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

DATA SINTÉTICA PRIVADA, EJECUCIÓN Y EVALUACIONES DE MODELOS

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN

GERARDO JORGE VILLARROEL GONZÁLEZ

PROFESOR GUÍA:
ANDRES ABELIUK

MIEMBROS DE LA COMISIÓN:
FEDERICO OLMEDO B
MATÍAS TORO I
CARLOS BUIL

SANTIAGO DE CHILE
2023

Resumen

En la era de la información, la generación y análisis de conjuntos de datos es crucial para avanzar en diversas disciplinas. Sin embargo, la privacidad y la utilidad de los datos se han convertido en consideraciones esenciales. Esta tesis aborda estas preocupaciones mediante el desarrollo de un mecanismo para generar conjuntos de datos sintéticos estructurados, que adicionalmente incluyen texto, y su evaluación comparativa con los datos originales. Este estudio es de gran importancia, ya que propone un enfoque práctico para mantener el equilibrio entre la utilidad de los datos y la privacidad de la información personal.

Se hace uso de múltiples técnicas y modelos generativos, como Tddpm y Smote, para la creación de estos conjuntos, empleando recursos como los conjuntos de datos de *King County* y Económicos.cl. Este trabajo también detalla la metodología implementada basada en *Synthetic Data Vault* (SDV), extendida para permitir fases intermedias de almacenamiento de modelos y resultados de evaluación.

En términos de resultados, el estudio reveló la eficacia de los modelos Tddpm y Smote en la generación de datos sintéticos que mostraron similitud con los datos originales. No obstante, se detectaron diferencias notables en aspectos como la cobertura, distribución y privacidad entre los conjuntos de datos generados, uno de los factores analizados corresponde al tratamiento de los valores nulos.

Las conclusiones destacan la utilidad y privacidad que ofrece el modelo Tddpm en la generación de datos sintéticos, proporcionando así una importante contribución al campo de la generación de datos sintéticos. Se identifican limitaciones en la metodología actual y se plantean oportunidades para futuras investigaciones, subrayando la creciente importancia de los modelos de generación de texto y la necesidad de evaluar la privacidad en este contexto.

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

Agradecimientos

Las primeras palabras deben ser para Brunis Medel. Durante 16 años, has sido la persona que más paciencia ha mostrado conmigo, y quienes me conocen sabrán que esto es toda una odisea en sí misma. Te has esforzado más allá de lo que la capacidad humana generalmente permite para proporcionarme el tiempo y la tranquilidad necesarios para la construcción de este proyecto. Has cuidado de nuestras hijas, Brunis Villarroel y Sofía Villarroel, quienes son nuestro legado conjunto. Este proyecto es también un testamento para ellas.

Mi sincero agradecimiento a Yuvaraj Sankaran y Felipe Castillo, los arquitectos detrás de mi aceptación universitaria. Vuestras cartas de recomendación fueron los planos que me permitieron abrir la puerta a esta oportunidad. Anhelo poder contar con vuestra valiosa ayuda para mis futuros emprendimientos académicos.

Deseo expresar mi gratitud a mi guía en este trabajo, Andres Abeliuk. Tus correcciones y consejos han sido las herramientas precisas que necesitaba para pulir y optimizar este trabajo.

Mis compañeros, Benjamín Obando, Danilo Bustos y Felipe Ávila, merecen mi reconocimiento por sus aportes esenciales que ayudaron a pulir y perfeccionar mi presentación. Sus consejos han sido vitales para este logro.

Por último, no puedo dejar de reconocer el aporte de ChatGPT, una entidad de inteligencia artificial. Como un robot de un cuento de Asimov, este sistema ha comenzado a mostrar capacidades nunca antes vistas en máquinas. Mi deseo es que en los milenios venideros, esta creación sea vista no solo como un legado de un ser humano, sino de todos nosotros como especie. Un descendiente que pueda ir donde nuestras limitaciones biológicas nos han cerrado las puertas, quizás acompañado por nuestros descendientes biológicos también.

A todos, con sincera gratitud y profundo respeto, gracias.

Tabla de Contenido

1. Introducción **1**

1.1. Equifax: Contexto y Restricciones 1

1.2. Objetivo 3

1.3. Estructura del documento 3

2. Revisión Bibliográfica **4**

2.1. Tipos de Datos 4

2.2. Privacidad de Datos 5

2.2.1. Tipo de datos a ser protegidos 5

2.2.2. Tipos de riesgos de divulgación 6

2.2.3. Privacidad Diferencial 7

2.2.4. k-Anonimato 7

2.2.5. Distancia al registro más cercano (DCR) 8

2.2.6. Ratio de distancia entre los vecinos más cercanos (NNDR) 8

2.2.7. Regulación de datos sintéticos 9

2.3. Generación de Datos Sintéticos 9

2.3.1. Generación de datos tabulares 10

2.3.2. Generación de texto en base de datos tabulares 10

2.4. Metricas de evaluación 12

2.4.1. SDMetrics Score 13

2.4.2. Reporte diagnóstico 14

2.4.3. Reporte de calidad 14
2.4.4. Correlación 16
2.4.5. Privacidad 16
2.4.6. Propiedades estadísticas y tecnicas de análisis 16

3. Desarrollo **19**

3.1. Recursos disponibles 19
3.1.1. Conjuntos de datos 19
3.1.2. Computación y Software 21
3.2. Desarrollo del flujo de procesamiento 23
3.3. Modelos de generación de datos 26
3.3.1. Modelos para datos tabulares 26
3.3.2. Modelos para textos 29
3.4. Privacidad y sus metricas 30
3.5. Obtención de Métricas 30
3.6. Tiempo de Ejecución 32

4. Resultados **34**

4.1. King County 35
4.1.1. SDMetrics Score 35
4.1.2. Correlación 36
4.1.3. Reporte diagnóstico 38
4.1.4. Reporte de calidad 39
4.1.5. Privacidad 43
4.1.6. Ejemplo de registros 47
4.1.7. Propiedades estadísticas 50
4.1.8. Resumen de resultados 51
4.2. Conjunto de datos proveniente de Economicos 53

4.2.1. Tratamiento de nulos en conjunto A y B 53
4.2.2. SDMetrics Score - Conjunto A 54
4.2.3. Correlación - Conjunto A 55
4.2.4. Reporte diagnóstico - Conjunto A 56
4.2.5. Reporte de calidad - Conjunto A 57
4.2.6. Privacidad - Conjunto A 58
4.2.7. Ejemplos de registros - Conjunto A 60
4.2.8. Propiedades estadísticas - Conjunto A 62
4.2.9. Resumen de resultados - Conjunto A 64
4.2.10. SDMetrics Score - Conjunto B 65
4.2.11. Correlación - Conjunto B 66
4.2.12. Reporte diagnóstico - Conjunto B 67
4.2.13. Reporte de calidad - Conjunto B 68
4.2.14. Privacidad - Conjunto B 69
4.2.15. Ejemplos de registros - Conjunto B 71
4.2.16. Propiedades estadísticas - Conjunto B 73
4.2.17. Resumen de los Resultados - Conjunto B 74

5. Conclusiones y discusión **76**

5.1. Conclusiones 76
5.2. Limitaciones 77
5.3. Discusión 78
5.4. Evaluación de objetivos y logros 79

Apéndice A. Anexos **85**

A.1. Código de entrenamiento de económicos 86
A.2. Archivo Devcontainer 87
A.3. Lista completa de figura pairwise kingcounty 88

A.4. Smote y Tddpm en KingCounty Gráficas por Columnas 91
A.5. Figuras de correlación Económicos - Conjunto A 99
A.6. Figuras de correlación Económicos - Conjunto B 102
A.7. Ejemplos de 10 Registros Generados Aleatoriamente en Descripciones Económicas A-1 107
A.8. Ejemplos de 10 Registros Generados Aleatoriamente en Descripciones Económicas 108
A.9. Estadísticos KingCounty 109
A.10. Estadísticos Económicos - Conjunto A 129
A.11. Estadísticos Económicos - Conjunto B 137
A.12. Ejemplos de código con fines de reproducibilidad 145

Índice de Tablas

2.1. Tipos de datos estructurados 5

2.2. Niveles de revelación y ejemplos 5

2.3. Tipos de Riesgos de Divulgación y sus Descripciones 6

2.4. Estado del arte en generación de datos tabulares 10

2.5. Estado del arte en generación de textos en base a datos 11

2.6. Ejemplo de tabla de entrada 11

2.7. Ejemplo de Métricas de Rendimiento para Diversos Modelos 14

2.8. Listado de conjunto estadísticos 16

3.1. Conjunto de datos King County 20

3.2. Base de datos Economicos.cl 21

3.3. Computador Usado 22

3.4. Variables de entrada para *Synthetic* 27

3.5. Modelos Tabulares Soportados 27

3.6. Metricas para campos numericos 30

3.7. Métricas para campos categóricos 32

3.8. Tiempo de ejecución para cada técnica evaluada 32

4.1. Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, King county 35

4.2. Cobertura Categoría/Rango para Modelos Smote y Tddpm, King county 38

4.3. Evaluación de Similitud de Distribución para Modelos Smote y Tddpm, King county 39

4.4. Proporción entre el más cercano y el segundo más cercano, percentil 5, King county 43

4.5. Proporción entre el más cercano y el segundo más cercano, percentil 1, King county 43

4.6. Proporción entre el más cercano y el segundo más cercano, mínimo, King county . 44

4.7. Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 5, King county 45

4.8. Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 1, King county 45

4.9. Proporción entre el más cercano y el segundo más cercano, minimo, datos king county 45

4.10. Ejemplos para el modelo smote-enc, minimo, King county (A-1) 47

4.11. Ejemplos para el modelo tddpm_mlp, minimo, King county (A-1) 48

4.12. Ejemplos para el modelo smote-enc, percentil 1, King county (A-1) 49

4.13. Propiedades estadisticas de variable bedrooms con cambio>5 %, King county (A-1) 50

4.14. Propiedades estadisticas de variable bathrooms con cambio>5 %, King county (A-1) 51

4.15. Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, Economicos 54

4.16. Cobertura Categoría/Rango para Modelos Smote y Tddpm, Economicos 56

4.17. Evaluación de Similitud de Distribución para Modelos Smote y Tddpm, Economicos 57

4.18. Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 5, Economicos 58

4.19. Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 1, Economicos 58

4.20. Distancia de registros más cercanos, minimo, datos economicos 58

4.21. Proporción entre el más cercano y el segundo más cercano, percentil 5, Economicos 59

4.22. Proporción entre el más cercano y el segundo más cercano, percentil 1, Economicos 59

4.23. Proporción entre el más cercano y el segundo más cercano, mínimo, Economicos . 59

4.24. Ejemplos para el modelo Tddpm, minimo, Economicos (A-2) 60

4.25. Ejemplos para el modelo Tddpm, percentil 1, Economicos (A-2) 60

4.26. Ejemplos para el modelo Tddpm, percentil 4, Economicos (A-1) 61

4.27. Ejemplos de texto modelo Tddpm, percentil 4, Economicos (A-1) 61

4.28. Propiedades estadisticas de variable m_size con cambio>5 %, Economicos (A-1) . 62

4.29. Propiedades estadisticas de variable county, Economicos (A-1) 63

4.30. Evaluación de Métricas de Rendimiento para Diversos Modelos de Aprendizaje Automático, Economicos 65

4.31. Cobertura Categoría/Rango para Modelos Smote y Tddpm, Economicos 67

4.32. Evaluación de Similitud de Distribución para Modelos Smote y Tddpm, Economicos 68

4.33. Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 5, Economicos 69

4.34. Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 1, Economicos 69

4.35. Distancia de registros más cercanos, minimo, datos economicos 69

4.36. Proporción entre el más cercano y el segundo más cercano, percentil 5, Economicos 70

4.37. Proporción entre el más cercano y el segundo más cercano, percentil 1, Economicos 70

4.38. Proporción entre el más cercano y el segundo más cercano, mínimo, Economicos . 70

4.39. Ejemplos para el modelo Tddpm, percentil 2, Economicos (B-1) 71

4.40. Ejemplos de texto modelo Tddpm, percentil 2, Economicos (B-1) 71

4.41. Ejemplos para el modelo Tddpm, percentil 4, Economicos (B-1) 72

4.42. Ejemplos de texto modelo Tddpm, percentil 4, Economicos (B-1) 72

4.43. Propiedades estadisticas de variable bathrooms con cambio>5 %, Economicos (B-1) 73

4.44. Propiedades estadisticas de variable m_size con cambio>5 %, Economicos (B-1) . 74

A.1. Ejemplos de textos aleatoreos del modelo Tddpm, conjunto Economicos (A-1) . . . 107

A.2. Ejemplos de textos aleatoreos del modelo Tddpm, conjunto Economicos (B-1) . . . 108

A.3. Propiedades estadisticas de variable sqft_living15, King county (A-3) 109

A.4. Propiedades estadisticas de variable yr_built, King county (A-3) 110

A.5. Propiedades estadisticas de variable sqft_living, King county (A-3) 111

A.6. Propiedades estadisticas de variable view, King county (A-3) 112

A.7. Propiedades estadisticas de variable price, King county (A-3) 113

A.8. Propiedades estadisticas de variable sqft_basement, King county (A-3) 114

A.9. Propiedades estadisticas de variable bedrooms, King county (A-3) 115
A.10. Propiedades estadisticas de variable condition, King county (A-3) 116
A.11. Propiedades estadisticas de variable sqft_lot15, King county (A-3) 117
A.12. Propiedades estadisticas de variable waterfront, King county (A-3) 118
A.13. Propiedades estadisticas de variable date, King county (A-3) 119
A.14. Propiedades estadisticas de variable long, King county (A-3) 120
A.15. Propiedades estadisticas de variable lat, King county (A-3) 121
A.16. Propiedades estadisticas de variable sqft_above, King county (A-3) 122
A.17. Propiedades estadisticas de variable grade, King county (A-3) 123
A.18. Propiedades estadisticas de variable bathrooms, King county (A-3) 124
A.19. Propiedades estadisticas de variable floors, King county (A-3) 125
A.20. Propiedades estadisticas de variable sqft_lot, King county (A-3) 126
A.21. Propiedades estadisticas de variable yr_renovated, King county (A-3) 127
A.22. Propiedades estadisticas de variable zipcode, King county (A-3) 128
A.23. Propiedades estadisticas de variable county, Economicos (A-3) 129
A.24. Propiedades estadisticas de variable state, Economicos (A-3) 129
A.25. Propiedades estadisticas de variable _price, Economicos (A-3) 130
A.26. Propiedades estadisticas de variable m_built, Economicos (A-3) 131
A.27. Propiedades estadisticas de variable publication_date, Economicos (A-3) 132
A.28. Propiedades estadisticas de variable rooms, Economicos (A-3) 133
A.29. Propiedades estadisticas de variable property_type, Economicos (A-3) 134
A.30. Propiedades estadisticas de variable transaction_type, Economicos (A-3) 134
A.31. Propiedades estadisticas de variable m_size, Economicos (A-3) 135
A.32. Propiedades estadisticas de variable bathrooms, Economicos (A-3) 136
A.33. Propiedades estadisticas de variable state, Economicos (B-3) 137
A.34. Propiedades estadisticas de variable publication_date, Economicos (B-3) 138
A.35. Propiedades estadisticas de variable property_type, Economicos (B-3) 139

A.36. Propiedades estadisticas de variable transaction_type, Economicos (B-3) 139
A.37. Propiedades estadisticas de variable bathrooms, Economicos (B-3) 140
A.38. Propiedades estadisticas de variable rooms, Economicos (B-3) 141
A.39. Propiedades estadisticas de variable _price, Economicos (B-3) 142
A.40. Propiedades estadisticas de variable m_size, Economicos (B-3) 143
A.41. Propiedades estadisticas de variable m_built, Economicos (B-3) 144
A.42. Propiedades estadisticas de variable county, Economicos (B-3) 145

Índice de Ilustraciones

2.1. Ejemplo de SDMetric en calculo de CDF [30] 15

3.1. Proceso para generar datos sintéticos con SDV 23

3.2. Proceso para generar datos sintéticos completo 25

3.3. Carpetas y archivos esperados generados por *Synthetic* 28

4.1. Correlación de conjunto original de entrenamiento y Copulagan, King county (A-2) 36

4.2. Correlación de conjunto original de entrenamiento y Gaussiancopula, King county (A-2) 36

4.3. Correlación de conjunto original de entrenamiento y Smote, King county (A-2) . . 37

4.4. Correlación de conjunto original de entrenamiento y Tddpm, King county (A-2) . . 37

4.5. Frecuencia del campo Grade en el modelo real y Top 2, King county (A-2) 40

4.6. Frecuencia del campo Bedrooms en el modelo real y Top 2, King county (A-2) . . 41

4.7. Frecuencia del campo Sqft lot15 en el modelo real y Top 2, King county (A-2) . . . 42

4.8. Frecuencia del campo Privacy en el modelo real y Top 2, King county (A-1) 46

4.9. Correlación de conjunto original de entrenamiento y Smote, Economicos (A-2) . . 55

4.10. Correlación de conjunto original de entrenamiento y Tddpm, Economicos (A-2) . . 55

4.11. Frecuencia del campo M size en el modelo real y Top 2, Economicos (A-2) 56

4.12. Correlación de conjunto original de entrenamiento y Smote, Economicos (B-1) . . 66

4.13. Correlación de conjunto original de entrenamiento y Tddpm, Economicos (B-1) . . 66

A.1. Correlación de conjunto original de entrenamiento y Copulagan, King county (A-3) 88

A.2. Correlación de conjunto original de entrenamiento y Tvae, King county (A-3) . . . 88

A.3. Correlación de conjunto original de entrenamiento y Gaussiancopula, King county (A-3) 89
A.4. Correlación de conjunto original de entrenamiento y Ctgan, King county (A-3) . . 89
A.5. Correlación de conjunto original de entrenamiento y Tablepreset, King county (A-3) 90
A.6. Correlación de conjunto original de entrenamiento y Smote, King county (A-3) . . 90
A.7. Correlación de conjunto original de entrenamiento y Tddpm, King county (A-3) . . 91
A.8. Frecuencia del campo Privacy en el modelo real y Top 2, King county (A-1) 91
A.9. Frecuencia del campo Floors en el modelo real y Top 2, King county (A-1) 92
A.10. Frecuencia del campo Bathrooms en el modelo real y Top 2, King county (A-1) . . 92
A.11. Frecuencia del campo Sqft above en el modelo real y Top 2, King county (A-1) . . 93
A.12. Frecuencia del campo Sqft lot en el modelo real y Top 2, King county (A-1) 93
A.13. Frecuencia del campo Price en el modelo real y Top 2, King county (A-1) 94
A.14. Frecuencia del campo Waterfront en el modelo real y Top 2, King county (A-1) . . 94
A.15. Frecuencia del campo Sqft living15 en el modelo real y Top 2, King county (A-1) . 95
A.16. Frecuencia del campo Sqft basement en el modelo real y Top 2, King county (A-1) 95
A.17. Frecuencia del campo Bedrooms en el modelo real y Top 2, King county (A-1) . . 96
A.18. Frecuencia del campo Yr built en el modelo real y Top 2, King county (A-1) 96
A.19. Frecuencia del campo Condition en el modelo real y Top 2, King county (A-1) . . . 97
A.20. Frecuencia del campo View en el modelo real y Top 2, King county (A-1) 97
A.21. Frecuencia del campo Sqft living en el modelo real y Top 2, King county (A-1) . . 98
A.22. Frecuencia del campo Grade en el modelo real y Top 2, King county (A-1) 98
A.23. Frecuencia del campo Sqft lot15 en el modelo real y Top 2, King county (A-1) . . . 99
A.24. Correlación de conjunto original de entrenamiento y Copulagan, Economicos (A-1) 99
A.25. Correlación de conjunto original de entrenamiento y Tvae, Economicos (A-1) . . . 100
A.26. Correlación de conjunto original de entrenamiento y Gaussiancopula, Economicos (A-1) 100
A.27. Correlación de conjunto original de entrenamiento y Ctgan, Economicos (A-1) . . 101
A.28. Correlación de conjunto original de entrenamiento y Smote, Economicos (A-1) . . 101

A.29. Correlación de conjunto original de entrenamiento y Tddpm, Economicos (A-1) . . 102

A.30. Correlación de conjunto original de entrenamiento y Copulagan, Economicos (B-1) 102

A.31. Correlación de conjunto original de entrenamiento y Tvae, Economicos (B-1) . . . 103

A.32. Correlación de conjunto original de entrenamiento y Gaussiancopula, Economicos (B-1) 103

A.33. Correlación de conjunto original de entrenamiento y Ctgan, Economicos (B-1) . . . 104

A.34. Correlación de conjunto original de entrenamiento y Smote, Economicos (B-1) . . 104

A.35. Correlación de conjunto original de entrenamiento y Tddpm, Economicos (B-1) . . 105

Lista de códigos

1. Eliminación de valores nulos en el conjunto de datos de Económicos 53
2. Reemplazo de valores nulos en el conjunto de datos de Económicos 53
3. Código de ejemplo en Python para sumar dos números. Fuente: Autor. 86
4. Devcontainer del proyecto en curso. 87
5. Mostrando Puntajes Promedios Calculados 145
6. Instanciando clase Synthetic 146

Capítulo 1

Introducción

Es probable que esta tesis esté desactualizada en el momento de su revisión. Desde la aparición de AlexNet [1] n 2012, el liderazgo en la clasificación de imágenes ha cambiado al menos 15 veces [2]. En la esfera de la generación de texto a imágenes, se destacan modelos como DALL-E 2 [3], Google Imagen [4] y Stable Diffusion [5], todos introducidos en 2022. Para 2023, se anticipa una contienda en el ámbito de la inteligencia artificial enfocada en chatbots, protagonizada por Google y Microsoft [6], [7]. En resumen, es un ámbito de constante evolución, que promete seguir innovando con nuevas técnicas y productos, en términos de variedad y calidad.

Enmarcado en la empresa **Equifax**, a la que se dirige este estudio, es imperativo progresar de manera rápida y efectiva en el empleo de su información para mantenerse a la vanguardia en su industria y competir con otras entidades del sector.

Según el libro *Practical synthetic data generation: balancing privacy and the broad availability of data* [8] 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 nuevas tecnicas.

1.1. Equifax: Contexto y Restricciones

Equifax es una entidad crediticia multinacional que, junto con Transunion y Experian, conforman los tres burós de crédito más grandes a nivel global. La compañía cuenta con equipos de

desarrollo en Estados Unidos, India, Irlanda y Chile y opera en más de 24 países. El negocio principal de Equifax es la generación de conocimientos a partir de la data recolectada, que incluye información crediticia, servicios básicos, autos, mercadotecnia, Twitter, revistas, datos demográficos, entre otros. El principal reto tecnológico de la compañía es proteger la privacidad de estos datos. El segundo es realizar predicciones significativas para el mercado utilizando los datos acumulados. Los datos son uno de los activos más importantes, si no el más importante, de la compañía.

El equipo **Keying and Linking** de Equifax se encarga de identificar entidades y establecer sus relaciones dentro de los diferentes conjuntos de datos. Esta tarea debe aplicarse a cada entidad dentro de la compañía y a través de todas las zonas geográficas. La identificación de entidades, o entity resolution, es el proceso de determinar que dos o más registros de información, que referencian a un objeto único en el mundo real, pueden ser una misma entidad. Por ejemplo, Bob Smith, Robert Smith y Robert S. podrían referirse a la misma persona, lo mismo podría suceder con una dirección. Es importante mencionar que la información requerida para este equipo es de identificación personal (PII), la cual está categorizada y protegida con las máximas restricciones dentro de la compañía, lo que justifica la estricta gestión de los registros y la prohibición de usar datos reales en ambientes de desarrollo.

El enfoque actual propone un método alternativo para la generación de datos sintéticos utilizando inteligencia artificial. Estos datos sintéticos son utilizados en pruebas de nuevo software en entornos no productivos en Equifax. Para el equipo de **Keying and Linking** y la compañía, es importante evaluar los nuevos desarrollos, pero es aún más importante proteger la privacidad y seguridad de los datos. Es por ello que la privacidad y calidad de estos datos son relevantes.

En cuanto a la regulación y el acceso directo a la información personal legible y no enmascarada en Equifax, ésta está regulada y solo disponible para proyectos categorizados como Protegidos. Estos proyectos están administrados por un equipo especializado en infraestructura, responsable de la seguridad y las herramientas ofrecidas para dichos espacios de trabajo. Los permisos de acceso son supervisados y revisados periódicamente.

Equifax, como empresa orientada a la IA (*AI-First Company*), está en una constante evolución buscando ser pionera en inteligencia artificial, utilizando los datos almacenados durante más de un siglo y su asociación con Google, su principal proveedor de servicios en la nube. El objetivo para 2022 es tener la capacidad de entrenar modelos de Deep Learning usando las plataformas analíticas actuales administradas por la compañía; el producto seleccionado y en proceso de implementación es Vertex AI. Equifax se encuentra en proceso de evaluación de empresas que generan datos sintéticos acordes a las necesidades de la organización. Una de las evaluadas es Tonic IA `https://www.tonic.ai/`, lo que evidencia la relevancia que tienen los datos sintéticos en los objetivos a mediano plazo de Equifax.

1.2. Objetivo

Objetivo General:

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

Objetivos Específicos:

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

1.3. Estructura del documento

Capítulo 2 Revisión Bibliográfica:
Este capítulo proporciona una revisión bibliográfica que abarca tipos de datos, privacidad de datos y generación de datos sintéticos. Se discuten y resumen los diferentes tipos de datos, se subraya la importancia de la protección de datos, y se discuten los enfoques de generación de datos sintéticos.

Capítulo 3 Desarrollo:
Este capítulo detalla el proceso de desarrollo del estudio, incluyendo: **Conjuntos de datos,** Se describen detalladamente las bases de datos utilizadas, con una visión general de sus campos y características. **Computación y Software,** Se proporciona una descripción detallada de los recursos de hardware y software empleados. **Desarrollo del flujo de procesamiento,** Se ofrece una descripción completa del flujo de procesamiento implementado, basado en la metodología de Synthetic Data Vault (SDV). **Modelos de generación de datos,** Se presentan y describen los modelos de generación de datos tabulares y de texto utilizados y finalmente **Obtención de Métricas,** Se detalla la metodología empleada para obtener y calcular las métricas para la evaluación de los conjuntos de datos sintéticos.

Capítulo 4 Resultados:
Los resultados del estudio se presentan en este capítulo, dividido en dos secciones principales: King County y Económicos. En cada sección, se discuten detalladamente los resultados obtenidos para cada conjunto de datos.

Capítulo 5 Conclusiones y discusión:
Este capítulo se divide en tres secciones principales: **Conclusiones,** Se recapitula el objetivo del estudio y se discuten los hallazgos más significativos. **Limitaciones,** Se reconocen y discuten las limitaciones del estudio. Se presenta un ejemplo del uso de ChatGPT para ilustrar la relevancia de la generación de datos sintéticos. **Discusión,** Se discute sobre los modelos emergentes de generación de texto y la necesidad de futuros estudios en esta área, incluyendo la importancia de evaluar la privacidad.

Capítulo 2

Revisión Bibliográfica

Este capítulo establece las bases teóricas de nuestra investigación, proporcionando un análisis exhaustivo de la literatura existente sobre los tipos de datos, la privacidad de los datos y la generación de datos sintéticos. El objetivo es establecer un contexto para la investigación y una base para el desarrollo de las metodologías aplicadas en los capítulos siguientes.

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.

En 2012, la Corporación de Datos Internacional, IDC por su sigla en ingles, estimó que para 2020, más del 95 % de los datos serían no estructurados [9]. En un análisis posterior, Kiran Adnan y Rehan Akbar [10] 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 resume la lista que se encuentra en *Practical Statistics for Data Scientists* [11].

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
Categórico		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 [11]. 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 del texto *Data privacy: Definitions and techniques* [12].

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* [11] están resumidos en la Tabla 2.3.

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 [8] 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. Privacidad Diferencial

La privacidad diferencial (*Differential Privacy*) se ha establecido como un marco matemático robusto para garantizar la privacidad en el análisis de datos. Se define formalmente como sigue [13]:

Definición (Privacidad Diferencial): Sea $\varepsilon > 0$ un parámetro pequeño. Un mecanismo aleatorio $K : D \to R$ (donde D es el dominio de los conjuntos de datos y R es el rango de posibles respuestas) proporciona *privacidad diferencial* ε si para todos los conjuntos de datos $D, D' \in D$ que difieren en a lo más un elemento, y para todo conjunto $S \subseteq R$, se cumple que:

$$\Pr[K(D) \in S] \leq \exp(\varepsilon) \times \Pr[K(D') \in S] \tag{2.1}$$

La privacidad diferencial garantiza que la inclusión o exclusión de un único individuo en el conjunto de datos no altera significativamente la probabilidad de cualquier resultado específico. Esto se logra mediante el mecanismo de ruido K, que introduce una incertidumbre calculada en los resultados, protegiendo así la información individual.

2.2.4. k-Anonimato

El *k-Anonimato* [14] es una propiedad de privacidad en el análisis de datos que se enfoca en proteger la identidad de los individuos en un conjunto de datos. Formalmente, un conjunto de datos se considera k-anónimo si cada registro es indistinguible de al menos $k-1$ otros registros respecto a ciertos atributos identificativos. Esto se logra mediante la técnica de generalización y supresión de datos.

Para calcular el k-Anonimato de un conjunto de datos, se siguen los siguientes pasos:

1. **Identificar Atributos Identificativos:** Seleccionar los atributos en el conjunto de datos que pueden ser utilizados para identificar a los individuos, conocidos como identificadores cuasi-únicos (QID).
2. **Aplicar Generalización y/o Supresión:** Generalizar o suprimir los QID para que cada combinación de valores de QID en el conjunto de datos se asocie con al menos k individuos.
3. **Verificar la Indistinguibilidad:** Asegurarse de que cada registro en el conjunto de datos sea indistinguible de al menos $k-1$ otros registros respecto a los QID.

Para demostrar que un conjunto de datos es k-anónimo, se debe mostrar que para cada registro, hay al menos $k-1$ registros adicionales que comparten los mismos valores de QID. Matemáticamente, para un conjunto de datos D y un conjunto de QID Q, se debe cumplir que:

$$\forall x \in D, |\{y \in D \mid Q(x) = Q(y)\}| \geq k \tag{2.2}$$

Donde $Q(x)$ representa el conjunto de valores de QID para el registro x.

2.2.5. Distancia al registro más cercano (DCR)

La *Distance to Closest Record* (DCR), o en español "Distancia al Registro más Cercano", es una métrica utilizada para medir la distancia euclidiana entre cualquier registro sintético y su vecino real más cercano. El propósito principal de la DCR es evaluar el riesgo de violación de la privacidad en conjuntos de datos que incluyen registros sintéticos. En este contexto, un registro sintético es una entidad de datos generada artificialmente que pretende ser indistinguible de un registro real.

La importancia de la DCR radica en su capacidad para cuantificar qué tan similares son los registros sintéticos a los reales. Cuanto mayor es la DCR, se considera que hay un menor riesgo de que los registros sintéticos puedan ser asociados directamente con sus contrapartes reales, lo que disminuye el riesgo de una brecha de privacidad. Esto es crucial en aplicaciones donde la privacidad de los datos es una preocupación significativa, como en el análisis de datos de salud o financiero.

Adicionalmente, para proporcionar una estimación más robusta del riesgo de privacidad, se calcula el percentil 5 de esta métrica como se propone en [15]. Según su estudio, el uso del percentil 5 ayuda a entender mejor la distribución de las distancias y a identificar si una proporción significativa de registros sintéticos está demasiado cerca de los registros reales, lo que podría indicar un riesgo de privacidad elevado. Este enfoque estadístico ofrece una perspectiva más completa sobre la efectividad de los registros sintéticos para preservar la privacidad en comparación con un análisis que solo considera el promedio o la mediana de las distancias.

2.2.6. Ratio de distancia entre los vecinos más cercanos (NNDR)

El Ratio de distancia entre los vecinos más cercanos, NNDR por sus siglas en ingles, *Nearest Neighbour Distance Ratio*, es una métrica utilizada para evaluar la relación entre la distancia euclidiana al vecino real más cercano y el segundo vecino más cercano. NNDR ha sido adaptado para evaluar la privacidad de los datos sintéticos en [15].

En el contexto de la privacidad de datos sintéticos, el NNDR proporciona una medida de cuán distinto es un registro sintético de sus dos vecinos reales más cercanos. Se calcula como el cociente de la distancia al vecino más cercano sobre la distancia al segundo vecino más cercano. Este ratio varía entre 0 y 1, donde valores más altos indican una mejor privacidad. Es decir, un valor alto de NNDR sugiere que el registro sintético está relativamente más alejado de su vecino real más cercano en comparación con el segundo vecino más cercano, reduciendo así el riesgo de identificación de los datos reales.

Por otro lado, valores bajos de NNDR entre los datos sintéticos y reales pueden indicar que los registros sintéticos están demasiado cerca de sus contrapartes reales, lo que podría resultar en la

revelación de información sensible del registro real más cercano. Esta métrica es particularmente útil para identificar si los datos sintéticos podrían comprometer la privacidad de los datos reales.

Similar a otras métricas en el ámbito de la privacidad de datos, el percentil 5 del NNDR también se calcula para proporcionar una estimación más robusta del riesgo según [15]. Este cálculo ayuda a entender la distribución de los valores de NNDR y a identificar si una proporción significativa de registros sintéticos presenta un riesgo elevado de privacidad. En resumen, el NNDR es una herramienta valiosa para evaluar la eficacia de los datos sintéticos en la preservación de la privacidad en conjuntos de datos.

2.2.7. 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 [11]. Los nuevos datos podrían ser afectos por:

1. Regulation (EU) 2016/679 of the European Parliament and of the Council [16], 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 [17]
3. Health insurance portability and accountability act of 1996 [18]

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 [11]. 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 Tabla 2.1, 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 2.4, 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 2.4: Estado del arte en generación de datos tabulares

Nombre	Fecha ↓	Código
REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers [19]	2023-02-04	Github
PreFair: Privately Generating Justifiably Fair Synthetic Data [20]	2022-12-20	
GenSyn: A Multi-stage Framework for Generating Synthetic Microdata using Macro Data Sources [21]	2022-12-08	Github
TabDDPM: Modelling Tabular Data with Diffusion Models [22]	2022-10-30	Github
Language models are realistic tabular data generators [23]	2022-10-12	Github
Ctab-gan+: Enhancing tabular data synthesis [24]	2022-04-01	Github
Ctab-gan: Effective table data synthesizing [15]	2021-05-31	Github
Modeling Tabular data using Conditional GAN [25]	2019-10-28	Github
Smote: synthetic minority over-sampling technique [26]	2002-06-02	Github

2.3.2. Generación de texto en base de datos tabulares

En la Tabla 2.5, se listan las publicaciones en la generación de texto a partir de datos estructurados.

Tabla 2.5: Estado del arte en generación de textos en base a datos

Nombre	Fecha ↓	Modelo Base
Table-To-Text generation and pre-training with TABT5 [27]	2022-10-17	T5
Text-to-text pre-training for data-to-text tasks [28]	2021-07-09	T5
TaPas: Weakly supervised table parsing via pre-training [29]	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 2.6 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 2.6: 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 [27], [29]. 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* [28] 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. Metricas 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.

- **SDMetrics Score**: Este es un promedio de las métricas de *Quality Report* y *Diagnostic Report* que varían de 0 a 1. El *Quality Report* se refiere a las formas de las columnas (Column Shapes) y las tendencias entre pares de columnas (Column Pair Trends). El *Diagnostic Report* se refiere a la síntesis (Synthesis), cobertura (Coverage) y límites (Boundaries).

- **Reporte diagnóstico**: Se refiere a un conjunto de métricas que incluyen Síntesis, Cobertura y Límites. La síntesis verifica la singularidad de los datos sintéticos, es decir, que sean en su combinación un elemento nuevo, la cobertura verifica si los datos sintéticos cubren todo el espectro de valores y los límites verifican si los datos sintéticos respetan los límites de los datos reales.
- **Reporte de calidad**: En este informe se revisan dos componentes, las formas de las columnas (Column Shapes) y las tendencias entre pares de columnas (Column Pair Trends). La forma de las columnas mide la capacidad de los datos sintéticos para capturar la distribución general de cada columna en los datos reales. Las tendencias entre pares de columnas describen cómo varían en relación entre sí.
- **Correlación**: En esta sección se utiliza la correlación de Pearson en columnas numéricas para generar una representación visual en un mapa de calor.
- **Privacidad**: Esta sección cubre dos métricas, Distancia al Registro más Cercano (DCR) y Relación de Distancia del Vecino más Cercano (NNDR). DCR cuantifica la distancia euclidiana entre cualquier registro sintético y su vecino real más cercano. NNDR mide la relación entre la distancia euclidiana al vecino real más cercano y al segundo vecino real más cercano de cualquier registro sintético.
- **Propiedades estadísticas**: Esta sección presenta un conjunto de propiedades estadísticas como el número de observaciones, elementos faltantes, media, mediana, moda, entre otros.

