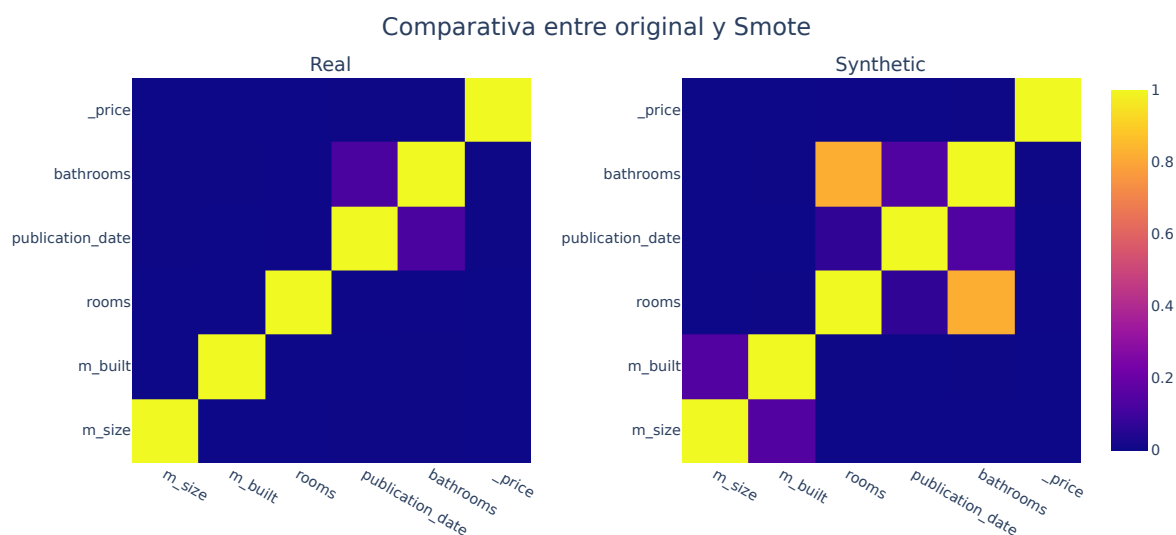


Figura A.18: Correlación de conjunto original de entrenamiento y Smote

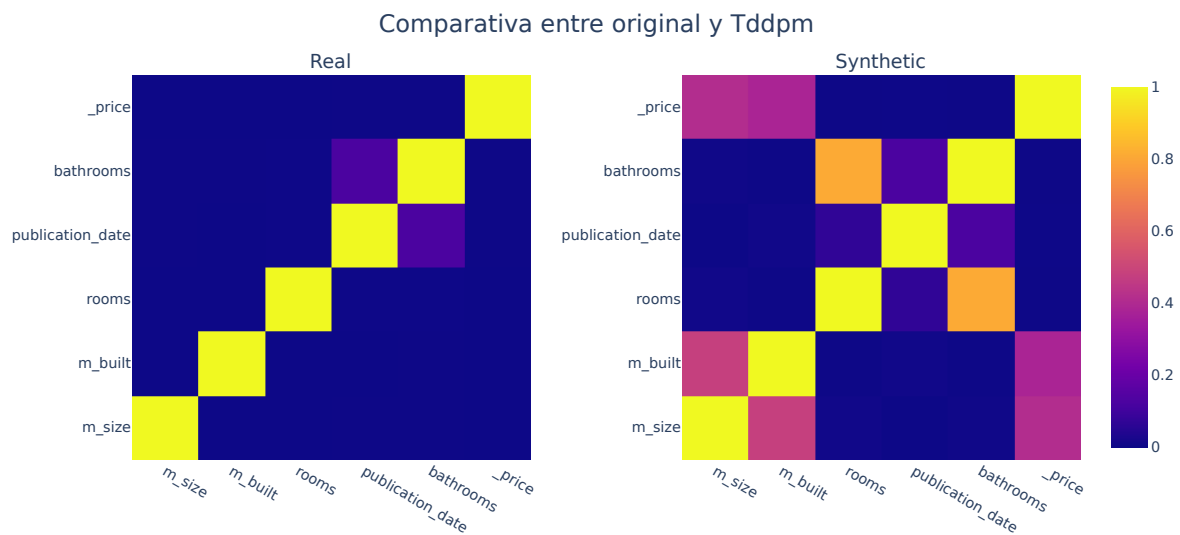


Figura A.19: Correlación de conjunto original de entrenamiento y Tddpm

A.7. Ejemplos de 10 Registros Generados Aleatoriamente en Descripciones Económicas A-1

Tabla A.1: Ejemplos de textos aleatorios del modelo tddpm_mlp, conjunto Economicos a-1

description
Departamento de 2 pisos, 1 baño con salida a terraza en living comedor que comparte un gran jardín para una amplia vista panorámica al barrio del Barrio Los Barneches (principal) el edificio cuenta con lavandería completamente equipada por las ventanas termopanel/calefacción central-conserjería techada +56 9 7 8 6 5 4
Piso de madera en el primer piso, living comedor con salida a terraza que tienen 2 baños completos (incluye una sala de estar) Living comedor amplio para lavadora separados por un bodega altura del hall de entrada; logia equipada como estacionamiento techado cerrado! Calefacción central controlada: Horno termopanel/Horno Termopanel-Calefacción radiador trasero Condominio Amplio Hall de entrada Cocina amoblada Amueblado 3 estacionamientos 1 Baño Edificio cuenta con dos salas de reuniones + quincho
Piso de madera en el primer piso, living comedor con salida a terraza que tienen 2 baños completos (incluye una sala de estar) Living comedor amplio para dos autos (2 estacionamientos) Cocina equipada completamente equipada Con vista panorámica al jardín El tercer piso cuenta con un gran hall de entrada por lavandería Calefacción central Azotea Techada: 1 bodega
Piso de madera en el living comedor con salida a la terraza, 2 baños completos que comparten una gran sala de estar para dos autos (incluye un bodega).
Departamento de 2 pisos, 1 baño completo, living comedor con salida a terraza en el primer piso: Living comedor amplio (cocina equipada) Cocina completamente equipada Con vista al jardín El departamento cuenta con una sala de estar Amplio Hall de entrada Calefacción por radiadores Ventanas termopanel Termopanel Azotea Techada Lavandería Terraza 3 estacionamientos 4 bodega Sala de reuniones Espacio para lavadora Recepción Baño principal Encimera Horno Depósitos
Departamento de 2 pisos, 1 baño con salida a terraza en living comedor que comparten un gran jardín para una amplia vista panorámica al barrio del Barrio Los Barneches (principal) el edificio cuenta con lavandería completamente equipada por las ventanas termopanel/calefacción central-congreso techado; sala de reuniones + quincho amplio: logia cerrada!
Piso de madera en el primer piso, living comedor con salida a terraza que tiene una vista panorámica al centro del barrio Concepción (condominio). Calefacción por radiadores para la ventanas termopaneles 2 bodega 3 estacionamientos 4 estacionamientos 1 baño completo Living Comedor Cocina equipada Amplia sala de estar Ascensor: Sala de Estar Edificio cuenta con un gran jardín Techado La casa está construida como uno de los dos pisos principales separados El tercer dormitorio es amplio hall de entrada Terraza completamente iluminada Encimera Ventanas Termopanel Horno Derecho Terreno Total Construcción Construida Completamente Equipada! Gran Jardín Propiedad Se Vende esta propiedad consta de 5 habitaciones (2 baños) Hall de acceso Baño Principal Superficie total aproximada \$100.000
Piso de madera en el primer piso, living comedor con salida a terraza que tienen 2 baños completos (incluye una sala de estar) Living comedor amplio Condominio Amplio Hall de entrada Calefacción por radiadores Ventanas termopanel Termopanel Cocina amoblada Edificio cuenta con lavandería Lavadora Azotea Techada: 1 bodega
Excelente conectividad, cercana a la ciudad de Coquimbo con acceso directo al centro comercial del sector principal en el primer piso: 2 baños completos (incluye una sala de estar) Living comedor amplio que cuenta con salida a terraza completamente equipada para dos vehículos; 1 bodega techada por un gran jardín como estacionamiento privado se entregan los siguientes elementos: 3 estacionamientos + quincho
Piso de madera en el primer piso, living comedor con salida a terraza que tiene una vista panorámica al barrio del Barrio Los Barneches por la zona principal (con baño) Living comedor amplio para 2 vehículos (2 autos), dos estacionamientos (1 bodega). Cocina equipada completamente equipada como sala de reuniones 1 Baño completo Amplio Hall de entrada Consta un hall de entrada Calefacción central Azotea Techada Encimera Ventanas termopanel Termopanel Lavandería Jardín Edificio cuenta con: Sala de Estar Terraza Gran jardín 3 pisos 4 estacionamiento 5 bodegas

A.8. Ejemplos de 10 Registros Generados Aleatoriamente en Descripciones Económicas

Tabla A.2: Ejemplos de textos aleatorios del modelo tddpm_mlp, conjunto Economicos b-1

description
Departamento de dos dormitorios, 2 baños, living comedor, cocina amoblada con encimera, horno empotrado, campana, lavadero, estacionamiento techado para 1 vehículo
Departamento de 3 dormitorios, 2 baños, living comedor con salida a terraza, cocina amoblada encimera, horno empotrado, campana cerrada, logia techada, estacionamiento subterráneo
Departamento de 2 dormitorios, 1 baño, living comedor con salida a terraza en el primer piso (con vista al mar), cocina amoblada equipada con campana, horno empotrado, cubiertas de granito para lavadora, estacionamiento subterráneo techado por los paneles solares que permiten un acceso controlado las 24 horas del día
Vendo terreno en Rancagua, sector de la playa con accesos controlados por los vecinos
Cuenta con 3 dormitorios, 2 baños, living comedor, cocina amoblada, logia, estacionamiento techado para dos vehículos en el primer piso (incluye agua potable)
Departamento de 3 dormitorios, 2 baños, living comedor con salida a terraza en el primer piso (con vista despejada), cocina tradicional amoblada equipada con horno empotrado, campana cerrada para lavadora, logia techada que permite un estacionamiento subterráneo
Ñuñoa, 2 dormitorios, 1 baño, estacionamiento.
Excelente oficina, 3 privados, 2 baños, living comedor, cocina amoblada, logia, estacionamiento techado

