



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. Contexto Temporal/tecnológico	5
1.4. Objetivo	6
2. Revisión Bibliográfica	7
2.1. Tipos de Datos	7
2.2. Privacidad de Datos	8
2.2.1. Tipo de datos a ser protegidos	8
2.2.2. Tipos de riesgos de divulgación	9
2.2.3. Regulación de datos sintéticos	10
2.2.4. Protección de Privacidad	10
2.3. Generación de Datos Sintéticos	11
2.3.1. Generación de datos tabulares	12
2.3.2. Generación de texto en base de datos tabulares	13
2.4. Metricas de evaluación	16
2.4.1. SDMetrics	16
2.4.2. Conjuntos Estadísticos	18
3. Desarrollo	20
3.1. Recursos disponibles	20

3.1.1.	Conjuntos de datos	20
3.1.2.	Computación y Software	22
3.2.	Desarrollo del flujo de procesamiento	23
3.3.	Modelos	27
3.3.1.	Modelos para datos tabulares	27
3.3.2.	Modelos para textos	30
3.4.	Obtención de Métricas	31
4.	Resultados	35
4.1.	King County	36
4.1.1.	Reportes	36
4.2.	Economicos	51
4.2.1.	Reportes - Conjunto A	52
4.2.2.	Reportes - Conjunto B	57
5.	Conclusiones y discusión	62
5.1.	Conclusiones	62
5.2.	Discusión	63
	Bibliografía	66
	Apéndice A. Anexos	67
A.1.	Código de entrenamiento de economicos	68
A.2.	Lista completa de figura pairwise kingcounty	69
A.3.	Smote y TDDPM en KingCounty Graficas por Columnas	73
A.4.	Tabla de comparación de Top5 KingCounty	73
A.5.	Figuras de correlación Economicos - Conjunto A	73
A.6.	Figuras de correlación Economicos - Conjunto B	76
A.6.1.	Conjunto A	79

Índice de Tablas

2.1. Tipos de datos estructurados	7
2.2. Niveles de revelación y ejemplos	8
2.3. Tipos de Riesgos de Divulgación y sus Descripciones	9
2.4. Metricas de privacidad	10
2.5. Estado del arte en generación de datos tabulares	12
2.6. Estado del arte en generación de textos en base a datos	13
2.7. Ejemplo de tabla de entrada	14
2.8. Listado de conjunto estadísticos	18
3.1. Conjunto de datos King County	21
3.2. Conjunto de datos Economicos.cl	22
3.3. Computador Usado	23
3.4. Variables de entrada para <i>Synthetic</i>	28
3.5. Modelos Tabulares Soportados	29
3.6. Metricas para campos numericos	31
3.7. Métricas para campos categóricos	33
3.8. Ejemplo de scores promedios	34
4.1. Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, King County	36
4.2. Evaluación de Cobertura Categoría-Rango para Modelos SMOTE-ENC y TDDPM_MLP, King County	41

4.3. Evaluación de Similitud de Distribución para Modelos SMOTE-ENC y TDDPM_MLP, King County	42
4.4. Distancia de registros más cercanos, percentil 5, datos king county	46
4.5. Distancia de registros más cercanos, percentil 1, datos king county	46
4.6. Distancia de registros más cercanos, minimo, datos king county	46
4.7. Ejemplos para el modelo smote-enc, minimo	47
4.8. Ejemplos para el modelo tddpm_mlp, minimo	48
4.9. Proporción entre el más cercano y el segundo más cercano, percentil 5, datos king county	49
4.10. Proporción entre el más cercano y el segundo más cercano, percentil 1, datos king county	49
4.11. Proporción entre el más cercano y el segundo más cercano, minimo, datos king county	49
4.12. Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, Economicos	52
4.13. Distancia de registros más cercanos, percentil 5, datos economicos	53
4.14. Distancia de registros más cercanos, percentil 1, datos economicos	53
4.15. Distancia de registros más cercanos, minimo, datos economicos	53
4.16. Proporción entre el más cercano y el segundo más cercano, percentil 5, datos economicos	54
4.17. Proporción entre el más cercano y el segundo más cercano, percentil 1, datos economicos	54
4.18. Proporción entre el más cercano y el segundo más cercano, minimo, datos economicos	54
4.19. Evaluación de Cobertura Categoría-Rango para Modelos SMOTE-ENC y TDDPM_MLP, Economicos	55
4.20. Evaluación de Similitud de Distribución para Modelos SMOTE-ENC y TDDPM_MLP, Economicos	55
4.21. Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, Economicos	57
4.22. Distancia de registros más cercanos, percentil 5, datos economicos	58
4.23. Distancia de registros más cercanos, percentil 1, datos economicos	58
4.24. Distancia de registros más cercanos, minimo, datos economicos	58

4.25. Proporción entre el más cercano y el segundo más cercano, percentil 5, datos económicos	59
4.26. Proporción entre el más cercano y el segundo más cercano, percentil 1, datos económicos	59
4.27. Proporción entre el más cercano y el segundo más cercano, mínimo, datos económicos	59
4.28. Evaluación de Cobertura Categoría-Rango para Modelos SMOTE-ENC y TDDPM_MLP, Económicos	60
4.29. Evaluación de Similitud de Distribución para Modelos SMOTE-ENC y TDDPM_MLP, Económicos	60

Índice de Ilustraciones

3.1. Proceso para generar datos sintéticos con SDV	25
3.2. Proceso para generar datos sintéticos completo	26
3.3. Carpetas y archivos esperados generados por <i>Synthetic</i>	29
4.1. Correlación de conjunto Real y Modelo: copulagan	37
4.2. Correlación de conjunto Real y Modelo: gaussiancopula	38
4.3. Correlación de conjunto Real y Modelo: smote-enc	39
4.4. Correlación de conjunto Real y Modelo: tddpm_mlp	40
4.5. Frecuencia del campo grade en el modelo real y top2	43
4.6. Frecuencia del campo bedrooms en el modelo real y top2	44
4.7. Frecuencia del campo sqft lot15 en el modelo real y top2	45
4.8. Frecuencia del campo privacy en el modelo real y top2	50
4.9. Correlación de conjunto Real y Modelo: smote-enc	56
4.10. Correlación de conjunto Real y Modelo: tddpm_mlp	56
4.11. Correlación de conjunto Real y Modelo: smote-enc	61
4.12. Correlación de conjunto Real y Modelo: tddpm_mlp	61
A.1. Correlación de conjunto Real y Modelo: copulagan	69
A.2. Correlación de conjunto Real y Modelo: tvae	69
A.3. Correlación de conjunto Real y Modelo: gaussiancopula	70
A.4. Correlación de conjunto Real y Modelo: ctgan	70
A.5. Correlación de conjunto Real y Modelo: tablepreset	71

A.6. Correlación de conjunto Real y Modelo: smote-enc	71
A.7. Correlación de conjunto Real y Modelo: tddpm_mlp	72
A.8. Correlación de conjunto Real y Modelo: copulagan	73
A.9. Correlación de conjunto Real y Modelo: tvae	74
A.10. Correlación de conjunto Real y Modelo: gaussiancopula	74
A.11. Correlación de conjunto Real y Modelo: ctgan	75
A.12. Correlación de conjunto Real y Modelo: smote-enc	75
A.13. Correlación de conjunto Real y Modelo: tddpm_mlp	76
A.14. Correlación de conjunto Real y Modelo: copulagan	76
A.15. Correlación de conjunto Real y Modelo: tvae	77
A.16. Correlación de conjunto Real y Modelo: gaussiancopula	77
A.17. Correlación de conjunto Real y Modelo: ctgan	78
A.18. Correlación de conjunto Real y Modelo: smote-enc	78
A.19. Correlación de conjunto Real y Modelo: tddpm_mlp	79

Lista de códigos

1.	Devcontainer del actual proyecto.	23
2.	Instanciando clase Synthetic	28
3.	Obtención de métricas en economicos_run-a.py	31
4.	Mostrando Scores Promedios Calculados	34
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.	68

Capítulo 1

Introducción

Cuando se revise esta tesis, estará desactualizada. Desde AlexNet [27] en 2012, el liderazgo en el problema de clasificación de imágenes ha cambiado al menos 15 veces [6]. En el campo de texto a imágenes, modelos como DALL-E 2 [1], Google Imagen [3] y Stable Diffusion [8] 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 [28, 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* [19] 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. Contexto Temporal/tecnológico

Usando ChatGPT en el marzo 2023 [29].

Prompt

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

ChatGPT-4

La generación de datos sintéticos ha surgido como una técnica innovadora y prometedora en el ámbito de la inteligencia artificial, la ciencia de datos y el aprendizaje automático. Esta tesis aborda la relevancia de la generación de datos sintéticos y su impacto en la investigación y el desarrollo de soluciones tecnológicas. La generación de datos sintéticos es esencial debido a diversas razones, entre las que destacan la privacidad, la escasez de datos y la mejora del rendimiento de los modelos.

En primer lugar, la privacidad de los datos es un tema de creciente preocupación en la era digital. La generación de datos sintéticos permite abordar este problema al crear datos que imitan las características y la distribución de los datos reales sin revelar información sensible o identificable. Esto es especialmente relevante en campos como la medicina, las finanzas o la investigación social, donde la protección de la privacidad de los individuos es de suma importancia.

En segundo lugar, la escasez de datos es un desafío común en diversas aplicaciones de aprendizaje automático y ciencia de datos. La generación de datos sintéticos puede mitigar este problema al complementar conjuntos de datos limitados o desequilibrados. Esto permite a los investigadores y profesionales desarrollar y evaluar modelos más sólidos y precisos, mejorando así la calidad y la confiabilidad de las soluciones propuestas.

Además, la generación de datos sintéticos contribuye a la mejora del rendimiento de los modelos de aprendizaje automático. Al ampliar y enriquecer conjuntos de datos existentes, los modelos pueden aprender patrones y relaciones más complejas y generalizables, lo que se traduce en una mejor capacidad de predicción y clasificación.

Esta tesis examinará las técnicas y enfoques actuales en la generación de datos sintéticos, así como las aplicaciones y desafíos asociados a su implementación en diferentes contextos. También se analizará el papel de los datos sintéticos en la ética y la privacidad de los datos y su impacto en la toma de decisiones basada en datos en el mundo real.

1.4. Objetivo

Objetivo General:

El objetivo general es definir un mecanismo para generar conjuntos de datos sintéticos estructurados, que incluyen textos y comparar, mediante modelos generativos y su contraparte original.

Objetivos Específicos:

1. Elaborar modelos generativos para sintetizar nuevos conjuntos de datos, a partir de los originales que incluyen textos.
2. Comparar los conjuntos de datos sintéticos y originales en 2 casos, propiedades estadísticas, distribuciones y privacidad.

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 [20]. En un análisis posterior, Kiran Adnan y Rehan Akbar [11] 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* [16].

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 [16]. 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* [18].

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* [16] 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 [19] 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 [16]. Los nuevos datos podrían ser afectos por:

1. Regulation (EU) 2016/679 of the European Parliament and of the Council [34], 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 [31]
3. Health insurance portability and accountability act of 1996 [10]

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. [37]
<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. [37]

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 [16]. 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 [35]	2023-02-04	Github
PreFair: Privately Generating Justifiably Fair Synthetic Data [33]	2022-12-20	
GenSyn: A Multi-stage Framework for Generating Synthetic Microdata using Macro Data Sources [9]	2022-12-08	Github
TabDDPM: Modelling Tabular Data with Diffusion Models [26]	2022-10-30	Github
Language models are realistic tabular data generators [15]	2022-10-12	Github
Ctab-gan+: Enhancing tabular data synthesis [38]	2022-04-01	Github
Ctab-gan: Effective table data synthesizing [37]	2021-05-31	Github
Modeling Tabular data using Conditional GAN [36]	2019-10-28	Github
SMOTE: synthetic minority over-sampling technique [17]	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 [13]	2022-10-17	T5
Text-to-text pre-training for data-to-text tasks [24]	2021-07-09	T5
TaPas: Weakly supervised table parsing via pre-training [21]	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 [13, 21]. 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* [24] 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 [7].

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 [23] 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 [25, 32].

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:

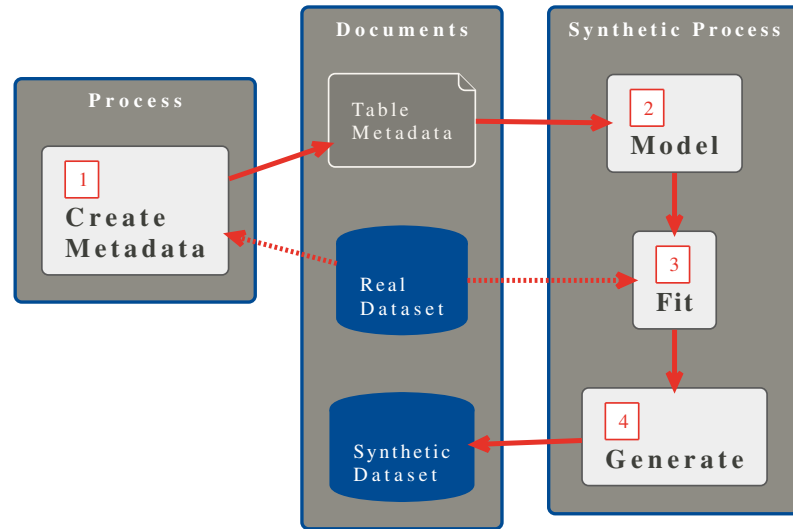


Figura 3.1: Proceso para generar datos sintéticos con SDV

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.