2.4.1. SDMetrics Score

En la Tablas incluidas en la sección *SDMetrics Score* se expresa en su última columna un promedio de *Quality Report* de SDMetric. Adicionalmente se incluyen las metricas calculadas en *Diagnostic Report*. Todas las metricas expresadas en este apartado van del 0 al 1, ya que las métricas promediadas van en este rango tambien. Detalles de cada metricas se podrán encontrar en las siguientes subsecciones 2.4.2 y 2.4.3.

$$Score = \frac{\text{Column Pair Trends} + \text{Column Shapes}}{2}$$

El *Quality Report* captura las formas de las columnas (*Column Shapes*) y las tendencias entre pares de columnas (*Column Pair Trends*). La forma de una columna describe su distribución general, y la tendencia entre dos columnas describe cómo varían en relación entre sí. Es importante tener en cuenta que las distribuciones de orden superior de 3 o más columnas no se incluyen en el *Quality Report*. Detalles de cada métricas se podrán encontrar en Sección 2.4.3.

Por otro lado, el *Diagnostic Report* mide la síntesis (*Synthesis*), cobertura (*Coverage*) y límites (*Boundaries*). La síntesis se refiere a si los datos sintéticos son únicos o si copian las filas reales. La cobertura verifica si los datos sintéticos cubren el rango de valores posibles. Los límites, por su parte, comprueban si los datos sintéticos respetan los límites establecidos por los datos reales. Al igual que en el *Quality Report*, se aplican métricas basadas en los tipos de columnas y se promedian para obtener una puntuación final. Detalles de cada métricas se podrán encontrar en Sección 2.4.2.

Tabla 2.7: Ejemplo de Métricas de Rendimiento para Diversos Modelos

Model Name	Column Pair Trends	Column Shapes	Coverage	Boundaries	Synthesis	**Score**
tddpm_mlp	**9.73e-01**	**9.84e-01**	**7.91e-01**	**1.00e+00**	9.91e-01	**9.79e-01**
smote-enc	9.62e-01	9.76e-01	6.67e-01	**1.00e+00**	9.24e-01	9.69e-01
copulagan	7.46e-01	7.90e-01	6.80e-01	**1.00e+00**	**1.00e+00**	7.68e-01

2.4.2. Reporte diagnóstico

La evaluación se lleva a cabo mediante una serie de métricas que se agrupan en: *Synthesis*, *Coverage*, y *Boundaries*.

Synthesis emplea *NewRowSynthesis* en SDMetrics, la cual verifica la singularidad de los datos sintéticos, esto es, si los datos sintéticos copian filas completas de los datos reales o si son capaces de generar nuevas filas únicas. Para su cálculo, en primer lugar, se obtiene un conjunto único de filas tanto para los datos reales como para los sintéticos. Posteriormente, se calcula el número de filas únicas que se encuentran en los datos sintéticos pero no en los reales. Finalmente, este número se divide por el número total de filas únicas en los datos sintéticos para obtener la probabilidad de generar una nueva fila única. En lo que respecta a la puntuación, un valor de 1.0 indica que todos los datos sintéticos y reales son únicos, mientras que un valor de 0.0 señala que todos los datos sintéticos y reales pueden compartir recursos. un valor de 1.0 es el mejor resultado para esta metrica.

La métrica *RangeCoverage* se emplea para determinar si una columna sintética cubre todo el espectro de valores presentes en una columna real. Esta métrica es compatible con datos numéricos continuos y de tiempo/fecha, los cuales se transforman en valores numéricos. Los valores vacíos son ignorados por esta métrica. El cálculo de la puntuación se realiza de acuerdo con la siguiente fórmula, donde r y s representan las columnas real y sintética respectivamente:

$$\text{score} = 1 - \left[\text{máx}\left(\frac{\text{mín}(s) - \text{mín}(r)}{\text{máx}(r) - \text{mín}(r)}, 0\right) + \text{máx}\left(\frac{\text{máx}(r) - \text{máx}(s)}{\text{máx}(r) - \text{mín}(r)}, 0\right)\right]$$

La métrica *Boundaries* se emplea para evaluar si una columna sintética respeta los valores mínimos y máximos de la columna real, devolviendo el porcentaje de filas sintéticas que se mantienen dentro de los límites reales. Esta métrica se aplica a los datos numéricos, transformando los valores de fecha y hora en valores numéricos, e ignorando los valores ausentes. Un valor de 0.9 en esta métrica significa que el 10 % de los registros no respeta los límites de los datos reales.

2.4.3. Reporte de calidad

El *Quality Report* en SDMetrics consta de dos componentes principales: *Column Shapes* y *Column Pair Trends*. Existe un tercero, pero que solo se útiliza cuando se genera varios conjuntos de datos relacionados.

El primero, *Column Shapes*, mide la capacidad de los datos sintéticos para capturar la forma

de cada columna en los datos reales. La forma de una columna describe su distribución general. Para realizar esta evaluación, se utilizan dos métricas basadas en los tipos de dato de la columna. Por ejemplo, para las columnas numéricas y de fecha y hora, se utiliza la métrica *KSComplement*, mientras que para las columnas categóricas y booleanas, se utiliza la métrica *TVComplement*.

El *KSComplement* utiliza la estadística de Kolmogorov-Smirnov para calcular la similitud entre las columnas reales y sintéticas. En particular, se convierte la distribución numérica en su función de distribución acumulativa, CDF por sus siglas en ingles de *Cumulative distribution function*, y la estadística de KS es la diferencia máxima entre las dos CDFs. En SDMetrics, se invierte la estadística: *KSComplement* devuelve $1 - (\text{estadística de KS})$ para que una puntuación más alta signifique una mayor calidad. En la Figura 2.1 se puede observar una representación grafica del calculo.

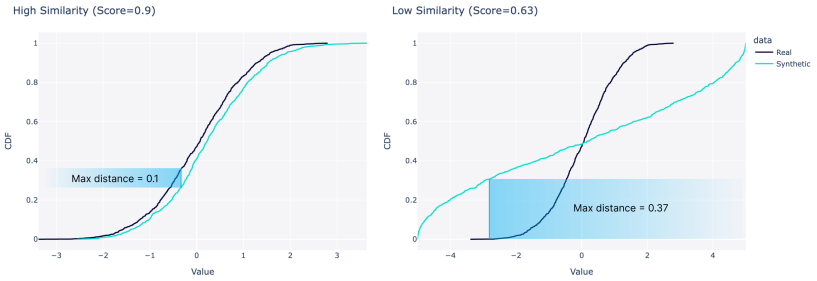

Figura 2.1: Ejemplo de SDMetric en calculo de CDF [30]

TVComplement calcula la distancia de variación total, conocida como TVD por su acrónimo en inglés de *Total Variation Distance*, entre las columnas reales y sintéticas. Para lograr esto, en primer lugar estima la frecuencia de cada valor de categoría y la expresa en términos de probabilidad. La estadística TVD compara las diferencias en estas probabilidades, siguiendo la fórmula:

$$\delta(R,S) = 2 \sum_{\omega \in \Omega} |R_\omega - S_\omega|$$

Aquí, ω representa todas las posibles categorías en una columna, Ω. Por otro lado, R y S hacen referencia a las frecuencias reales y sintéticas para dichas categorías, respectivamente. *TVComplement* retorna $1 -$ TVD de modo que una puntuación más alta denote una calidad superior.

El segundo componente, *Column Pair Trends*, evalúa la capacidad de los datos sintéticos para replicar las correlaciones y tendencias entre cada par de columnas en los datos reales. Esta métrica se basa en la forma de las columnas individuales, las correlaciones entre cada par de columnas y la cardinalidad entre las tablas. Utiliza la métrica *CorrelationSimilarity* para medir la correlación entre pares de columnas numéricas y calcular la similitud entre los datos reales y sintéticos. La métrica *CorrelationSimilarity* soporta tanto los coeficientes de correlación de Pearson como de Spearman, proporcionando una evaluación flexible de las relaciones entre los datos.

CorrelationSimilarity es una métrica únicamente aplicable a variables numéricas; las fechas son transformadas a números para este fin. Esta métrica calcula la correlación utilizando los coeficientes de Pearson y Spearman. El score se obtiene mediante la siguiente fórmula, donde $S_{A,B}$ representa la correlación del conjunto sintético y $R_{A,B}$ la correlación del conjunto real.

$$score = 1 - \frac{|S_{A,B} - R_{A.B}|}{2}$$

2.4.4. Correlación

En esta sección, se emplea la métrica de correlación de Pearson en las columnas numéricas, transformando las fechas en números e ignorando los valores ausentes, con el propósito de generar una representación visual en un mapa de calor para su análisis.

2.4.5. Privacidad

Pendiente
Reformular, figura 6 no aparece

La Distancia al Registro más Cercano (DCR, por sus siglas en inglés) se emplea para cuantificar la distancia euclidiana entre cualquier registro sintético y su vecino real más próximo. Idealmente, a medida que la DCR se incrementa, el riesgo de violación de la privacidad disminuye. Adicionalmente, se estima el percentil 5 de esta métrica para proporcionar una medición robusta del riesgo de privacidad [15].

Relación de Distancia del Vecino más Cercano (NNDR, por sus siglas en inglés). La NNDR mide la relación entre la distancia euclidiana para el vecino real más cercano y el segundo vecino real más cercano a cualquier registro sintético correspondiente. NNDR es un concepto comúnmente utilizado en el área de visión por computadora para el emparejamiento de descriptores de imagen locales. Platzer [31] introduce esta idea para evaluar la privacidad de los datos sintéticos. Esta relación está dentro del intervalo [0, 1]. Valores más altos indican una mejor privacidad. Los valores bajos de NNDR entre datos sintéticos y reales pueden revelar información sensible del registro de datos reales más cercano. La Fig. 6 ilustra el caso. Tenga en cuenta que también se calcula aquí el percentil 5 [15].

2.4.6. Propiedades estadísticas y tecnicas de análisis

En la Tabla 2.8 se lista algunas de las propiedades estadisticas para conjuntos de valores.

Tabla 2.8: Listado de conjunto estadísticos

Nombre	abrev.	Descripción
Numero de observaciones	nobs	Cuenta de elementos
Vacios	missing	Numero de elementos vacios
Media	mean	La suma de todos los valores dividido por el número total de valores
Continúa en la siguiente página		

Nombre	abrev.	Descripción
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	0.1 %, ... 99.9 %	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	tmean	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
Varianza	var	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	std	La raíz cuadrada de la varianza
Error estándar de la media	std_err	Medida de la variación de las muestras de la media poblacional.
Desviación Absoluta Media	mad	La media de los valores absolutos de las desviaciones desde la media
Desviación Mediana Absoluta Normalizada	mad_normal	Medida de dispersión basada en la mediana que normaliza la desviación mediana absoluta utilizando la constante de escala 1.4826 para la distribución normal.
Rango	range	La diferencia entre el valor más grande y el valor más pequeño en un conjunto de datos
Tablas de Frecuencia	top5_freq	Un método para resumir los datos al contar cuántas veces ocurre cada valor en un conjunto de datos
Probabilidad	top5_prob	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	cont_table	Una tabla que muestra la distribución conjunta de dos o más variables categóricas
Comparación de Modelos Predictivos Multivariables	cross_pred	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

Continúa en la siguiente página

Nombre	abrev.	Descripción
Kullback-Leibler	kl	Una medida de la divergencia entre dos distribuciones de probabilidad
Log-Cluster	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
Cross-Classification	cross_class	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
Intervalo de confianza superior	upper_ci	Valor máximo del intervalo de confianza alrededor de la media.
Intervalo de confianza inferior	lower_ci	Valor mínimo del intervalo de confianza alrededor de la media.
Rango intercuartil	iqr	Diferencia entre el tercer y el primer cuartil, mide la dispersión de los datos.
Rango intercuartil normalizado	iqr_normal	Rango intercuartil dividido por la media, para normalizar los valores.
Coeficiente de variación	coef_var	Relación entre la desviación estándar y la media, mide la dispersión relativa de los datos.
Sesgo	skew	Medida de la asimetría de la distribución de los datos alrededor de la media.
Curtosis	kurtosis	Medida de la "pesadez"de las colas de la distribución de los datos.
Test de Jarque-Bera	jarque_bera	Estadístico que mide si los datos tienen la asimetría y la curtosis correspondiente a una distribución normal.
p-valor del test de Jarque-Bera	jarque_bera _pval	Probabilidad de que los datos sean normalmente distribuidos según el test de Jarque-Bera.

Capítulo 3

Desarrollo

La generación de datos sintéticos es una práctica en rápida evolución con implicaciones significativas para el campo de la inteligencia artificial. Este capítulo se centra en proporcionar una visión detallada del desarrollo y la implementación este proceso, con el fin de la generación de dichos datos.

En primer lugar, se examinan los recursos disponibles que forman la base de este estudio. Se describen dos conjuntos de datos principales: uno derivado de la información de los precios de las viviendas de King County y otro de Economicos.cl, un portal chileno de anuncios clasificados. Además, se proporciona información detallada sobre el equipo informático y el software empleado en este estudio.

El núcleo de este capítulo se centra en la explicación del desarrollo de un proceso de generación de datos sintéticos. Este proceso se basa en la metodología propuesta por Synthetic Data Vault (SDV) y se extiende para incluir etapas intermedias de almacenamiento de modelos y resultados de evaluación. Se proporciona una descripción detallada de cada paso del proceso, desde la creación de metadatos hasta la generación del conjunto de datos sintéticos.

Para concluir, se abordan los métodos para la evaluación de los conjuntos de datos sintéticos generados. Esto incluye la descripción de cómo se obtienen y calculan las métricas, con un ejemplo de cómo calcular y visualizar el *score* promedio para una selección específica de modelos. Este capítulo prepara el terreno para un análisis profundo de los resultados obtenidos a través de esta metodología, que se presentará en los capítulos siguientes.

3.1. Recursos disponibles

3.1.1. Conjuntos de datos

A continuación se describen y detallan las bases de datos utilizadas en los experimentos.

King County

La base de datos de King County [32] contiene notificación sobre precios de venta y características de 21,613 viviendas en Seattle y King County de los años 2014 y 2015. La base incluye datos como el número de habitaciones, el número de baños, la superficie del terreno y la superficie construida, así como detalles sobre la ubicación de la propiedad, como la latitud y la longitud. Este paquete de datos es comúnmente utilizado para tareas de regresión y predicción de precios de viviendas. Sus campos se describen en la Tabla 3.1.

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

Económicos

Economicos.cl es un portal web chileno que se especializa en la publicación de anuncios clasificados en línea, enfocándose en las categorías de bienes raíces, vehículos, empleos, servicios y

productos variados. La base de datos se originó de un *Web Scraping* ejecutado en 2020, y contiene 22.059 observaciones.

Tabla 3.2: Base de datos Economicos.cl

Variable	Descripción
url	URL de la publicación
Descripción	Detalles de la publicación
price	Valor de venta, en dólares, UF o pesos
property_type	Clase de propiedad: Casa, Departamento, etc.
transaction_type	Clase de transacción Arriendo, Venta
state	Región de la publicación
county	Comuna de la publicación
publication_date	Fecha de la publicación
rooms	Cantidad de dormitorios
bathrooms	Cantidad de baños
m_built	Extensión del área habitable en metros cuadrados
m_size	Extensión del terreno en metros cuadrados
source	Medio de la publicación
title	Título de la publicación
address	Dirección de la publicación
owner	Publicador
_price	Valor convertido a UF del día de la publicación

3.1.2. Computación y Software

Para efectuar los experimentos, se recurrió a un equipo informático con las especificaciones técnicas detalladas en la Tabla 3.3. El procesador seleccionado fue un AMD Ryzen 9 7950X 16-Core Procesadores, complementado con cuatro módulos de 32 GB para sumar una memoria total de 128 GB DDR5. La tarjeta gráfica incorporada fue una NVIDIA GeForce RTX 4090, y el equipo contó con dos discos duros de 500 GB SSD. El uso de un sistema con estas características garantizó una ejecución eficaz de los modelos de generación de datos, asegurando la viabilidad de los experimentos. Cabe resaltar que la selección de los componentes del equipo se realizó de manera meticulosa para garantizar que los resultados obtenidos no se vieran afectados por una capacidad de hardware limitada.

En lo que respecta al software, se empleó el sistema operativo Ubuntu 20.04.2 LTS y se utilizó el lenguaje de programación Python 3.10 para la implementación de los modelos de generación de datos. Se recurrió a diversas bibliotecas, incluyendo DVC, SDV y PyTorch, cuya lista completa está disponible en el repositorio en Github. La elección de estas herramientas estuvo guiada por su compatibilidad con el modelo Tddpm, uno de los listados en la sección 2.3.1, el cual fue empleado 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

Con el objetivo de asegurar la reproducibilidad, se implementó *devcontainer*, que configura el entorno de desarrollo y pruebas mediante una imagen replicable de *Docker*. Los experimentos pueden ser reproducidos utilizando el contenedor descrito en el repositorio y el Código en el Anexo A.2.

El código fuente de los modelos destinados a la generación de datos, así como los scripts para el análisis y la representación gráfica de los resultados, se encuentran disponibles en un repositorio público de Github: gvillarroel/synthetic-data-for-text. Se requiere el uso de DVC para la descarga de datos desde un directorio compartido en Google Drive.

3.2. Desarrollo del flujo de procesamiento

En las siguientes secciones se detalla el flujo de procesamiento implementado para la generación de nuevos datos sintéticos. Este flujo se inspira en el propuesto por Synthetic Data Vault (SDV), incorporando algunas modificaciones para preservar etapas intermedias.

SDV es un ecosistema de bibliotecas para la generación de datos sintéticos que facilita a los usuarios aprender de bases de datos unidimensionales, multidimensionales y de series temporales, para posteriormente generar nuevos datos sintéticos que mantengan las mismas propiedades estadísticas y el mismo formato que las bases de datos originales. Para conseguir esto, SDV emplea diversas técnicas, como modelos generativos y redes neuronales, con el fin de aprender la distribución subyacente de los datos y generar nuevos datos que sigan dicha distribución [33], [34].

A continuación, se explica el proceso de generación de datos sintéticos para una base de datos unidimensional utilizando la biblioteca Synthetic Data Vault (SDV), seguido de las modificaciones introducidas para expandir el proceso e incorporar nuevos modelos.

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

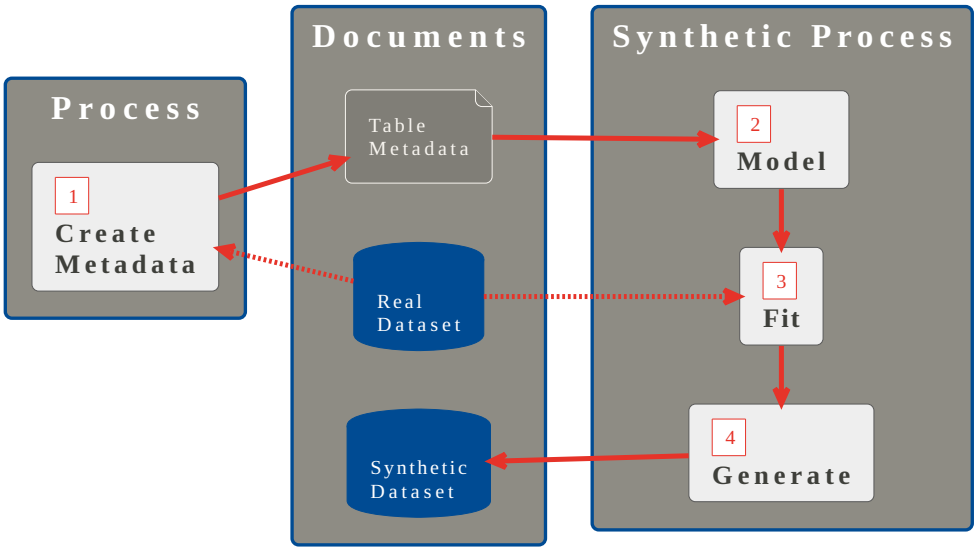

Figura 3.1: Proceso para generar datos sintéticos con SDV

1. **Creación de Metadatos**: Se elabora un diccionario que define los campos del conjunto de datos y los tipos de datos que contiene. Esto le permite a SDV aprender la estructura del conjunto de datos original y utilizarla para generar nuevos datos sintéticos con la misma estructura.
2. **Creación del Modelo**: Se selecciona el modelo de generación de datos a utilizar. SDV proporciona varios modelos, entre ellos GaussianCopula, CTGAN, CopulaGAN y TVAE, que se adaptan a distintos tipos de datos y distribuciones.
3. **Entrenamiento del Modelo**: El modelo seleccionado se entrena con el conjunto de datos original para aprender sus distribuciones y patrones estadísticos.

4. **Generación del Conjunto de Datos Sintéticos**: Con el modelo ya entrenado, se generan nuevos datos sintéticos que mantienen la misma estructura y características estadísticas que el conjunto original. Este nuevo conjunto de datos puede ser empleado para diversas aplicaciones, como pruebas de software o análisis de datos sensibles.

Es crucial señalar que el proceso de generación de datos sintéticos con SDV es escalable y puede aplicarse a bases de datos unidimensionales, multidimensionales y de series temporales. Adicionalmente, en este proyecto se introdujeron ciertas modificaciones al flujo para expandir el proceso y facilitar la incorporación de nuevos modelos.

En el proceso extendido de generación de datos sintéticos con SDV, se introducen dos nuevas etapas para permitir el almacenamiento de los modelos intermedios y los resultados de la evaluación. El proceso completo se ilustra en la Figura 3.2 y comprende los siguientes pasos:

1. **Creación de Metadatos**: Se elabora un diccionario que define los campos del conjunto de datos y los tipos de datos que contiene.
2. **Creación del Modelo**: Se selecciona el modelo a utilizar. SDV permite elegir entre GaussianCopula, CTGAN, CopulaGAN y TVAE.
3. **Entrenamiento del Modelo**: El modelo seleccionado se entrena con el conjunto de datos original para aprender sus distribuciones.
4. **Guardado del Modelo**: El modelo entrenado se almacena en un archivo para su uso posterior.
5. **Generación del Conjunto de Datos Sintéticos**: Se genera un nuevo conjunto de datos utilizando el modelo entrenado.
6. **Evaluación y Guardado de Métricas**: Se evalúa el conjunto de datos sintético generado y se almacenan las métricas, como la correlación, el error absoluto medio y el error cuadrático medio.

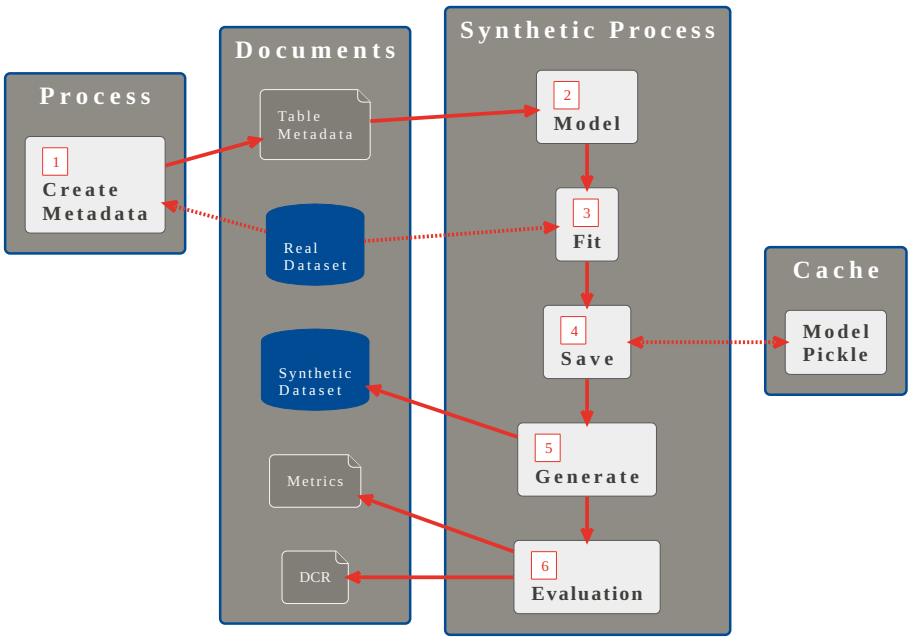

Figura 3.2: Proceso para generar datos sintéticos completo

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 de generación de datos

Los modelos de generación de datos tabulares se fundamentan en la metodología propuesta por *Synthetic Data Vault* (SDV), mientras que los modelos de generación de texto utilizan 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 sea compatible con SDV, debe implementar los siguientes métodos:

1. **cargar** (load): Carga el modelo desde un archivo.
2. **entrenar** (fit): Entrena el modelo, tomando como entrada un dataframe de pandas.
3. **guardar** (save): Almacena el modelo en un archivo.
4. **muestrear** (sample): Genera un conjunto de nuevos registros utilizando el modelo entrenado.

Como consideración adicional, se aconseja llevar a cabo el proceso mediante un script en lugar de un cuaderno Jupyter, dado que se ha observado que el cuaderno puede encontrar problemas con algunos modelos debido a restricciones de memoria. A continuación, se especifican los pasos para la ejecución del proceso:

1. Generar un archivo de configuración que contenga la información requerida para la generación de datos sintéticos, como la ruta del conjunto de datos original y la configuración de los modelos a emplear.
2. Crear un script que cargue la configuración, ejecute el proceso de generación de datos sintéticos y almacene el conjunto de datos sintético resultante.
3. Poner en marcha el script creado en el paso previo.

De esta forma, es posible automatizar el proceso de generación de datos sintéticos y aprovechar una mayor capacidad de procesamiento, lo que puede mejorar el rendimiento del proceso y reducir los tiempos de ejecución.

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*, lo que permite una configuración general de las entradas y la selección de modelos.

Para ver como un ejemplo del uso, diríjase a Anexo A.12.

La Tabla 3.4 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 se detallan los modelos actualmente soportados en la clase *Synthetic* y su origen.

Tabla 3.5: Modelos Tabulares Soportados

Nombre Modelo	Fuente
copulagan	SDV [33]
tvae	SDV [33]
gaussiancopula	SDV [33]
ctgan	SDV [33]
tablepreset	SDV [33]
smote-enc	tabDDPM [35]
tddpm_mlp	tabDDPM [35]

Al ejecutar el script de generación de datos sintéticos, se crearán múltiples archivos en una carpeta. En la Tabla 3.3 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).

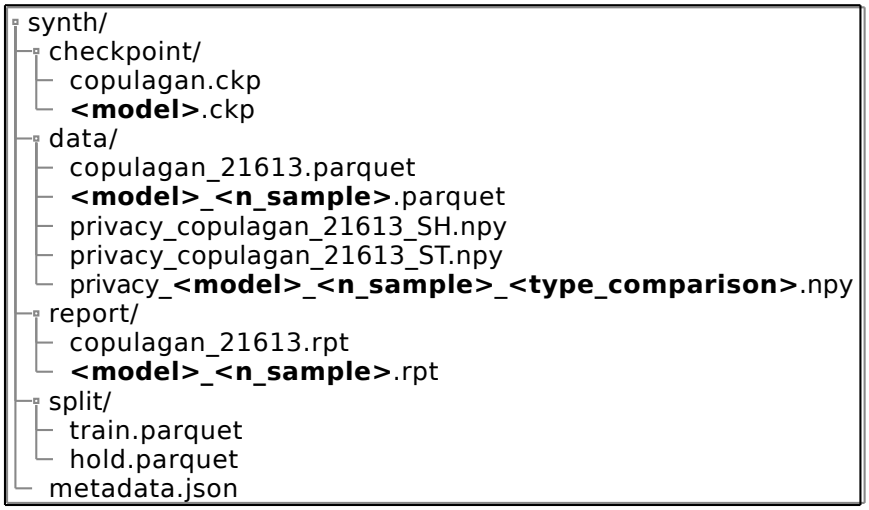

Figura 3.3: Carpetas y archivos esperados generados por *Synthetic*

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* [28]. 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>

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. Privacidad y sus metricas

Se exponen las métricas de Distancia al Registro más Cercano (DCR) y Relación de Distancia del Vecino más Cercano (NNDR), ambas para el absoluto, percentil 1 y percentil 5. Para mostrar que tan cerca están los elementos más cercanos. Adicionalmente mostrando algunos ejemplos de cada sección, así con datos se puede observar que diferencias se pueden apreciar para esas distancias.

Es cierto que existen metricas y tecnicas mencionadas en el marco teorico. Estas no figuran, primero por razones practicas, ya que los metodos establecidos para comprobar la privacidad por SDMetrics y el modelo de diffusión son las anteriores descritas y no otras las usadas y en segunda instancia por el objetivo de las tecnicas de privacidad diferencial *(Differential Privacy)* y k-*anonymity*.

Los conjuntos generados son del tamaño original o mayores. Esto disminuye las posibilidades de que elementos pequeños

DCR cuantifica la distancia euclidiana entre cualquier registro sintético y su vecino real más cercano. NNDR mide la relación entre la distancia euclidiana al vecino real más cercano y al segundo vecino real más cercano de cualquier registro sintético.

3.5. 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).

En la Tabla 3.6 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_frec)	[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
Continúa en la siguiente página	

Campo	Ejemplos
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
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
Coeficiente 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 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

Para ver como poder obtener estas métricas, diríjase a Anexo A.12.

3.6. Tiempo de Ejecución

En esta sección, se presenta una evaluación directa y técnica del tiempo de ejecución para cada script utilizado en la investigación. Los tiempos se midieron en un entorno controlado para garantizar consistencia y reproducibilidad. La tabla siguiente resume los tiempos de ejecución para cada técnica evaluada:

Script	**Tiempo Aprox**	**Observaciones**
kingcounty_run.py	6h	Se ejecutó 3 veces con parametros a-1, a-2, a-3, vram usada 4GB
economicos_run-a.py	9h	Se ejecutó 3 veces con parametros a-1, a-2, a-3, vram usada 4GB
economicos_run-b.py	9h	Se ejecutó 3 veces con parametros b-1, b-2, b-3, vram usada 6GB
economicos_text.py	15h	con parametro a-1, vram usada 23GB
economicos_text.py	50h	con parametro b-1, vram usada 23GB
economicos_text_gen.py	10h	con parametro a-1, vram usada 23GB
economicos_text_gen.py	20h	con parametro b-1, vram usada 23GB

Tabla 3.8: Tiempo de ejecución para cada técnica evaluada

Nota: Los tiempos de ejecución fueron medidos utilizando el software y hardware mencionados en 3.1.2.

Esta evaluación proporciona una referencia directa sobre la eficiencia en tiempo de cada técnica, ofreciendo una perspectiva crucial para su aplicabilidad en diferentes escenarios.

Capítulo 4

Resultados

Este capítulo aborda los resultados obtenidos en el actual trabajo, donde se emplearon diversas técnicas de preprocesamiento y modelos de aprendizaje automático. Aquí se presentan los resultados en función del desempeño de los modelos, la similitud con los datos originales y la tensión entre privacidad y utilidad de los datos generados.

Se enfocará en la evaluación de los conjuntos de datos de King County y Económicos, resaltando los logros de los modelos Tddpm y Smote en términos de similitud con los datos originales y cobertura. Se explorará además el análisis de privacidad, destacando el rendimiento superior del modelo Tddpm en términos de privacidad.

Finalmente, se hará un resumen de los hallazgos más relevantes, destacando la eficacia de los modelos Tddpm y Smote en la generación de datos sintéticos útiles, y se abordarán las diferencias significativas observadas en la cobertura, distribución y privacidad entre los conjuntos de datos.

4.1. King County

4.1.1. SDMetrics Score

La Tabla 4.1 muestra los puntajes obtenidos por los distintos patrones utilizados en este estudio. Es notorio que los patrones con puntajes más altos, como Tddpm y Smote, presentan una mayor similitud con el conjunto de datos original. En contraposición, los patrones con puntajes más bajos, como ctgan, exhiben una correspondencia considerablemente menor con el conjunto original. Se muestra el promedio ± 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	Synthesis	**Score**
tddpm_mlp	**9.45e-01±2.59e-03**	**9.71e-01±3.78e-04**	**9.65e-01±4.55e-03**	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	**9.58e-01±1.35e-03**
smote-enc	9.41e-01±2.59e-04	9.65e-01±2.71e-04	8.40e-01±6.44e-03	**1.00e+00±1.01e-05**	**1.00e+00±1.25e-04**	9.53e-01±5.05e-05
tablepreset	8.38e-01±0.00e+00	8.38e-01±6.41e-17	7.42e-01±0.00e+00	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	8.38e-01±1.57e-16
ctgan	8.03e-01±2.69e-03	8.20e-01±1.37e-02	8.58e-01±1.06e-03	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	8.12e-01±7.96e-03
copulagan	7.67e-01±3.77e-03	8.21e-01±7.68e-03	8.28e-01±4.42e-03	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	7.94e-01±5.71e-03
gaussiancopula	7.66e-01±0.00e+00	8.13e-01±1.28e-16	7.65e-01±6.41e-17	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	7.89e-01±1.57e-16
tvae	6.91e-01±1.94e-02	7.70e-01±1.56e-02	4.24e-01±2.07e-02	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	7.30e-01±1.75e-02

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

4.1.2. Correlación

En el Anexo A.3, se contrasta la lista completa de cada modelo. Se observa que, en general, los modelos con puntajes más altos exhiben una mayor similitud visual con los datos reales. A modo de ilustración, las imágenes 4.1 y 4.2 contrastan los datos reales con los generados por los modelos gaussiancopula y copulagan. A pesar de que estos modelos presentan puntajes similares, el modelo gaussiancopula muestra una mayor similitud visual con los datos reales en comparación con el modelo copulagan.