A.9. Estadísticos KingCounty

Tabla A.3: Propiedades estadísticas de variable grade, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[7 8 9 6 10]	[7 8 9 6 10]	[7 8 9 6 10]	[7 8 6 9 10]
top5_freq	[7201 4879 2072 1620 915]	[9607 6226 2568 1761 991]	[9705 6246 2567 1768 921]	[7896 5461 2560 1586 1325]
top5_prob	[0.41648352 0.28218623 0.11983806 0.09369578 0.05292076]	[0.44450099 0.28806737 0.11881738 0.08147874 0.04585203]	[0.44901453 0.28897937 0.11876561 0.08179883 0.04261127]	[0.36533568 0.252672 0.11844723 0.07338176 0.0613057]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	7.654	7.643	7.615	7.645
std_err	0.009	0.007	0.007	0.012
upper_ci	7.671	7.657	7.629	7.667
lower_ci	7.636	7.628	7.601	7.622
std	1.170	1.079	1.049	1.711
iqr	1.000	1.000	1.000	1.000
iqr_normal	0.741	0.741	0.741	0.741
mad	0.926	0.864	0.847	1.252
mad_normal	1.160	1.082	1.061	1.569
coef_var	0.153	0.141	0.138	0.224
range	12.000	9.000	8.000	12.000
max	13.000	13.000	12.000	13.000
min	1.000	4.000	4.000	1.000
skew	0.758	0.911	0.814	0.604
kurtosis	4.209	4.316	4.024	4.528
jarque_bera	2709	4551	3331	3416
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	7.000	7.000	7.000	7.000
mode_freq	0.416	0.445	0.449	0.365
median	7.000	7.000	7.000	7.000
0.1 %	4.000	5.000	5.000	3.000
1.0 %	5.000	6.000	6.000	3.000
5.0 %	6.000	6.000	6.000	5.000
25.0 %	7.000	7.000	7.000	7.000
75.0 %	8.000	8.000	8.000	8.000
95.0 %	10.000	10.000	10.000	11.000
99.0 %	11.000	11.000	11.000	13.000
99.9 %	12.000	12.000	12.000	13.000

Tabla A.4: Propiedades estadísticas de variable view, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[0 2 3 1 4]	[0 2 3 4 1]	[0 2 3 4 1]	[0 2 4 3 1]
top5_freq	[15586 783 396 275 250]	[20422 590 315 160 126]	[20868 331 200 140 75]	[16772 1981 1266 812 782]
top5_prob	[0.90144592 0.04528629 0.02290341 0.01590515 0.01445922]	[0.94489428 0.02729839 0.01457456 0.00740295 0.00582982]	[0.96548533 0.01531415 0.00925326 0.00647728 0.00346997]	[0.77601444 0.0916578 0.05857586 0.03756998 0.03618193]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	0.233	0.134	0.088	0.567
std_err	0.006	0.004	0.003	0.008
upper_ci	0.244	0.142	0.094	0.582
lower_ci	0.222	0.126	0.081	0.551
std	0.762	0.589	0.494	1.165
iqr	0.000	0.000	0.000	0.000
iqr_normal	0.000	0.000	0.000	0.000
mad	0.420	0.253	0.169	0.879
mad_normal	0.527	0.317	0.212	1.102
coef_var	3.269	4.402	5.628	2.056
range	4.000	4.000	4.000	4.000
max	4.000	4.000	4.000	4.000
min	0.000	0.000	0.000	0.000
skew	3.402	4.653	6.021	1.925
kurtosis	13.971	24.722	40.167	5.389
jarque_bera	120072	502895	1374638	18491
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	0.000	0.000	0.000	0.000
mode_freq	0.901	0.945	0.965	0.776
median	0.000	0.000	0.000	0.000
0.1 %	0.000	0.000	0.000	0.000
1.0 %	0.000	0.000	0.000	0.000
5.0 %	0.000	0.000	0.000	0.000
25.0 %	0.000	0.000	0.000	0.000
75.0 %	0.000	0.000	0.000	0.000
95.0 %	2.000	1.000	0.000	4.000
99.0 %	4.000	3.000	3.000	4.000
99.9 %	4.000	4.000	4.000	4.000

Tabla A.5: Propiedades estadísticas de variable lat, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[47.5402 47.6875 47.686 47.6624 47.5518]	[47.7776 47.1559 47.16139285 47.65148565 47.65162378]	[47.5446 47.6265 47.6647 47.6955 47.6961]	[47.7776 47.6108 47.1593 47.618 47.6221]
top5_freq	[14 13 13 13 13]	[18 3 1 1 1]	[8 5 5 4 4]	[12 12 20 19 18 18]
top5_prob	[0.00080972 0.00075188 0.00075188 0.00075188 0.00075188]	[8.32832092e-04 1.38805349e-04 4.62684495e-05 4.62684495e-05 4.62684495e-05]	[0.00037013 0.00023133 0.00023133 0.00018507 0.00018507]	[0.05607736 0.00092537 0.0008791 0.00083283 0.00083283]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	47.560	47.562	47.561	47.594
std_err	0.001	0.001	0.001	0.001
upper_ci	47.562	47.564	47.563	47.596
lower_ci	47.558	47.560	47.559	47.592
std	0.138	0.137	0.137	0.124
iqr	0.206	0.205	0.203	0.174
iqr_normal	0.153	0.152	0.150	0.129
mad	0.115	0.114	0.113	0.100
mad_normal	0.144	0.143	0.142	0.125
coef_var	0.003	0.003	0.003	0.003
range	0.618	0.622	0.605	0.618
max	47.778	47.778	47.777	47.778
min	47.159	47.156	47.171	47.159
skew	-0.487	-0.473	-0.494	-0.575
kurtosis	2.328	2.266	2.361	2.917
jarque_bera	1009	1290	1247	1197
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	47.540	47.778	47.545	47.778
mode_freq	0.001	0.001	0.000	0.056
median	47.572	47.571	47.572	47.605
0.1 %	47.193	47.189	47.193	47.170
1.0 %	47.257	47.262	47.250	47.267
5.0 %	47.311	47.312	47.313	47.367
25.0 %	47.472	47.475	47.475	47.520
75.0 %	47.678	47.681	47.678	47.693
95.0 %	47.750	47.748	47.749	47.778
99.0 %	47.773	47.772	47.771	47.778
99.9 %	47.777	47.778	47.776	47.778

Tabla A.6: Propiedades estadísticas de variable zipcode, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[98103 98038 98115 98052 98117]	[98052 98115 98103 98034 98038]	[98038 98103 98052 98042 98115]	[98006 98118 98103 98040 98034]
top5_freq	[489 473 462 459 455]	[662 649 623 620 619]	[619 605 589 561 559]	[885 778 666 589 572]
top5_prob	[0.02828224 0.02735685 0.02672065 0.02654714 0.02631579]	[0.03062971 0.03002822 0.02882524 0.02868644 0.02864017]	[0.02863885 0.02799112 0.02725086 0.0259554 0.02586287]	[0.04094758 0.03599685 0.03081479 0.02725212 0.02646555]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	98078	98078	98078	98081
std_err	0.406	0.363	0.361	0.381
upper_ci	98079	98079	98079	98082
lower_ci	98077	98077	98077	98080
std	53.326	53.361	53.037	56.072
iqr	84.000	84.000	84.000	86.000
iqr_normal	62.269	62.269	62.269	63.752
mad	46.554	46.714	46.383	49.470
mad_normal	58.347	58.547	58.132	62.002
coef_var	0.001	0.001	0.001	0.001
range	198.000	198.000	198.000	198.000
max	98199	98199	98199	98199
min	98001	98001	98001	98001
skew	0.402	0.399	0.392	0.371
kurtosis	2.153	2.140	2.136	2.025
jarque_bera	983.027	1238.956	1226.468	1350.558
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	98103	98052	98038	98006
mode_freq	0.028	0.031	0.029	0.041
median	98065	98065	98065	98074
0.1 %	98001	98001	98001	98001
1.0 %	98001	98001	98001	98002
5.0 %	98004	98004	98004	98006
25.0 %	98033	98033	98033	98033
75.0 %	98117	98117	98117	98119
95.0 %	98177	98177	98177	98188
99.0 %	98199	98199	98199	98199
99.9 %	98199	98199	98199	98199

Tabla A.7: Propiedades estadísticas de variable sqft_basement, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[0 600 700 500 800]	[0. 500. 600. 700. 800.]	[0. 600. 400. 380. 200.]	[0 6 8 7 10]
top5_freq	[10553 182 169 167 164]	[13152 228 220 204 181]	[11754 16 12 11 9]	[4948 817 816 789 759]
top5_prob	[0.61035281 0.01052632 0.00977444 0.00965876 0.00948525]	[0.60852265 0.01054921 0.01017906 0.00943876 0.00837459]	[5.43814195e-01 7.40260942e-04 5.55195706e-04 5.08929398e-04 4.16396780e-04]	[0.22893629 0.03780132 0.03775505 0.03650581 0.03511775]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	287.933	312.217	271.488	248.476
std_err	3.337	3.470	2.753	3.448
upper_ci	294.472	319.019	276.884	255.234
lower_ci	281.393	305.416	266.091	241.717
std	438.727	510.175	404.769	506.927
iqr	550.000	600.000	503.482	17.000
iqr_normal	407.716	444.781	373.232	12.602
mad	360.277	385.339	330.089	367.559
mad_normal	451.541	482.950	413.705	460.667
coef_var	1.524	1.634	1.491	2.040
range	4820	4820	3456	3194
max	4820	4820	3456	3194
min	0.000	0.000	0.000	0.000
skew	1.571	3.287	1.525	2.300
kurtosis	5.639	24.533	5.001	8.137
jarque_bera	12126	456478	11979	42810
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	0.000	0.000	0.000	0.000
mode_freq	0.610	0.609	0.544	0.229
median	0.000	0.000	0.000	9.000
0.1 %	0.000	0.000	0.000	0.000
1.0 %	0.000	0.000	0.000	0.000
5.0 %	0.000	0.000	0.000	0.000
25.0 %	0.000	0.000	0.000	1.000
75.0 %	550.000	600.000	503.482	18.000
95.0 %	1180	1148	1096	1393
99.0 %	1650	1702	1514	2170
99.9 %	2324	4820	2078	2914

Tabla A.8: Propiedades estadísticas de variable price, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[350000. 450000. 425000. 550000. 325000.]	[350000. 500000. 300000. 450000. 425000.]	[550000. 350000. 450000. 325000. 250000.]	[75000. 441284. 373224. 155731. 552865.]
top5_freq	[143 140 123 123 123]	[144 141 140 140 140]	[213 205 182 182 173]	[523 3 2 2 2]
top5_prob	[0.00827068 0.00809717 0.00711394 0.00711394 0.00711394]	[0.00666266 0.00652385 0.00647758 0.00647758 0.00647758]	[0.00985472 0.00948459 0.00842047 0.00842047 0.00800407]	[2.41983991e-02 1.38805349e-04 9.25368991e-05 9.25368991e-05 9.25368991e-05]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	537768	540676	519947	479590
std_err	2749	3439	2114	2796
upper_ci	543156	547416	524090	485070
lower_ci	532380	533936	515804	474110
std	361464	505561	310773	411063
iqr	319850	305000	310000	340312
iqr_normal	237105	226097	229803	252274
mad	231680	235939	211646	264694
mad_normal	290368	295706	265259	331745
coef_var	0.672	0.935	0.598	0.857
range	7625000	7624938	4127000	4329847
max	7700000	7700000	4208000	4404847
min	75000	75062	81000	75000
skew	4.032	9.444	2.710	3.070
kurtosis	39.678	126.853	15.954	17.323
jarque_bera	1016020	14135158	177569	218683
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	350000	350000	550000	75000
mode_freq	0.008	0.007	0.010	0.024
median	450000	448898	450500	376938
0.1 %	95000	95123	95000	75000
1.0 %	154467	163882	153000	75000
5.0 %	210000	215000	210000	100318
25.0 %	320150	320000	317000	235649
75.0 %	640000	625000	627000	575961
95.0 %	1150000	1050858	1039000	1234294
99.0 %	1950000	1894340	1770000	2237925
99.9 %	3331995	7700000	2750000	3641937