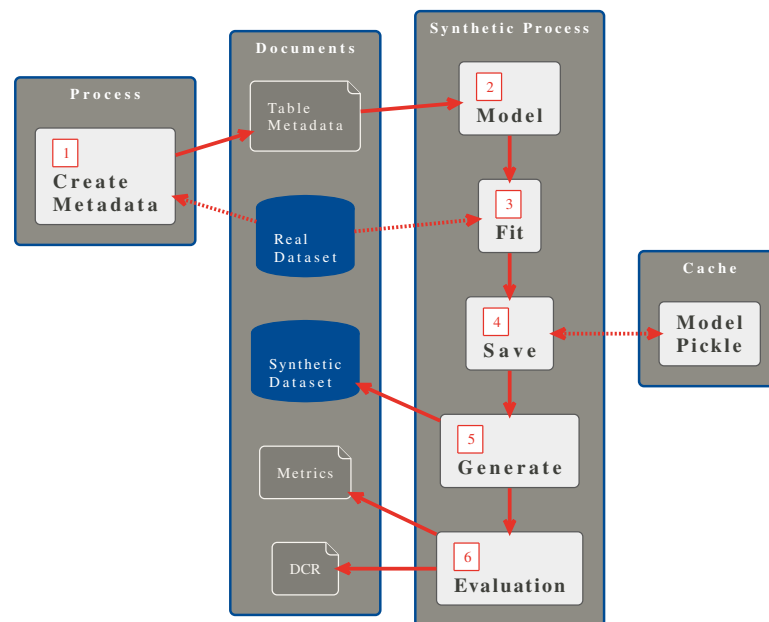


Figura 3.2: Proceso para generar datos sintéticos completo

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 [25]
tvae	SDV [25]
gaussiancopula	SDV [25]
ctgan	SDV [25]
tablepreset	SDV [25]
smote-enc	tabDDPM [12]
tddpm_mlp	tabDDPM [12]

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).