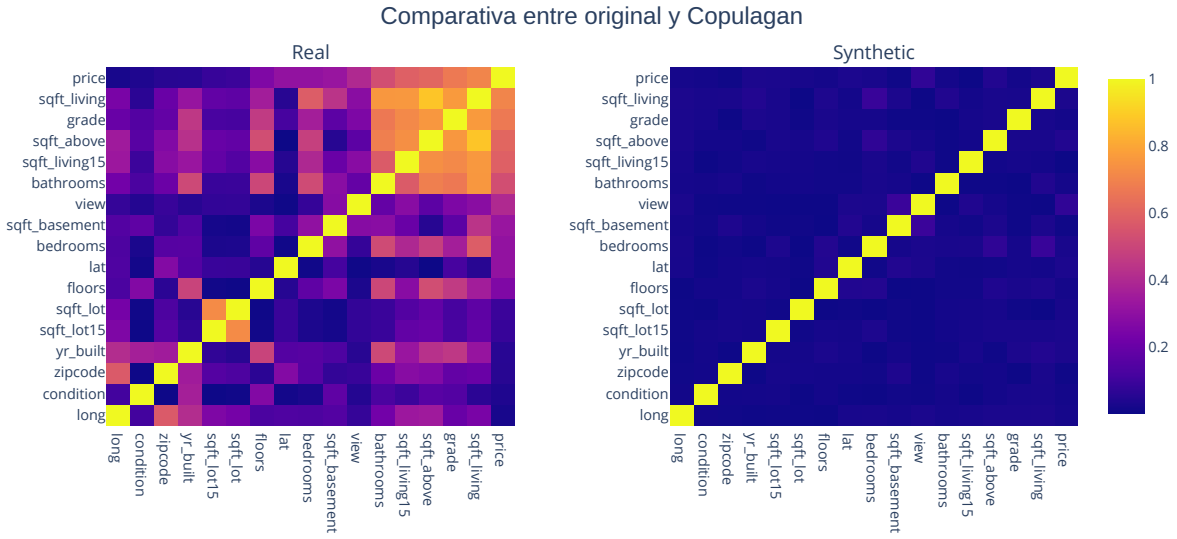

Figura 4.1: Correlación de conjunto original de entrenamiento y Copulagan, King county (A-2)

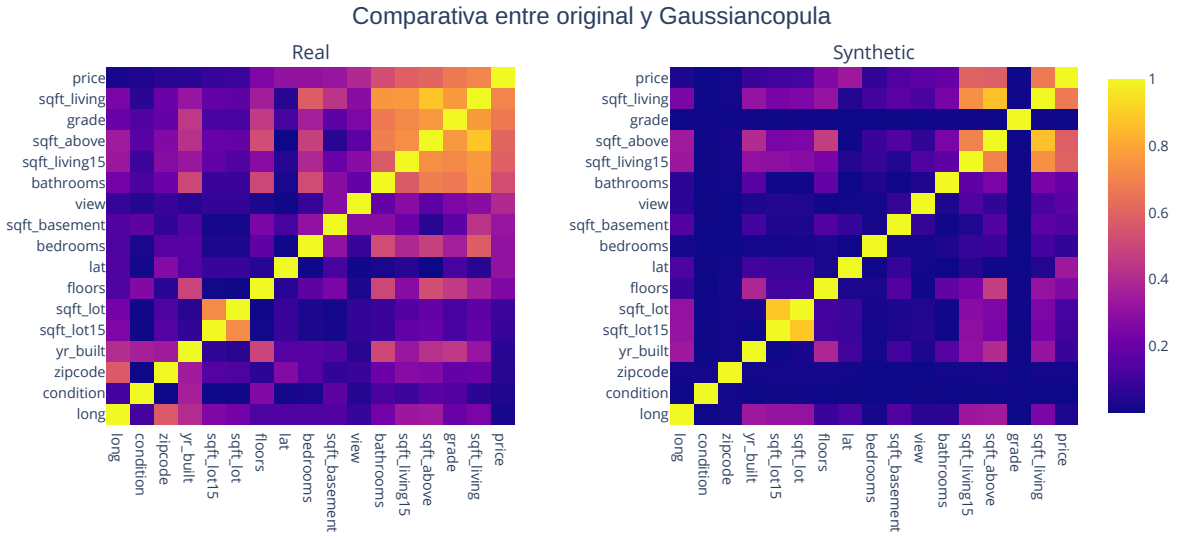

Figura 4.2: Correlación de conjunto original de entrenamiento y Gaussiancopula, King county (A-2)

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

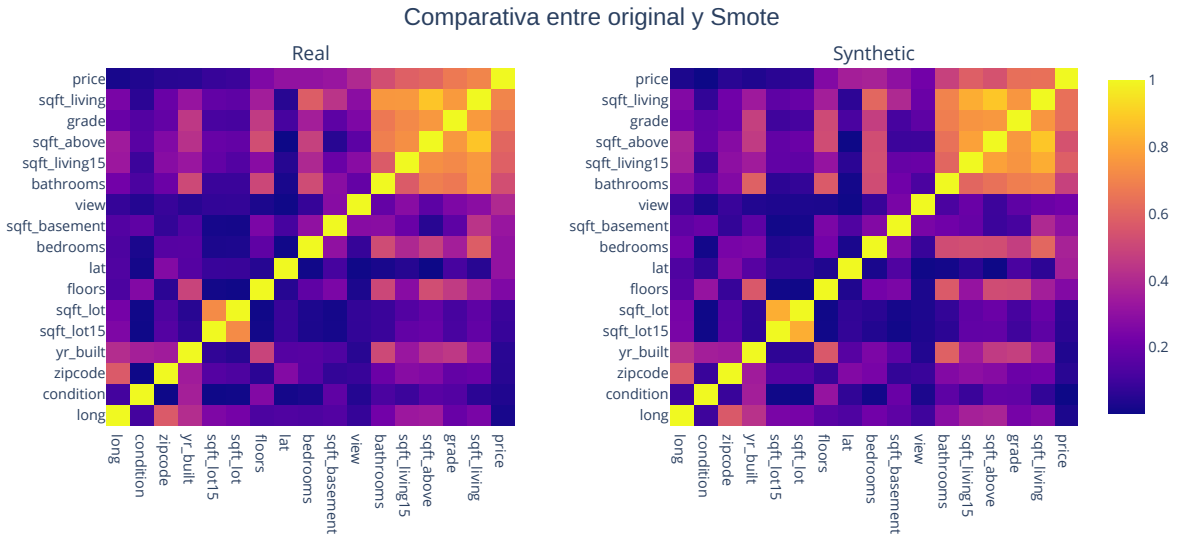

Figura 4.3: Correlación de conjunto original de entrenamiento y Smote, King county (A-2)

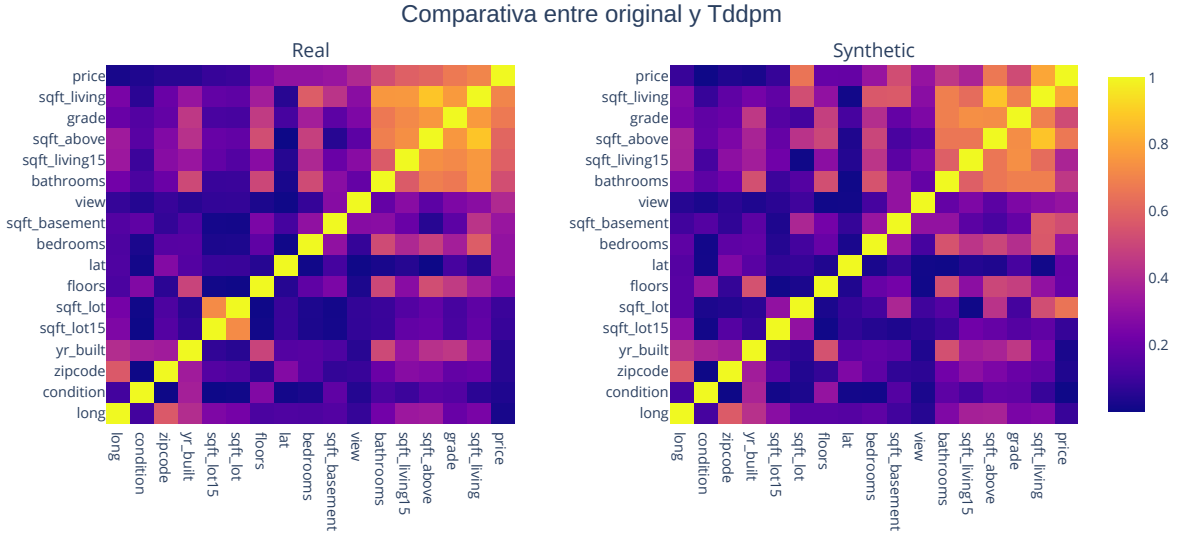

Figura 4.4: Correlación de conjunto original de entrenamiento y Tddpm, King county (A-2)

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

4.1.3. Reporte diagnóstico

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

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

Columna	Metrica	smote-enc	tddpm_mlp
bathrooms	CategoryCoverage	6.56e-01±1.57e-02	**8.33e-01±4.71e-02**
bedrooms	CategoryCoverage	5.64e-01±7.25e-02	**6.92e-01±6.28e-02**
condition	CategoryCoverage	8.00e-01±1.11e-16	**1.00e+00±0.00e+00**
date	CategoryCoverage	**9.70e-01±1.27e-03**	9.36e-01±1.10e-02
floors	CategoryCoverage	8.33e-01±0.00e+00	**1.00e+00±0.00e+00**
grade	CategoryCoverage	7.50e-01±0.00e+00	**8.33e-01±0.00e+00**
id	RangeCoverage	9.94e-01±2.27e-04	**1.00e+00±1.95e-04**
lat	RangeCoverage	9.67e-01±1.26e-02	**9.95e-01±6.95e-03**
long	RangeCoverage	9.91e-01±1.79e-03	**1.00e+00±1.60e-04**
price	RangeCoverage	5.06e-01±2.52e-02	**1.00e+00±4.83e-05**
sqft_above	RangeCoverage	8.85e-01±3.90e-02	**1.00e+00±4.84e-04**
sqft_basement	RangeCoverage	7.35e-01±5.73e-02	**1.00e+00±0.00e+00**
sqft_living	RangeCoverage	6.61e-01±6.17e-02	**1.00e+00±3.89e-04**
sqft_living15	RangeCoverage	9.19e-01±2.02e-02	**1.00e+00±6.67e-04**
sqft_lot	RangeCoverage	5.32e-01±4.96e-02	**1.00e+00±2.54e-05**
sqft_lot15	RangeCoverage	8.87e-01±8.32e-02	**9.83e-01±2.19e-02**
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±2.32e-04**	**1.00e+00±0.00e+00**
yr_renovated	RangeCoverage	**1.00e+00±5.80e-05**	**1.00e+00±0.00e+00**
zipcode	CategoryCoverage	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**

4.1.4. Reporte de calidad

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

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

Columna	Metrica	smote-enc	tddpm_mlp
bathrooms	TVComplement	8.84e-01±3.58e-03	**9.45e-01±3.05e-03**
bedrooms	TVComplement	9.20e-01±2.30e-03	**9.46e-01±2.27e-03**
condition	TVComplement	9.35e-01±3.21e-03	**9.61e-01±6.02e-04**
date	TVComplement	**9.38e-01±2.83e-03**	9.28e-01±1.57e-03
floors	TVComplement	9.65e-01±1.98e-03	**9.67e-01±1.24e-03**
grade	TVComplement	9.57e-01±4.16e-04	**9.63e-01±2.16e-03**
id	KSComplement	**9.87e-01±3.42e-04**	9.80e-01±1.51e-03
lat	KSComplement	**9.92e-01±6.06e-04**	9.87e-01±1.68e-03
long	KSComplement	**9.88e-01±2.57e-03**	9.78e-01±6.00e-04
price	KSComplement	**9.80e-01±1.15e-03**	9.78e-01±3.15e-03
sqft_above	KSComplement	9.76e-01±7.99e-04	**9.83e-01±1.72e-03**
sqft_basement	KSComplement	9.34e-01±4.03e-03	**9.77e-01±6.61e-03**
sqft_living	KSComplement	9.78e-01±5.10e-04	**9.82e-01±2.84e-04**
sqft_living15	KSComplement	9.81e-01±7.03e-04	**9.82e-01±1.29e-03**
sqft_lot	KSComplement	**9.83e-01±1.06e-03**	9.78e-01±4.42e-03
sqft_lot15	KSComplement	**9.85e-01±1.36e-03**	9.79e-01±8.72e-04
view	TVComplement	9.36e-01±1.33e-03	**9.53e-01±1.19e-03**
waterfront	TVComplement	9.94e-01±2.73e-04	**9.95e-01±4.16e-04**
yr_built	KSComplement	**9.83e-01±1.49e-03**	9.73e-01±2.48e-03
yr_renovated	KSComplement	**9.92e-01±8.12e-04**	**9.92e-01±1.38e-03**
zipcode	TVComplement	**9.75e-01±2.48e-03**	9.56e-01±2.36e-03

Al examinar las variables de los conjuntos de datos completos, como se ilustra en la lista Anexa A.3, se observa una similitud entre los tres conjuntos analizados: Real, Smote y Tddpm. Sin embargo, también surgen diferencias significativas. Es relevante mencionar que los conjuntos de datos generados son aproximadamente un 20 % más grandes que el conjunto real. En varias columnas, la distribución de datos en los tres conjuntos es similar, como se evidencia en los casos de bathrooms, sqft_lot, sqft_above, price, sqft_living, sqft_basement, yr_built, sqft_living15 y grade. Este patrón se puede apreciar en la Figura A.22.

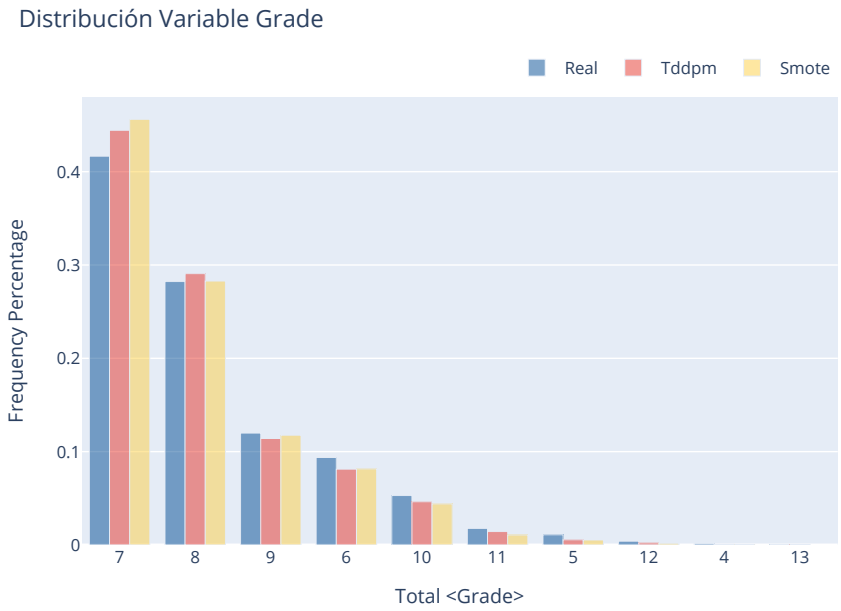

Figura 4.5: Frecuencia del campo Grade en el modelo real y Top 2, King county (A-2)

Por otra parte, la distribución de los atributos bedrooms, condition, view y floors en el conjunto de datos generado por el modelo Tddpm presenta una particularidad: contiene un mayor número de elementos menos frecuentes comparado con los demás conjuntos. Al considerar la columna *bedrooms* como ejemplo (refiérase a Figura A.17), la distribución de valores en el conjunto Tddpm se desvía de la del conjunto Smote. En específico, se registra un aumento en la cantidad de registros correspondientes a los valores 6 y 1.

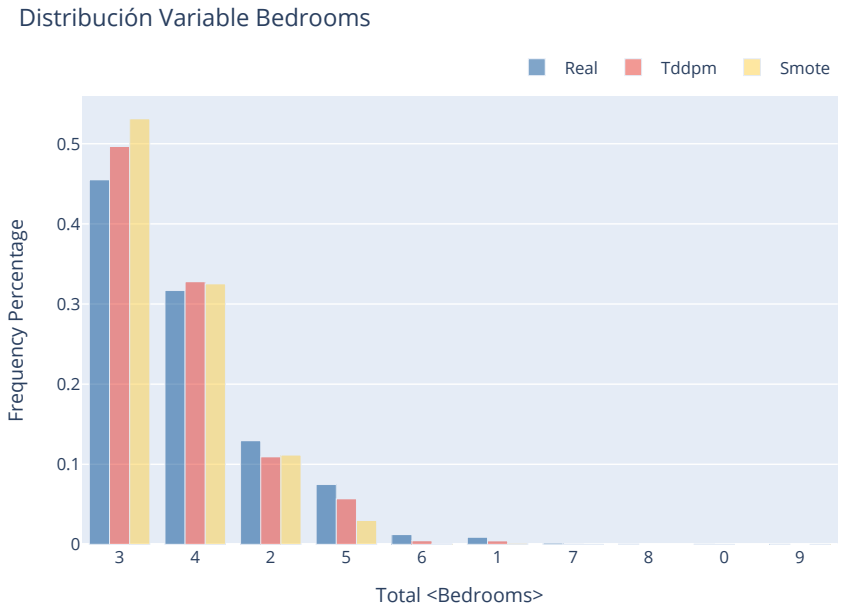

Figura 4.6: Frecuencia del campo Bedrooms en el modelo real y Top 2, King county (A-2)

En el caso de la variable *sqft_lot15*, la distribución generada por el modelo Smote resulta ser más similar a la del conjunto de datos real, como se puede apreciar en la figura A.23.

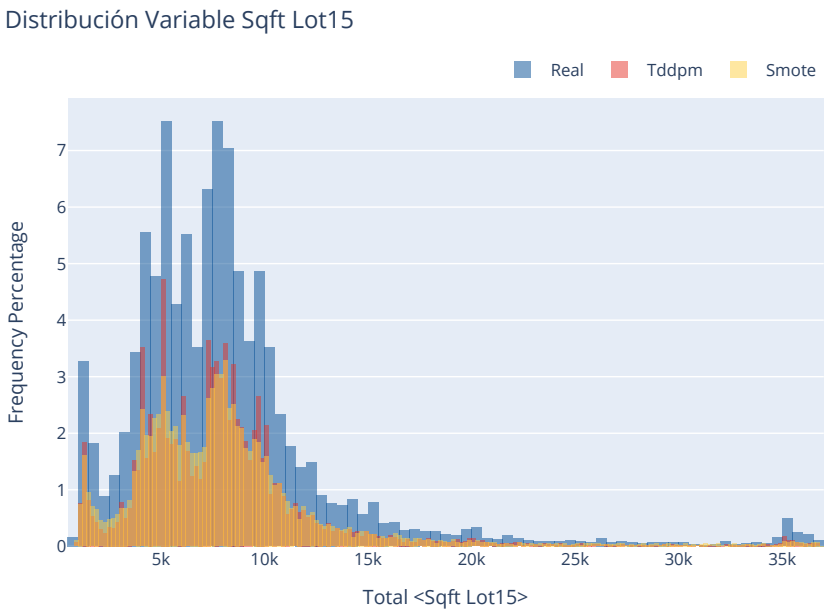

Figura 4.7: Frecuencia del campo Sqft lot15 en el modelo real y Top 2, King county (A-2)

4.1.5. Privacidad

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

Se presentan a continuación tablas comparativas que incluyen tres conjuntos de datos: el conjunto sintético frente al utilizado para el entrenamiento (denominado ST, por Synthetic-Train), la comparación del conjunto sintético con el conjunto de reserva, no utilizado en el entrenamiento (SH, Synthetic-Hold), y la comparación entre el conjunto de entrenamiento y el conjunto de reserva (TH, Train-Hold). Cabe destacar que tanto el conjunto de entrenamiento como el de reserva son conjuntos reales.

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

Modelo	NNDR ST	NNDR SH	NNDR TH	**Score**
tddpm_mlp	6.12e-01±3.23e-03	6.05e-01±4.01e-03	**3.76e-01±0.00e+00**	**9.58e-01±1.35e-03**
smote-enc	1.92e-01±7.02e-03	4.19e-01±4.47e-03	**3.76e-01±0.00e+00**	9.53e-01±5.05e-05
ctgan	8.10e-01±1.98e-03	8.16e-01±4.08e-03	**3.76e-01±0.00e+00**	8.12e-01±7.96e-03
tablepreset	**8.29e-01±6.41e-17**	8.13e-01±0.00e+00	**3.76e-01±0.00e+00**	8.38e-01±9.06e-17
copulagan	8.20e-01±3.87e-03	**8.21e-01±3.94e-03**	**3.76e-01±0.00e+00**	7.94e-01±5.71e-03
gaussiancopula	7.61e-01±9.06e-17	7.59e-01±0.00e+00	**3.76e-01±0.00e+00**	7.89e-01±1.57e-16
tvae	7.33e-01±4.95e-03	6.90e-01±8.75e-03	**3.76e-01±0.00e+00**	7.30e-01±1.75e-02

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

Modelo	NNDR ST	NNDR SH	NNDR TH	**Score**
tddpm_mlp	4.53e-01±4.15e-03	4.50e-01±8.01e-03	**7.98e-02±8.01e-18**	**9.58e-01±1.35e-03**
smote-enc	6.30e-02±1.81e-03	1.92e-01±6.65e-03	**7.98e-02±8.01e-18**	9.53e-01±5.05e-05
ctgan	7.17e-01±4.34e-03	7.18e-01±3.84e-03	**7.98e-02±8.01e-18**	8.12e-01±7.96e-03
tablepreset	7.19e-01±6.41e-17	7.11e-01±1.11e-16	**7.98e-02±8.01e-18**	8.38e-01±9.06e-17
copulagan	**7.44e-01±5.92e-03**	**7.47e-01±4.75e-03**	**7.98e-02±8.01e-18**	7.94e-01±5.71e-03
gaussiancopula	6.51e-01±0.00e+00	6.51e-01±9.06e-17	**7.98e-02±8.01e-18**	7.89e-01±1.57e-16
tvae	5.89e-01±6.63e-03	5.69e-01±8.57e-03	**7.98e-02±8.01e-18**	7.30e-01±1.75e-02

Tabla 4.6: Proporción entre el más cercano y el segundo más cercano, mínimo, King county

Modelo	NNDR ST	NNDR SH	NNDR TH	**Score**
tddpm_mlp	1.43e-01±1.03e-02	1.44e-01±3.92e-02	**0.00e+00±0.00e+00**	**9.58e-01±1.35e-03**
smote-enc	0.00e+00±0.00e+00	8.62e-03±2.30e-03	**0.00e+00±0.00e+00**	9.53e-01±5.05e-05
ctgan	4.12e-01±7.51e-02	4.22e-01±6.20e-03	**0.00e+00±0.00e+00**	8.12e-01±7.96e-03
tablepreset	3.97e-01±3.20e-17	3.59e-01±6.41e-17	**0.00e+00±0.00e+00**	8.38e-01±9.06e-17
copulagan	**5.43e-01±1.28e-02**	**5.18e-01±2.09e-02**	**0.00e+00±0.00e+00**	7.94e-01±5.71e-03
gaussiancopula	3.81e-01±0.00e+00	3.58e-01±4.53e-17	**0.00e+00±0.00e+00**	7.89e-01±1.57e-16
tvae	3.27e-01±2.64e-02	2.89e-01±3.13e-02	**0.00e+00±0.00e+00**	7.30e-01±1.75e-02

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

Tabla 4.7: Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 5, King county

Modelo	DCR ST	DCR SH	DCR TH	**Score**
tddpm_mlp	5.82e-02±6.47e-04	7.62e-02±5.61e-04	**3.57e-02±0.00e+00**	**9.58e-01±1.35e-03**
smote-enc	7.12e-03±2.73e-04	3.72e-02±4.54e-04	**3.57e-02±0.00e+00**	9.53e-01±5.05e-05
ctgan	2.21e-01±3.48e-03	2.45e-01±4.59e-03	**3.57e-02±0.00e+00**	8.12e-01±7.96e-03
tablepreset	1.78e-01±0.00e+00	1.97e-01±0.00e+00	**3.57e-02±0.00e+00**	8.38e-01±9.06e-17
copulagan	**3.71e-01±1.23e-02**	**4.11e-01±1.53e-02**	**3.57e-02±0.00e+00**	7.94e-01±5.71e-03
gaussiancopula	2.69e-01±3.20e-17	3.12e-01±0.00e+00	**3.57e-02±0.00e+00**	7.89e-01±1.57e-16
tvae	7.94e-02±3.74e-04	9.45e-02±1.12e-03	**3.57e-02±0.00e+00**	7.30e-01±1.75e-02

Tabla 4.8: Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 1, King county

Modelo	DCR ST	DCR SH	DCR TH	**Score**
tddpm_mlp	4.00e-02±5.96e-04	5.19e-02±8.09e-04	**7.94e-03±0.00e+00**	**9.58e-01±1.35e-03**
smote-enc	1.94e-03±6.88e-05	1.50e-02±2.48e-04	**7.94e-03±0.00e+00**	9.53e-01±5.05e-05
ctgan	1.72e-01±2.57e-03	1.93e-01±3.65e-03	**7.94e-03±0.00e+00**	8.12e-01±7.96e-03
tablepreset	1.39e-01±1.60e-17	1.58e-01±0.00e+00	**7.94e-03±0.00e+00**	8.38e-01±9.06e-17
copulagan	**3.25e-01±9.19e-03**	**3.58e-01±1.12e-02**	**7.94e-03±0.00e+00**	7.94e-01±5.71e-03
gaussiancopula	2.11e-01±1.60e-17	2.48e-01±1.60e-17	**7.94e-03±0.00e+00**	7.89e-01±1.57e-16
tvae	6.13e-02±5.17e-04	7.29e-02±5.38e-04	**7.94e-03±0.00e+00**	7.30e-01±1.75e-02

Tabla 4.9: Proporción entre el más cercano y el segundo más cercano, minimo, datos king county

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

En la Figura A.8 solo se consideran los modelos Tddpm y Smote para su comparación. En ambos casos, existe una distancia mayor a cero. Sin embargo, esta distancia es mayor en el caso de Tddpm, lo que sugiere que este conjunto puede ser considerado superior en términos de privacidad.

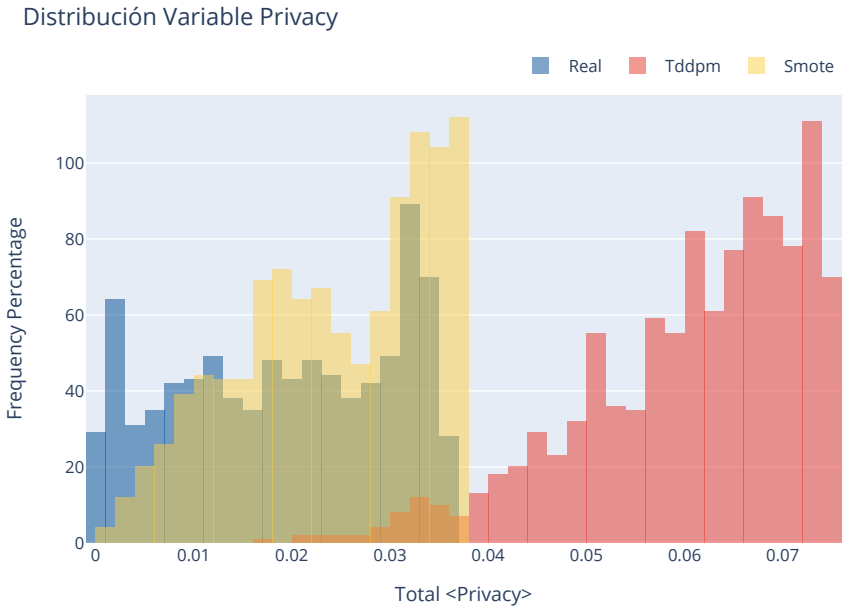

Figura 4.8: Frecuencia del campo Privacy en el modelo real y Top 2, King county (A-1)

4.1.6. Ejemplo de registros

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

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

Tabla 4.10: Ejemplos para el modelo smote-enc, minimo, King county (A-1)

Variable/Distancia	Sintético	DCR1 d(0.00e+00)	DCR2 d(9.23e-03)
sqft_lot15	9324.000000	9324.000000	9420.000000
id	1788800630.000000	1788800630.000000	1788900230.000000
sqft_above	840.000000	840.000000	840.000000
sqft_lot	12091.000000	12091.000000	9480.000000
view	0	0	0
yr_built	1959.000000	1959.000000	1960.000000
long	-122.343000	-122.343000	-122.341000
bathrooms	1.000000	1.000000	1.000000
zipcode	98023	98023	98023
waterfront	0	0	0
grade	6	6	6
sqft_living	840.000000	840.000000	840.000000
condition	3	3	3
floors	1.000000	1.000000	1.000000
bedrooms	3	3	3
yr_renovated	0.000000	0.000000	0.000000
date	20140722T000000	20150225T000000	20150403T000000
price	185000.000000	185000.000000	199950.000000
sqft_basement	0.000000	0.000000	0.000000
lat	47.328100	47.328100	47.327700
sqft_living15	840.000000	840.000000	840.000000

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

Tabla 4.11: Ejemplos para el modelo tddpm_mlp, minimo, King county (A-1)

Variable/Distancia	Sintético	DCR1 d(1.31e-02)	DCR2 d(1.76e-02)
id	8644685321.218300	8682231210.000000	8682261140.000000
sqft_living	1820.000000	1870.000000	1690.000000
sqft_lot	5804.228950	5580.000000	4500.000000
sqft_above	1770.000000	1870.000000	1690.000000
sqft_basement	0.000000	0.000000	0.000000
yr_built	2004.000000	2004.000000	2004.000000
yr_renovated	0.000000	0.000000	0.000000
lat	47.709128	47.710100	47.713300
long	-122.033000	-122.031000	-122.031000
sqft_living15	1680.000000	1670.000000	1640.000000
sqft_lot15	5200.000000	4500.000000	4500.000000
price	525000.000000	554000.000000	564000.000000
date	20140826T000000	20140805T000000	20140618T000000
bedrooms	2	2	2
bathrooms	2.000000	2.000000	2.000000
floors	1.000000	1.000000	1.000000
waterfront	0	0	0
view	0	0	0
condition	3	3	3
grade	8	8	8
zipcode	98053	98053	98053

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

Tabla 4.12: Ejemplos para el modelo smote-enc, percentil 1, King county (A-1)

Variable/Distancia	Sintético	DCR1 d(2.03e-03)	DCR2 d(2.66e-03)
sqft_lot15	10520.528443	10050.000000	10565.000000
id	1771000590.684903	1771000690.000000	1771000440.000000
sqft_above	1160.000000	1160.000000	1160.000000
sqft_lot	9506.898249	11776.000000	9750.000000
view	0	0	0
yr_built	1967.879322	1968.000000	1968.000000
long	-122.073000	-122.074000	-122.072000
bathrooms	1.000000	1.000000	1.000000
zipcode	98077	98077	98077
waterfront	0	0	0
grade	7	7	7
sqft_living	1160.000000	1160.000000	1160.000000
condition	3	3	3
floors	1.000000	1.000000	1.000000
bedrooms	3	3	3
yr_renovated	0.000000	0.000000	0.000000
date	20140528T000000	20140528T000000	20140904T000000
price	305000.000000	305000.000000	322500.000000
sqft_basement	0.000000	0.000000	0.000000
lat	47.742955	47.742700	47.742900
sqft_living15	1160.000000	1160.000000	1160.000000

4.1.7. Propiedades estadísticas

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

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

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

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
nobs	17290	**21613**	21614	**21613**
mean	3.368	**3.339**	3.273	3.755
std_err	0.007	**0.005**	0.005	0.021
upper_ci	3.382	**3.349**	3.283	3.795
lower_ci	3.354	**3.328**	3.264	3.714
std	0.931	**0.785**	0.705	3.036
mad	0.734	**0.650**	0.581	1.327
mad_normal	0.920	**0.815**	0.728	1.663
coef_var	0.277	**0.235**	0.215	0.809
range	33.000	7.000	8.000	**33.000**
max	33.000	7.000	9.000	**33.000**
min	0.000	**0.000**	1.000	**0.000**
skew	2.304	**0.254**	0.137	7.567
kurtosis	63.268	3.416	3.222	**72.734**
jarque_bera	2631992	389	112	**4585387**
jarque_bera_pval	0.000	0.000	0.000	**0.000**
mode_freq	0.455	0.497	0.531	**0.414**
95.0 %	5.000	**5.000**	4.000	6.000
99.0 %	6.000	**5.000**	**5.000**	9.000
99.9 %	7.000	**6.000**	5.000	33.000

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

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

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
nobs	17290	**21613**	21614	**21613**
mean	2.114	**2.073**	2.007	2.364
std_err	0.006	0.005	0.005	**0.006**
upper_ci	2.125	**2.083**	2.016	2.376
lower_ci	2.102	**2.064**	1.997	2.351
std	0.767	**0.720**	0.703	0.935
mad	0.615	0.586	**0.588**	0.732
mad_normal	0.771	0.734	**0.737**	0.918
coef_var	0.363	0.347	**0.351**	0.396
range	8.000	**8.000**	5.250	**8.000**
max	8.000	**8.000**	6.000	**8.000**
min	0.000	**0.000**	0.750	**0.000**
skew	0.464	0.335	0.142	**0.381**
kurtosis	3.989	**3.881**	2.798	3.556
jarque_bera	1326	**1103**	110	801
jarque_bera_pval	0.000	**0.000**	0.000	0.000
mode_freq	0.251	**0.282**	0.302	0.204
median	2.250	**2.250**	2.000	2.500
0.1 %	0.750	**0.750**	1.000	0.000
95.0 %	3.500	**3.250**	**3.250**	4.000
99.0 %	4.250	**4.000**	3.500	4.750
99.9 %	5.428	5.097	4.500	**5.750**

4.1.8. Resumen de resultados

En esta sección, se proporciona un resumen de los hallazgos más significativos tras el análisis de los resultados obtenidos de los modelos Tddpm y Smote.

1. Los modelos Tddpm y Smote obtienen los puntajes más altos en la evaluación de métricas de rendimiento, mostrando su eficacia para la generación de datos sintéticos beneficiosos para este conjunto de datos (Sección 4.1.1).
2. A pesar de no presentar diferencias visuales destacables, Tddpm exhibe una mayor cobertura de valores distintos en comparación con Smote (Sección 4.1.2 y Sección 4.1.3).
3. La distribución de los datos generados por Tddpm y Smote se aproxima a la distribución real en la mayoría de las variables, demostrando su utilidad en la simulación de los patrones de los datos reales (Sección 4.1.4).
4. Tddpm ofrece una mayor privacidad en comparación con Smote, como se evidencia por su mayor distancia entre el vecino más cercano y el segundo vecino más cercano (Sección 4.1.5).

5. En términos de similitud con las propiedades estadísticas del conjunto original, Tddpm se destaca en las variables *bedrooms* y *bathrooms*, con excepciones notables en algunas métricas específicas (Sección 4.1.7).

4.2. Conjunto de datos proveniente de Economicos

4.2.1. Tratamiento de nulos en conjunto A y B

El conjunto de Económicos, a diferencia del conjunto de datos de King County que fue filtrado y preprocesado para evitar valores nulos, contiene elementos nulos. A continuación se describen dos tratamientos de estos elementos nulos. El primer enfoque simplemente elimina todos los registros que contienen un registro vacío utilizando el método 'dropna', como se muestra en el Código 1; 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 2; este será considerado como el Conjunto B.