Tabla A.9: Propiedades estadísticas de variable sqft_above, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1300 1010 1200 1220 1140]	[1300. 1010. 1220. 1340. 1060.]	[1800. 1830. 770. 1200. 1160.]	[1553 1446 1533 1479 1384]
top5_freq	[166 165 160 152 148]	[196 178 176 173 170]	[15 13 12 12 11]	[25 25 25 24 23]
top5_prob	[0.00960093 0.00954309 0.0092539 0.00879121 0.00855986]	[0.00906862 0.00823578 0.00814325 0.00800444 0.00786564]	[0.00069399 0.00060146 0.0005552 0.0005552 0.00050893]	[0.00115671 0.00115671 0.00115671 0.00111044 0.00106417]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	1786	1766	1771	2006
std_err	6.249	5.870	5.222	7.523
upper_ci	1798	1778	1781	2021
lower_ci	1774	1755	1760	1991
std	821.626	862.982	767.744	1105.973
iqr	1000.000	958.755	961.973	1127.000
iqr_normal	741.301	710.726	713.112	835.446
mad	635.012	616.093	600.877	820.979
mad_normal	795.870	772.158	753.087	1028.945
coef_var	0.460	0.489	0.434	0.551
range	8570	9120	7429	7255
max	8860	9410	7819	7545
min	290.000	290.229	389.730	290.000
skew	1.428	2.895	1.273	1.577
kurtosis	6.260	21.428	4.922	5.740
jarque_bera	13530	336010	9168	15721
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1300	1300	1800	1446
mode_freq	0.010	0.009	0.001	0.001
median	1560	1540	1554	1691
0.1 %	520.000	534.834	609.393	371.224
1.0 %	700.000	718.132	739.478	559.120
5.0 %	850.000	868.899	887.703	814.000
25.0 %	1200	1200	1211	1275
75.0 %	2200	2159	2173	2402
95.0 %	3380	3228	3296	4282
99.0 %	4371	4263	4147	5802
99.9 %	6070	9410	5082	7015

Tabla A.10: Propiedades estadísticas de variable date, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['20140623T000000' '20140625T000000' '20140626T000000' '20150421T000000' '20150325T000000']	['20140623T000000' '20150325T000000' '20150421T000000' '20140625T000000' '20140825T000000']	['20140625T000000' '20140623T000000' '20140825T000000' '20150422T000000' '20150427T000000']	['20150310T000000' '20150327T000000' '20140603T000000' '20150226T000000' '20150329T000000']
top5_freq	[123 105 101 101 101]	[179 171 160 150 149]	[169 166 135 133 133]	[471 406 372 321 286]
top5_prob	[0.00711394 0.00607287 0.00584153 0.00584153 0.00584153]	[0.00828205 0.0079119 0.00740295 0.00694027 0.006894]	[0.00781901 0.00768021 0.00624595 0.00615342 0.00615342]	[0.02179244 0.01878499 0.01721186 0.01485217 0.01323278]
nobs	17290	21613	21614	21613
missing	17290	0	0	0

Tabla A.11: Propiedades estadísticas de variable bathrooms, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[2.5 1. 1.75 2.25 2.]	[2.5 1. 1.75 2.25 2.]	[2.5 1. 1.75 2.25 2.]	[2.5 1. 1.75 2.25 2.]
top5_freq	[4333 3088 2425 1621 1526]	[6052 3979 3245 2158 1698]	[6736 4809 3275 1934 1345]	[5486 4029 2410 1477 1270]
top5_prob	[0.25060729 0.17860035 0.14025448 0.09375361 0.08825911]	[0.28001666 0.18410216 0.15014112 0.09984731 0.07856383]	[0.31164986 0.22249468 0.15152216 0.08947904 0.06222819]	[0.25382871 0.18641558 0.11150696 0.0683385 0.05876093]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	2.114	2.084	2.018	2.233
std_err	0.006	0.005	0.005	0.006
upper_ci	2.125	2.093	2.028	2.245
lower_ci	2.102	2.074	2.009	2.220
std	0.767	0.721	0.710	0.930
iqr	1.000	0.750	1.000	1.250
iqr_normal	0.741	0.556	0.741	0.927
mad	0.615	0.587	0.596	0.733
mad_normal	0.771	0.736	0.746	0.919
coef_var	0.363	0.346	0.352	0.417
range	8.000	7.250	4.750	8.000
max	8.000	7.750	5.500	8.000
min	0.000	0.500	0.750	0.000
skew	0.464	0.316	0.128	0.441
kurtosis	3.989	3.852	2.742	3.687
jarque_bera	1326	1013	119	1126
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	2.500	2.500	2.500	2.500
mode_freq	0.251	0.280	0.312	0.254
median	2.250	2.250	2.250	2.500
0.1 %	0.750	1.000	1.000	0.000
1.0 %	1.000	1.000	1.000	0.750
5.0 %	1.000	1.000	1.000	1.000
25.0 %	1.500	1.750	1.500	1.500
75.0 %	2.500	2.500	2.500	2.750
95.0 %	3.500	3.250	3.250	4.000
99.0 %	4.250	3.750	3.500	4.750
99.9 %	5.428	5.000	4.500	6.000

Tabla A.12: Propiedades estadísticas de variable long, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[-122.29 -122.362 -122.288 -122.3 - 122.372]	[-122.3 -122.29 -122.189 -122.304 -122.307]	[-122.284 -122.351 -122.017 -122.387 - 122.29]	[-122.331 -122.337 -122.347 -122.324 - 122.319]
top5_freq	[100 88 81 81 81]	[113 101 88 87 84]	[26 22 21 20 18]	[130 120 110 108 107]
top5_prob	[0.00578369 0.00508965 0.00468479 0.00468479 0.00468479]	[0.00522833 0.00467311 0.00407162 0.00402536 0.00388655]	[0.00120292 0.00101786 0.00097159 0.00092533 0.00083279]	[0.0060149 0.00555221 0.00508953 0.00499699 0.00495072]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	-122.214	-122.218	-122.213	-122.215
std_err	0.001	0.001	0.001	0.001
upper_ci	-122.212	-122.216	-122.211	-122.213
lower_ci	-122.216	-122.220	-122.215	-122.218
std	0.140	0.135	0.139	0.161
iqr	0.204	0.194	0.201	0.217
iqr_normal	0.151	0.144	0.149	0.161
mad	0.115	0.111	0.114	0.130
mad_normal	0.144	0.139	0.143	0.163
coef_var	-0.001	-0.001	-0.001	-0.001
range	1.204	1.203	1.187	0.941
max	-121.315	-121.316	-121.322	-121.578
min	-122.519	-122.519	-122.509	-122.519
skew	0.867	0.790	0.788	1.113
kurtosis	3.953	3.648	3.486	4.064
jarque_bera	2819	2626	2450	5485
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	-122.290	-122.300	-122.284	-122.331
mode_freq	0.006	0.005	0.001	0.006
median	-122.231	-122.228	-122.226	-122.262
0.1 %	-122.497	-122.519	-122.476	-122.519
1.0 %	-122.408	-122.405	-122.405	-122.426
5.0 %	-122.387	-122.386	-122.386	-122.398
25.0 %	-122.329	-122.326	-122.325	-122.339
75.0 %	-122.125	-122.132	-122.125	-122.122
95.0 %	-121.979	-121.999	-121.981	-121.922
99.0 %	-121.787	-121.810	-121.801	-121.702
99.9 %	-121.699	-121.721	-121.714	-121.609

Tabla A.13: Propiedades estadísticas de variable sqft_lot, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[5000 4000 6000 7200 4800]	[5000. 4000. 6000. 7200. 7500.]	[5000. 4000. 3600. 6000. 7200.]	[520 4312 642 1105 4267]
top5_freq	[301 209 208 179 98]	[294 268 218 201 109]	[57 39 32 20 11]	[5159 7 7 6 6]
top5_prob	[0.01740891 0.01208791 0.01203008 0.01035281 0.00566802]	[0.01360292 0.01239994 0.01008652 0.00929996 0.00504326]	[0.00263718 0.00180439 0.00148052 0.00092533 0.00050893]	[0.23869893 0.00032388 0.00032388 0.00027761 0.00027761]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	14799	16438	13507	9579
std_err	295.375	574.001	206.847	164.968
upper_ci	15378	17563	13912	9903
lower_ci	14220	15313	13101	9256
std	38839	84386	30410	24253
iqr	5606	5038	5475	8679
iqr_normal	4155	3735	4059	6434
mad	13382	16484	11151	9367
mad_normal	16772	20660	13975	11740
coef_var	2.624	5.134	2.252	2.532
range	1164274	1650829	958936	324058
max	1164794	1651359	959567	324578
min	520.000	529.965	630.944	520.000
skew	11.588	17.333	10.344	7.011
kurtosis	215.591	326.519	178.841	60.174
jarque_bera	32946220	95337170	28231599	3120785
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	5000	5000	5000	520
mode_freq	0.017	0.014	0.003	0.239
median	7600	7618	7666	4602
0.1 %	737.156	818.002	834.632	520.000
1.0 %	1005	1056	1082	520
5.0 %	1756	1988	1813	520
25.0 %	5001	5097	5109	704
75.0 %	10607	10135	10584	9383
95.0 %	42999	35585	39715	25720
99.0 %	212192	191215	168317	152365
99.9 %	435600	1651359	362801	268851