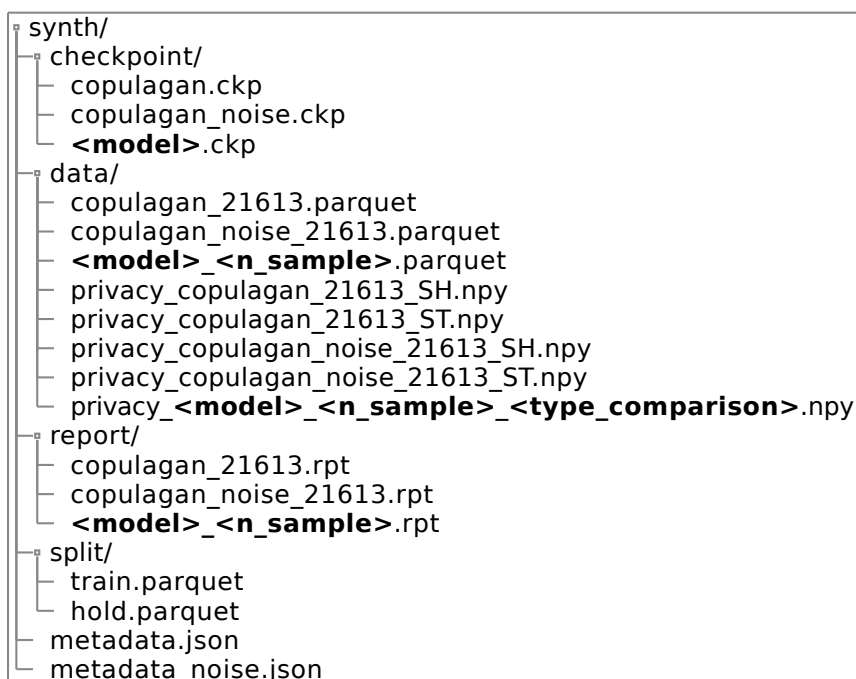


Figura 3.3: Carpetas y archivos esperados generados por *Synthetic*

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* [24]. 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>
<título, 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.current_metrics())
```

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. Reportes

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_mlp y smote-enc, 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.

Correlación pairwise

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 Figura A.1 Correlación de conjunto Real y Modelo: copulagan y Figura A.3 Correlación de conjunto Real y Modelo: gaussiancopula 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.

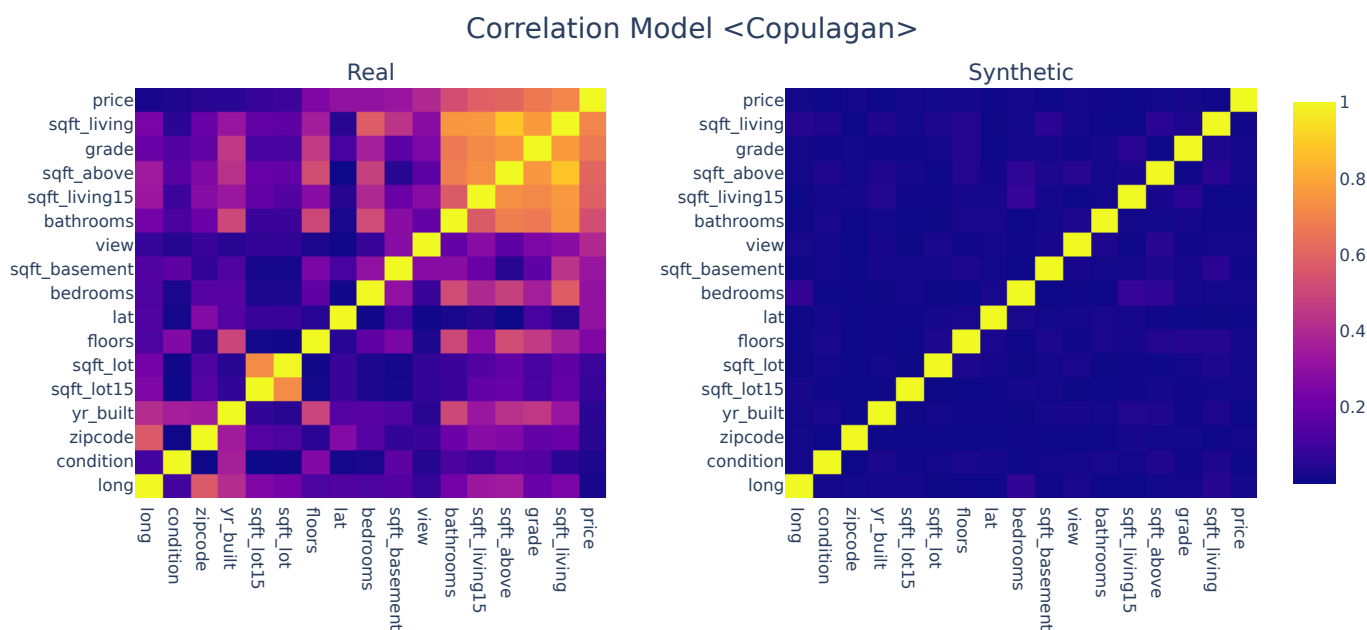


Figura 4.1: Correlación de conjunto Real y Modelo: copulagan

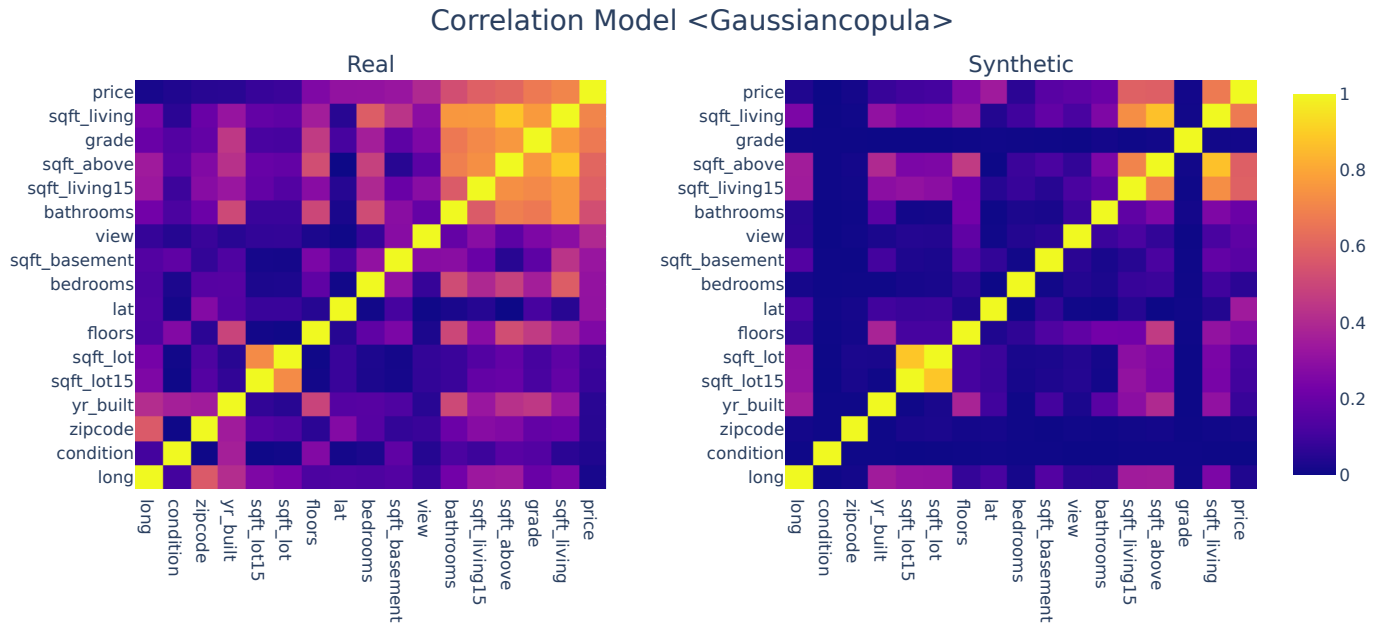


Figura 4.2: Correlación de conjunto Real y Modelo: gaussiancopula

Cabe resaltar 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. Esto se aprecia en las figuras Figura A.6 Correlación de conjunto Real y Modelo: smote-enc y Figura A.7 Correlación de conjunto Real y Modelo: tddpm_mlp, donde se contrastan los datos reales con los generados por los modelos smote-enc y tddpm_mlp, respectivamente. Ambos modelos ostentan puntajes por encima del 90 %, y la correspondencia visual entre los datos reales y los generados es notoriamente alta en ambos casos.

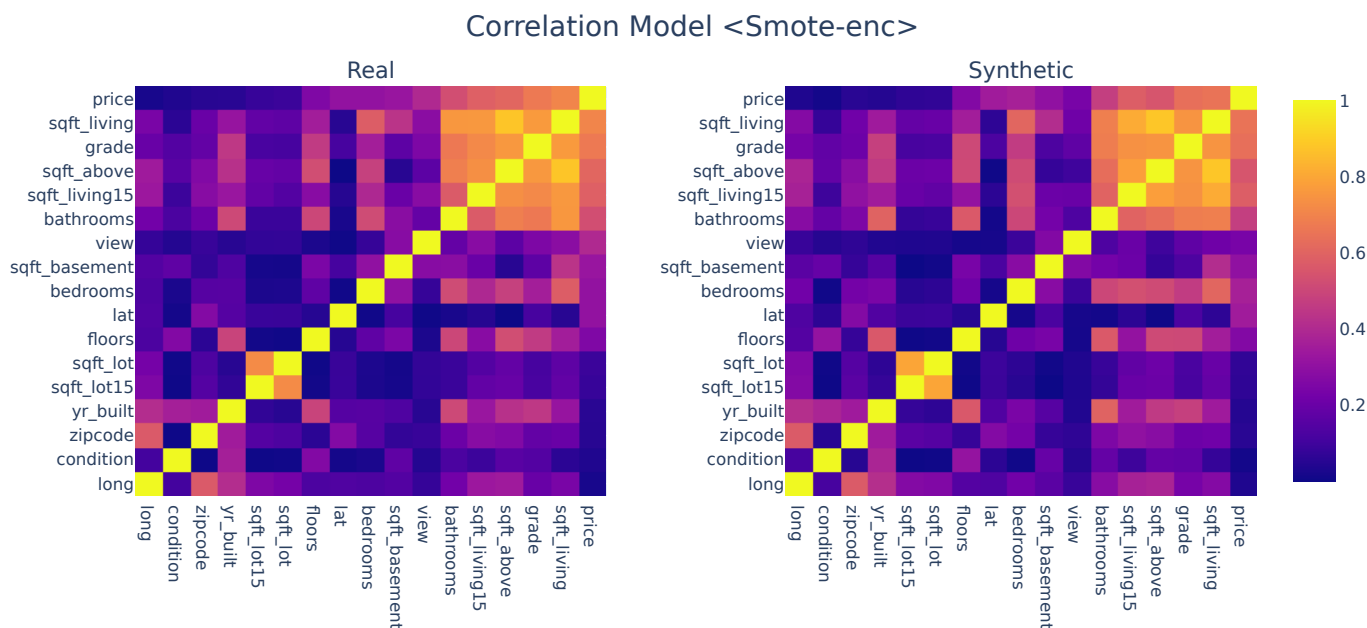


Figura 4.3: Correlación de conjunto Real y Modelo: smote-enc

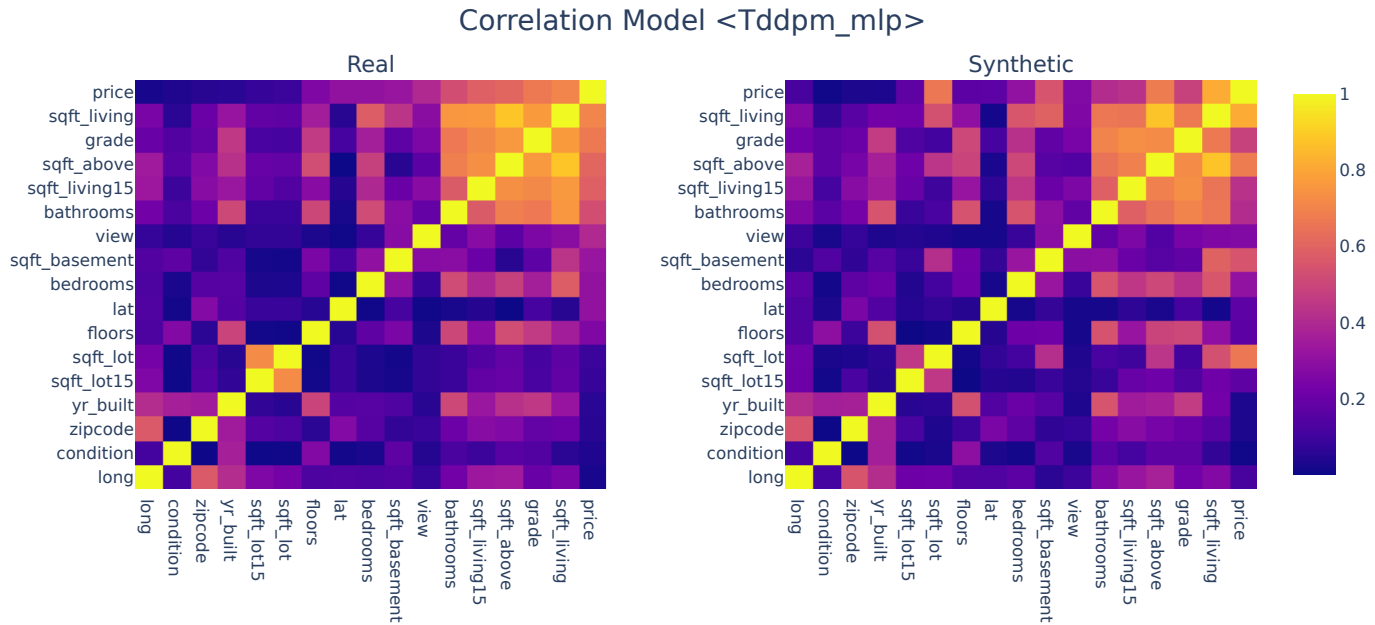


Figura 4.4: Correlación de conjunto Real y Modelo: tddpm_mlp

En la evaluación mediante SDMetrics y en la comparación visual a través de la correlación de Wise, 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.

Revisión de Columnas

La Tabla 4.2 evidencia la predominancia del modelo TDDPM en términos de cobertura de los diferentes valores en general, aunque se presentan casos en los que ambos modelos no logran una cobertura completa. Por ejemplo, en la columna correspondientes a *bedrooms*, TDDPM solo logra una cobertura que supera el 71.8 %, sin embargo, aún así supera a SMOTE. Por su parte, SMOTE solo alcanzan una cobertura del 51 % en la misma columna.

Tabla 4.2: Evaluación de Cobertura Categoría-Rango para Modelos SMOTE-ENC y TDDPM_MLP, 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

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-ENC y TDDPM_MLP, 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