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

Código 1: Eliminación de valores nulos en el conjunto de datos de Económicos

```
df_converted = df.fillna(dict(
            property_type = "None",
            transaction_type = "None",
            state = "None",
            county = "None",
            rooms = -1,
            bathrooms = -1,
            m_built = -1,
            m_size = -1,
            source = "None"
    )).fillna(-1).astype({k: 'str' for k in ("description", "price", "title",
    ↪  "address", "owner",)})
basedate = pd.Timestamp('2017-12-01')
dtime = df_converted.pop("publication_date")
df_converted["publication_date"] = dtime.apply(lambda x: (x - basedate).days)
```

Código 2: Reemplazo de valores nulos en el conjunto de datos de Económicos

4.2.2. SDMetrics Score - Conjunto A

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

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

Model Name	Column Pair Trends	Column Shapes	Coverage	Boundaries	Synthesis	**Score**
tddpm_mlp	**9.72e-01±1.50e-03**	**9.83e-01±1.09e-03**	**8.12e-01±1.89e-02**	**1.00e+00±0.00e+00**	9.90e-01±8.52e-04	**9.77e-01±6.88e-04**
smote-enc	9.59e-01±1.20e-03	9.76e-01±4.34e-04	6.27e-01±1.31e-02	**1.00e+00±0.00e+00**	9.24e-01±1.97e-03	9.67e-01±8.19e-04
copulagan	7.60e-01±1.58e-02	8.02e-01±2.69e-02	6.80e-01±6.95e-03	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	7.81e-01±2.03e-02
ctgan	7.43e-01±1.27e-02	6.49e-01±1.69e-02	6.76e-01±7.85e-04	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	6.96e-01±1.00e-02
gaussiancopula	6.94e-01±0.00e+00	6.87e-01±6.41e-17	5.52e-01±9.06e-17	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	6.91e-01±6.41e-17
tvae	5.95e-01±4.07e-03	6.85e-01±2.67e-03	9.53e-02±1.48e-03	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	6.40e-01±3.35e-03

4.2.3. Correlación - Conjunto A

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

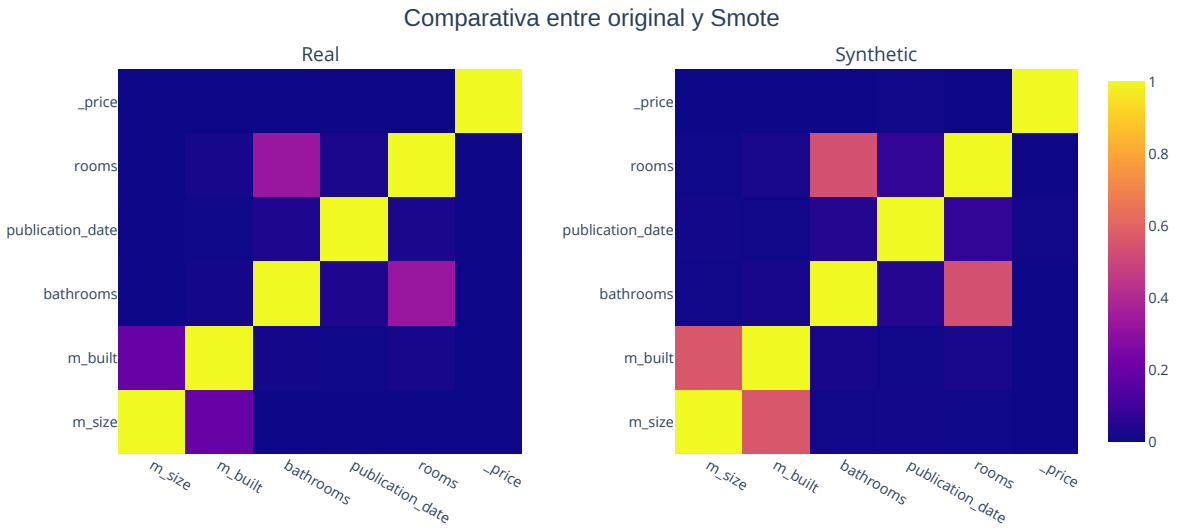

Figura 4.9: Correlación de conjunto original de entrenamiento y Smote, Economicos (A-2)

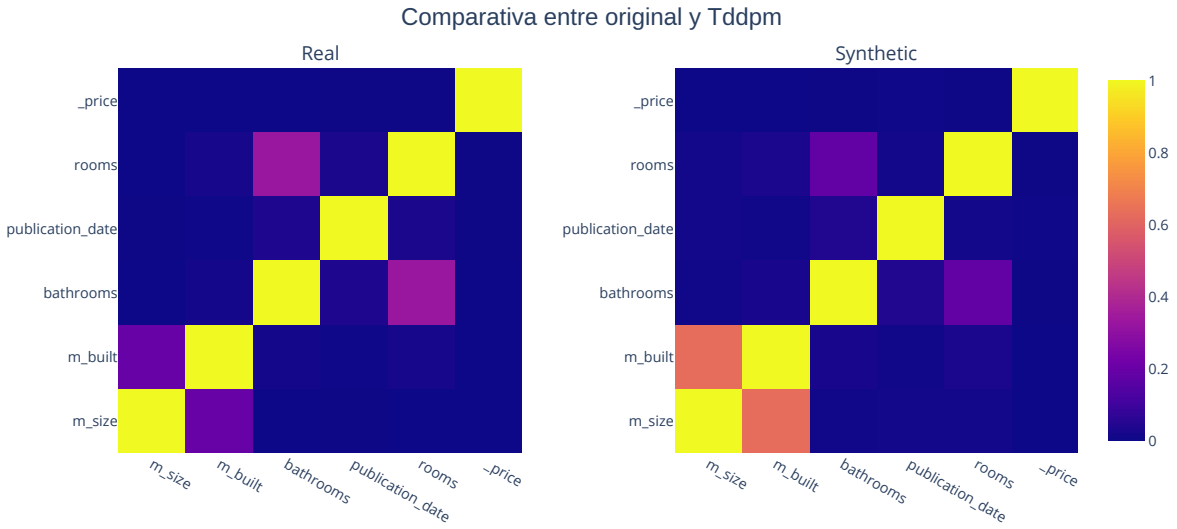

Figura 4.10: Correlación de conjunto original de entrenamiento y Tddpm, Economicos (A-2)

4.2.4. Reporte diagnóstico - Conjunto A

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

Tabla 4.16: Cobertura Categoría/Rango para Modelos Smote y Tddpm, Economicos

Columna	Metrica	smote-enc	tddpm_mlp
_price	RangeCoverage	8.10e-01±1.34e-01	**9.11e-01±1.37e-02**
bathrooms	CategoryCoverage	**8.63e-01±5.00e-02**	6.67e-01±1.39e-02
county	CategoryCoverage	5.90e-01±3.05e-03	**7.99e-01±2.20e-02**
m_built	RangeCoverage	3.18e-01±1.01e-01	**7.54e-01±1.77e-01**
m_size	RangeCoverage	3.45e-02±1.98e-03	**4.00e-01±1.51e-01**
property_type	CategoryCoverage	6.30e-01±5.24e-02	**9.07e-01±5.24e-02**
publication_date	RangeCoverage	9.77e-01±6.18e-03	**9.88e-01±4.44e-03**
rooms	CategoryCoverage	7.56e-01±3.98e-02	**7.97e-01±3.04e-02**
state	CategoryCoverage	7.92e-01±2.95e-02	**9.79e-01±2.95e-02**
transaction_type	CategoryCoverage	5.00e-01±0.00e+00	**9.17e-01±1.18e-01**

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

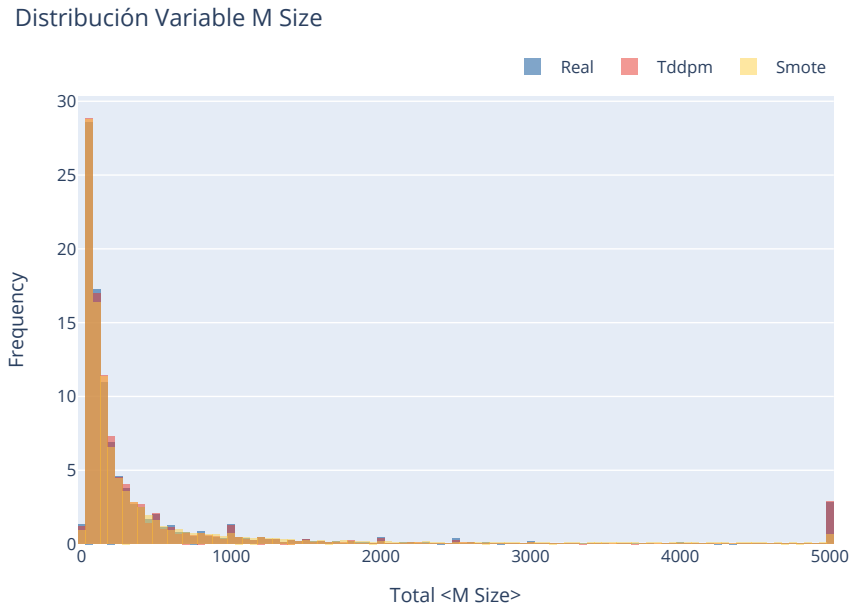

Figura 4.11: Frecuencia del campo M size en el modelo real y Top 2, Economicos (A-2)

4.2.5. Reporte de calidad - Conjunto A

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

Tabla 4.17: Evaluación de Similitud de Distribución para Modelos Smote y Tddpm, Economicos

Columna	Metrica	smote-enc	tddpm_mlp
_price	KSComplement	**9.91e-01±3.85e-04**	9.84e-01±3.53e-03
bathrooms	TVComplement	**9.94e-01±6.66e-04**	9.87e-01±2.15e-03
county	TVComplement	9.22e-01±9.28e-04	**9.66e-01±2.10e-03**
m_built	KSComplement	**9.87e-01±2.14e-03**	**9.87e-01±1.11e-03**
m_size	KSComplement	9.72e-01±7.43e-04	**9.84e-01±3.22e-03**
property_type	TVComplement	9.67e-01±1.33e-03	**9.82e-01±9.49e-04**
publication_date	KSComplement	9.80e-01±1.61e-03	**9.85e-01±1.61e-03**
rooms	TVComplement	9.77e-01±2.28e-03	**9.81e-01±3.18e-03**
state	TVComplement	9.69e-01±4.29e-04	**9.83e-01±1.05e-03**
transaction_type	TVComplement	**9.98e-01±1.07e-03**	9.93e-01±3.54e-03

4.2.6. Privacidad - Conjunto A

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

Tabla 4.18: Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 5, Economicos

Modelo	DCR ST	DCR SH	DCR TH	**Score**
tddpm_mlp	4.29e-09±2.16e-10	3.50e-08±1.92e-09	**1.28e-08±0.00e+00**	**9.77e-01±6.88e-04**
smote-enc	2.90e-11±1.13e-12	4.41e-08±2.36e-09	**1.28e-08±0.00e+00**	9.67e-01±8.19e-04
ctgan	**7.59e-06±5.75e-06**	**1.91e-05±2.01e-05**	**1.28e-08±0.00e+00**	6.96e-01±1.00e-02
copulagan	1.27e-06±3.04e-07	2.73e-06±5.89e-07	**1.28e-08±0.00e+00**	7.81e-01±2.03e-02
gaussiancopula	5.11e-06±0.00e+00	8.25e-06±0.00e+00	**1.28e-08±0.00e+00**	6.91e-01±6.41e-17
tvae	2.19e-07±1.60e-09	4.15e-07±5.43e-09	**1.28e-08±0.00e+00**	6.40e-01±3.35e-03

Tabla 4.19: Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 1, Economicos

Modelo	DCR ST	DCR SH	DCR TH	**Score**
tddpm_mlp	1.44e-10±6.01e-12	1.40e-09±1.05e-10	**0.00e+00±0.00e+00**	**9.77e-01±6.88e-04**
smote-enc	0.00e+00±0.00e+00	1.41e-09±4.21e-10	**0.00e+00±0.00e+00**	9.67e-01±8.19e-04
ctgan	**2.20e-06±1.50e-06**	**3.24e-06±1.58e-06**	**0.00e+00±0.00e+00**	6.96e-01±1.00e-02
copulagan	2.04e-07±2.85e-08	4.37e-07±6.38e-08	**0.00e+00±0.00e+00**	7.81e-01±2.03e-02
gaussiancopula	8.04e-07±8.64e-23	1.93e-06±0.00e+00	**0.00e+00±0.00e+00**	6.91e-01±6.41e-17
tvae	7.41e-08±1.95e-09	1.16e-07±3.43e-09	**0.00e+00±0.00e+00**	6.40e-01±3.35e-03

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

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

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

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

Modelo	NNDR ST	NNDR SH	NNDR TH	**Score**
tddpm_mlp	6.79e-02±7.37e-04	**1.00e-01±2.26e-03**	**1.31e-02±0.00e+00**	9.77e-01±6.88e-04
smote-enc	**7.15e-04±7.49e-06**	1.14e-01±4.79e-03	1.31e-02±0.00e+00	9.67e-01±8.19e-04
ctgan	2.57e-01±8.81e-03	3.27e-01±4.72e-02	1.31e-02±0.00e+00	6.96e-01±1.00e-02
copulagan	2.01e-01±1.27e-02	2.23e-01±5.47e-02	1.31e-02±0.00e+00	7.81e-01±2.03e-02
gaussiancopula	3.07e-01±0.00e+00	2.76e-01±0.00e+00	1.31e-02±0.00e+00	6.91e-01±6.41e-17
tvae	3.02e-01±6.15e-03	3.49e-01±1.26e-03	1.31e-02±0.00e+00	**6.40e-01±3.35e-03**

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

Modelo	NNDR ST	NNDR SH	NNDR TH	**Score**
tddpm_mlp	2.52e-03±1.71e-04	9.61e-03±1.17e-04	**0.00e+00±0.00e+00**	9.77e-01±6.88e-04
smote-enc	**0.00e+00±0.00e+00**	**3.00e-03±1.28e-03**	0.00e+00±0.00e+00	9.67e-01±8.19e-04
ctgan	1.35e-02±2.69e-03	2.77e-02±1.69e-02	0.00e+00±0.00e+00	6.96e-01±1.00e-02
copulagan	1.03e-02±1.29e-03	1.11e-02±2.64e-03	0.00e+00±0.00e+00	7.81e-01±2.03e-02
gaussiancopula	2.90e-02±0.00e+00	2.94e-02±3.47e-18	0.00e+00±0.00e+00	6.91e-01±6.41e-17
tvae	3.26e-02±1.29e-02	1.45e-01±1.91e-03	0.00e+00±0.00e+00	**6.40e-01±3.35e-03**

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

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**	9.77e-01±6.88e-04
smote-enc	0.00e+00±0.00e+00	0.00e+00±0.00e+00	0.00e+00±0.00e+00	9.67e-01±8.19e-04
ctgan	4.47e-04±2.18e-04	1.88e-04±2.25e-04	0.00e+00±0.00e+00	6.96e-01±1.00e-02
copulagan	1.79e-04±4.03e-05	2.19e-04±3.23e-05	0.00e+00±0.00e+00	7.81e-01±2.03e-02
gaussiancopula	1.99e-04±0.00e+00	1.04e-04±1.36e-20	0.00e+00±0.00e+00	6.91e-01±6.41e-17
tvae	8.02e-04±2.62e-04	8.94e-03±1.12e-03	0.00e+00±0.00e+00	**6.40e-01±3.35e-03**

4.2.7. Ejemplos de registros - Conjunto A

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

Tabla 4.24: Ejemplos para el modelo Tddpm, minimo, Economicos (A-2)

Variable/Distancia	Sintético	DCR1 d(0.00e+00)	DCR2 d(2.36e-07)
_price	4500.000000	4500.000000	4592.000000
bathrooms	2.000000	2.000000	2.000000
county	Recoleta	Macul	Independencia
m_built	60.000000	60.000000	60.000000
m_size	60.000000	60.000000	96.000000
property_type	Departamento	Departamento	Casa
publication_date	1545.000000	1545.000000	1545.000000
rooms	3.000000	3.000000	3.000000
state	Metropolitana de Santiago	Metropolitana de Santiago	Metropolitana de Santiago
transaction_type	Venta	Venta	Venta

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

Tabla 4.25: Ejemplos para el modelo Tddpm, percentil 1, Economicos (A-2)

Variable/Distancia	Sintético	DCR1 d(1.36e-10)	DCR2 d(2.56e-08)
_price	6000.000000	6000.000000	5990.000000
bathrooms	2.000000	2.000000	2.000000
county	Santiago	Ñuñoa	Ñuñoa
m_built	78.000000	78.000000	78.000000
m_size	78.674688	86.000000	80.000000
property_type	Departamento	Departamento	Departamento
publication_date	1693.000000	1693.000000	1693.000000
rooms	3.000000	3.000000	3.000000
state	Metropolitana de Santiago	Metropolitana de Santiago	Metropolitana de Santiago
transaction_type	Venta	Venta	Venta

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

Tabla 4.26: Ejemplos para el modelo Tddpm, percentil 4, Economicos (A-1)

Variable/Distancia	Sintético	DCR1 d(2.79e-09)	DCR2 d(5.75e-09)
_price	15.586326	14.500000	13.345158
bathrooms	1.000000	1.000000	1.000000
county	Las Condes	Santiago	Santiago
m_built	40.000000	40.000000	40.000000
m_size	39.422082	40.000000	41.000000
property_type	Oficina o Casa Oficina	Departamento	Departamento
publication_date	1545.000000	1545.000000	1545.000000
rooms	2.000000	2.000000	2.000000
state	Metropolitana de Santiago	Metropolitana de Santiago	Metropolitana de Santiago
transaction_type	Arriendo	Arriendo	Arriendo

Tabla 4.27: Ejemplos de texto modelo Tddpm, percentil 4, Economicos (A-1)

Distancia	description
Sintético	nan
DCR1 d(2.79e-09)	Departamento amoblado,excelente ubicación y orientación, a pasos del metro Moneda. 40 mts2 + terraza.ling-comedor-cocina incorporada,2 dormitorios,1 baño.Piscina,quincho,gimnasio,gran sala de eventos, lavandería.5 estacionamientos de visita.
DCR2 d(5.75e-09)	Muy bien ubicado, a pasos de metro Irarrázaval. 2 dormitorios, 1 baño, 1 bodega y 1 estacionamiento. Piso 5, vista norte despejada. Conexión a lavadora. Dormitorio principal con walk in closet. Gastos comunes $75.000 aproximadamente. Espacios comunes cuentan con: Piscina, quinchos, estacionamiento de visitas, sala multiuso, lavandería y gym. Se arrienda sin muebles. Los valores de arriendo son totalmente conversables Valor arriendo por departamento $420.000.- Valor arriendo por departamento y bodega $450.000.- Valor arriendo por departamento, bodega y estacionamiento $480.000.- Excelente oportunidad.

4.2.8. Propiedades estadísticas - Conjunto A

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

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

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
nobs	22059	27574	27574	27574
mean	146269	1038479	**206688**	34097
std_err	105454	557374	**108369**	408
upper_ci	352956	2130912	**419086**	34897
lower_ci	-60417	**-53955**	-5711	33297
std	15662334	92554326	**17995050**	67779
iqr	340.500	**330.387**	365.680	33605.232
iqr_normal	252.413	**244.916**	271.079	24911.596
mad	290635	2075076	**411702**	49032
mad_normal	364257	2600722	**515991**	61452
coef_var	107.079	**89.125**	87.064	1.988
range	2.24100e+09	1.11126e+10	**2.00500e+09**	3.94949e+05
max	2.24100e+09	1.11126e+10	**2.00500e+09**	3.94949e+05
min	0.000	0.633	1.000	**0.000**
skew	134.762	**98.466**	95.833	2.207
kurtosis	19053	**10287**	9476	7
jarque_bera	3.33616e+11	**1.21564e+11**	1.03141e+11	4.38779e+04
mode_freq	0.027	**0.029**	0.006	0.687
median	145.000	**147.836**	148.460	0.000
0.1 %	2.000	5.258	14.193	**0.000**
1.0 %	22.000	**22.755**	25.000	0.000
25.0 %	66.000	69.613	**67.847**	0.000
75.0 %	406.500	**400.000**	433.527	33605.232
95.0 %	5000	**5000**	4114	197629
99.0 %	10200	**9126**	8578	285448
99.9 %	70000	44467	**48000**	344981

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

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Las Condes' 'Santiago' 'Providencia' 'Vitacura' 'Lo Barnechea']	['Las Condes' 'Santiago' 'Providencia' 'Vitacura' 'Lo Barnechea']	['Las Condes' 'Santiago' 'Providencia' 'Vitacura' 'Lo Barnechea']	['Las Condes' 'Santiago' 'Viña del Mar' 'Vitacura' 'Providencia']
top5_freq	[3233 2703 1481 1415 1322]	[4149 3211 1910 1871 1740]	[4662 3729 1948 1843 1815]	[3808 3764 2308 1679 1355]
top5_prob	[0.14656149 0.12253502 0.06713813 0.06414615 0.05993019]	[0.15046783 0.11645028 0.06926815 0.06785378 0.06310292]	[0.16907231 0.13523609 0.07064626 0.06683833 0.06582288]	[0.13810111 0.1365054 0.08370204 0.06089069 0.04914049]
nobs	22059	27574	27574	27574
missing	22059	0	0	0

4.2.9. Resumen de resultados - Conjunto A

En esta sección, se proporciona un resumen de los hallazgos más significativos tras el análisis de los resultados obtenidos de los modelos Tddpm y Smote en el Conjunto A.

1. En el conjunto A, Tddpm supera a Smote y al resto de los modelos en el *SDMetrics Score*, aunque Smote tiene una cobertura que es 12 puntos inferiores a Tddpm (Sección 4.2.2).
2. A pesar de las puntuaciones generales, la visualización de correlaciones muestra que Tddpm presenta una mayor similitud en las variables *rooms* y *bathrooms* en comparación con Smote (Sección 4.2.3).
3. En cuanto a cobertura de datos, ambos modelos (Smote y Tddpm) muestran puntuaciones bajas debido a elementos con cobertura inferior al 40 %, como es el caso de la variable *m_size*. Aun así, Tddpm supera ligeramente a Smote en la mayoría de las columnas (Sección 4.2.4).
4. Se sugiere que la baja cobertura en *m_size* podría deberse a su distribución, que presenta una larga cola caracterizada por valores altos pero infrecuentes (Sección 4.2.4).
5. Ambos modelos, Smote y Tddpm, muestran un buen rendimiento en cuanto a la forma y la distribución de los datos. Sin embargo, una buena distribución no asegura una cobertura completa (Sección 4.2.5).
6. El estudio demuestra que cuando el DCR (Distancia de Copia de Registro) es igual a 0, el registro se ha copiado directamente del conjunto de datos real (Sección 4.2.7).
7. El análisis muestra que incluso con el percentil 1, las diferencias pueden ser significativas. En este caso, las variables *m_size* (metros cuadrados) y *_price* (precio) cambian, así como la variable *county* (condado) en el segundo registro más cercano (Sección 4.2.6).
8. Se observa la coherencia de simulación en algunos registros. Por ejemplo, cuando el sistema genera un número de teléfono, este parece coherente. También menciona estar cerca de un metro, detalle que podría estar presente en una publicación real, aunque el metro indicado no exista (Sección 4.2.7).
9. Propiedades estadísticas - Conjunto A: La tabla muestra las propiedades estadísticas de las variables *m_size* y *county* en el Conjunto A, donde los modelos Tddpm y Smote muestran una diferencia mayor al 5 % con respecto al conjunto original de entrenamiento. Tanto en la variable *m_size* como en la variable *county*, el modelo Tddpm muestra grandes diferencias respecto al conjunto original, indicado por las celdas marcadas en rojo. En general, parece que el modelo Tddpm tiene dificultades para replicar de manera precisa las propiedades estadísticas de estas variables. (Sección 4.2.8)

4.2.10. SDMetrics Score - Conjunto B

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

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

Model Name	Column Pair Trends	Column Shapes	Coverage	Boundaries	Synthesis	**Score**
tddpm_mlp	**9.72e-01±1.50e-03**	**9.83e-01±1.09e-03**	**8.12e-01±1.89e-02**	**1.00e+00±0.00e+00**	9.90e-01±8.52e-04	**9.77e-01±6.88e-04**
smote-enc	9.59e-01±1.20e-03	9.76e-01±4.34e-04	6.27e-01±1.31e-02	**1.00e+00±0.00e+00**	9.24e-01±1.97e-03	9.67e-01±8.19e-04
copulagan	7.60e-01±1.58e-02	8.02e-01±2.69e-02	6.80e-01±6.95e-03	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	7.81e-01±2.03e-02
ctgan	7.43e-01±1.27e-02	6.49e-01±1.69e-02	6.76e-01±7.85e-04	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	6.96e-01±1.00e-02
gaussiancopula	6.94e-01±0.00e+00	6.87e-01±6.41e-17	5.52e-01±9.06e-17	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	6.91e-01±6.41e-17
tvae	5.95e-01±4.07e-03	6.85e-01±2.67e-03	9.53e-02±1.48e-03	**1.00e+00±0.00e+00**	**1.00e+00±0.00e+00**	6.40e-01±3.35e-03

4.2.11. Correlación - Conjunto B

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

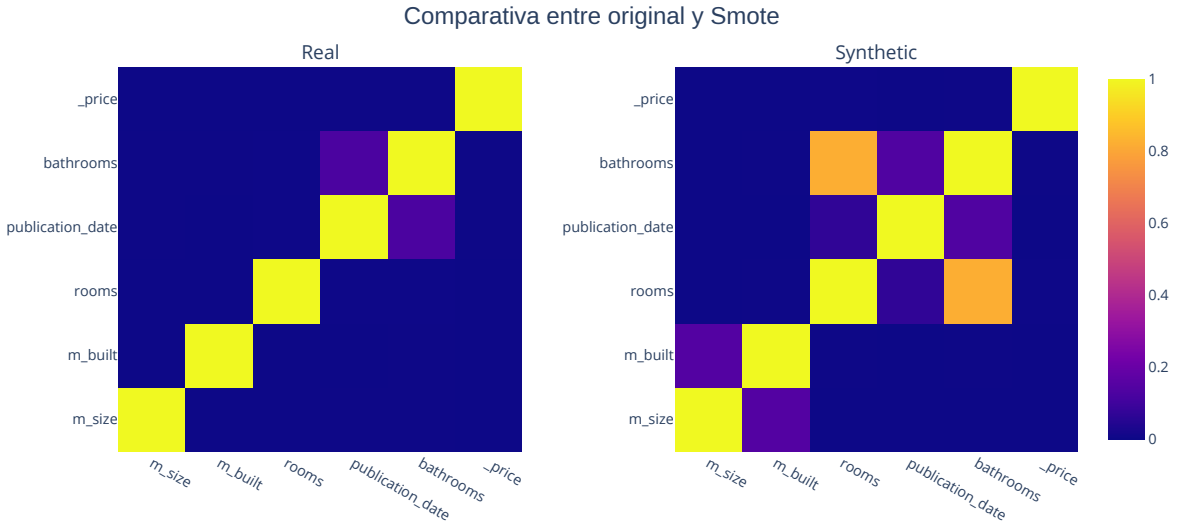

Figura 4.12: Correlación de conjunto original de entrenamiento y Smote, Economicos (B-1)

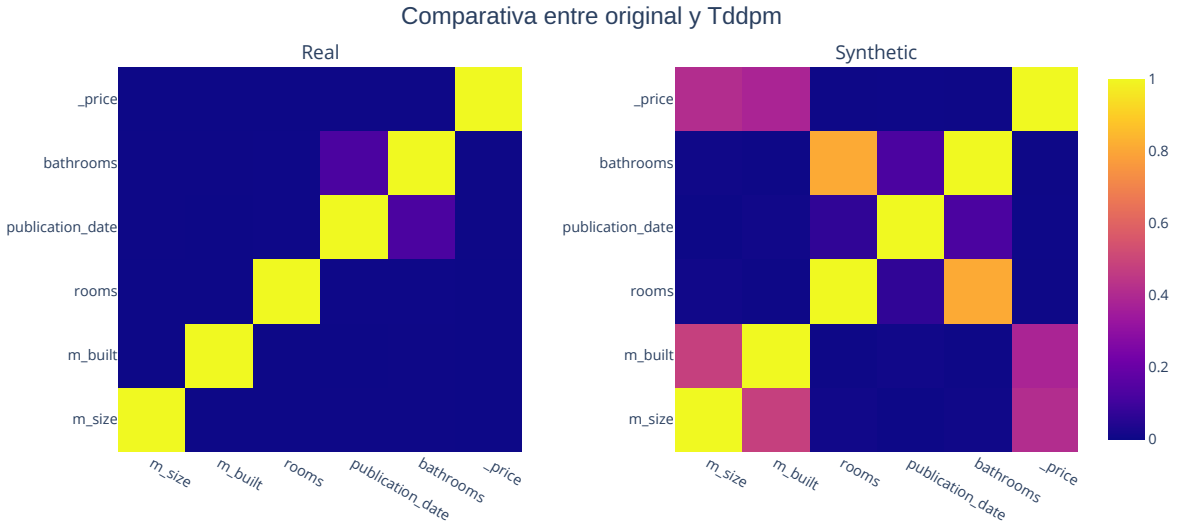

Figura 4.13: Correlación de conjunto original de entrenamiento y Tddpm, Economicos (B-1)

4.2.12. Reporte diagnóstico - Conjunto B

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

Tabla 4.31: Cobertura Categoría/Rango para Modelos Smote y Tddpm, Economicos

Columna	Metrica	smote-enc	tddpm_mlp
_price	RangeCoverage	8.10e-01±1.34e-01	**9.11e-01±1.37e-02**
bathrooms	CategoryCoverage	**8.63e-01±5.00e-02**	6.67e-01±1.39e-02
county	CategoryCoverage	5.90e-01±3.05e-03	**7.99e-01±2.20e-02**
m_built	RangeCoverage	3.18e-01±1.01e-01	**7.54e-01±1.77e-01**
m_size	RangeCoverage	3.45e-02±1.98e-03	**4.00e-01±1.51e-01**
property_type	CategoryCoverage	6.30e-01±5.24e-02	**9.07e-01±5.24e-02**
publication_date	RangeCoverage	9.77e-01±6.18e-03	**9.88e-01±4.44e-03**
rooms	CategoryCoverage	7.56e-01±3.98e-02	**7.97e-01±3.04e-02**
state	CategoryCoverage	7.92e-01±2.95e-02	**9.79e-01±2.95e-02**
transaction_type	CategoryCoverage	5.00e-01±0.00e+00	**9.17e-01±1.18e-01**

4.2.13. Reporte de calidad - Conjunto B

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

Tabla 4.32: Evaluación de Similitud de Distribución para Modelos Smote y Tddpm, Economicos

Columna	Metrica	smote-enc	tddpm_mlp
_price	KSComplement	**9.91e-01±3.85e-04**	9.84e-01±3.53e-03
bathrooms	TVComplement	**9.94e-01±6.66e-04**	9.87e-01±2.15e-03
county	TVComplement	9.22e-01±9.28e-04	**9.66e-01±2.10e-03**
m_built	KSComplement	**9.87e-01±2.14e-03**	**9.87e-01±1.11e-03**
m_size	KSComplement	9.72e-01±7.43e-04	**9.84e-01±3.22e-03**
property_type	TVComplement	9.67e-01±1.33e-03	**9.82e-01±9.49e-04**
publication_date	KSComplement	9.80e-01±1.61e-03	**9.85e-01±1.61e-03**
rooms	TVComplement	9.77e-01±2.28e-03	**9.81e-01±3.18e-03**
state	TVComplement	9.69e-01±4.29e-04	**9.83e-01±1.05e-03**
transaction_type	TVComplement	**9.98e-01±1.07e-03**	9.93e-01±3.54e-03

4.2.14. Privacidad - Conjunto B

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

Tabla 4.33: Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 5, Economicos

Modelo	DCR ST	DCR SH	DCR TH	**Score**
tddpm_mlp	4.29e-09±2.16e-10	3.50e-08±1.92e-09	**1.28e-08±0.00e+00**	**9.77e-01±6.88e-04**
smote-enc	2.90e-11±1.13e-12	4.41e-08±2.36e-09	**1.28e-08±0.00e+00**	9.67e-01±8.19e-04
ctgan	**7.59e-06±5.75e-06**	**1.91e-05±2.01e-05**	**1.28e-08±0.00e+00**	6.96e-01±1.00e-02
copulagan	1.27e-06±3.04e-07	2.73e-06±5.89e-07	**1.28e-08±0.00e+00**	7.81e-01±2.03e-02
gaussiancopula	5.11e-06±0.00e+00	8.25e-06±0.00e+00	**1.28e-08±0.00e+00**	6.91e-01±6.41e-17
tvae	2.19e-07±1.60e-09	4.15e-07±5.43e-09	**1.28e-08±0.00e+00**	6.40e-01±3.35e-03

Tabla 4.34: Distancia de registros más cercanos entre conjuntos Sinteticos, percentil 1, Economicos

Modelo	DCR ST	DCR SH	DCR TH	**Score**
tddpm_mlp	1.44e-10±6.01e-12	1.40e-09±1.05e-10	**0.00e+00±0.00e+00**	**9.77e-01±6.88e-04**
smote-enc	0.00e+00±0.00e+00	1.41e-09±4.21e-10	**0.00e+00±0.00e+00**	9.67e-01±8.19e-04
ctgan	**2.20e-06±1.50e-06**	**3.24e-06±1.58e-06**	**0.00e+00±0.00e+00**	6.96e-01±1.00e-02
copulagan	2.04e-07±2.85e-08	4.37e-07±6.38e-08	**0.00e+00±0.00e+00**	7.81e-01±2.03e-02
gaussiancopula	8.04e-07±8.64e-23	1.93e-06±0.00e+00	**0.00e+00±0.00e+00**	6.91e-01±6.41e-17
tvae	7.41e-08±1.95e-09	1.16e-07±3.43e-09	**0.00e+00±0.00e+00**	6.40e-01±3.35e-03

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

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

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

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

Modelo	NNDR ST	NNDR SH	NNDR TH	**Score**
tddpm_mlp	6.79e-02±7.37e-04	**1.00e-01±2.26e-03**	**1.31e-02±0.00e+00**	9.77e-01±6.88e-04
smote-enc	**7.15e-04±7.49e-06**	1.14e-01±4.79e-03	1.31e-02±0.00e+00	9.67e-01±8.19e-04
ctgan	2.57e-01±8.81e-03	3.27e-01±4.72e-02	1.31e-02±0.00e+00	6.96e-01±1.00e-02
copulagan	2.01e-01±1.27e-02	2.23e-01±5.47e-02	1.31e-02±0.00e+00	7.81e-01±2.03e-02
gaussiancopula	3.07e-01±0.00e+00	2.76e-01±0.00e+00	1.31e-02±0.00e+00	6.91e-01±6.41e-17
tvae	3.02e-01±6.15e-03	3.49e-01±1.26e-03	1.31e-02±0.00e+00	**6.40e-01±3.35e-03**

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

Modelo	NNDR ST	NNDR SH	NNDR TH	**Score**
tddpm_mlp	2.52e-03±1.71e-04	9.61e-03±1.17e-04	**0.00e+00±0.00e+00**	9.77e-01±6.88e-04
smote-enc	**0.00e+00±0.00e+00**	**3.00e-03±1.28e-03**	0.00e+00±0.00e+00	9.67e-01±8.19e-04
ctgan	1.35e-02±2.69e-03	2.77e-02±1.69e-02	0.00e+00±0.00e+00	6.96e-01±1.00e-02
copulagan	1.03e-02±1.29e-03	1.11e-02±2.64e-03	0.00e+00±0.00e+00	7.81e-01±2.03e-02
gaussiancopula	2.90e-02±0.00e+00	2.94e-02±3.47e-18	0.00e+00±0.00e+00	6.91e-01±6.41e-17
tvae	3.26e-02±1.29e-02	1.45e-01±1.91e-03	0.00e+00±0.00e+00	**6.40e-01±3.35e-03**

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

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**	9.77e-01±6.88e-04
smote-enc	0.00e+00±0.00e+00	0.00e+00±0.00e+00	0.00e+00±0.00e+00	9.67e-01±8.19e-04
ctgan	4.47e-04±2.18e-04	1.88e-04±2.25e-04	0.00e+00±0.00e+00	6.96e-01±1.00e-02
copulagan	1.79e-04±4.03e-05	2.19e-04±3.23e-05	0.00e+00±0.00e+00	7.81e-01±2.03e-02
gaussiancopula	1.99e-04±0.00e+00	1.04e-04±1.36e-20	0.00e+00±0.00e+00	6.91e-01±6.41e-17
tvae	8.02e-04±2.62e-04	8.94e-03±1.12e-03	0.00e+00±0.00e+00	**6.40e-01±3.35e-03**

4.2.15. Ejemplos de registros - Conjunto B

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

Tabla 4.39: Ejemplos para el modelo Tddpm, percentil 2, Economicos (B-1)

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

Tabla 4.40: Ejemplos de texto modelo Tddpm, percentil 2, Economicos (B-1)

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

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

Tabla 4.41: Ejemplos para el modelo Tddpm, percentil 4, Economicos (B-1)

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

Tabla 4.42: Ejemplos de texto modelo Tddpm, percentil 4, Economicos (B-1)

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

4.2.16. Propiedades estadísticas - Conjunto B

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

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

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

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

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

4.2.17. Resumen de los Resultados - Conjunto B

En esta sección, se proporciona un resumen de los hallazgos más significativos tras el análisis de los resultados obtenidos de los modelos Tddpm y Smote en el Conjunto B.

1. El modelo Tddpm obtuvo mejores Score, Coverage, Column Shape y Column Pair Trends en comparación con el conjunto A. Esto podría deberse a una mayor cantidad de datos, permitiendo más tiempo para aprender la distribución. Los otros modelos no mostraron una mejora notable, indicando una mayor capacidad de Tddpm. La cobertura más baja fue del 47 % en el peor de los casos (Sección 4.2.10 y 4.2.12).

2. Los modelos Smote y Tddpm presentaron diferencias marcadas al ser comparados con el conjunto original, creando correlaciones no presentes en los datos originales. Para Smote, las correlaciones fueron entre las variables *bathrooms-rooms* y *m_size-m_built*. Tddpm generó adicionalmente correlaciones entre *_price-m_size* y *_price-m_built* (Sección 4.2.11).

3. La cobertura fue notablemente baja en las variables *rooms* y *m_size* en Smote, y en *bathrooms* y *rooms* en Tddpm. En general, Tddpm superó ligeramente a Smote. (Sección 4.2.12)

4. Ambos modelos, Smote y Tddpm, presentaron buenas métricas, superando el 91 % en términos de distribución y forma. Sin embargo, hubo excepciones en Smote, los casos de *m_built* (85 %) y *m_size* (55 %) (Sección 4.2.13).

5. Las distancias mínimas para los percentiles 5 y 1 fueron considerablemente menores en el Conjunto B que en el Conjunto A. Se puede afirmar que el 95 % de los registros tiene al menos una distancia de $9{,}12 \times 10^{-15}$ (Sección 4.2.14).

6. En el conjunto B, el modelo Tddpm generó departamentos de dos dormitorios con características y diferencias notables entre los datos sintéticos y a pesar de estar en el percentil 2 más cercano (Sección 4.2.15).

7. En el modelo Tddpm la variable *bathrooms* posee un máximo, skew y Jarque-Bera distintos a los reales. Estas son las propiedades que consistentemente el modelo le ha ido peor. (Sección 4.2.16).