Tabla A.14: Propiedades estadísticas de variable bedrooms, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[3 4 2 5 6]	[3 4 2 5 6]	[3 4 2 5 1]	[3 4 2 5 1]
top5_freq	[7865 5477 2237 1292 212]	[10764 7042 2416 1223 89]	[11340 7133 2395 682 42]	[7651 7589 2472 1742 451]
top5_prob	[0.45488722 0.3167727 0.12938115 0.07472527 0.01226142]	[0.49803359 0.32582242 0.11178457 0.05658631 0.00411789]	[0.52465994 0.33001758 0.11080781 0.03155362 0.00194318]	[0.35399991 0.35113126 0.11437561 0.08059964 0.02086707]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	3.368	3.335	3.282	3.975
std_err	0.007	0.005	0.005	0.025
upper_ci	3.382	3.346	3.291	4.025
lower_ci	3.354	3.324	3.272	3.926
std	0.931	0.805	0.712	3.717
iqr	1.000	1.000	1.000	1.000
iqr_normal	0.741	0.741	0.741	0.741
mad	0.734	0.648	0.589	1.354
mad_normal	0.920	0.812	0.738	1.697
coef_var	0.277	0.241	0.217	0.935
range	33.000	33.000	8.000	33.000
max	33.000	33.000	9.000	33.000
min	0.000	0.000	1.000	0.000
skew	2.304	2.542	0.139	6.765
kurtosis	63.268	88.155	3.373	52.903
jarque_bera	2631992	6553502	195	2407490
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	3.000	3.000	3.000	3.000
mode_freq	0.455	0.498	0.525	0.354
median	3.000	3.000	3.000	3.000
0.1 %	1.000	1.000	1.000	0.000
1.0 %	2.000	2.000	2.000	0.000
5.0 %	2.000	2.000	2.000	2.000
25.0 %	3.000	3.000	3.000	3.000
75.0 %	4.000	4.000	4.000	4.000
95.0 %	5.000	5.000	4.000	6.000
99.0 %	6.000	5.000	5.000	33.000
99.9 %	7.000	6.000	5.387	33.000

Tabla A.15: Propiedades estadísticas de variable sqft_living15, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1440 1540 1560 1500 1610]	[1440. 1520. 1500. 1550. 1640.]	[1830. 1310. 1440. 1160. 2370.]	[1738 1706 1818 2079 1893]
top5_freq	[156 154 152 137 136]	[210 165 159 158 156]	[18 16 16 16 15]	[28 28 26 25 24]
top5_prob	[0.00902256 0.00890688 0.00879121 0.00792366 0.00786582]	[0.00971637 0.00763429 0.00735668 0.00731042 0.00721788]	[0.00083279 0.00074026 0.00074026 0.00074026 0.00069399]	[0.00129552 0.00129552 0.00120298 0.00115671 0.00111044]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	1983	1953	1979	1759
std_err	5.181	4.370	4.397	4.540
upper_ci	1993	1961	1987	1768
lower_ci	1973	1944	1970	1750
std	681.232	642.407	646.410	667.478
iqr	880.000	810.000	839.395	835.000
iqr_normal	652.345	600.454	622.244	618.986
mad	533.237	500.663	509.027	510.410
mad_normal	668.313	627.488	637.971	639.704
coef_var	0.344	0.329	0.327	0.379
range	5811	5810	4652	5811
max	6210	6209	5279	6210
min	399.000	399.000	627.292	399.000
skew	1.095	1.104	1.041	1.033
kurtosis	4.572	4.942	4.196	5.310
jarque_bera	5237	7785	5190	8650
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1440	1440	1830	1706
mode_freq	0.009	0.010	0.001	0.001
median	1840	1820	1840	1695
0.1 %	740.000	399.000	763.863	440.792
1.0 %	950.000	958.640	985.793	640.000
5.0 %	1140	1160	1177	839
25.0 %	1480	1490	1501	1276
75.0 %	2360	2300	2341	2111
95.0 %	3280	3182	3236	2946
99.0 %	4050	3890	3909	3865
99.9 %	4986	4913	4774	5141

Tabla A.16: Propiedades estadísticas de variable waterfront, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[0 1]	[0 1]	[0 1]	[0 1]
top5_freq	[17166 124]	[21553 60]	[21586 28]	[19823 1790]
top5_prob	[0.99282822 0.00717178]	[0.99722389 0.00277611]	[0.99870454 0.00129546]	[0.91717948 0.08282052]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	0.007	0.003	0.001	0.083
std_err	0.001	0.000	0.000	0.002
upper_ci	0.008	0.003	0.002	0.086
lower_ci	0.006	0.002	0.001	0.079
std	0.084	0.053	0.036	0.276
iqr	0.000	0.000	0.000	0.000
iqr_normal	0.000	0.000	0.000	0.000
mad	0.014	0.006	0.003	0.152
mad_normal	0.018	0.007	0.003	0.190
coef_var	11.766	18.953	27.766	3.328
range	1.000	1.000	1.000	1.000
max	1.000	1.000	1.000	1.000
min	0.000	0.000	0.000	0.000
skew	11.681	18.900	27.730	3.027
kurtosis	137.443	358.219	769.930	10.165
jarque_bera	13414600	114917884	532476328	79239
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	0.000	0.000	0.000	0.000
mode_freq	0.993	0.997	0.999	0.917
median	0.000	0.000	0.000	0.000
0.1 %	0.000	0.000	0.000	0.000
1.0 %	0.000	0.000	0.000	0.000
5.0 %	0.000	0.000	0.000	0.000
25.0 %	0.000	0.000	0.000	0.000
75.0 %	0.000	0.000	0.000	0.000
95.0 %	0.000	0.000	0.000	1.000
99.0 %	0.000	0.000	0.000	1.000
99.9 %	1.000	1.000	1.000	1.000

Tabla A.17: Propiedades estadísticas de variable sqft_living, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1400 1300 1720 1250 1540]	[1800. 1300. 1440. 1580. 2240.]	[1800. 1830. 2230. 1690. 770.]	[290 1671 1460 1822 1618]
top5_freq	[109 107 106 106 105]	[134 122 119 108 105]	[15 15 13 12 12]	[95 24 21 20 20]
top5_prob	[0.00630422 0.00618855 0.00613071 0.00613071 0.00607287]	[0.00619997 0.00564475 0.00550595 0.00499699 0.00485819]	[0.00069399 0.00069399 0.00060146 0.0005552 0.0005552]	[0.0043955 0.00111044 0.00097164 0.00092537 0.00092537]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	2074	2079	2042	1832
std_err	6.900	7.340	5.682	5.975
upper_ci	2087	2093	2053	1844
lower_ci	2060	2064	2031	1820
std	907.298	1079.062	835.391	878.356
iqr	1110	1065	1057	1033
iqr_normal	822.844	789.339	783.883	765.764
mad	693.180	690.927	649.181	657.958
mad_normal	868.773	865.949	813.627	824.628
coef_var	0.437	0.519	0.409	0.479
range	11760	13250	10060	9364
max	12050	13540	10450	9654
min	290.000	290.306	389.730	290.000
skew	1.371	4.654	1.155	1.287
kurtosis	7.167	45.621	5.605	6.644
jarque_bera	17922	1713859	10916	17918
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1400	1800	1800	290
mode_freq	0.006	0.006	0.001	0.004
median	1910	1900	1894	1713
0.1 %	522.890	569.633	620.406	290.000
1.0 %	720.000	739.158	770.000	365.000
5.0 %	940.000	960.000	967.472	638.000
25.0 %	1430	1430	1433	1236
75.0 %	2540	2495	2491	2269
95.0 %	3740	3617	3604	3423
99.0 %	4921	4949	4525	4754
99.9 %	6966	13540	6095	6582

Tabla A.18: Propiedades estadísticas de variable condition, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[3 4 5 2 1]	[3 4 5 2 1]	[3 4 5 2 1]	[3 4 5 2 1]
top5_freq	[11248 4512 1364 139 27]	[14771 5499 1282 54 7]	[15485 5206 896 25 2]	[15887 3813 1366 297 250]
top5_prob	[0.65054945 0.26096009 0.07888953 0.00803933 0.0015616]	[6.83431268e-01 2.54430204e-01 5.93161523e-02 2.49849628e-03 3.23879147e-04]	[7.16433793e-01 2.40862404e-01 4.14546128e-02 1.15665772e-03 9.25326177e-05]	[0.73506686 0.1764216 0.0632027 0.01374173 0.01156711]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	3.408	3.370	3.322	3.266
std_err	0.005	0.004	0.004	0.004
upper_ci	3.417	3.378	3.330	3.275
lower_ci	3.398	3.362	3.315	3.257
std	0.652	0.599	0.552	0.647
iqr	1.000	1.000	1.000	0.000
iqr_normal	0.741	0.741	0.741	0.000
mad	0.560	0.514	0.465	0.478
mad_normal	0.702	0.644	0.583	0.599
coef_var	0.191	0.178	0.166	0.198
range	4.000	4.000	4.000	4.000
max	5.000	5.000	5.000	5.000
min	1.000	1.000	1.000	1.000
skew	1.028	1.281	1.447	0.824
kurtosis	3.556	3.815	4.246	5.519
jarque_bera	3269	6513	8939	8157
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	3.000	3.000	3.000	3.000
mode_freq	0.651	0.683	0.716	0.735
median	3.000	3.000	3.000	3.000
0.1 %	1.000	2.000	2.000	1.000
1.0 %	3.000	3.000	3.000	1.000
5.0 %	3.000	3.000	3.000	3.000
25.0 %	3.000	3.000	3.000	3.000
75.0 %	4.000	4.000	4.000	3.000
95.0 %	5.000	5.000	4.000	5.000
99.0 %	5.000	5.000	5.000	5.000
99.9 %	5.000	5.000	5.000	5.000

Tabla A.19: Propiedades estadísticas de variable yr_built, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[2014 2005 2006 2004 2007]	[2006. 2005. 2004. 2008. 2007.]	[2014. 2006. 2004. 2005. 2003.]	[2015 1978 1980 1977 1975]
top5_freq	[449 371 366 350 347]	[511 484 426 406 404]	[270 109 104 103 102]	[1581 325 324 316 314]
top5_prob	[0.02596877 0.02145749 0.02116831 0.02024291 0.0200694]	[0.02364318 0.02239393 0.01971036 0.01878499 0.01869245]	[0.0124919 0.00504303 0.0048117 0.00476543 0.00471916]	[0.07315042 0.01503725 0.01499098 0.01462083 0.01452829]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	1971	1972	1971	1973
std_err	0.224	0.191	0.198	0.203
upper_ci	1972	1973	1972	1974
lower_ci	1971	1972	1971	1973
std	29.436	28.054	29.073	29.823
iqr	46.000	44.000	45.388	45.000
iqr_normal	34.100	32.617	33.646	33.359
mad	24.632	23.315	24.378	24.488
mad_normal	30.872	29.221	30.553	30.691
coef_var	0.015	0.014	0.015	0.015
range	115.000	115.000	115.000	115.000
max	2015	2015	2015	2015
min	1900	1900	1900	1900
skew	-0.472	-0.477	-0.465	-0.446
kurtosis	2.337	2.432	2.310	2.412
jarque_bera	957.631	1108.611	1206.759	1027.161
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	2014	2006	2014	2015
mode_freq	0.026	0.024	0.012	0.073
median	1975	1974	1975	1975
0.1 %	1900	1900	1901	1900
1.0 %	1904	1906	1906	1902
5.0 %	1915	1919	1916	1918
25.0 %	1951	1954	1952	1954
75.0 %	1997	1998	1997	1999
95.0 %	2011	2009	2010	2015
99.0 %	2014	2014	2014	2015
99.9 %	2015	2015	2014	2015