En la revisión de las columnas de los conjuntos de datos completos, se puede observar en la lista Sección A.2 Lista completa de figura pairwise kingcounty una similitud entre los tres conjuntos analizados: Real, SMOTE y TDDPM. Sin embargo, también se pueden notar diferencias significativas. Es importante 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, tal como ocurre en los casos de bathrooms, sqft_lot, sqft_above, price, sqft_living, sqft_basement, yr_built, sqft_living15 y grade. Un ejemplo de esto se puede apreciar en la Figura 4.5.

Attribute <Grade> Distribution

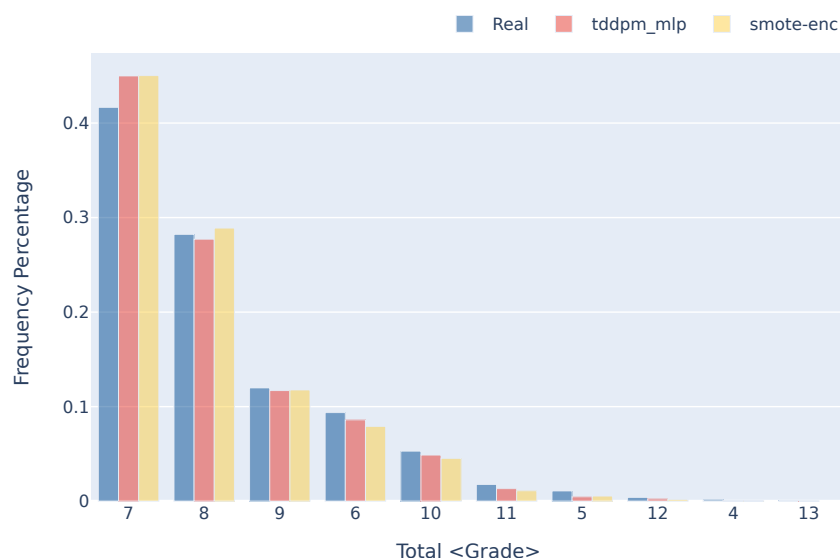


Figura 4.5: Frecuencia del campo grade en el modelo real y top2

La distribución de los atributos bedrooms, condition, view y floors en el conjunto de datos generado por el modelo TDDPM presenta una característica notable: contiene un mayor número de elementos menos frecuentes en comparación con los otros conjuntos. Tomando como ejemplo la columna *bedrooms* (véase Figura 4.6 Frecuencia del campo bedrooms en el modelo real y top2), la distribución de valores en el conjunto TDDPM difiere de la del conjunto SMOTE. En particular, se observa un incremento en la cantidad de registros correspondientes a los valores 6 y 1.

Attribute <Bedrooms> Distribution

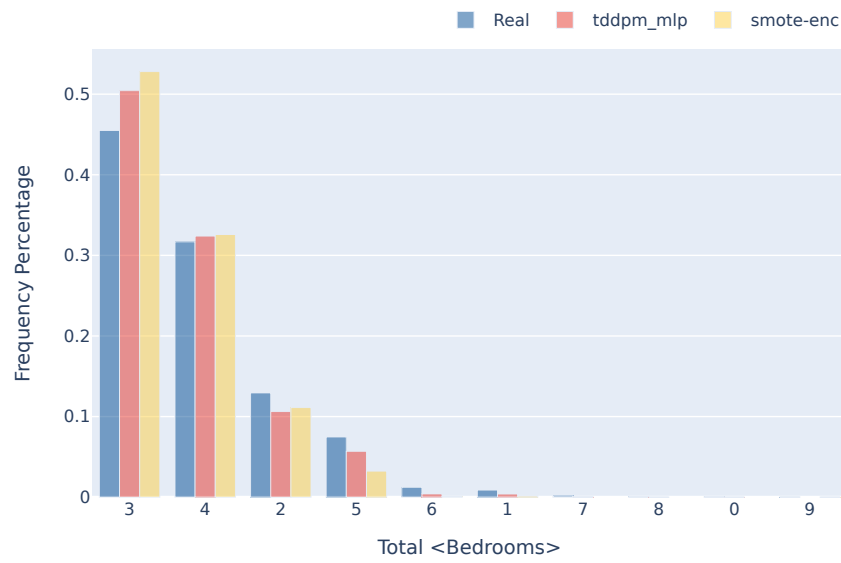


Figura 4.6: Frecuencia del campo bedrooms en el modelo real y top2

Por otro lado, en el caso de la columna *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 Figura 4.7 Frecuencia del campo *sqft lot15* en el modelo real y top2.

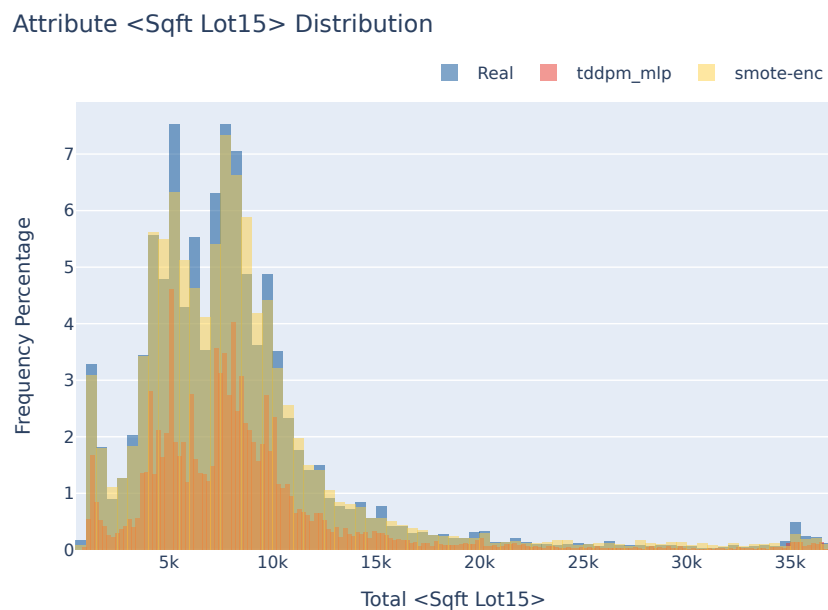


Figura 4.7: Frecuencia del campo *sqft lot15* en el modelo real y top2

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: 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±0.00e+00	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 4.5: 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±1.96e-17	2.40e-01±1.96e-17	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 4.6: 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±0.00e+00	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

Tabla 4.7: Ejemplos para el modelo smote-enc, minimo

Variable/Distancia	0	0.00e+00	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

Tabla 4.8: Ejemplos para el modelo tddpm_mlp, minimo

Variable/Distancia	0	9.90e-03	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