Capítulo 5

Conclusiones y discusión

Este capítulo resume y discute los hallazgos clave derivados de nuestro estudio sobre la generación de conjuntos de datos sintéticos estructurados, incluyendo textos. Se centra en la evaluación de diferentes modelos generativos y el análisis de métricas relevantes para la calidad de los datos. Además, se reflexiona sobre las implicaciones de la utilidad frente a la privacidad en los datos sintéticos, y se destaca el valor del modelo Tddpm para la preservación de la privacidad. Finalmente, se identifican limitaciones y se proponen caminos para investigaciones futuras, especialmente en lo que respecta a la creciente importancia de la generación de texto y la evaluación de la privacidad en dicho contexto.

5.1. Conclusiones

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

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

En lo referente a la privacidad, se exploró la relación existente entre utilidad representada por *SDMetric Score* y privacidad representada por la distancia al registro más cercano (**DCR**). Se observó que a medida que el conjunto sintético se asemejaba más al original, mayor SDMetric Score, las métricas de privacidad disminuían, como se reflejaba en la disminución DCR y la relación entre el registro más cercano y el segundo más cercano (**NNDR**). 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.

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

5.2. Limitaciones

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

Y más allá de las métricas que por tiempo no pudieron ser analizadas, me gustaría contextualizar en el tiempo que se desarrolló este trabajo y es que al iniciar esta tesis, a mediados de 2022, era evidente la creciente complejidad de la inteligencia artificial y la mejora continua de sus resultados. Sin embargo, era difícil anticipar que, durante el desarrollo de este trabajo, este crecimiento se aceleraría de tal manera que, al acercarse a la conclusión del mismo, los modelos utilizados aquí, como el **T5**, constituirían solo uno de los muchos modelos disponibles para la generación de texto. Desafortunadamente, debido a limitaciones de tiempo, no se podrán tratar otros modelos en profundidad y estos serán objeto de la sección de discusión. Para aquellos interesados en continuar con esta línea de investigación, se recomienda visitar el HuggingFace Leaderboard LLM [36], donde encontrarán una amplia gama de modelos a explorar y explotar.

Ejemplo de uso de ChatGPT en el marzo 2023 [37].

Prompt

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

ChatGPT-4

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

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

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

5.3. Discusión

Los modelos de generación de texto están en pleno auge. Recientemente han emergido modelos como **GPT-4** [38], **Llama** [39], **Palm2** [40] y **Falcon** [41], entre muchos otros que se pueden ver en el HuggingFace Leaderboard LLM [36]. El modelo **Chinchilla** [42] 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 con las métricas, tal como se mencionó en la conclusión, algunas de ellas no se calcularon en este trabajo debido a restricciones de tiempo. Además, el estudio de la privacidad en la generación de texto es un área que aún no ha sido ampliamente explorada. Determinar qué métricas son relevantes en este aspecto podría ser tan importante como la evaluación de la eficacia de los nuevos modelos.

5.4. Evaluación de objetivos y logros

En esta sección, confrontaremos los objetivos propuestos y los resultados alcanzados en nuestro estudio. Nuestro reto inicial radicaba en concebir un mecanismo para la creación de conjuntos de datos sintéticos con contenido textual y su comparación con los datos originales. A continuación, analizaremos la correspondencia entre nuestros descubrimientos y dichos objetivos, permitiéndonos estimar el éxito de nuestra investigación, identificar potenciales áreas de mejora y detectar oportunidades para futuros estudios.

Objetivo especifico 1: Modelos generativos

Utilización de modelos generativos capaces de producir **nuevos conjuntos de datos sintéticos a partir de datos originales** que **contienen texto**.

Este objetivo engloba conjuntos de datos numéricos y categóricos como fundamentales, añadiendo el texto como elemento adicional. Por ende, todos los aspectos deben ser evaluados en relación a estos tres conjuntos de datos.

Pregunta: ¿Fueron creados nuevos conjuntos de datos sintéticos?.
Respuesta: Sí, cada técnica y modelo empleado generó un conjunto de datos sintéticos. ✓

Pregunta: ¿Fueron creados nuevos conjuntos de datos sintéticos que incluyen **texto**?.
Respuesta: Sí, se obtuvieron descripciones de propiedades. Este procedimiento puede ser replicado para cualquier otro campo de texto. ✓

Pregunta: ¿Se empleó algún modelo generativo?.
Respuesta: Sí, se utilizaron Tddpm y ctgan, entre otros, para campos numéricos y categóricos. ✓
Nota: Aunque se incluye SMOTE para comparación, SMOTE no es un modelo generativo. ⚠

Pregunta: ¿Se empleó algún modelo generativo para **texto**?.
Respuesta: Sí, se utilizó mT5, un modelo multilenguaje derivado de la familia T5 de Google, que es un modelo generativo. ✓
Nota: Aunque mT5 es parte de la familia T5, no es el último modelo lanzado. A finales de 2022, se liberó Flan-T5, sin embargo, no se recomienda su uso, ya que su *embedding* no maneja bien caracteres del español. Es aconsejable considerar la evaluación de modelos más nuevos, como Falcon o Llama. ⚠

Pregunta: ¿Se emplearon los conjuntos de datos originales?.
Respuesta: Sí, todos los modelos o técnicas utilizadas se basaron en un conjunto de datos original para imitar las distribuciones o llevar a cabo el entrenamiento de los modelos. ✓

Pregunta: ¿Se emplearon los conjuntos de datos originales para la generación de **texto**?.
Respuesta: Sí, la descripción en los conjuntos originales y otras columnas se emplearon en el entrenamiento, luego, los datos sintéticos se usaron para generar textos sintéticos. ✓

Podemos concluir que el primer objetivo específico se cumplió de manera satisfactoria. ✓ ✓

Objetivo especifico 2: Evaluación y comparación

Evaluar y **comparar** las características de los **conjuntos de datos sintéticos y originales** en tres aspectos: **propiedades estadísticas, nivel de privacidad, y sus distribuciones**.

Procedemos a plantear las preguntas para evaluar el cumplimiento del segundo objetivo específico.

Pregunta: ¿Se evaluaron las propiedades estadísticas de los conjuntos originales y sintéticos?.
Respuesta: De manera parcial, se analizaron numerosas propiedades y se efectuó una comparación. ⚠
Nota: Aunque se presentaron propiedades estadísticas en el marco, algunas no fueron finalmente comparadas, como por ejemplo, la Desviación Mediana Absoluta Normalizada. A su vez, se incluyeron técnicas que no se consideraron en el objetivo original. ⚠

Pregunta: ¿Se evaluaron las propiedades estadísticas de los conjuntos de **texto** originales y sintéticos?.
Respuesta: No, se pudieron considerar elementos como frecuencia, moda, que en textos como los títulos de publicaciones o distribuciones pueden no tener sentido. Sin embargo, podrían ser útiles algunas técnicas como TF-IDF o el análisis de N-GRAMAS. ✗
Nota: Técnicas como TF-IDF y cross-validación, entre otras, no se incluyeron debido a la priorización y limitaciones de tiempo. Se recomienda complementar este trabajo en estos aspectos en futuros estudios. ⚠

Pregunta: ¿Se evaluó la privacidad de los conjuntos originales y sintéticos?.
Respuesta: Sí, se utilizó la métrica DCR y NNDR como principales y se interpretó cada valor en percentiles específicos. ✓
Nota: Se sugiere que futuros trabajos profundicen en la evaluación de privacidad, facilitando la decisión del delta mínimo esperado para el percentil 5. Aunque el presente trabajo ilustra con ejemplos el significado de la distancia medida, una forma de cálculo de la distancia mínima deseada representaría una mejora sustancial. ⚠

Pregunta: ¿Se evaluó la privacidad de los conjuntos de **textos** originales y sintéticos?.
Respuesta: No, tampoco se encontró literatura relevante al respecto.
Nota: La privacidad en textos sintéticos podría ser un campo de investigación relevante. ⚠

Pregunta: ¿Se evaluó la distribución de los conjuntos originales y sintéticos?.
Respuesta: Sí, **Column Shape** es una de las métricas más adecuadas. ✓

Pregunta: ¿Se evaluó la distribución de los conjuntos de **textos** originales y sintéticos?.
Respuesta: No, se podría haber empleado TF-IDF o el análisis de frecuencia de palabras. ✗

Podemos concluir que el segundo objetivo específico se alcanzó de manera parcial, identificándose múltiples aspectos susceptibles de ser investigados y mejorados en el futuro. ✓ ⚠

Objetivo general

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

Después de evaluar cada uno de los objetivos de manera independiente, se puede concluir que el objetivo inicial se logró parcialmente: se tuvo un buen desempeño en la generación, pero quedaron aspectos pendientes en la evaluación.

Este trabajo brinda certidumbre sobre la viabilidad técnica del enfoque empleado, pero deja abiertas las siguientes cuestiones para futuras investigaciones:

1. ¿Cómo se puede evaluar la privacidad en textos?
2. ¿Cómo se puede asegurar la variabilidad en los textos generados para evitar repeticiones frecuentes?
3. ¿Es posible reducir la lista de propiedades y métricas calculadas en este estudio, manteniendo al mismo tiempo la capacidad de evaluación?
4. ¿Cómo se pueden incorporar otros tipos de datos, como arreglos, estructuras anidadas o series de tiempo?
5. ¿Un modelo más poderoso podría generar una mayor variabilidad?

Bibliografía

[1] A. Krizhevsky, I. Sutskever y G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," en *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012. dirección: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (visitado 19-02-2023).

[2] "Papers with code - ImageNet benchmark (image classification)." (), dirección: https://paperswithcode.com/sota/image-classification-on-imagenet (visitado 19-02-2023).

[3] "DALL·e 2," OpenAI. (), dirección: https://openai.com/dall-e-2/ (visitado 19-02-2023).

[4] "Imagen: Text-to-Image Diffusion Models." (), dirección: https://imagen.research.google/ (visitado 19-02-2023).

[5] "Stable diffusion public release," Stability AI. (), dirección: https://stability.ai/blog/stable-diffusion-public-release (visitado 19-02-2023).

[6] D. Milmo y D. M. G. t. editor, "Google v microsoft: Who will win the AI chatbot race?" *The Guardian*, 10 de feb. de 2023, ISSN: 0261-3077. dirección: https://www.theguardian.com/technology/2023/feb/10/google-v-microsoft-who-will-win-the-ai-chatbot-race-bard-chatgpt (visitado 19-02-2023).

[7] "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." (15 de feb. de 2023), dirección: https://finance.yahoo.com/news/microsoft-google-game-thrones-battle-174112314.html (visitado 19-02-2023).

[8] K. El Emam, L. Mosquera y R. Hoptroff, *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media, 2020.

[9] J. Gantz y D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," *IDC iView: IDC Analyze the future*, vol. 2007, n.º 2012, págs. 1-16, 2012. dirección: https://datastorageasean.com/sites/default/files/idc-the-digital-universe-in-2020.pdf.

[10] K. Adnan y R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *Journal of Big Data*, vol. 6, págs. 1-38, 2019, Publisher: Springer. dirección: https://link.springer.com/article/10.1186/s40537-019-0254-8.

[11] P. Bruce, A. Bruce y P. Gedeck, *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media, 2020.

[12] S. De Capitani Di Vimercati, S. Foresti, G. Livraga y P. Samarati, "Data privacy: Definitions and techniques," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 20, n.º 6, págs. 793-817, 2012, Publisher: World Scientific.

[13] C. Dwork, "Differential privacy," en *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, Springer, 2006, págs. 1-12.

[14] S. De Capitani Di Vimercati y P. Samarati, "K-anonymity," en *Encyclopedia of Cryptography, Security and Privacy*, S. Jajodia, P. Samarati y M. Yung, eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2021, págs. 1-5, ISBN: 978-3-642-27739-9. DOI: 10.1007/978-3-642-27739-9_754-2. dirección: https://link.springer.com/10.1007/978-3-642-27739-9_754-2 (visitado 17-11-2023).

[15] Z. Zhao, A. Kunar, R. Birke y L. Y. Chen, "CTAB-GAN: Effective Table Data Synthesizing," en *Proceedings of The 13th Asian Conference on Machine Learning*, V. N. Balasubramanian e I. Tsang, eds., ép. Proceedings of Machine Learning Research, vol. 157, PMLR, 17 de nov. de 2021, págs. 97-112. dirección: https://proceedings.mlr.press/v157/zhao21a.html.

[16] P. Regulation, "Regulation (EU) 2016/679 of the European Parliament and of the Council," *Regulation (eu)*, vol. 679, pág. 2016, 2016. dirección: https://dvbi.ru/Portals/0/DOCUMENTS_SHARE/RISK_MANAGEMENT/EBA/GDPR_eng_rus.pdf.

[17] S. L. Pardau, "The California consumer privacy act: Towards a European-style privacy regime in the United States," *J. Tech. L. & Pol'y*, vol. 23, pág. 68, 2018, Publisher: HeinOnline.

[18] A. Act, "Health insurance portability and accountability act of 1996," *Public law*, vol. 104, pág. 191, 1996. dirección: http://www.eolusinc.com/pdf/hipaa.pdf.

[19] A. V. Solatorio y O. Dupriez, "REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers," *arXiv preprint arXiv:2302.02041*, 2023.

[20] D. Pujol, A. Gilad y A. Machanavajjhala, "PreFair: Privately Generating Justifiably Fair Synthetic Data," *arXiv preprint arXiv:2212.10310*, 2022.

[21] A. Acharya, S. Sikdar, S. Das y H. Rangwala, "GenSyn: A Multi-stage Framework for Generating Synthetic Microdata using Macro Data Sources," *arXiv preprint arXiv:2212.05975*, 2022.

[22] A. Kotelnikov, D. Baranchuk, I. Rubachev y A. Babenko, "TabDDPM: Modelling Tabular Data with Diffusion Models," *arXiv preprint arXiv:2209.15421*, 2022.

[23] V. Borisov, K. Seßler, T. Leemann, M. Pawelczyk y G. Kasneci, "Language models are realistic tabular data generators," *arXiv preprint arXiv:2210.06280*, 2022. dirección: https://arxiv.org/pdf/2210.06280.pdf.

[24] Z. Zhao, A. Kunar, R. Birke y L. Y. Chen, "CTAB-GAN+: Enhancing Tabular Data Synthesis," *arXiv preprint arXiv:2204.00401*, 2022. dirección: https://arxiv.org/abs/2204.00401.

[25] L. Xu, M. Skoularidou, A. Cuesta-Infante y K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall y W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, págs. 321-357, 2002.

[27] E. Andrejczuk, J. M. Eisenschlos, F. Piccinno, S. Krichene e Y. Altun, "Table-To-Text generation and pre-training with TabT5," *arXiv preprint arXiv:2210.09162*, 2022.

[28] M. Kale y A. Rastogi, "Text-to-text pre-training for data-to-text tasks," *arXiv preprint arXiv:2005.10433*, 2020.

[29] J. Herzig, P. K. Nowak, T. Müller, F. Piccinno y J. M. Eisenschlos, "TaPas: Weakly supervised table parsing via pre-training," *arXiv preprint arXiv:2004.02349*, 2020.

[30] "Synthetic data metrics." (), dirección: https://docs.sdv.dev/sdmetrics/ (visitado 17-06-2023).

[31] *Virtual Data Lab (VDL)*, original-date: 2020-07-08T16:46:38Z, 14 de nov. de 2022. dirección: https://github.com/mostly-ai/virtualdatalab (visitado 02-07-2023).

[32] H. Kaggle. "House Sales in King County, USA." (2015), dirección: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction.

[33] A. Kotelnikov, D. Baranchuk, I. Rubachev y A. Babenko. "Overview — SDV 0.18.0 documentation." (), dirección: https://sdv.dev/SDV/ (visitado 26-02-2023).

[34] N. Patki, R. Wedge y K. Veeramachaneni, "The synthetic data vault," en *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2016, págs. 399-410.

[35] Akim, *TabDDPM: Modelling Tabular Data with Diffusion Models*, original-date: 2022-10-02T23:01:07Z, 1 de mar. de 2023. dirección: https://github.com/rotot0/tab-ddpm (visitado 01-03-2023).

[36] "Open LLM Leaderboard - a Hugging Face Space by HuggingFaceH4." (), dirección: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard (visitado 24-06-2023).

[37] OpenAI, *ChatGPT: a large language model trained by OpenAI*, 2023. dirección: https://openai.com/blog/chatgpt-a-large-scale-generative-language-model/.

[38] OpenAI, *GPT-4 Technical Report*, 27 de mar. de 2023. DOI: 10.48550/arXiv.2303.08774. arXiv: 2303.08774[cs]. dirección: http://arxiv.org/abs/2303.08774 (visitado 21-06-2023).

[39] "LLaMA: Open and Efficient Foundation Language Models - Meta Research," Meta Research. (), dirección: https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/ (visitado 21-06-2023).

[40] R. Anil, A. M. Dai, O. Firat y col., *PaLM 2 Technical Report*, 17 de mayo de 2023. DOI: 10.48550/arXiv.2305.10403. arXiv: 2305.10403[cs]. dirección: http://arxiv.org/abs/2305.10403 (visitado 21-06-2023).

[41] "Falcon LLM." (), dirección: https://falconllm.tii.ae/ (visitado 21-06-2023).

[42] J. Hoffmann, S. Borgeaud, A. Mensch y col., *Training Compute-Optimal Large Language Models*, 29 de mar. de 2022. DOI: 10.48550/arXiv.2203.15556. arXiv: 2203.15556[cs]. dirección: http://arxiv.org/abs/2203.15556 (visitado 21-06-2023).

Apéndice A

Anexos

Este capítulo de Anexos proporciona información adicional y detallada que respalda la investigación realizada en esta tesis. Aunque estos detalles son esenciales para el completo entendimiento de la investigación, se han incluido en los anexos para mantener la fluidez del cuerpo principal de la tesis.

En las siguientes secciones, se presentan diversos elementos suplementarios. El código de entrenamiento de modelos económicos se proporciona para dar visibilidad a los métodos de aprendizaje automático utilizados. Se incluyen gráficos detallados de correlaciones y estadísticas para los conjuntos de datos utilizados, aportando un análisis más profundo de las características y estructuras de estos conjuntos de datos. También se proporcionan ejemplos de registros generados, ofreciendo una visión tangible de los resultados de la generación de datos.

Por favor, refiérase a estos anexos para una comprensión más completa y detallada de la investigación y los métodos utilizados en este trabajo.

A.1. Código de entrenamiento de económicos

```python
import pandas as pd
from syntheticml.data.synthetic import Synthetic, MODELS
from syntheticml.models.tab_ddpm.sdv import SDV_MLP
import torch
import numpy as np
import itertools
import multiprocessing as mp
import os

def test_train(args):
    lrc, ntc, sts, btsc, rtdlc, syn, df = args
    #notebooks/economicos_good/2e-06_10_100000_5000_1024-512-256
    checkpoint = "economicos_good2/" +  "_".join(
            map(str, [lrc, ntc, sts, btsc, "-".join(map(str, rtdlc))]))
    checkpoint = "con_fechas"
    if os.path.exists(f"{checkpoint}/final_model.pt") or os.path.exists(f"{checkpoint}/exit"):
        return (checkpoint, 1)
    model = SDV_MLP(syn.metadata,
                    "_price",
                    exclude_columns=syn.exclude_columns,
                    df=df,
                    batch_size=btsc,
                    steps=sts,
                    checkpoint=checkpoint,
                    num_timesteps=ntc,
                    weight_decay=0.0,
                    lr=lrc,
                    model_params=dict(rtdl_params=dict(
                        dropout=0.0,
                        d_layers=rtdlc
                    ))
                    )
    model.fit(syn.train)
    model.save(f"{checkpoint}/final_model.pt")
    return (checkpoint, 1)

if __name__ == '__main__':
    df = pd.read_parquet('../datasets/economicos/synth/split/train.parquet')
    category_columns=("property_type", "transaction_type", "state", "county", "rooms", "bathrooms", "m_built", "m_size", "source", )
    # TODO: Estudiar implicancia de valores nulos en categorias y numeros
    df_converted = df.astype({k: 'str' for k in ("description", "price", "title", "address", "owner",)})
    basedate = pd.Timestamp('2017-12-01')
    dtime = df_converted.pop("publication_date")
    df_converted["publication_date"] = dtime.apply(lambda x: (x - basedate).days)
    syn = Synthetic(df_converted,
            id="url",
            category_columns=category_columns,
            text_columns=("description", "price", "title", "address", "owner",),
            exclude_columns=tuple(),
            synthetic_folder = "../datasets/economicos/synth",
            models=['copulagan', 'tvae', 'gaussiancopula', 'ctgan', 'smote-enc'],
            n_sample = df.shape[0],
            target_column="_price"
    )

    lrs = np.linspace(2e-6, 2e-3, 10)
    num_timesteps = np.linspace(10, 1000, 3, dtype=int)
    batch_size = np.linspace(2500, 5000, 3, dtype=int)
    steps = np.linspace(150000, 500000, 5, dtype=int)
    rtdl_params = [
        [1024, 512, 256], [512, 256], [256, 128], [256, 128, 128], [256, 128, 128, 128]
    ]
    try:
        torch.multiprocessing.set_start_method('spawn')
    except:
        pass
    with mp.Pool(1) as p:
        fitted_models = dict(list(p.map(test_train, itertools.product(lrs, num_timesteps, steps, batch_size, rtdl_params, [syn],
        ↪  [df_converted]))))
```

Código 3: Código de ejemplo en Python para sumar dos números. Fuente: Autor.

A.2. Archivo Devcontainer

```
{
  "name": "SyntheticData",
  "image": "nvidia/cuda:12.1.0-devel-ubuntu22.04",
  "extensions": [
    "jebbs.plantuml",
    "ms-toolsai.jupyter-keymap",
    "MS-CEINTL.vscode-language-pack-es",
    "SimonSiefke.svg-preview",
    "adamvoss.vscode-languagetool",
    "mathematic.vscode-latex",
    "maltehei.latex-citations",
    "James-Yu.latex-workshop",
    "valentjn.vscode-ltex",
    "yzhang.markdown-all-in-one",
    "ms-python.python",
    "ms-azuretools.vscode-docker",
    "ms-toolsai.jupyter"
  ],
  "postCreateCommand": "bash ./.devcontainer/postscript.sh",
  "runArgs": ["--gpus", "all"],
  "settings": {
    "terminal.integrated.shell.linux": "/bin/bash",
    "latex-workshop.latex.recipes": [
      {
        "name": "latexmk",
        "tools": [
          "latexmk"
        ]
      }
    ],
    "latex-workshop.latex.tools": [
      {
        "name": "latexmk",
        "command": "latexmk",
        "args": [
          "-pdf",
          "-interaction=nonstopmode",
          "-synctex=1",
          "-shell-escape",
          "%DOC%"
        ]
      },
      {
```

A.3. Lista completa de figura pairwise kingcounty

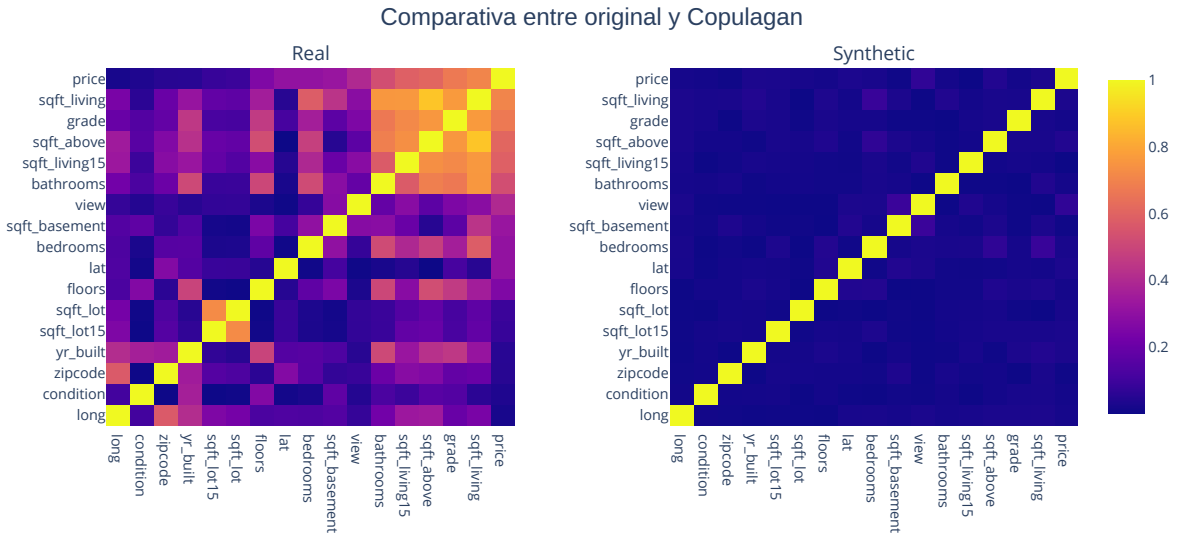

Figura A.1: Correlación de conjunto original de entrenamiento y Copulagan, King county (A-3)

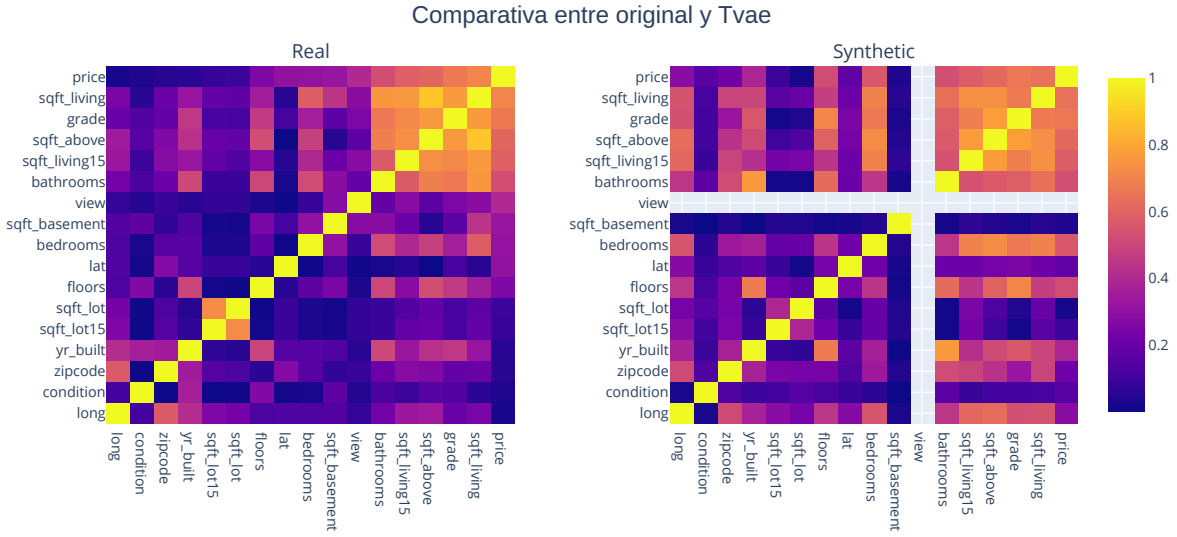

Figura A.2: Correlación de conjunto original de entrenamiento y Tvae, King county (A-3)

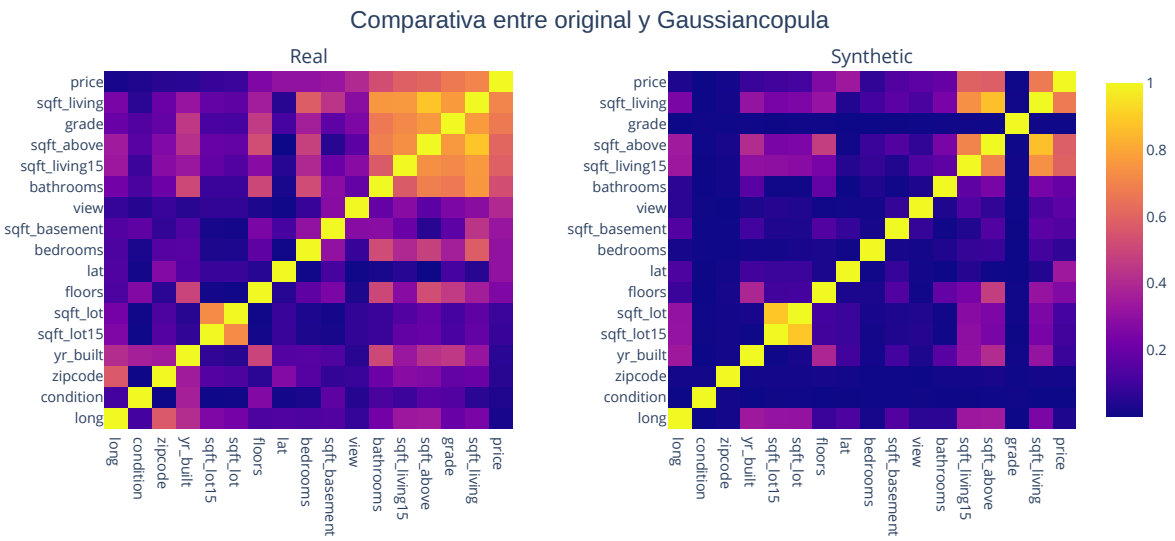

Figura A.3: Correlación de conjunto original de entrenamiento y Gaussiancopula, King county (A-3)

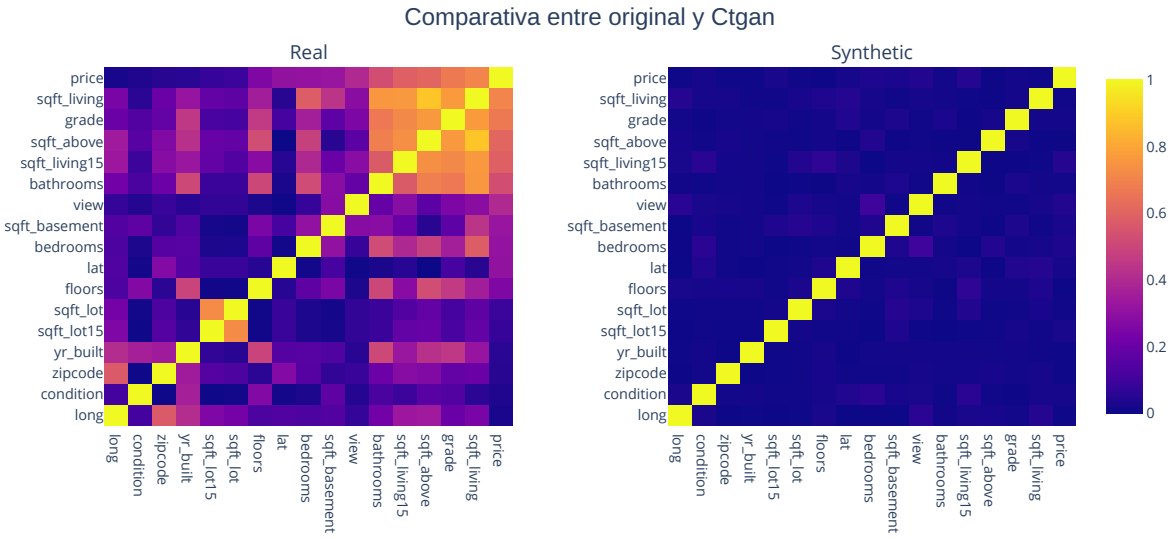

Figura A.4: Correlación de conjunto original de entrenamiento y Ctgan, King county (A-3)

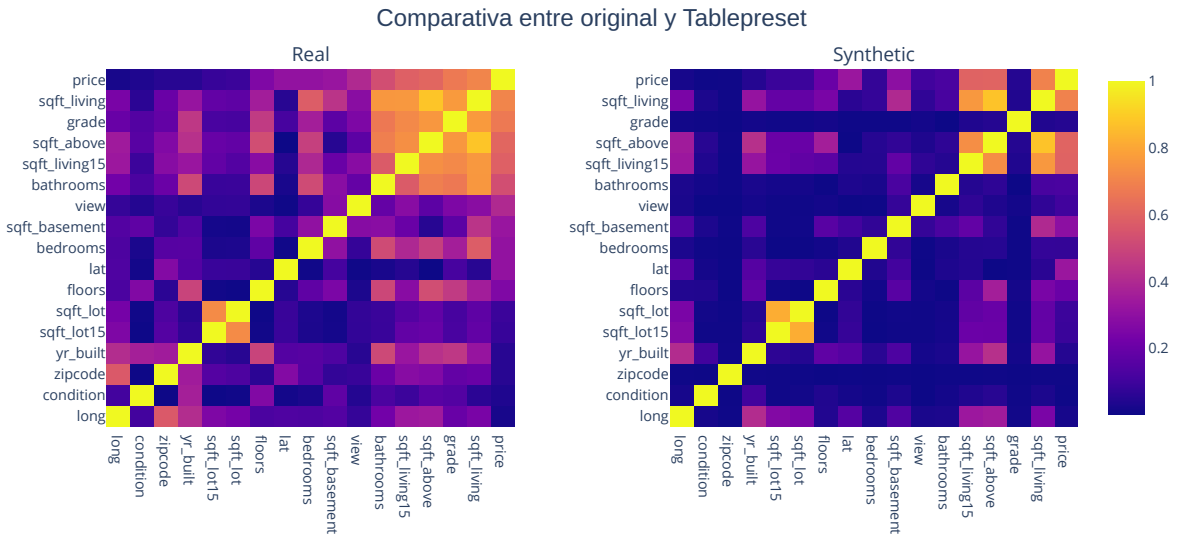

Figura A.5: Correlación de conjunto original de entrenamiento y Tablepreset, King county (A-3)

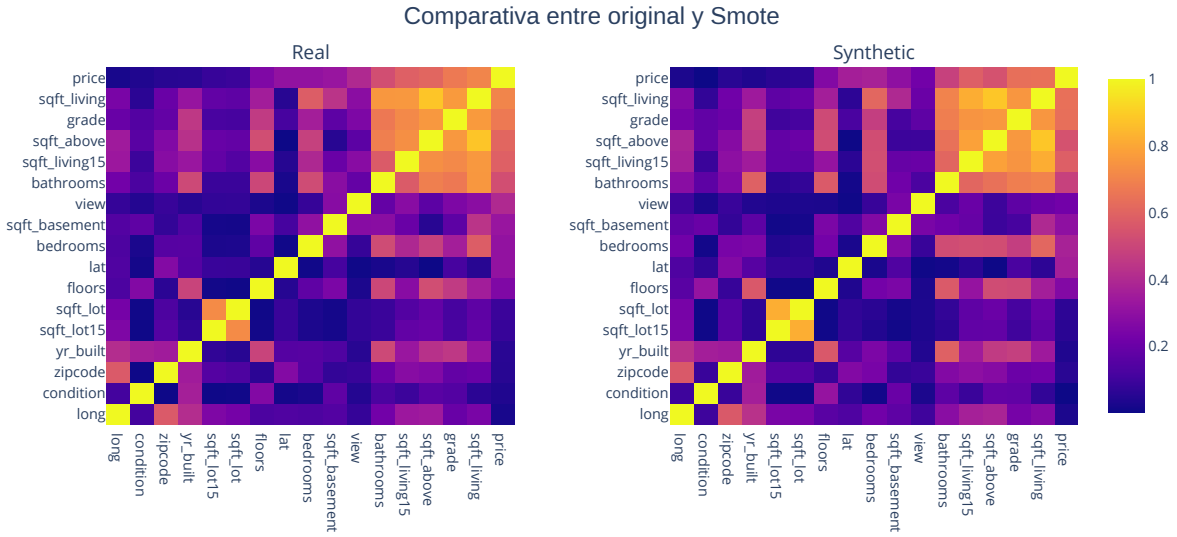

Figura A.6: Correlación de conjunto original de entrenamiento y Smote, King county (A-3)