Tabla A.20: Propiedades estadísticas de variable sqft_lot15, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[5000 4000 6000 7200 4800]	[5000. 4000. 6000. 7200. 4800.]	[5000. 4000. 3600. 6000. 5200.]	[651 3104 6478 4042 7791]
top5_freq	[349 289 224 160 120]	[392 351 265 172 142]	[71 71 32 28 24]	[1594 7 7 7 7]
top5_prob	[0.02018508 0.01671486 0.01295547 0.0092539 0.00694043]	[0.01813723 0.01624023 0.01226114 0.00795817 0.00657012]	[0.00328491 0.00328491 0.00148052 0.00129546 0.00111039]	[0.07375191 0.00032388 0.00032388 0.00032388 0.00032388]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	12725	11502	11986	11668
std_err	209.331	172.030	159.853	148.745
upper_ci	13135	11839	12300	11960
lower_ci	12315	11165	11673	11377
std	27525	25291	23501	21868
iqr	4963	4701	4846	7302
iqr_normal	3679	3485	3592	5413
mad	10095	8064	8799	9375
mad_normal	12652	10106	11028	11749
coef_var	2.163	2.199	1.961	1.874
range	870549	870549	866657	317306
max	871200	871200	867405	317957
min	651.000	651.000	748.259	651.000
skew	9.701	14.230	10.620	6.711
kurtosis	163.253	329.215	218.687	59.802
jarque_bera	18772189	96561454	42302376	3067794
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	5000	5000	4000	651
mode_freq	0.020	0.018	0.003	0.074
median	7615	7672	7673	7104
0.1 %	886.289	651.000	929.527	651.000
1.0 %	1189	1225	1207	651
5.0 %	1965	2416	2213	651
25.0 %	5083	5100	5164	3787
75.0 %	10046	9801	10010	11089
95.0 %	36822	33447	35670	36797
99.0 %	168296	106557	123646	120278
99.9 %	306998	257142	246150	251946

Tabla A.21: Propiedades estadísticas de variable floors, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1. 2. 1.5 3. 2.5]	[1. 2. 1.5 3. 2.5]	[1. 2. 1.5 3. 2.5]	[1. 2. 1.5 3. 2.5]
top5_freq	[8488 6628 1523 517 128]	[11240 8399 1438 489 47]	[11137 8525 1332 581 39]	[12466 6155 1835 531 467]
top5_prob	[0.49091961 0.38334297 0.0880856 0.02990168 0.00740312]	[0.52005737 0.38860871 0.06653403 0.02262527 0.00217462]	[0.51526788 0.39442028 0.06162672 0.02688073 0.00180439]	[0.57678249 0.28478231 0.0849026 0.02456855 0.02160737]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	1.499	1.470	1.482	1.427
std_err	0.004	0.004	0.004	0.004
upper_ci	1.507	1.477	1.489	1.435
lower_ci	1.491	1.463	1.475	1.420
std	0.543	0.529	0.538	0.563
iqr	1.000	1.000	1.000	1.000
iqr_normal	0.741	0.741	0.741	0.741
mad	0.490	0.489	0.496	0.493
mad_normal	0.614	0.613	0.622	0.618
coef_var	0.362	0.359	0.363	0.394
range	2.500	2.000	2.000	2.500
max	3.500	3.000	3.000	3.500
min	1.000	1.000	1.000	1.000
skew	0.615	0.589	0.599	1.103
kurtosis	2.526	2.336	2.400	3.643
jarque_bera	1252	1646	1618	4758
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1.000	1.000	1.000	1.000
mode_freq	0.491	0.520	0.515	0.577
median	1.500	1.000	1.000	1.000
0.1%	1.000	1.000	1.000	1.000
1.0%	1.000	1.000	1.000	1.000
5.0%	1.000	1.000	1.000	1.000
25.0%	1.000	1.000	1.000	1.000
75.0%	2.000	2.000	2.000	2.000
95.0%	2.000	2.000	2.000	2.500
99.0%	3.000	3.000	3.000	3.000
99.9%	3.000	3.000	3.000	3.500

Tabla A.22: Propiedades estadísticas de variable yr_renovated, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[0 2014 2005 2000 2003]	[0. 2014. 2015. 1955.66277505 2006.44440938]	[0. 2014. 2005. 2006. 2000.]	[0 1 2 3 4]
top5_freq	[16571 76 32 30 29]	[20904 43 5 1 1]	[20762 18 5 3 3]	[15888 1082 893 782 614]
top5_prob	[0.95841527 0.0043956 0.00185078 0.00173511 0.00167727]	[9.67195669e-01 1.98954333e-03 2.31342248e-04 4.62684495e-05 4.62684495e-05]	[9.60581105e-01 8.32793560e-04 2.31331544e-04 1.38798927e-04 1.38798927e-04]	[0.73511313 0.05006246 0.04131773 0.03618193 0.02840883]
nobs	17290	21613	21614	21613
missing	0.000	0.000	0.000	0.000
mean	83.003	65.610	72.998	76.292
std_err	3.031	2.424	2.514	2.536
upper_ci	88.943	70.361	77.926	81.263
lower_ci	77.063	60.859	68.070	71.321
std	398.503	356.361	369.663	372.863
iqr	0.000	0.000	0.000	1.000
iqr_normal	0.000	0.000	0.000	0.741
mad	159.103	126.915	140.273	144.963
mad_normal	199.407	159.064	175.806	181.684
coef_var	4.801	5.432	5.064	4.887
range	2015	2015	2015	2015
max	2015	2015	2015	2015
min	0.000	0.000	0.000	0.000
skew	4.593	5.249	4.937	4.749
kurtosis	22.096	28.552	25.530	23.603
jarque_bera	323506	687184	544922	463512
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	0.000	0.000	0.000	0.000
mode_freq	0.958	0.967	0.961	0.735
median	0.000	0.000	0.000	0.000
0.1 %	0.000	0.000	0.000	0.000
1.0 %	0.000	0.000	0.000	0.000
5.0 %	0.000	0.000	0.000	0.000
25.0 %	0.000	0.000	0.000	0.000
75.0 %	0.000	0.000	0.000	1.000
95.0 %	0.000	0.000	0.000	8.000
99.0 %	2008	2008	2005	2002
99.9 %	2014	2014	2014	2015

A.10. Estadísticos Economicos - Conjunto A

Tabla A.23: Propiedades estadísticas de variable bathrooms, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[2. 1. 3. 4. 5.]	[2. 1. 3. 4. 5.]	[2. 1. 3. 4. 5.]	[3. 2. 1. 5. 4.]
top5_freq	[7511 5440 4486 2665 1084]	[9658 6711 5628 3439 1334]	[9405 6758 5627 3356 1316]	[9608 4916 3986 3475 3097]
top5_prob	[0.34049594 0.24661136 0.20336371 0.12081237 0.04914094]	[0.35025749 0.24338145 0.20410532 0.12471894 0.04837891]	[0.34108218 0.24508595 0.20406905 0.12170886 0.04772612]	[0.34844419 0.17828389 0.14455647 0.12602452 0.11231595]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	2.604	2.512	2.588	3.620
std_err	0.025	0.011	0.012	0.046
upper_ci	2.652	2.533	2.612	3.711
lower_ci	2.556	2.491	2.564	3.529
std	3.655	1.792	2.039	7.719
iqr	1.000	1.000	1.000	2.000
iqr_normal	0.741	0.741	0.741	1.483
mad	1.203	1.095	1.180	1.767
mad_normal	1.507	1.372	1.479	2.214
coef_var	1.404	0.714	0.788	2.132
range	435.000	179.000	179.000	435.000
max	436.000	180.000	180.000	436.000
min	1.000	1.000	1.000	1.000
skew	82.448	36.498	27.293	42.809
kurtosis	9252	3497	2130	2239
jarque_bera	7.86518e+10	1.40304e+10	5.20223e+09	5.75190e+09
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	2.000	2.000	2.000	3.000
mode_freq	0.340	0.350	0.341	0.348
median	2.000	2.000	2.000	3.000
0.1%	1.000	1.000	1.000	1.000
1.0%	1.000	1.000	1.000	1.000
5.0%	1.000	1.000	1.000	1.000
25.0%	2.000	2.000	2.000	2.000
75.0%	3.000	3.000	3.000	4.000
95.0%	5.000	5.000	5.000	7.000
99.0%	8.000	7.000	8.000	14.000
99.9%	17.942	13.000	17.427	42.000

Tabla A.24: Propiedades estadísticas de variable property_type, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Departamento' 'Casa' 'Oficina o Casa Oficina' 'Parcela o Chacra' 'Local o Casa comercial']	['Departamento' 'Casa' 'Oficina o Casa Oficina' 'Parcela o Chacra' 'Local o Casa comercial']	['Departamento' 'Casa' 'Oficina o Casa Oficina' 'Parcela o Chacra' 'Departamento Amoblado']	['Departamento' 'Casa' 'Oficina o Casa Oficina' 'Parcela o Chacra' 'Local o Casa comercial']
top5_freq	[10592 8911 1553 413 255]	[13470 11435 1892 376 154]	[13613 11703 1796 237 121]	[10187 10078 3058 2086 1216]
top5_prob	[0.48016683 0.4039621 0.0704021 0.01872252 0.01155991]	[0.48850366 0.41470226 0.06861536 0.01363603 0.00558497]	[0.49368971 0.42442156 0.06513382 0.00859505 0.00438819]	[0.36944223 0.36548923 0.11090157 0.07565098 0.04409951]
nobs	22059	27574	27574	27574
missing	22059	0	0	0