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 4.9: 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±1.11e-16	8.18e-01±0.00e+00	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±1.11e-16	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 4.10: 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±0.00e+00	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±1.36e-16	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 4.11: 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±5.55e-17	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±3.93e-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

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.

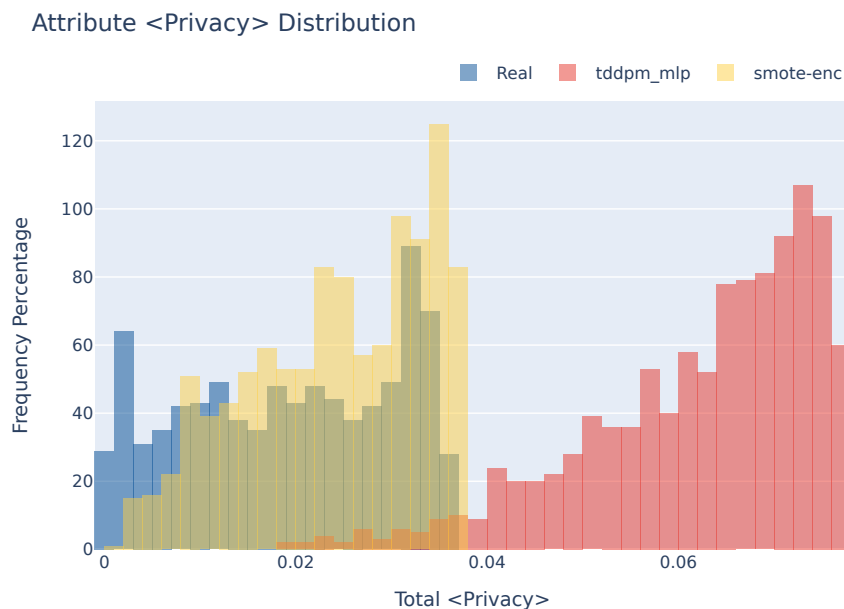


Figura 4.8: Frecuencia del campo privacy en el modelo real y top2

4.2. Economicos

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

4.2.1. Reportes - Conjunto A

Para el conjunto A, como se muestra en la Tabla 4.12, 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 4.12: 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

Es interesante observar que, para el percentil 1 y el 5, el modelo TDDPM muestra que la cercanía de los registros más cercanos es predominante al comparar el conjunto sintético y el conjunto de retención (*Hold*). Este comportamiento no se observa con ninguna otra comparación. También se nota que las diferencias mínimas alcanzan cero en los dos mejores modelos (TDDPM y SMOTE), y que los valores de distancia son muy pequeños. En el percentil 5, TDDPM registra una distancia de $4,48 \times 10^{-9}$.

Tabla 4.13: 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 4.14: 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±0.00e+00	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 4.15: 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 4.16: 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 4.17: 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±4.25e-18	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 4.18: 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±8.47e-22	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

En las tablas detalles de cobertura 4.19 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 4.19: 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

Ambos modelos muestran un buen rendimiento en cuanto a la forma y la distribución de los datos, como se evidencia en la Tabla 4.20.

Tabla 4.20: 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

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*.

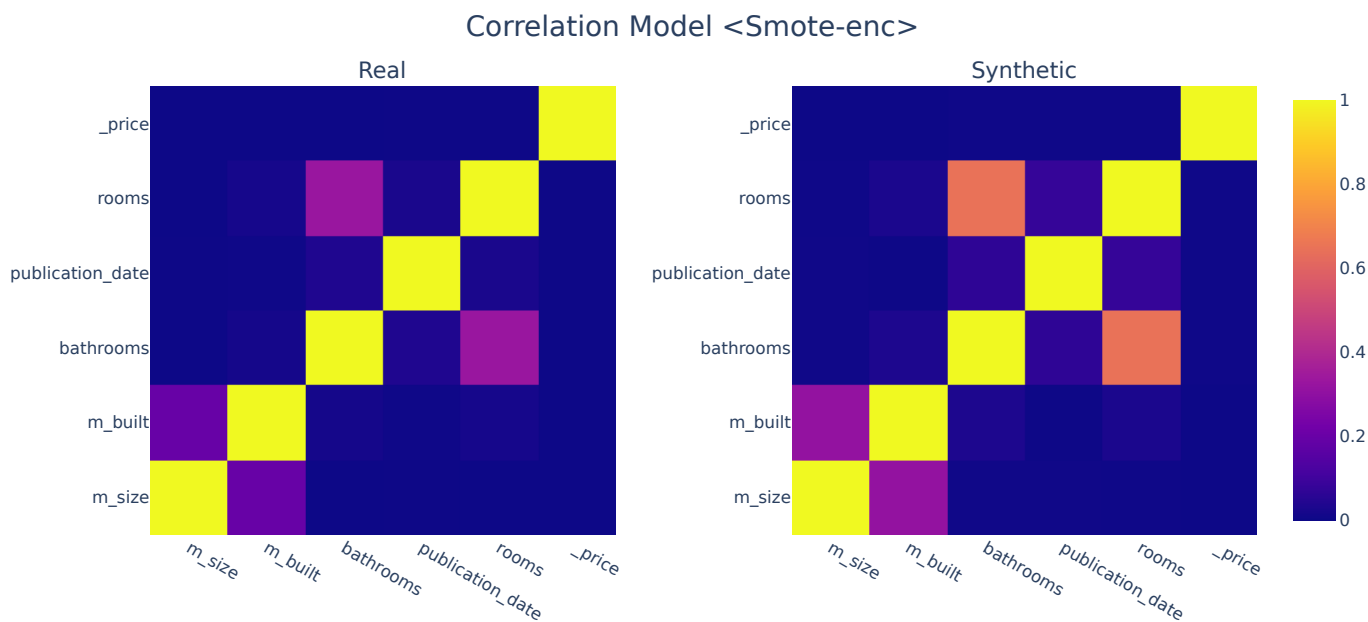