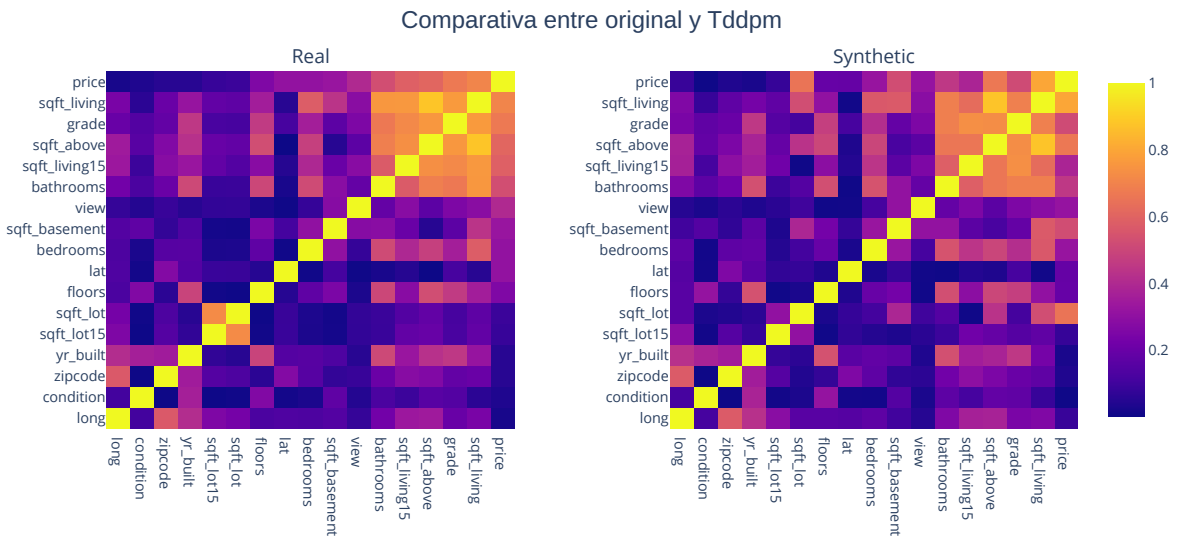

Figura A.7: Correlación de conjunto original de entrenamiento y Tddpm, King county (A-3)

A.4. Smote y Tddpm en KingCounty Gráficas por Columnas

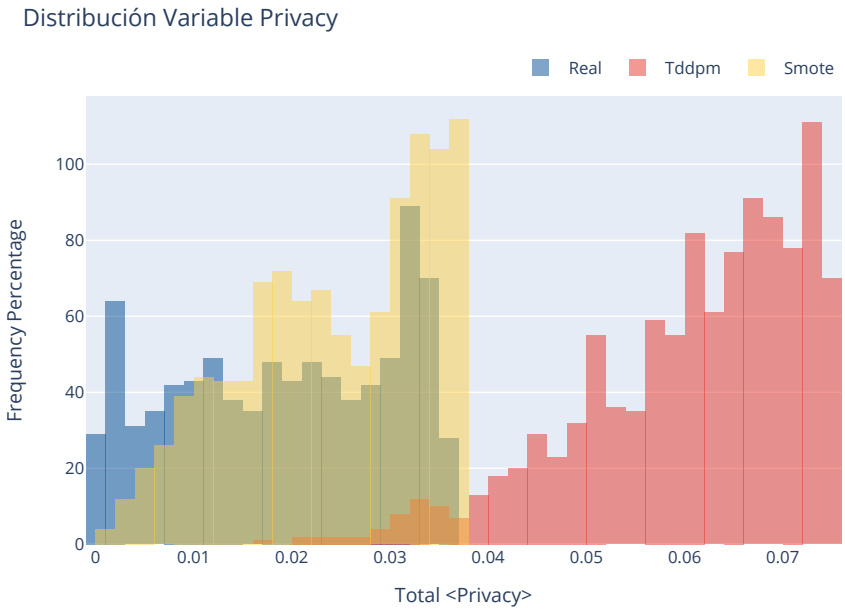

Figura A.8: Frecuencia del campo Privacy en el modelo real y Top 2, King county (A-1)

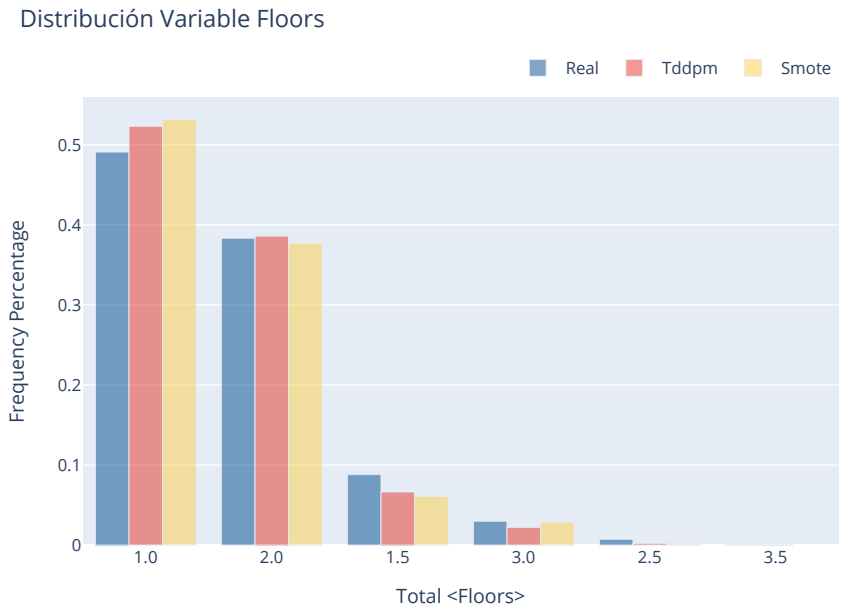

Figura A.9: Frecuencia del campo Floors en el modelo real y Top 2, King county (A-1)

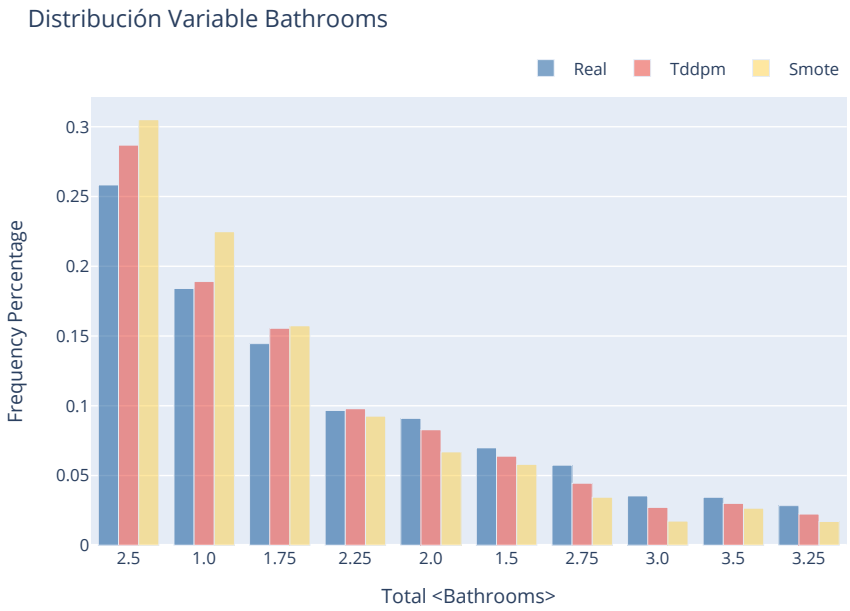

Figura A.10: Frecuencia del campo Bathrooms en el modelo real y Top 2, King county (A-1)

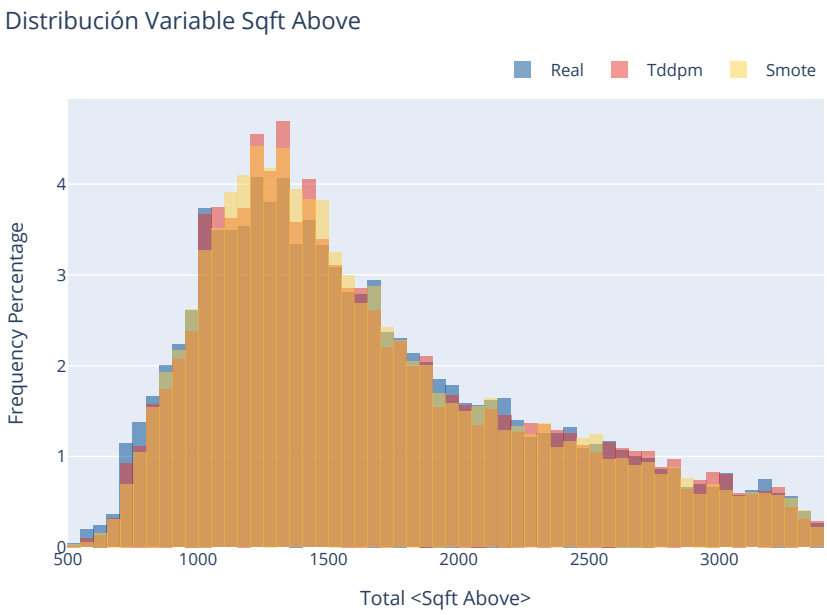

Figura A.11: Frecuencia del campo Sqft above en el modelo real y Top 2, King county (A-1)

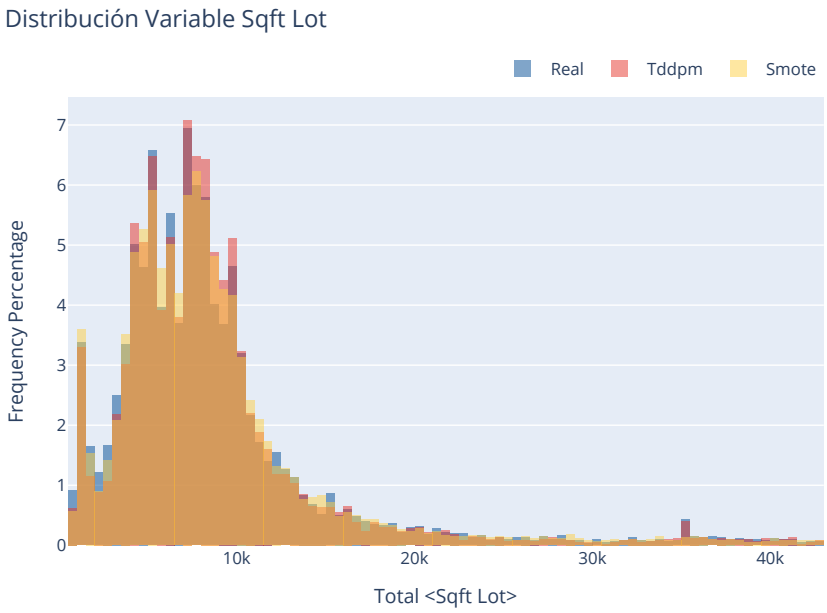

Figura A.12: Frecuencia del campo Sqft lot en el modelo real y Top 2, King county (A-1)

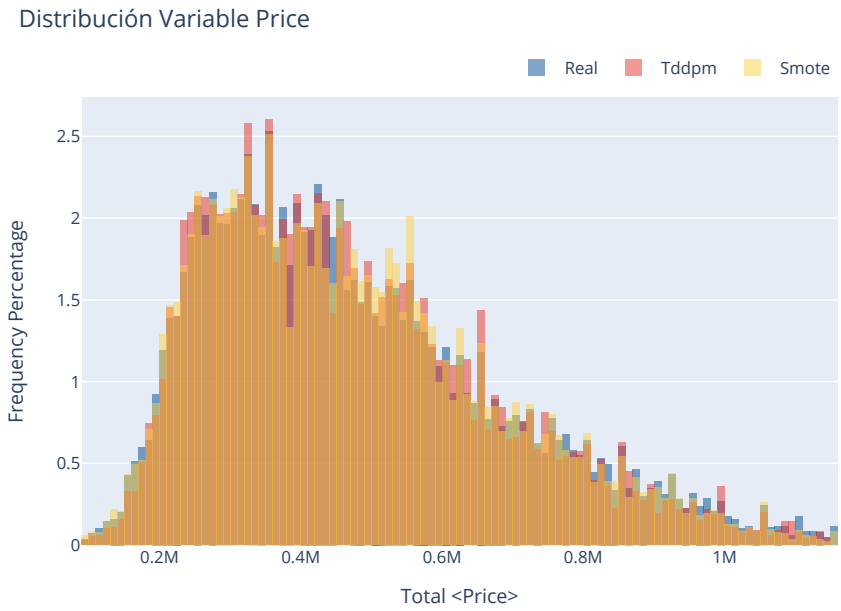

Figura A.13: Frecuencia del campo Price en el modelo real y Top 2, King county (A-1)

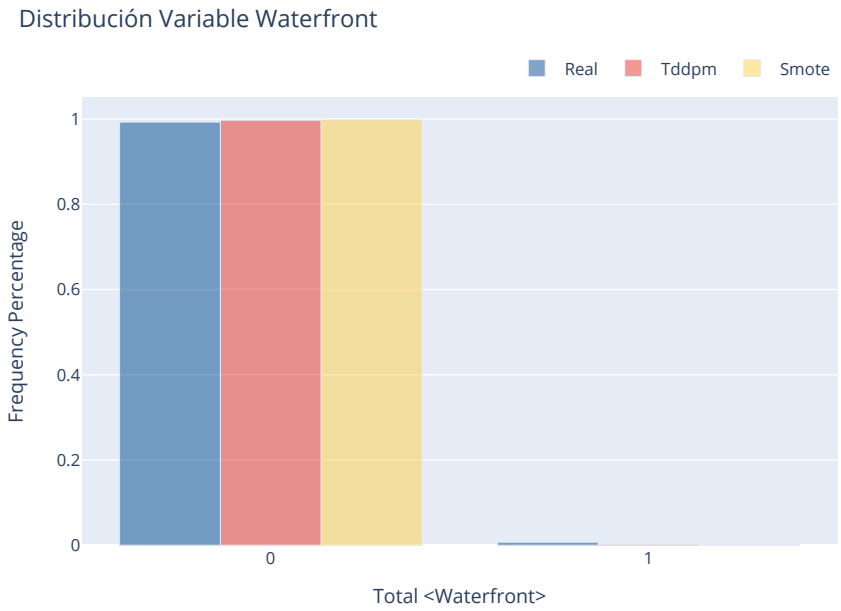

Figura A.14: Frecuencia del campo Waterfront en el modelo real y Top 2, King county (A-1)

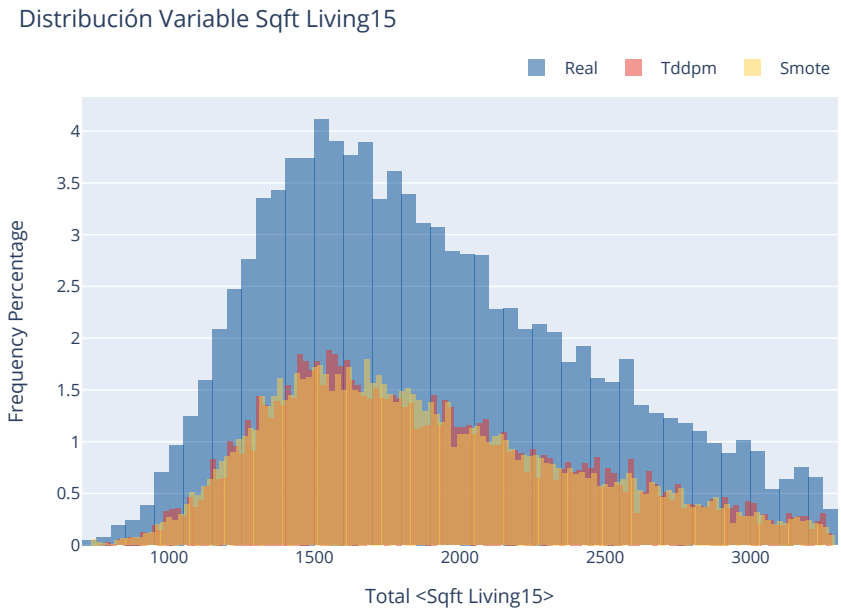

Figura A.15: Frecuencia del campo Sqft living15 en el modelo real y Top 2, King county (A-1)

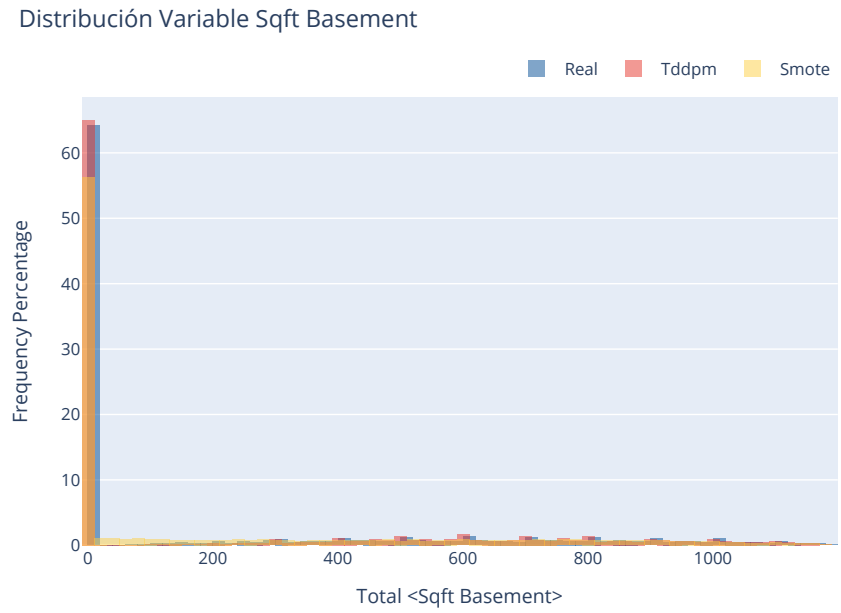

Figura A.16: Frecuencia del campo Sqft basement en el modelo real y Top 2, King county (A-1)

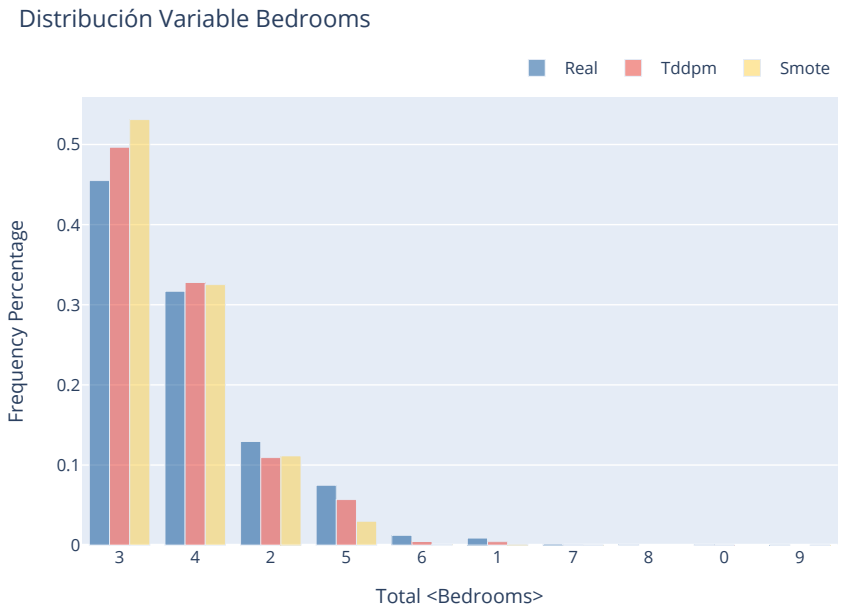

Figura A.17: Frecuencia del campo Bedrooms en el modelo real y Top 2, King county (A-1)

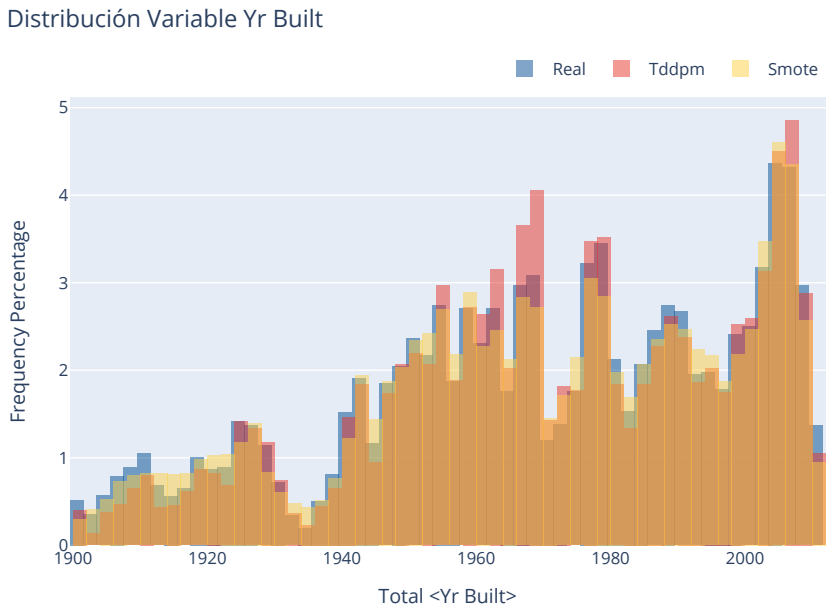

Figura A.18: Frecuencia del campo Yr built en el modelo real y Top 2, King county (A-1)

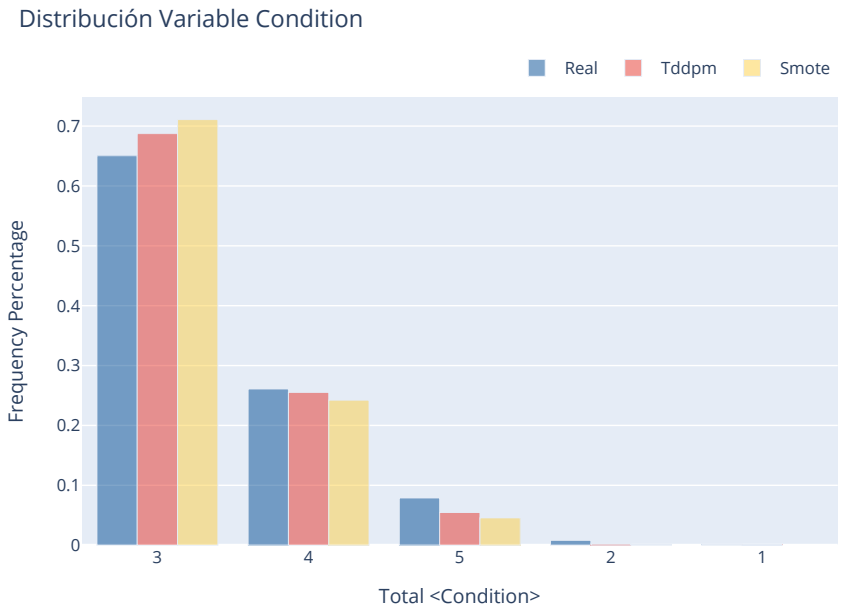

Figura A.19: Frecuencia del campo Condition en el modelo real y Top 2, King county (A-1)

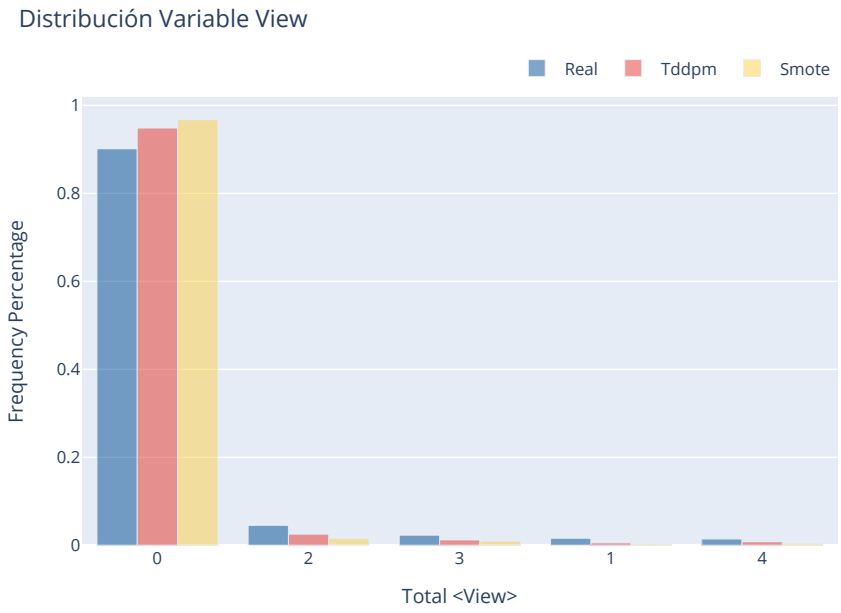

Figura A.20: Frecuencia del campo View en el modelo real y Top 2, King county (A-1)

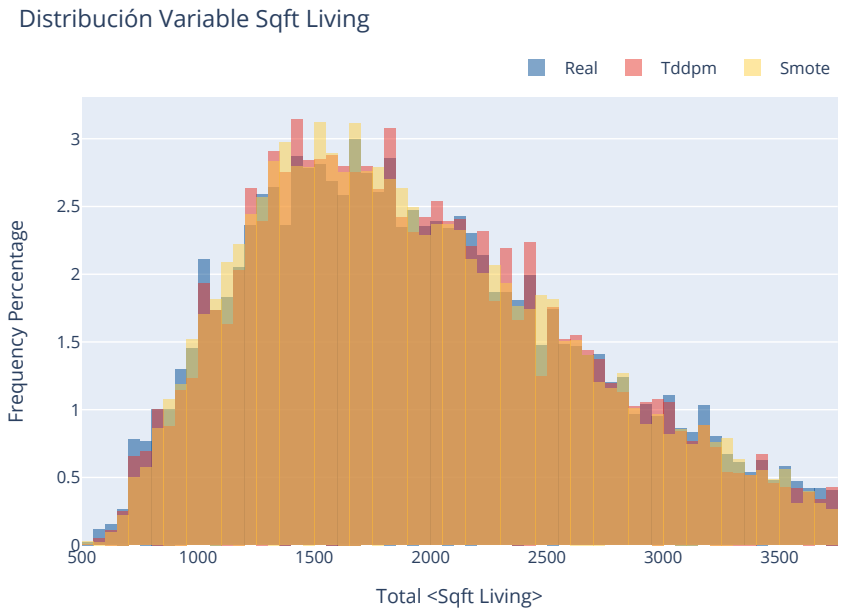

Figura A.21: Frecuencia del campo Sqft living en el modelo real y Top 2, King county (A-1)

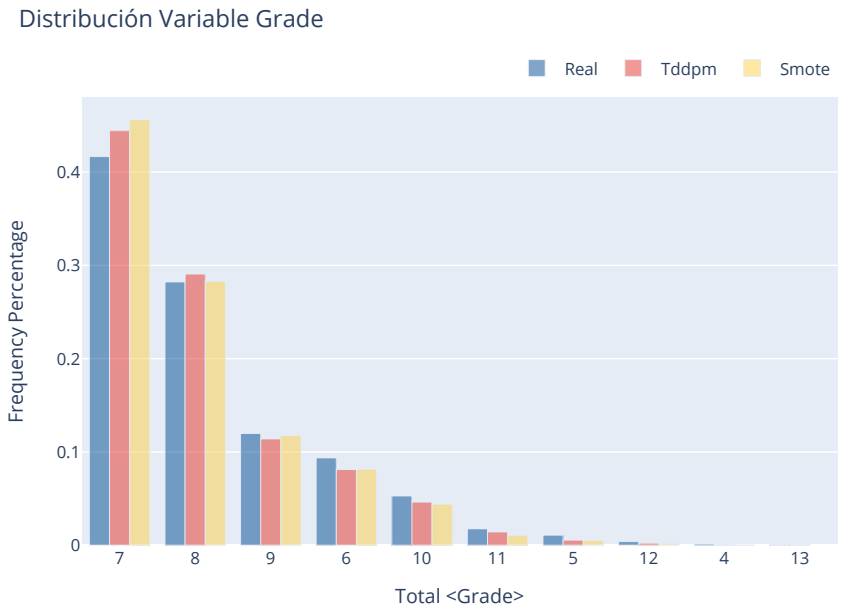

Figura A.22: Frecuencia del campo Grade en el modelo real y Top 2, King county (A-1)

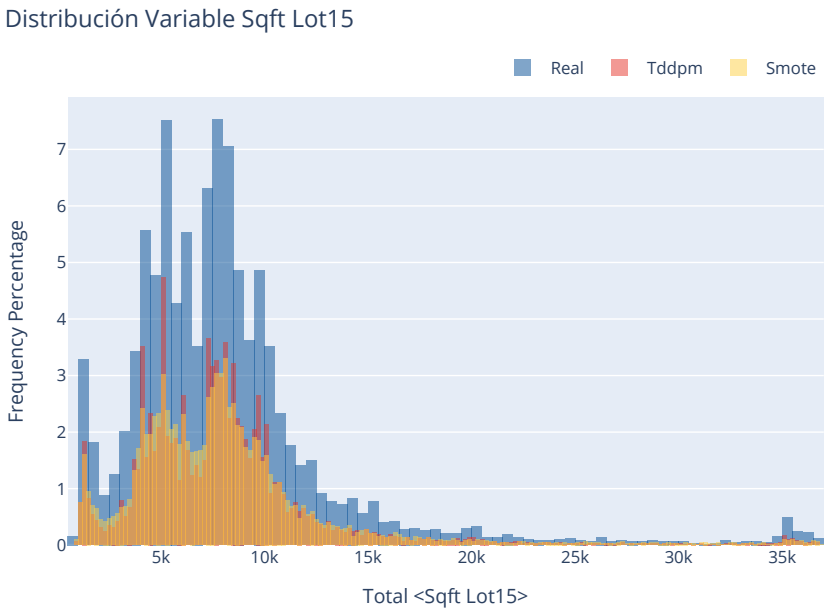

Figura A.23: Frecuencia del campo Sqft lot15 en el modelo real y Top 2, King county (A-1)

A.5. Figuras de correlación Económicos - Conjunto A

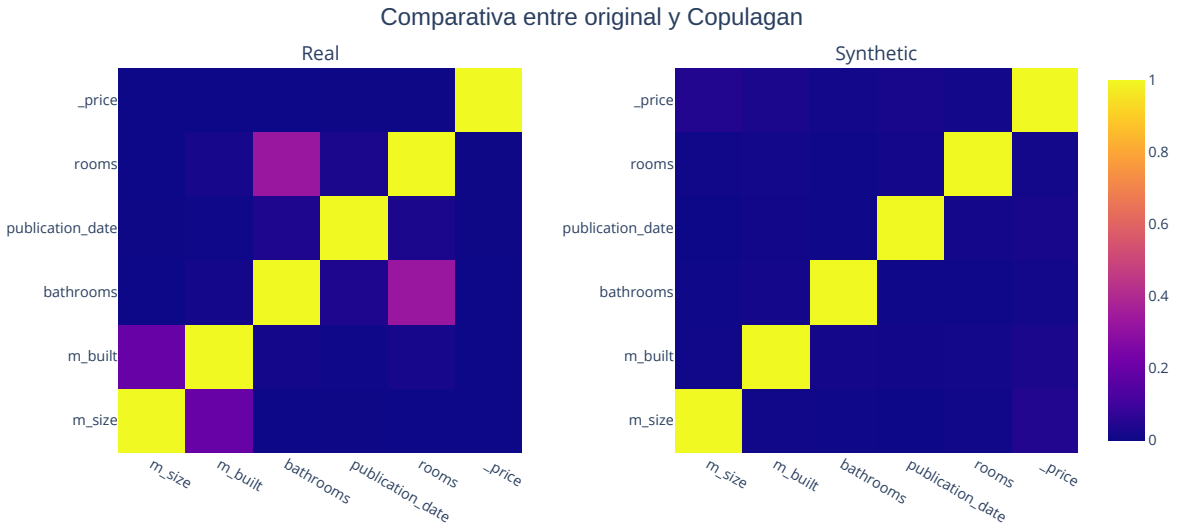

Figura A.24: Correlación de conjunto original de entrenamiento y Copulagan, Economicos (A-1)

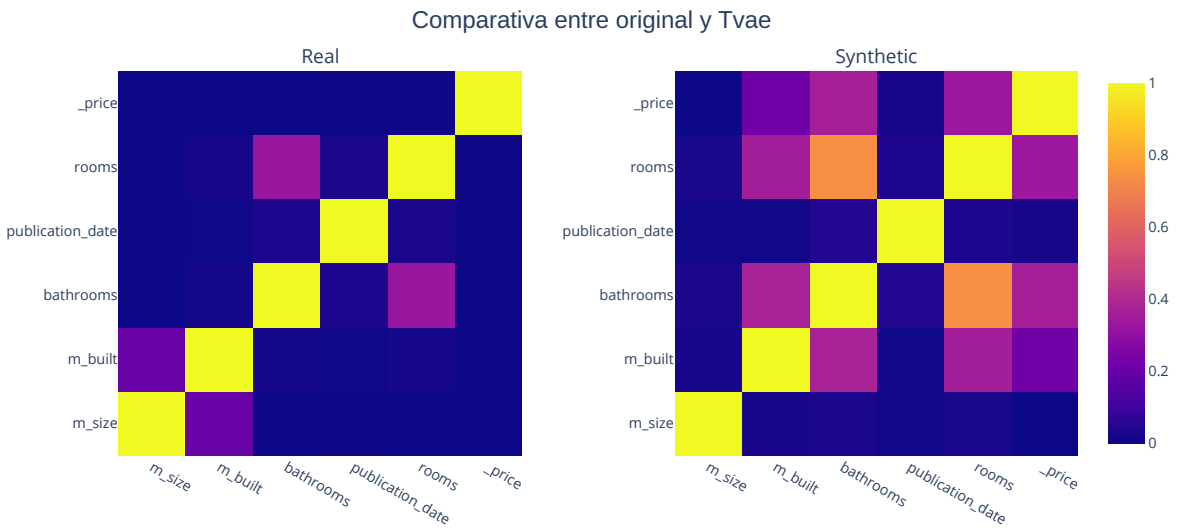

Figura A.25: Correlación de conjunto original de entrenamiento y Tvae, Economicos (A-1)

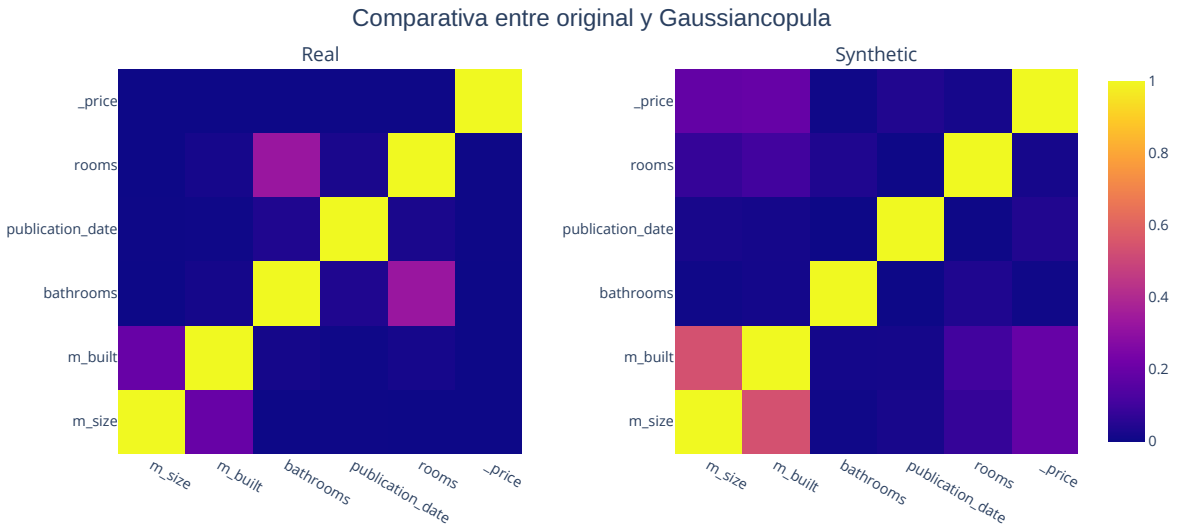

Figura A.26: Correlación de conjunto original de entrenamiento y Gaussiancopula, Economicos (A-1)