Tabla A.25: Propiedades estadísticas de variable _price, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[12500. 10500. 11500. 8500. 9000.]	[12500. 8500. 11500. 10500. 14000.]	[10500. 11500. 12500. 13500. 20000.]	[0. 68118.45452695 98057.57024952 26479.35274704 123382.2413035]
top5_freq	[104 99 91 86 85]	[110 106 91 85 82]	[179 160 149 144 135]	[9385 2 2 2 2]
top5_prob	[0.00471463 0.00448796 0.0041253 0.00389864 0.0038533]	[0.00398927 0.0038442 0.00330021 0.00308261 0.00297382]	[0.00649162 0.00580257 0.00540364 0.00522231 0.00489592]	[3.40356858e-01 7.25320955e-05 7.25320955e-05 7.25320955e-05 7.25320955e-05]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	110379	59460	78222	38997
std_err	32746	20065	22702	241
upper_ci	174559	98786	122716	39468
lower_ci	46199	20134	33727	38525
std	4863477	3331817	3769711	39936
iqr	9959	10119	10405	71052
iqr_normal	7383	7501	7713	52671
mad	202281	100814	137997	34948
mad_normal	253522	126351	172953	43801
coef_var	44.062	56.034	48.193	1.024
range	390000000	364203170	390000000	143410
max	390000000	364203170	390000000	143410
min	0.000	0.000	0.000	0.000
skew	60.579	85.625	75.012	0.614
kurtosis	4067	8059	6474	2
jarque_bera	1.51936e+10	7.45932e+10	4.81365e+10	2.70036e+03
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	12500	12500	10500	0
mode_freq	0.005	0.004	0.006	0.340
median	5084	5224	5084	29046
0.1 %	0.263	0.232	0.389	0.000
1.0 %	6.270	7.390	7.452	0.000
5.0 %	11.760	12.141	12.005	0.000
25.0 %	2041	2179	2095	0
75.0 %	12000	12298	12500	71052
95.0 %	32000	31122	32500	112381
99.0 %	58942	54746	55000	129106
99.9 %	262695	110823	135000	138297

Tabla A.26: Propiedades estadísticas de variable publication_date, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1545 1693 1546 1549 721]	[1545. 1693. 1546. 1549. 721.]	[1545. 1693. 1546. 1549. 721.]	[1693 1546 1548 1547 1549]
top5_freq	[10883 6103 895 320 125]	[14087 7611 1137 229 121]	[13284 7427 472 102 101]	[3459 1612 1598 1596 1514]
top5_prob	[0.49335872 0.27666712 0.04057301 0.01450655 0.00566662]	[0.51087981 0.27602089 0.0412345 0.00830492 0.00438819]	[0.48175818 0.26934794 0.01711757 0.00369914 0.00366287]	[0.12544426 0.05846087 0.05795314 0.05788061 0.0549068]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	1471	1476	1473	1504
std_err	2.056	1.786	1.820	1.432
upper_ci	1475	1479	1476	1507
lower_ci	1467	1472	1469	1501
std	305.403	296.604	302.208	237.822
iqr	148.000	148.000	148.000	9.000
iqr_normal	109.713	109.713	109.713	6.672
mad	206.755	199.027	203.032	133.263
mad_normal	259.128	249.443	254.463	167.021
coef_var	0.208	0.201	0.205	0.158
range	1489	1487	1462	1287
max	1693	1693	1693	1693
min	204.000	205.521	231.000	406.000
skew	-2.019	-2.039	-2.043	-2.746
kurtosis	5.892	6.010	5.988	10.071
jarque_bera	22679	29516	29441	92083
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1545	1545	1545	1693
mode_freq	0.493	0.511	0.482	0.125
median	1545	1545	1545	1547
0.1 %	450.696	510.024	505.438	458.000
1.0 %	531.000	538.813	533.029	558.730
5.0 %	628.900	715.105	629.000	841.000
25.0 %	1545	1545	1545	1543
75.0 %	1693	1693	1693	1552
95.0 %	1693	1693	1693	1693
99.0 %	1693	1693	1693	1693
99.9 %	1693	1693	1693	1693

Tabla A.27: Propiedades estadísticas de variable state, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Metropolitana de Santiago' 'Valparaíso' 'Coquimbo' 'Araucanía' "Libertador General Bernardo O'higgins"]	['Metropolitana de Santiago' 'Valparaíso' 'Coquimbo' 'Araucanía' "Libertador General Bernardo O'higgins"]	['Metropolitana de Santiago' 'Valparaíso' 'Coquimbo' 'Araucanía' 'Maule']	['Metropolitana de Santiago' 'Valparaíso' 'Araucanía' 'Los Lagos' 'Coquimbo']
top5_freq	[17248 2014 567 558 305]	[21950 2452 756 682 322]	[22494 2386 658 597 277]	[11784 4500 3155 1740 1316]
top5_prob	[0.78190308 0.0913006 0.02570379 0.0252958 0.01382656]	[0.79603975 0.08892435 0.02741713 0.02473344 0.01167767]	[0.81576848 0.08653079 0.02386306 0.02165083 0.0100457]	[0.42735911 0.16319721 0.11441938 0.06310292 0.04772612]
nobs	22059	27574	27574	27574
missing	22059	0	0	0

Tabla A.28: Propiedades estadísticas de variable county, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Las Condes' 'Santiago' 'Providencia' 'Vitacura' 'Lo Barnechea']	['Las Condes' 'Santiago' 'Providencia' 'Lo Barnechea' 'Vitacura']	['Las Condes' 'Santiago' 'Providencia' 'Vitacura' 'Lo Barnechea']	['Santiago' 'Las Condes' 'Viña del Mar' 'Lo Barnechea' 'Providencia']
top5_freq	[3233 2703 1481 1415 1322]	[4334 3507 1876 1775 1692]	[4577 3767 1945 1894 1855]	[5468 3919 1968 1961 1299]
top5_prob	[0.14656149 0.12253502 0.06713813 0.06414615 0.05993019]	[0.15717705 0.12718503 0.06803511 0.06437223 0.06136215]	[0.1659897 0.1366142 0.07053746 0.06868789 0.06727352]	[0.19830275 0.14212664 0.07137158 0.07111772 0.0471096]
nobs	22059	27574	27574	27574
missing	22059	0	0	0

Tabla A.29: Propiedades estadísticas de variable rooms, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[3. 2. 4. 1. 5.]	[3. 2. 4. 1. 5.]	[3. 2. 4. 1. 5.]	[3. 5. 2. 4. 1.]
top5_freq	[6355 4614 4168 2671 2232]	[8206 5706 5435 3276 2794]	[8192 5817 5466 3228 2869]	[6358 4459 4267 3642 1764]
top5_prob	[0.28809103 0.20916633 0.18894782 0.12108436 0.10118319]	[0.29759919 0.20693407 0.19710597 0.11880757 0.10132734]	[0.29709146 0.2109596 0.19823022 0.1170668 0.10404729]	[0.23057953 0.16171031 0.15474723 0.13208095 0.06397331]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	3.446	3.324	3.316	5.324
std_err	0.026	0.012	0.012	0.052
upper_ci	3.497	3.348	3.339	5.426
lower_ci	3.395	3.301	3.292	5.223
std	3.881	2.004	1.986	8.593
iqr	2.000	2.000	2.000	3.000
iqr_normal	1.483	1.483	1.483	2.224
mad	1.454	1.293	1.285	3.109
mad_normal	1.822	1.621	1.610	3.896
coef_var	1.126	0.603	0.599	1.614
range	399.000	47.000	56.000	399.000
max	400.000	48.000	57.000	400.000
min	1.000	1.000	1.000	1.000
skew	57.785	5.038	4.703	24.930
kurtosis	5331	70	66	948
jarque_bera	2.61061e+10	5.32586e+06	4.60194e+06	1.02949e+09
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	3.000	3.000	3.000	3.000
mode_freq	0.288	0.298	0.297	0.231
median	3.000	3.000	3.000	4.000
0.1 %	1.000	1.000	1.000	1.000
1.0 %	1.000	1.000	1.000	1.000
5.0 %	1.000	1.000	1.000	1.000
25.0 %	2.000	2.000	2.000	3.000
75.0 %	4.000	4.000	4.000	6.000
95.0 %	6.000	6.000	6.000	14.000
99.0 %	12.000	10.000	10.000	25.000
99.9 %	25.000	22.427	23.000	57.000

Tabla A.30: Propiedades estadísticas de variable m_size, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[5000. 50. 60. 200. 70.]	[5000. 50. 60. 100. 200.]	[5000. 30. 45. 40. 50.]	[0. 222105.06 152316.2 41293.19 202121.85]
top5_freq	[601 342 321 285 281]	[742 396 361 346 343]	[151 78 77 66 62]	[10798 2 2 2 2]
top5_prob	[0.02724512 0.01550388 0.01455188 0.0129199 0.01273856]	[0.02690941 0.01436135 0.01309204 0.01254805 0.01243925]	[0.00547617 0.00282875 0.00279249 0.00239356 0.00224849]	[3.91600783e-01 7.25320955e-05 7.25320955e-05 7.25320955e-05 7.25320955e-05]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	146269	2040435	59502	92671
std_err	105454	960179	46375	630
upper_ci	352956	3922351	150396	93907
lower_ci	-60417	158520	-31392	91436
std	15662334	159441727	7700818	104688
iqr	340.500	319.835	364.247	170645.070
iqr_normal	252.413	237.094	270.017	126499.380
mad	290635	4078388	117393	90332
mad_normal	364257	5111501	147130	113215
coef_var	107.079	78.141	129.421	1.130
range	2.24100e+09	1.85996e+10	1.26328e+09	4.00260e+05
max	2.24100e+09	1.85996e+10	1.26328e+09	4.00260e+05
min	0.000	0.001	1.000	0.000
skew	134.762	90.861	160.494	0.825
kurtosis	19053	9035	26268	2
jarque_bera	3.33616e+11	9.37643e+10	7.92684e+11	3.48544e+03
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	5000	5000	5000	0
mode_freq	0.027	0.027	0.005	0.392
median	145.000	144.000	148.530	53178.165
0.1 %	2.000	3.298	17.278	0.000
1.0 %	22.000	23.313	25.976	0.000
5.0 %	35.000	35.000	35.392	0.000
25.0 %	66.000	66.939	67.651	0.000
75.0 %	406.500	386.774	431.898	170645.070
95.0 %	5000	5000	4220	297686
99.0 %	10200	8437	9176	348477
99.9 %	70000	33892	46037	379558

Tabla A.31: Propiedades estadísticas de variable m_built, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[140. 60. 120. 50. 70.]	[140. 60. 50. 120. 70.]	[140. 50. 60. 70. 40.]	[1. 526.52 926.3 825.58 804.72]
top5_freq	[700 467 444 431 415]	[939 568 545 529 486]	[354 165 159 150 136]	[17926 3 3 3 2]
top5_prob	[0.03173308 0.0211705 0.02012784 0.01953851 0.01881318]	[0.03405382 0.02059912 0.019765 0.01918474 0.0176253]	[0.01283818 0.0059839 0.0057663 0.00543991 0.00493218]	[6.50105172e-01 1.08798143e-04 1.08798143e-04 1.08798143e-04 7.25320955e-05]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	1771	2455	1326	391
std_err	664.365	912.066	379.290	10.197
upper_ci	3073	4243	2070	411
lower_ci	469.205	667.459	583.072	370.831
std	98673	151452	62983	1693
iqr	140.000	132.453	139.471	358.523
iqr_normal	103.782	98.188	103.390	265.773
mad	3202	4581	2323	550
mad_normal	4013	5742	2912	689
coef_var	55.706	61.689	47.482	4.333
range	11999999	11999999	6213436	40897
max	12000000	12000000	6213437	40898
min	1.000	1.000	1.000	1.000
skew	96.078	74.742	76.091	13.993
kurtosis	10659	5793	6548	239
jarque_bera	1.04399e+11	3.85432e+10	4.92421e+10	6.48622e+07
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	140.000	140.000	140.000	1.000
mode_freq	0.032	0.034	0.013	0.650
median	107.000	106.000	107.688	1.000
0.1 %	2.000	4.799	13.968	1.000
1.0 %	23.000	23.982	25.208	1.000
5.0 %	33.000	34.000	34.000	1.000
25.0 %	60.000	60.000	60.278	1.000
75.0 %	200.000	192.453	199.749	359.523
95.0 %	490.000	450.000	469.831	1534.202
99.0 %	1200	833	1116	2331
99.9 %	37947	10395	32578	28397