Figura 4.9: Correlación de conjunto Real y Modelo: smote-enc

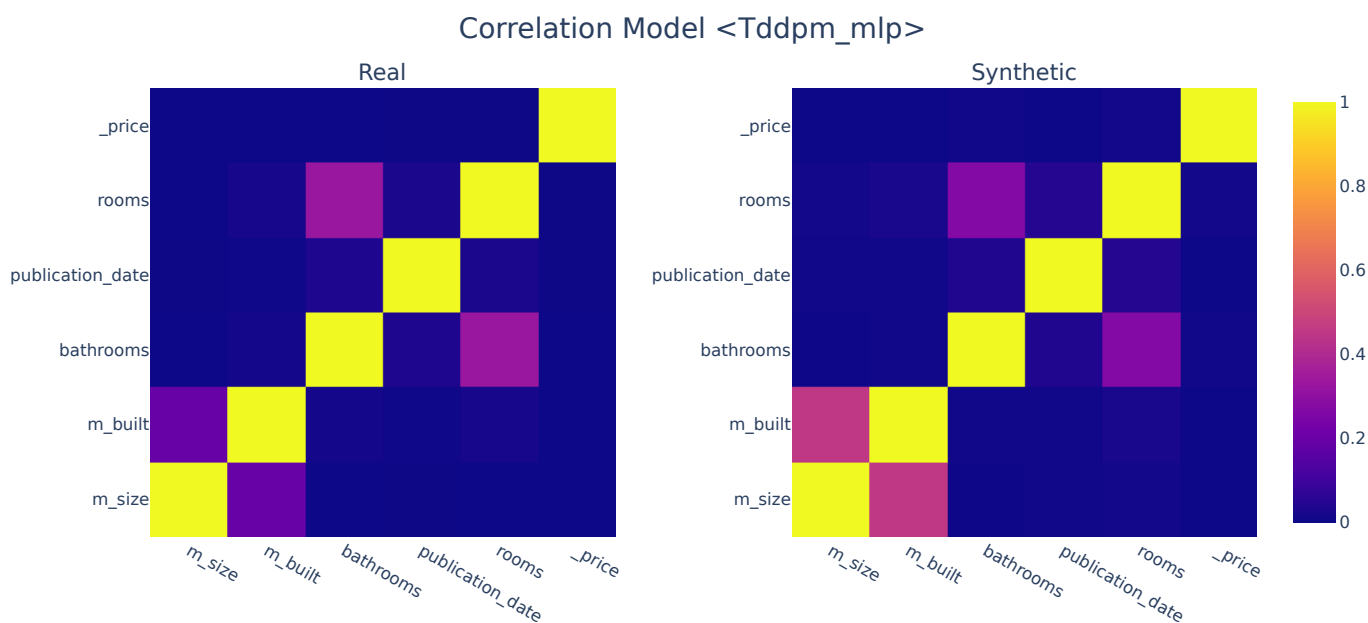


Figura 4.10: Correlación de conjunto Real y Modelo: tddpm_mlp

4.2.2. Reportes - Conjunto B

Iniciaría contrastando los resultados entre ambos conjuntos para el modelo TDDPM La Tabla 4.21 muestra mejores *Score*, *Coverage*, *Column Shape* y *Column Pair Trends* comparadas con la Tabla 4.12. 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 4.21: 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

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 4.27 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 4.22: 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 4.23: 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 4.24: 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

Se puede observar que la distancia del segundo registro es al menos tres veces la distancia del primer registro.

Tabla 4.25: 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 4.26: 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 4.27: 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

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 4.28: 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

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 4.29: 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

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*.

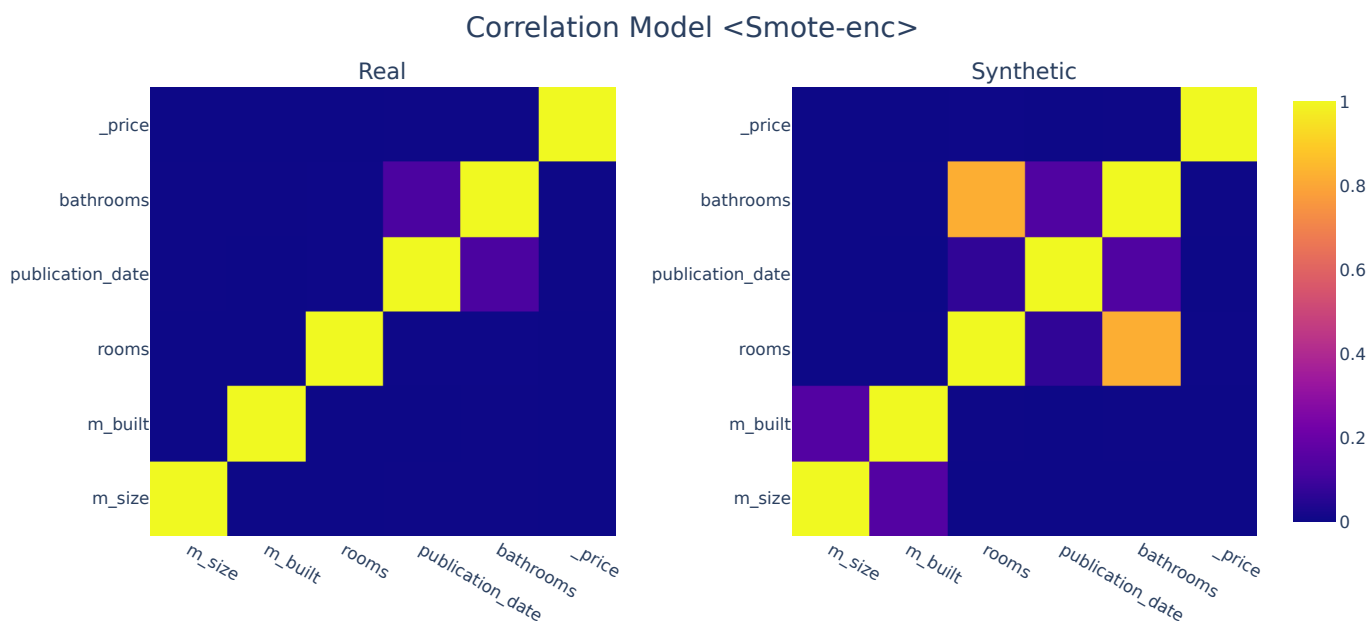


Figura 4.11: Correlación de conjunto Real y Modelo: smote-enc

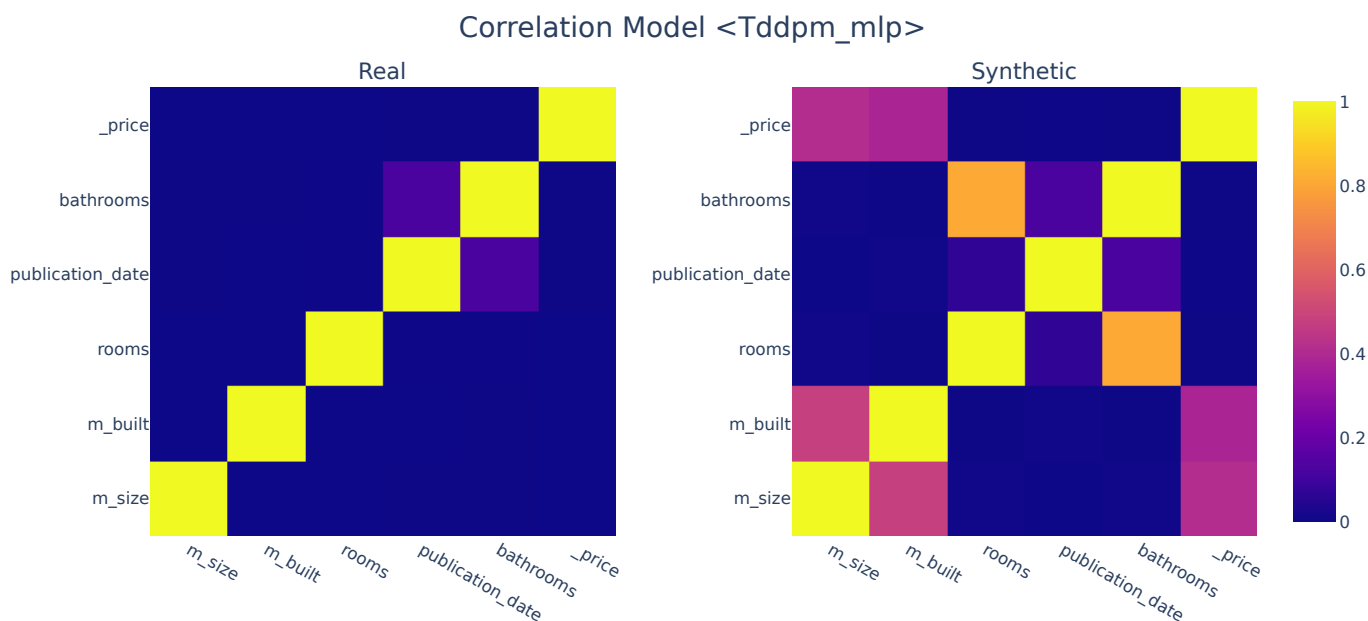


Figura 4.12: Correlación de conjunto Real y Modelo: tddpm_mlp

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 generó 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 generar textos coherentes basados en las entradas proporcionadas.

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 proporciona 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.

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.

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. Discusión

Los modelos de generación de texto están en pleno auge. Recientemente han emergido modelos como **GPT-4** [30], **Llama** [4], **Palm2** [14] y **Falcon** [2], entre otros. El modelo **Chinchilla** [22] 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 efectividad 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] Papers with code - ImageNet benchmark (image classification).
- [7] SDMetrics.
- [8] Stable diffusion public release.