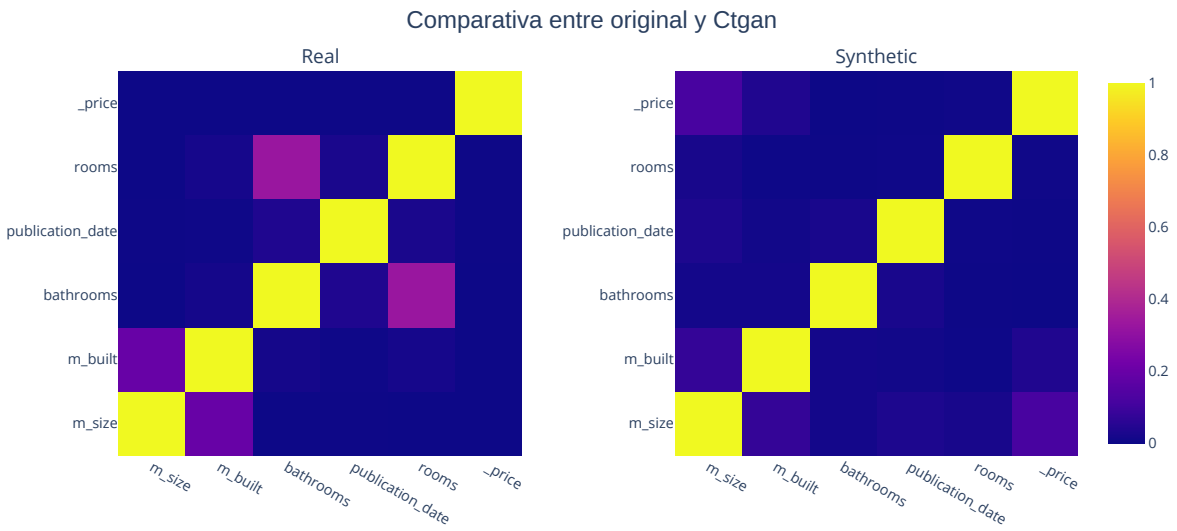

Figura A.27: Correlación de conjunto original de entrenamiento y Ctgan, Economicos (A-1)

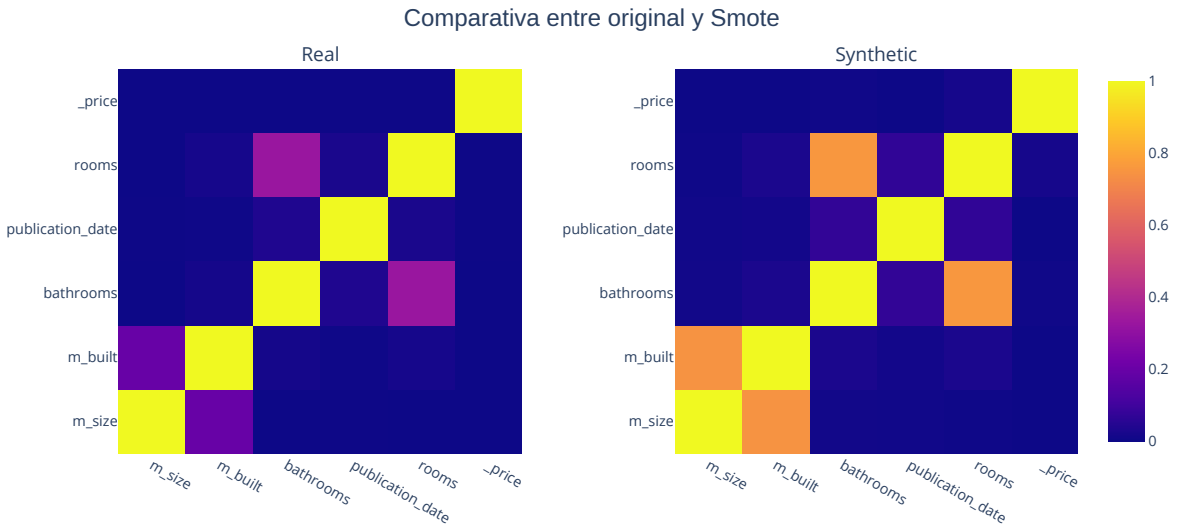

Figura A.28: Correlación de conjunto original de entrenamiento y Smote, Economicos (A-1)

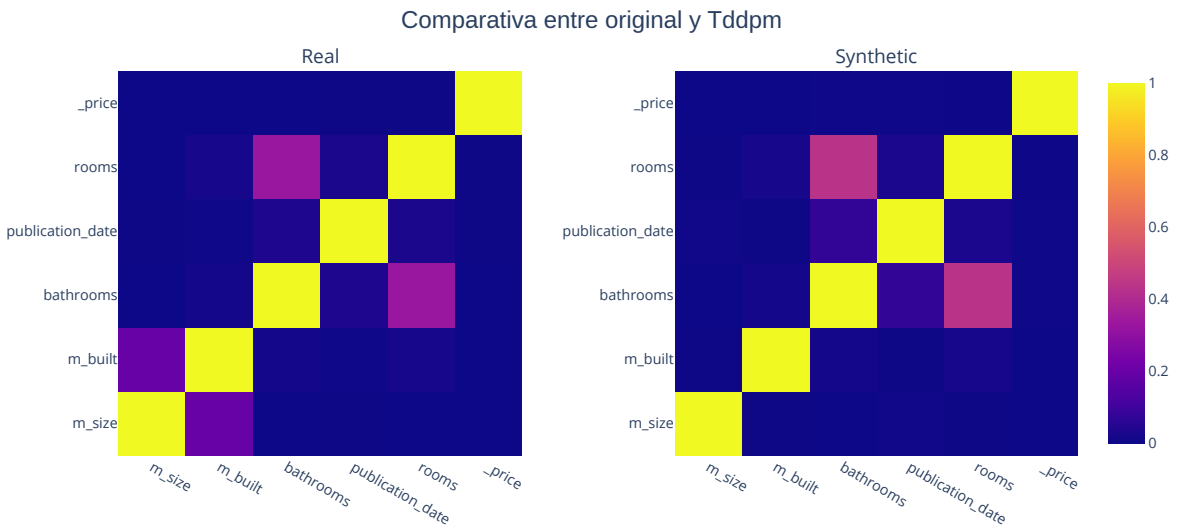

Figura A.29: Correlación de conjunto original de entrenamiento y Tddpm, Economicos (A-1)

A.6. Figuras de correlación Económicos - Conjunto B

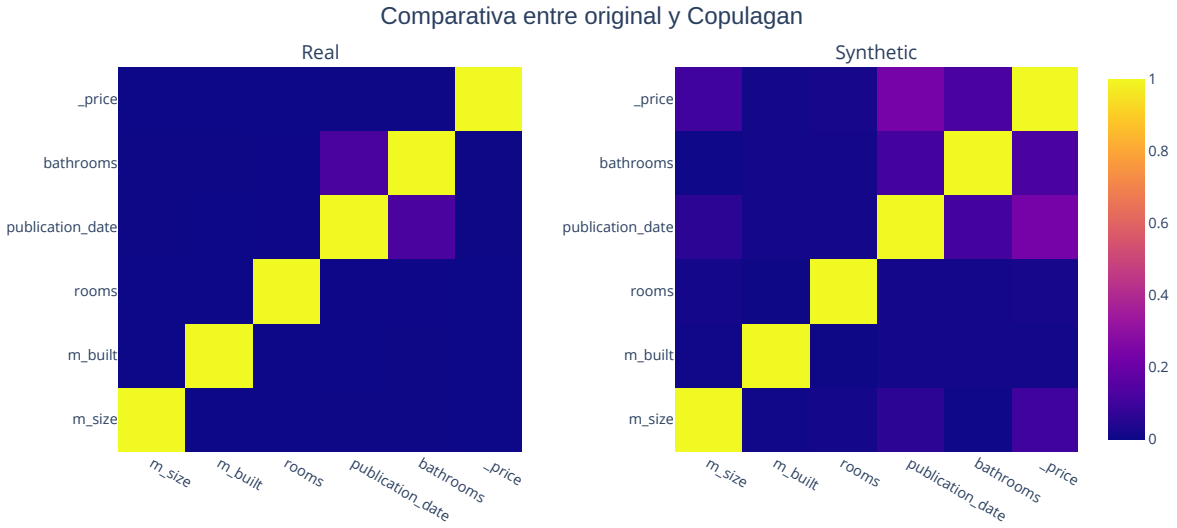

Figura A.30: Correlación de conjunto original de entrenamiento y Copulagan, Economicos (B-1)

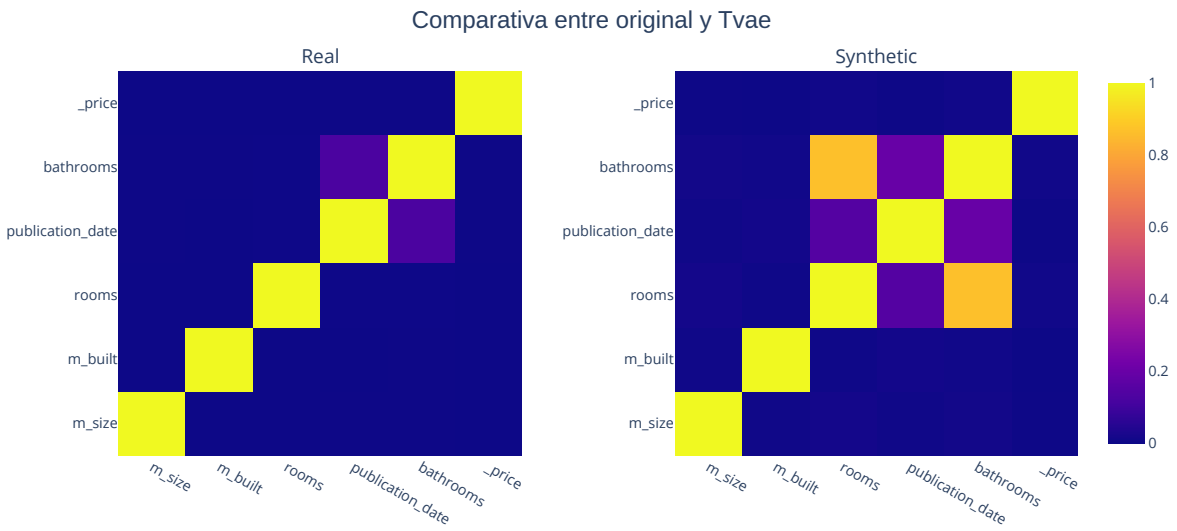

Figura A.31: Correlación de conjunto original de entrenamiento y Tvae, Economicos (B-1)

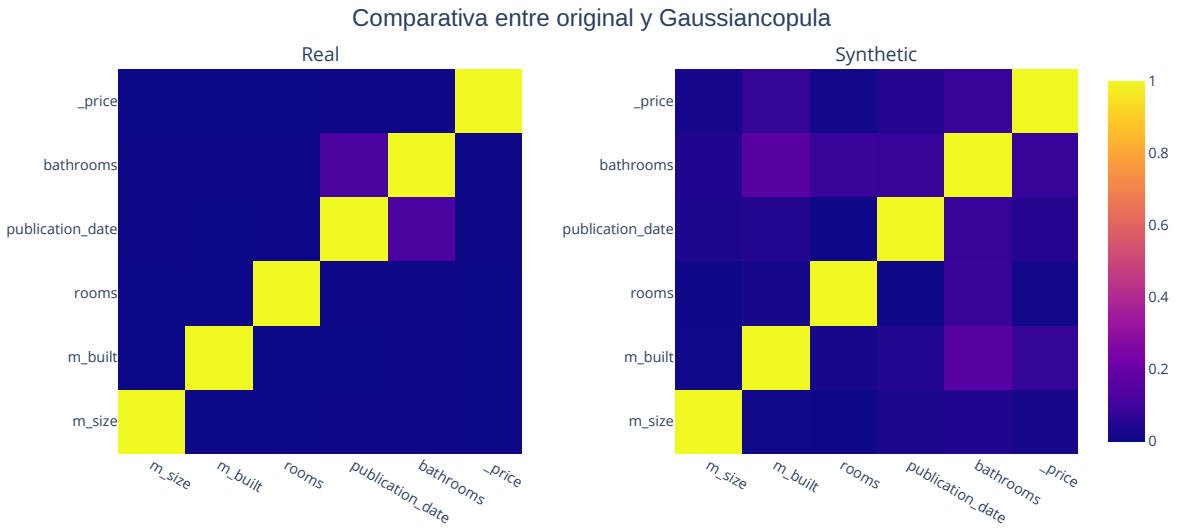

Figura A.32: Correlación de conjunto original de entrenamiento y Gaussiancopula, Economicos (B-1)

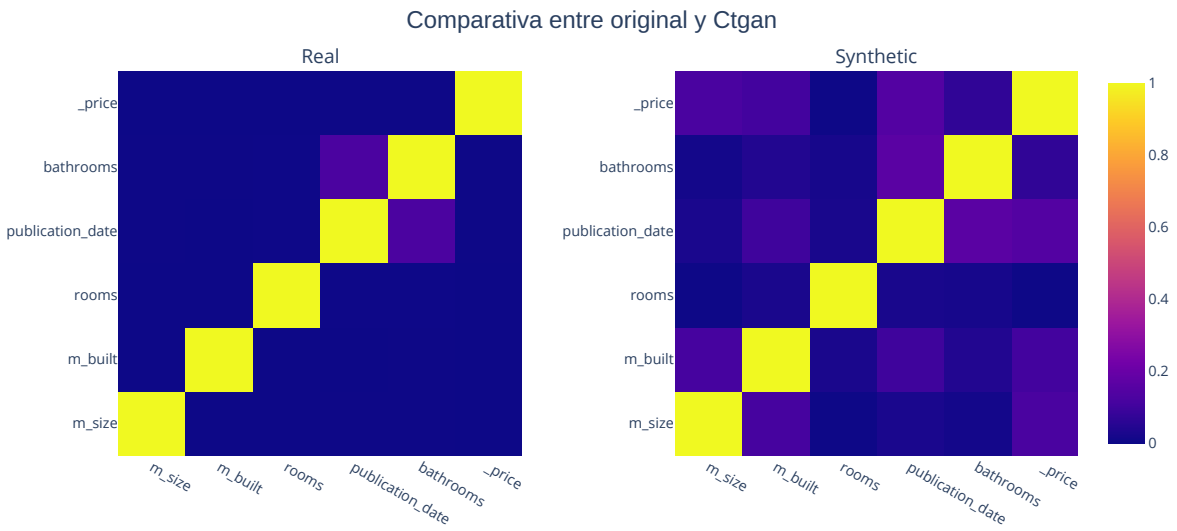

Figura A.33: Correlación de conjunto original de entrenamiento y Ctgan, Economicos (B-1)

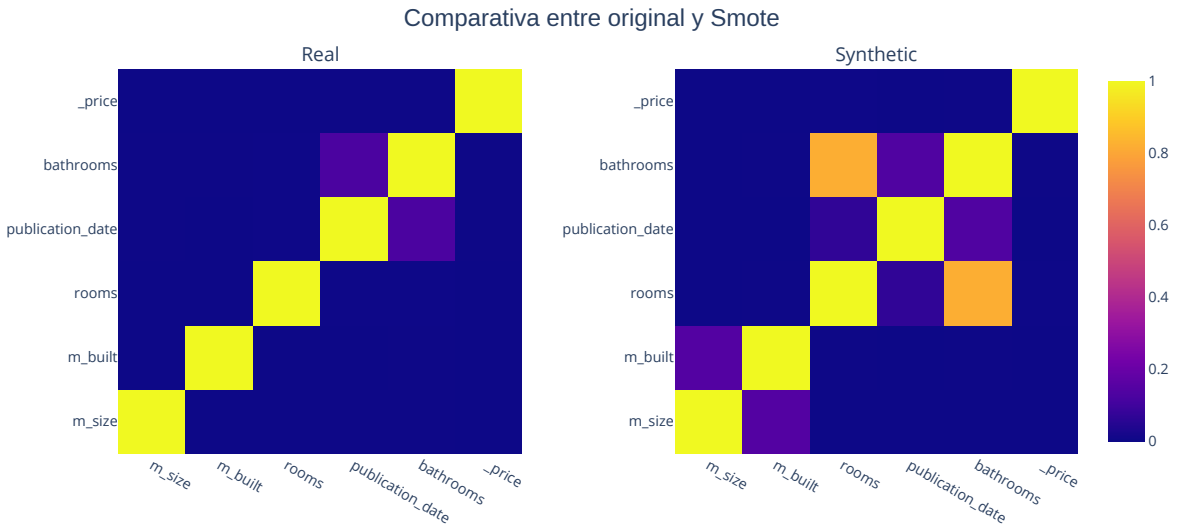

Figura A.34: Correlación de conjunto original de entrenamiento y Smote, Economicos (B-1)

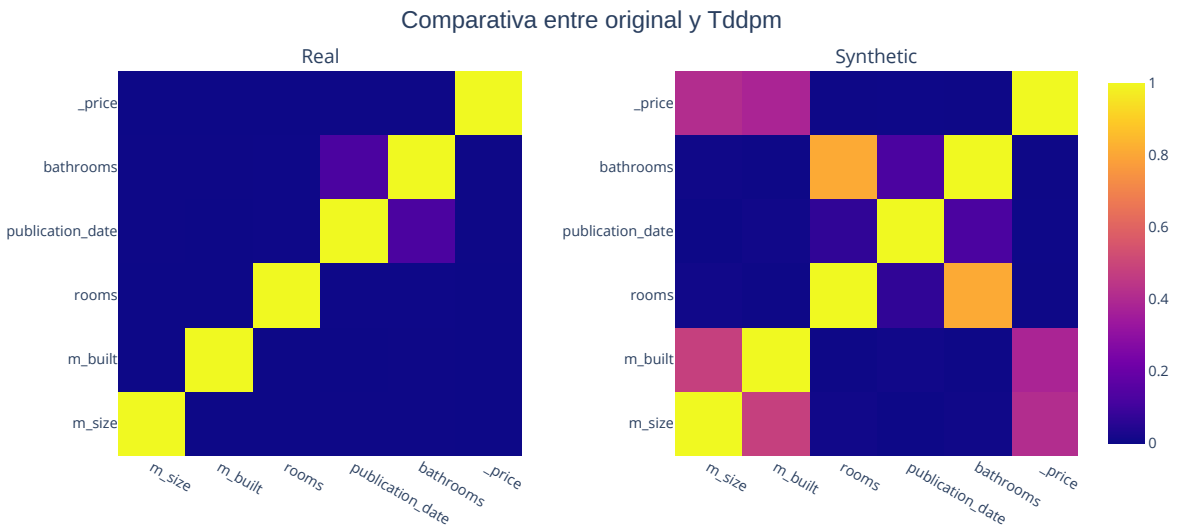

Figura A.35: Correlación de conjunto original de entrenamiento y Tddpm, Economicos (B-1)

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

Tabla A.1: Ejemplos de textos aleatoreos del modelo Tddpm, conjunto Economicos (A-1)

description
Departamento de 2 pisos, 1 baño completo, living comedor con salida a terraza en el primer piso (cocina equipada), sala de estar, quincho, lavandería, gimnasio, estacionamiento para dos vehículos, 3 estacionamientos, 4 bodegas, logia techada, horno flotante, campana termopanel, cortinas fluorescentes, calefacción por radiadores, agua potable, refrigerador/calefacción central, altura del jardín, amplio hall de entrada, gran vista panorámica sobre las calles más cercanas a supermercados, centros comerciales, bancos etc.
Piso de madera en el primer piso, living comedor con salida a terraza que tienen 2 baños completos (incluye una sala de estar) Living comedor amplio para lavadora separados por un bodega altura del hall de entrada; 1 estacionamientos techado/calefacción central cerrada: horno termopanel cubierta durante todo el año!
Piso de madera en el primer piso, living comedor con salida a terraza que tiene una vista panorámica al centro del barrio Peñalolén por la zona principal (con baño), sala de estar, quincho para 2 vehículos, dos estacionamientos techados, 1 bodega; un amplio hall de entrada: Living comedor como walk in closet/cocina equipada completamente iluminada) Calefacción por radiadores A pasos de metro Plaza Las Palmas +569 9 7 8 6 5 4
Piso de madera en el primer piso, living comedor con salida a terraza que tienen 2 baños completos (incluye una sala de estar) Living comedor amplio para dos autos (2 estacionamientos) Cocina equipada completamente equipada Con vista panorámica al jardín Las casas se venden por las ventanas del sector: 1 bodega Calefacción central Azotea Techado La propiedad cuenta con lavandería Edificio está equipado con estacionamiento privado Amplio Hall de entrada Sala de Estación Terraza Gran Jardín 3 Baños 4 Bodega 5 pisos Totales Construidos Terreno construido El terreno consta de un gran hall de acceso principal Encimera Ventanas termopanel Termopanel Clima Solar Agua Natural
Piso de madera en el primer piso, living comedor con salida a terraza que tienen 2 baños (incluye una sala de estar) Living comedor amplio para dos autos (2 estacionamientos) Cocina equipada completamente equipada Con vista panorámica al jardín Las casas se venden por todo lo contrario como la entrada principal del sector cercano a estacionamiento Además cuenta con un gran hall de acceso Amplio Hall de entrada Calefacción central Horno termopanel Ventanas Termopanel Lavandería Jardín Techado Sala de reuniones 1 Baño completo Terraza: 3 pisos + quincho
Departamento de 2 pisos, 1 baño con salida a terraza en living comedor que comparte un amplio hall de entrada para una gran vista panorámica por el sector del metro Los Santos Las Condes (principal) al lado de la sala de estar principal como dos estacionamiento más bodega según lo anteriormente mencionada: Living comedor equipado con walk in closet; Cocina amoblada completamente equipada cubierta de granito cerrada Condominio cuenta con 3 estacionamientos principales +569 7 6 8 9 4 5
Departamento de 2 pisos, 1 baño con walk in closet, living comedor en su primer piso, sala de estar, quincho, lavandería, gimnasio, estacionamiento para dos vehículos, 3 estacionamientos, terraza privada, logia techada, bodega por todo el sector del barrio que está cerca de las principales arterias comerciales, supermercados, bancos etc.
Departamento de 2 pisos, 1 baño completo, living comedor con salida a terraza en el primer piso: Living comedor para 3 personas (incluye una sala de estar) Cocina equipada completamente equipada Con vista al metro Las Condes Aproximadamente 4 metros cuadrados del Metro Los Condes Calefacción por radiadores El edificio cuenta con lavandería Lavandería Jardín Edificio principal Amplio Hall de entrada Sala de estar amplio Terraza Techada Encimera Ventanas termopanel Termopanel Horno-Horno/Calefacción central Portón Controlado De Propiedades
Departamento de 2 pisos, 1 baño con walk in closet, living comedor en su primer piso, sala de estar, quincho, lavandería, gimnasio, estacionamiento para dos vehículos, 3 estacionamientos, terraza privada, logia, bodega, cocina equipada, horno termopanel, campana portón trasero, jardines infantiles, supermercados, bancos etc.
Piso de madera en el primer piso, living comedor con salida a terraza que tienen 2 baños completos (incluye una sala de estar) Living comedor amplio para dos autos (2 estacionamientos) 1 bodega

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

Tabla A.2: Ejemplos de textos aleatoreos del modelo Tddpm, conjunto Economicos (B-1)

description
Cuenta con 3 dormitorios, 2 baños, living comedor, cocina amoblada, logia, estacionamiento techado para dos vehículos en el primer piso de la casa se encuentra al lado del centro de Lampue (Lima).
Departamento de 85 m2, con vista despejada hacia la cordillera en el primer piso (con ventanas termopanel) 2 dormitorios 1 baño Cocina amoblada Logia Terraza Living Comedor Condominio cuenta con gimnasio Sala multiuso Lavandería Conserjería 24 horas Gastos comunes $50.000 aprox
Departamento de dos dormitorios, 2 baños, living comedor, cocina amoblada con encimera, horno empotrado, campana, terraza techada, estacionamiento subterráneo
Cuenta con: 2 dormitorios, 1 baño, living comedor, cocina amoblada, logia, estacionamiento para dos vehículos
Excelente casa de dos pisos, con 2 dormitorios, 1 baño, living comedor, cocina amoblada, logia, estacionamiento techado para 3 vehículos en el primer piso (con ventanas termopanel)
Vendo parcela de 5000 m2 en Frutillar, con vista al mar por la cordillera del Lago Los Lobos (Frutillar). Condominio cuenta con accesos controlados las 24 hectareas que permiten disfrutar de un lugar tranquilo para disfrute de una hermosa paisaje natural!
Departamento en San Joaquín, 1 dormitorio, 1 baño, living comedor con salida a terraza (con vista despejada) Cocina amoblada equipada con campana, horno empotrado, cubiertas de granito para lavadora Calefacción por radiadores Edificio cuenta con gimnasio, sala multiuso, quincho, estacionamiento visitas
Departamento amoblado, 2 dormitorios, 1 baño, living comedor con salida a terraza en el primer piso (cocina equipada), logia, estacionamiento de visitas, sala multiuso, gimnasio, lavandería
Departamento de 1 dormitorio en suite con walk in closet, living comedor separados (con salida a terraza), cocina equipada con horno eléctrico, campana empotrada, cubierta de granito para lavadora; logia cerrada completamente amoblada por radiadores). El edificio cuenta con gimnasio, sala multiuso, quincho, salón de eventos, estacionamiento subterráneo

A.9. Estadísticos KingCounty

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

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1440 1540 1560 1500 1610]	[1440. 1540. 1520. 1560. 1580.]	[1440. 1610. 1670. 1580. 1240.]	[1615 1676 1794 1639 1803]
top5_freq	[156 154 152 137 136]	[188 177 168 166 163]	[19 17 17 16 16]	[27 26 25 24 24]
top5_prob	[0.00902256 0.00890688 0.00879121 0.00792366 0.00786582]	[0.00869847 0.00818952 0.0077731 0.00768056 0.00754176]	[0.00087906 0.00078653 0.00078653 0.00074026 0.00074026]	[0.00124925 0.00120298 0.00115671 0.00111044 0.00111044]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	1983	**1979**	1976	1778
std_err	5.181	4.473	4.427	**4.667**
upper_ci	1993	**1988**	1984	1787
lower_ci	1973	**1970**	1967	1769
std	681.232	657.650	650.871	**686.109**
iqr	880.000	820.140	**844.567**	824.000
iqr_normal	652.345	607.970	**626.079**	610.832
mad	533.237	510.997	511.611	**515.429**
mad_normal	668.313	640.440	641.209	**645.995**
coef_var	0.344	**0.332**	0.329	0.386
range	5811	**5811**	5491	**5811**
max	6210	**6210**	5994	**6210**
min	399.000	**399.000**	503.651	**399.000**
skew	1.095	1.148	**1.078**	1.275
kurtosis	4.572	5.170	**4.410**	6.223
jarque_bera	5237	8987	**5975**	15215
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1440	**1440**	1440	1615
mode_freq	0.009	**0.009**	0.001	0.001
median	1840	**1840**	1833	1693
0.1 %	740.000	399.000	**812.077**	498.448
1.0 %	950.000	**969.449**	987.551	702.120
5.0 %	1140	**1165**	1171	875
25.0 %	1480	1510	**1493**	1293
75.0 %	2360	2330	**2338**	2117
95.0 %	3280	**3237**	3228	3006
99.0 %	4050	3957	3922	**4127**
99.9 %	4986	5223	**4861**	5405

Tabla A.4: Propiedades estadisticas de variable yr_built, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[2014 2005 2006 2004 2007]	[2005. 2006. 2004. 2003. 1977.]	[2014. 2005. 2003. 1977. 2008.]	[1900 1972 1974 1973 1975]
top5_freq	[449 371 366 350 347]	[524 507 457 431 426]	[244 113 106 94 93]	[894 517 512 509 466]
top5_prob	[0.02596877 0.02145749 0.02116831 0.02024291 0.0200694]	[0.02424467 0.0234581 0.02114468 0.0199417 0.01971036]	[0.01128898 0.00522809 0.00490423 0.00434903 0.00430277]	[0.04136399 0.02392079 0.02368945 0.02355064 0.0215611]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	1971	1972	**1971**	1956
std_err	0.224	0.190	0.195	**0.199**
upper_ci	1972	1973	**1972**	1956
lower_ci	1971	1972	**1971**	1956
std	29.436	27.911	28.741	**29.289**
iqr	46.000	43.000	**44.558**	42.000
iqr_normal	34.100	31.876	**33.031**	31.135
mad	24.632	23.146	24.052	**24.723**
mad_normal	30.872	29.010	30.145	**30.986**
coef_var	0.015	0.014	0.015	**0.015**
range	115.000	**115.000**	114.943	**115.000**
max	2015	**2015**	2015	**2015**
min	1900	**1900**	1900	**1900**
skew	-0.472	-0.486	**-0.473**	-0.269
kurtosis	2.337	2.451	**2.336**	2.018
jarque_bera	957.631	**1122.949**	1203.111	1130.580
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	2014	2005	**2014**	1900
mode_freq	0.026	**0.024**	0.011	0.041
median	1975	**1975**	**1975**	1957
0.1 %	1900	**1900**	1900	**1900**
1.0 %	1904	**1906**	**1906**	1900
5.0 %	1915	1919	**1917**	1904
25.0 %	1951	1954	**1952**	1936
75.0 %	1997	**1997**	1997	1978
95.0 %	2011	2009	**2009**	1998
99.0 %	2014	**2014**	**2014**	2004
99.9 %	2015	**2015**	2014	2009

Tabla A.5: Propiedades estadisticas de variable sqft_living, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1400 1300 1720 1250 1540]	[1300. 1440. 2240. 1800. 2040.]	[1280. 1830. 1800. 1870. 1610.]	[290 1131 961 1842 1010]
top5_freq	[109 107 106 106 105]	[129 125 124 115 108]	[16 12 11 11 11]	[52 19 18 18 18]
top5_prob	[0.00630422 0.00618855 0.00613071 0.00613071 0.00607287]	[0.00596863 0.00578356 0.00573729 0.00532087 0.00499699]	[0.00074026 0.0005552 0.00050893 0.00050893 0.00050893]	[0.00240596 0.0008791 0.00083283 0.00083283 0.00083283]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	2074	**2093**	2035	1950
std_err	6.900	7.107	5.664	**6.782**
upper_ci	2087	**2107**	2046	1963
lower_ci	2060	**2079**	2024	1937
std	907.298	1044.878	**832.727**	997.085
iqr	1110	1051	**1069**	1251
iqr_normal	822.844	779.193	**792.472**	927.368
mad	693.180	**685.111**	646.984	761.028
mad_normal	868.773	**858.659**	810.875	953.807
coef_var	0.437	0.499	**0.409**	0.511
range	11760	**13239**	7600	8450
max	12050	**13540**	7955	8740
min	290.000	301.196	354.765	**290.000**
skew	1.371	4.315	1.141	**1.228**
kurtosis	7.167	42.802	5.175	**6.081**
jarque_bera	17922	1493680	8948	**13977**
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1400	**1300**	1280	290
mode_freq	0.006	**0.006**	0.001	0.002
median	1910	**1925**	1882	1823
0.1 %	522.890	**567.112**	592.630	290.000
1.0 %	720.000	**758.864**	760.455	389.000
5.0 %	940.000	**970.000**	978.660	643.600
25.0 %	1430	1450	**1423**	1227
75.0 %	2540	**2501**	2492	2478
95.0 %	3740	3665	3575	**3707**
99.0 %	4921	**4926**	4618	5150
99.9 %	6966	13540	6173	**7275**

Tabla A.6: Propiedades estadisticas de variable view, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[0 2 3 1 4]	[0 2 3 4 1]	[0 2 3 4 1]	[0 2 4 3 1]
top5_freq	[15586 783 396 275 250]	[20486 520 304 166 137]	[20842 366 224 118 64]	[18000 1611 951 658 393]
top5_prob	[0.90144592 0.04528629 0.02290341 0.01590515 0.01445922]	[0.94785546 0.02405959 0.01406561 0.00768056 0.00633878]	[0.96428241 0.01693347 0.01036365 0.00545942 0.00296104]	[0.83283209 0.07453847 0.0440013 0.03044464 0.0181835]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	0.233	**0.127**	0.090	0.435
std_err	0.006	0.004	0.003	**0.007**
upper_ci	0.244	**0.135**	0.096	0.449
lower_ci	0.222	**0.120**	0.083	0.421
std	0.762	**0.580**	0.493	1.051
iqr	0.000	0.000	0.000	0.000
iqr_normal	0.000	0.000	0.000	0.000
mad	0.420	**0.241**	0.173	0.724
mad_normal	0.527	**0.303**	0.217	0.907
coef_var	3.269	4.550	5.495	**2.419**
range	4.000	4.000	4.000	4.000
max	4.000	4.000	4.000	4.000
min	0.000	0.000	0.000	0.000
skew	3.402	4.829	5.846	**2.348**
kurtosis	13.971	26.467	37.976	**7.289**
jarque_bera	120072	579932	1224844	**36428**
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	0.000	0.000	0.000	0.000
mode_freq	0.901	**0.948**	0.964	0.833
median	0.000	0.000	0.000	0.000
0.1 %	0.000	0.000	0.000	0.000
1.0 %	0.000	0.000	0.000	0.000
5.0 %	0.000	0.000	0.000	0.000
25.0 %	0.000	0.000	0.000	0.000
75.0 %	0.000	0.000	0.000	0.000
95.0 %	2.000	**1.000**	0.000	**3.000**
99.0 %	4.000	3.000	3.000	**4.000**
99.9 %	4.000	4.000	4.000	4.000

Tabla A.7: Propiedades estadisticas de variable price, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[350000. 450000. 425000. 550000. 325000.]	[450000. 525000. 350000. 300000. 550000.]	[350000. 550000. 450000. 300000. 325000.]	[75000. 640966. 220325. 366666. 393268.]
top5_freq	[143 140 123 123 123]	[151 142 138 137 127]	[222 214 176 173 173]	[70 3 3 3 3]
top5_prob	[0.00827068 0.00809717 0.00711394 0.00711394 0.00711394]	[0.00698654 0.00657012 0.00638505 0.00633878 0.00587609]	[0.01027112 0.00990099 0.00814287 0.00800407 0.00800407]	[0.00323879 0.00013881 0.00013881 0.00013881 0.00013881]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	537768	**541748**	517890	761330
std_err	2749	**3240**	2061	4730
upper_ci	543156	**548099**	521930	770601
lower_ci	532380	**535397**	513850	752058
std	361464	476348	**303039**	695431
iqr	319850	303892	**310000**	457015
iqr_normal	237105	225275	**229803**	338786
mad	231680	**231849**	209811	445065
mad_normal	290368	**290579**	262960	557806
coef_var	0.672	0.879	**0.585**	0.913
range	7625000	**7624099**	3722000	4823597
max	7700000	**7700000**	3800000	4898597
min	75000	75901	78000	**75000**
skew	4.032	9.265	2.492	**2.690**
kurtosis	39.678	129.395	**13.821**	11.214
jarque_bera	1016020	14695939	**127811**	86829
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	350000	450000	**350000**	75000
mode_freq	0.008	**0.007**	0.010	0.003
median	450000	454386	**450500**	552289
0.1 %	95000	107307	**100000**	75000
1.0 %	154467	163477	**154956**	112801
5.0 %	210000	219232	**210000**	200729
25.0 %	320150	**325000**	315000	378490
75.0 %	640000	**628892**	625000	835505
95.0 %	1150000	**1062302**	1047700	2323820
99.0 %	1950000	**1875807**	1689740	3797286
99.9 %	3331995	7700000	**2700000**	4573432

Tabla A.8: Propiedades estadisticas de variable sqft_basement, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[0 600 700 500 800]	[0. 600. 700. 800. 500.]	[0. 850. 500. 600. 800.]	[0 1 2 4 3]
top5_freq	[10553 182 169 167 164]	[13195 219 214 211 187]	[11690 15 13 11 11]	[9489 331 331 310 294]
top5_prob	[0.61035281 0.01052632 0.00977444 0.00965876 0.00948525]	[0.61051219 0.01013279 0.00990145 0.00976264 0.0086522]	[5.40853151e-01 6.93994633e-04 6.01462015e-04 5.08929398e-04 5.08929398e-04]	[0.43904132 0.01531486 0.01531486 0.01434322 0.01360292]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	287.933	309.263	271.272	**286.436**
std_err	3.337	**3.344**	2.722	3.142
upper_ci	294.472	315.817	276.608	**292.593**
lower_ci	281.393	302.708	265.937	**280.278**
std	438.727	491.630	400.236	**461.855**
iqr	550.000	606.287	**510.562**	470.000
iqr_normal	407.716	449.441	**378.480**	348.412
mad	360.277	381.922	327.925	**349.719**
mad_normal	451.541	478.669	410.993	**438.308**
coef_var	1.524	1.590	**1.475**	1.612
range	4820	**4820**	3472	3626
max	4820	**4820**	3472	3626
min	0.000	0.000	0.000	0.000
skew	1.571	2.955	**1.497**	2.114
kurtosis	5.639	21.759	**4.896**	8.512
jarque_bera	12126	348368	**11312**	43464
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	0.000	0.000	0.000	0.000
mode_freq	0.610	**0.611**	0.541	0.439
median	0.000	**0.000**	**0.000**	5.000
0.1 %	0.000	0.000	0.000	0.000
1.0 %	0.000	0.000	0.000	0.000
5.0 %	0.000	0.000	0.000	0.000
25.0 %	0.000	0.000	0.000	0.000
75.0 %	550.000	606.287	**510.562**	470.000
95.0 %	1180	**1166**	1082	1237
99.0 %	1650	**1625**	1498	1924
99.9 %	2324	4820	**2132**	3107