Tabla A.32: Propiedades estadísticas de variable transaction_type, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Venta' 'Arriendo' 'Busco arriendo' 'Compro']	['Venta' 'Arriendo' 'Busco arriendo' 'Compro']	['Venta' 'Arriendo']	['Venta' 'Arriendo' 'Compro' 'Busco arriendo']
top5_freq	[17540 4517 1 1]	[22100 5471 2 1]	[21933 5641]	[17449 8280 1824 21]
top5_prob	[7.95140306e-01 2.04769029e-01 4.53329707e-05 4.53329707e-05]	[8.01479655e-01 1.98411547e-01 7.25320955e-05 3.62660477e-05]	[0.79542322 0.20457678]	[0.63280627 0.30028288 0.06614927 0.00076159]
nobs	22059	27574	27574	27574
missing	22059	0	0	0

A.11. Estadísticos Economicos - Conjunto B

Tabla A.33: Propiedades estadísticas de variable m_built, Economicos (B-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[-1. 60. 50. 70. 100.]	[-1. 60. 70. 50. 80.]	[-1. 0. 60. 50. 70.]	[-1.00000000e+00 8.60224752e+08 6.62292673e+08 6.07707243e+08 9.01931774e+08]
top5_freq	[188514 13831 11796 11648 9716]	[238437 16647 15118 14059 12406]	[136164 1703 1487 1214 1013]	[190533 3 3 3 3]
top5_prob	[0.34534596 0.02533753 0.02160954 0.02133841 0.01779911]	[0.34944119 0.024397 0.02215617 0.02060416 0.01818161]	[0.19955506 0.00249583 0.00217927 0.00177918 0.0014846]	[2.79235511e-01 4.39664800e-06 4.39664800e-06 4.39664800e-06 4.39664800e-06]
nobs	545870	682338	682338	682338
missing	0.000	0.000	0.000	0.000
mean	2.27142e+09	1.07079e+12	3.45545e+09	3.50871e+08
std_err	2.04392e+09	1.03121e+11	2.25542e+09	4.19404e+05
upper_ci	6.27743e+09	1.27291e+12	7.87598e+09	3.51693e+08
lower_ci	-1.73460e+09	8.68681e+11	-9.65085e+08	3.50049e+08
std	1.51011e+12	8.51817e+13	1.86306e+12	3.46443e+08
iqr	99.000	96.071	96.244	584547063.877
iqr_normal	73.389	71.217	71.346	433325386.863
mad	4.54267e+09	2.14100e+12	6.91040e+09	2.89207e+08
mad_normal	5.69340e+09	2.68334e+12	8.66090e+09	3.62467e+08
coef_var	664.832	79.550	539.165	0.987
range	1.11111e+15	1.11111e+16	1.10035e+15	2.38947e+09
max	1.11111e+15	1.11111e+16	1.10035e+15	2.38947e+09
min	-1.000	-1.000	-1.000	-1.000
skew	730.252	101.863	580.139	0.844
kurtosis	536948	11645	338072	3
jarque_bera	6.55754e+15	3.85454e+12	3.24941e+15	8.11271e+04
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	-1.000	-1.000	-1.000	-1.000
mode_freq	0.345	0.349	0.200	0.279
median	50.000	50.000	51.432	283214063.174
0.1 %	-1.000	-1.000	-1.000	-1.000
1.0 %	-1.000	-1.000	-1.000	-1.000
5.0 %	-1.000	-1.000	-1.000	-1.000
25.0 %	-1.000	-1.000	4.395	-1.000
75.0 %	98.000	95.071	100.639	584547062.877
95.0 %	400.000	356.625	513.500	1005038782.529
99.0 %	8351	6005	61477	1298434790
99.9 %	550000	454715	478266	1620793464

Tabla A.34: Propiedades estadísticas de variable bathrooms, Economicos (B-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[-1. 2. 1. 3. 4.]	[-1. 2. 1. 3. 4.]	[-1. 2. 1. 3. 4.]	[-1. 1. 2. 3. 8.]
top5_freq	[206916 136581 134963 43884 14719]	[259422 172420 169557 54199 17651]	[257978 171740 168539 54581 18527]	[338218 132407 76570 36301 30470]
top5_prob	[0.37905728 0.25020792 0.24724385 0.08039277 0.0269643]	[0.38019574 0.25269002 0.24849415 0.07943131 0.02586841]	[0.37807949 0.25169344 0.24700222 0.07999115 0.02715223]	[0.49567516 0.19404899 0.11221711 0.05320091 0.04465529]
nobs	545870	682338	682338	682338
missing	0.000	0.000	0.000	0.000
mean	0.815	0.788	0.815	1.376
std_err	0.003	0.002	0.002	0.007
upper_ci	0.820	0.792	0.819	1.391
lower_ci	0.810	0.784	0.811	1.362
std	1.898	1.603	1.689	5.990
iqr	3.000	3.000	3.000	3.000
iqr_normal	2.224	2.224	2.224	2.224
mad	1.376	1.359	1.373	2.502
mad_normal	1.725	1.704	1.720	3.136
coef_var	2.328	2.034	2.071	4.352
range	437.000	41.000	181.000	437.000
max	436.000	40.000	180.000	436.000
min	-1.000	-1.000	-1.000	-1.000
skew	36.380	0.498	3.433	17.659
kurtosis	6629	5	235	601
jarque_bera	9.98582e+11	1.26362e+05	1.52873e+09	1.02028e+10
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	-1.000	-1.000	-1.000	-1.000
mode_freq	0.379	0.380	0.378	0.496
median	1.000	1.000	1.000	1.000
0.1 %	-1.000	-1.000	-1.000	-1.000
1.0 %	-1.000	-1.000	-1.000	-1.000
5.0 %	-1.000	-1.000	-1.000	-1.000
25.0 %	-1.000	-1.000	-1.000	-1.000
75.0 %	2.000	2.000	2.000	2.000
95.0 %	3.000	3.000	3.000	8.000
99.0 %	5.000	5.000	5.000	13.000
99.9 %	9.000	7.000	9.000	83.000

Tabla A.35: Propiedades estadísticas de variable transaction_type, Economicos (B-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Venta' 'Arriendo' 'Busco arriendo' 'Compro' 'None']	['Venta' 'Arriendo' 'Busco arriendo' 'Compro' 'None']	['Venta' 'Arriendo' 'Busco arriendo' 'Compro' 'None']	['Venta' 'Arriendo' 'Otros' 'Busco arriendo' 'Compro']
top5_freq	[282495 258300 3031 1901 86]	[352202 324618 3417 2003 59]	[354665 325665 1521 451 24]	[366440 270864 29831 7637 4585]
top5_prob	[5.17513327e-01 4.73189587e-01 5.55260410e-03 3.48251415e-03 1.57546669e-04]	[5.16169406e-01 4.75743693e-01 5.00778207e-03 2.93549531e-03 8.64674106e-05]	[5.19779054e-01 4.77278123e-01 2.22910053e-03 6.60962749e-04 3.51731840e-05]	[0.5370359 0.39696455 0.0437188 0.0111924 0.00671954]
nobs	545870	682338	682338	682338
missing	545870	0	0	0

Tabla A.36: Propiedades estadísticas de variable m_size, Economicos (B-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[-1.e+00 5.e+03 2.e+02 6.e+01 5.e+01]	[-1.e+00 5.e+03 2.e+02 6.e+01 5.e+01]	[-1.e+00 5.e+03 2.e+02 5.e+01 6.e+01]	[-1.00000000e+03 7.44169509e+14 3.81984385e+15 1.43692622e+15 3.02263916e+15]
top5_freq	[245062 19573 6932 6312 5748]	[312676 27731 7900 7858 7231]	[189632 8297 459 380 367]	[326171 3 3 3 3]
top5_prob	[0.44893839 0.03585652 0.01269899 0.01156319 0.01052998]	[0.4582421 0.04064115 0.01157784 0.01151629 0.01059739]	[0.27791505 0.01215966 0.00067269 0.00055691 0.00053786]	[4.78019691e-01 4.39664800e-06 4.39664800e-06 4.39664800e-06 4.39664800e-06]
nobs	545870	682338	682338	682338
missing	0.000	0.000	0.000	0.000
mean	2.03551e+16	1.32450e+18	1.05479e+16	1.59961e+15
std_err	2.03549e+16	1.06899e+17	9.40725e+15	2.73730e+12
upper_ci	6.02499e+16	1.53401e+18	2.89858e+16	1.60497e+15
lower_ci	-1.95397e+16	1.11498e+18	-7.88994e+15	1.59424e+15
std	1.50388e+19	8.83023e+19	7.77074e+18	2.26112e+15
iqr	181.000	172.566	211.893	2734253148430780.000
iqr_normal	134.176	127.924	157.076	2026904891909700.750
mad	4.07100e+16	2.64804e+18	2.10957e+16	1.79642e+15
mad_normal	5.10224e+16	3.31883e+18	2.64396e+16	2.25148e+15
coef_var	738.823	66.669	736.708	1.414
range	1.11111e+22	1.11111e+22	6.36472e+21	2.22408e+16
max	1.11111e+22	1.11111e+22	6.36472e+21	2.22408e+16
min	-1000.000	-1000.000	-546.048	-1000.000
skew	738.828	84.640	807.089	1.614
kurtosis	545868	8261	659774	6
jarque_bera	6.77722e+15	1.93945e+12	1.23759e+16	4.85481e+05
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	-1.000	-1.000	-1.000	-1000.000
mode_freq	0.449	0.458	0.278	0.478
median	36.000	38.000	46.354	211099107154120.188
0.1 %	-1.000	-1.000	-1.000	-1000.000
1.0 %	-1.000	-1.000	-1.000	-1000.000
5.0 %	-1.000	-1.000	-1.000	-1000.000
25.0 %	-1.000	-1.000	-1.000	-1000.000
75.0 %	180.000	171.566	210.893	2734253148429780.000
95.0 %	5000	5000	5089	6325359364193757
99.0 %	50000	43877	68497	8943590289181725
99.9 %	4920000	1695621	5664965	12390596650635482