- [9] Angeela Acharya, Siddhartha Sikdar, Sanmay Das, and Huzefa Rangwala. GenSyn: A multi-stage framework for generating synthetic microdata using macro data sources.
- [10] Accountability Act. Health insurance portability and accountability act of 1996. 104:191.
- [11] Kiran Adnan and Rehan Akbar. An analytical study of information extraction from unstructured and multidimensional big data. 6:1–38. Publisher: Springer.
- [12] Akim. TabDDPM: Modelling tabular data with diffusion models. original-date: 2022-10-02T23:01:07Z.
- [13] Ewa Andrejczuk, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, and Yasemin Altun. Table-to-text generation and pre-training with TabT5.
- [14] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, 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.
- [15] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators.
 - [16] Peter Bruce, Andrew Bruce, and Peter Gedeck. *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media.
 - [17] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. 16:321–357.
 - [18] Sabrina De Capitani Di Vimercati, Sara Foresti, Giovanni Livraga, and Pierangela Samarati. Data privacy: Definitions and techniques. 20(6):793–817. Publisher: World Scientific.
 - [19] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media.
 - [20] 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.
 - [21] Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. TaPas: Weakly supervised table parsing via pre-training.
 - [22] 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.
 - [23] HARLFOXEM Kaggle. House sales in king county, USA.
 - [24] Mihir Kale and Abhinav Rastogi. Text-to-text pre-training for data-to-text tasks.
 - [25] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Overview — SDV 0.18.0 documentation.
 - [26] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models.

- [27] 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.
- [28] Dan Milmo and Dan Milmo Global technology editor. Google v microsoft: who will win the AI chatbot race?
- [29] OpenAI. ChatGPT: a large language model trained by OpenAI.
- [30] OpenAI. GPT-4 technical report.
- [31] Stuart L Pardau. The california consumer privacy act: Towards a european-style privacy regime in the united states. 23:68. Publisher: HeinOnline.
- [32] 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.
- [33] David Pujol, Amir Gilad, and Ashwin Machanavajjhala. PreFair: Privately generating justifiably fair synthetic data.