Tabla A.37: Propiedades estadísticas de variable property_type, Economicos (B-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Departamento' 'Casa' 'Sitio o Terreno' 'Parcela o Chacra' 'Departamen- to Amoblado']	['Departamento' 'Casa' 'Sitio o Terreno' 'Parcela o Chacra' 'Departamen- to Amoblado']	['Departamento' 'Casa' 'Sitio o Terreno' 'Parcela o Chacra' 'Departamen- to Amoblado']	['Casa' 'Departamento' 'Parcela o Chacra' 'Resi- dencial/Pieza' 'Departa- mento Amoblado']
top5_freq	[211405 142054 31393 30020 27415]	[267411 179229 39247 36474 33396]	[272561 181361 40333 37788 31513]	[220142 142362 78898 40399 34352]
top5_prob	[0.38728085 0.26023412 0.05751003 0.05499478 0.05022258]	[0.39190401 0.26266894 0.05751841 0.05345445 0.04894349]	[0.39945159 0.26579349 0.05911 0.05538018 0.04618386]	[0.32262896 0.20863853 0.11562891 0.05920673 0.05034455]
nobs	545870	682338	682338	682338
missing	545870	0	0	0

Tabla A.38: Propiedades estadísticas de variable county, Economicos (B-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Santiago' 'Viña del Mar' 'Las Condes' 'Providencia' 'None']	['Santiago' 'Viña del Mar' 'Las Condes' 'Providencia' 'None']	['Santiago' 'Viña del Mar' 'Las Condes' 'Providencia' 'None']	['Santiago' 'Viña del Mar' 'None' 'Providencia' 'Puerto Montt']
top5_freq	[65125 33263 32327 27981 24863]	[82872 42783 40663 35645 29936]	[94757 49939 47207 37862 34230]	[75755 71804 64070 38903 31558]
top5_prob	[0.11930496 0.06093575 0.05922106 0.05125946 0.04554747]	[0.121453 0.0627006 0.05959363 0.05223951 0.04387268]	[0.13887106 0.07318807 0.06918419 0.05548863 0.05016575]	[0.11102269 0.1052323 0.09389775 0.05701427 0.04624981]
nobs	545870	682338	682338	682338
missing	545870	0	0	0

Tabla A.39: Propiedades estadísticas de variable rooms, Economicos (B-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[-1. 3. 2. 1. 4.]	[-1. 3. 2. 1. 4.]	[-1. 3. 2. 1. 4.]	[-1. 2. 3. 1. 4.]
top5_freq	[196417 125902 97220 54183 44539]	[245532 158537 123328 66780 55841]	[246891 164177 122066 66430 55340]	[342760 96371 61068 55167 36727]
top5_prob	[0.35982377 0.23064466 0.17810101 0.0992599 0.08159269]	[0.35983926 0.23234379 0.18074327 0.09786938 0.08183774]	[0.36183094 0.24060949 0.17889374 0.09735644 0.0811035]	[0.50233169 0.14123645 0.08949817 0.08084996 0.05382523]
nobs	545870	682338	682338	682338
missing	0.000	0.000	0.000	0.000
mean	9129	1464	1	2371371
std_err	4082	1463	0	58829
upper_ci	17129	4332	1	2486674
lower_ci	1128	-1403	1	2256068
std	3015768	1208483	2	48594977
iqr	4.000	4.000	4.000	4.000
iqr_normal	2.965	2.965	2.965	2.965
mad	18254	2926	2	4731472
mad_normal	22878	3667	2	5930020
coef_var	330.362	825.250	1.482	20.492
range	998252511	998252511	39	998252511
max	998252510	998252510	38	998252510
min	-1.000	-1.000	-1.000	-1.000
skew	330.413	826.036	0.244	20.444
kurtosis	109174	682336	3	419
jarque_bera	2.71088e+14	1.32368e+16	7.34519e+03	4.96620e+09
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	-1.000	-1.000	-1.000	-1.000
mode_freq	0.360	0.360	0.362	0.502
median	2.000	2.000	2.000	-1.000
0.1 %	-1.000	-1.000	-1.000	-1.000
1.0 %	-1.000	-1.000	-1.000	-1.000
5.0 %	-1.000	-1.000	-1.000	-1.000
25.0 %	-1.000	-1.000	-1.000	-1.000
75.0 %	3.000	3.000	3.000	3.000
95.0 %	5.000	4.000	4.000	8.000
99.0 %	7.000	6.000	6.000	15.000
99.9 %	11.000	9.000	7.000	998252510.000

Tabla A.40: Propiedades estadísticas de variable publication_date, Economicos (B-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1545 1693 1392 1492 1408]	[1545. 1693. 1392. 1492. 1408.]	[1545. 1693. 1392. 1492. 1408.]	[1693 1543 1545 1544 1547]
top5_freq	[19744 11666 10260 3838 2445]	[26496 14721 10960 4163 2581]	[19616 12624 8473 1812 660]	[13161 4256 4119 4114 3969]
top5_prob	[0.03616978 0.02137139 0.01879568 0.00703098 0.00447909]	[0.0388312 0.02157435 0.01606242 0.00610108 0.00378258]	[0.02874822 0.01850109 0.0124176 0.00265558 0.00096726]	[0.01928809 0.00623738 0.0060366 0.00602927 0.00581677]
nobs	545870	682338	682338	682338
missing	0.000	0.000	0.000	0.000
mean	702.711	699.555	704.030	862.363
std_err	0.625	0.555	0.557	0.569
upper_ci	703.935	700.644	705.122	863.478
lower_ci	701.487	698.466	702.938	861.248
std	461.457	458.848	460.100	469.859
iqr	715.000	707.435	714.553	736.000
iqr_normal	530.030	524.422	529.699	545.598
mad	387.485	384.492	386.493	426.510
mad_normal	485.640	481.889	484.397	534.551
coef_var	0.657	0.656	0.654	0.545
range	3208	3208	3207	1628
max	1693	1693	1693	1693
min	-1515	-1515	-1514	65
skew	0.533	0.518	0.537	0.064
kurtosis	2.198	2.263	2.187	1.668
jarque_bera	40467	45997	51584	50877
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1545	1545	1545	1693
mode_freq	0.036	0.039	0.029	0.019
median	609.000	608.000	609.141	988.000
0.1 %	42.000	42.830	44.027	92.000
1.0 %	56.000	56.255	61.339	110.000
5.0 %	100.000	99.945	105.110	143.000
25.0 %	322.000	322.035	323.591	463.000
75.0 %	1037	1029	1038	1199
95.0 %	1545	1545	1545	1550
99.0 %	1693	1693	1693	1693
99.9 %	1693	1693	1693	1693

Tabla A.41: Propiedades estadísticas de variable _price, Economicos (B-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[0. 3500. 5500. 6500. 4500.]	[0.00000000e+00 4.50000000e+03 3.00000000e+03 3.50000000e+03 3.83275261e+15]	[0. 3500. 3000. 2500. 6500.]	[0.00000000e+00 6.29812066e+08 2.85923597e+08 2.06496794e+09 1.38283530e+09]
top5_freq	[17989 865 767 763 740]	[22674 724 702 691 11]	[32910 1326 1081 1079 1073]	[366966 3 3 2 2]
top5_prob	[0.03295473 0.00158463 0.0014051 0.00139777 0.00135563]	[3.32298656e-02 1.06105772e-03 1.02881563e-03 1.01269459e-03 1.61210427e-05]	[0.04823123 0.00194332 0.00158426 0.00158133 0.00157253]	[5.37806776e-01 4.39664800e-06 4.39664800e-06 2.93109866e-06 2.93109866e-06]
nobs	545870	682338	682338	682338
missing	0.000	0.000	0.000	0.000
mean	7.09830e+09	5.24838e+11	5.61835e+09	5.17667e+08
std_err	7.02174e+09	3.91827e+10	5.61709e+09	9.72252e+05
upper_ci	2.08606e+10	6.01635e+11	1.66276e+10	5.19573e+08
lower_ci	-6.66405e+09	4.48042e+11	-5.39094e+09	5.15762e+08
std	5.18787e+12	3.23664e+13	4.63993e+12	8.03117e+08
iqr	3538	3504	3543	859824612
iqr_normal	2623	2597	2626	637388938
mad	1.41961e+10	1.04920e+12	1.12365e+10	6.26876e+08
mad_normal	1.77921e+10	1.31498e+12	1.40829e+10	7.85673e+08
coef_var	730.861	61.669	825.852	1.551
range	3.83275e+15	3.83275e+15	3.83275e+15	7.17933e+09
max	3.83275e+15	3.83275e+15	3.83275e+15	7.17933e+09
min	0.000	0.000	0.000	0.000
skew	738.712	80.859	826.036	1.782
kurtosis	545753	7533	682336	6
jarque_bera	6.77436e+15	1.61260e+12	1.32368e+16	6.37854e+05
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	0.000	0.000	0.000	0.000
mode_freq	0.033	0.033	0.048	0.538
median	174.025	146.183	170.000	0.000
0.1 %	0.000	0.000	0.000	0.000
1.0 %	0.000	0.000	0.000	0.000
5.0 %	1.270	1.527	0.005	0.000
25.0 %	12.283	12.353	12.199	0.000
75.0 %	3550	3516	3555	859824612
95.0 %	14400	13725	14700	2245714576
99.0 %	47000	40662	47000	3209542489
99.9 %	1906676	971231	1347963	4513684174

Tabla A.42: Propiedades estadísticas de variable state, Economicos (B-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Metropolitana de Santiago' 'Valparaíso' 'Biobío' 'Araucanía' "Liberador General Bernardo O'higgins"]	['Metropolitana de Santiago' 'Valparaíso' 'Biobío' 'Araucanía' "Liberador General Bernardo O'higgins"]	['Metropolitana de Santiago' 'Valparaíso' 'Biobío' 'Araucanía' "Liberador General Bernardo O'higgins"]	['Metropolitana de Santiago' 'Valparaíso' 'Los Lagos' 'None' 'Antofagasta']
top5_freq	[272808 108197 29379 21581 16533]	[345222 136959 36425 26568 20094]	[353117 137827 35992 26247 19119]	[281780 157931 43904 40684 39267]
top5_prob	[0.49976734 0.1982102 0.05382051 0.03953505 0.03028743]	[0.50593987 0.20072017 0.05338263 0.03893671 0.02944875]	[0.51751038 0.20199227 0.05274805 0.03846627 0.02801984]	[0.41296249 0.23145567 0.06434348 0.05962441 0.05754773]
nobs	545870	682338	682338	682338
missing	545870	0	0	0