- [34] Protection Regulation. Regulation (EU) 2016/679 of the european parliament and of the council. 679:2016.
- [35] Aivin V Solatorio and Olivier Dupriez. REaLTabFormer: Generating realistic relational and tabular data using transformers.
- [36] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. 32.
- [37] 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.
- [38] 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

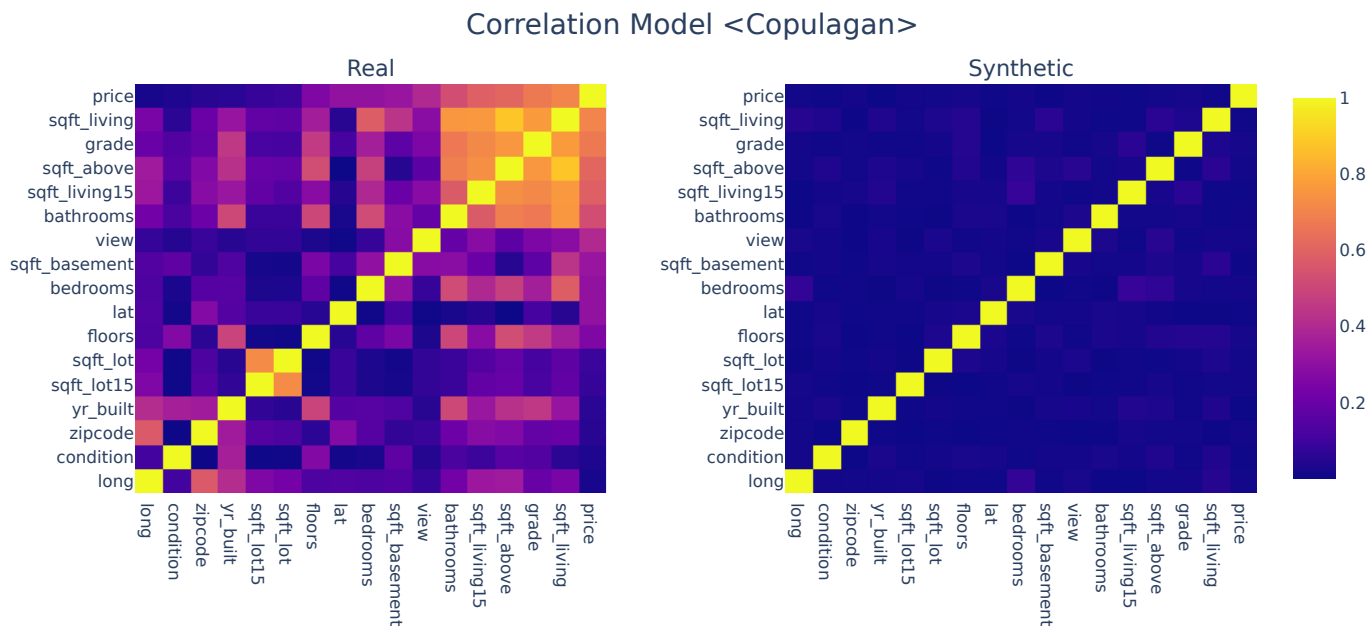


Figura A.1: Correlación de conjunto Real y Modelo: copulagan

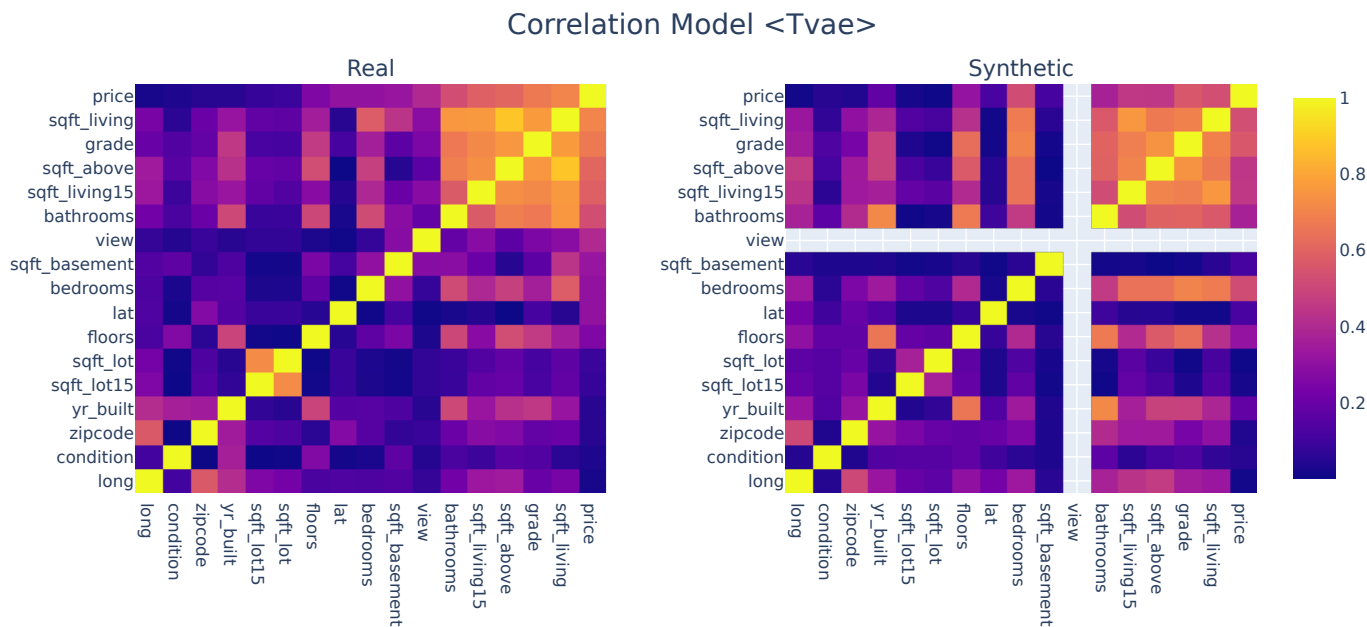


Figura A.2: Correlación de conjunto Real y Modelo: tvae

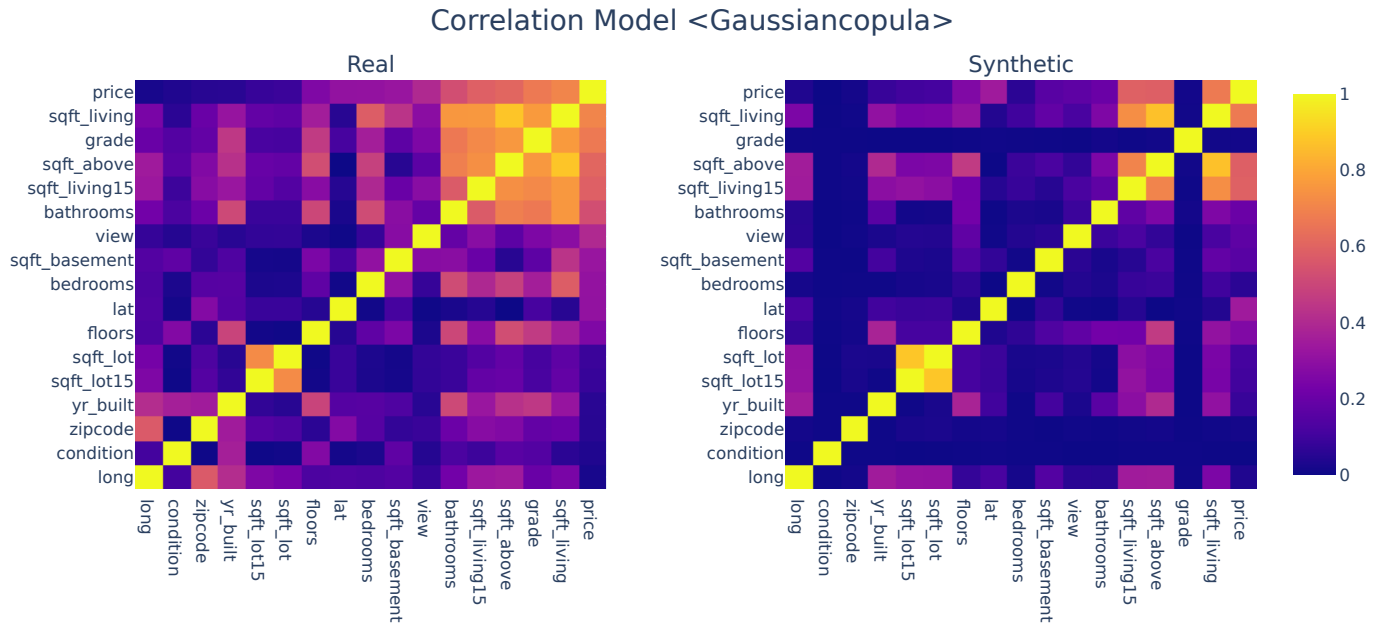


Figura A.3: Correlación de conjunto Real y Modelo: gaussiancopula

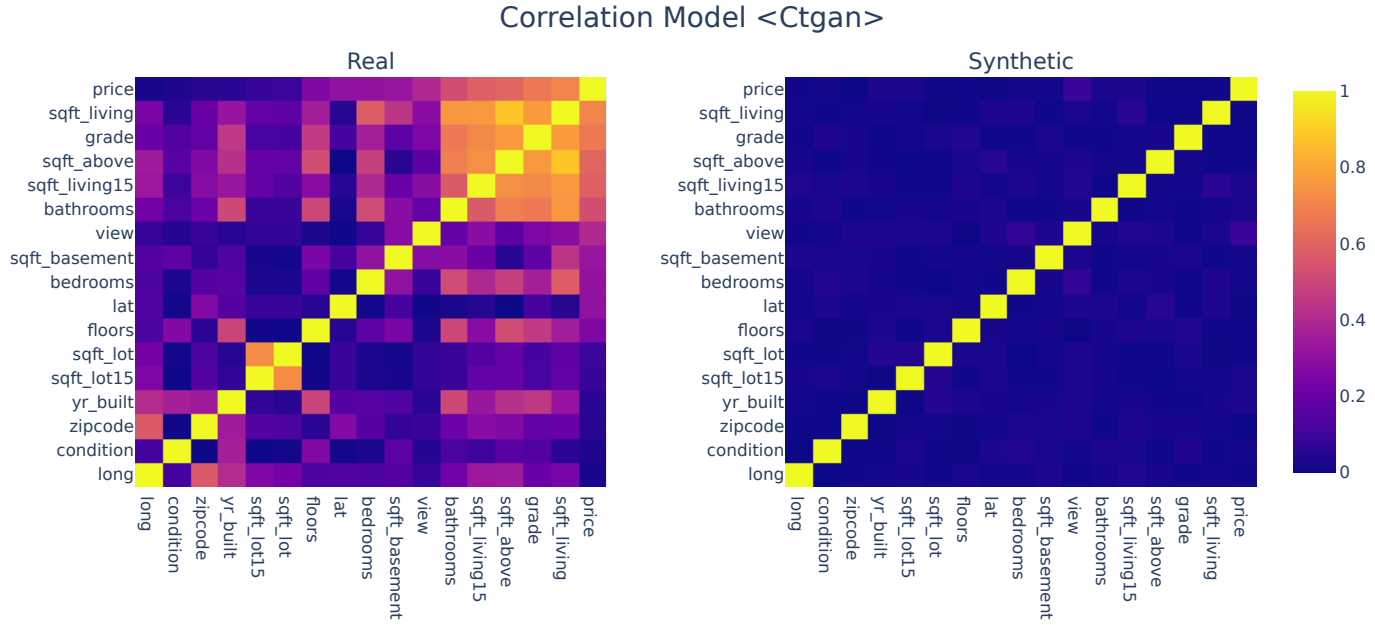


Figura A.4: Correlación de conjunto Real y Modelo: ctgan

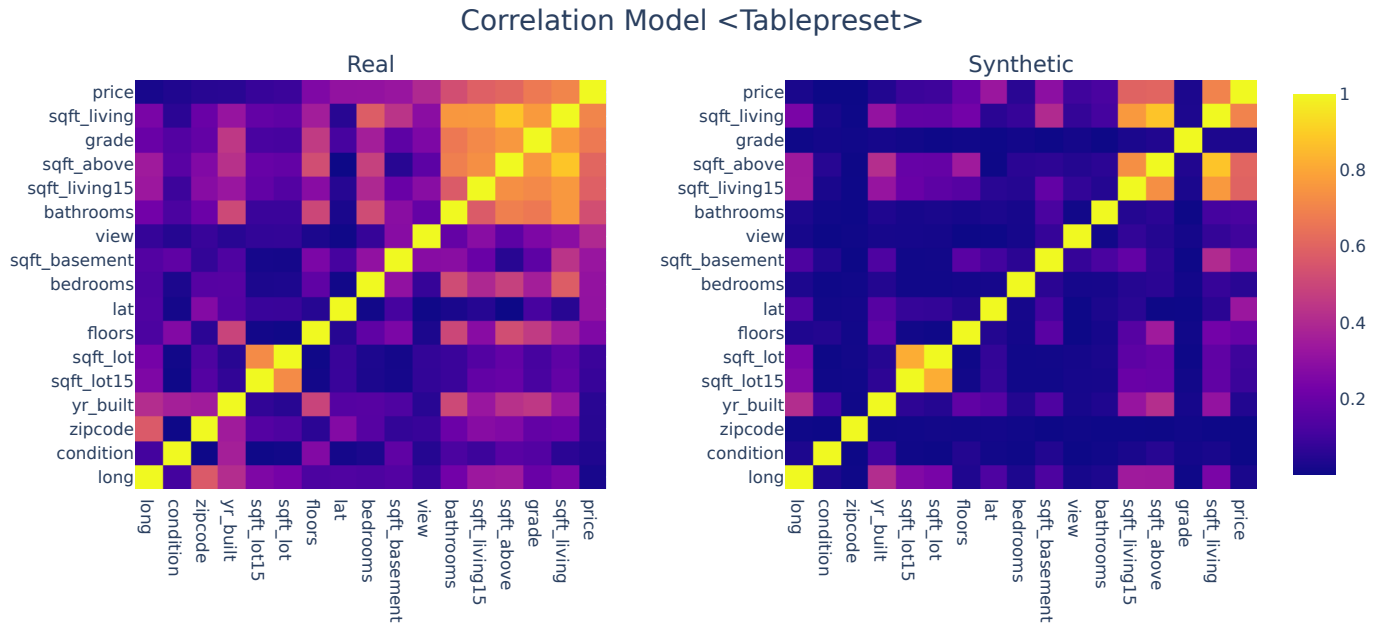


Figura A.5: Correlación de conjunto Real y Modelo: tablepreset

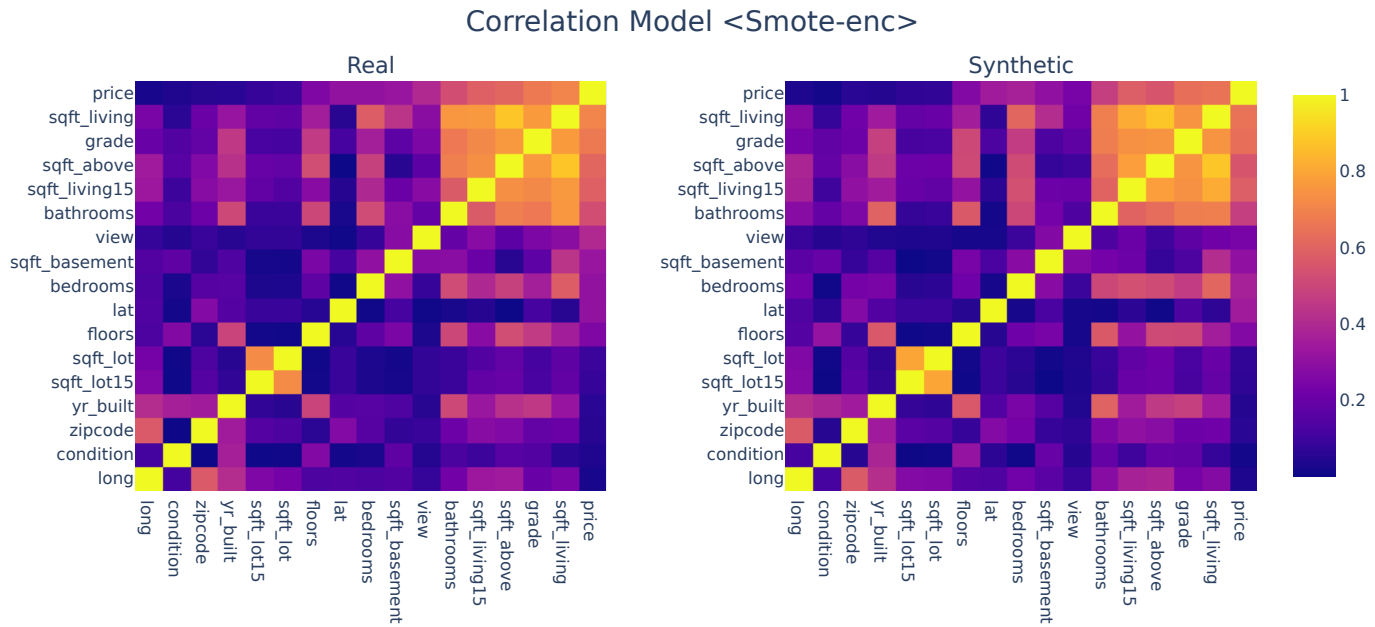


Figura A.6: Correlación de conjunto Real y Modelo: smote-enc

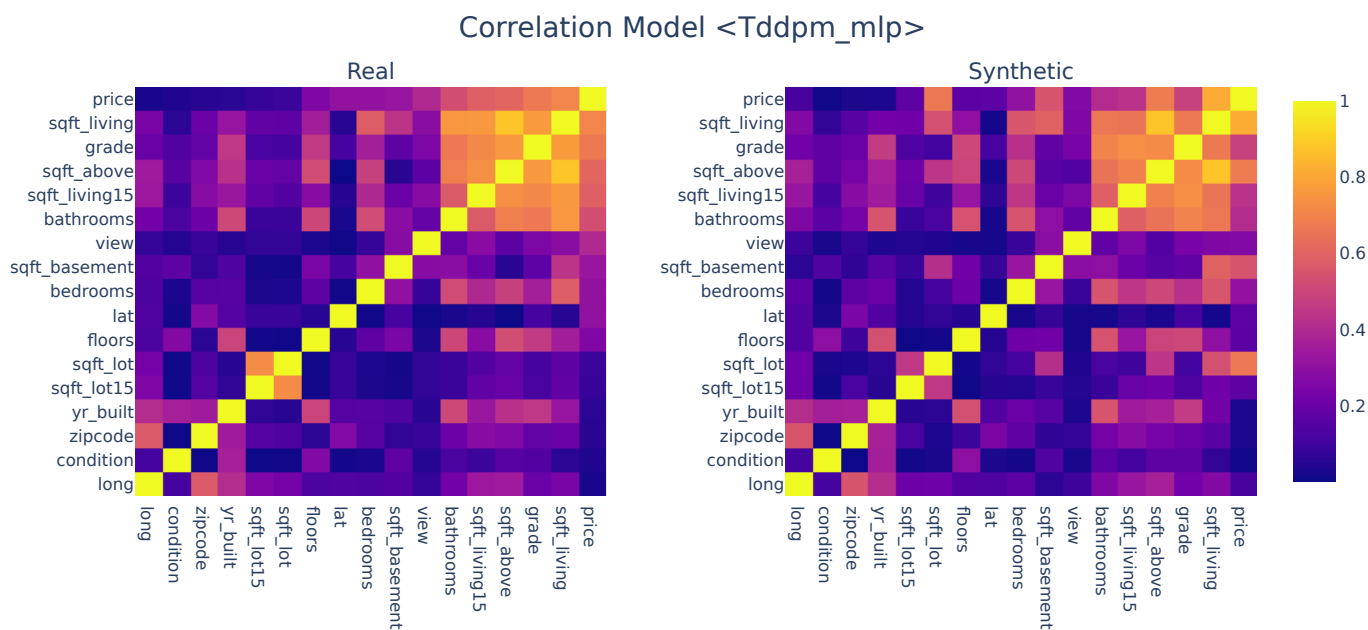


Figura A.7: Correlación de conjunto Real y Modelo: tddpm_mlp

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

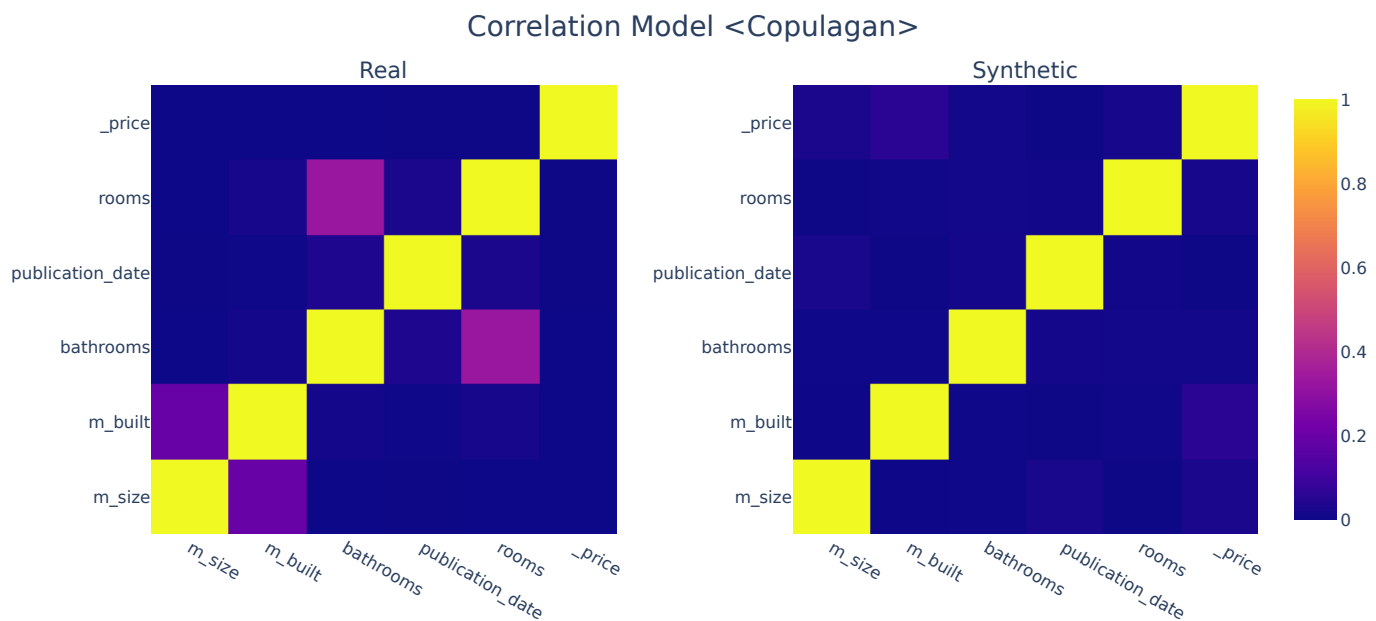


Figura A.8: Correlación de conjunto Real y Modelo: copulagan

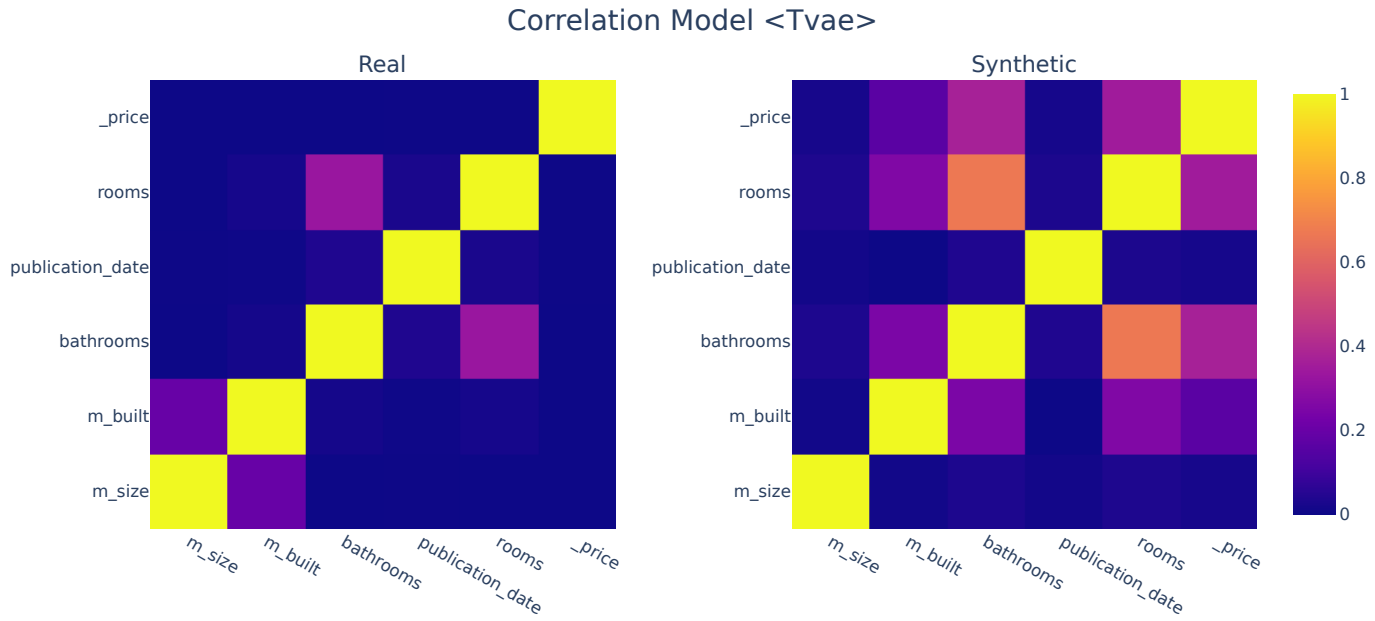


Figura A.9: Correlación de conjunto Real y Modelo: tvae

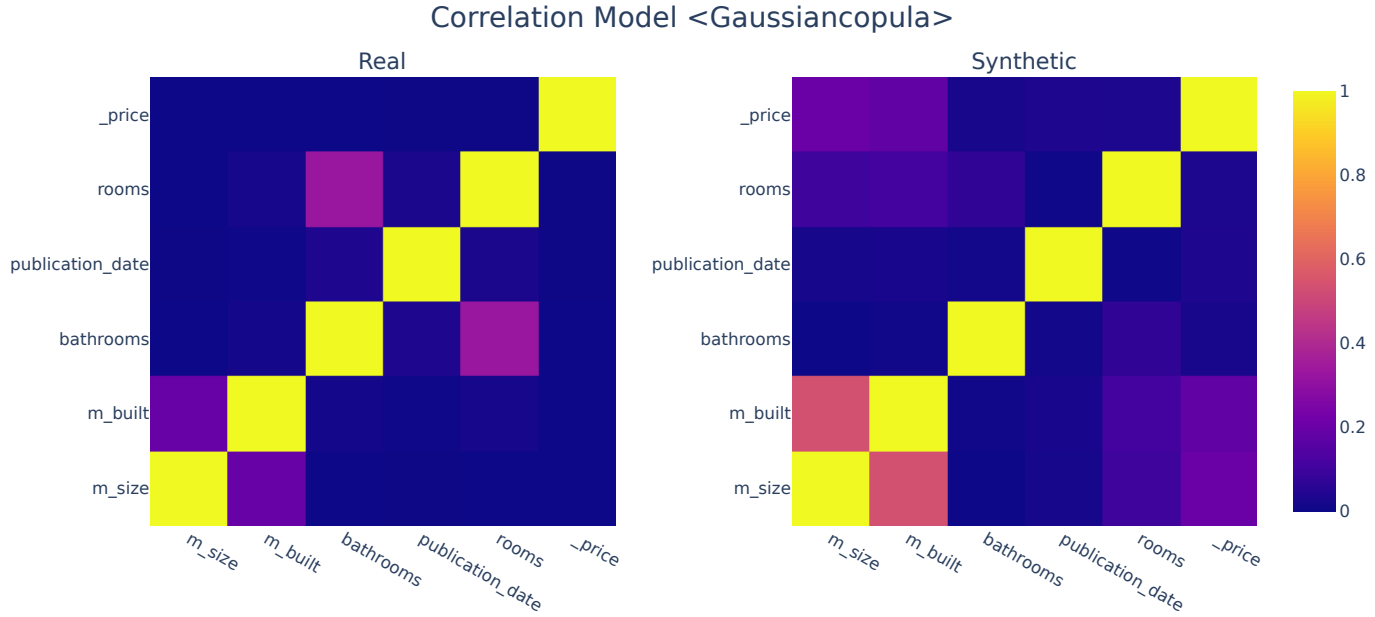


Figura A.10: Correlación de conjunto Real y Modelo: gaussiancopula

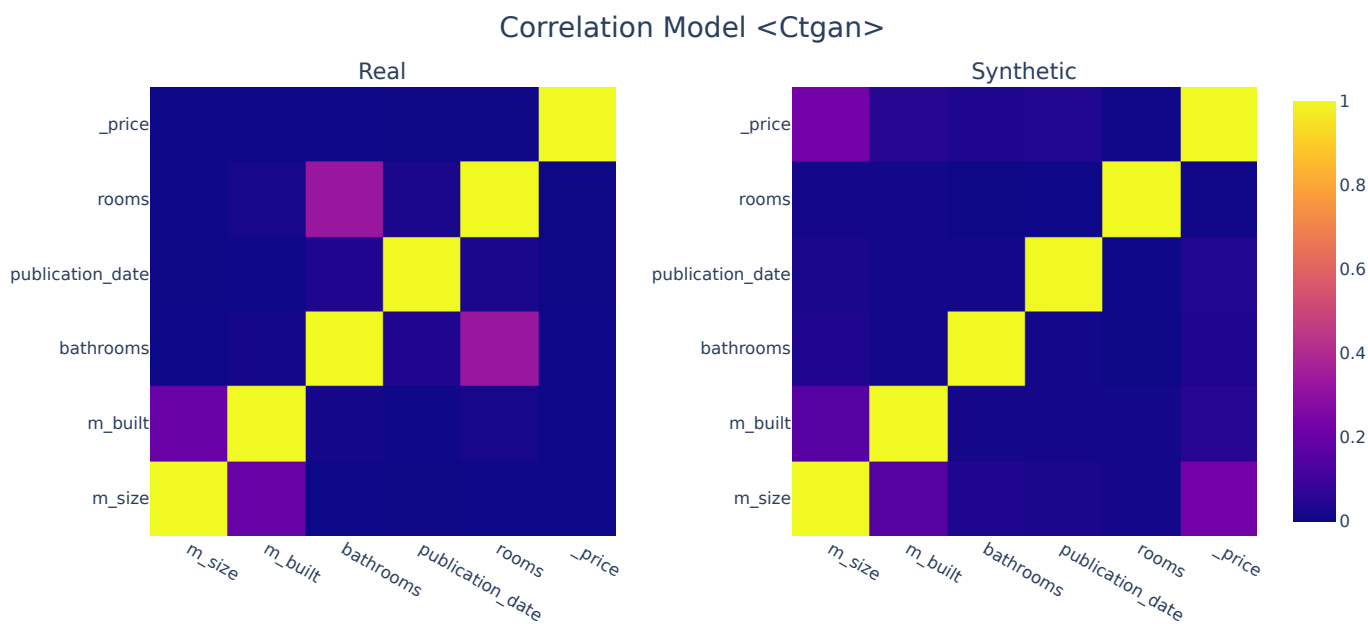


Figura A.11: Correlación de conjunto Real y Modelo: ctgan

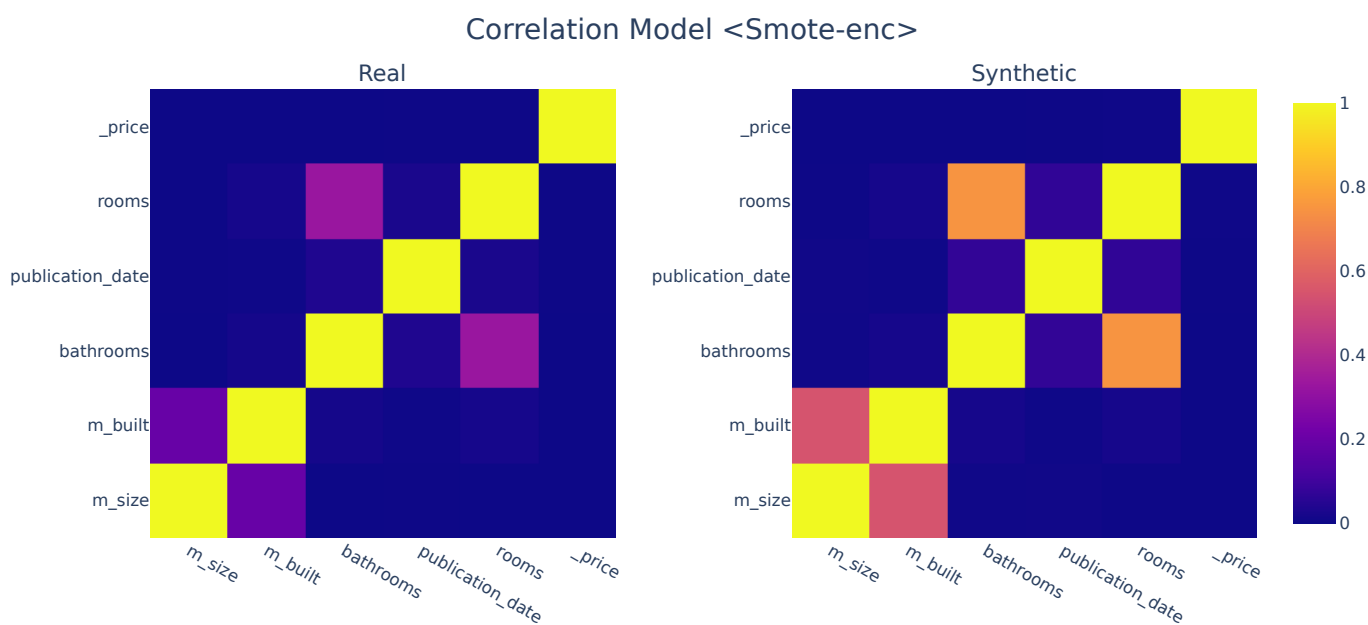


Figura A.12: Correlación de conjunto Real y Modelo: smote-enc

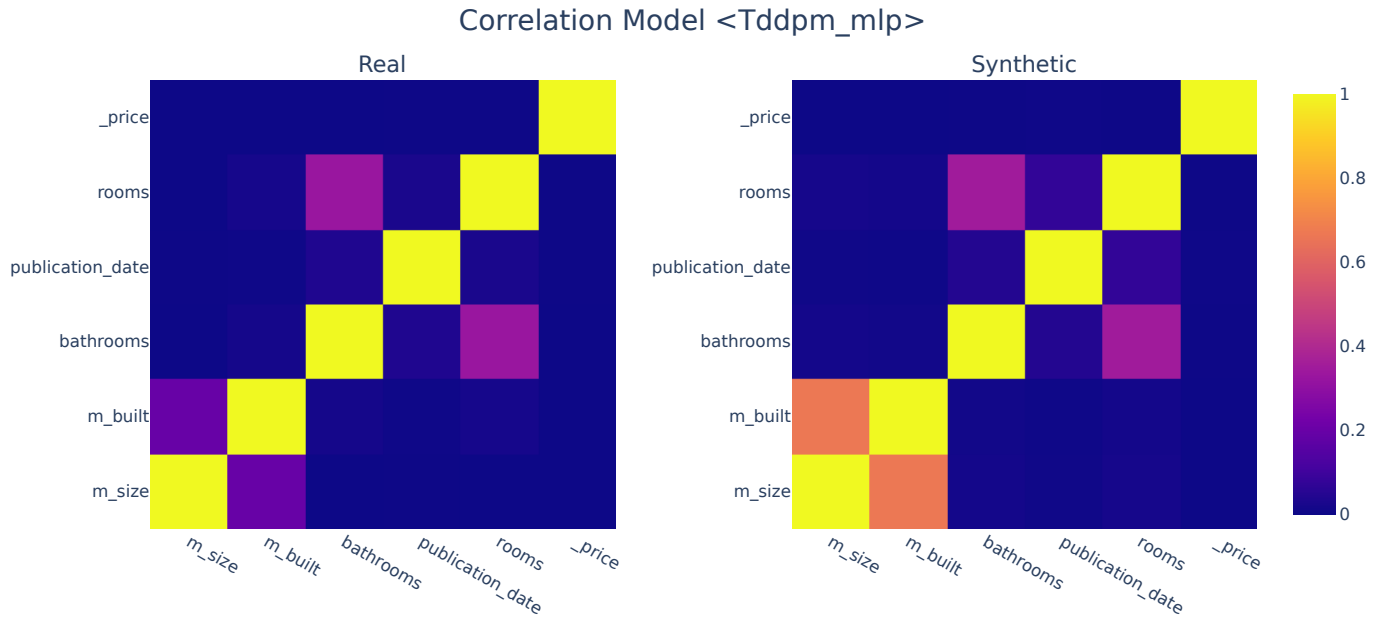


Figura A.13: Correlación de conjunto Real y Modelo: tddpm_mlp

A.6. Figuras de correlación Economicos - Conjunto B

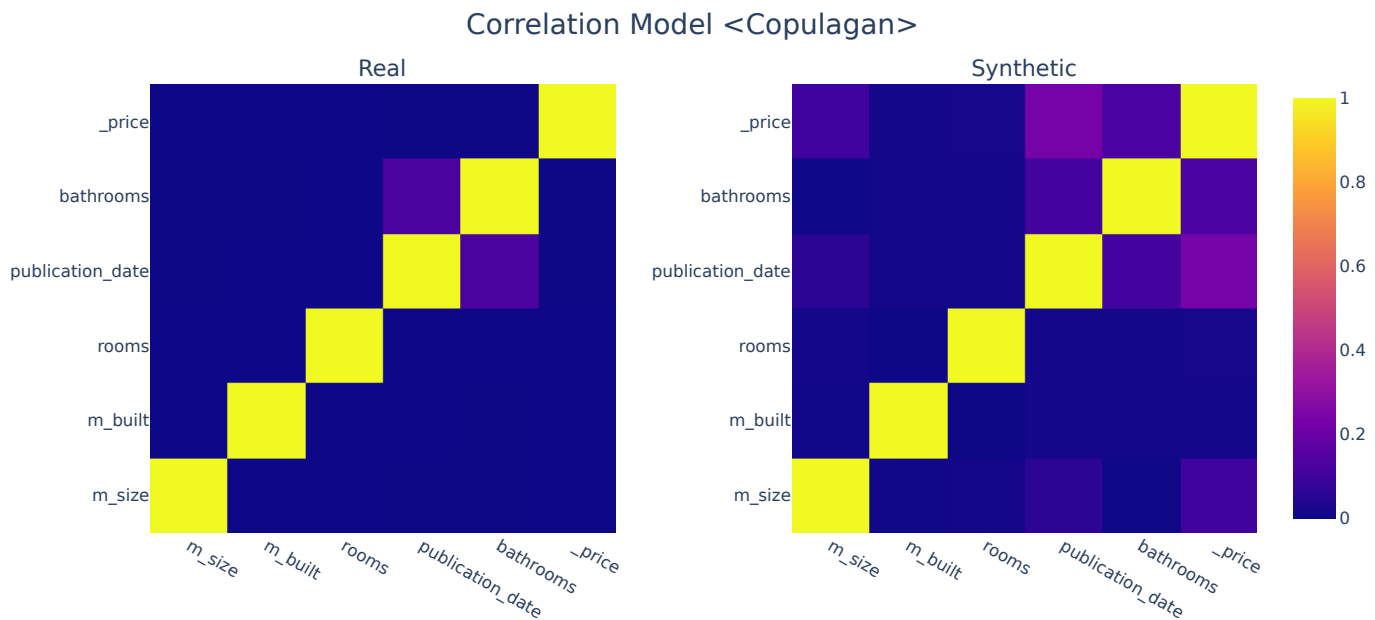


Figura A.14: Correlación de conjunto Real y Modelo: copulagan

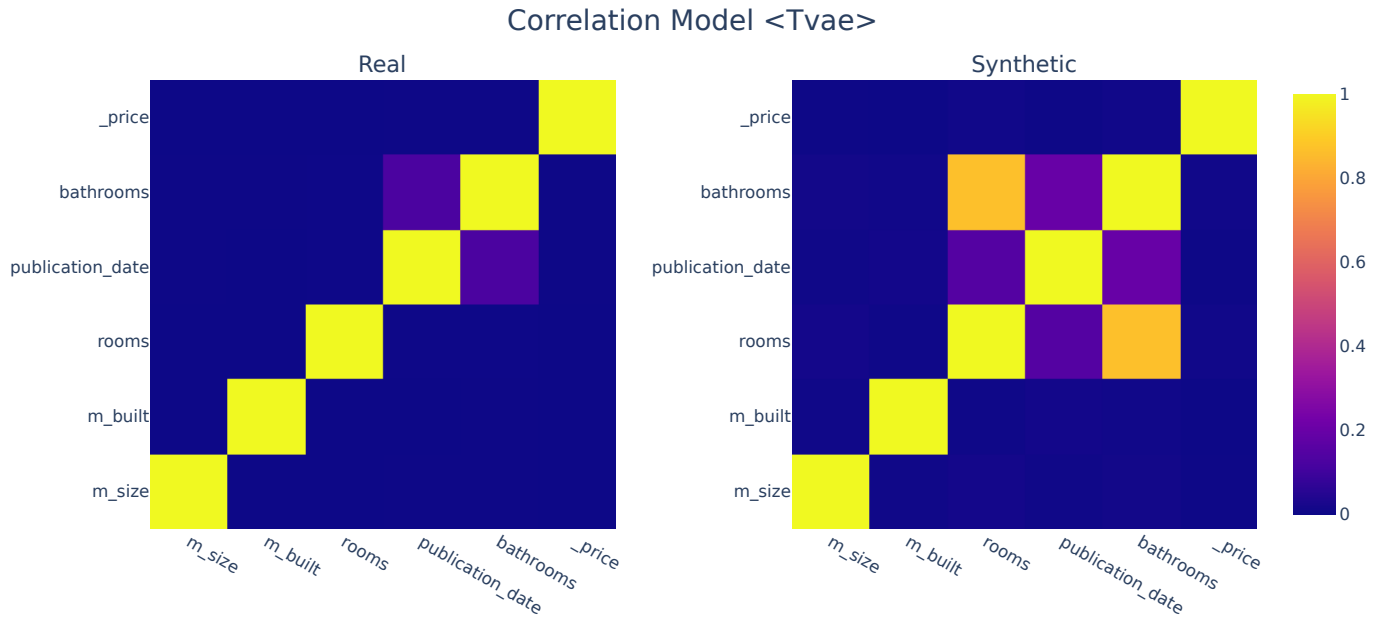


Figura A.15: Correlación de conjunto Real y Modelo: tvae

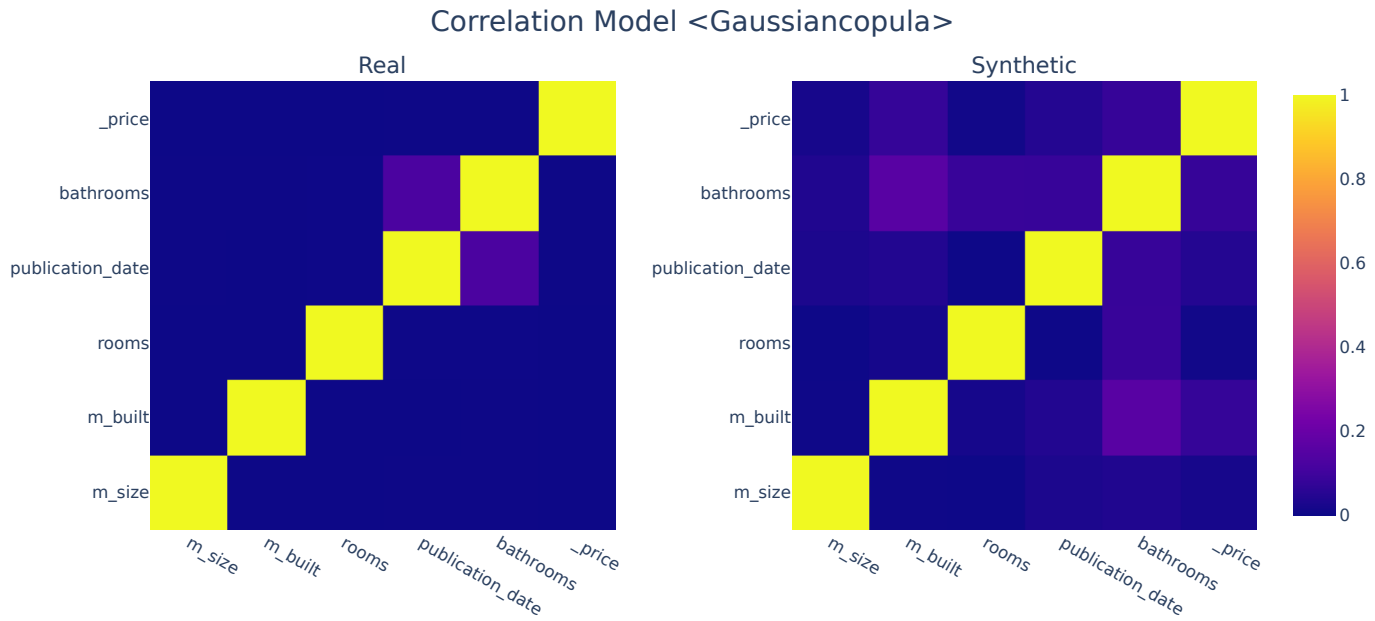


Figura A.16: Correlación de conjunto Real y Modelo: gaussiancopula

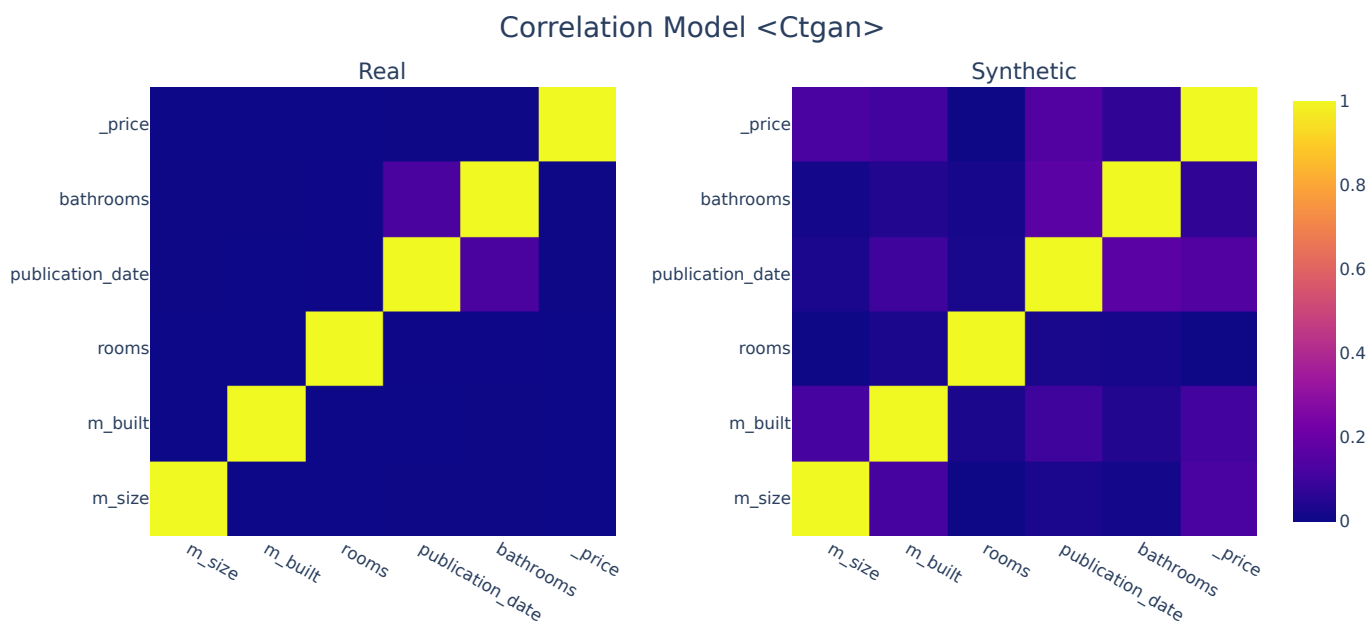


Figura A.17: Correlación de conjunto Real y Modelo: ctgan

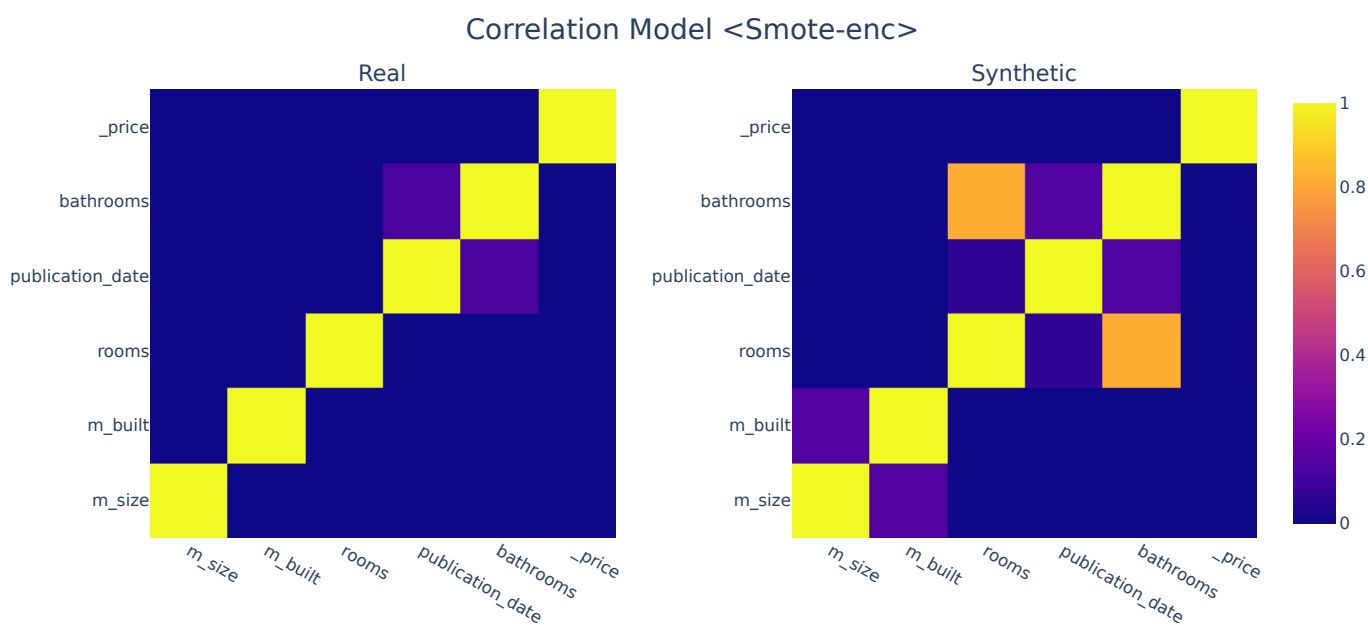


Figura A.18: Correlación de conjunto Real y Modelo: smote-enc

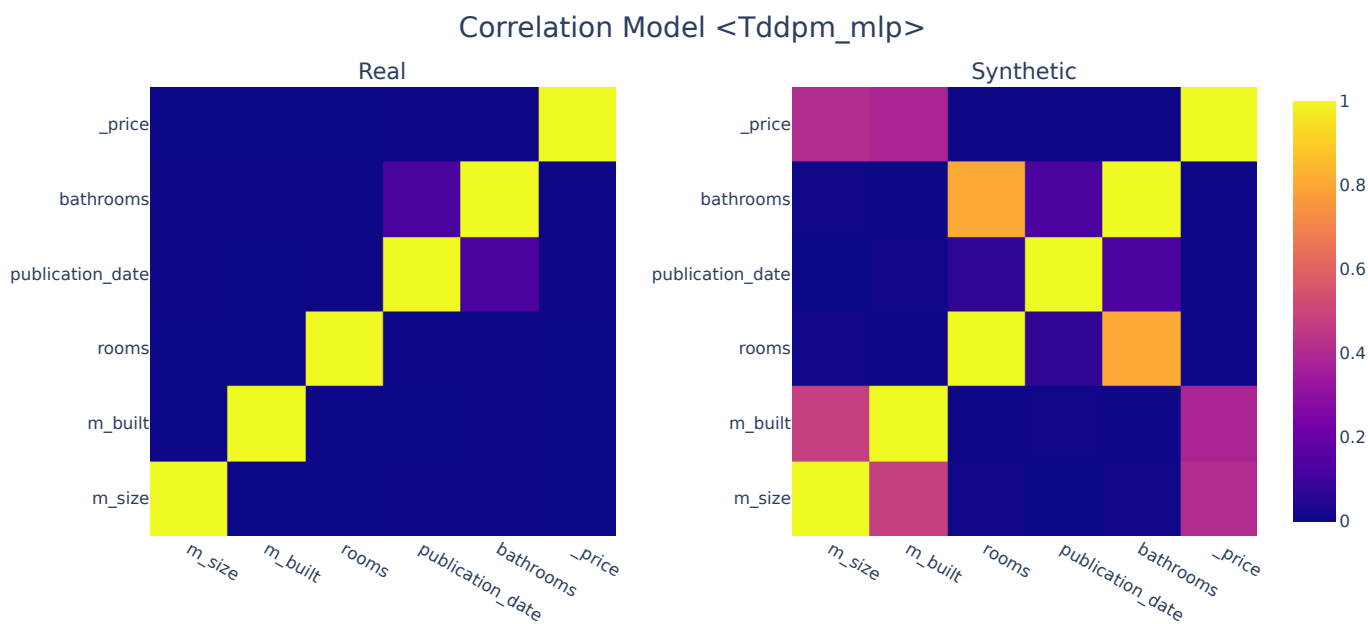


Figura A.19: Correlación de conjunto Real y Modelo: tddpm_mlp

A.6.1. Conjunto A