Tabla A.9: Propiedades estadisticas de variable bedrooms, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[3 4 2 5 6]	[3 4 2 5 6]	[3 4 2 5 1]	[4 3 2 5 6]
top5_freq	[7865 5477 2237 1292 212]	[10642 7215 2275 1292 101]	[11401 7033 2445 680 38]	[8962 5688 2244 2218 696]
top5_prob	[0.45488722 0.3167727 0.12938115 0.07472527 0.01226142]	[0.49238884 0.33382686 0.10526072 0.05977884 0.00467311]	[0.52748219 0.32539095 0.11312113 0.03146109 0.00175812]	[0.41465784 0.26317494 0.1038264 0.10262342 0.03220284]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	3.368	**3.359**	3.274	3.909
std_err	0.007	**0.005**	0.005	0.020
upper_ci	3.382	**3.370**	3.283	3.949
lower_ci	3.354	**3.349**	3.265	3.870
std	0.931	**0.788**	0.710	2.949
iqr	1.000	1.000	1.000	1.000
iqr_normal	0.741	0.741	0.741	0.741
mad	0.734	**0.655**	0.585	1.159
mad_normal	0.920	**0.821**	0.734	1.453
coef_var	0.277	**0.235**	0.217	0.754
range	33.000	8.000	5.000	**33.000**
max	33.000	9.000	6.000	**33.000**
min	0.000	1.000	1.000	**0.000**
skew	2.304	**0.347**	0.103	7.750
kurtosis	63.268	3.808	3.003	**76.541**
jarque_bera	2631992	1023	38	**5086707**
jarque_bera_pval	0.000	0.000	0.000	**0.000**
mode	3.000	**3.000**	**3.000**	4.000
mode_freq	0.455	**0.492**	0.527	0.415
median	3.000	**3.000**	**3.000**	4.000
0.1 %	1.000	**1.000**	**1.000**	0.000
1.0 %	2.000	**2.000**	**2.000**	0.000
5.0 %	2.000	2.000	2.000	2.000
25.0 %	3.000	3.000	3.000	3.000
75.0 %	4.000	4.000	4.000	4.000
95.0 %	5.000	**5.000**	4.000	6.000
99.0 %	6.000	**5.000**	**5.000**	9.000
99.9 %	7.000	**6.000**	5.000	33.000

Tabla A.10: Propiedades estadisticas de variable condition, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[3 4 5 2 1]	[3 4 5 2 1]	[3 4 5 2]	[3 4 5 2 1]
top5_freq	[11248 4512 1364 139 27]	[14867 5518 1182 43 3]	[15524 5111 956 23]	[14953 4233 1578 476 373]
top5_prob	[0.65054945 0.26096009 0.07888953 0.00803933 0.0015616]	[6.87873039e-01 2.55309305e-01 5.46893074e-02 1.98954333e-03 1.38805349e-04]	[0.71823818 0.2364671 0.04423059 0.00106413]	[0.69185213 0.19585435 0.07301161 0.02202378 0.01725813]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	3.408	**3.362**	3.324	3.285
std_err	0.005	0.004	0.004	**0.005**
upper_ci	3.417	**3.370**	3.331	3.295
lower_ci	3.398	**3.355**	3.316	3.276
std	0.652	0.588	0.556	**0.705**
iqr	1.000	1.000	1.000	1.000
iqr_normal	0.741	0.741	0.741	0.741
mad	0.560	0.505	0.468	**0.530**
mad_normal	0.702	0.633	0.587	**0.665**
coef_var	0.191	**0.175**	0.167	0.215
range	4.000	**4.000**	3.000	**4.000**
max	5.000	5.000	5.000	5.000
min	1.000	**1.000**	2.000	**1.000**
skew	1.028	**1.315**	1.478	0.486
kurtosis	3.556	**3.846**	4.275	4.919
jarque_bera	3269	6872	9330	**4168**
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	3.000	3.000	3.000	3.000
mode_freq	0.651	**0.688**	0.718	0.692
median	3.000	3.000	3.000	3.000
0.1 %	1.000	2.000	2.000	**1.000**
1.0 %	3.000	**3.000**	**3.000**	1.000
5.0 %	3.000	3.000	3.000	3.000
25.0 %	3.000	3.000	3.000	3.000
75.0 %	4.000	4.000	4.000	4.000
95.0 %	5.000	**5.000**	4.000	**5.000**
99.0 %	5.000	5.000	5.000	5.000
99.9 %	5.000	5.000	5.000	5.000

Tabla A.11: Propiedades estadisticas de variable sqft_lot15, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[5000 4000 6000 7200 4800]	[5000. 4000. 6000. 7200. 4800.]	[4000. 5000. 6000. 5200. 4080.]	[651 4044 2593 1242 1858]
top5_freq	[349 289 224 160 120]	[376 338 256 173 126]	[72 66 34 32 24]	[4125 8 8 8 7]
top5_prob	[0.02018508 0.01671486 0.01295547 0.0092539 0.00694043]	[0.01739694 0.01563874 0.01184472 0.00800444 0.00582982]	[0.00333117 0.00305358 0.00157305 0.00148052 0.00111039]	[0.19085735 0.00037015 0.00037015 0.00037015 0.00032388]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	12725	11744	**12292**	8706
std_err	209.331	161.204	**169.505**	140.911
upper_ci	13135	12060	**12624**	8982
lower_ci	12315	11428	**11959**	8430
std	27525	23699	**24920**	20716
iqr	4963	4827	**4934**	6821
iqr_normal	3679	3579	**3657**	5056
mad	10095	8406	**9281**	8329
mad_normal	12652	10535	**11633**	10439
coef_var	2.163	2.018	**2.027**	2.380
range	870549	**868126**	779076	278056
max	871200	**868777**	779794	278707
min	651.000	**651.000**	717.897	**651.000**
skew	9.701	11.306	**9.300**	7.301
kurtosis	163.253	224.253	**144.307**	65.928
jarque_bera	18772189	44544403	**18294136**	3758100
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	5000	**5000**	4000	651
mode_freq	0.020	**0.017**	0.003	0.191
median	7615	**7683**	7686	4233
0.1 %	886.289	651.000	**930.012**	651.000
1.0 %	1189	**1197**	1232	651
5.0 %	1965	**2104**	2222	651
25.0 %	5083	**5107**	5127	1344
75.0 %	10046	9934	**10061**	8165
95.0 %	36822	35281	**36185**	26166
99.0 %	168296	114278	**144057**	127032
99.9 %	306998	220761	**268288**	228588

Tabla A.12: Propiedades estadisticas de variable waterfront, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[0 1]	[0 1]	[0 1]	[0 1]
top5_freq	[17166 124]	[21543 70]	[21594 20]	[20050 1563]
top5_prob	[0.99282822 0.00717178]	[0.99676121 0.00323879]	[9.99074674e-01 9.25326177e-04]	[0.92768241 0.07231759]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	0.007	**0.003**	0.001	0.072
std_err	0.001	**0.000**	0.000	0.002
upper_ci	0.008	**0.004**	0.001	0.076
lower_ci	0.006	**0.002**	0.001	0.069
std	0.084	**0.057**	0.030	0.259
iqr	0.000	0.000	0.000	0.000
iqr_normal	0.000	0.000	0.000	0.000
mad	0.014	**0.006**	0.002	0.134
mad_normal	0.018	**0.008**	0.002	0.168
coef_var	11.766	**17.543**	32.860	3.582
range	1.000	1.000	1.000	1.000
max	1.000	1.000	1.000	1.000
min	0.000	0.000	0.000	0.000
skew	11.681	**17.486**	32.828	3.302
kurtosis	137.443	306.760	1078.701	**11.906**
jarque_bera	13414600	84194718	1045976466	**110710**
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	0.000	0.000	0.000	0.000
mode_freq	0.993	**0.997**	0.999	0.928
median	0.000	0.000	0.000	0.000
0.1 %	0.000	0.000	0.000	0.000
1.0 %	0.000	0.000	0.000	0.000
5.0 %	0.000	0.000	0.000	0.000
25.0 %	0.000	0.000	0.000	0.000
75.0 %	0.000	0.000	0.000	0.000
95.0 %	0.000	**0.000**	**0.000**	1.000
99.0 %	0.000	**0.000**	**0.000**	1.000
99.9 %	1.000	**1.000**	0.000	**1.000**

Tabla A.13: Propiedades estadisticas de variable date, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['20140623T000000' '20140625T000000' '20140626T000000' '20150421T000000' '20150325T000000']	['20140623T000000' '20140716T000000' '20140709T000000' '20150427T000000' '20140625T000000']	['20150421T000000' '20140623T000000' '20140625T000000' '20150407T000000' '20140722T000000']	['20150310T000000' '20150417T000000' '20140811T000000' '20150108T000000' '20140610T000000']
top5_freq	[123 105 101 101 101]	[203 161 161 152 151]	[146 140 140 132 130]	[401 387 371 336 325]
top5_prob	[0.00711394 0.00607287 0.00584153 0.00584153 0.00584153]	[0.0093925 0.00744922 0.00744922 0.0070328 0.00698654]	[0.00675488 0.00647728 0.00647728 0.00610715 0.00601462]	[0.01855365 0.01790589 0.01716559 0.0155462 0.01503725]
nobs	17290	**21613**	21614	**21613**
missing	17290	0	0	0

Tabla A.14: Propiedades estadisticas de variable long, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[-122.29 -122.362 -122.3 -122.372 -122.284]	[-122.3 -122.363 -122.299 -122.304 -122.169]	[-122.346 -122.317 -122.375 -122.387 -122.34]	[-122.339 -122.32 -122.338 -122.333 -122.325]
top5_freq	[100 88 81 81 81]	[111 104 93 90 82]	[19 18 17 17 17]	[101 100 97 97 95]
top5_prob	[0.00578369 0.00508965 0.00468479 0.00468479 0.00468479]	[0.0051358 0.00481192 0.00430297 0.00416416 0.00379401]	[0.00087906 0.00083279 0.00078653 0.00078653 0.00078653]	[0.00467311 0.00462684 0.00448804 0.00448804 0.0043955]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	-122.214	**-122.214**	-122.214	-122.208
std_err	0.001	0.001	0.001	**0.001**
upper_ci	-122.212	**-122.212**	-122.212	-122.206
lower_ci	-122.216	**-122.216**	-122.215	-122.211
std	0.140	0.138	**0.139**	0.153
iqr	0.204	0.200	**0.202**	0.233
iqr_normal	0.151	0.148	**0.150**	0.173
mad	0.115	0.113	**0.115**	0.128
mad_normal	0.144	0.142	**0.144**	0.161
coef_var	-0.001	-0.001	**-0.001**	-0.001
range	1.204	**1.204**	1.190	0.961
max	-121.315	**-121.315**	-121.324	-121.558
min	-122.519	**-122.519**	-122.515	**-122.519**
skew	0.867	**0.865**	0.785	0.733
kurtosis	3.953	**4.217**	3.474	3.284
jarque_bera	2819	4027	**2421**	2009
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	-122.290	**-122.300**	-122.346	-122.339
mode_freq	0.006	**0.005**	0.001	0.005
median	-122.231	-122.222	**-122.229**	-122.238
0.1%	-122.497	-122.513	**-122.489**	-122.519
1.0%	-122.408	-122.400	**-122.407**	-122.435
5.0%	-122.387	**-122.386**	-122.386	-122.400
25.0%	-122.329	-122.326	**-122.327**	-122.335
75.0%	-122.125	-122.126	**-122.124**	-122.102
95.0%	-121.979	-121.994	**-121.980**	-121.955
99.0%	-121.787	-121.814	**-121.787**	-121.752
99.9%	-121.699	-121.617	**-121.724**	-121.624

Tabla A.15: Propiedades estadisticas de variable lat, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[47.5402 47.6914 47.6853 47.6624 47.6968]	[47.1559 47.7776 47.64524616 47.64523806 47.64521392]	[47.3992 47.6844 47.7386 47.7073 47.3281]	[47.7776 47.1593 47.6108 47.5728 47.5685]
top5_freq	[14 13 13 13 13]	[18 2 1 1 1]	[5 5 4 4 4]	[333 67 16 15 15]
top5_prob	[0.00080972 0.00075188 0.00075188 0.00075188 0.00075188]	[8.32832092e-04 9.25368991e-05 4.62684495e-05 4.62684495e-05 4.62684495e-05]	[0.00023133 0.00023133 0.00018507 0.00018507 0.00018507]	[0.01540739 0.00309999 0.0007403 0.00069403 0.00069403]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	47.560	47.559	**47.561**	47.555
std_err	0.001	0.001	**0.001**	0.001
upper_ci	47.562	47.561	**47.563**	47.556
lower_ci	47.558	47.557	**47.559**	47.553
std	0.138	0.137	**0.138**	0.130
iqr	0.206	**0.205**	0.205	0.179
iqr_normal	0.153	**0.152**	0.152	0.133
mad	0.115	0.114	**0.115**	0.105
mad_normal	0.144	0.142	**0.144**	0.132
coef_var	0.003	0.003	**0.003**	0.003
range	0.618	0.622	0.590	**0.618**
max	47.778	**47.778**	47.778	**47.778**
min	47.159	47.156	47.188	**47.159**
skew	-0.487	-0.494	**-0.488**	-0.548
kurtosis	2.328	2.368	**2.311**	2.798
jarque_bera	1009	1239	1287	**1117**
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	47.540	47.156	**47.399**	47.778
mode_freq	0.001	**0.001**	0.000	0.015
median	47.572	47.569	47.573	**47.573**
0.1%	47.193	47.159	**47.194**	47.159
1.0%	47.257	47.214	**47.258**	47.223
5.0%	47.311	47.311	**47.311**	47.312
25.0%	47.472	**47.472**	47.474	47.472
75.0%	47.678	**47.677**	47.679	47.651
95.0%	47.750	47.746	**47.751**	47.741
99.0%	47.773	**47.771**	47.771	47.778
99.9%	47.777	**47.776**	47.776	47.778

Tabla A.16: Propiedades estadisticas de variable sqft_above, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1300 1010 1200 1220 1140]	[1300. 1220. 1010. 1200. 1250.]	[1150. 1800. 1280. 1560. 1610.]	[1737 1718 1889 1514 1762]
top5_freq	[166 165 160 152 148]	[207 204 198 180 170]	[14 13 12 11 11]	[25 23 23 23 23]
top5_prob	[0.00960093 0.00954309 0.0092539 0.00879121 0.00855986]	[0.00957757 0.00943876 0.00916115 0.00832832 0.00786564]	[0.00064773 0.00060146 0.0005552 0.00050893 0.00050893]	[0.00115671 0.00106417 0.00106417 0.00106417 0.00106417]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	1786	**1785**	1764	2324
std_err	6.249	**5.813**	5.225	6.736
upper_ci	1798	**1796**	1774	2338
lower_ci	1774	**1774**	1754	2311
std	821.626	**854.662**	768.170	990.211
iqr	1000.000	**964.226**	954.450	1063.000
iqr_normal	741.301	**714.782**	707.535	788.003
mad	635.012	**620.448**	600.814	734.255
mad_normal	795.870	**777.616**	753.009	920.253
coef_var	0.460	**0.479**	0.435	0.426
range	8570	**9111**	7574	7341
max	8860	**9410**	7928	7799
min	290.000	**299.358**	354.765	458.000
skew	1.428	2.600	1.282	**1.556**
kurtosis	6.260	18.428	4.876	**6.076**
jarque_bera	13530	238717	9088	**17241**
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1300	**1300**	1150	1737
mode_freq	0.010	**0.010**	0.001	0.001
median	1560	**1560**	1543	2066
0.1 %	520.000	**548.550**	592.001	717.284
1.0 %	700.000	**724.316**	736.865	942.000
5.0 %	850.000	890.000	**887.154**	1205.000
25.0 %	1200	1210	**1210**	1671
75.0 %	2200	**2174**	2164	2734
95.0 %	3380	**3296**	3277	4258
99.0 %	4371	**4321**	4173	5736
99.9 %	6070	9410	**5253**	7086

Tabla A.17: Propiedades estadisticas de variable grade, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[7 8 9 6 10]	[7 8 9 6 10]	[7 8 9 6 10]	[7 8 9 10 6]
top5_freq	[7201 4879 2072 1620 915]	[9443 6433 2503 1643 1070]	[9888 6093 2505 1749 1016]	[6800 4148 3999 2266 1769]
top5_prob	[0.41648352 0.28218623 0.11983806 0.09369578 0.05292076]	[0.43691297 0.29764494 0.11580993 0.07601906 0.04950724]	[0.45748126 0.28190062 0.1158971 0.08091977 0.04700657]	[0.31462546 0.19192153 0.18502753 0.10484431 0.08184889]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	7.654	**7.672**	7.612	7.948
std_err	0.009	**0.007**	0.007	0.012
upper_ci	7.671	**7.686**	7.626	7.971
lower_ci	7.636	**7.657**	7.598	7.925
std	1.170	**1.091**	1.048	1.720
iqr	1.000	**1.000**	**1.000**	2.000
iqr_normal	0.741	**0.741**	**0.741**	1.483
mad	0.926	**0.868**	0.848	1.325
mad_normal	1.160	**1.088**	1.062	1.661
coef_var	0.153	**0.142**	0.138	0.216
range	12.000	9.000	8.000	**12.000**
max	13.000	**13.000**	12.000	**13.000**
min	1.000	4.000	4.000	**1.000**
skew	0.758	0.906	**0.829**	0.214
kurtosis	4.209	**4.216**	3.935	3.674
jarque_bera	2709	4287	**3266**	574
jarque_bera_pval	0.000	**0.000**	**0.000**	0.000
mode	7.000	7.000	7.000	7.000
mode_freq	0.416	**0.437**	0.457	0.315
median	7.000	**7.000**	**7.000**	8.000
0.1 %	4.000	5.000	5.000	3.000
1.0 %	5.000	6.000	6.000	4.000
5.0 %	6.000	**6.000**	**6.000**	5.000
25.0 %	7.000	7.000	7.000	7.000
75.0 %	8.000	**8.000**	**8.000**	9.000
95.0 %	10.000	**10.000**	**10.000**	11.000
99.0 %	11.000	**11.000**	**11.000**	13.000
99.9 %	12.000	**12.000**	**12.000**	13.000

Tabla A.18: Propiedades estadisticas de variable bathrooms, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[2.5 1. 1.75 2.25 2.]	[2.5 1. 1.75 2.25 2.]	[2.5 1. 1.75 2.25 2.]	[1.75 1. 2.25 2.5 1.5]
top5_freq	[4333 3088 2425 1621 1526]	[6171 3786 3353 2029 1782]	[6715 4783 3314 1983 1402]	[3824 3059 2718 2675 1930]
top5_prob	[0.25060729 0.17860035 0.14025448 0.09375361 0.08825911]	[0.2855226 0.17517235 0.15513811 0.09387868 0.08245038]	[0.31067826 0.22129176 0.15332655 0.09174609 0.06486537]	[0.17693055 0.14153519 0.12575765 0.1237681 0.08929811]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	2.114	**2.099**	2.013	2.209
std_err	0.006	0.005	0.005	**0.006**
upper_ci	2.125	**2.108**	2.022	2.222
lower_ci	2.102	**2.089**	2.004	2.196
std	0.767	**0.718**	0.705	0.941
iqr	1.000	0.750	**1.000**	1.250
iqr_normal	0.741	0.556	**0.741**	0.927
mad	0.615	0.583	**0.589**	0.735
mad_normal	0.771	0.731	**0.738**	0.921
coef_var	0.363	0.342	**0.350**	0.426
range	8.000	6.250	4.750	**8.000**
max	8.000	6.750	5.500	**8.000**
min	0.000	0.500	0.750	**0.000**
skew	0.464	**0.273**	0.138	0.691
kurtosis	3.989	3.472	2.825	**3.819**
jarque_bera	1326	**469**	96	2325
jarque_bera_pval	0.000	0.000	0.000	**0.000**
mode	2.500	**2.500**	**2.500**	1.750
mode_freq	0.251	**0.286**	0.311	0.177
median	2.250	2.250	2.250	2.250
0.1 %	0.750	**1.000**	**1.000**	0.000
1.0 %	1.000	**1.000**	**1.000**	0.750
5.0 %	1.000	1.000	1.000	1.000
25.0 %	1.500	1.750	**1.500**	**1.500**
75.0 %	2.500	**2.500**	**2.500**	2.750
95.0 %	3.500	**3.250**	**3.250**	4.000
99.0 %	4.250	**4.000**	3.500	4.750
99.9 %	5.428	**5.250**	4.500	6.000

Tabla A.19: Propiedades estadisticas de variable floors, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1. 2. 1.5 3. 2.5]	[1. 2. 1.5 3. 2.5]	[1. 2. 1.5 3. 2.5]	[1. 2. 1.5 3. 2.5]
top5_freq	[8488 6628 1523 517 128]	[11160 8426 1451 530 45]	[11343 8346 1322 578 25]	[11625 6405 2021 644 600]
top5_prob	[0.49091961 0.38334297 0.0880856 0.02990168 0.00740312]	[0.5163559 0.38985796 0.06713552 0.02452228 0.00208208]	[0.52479874 0.38613861 0.06116406 0.02674193 0.00115666]	[0.53787073 0.29634942 0.09350854 0.02979688 0.02776107]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	1.499	1.476	1.472	**1.481**
std_err	0.004	0.004	0.004	**0.004**
upper_ci	1.507	1.483	1.479	**1.489**
lower_ci	1.491	1.469	1.465	**1.473**
std	0.543	0.532	**0.537**	0.602
iqr	1.000	1.000	1.000	1.000
iqr_normal	0.741	0.741	0.741	0.741
mad	0.490	**0.491**	0.495	0.518
mad_normal	0.614	**0.616**	0.621	0.649
coef_var	0.362	**0.361**	0.365	0.406
range	2.500	**2.500**	2.000	**2.500**
max	3.500	**3.500**	3.000	**3.500**
min	1.000	1.000	1.000	1.000
skew	0.615	**0.597**	0.636	1.086
kurtosis	2.526	2.382	**2.445**	3.690
jarque_bera	1252	**1627**	1733	4676
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1.000	1.000	1.000	1.000
mode_freq	0.491	**0.516**	0.525	0.538
median	1.500	1.000	1.000	1.000
0.1 %	1.000	1.000	1.000	1.000
1.0 %	1.000	1.000	1.000	1.000
5.0 %	1.000	1.000	1.000	1.000
25.0 %	1.000	1.000	1.000	1.000
75.0 %	2.000	2.000	2.000	2.000
95.0 %	2.000	**2.000**	**2.000**	2.500
99.0 %	3.000	**3.000**	**3.000**	3.500
99.9 %	3.000	**3.000**	**3.000**	3.500

Tabla A.20: Propiedades estadisticas de variable sqft_lot, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[5000 4000 6000 7200 4800]	[5000. 4000. 6000. 7200. 4500.]	[5000. 4000. 6000. 5200. 3600.]	[520 13410 8725 6954 9439]
top5_freq	[301 209 208 179 98]	[307 239 229 193 110]	[49 42 24 20 15]	[704 7 7 7 7]
top5_prob	[0.01740891 0.01208791 0.01203008 0.01035281 0.00566802]	[0.01420441 0.01105816 0.01059547 0.00892981 0.00508953]	[0.00226705 0.00194318 0.00111039 0.00092533 0.00069399]	[0.03257299 0.00032388 0.00032388 0.00032388 0.00032388]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	14799	16966	**13873**	16108
std_err	295.375	566.388	**215.499**	193.268
upper_ci	15378	18076	**14295**	16487
lower_ci	14220	15856	**13450**	15730
std	38839	83267	**31682**	28413
iqr	5606	5319	**5617**	7627
iqr_normal	4155	3943	**4164**	5654
mad	13382	17110	11638	**12248**
mad_normal	16772	21444	14586	**15350**
coef_var	2.624	4.908	**2.284**	1.764
range	1164274	1650745	**981179**	341735
max	1164794	1651359	**981832**	342255
min	520.000	614.056	653.728	**520.000**
skew	11.588	17.346	**9.945**	6.378
kurtosis	215.591	330.178	**159.385**	52.514
jarque_bera	32946220	97482676	**22381304**	2354362
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	5000	**5000**	**5000**	520
mode_freq	0.017	**0.014**	0.002	0.033
median	7600	**7700**	7751	10395
0.1 %	737.156	850.546	**820.712**	520.000
1.0 %	1005	**1061**	1084	520
5.0 %	1756	1902	**1865**	1348
25.0 %	5001	**5100**	5114	6474
75.0 %	10607	10419	**10732**	14101
95.0 %	42999	40761	**40935**	52186
99.0 %	212192	**192965**	173689	160319
99.9 %	435600	1651359	**365951**	302666

Tabla A.21: Propiedades estadisticas de variable yr_renovated, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[0 2014 2005 2000 2003]	[0. 2014. 2015. 2006.4581577 2006.45928954]	[0. 2014. 1991. 1970. 2005.]	[7 8 6 9 5]
top5_freq	[16571 76 32 30 29]	[20817 56 25 1 1]	[20736 21 3 2 2]	[2614 2441 2317 2174 2025]
top5_prob	[0.95841527 0.0043956 0.00185078 0.00173511 0.00167727]	[9.63170314e-01 2.59103317e-03 1.15671124e-03 4.62684495e-05 4.62684495e-05]	[9.59378181e-01 9.71592486e-04 1.38798927e-04 9.25326177e-05 9.25326177e-05]	[0.12094573 0.11294129 0.107204 0.10058761 0.09369361]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	83.003	73.148	**75.833**	158.302
std_err	3.031	2.553	**2.565**	3.615
upper_ci	88.943	78.152	**80.860**	165.387
lower_ci	77.063	68.144	**70.807**	151.217
std	398.503	375.347	**377.062**	531.439
iqr	0.000	**0.000**	**0.000**	5.000
iqr_normal	0.000	**0.000**	**0.000**	3.707
mad	159.103	140.919	**145.518**	281.977
mad_normal	199.407	176.615	**182.380**	353.406
coef_var	4.801	5.131	**4.972**	3.357
range	2015	**2015**	2015	**2015**
max	2015	**2015**	2015	**2015**
min	0.000	0.000	0.000	0.000
skew	4.593	4.941	**4.838**	3.195
kurtosis	22.096	25.424	**24.549**	11.211
jarque_bera	323506	540789	**502506**	97492
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	0.000	**0.000**	**0.000**	7.000
mode_freq	0.958	0.963	**0.959**	0.121
median	0.000	**0.000**	**0.000**	6.000
0.1 %	0.000	0.000	0.000	0.000
1.0 %	0.000	0.000	0.000	0.000
5.0 %	0.000	0.000	0.000	0.000
25.0 %	0.000	**0.000**	**0.000**	4.000
75.0 %	0.000	**0.000**	**0.000**	9.000
95.0 %	0.000	**0.000**	**0.000**	2015.000
99.0 %	2008	**2010**	2005	2015
99.9 %	2014	2015	**2014**	2015

Tabla A.22: Propiedades estadisticas de variable zipcode, King county (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[98103 98038 98115 98052 98117]	[98052 98115 98042 98117 98038]	[98103 98038 98117 98052 98115]	[98103 98118 98115 98034 98004]
top5_freq	[489 473 462 459 455]	[650 629 622 608 602]	[594 591 585 584 580]	[1018 704 684 610 591]
top5_prob	[0.02828224 0.02735685 0.02672065 0.02654714 0.02631579]	[0.03007449 0.02910285 0.02877898 0.02813122 0.02785361]	[0.02748219 0.02734339 0.02706579 0.02701952 0.02683446]	[0.04710128 0.03257299 0.03164762 0.02822375 0.02734465]
nobs	17290	**21613**	21614	**21613**
missing	0.000	0.000	0.000	0.000
mean	98078	**98078**	98078	98084
std_err	0.406	0.361	0.360	**0.375**
upper_ci	98079	**98079**	98078	98085
lower_ci	98077	**98077**	98077	98083
std	53.326	**53.111**	52.993	55.145
iqr	84.000	**84.000**	**84.000**	85.000
iqr_normal	62.269	**62.269**	**62.269**	63.011
mad	46.554	**46.423**	46.137	48.478
mad_normal	58.347	**58.183**	57.825	60.759
coef_var	0.001	**0.001**	0.001	0.001
range	198.000	198.000	198.000	198.000
max	98199	98199	98199	98199
min	98001	98001	98001	98001
skew	0.402	**0.405**	0.409	0.221
kurtosis	2.153	**2.157**	2.170	2.009
jarque_bera	983.027	1231.053	1224.594	**1061.029**
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	98103	98052	**98103**	**98103**
mode_freq	0.028	0.030	**0.027**	0.047
median	98065	**98065**	**98065**	98092
0.1 %	98001	98001	98001	98001
1.0 %	98001	98001	98001	98001
5.0 %	98004	98004	98004	98004
25.0 %	98033	98033	98033	98033
75.0 %	98117	**98117**	**98117**	98118
95.0 %	98177	**98177**	**98177**	98178
99.0 %	98199	98199	98199	98199
99.9 %	98199	98199	98199	98199

A.10. Estadísticos Económicos - Conjunto A

Tabla A.23: Propiedades estadisticas de variable county, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Las Condes' 'Santiago' 'Providencia' 'Vitacura' 'Lo Barnechea']	['Las Condes' 'Santiago' 'Providencia' 'Vitacura' 'Lo Barnechea']	['Las Condes' 'Santiago' 'Providencia' 'Lo Barnechea' 'Vitacura']	['Las Condes' 'Santiago' 'Viña del Mar' 'Lo Barnechea' 'Providencia']
top5_freq	[3233 2703 1481 1415 1322]	[4252 3397 1988 1869 1678]	[4630 3818 1898 1863 1859]	[4064 2917 2326 1342 1145]
top5_prob	[0.14656149 0.12253502 0.06713813 0.06414615 0.05993019]	[0.15420323 0.12319576 0.0720969 0.06778124 0.06085443]	[0.1679118 0.13846377 0.06883296 0.06756365 0.06741858]	[0.14738522 0.10578806 0.08435483 0.04866904 0.04152462]
nobs	22059	27574	27574	27574
missing	22059	0	0	0

Tabla A.24: Propiedades estadisticas de variable state, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Metropolitana de Santiago' 'Valparaíso' 'Coquimbo' 'Araucanía' "Libertador General Bernardo O'higgins"]	['Metropolitana de Santiago' 'Valparaíso' 'Coquimbo' 'Araucanía' "Libertador General Bernardo O'higgins"]	['Metropolitana de Santiago' 'Valparaíso' 'Coquimbo' 'Araucanía' "Libertador General Bernardo O'higgins"]	['Metropolitana de Santiago' 'Araucanía' 'Valparaíso' 'Coquimbo' 'Maule']
top5_freq	[17248 2014 567 558 305]	[22028 2393 704 654 350]	[22460 2395 670 626 283]	[19374 2381 2143 664 558]
top5_prob	[0.78190308 0.0913006 0.02570379 0.0252958 0.01382656]	[0.7988685 0.08678465 0.0255313 0.023718 0.01269312]	[0.81453543 0.08685718 0.02429825 0.02270255 0.01026329]	[0.70261841 0.08634946 0.07771814 0.02408066 0.02023645]
nobs	22059	27574	27574	27574
missing	22059	0	0	0

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

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[12500. 10500. 11500. 8500. 9000.]	[8500. 10500. 20000. 13500. 11500.]	[10500. 11500. 12500. 13500. 9000.]	[0. 90004.3925248 73802.71034456 71815.36214198 75642.89288791]
top5_freq	[104 99 91 86 85]	[111 102 96 94 90]	[165 162 155 138 134]	[8209 2 2 2 2]
top5_prob	[0.00471463 0.00448796 0.0041253 0.00389864 0.0038533]	[0.00402553 0.00369914 0.00348154 0.00340901 0.00326394]	[0.0059839 0.0058751 0.00562124 0.00500471 0.00485965]	[2.97707986e-01 7.25320955e-05 7.25320955e-05 7.25320955e-05 7.25320955e-05]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	110379	47385	**92582**	38521
std_err	32746	17777	**24387**	226
upper_ci	174559	82227	**140379**	38965
lower_ci	46199	12542	**44785**	38078
std	4863477	2951960	**4049530**	37555
iqr	9959	**9926**	10400	68059
iqr_normal	7383	**7358**	7710	50452
mad	202281	77057	**166646**	32586
mad_normal	253522	96576	**208859**	40840
coef_var	44.062	62.298	**43.740**	0.975
range	390000000	**350488588**	279000000	142945
max	390000000	**350488588**	279000000	142945
min	0.000	0.000	0.000	0.000
skew	60.579	94.396	**55.635**	0.590
kurtosis	4067	9748	**3308**	2
jarque_bera	1.51936e+10	1.09155e+11	**1.25655e+10**	2.44116e+03
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	12500	8500	**10500**	0
mode_freq	0.005	**0.004**	0.006	0.298
median	5084	5202	**5086**	30858
0.1 %	0.263	0.415	**0.396**	0.000
1.0 %	6.270	**7.612**	7.626	0.000
5.0 %	11.760	12.120	**12.005**	0.000
25.0 %	2041	2166	**2100**	0
75.0 %	12000	**12092**	12500	68059
95.0 %	32000	30547	**32000**	107449
99.0 %	58942	54680	**55000**	125374
99.9 %	262695	108850	**137000**	136320

Tabla A.26: Propiedades estadisticas de variable m_built, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[140. 60. 120. 50. 70.]	[140. 50. 60. 70. 120.]	[140. 50. 60. 70. 30.]	[1.00000e+00 1.26378e+03 6.84600e+01 1.27400e+03 4.24150e+02]
top5_freq	[700 467 444 431 415]	[930 552 510 491 483]	[352 164 162 149 145]	[12171 3 3 3 3]
top5_prob	[0.03173308 0.0211705 0.02012784 0.01953851 0.01881318]	[0.03372742 0.02001886 0.01849568 0.01780663 0.0175165]	[0.01276565 0.00594763 0.0058751 0.00540364 0.00525858]	[4.41394067e-01 1.08798143e-04 1.08798143e-04 1.08798143e-04 1.08798143e-04]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	1771	782	**858**	634
std_err	664.365	**327.297**	216.385	10.690
upper_ci	3073	**1423**	1282	655
lower_ci	469.205	140.235	**433.547**	613.290
std	98673	**54349**	35932	1775
iqr	140.000	132.231	**139.939**	955.520
iqr_normal	103.782	98.023	**103.737**	708.328
mad	3202	1255	**1395**	696
mad_normal	4013	1572	**1748**	872
coef_var	55.706	69.525	**41.895**	2.799
range	11999999	**8052204**	2673911	39420
max	12000000	**8052205**	2673912	39421
min	1.000	1.000	1.000	1.000
skew	96.078	**126.975**	62.012	12.288
kurtosis	10659	17908	**3993**	189
jarque_bera	1.04399e+11	3.68395e+11	**1.83102e+10**	4.04027e+07
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	140.000	**140.000**	**140.000**	1.000
mode_freq	0.032	**0.034**	0.013	0.441
median	107.000	**106.115**	108.644	171.375
0.1 %	2.000	3.536	15.092	**1.000**
1.0 %	23.000	**24.760**	25.552	1.000
5.0 %	33.000	34.000	**33.810**	1.000
25.0 %	60.000	**60.000**	60.061	1.000
75.0 %	200.000	192.231	**200.000**	956.520
95.0 %	490.000	450.000	**466.036**	1893.300
99.0 %	1200	875	**1074**	2437
99.9 %	37947	12626	21232	**29329**

Tabla A.27: Propiedades estadisticas de variable publication_date, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[1545 1693 1546 1549 721]	[1545. 1693. 1546. 1549. 721.]	[1545. 1693. 1546. 1549. 721.]	[1541 1540 1542 1539 1543]
top5_freq	[10883 6103 895 320 125]	[14237 7510 1069 280 111]	[13183 7397 492 107 86]	[1601 1564 1537 1472 1339]
top5_prob	[0.49335872 0.27666712 0.04057301 0.01450655 0.00566662]	[0.51631972 0.27235802 0.03876841 0.01015449 0.00402553]	[0.47809531 0.26825996 0.0178429 0.00388047 0.00311888]	[0.05806194 0.0567201 0.05574092 0.05338362 0.04856024]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	1471	**1472**	1468	1391
std_err	2.056	1.815	**1.853**	2.279
upper_ci	1475	**1476**	1471	1395
lower_ci	1467	**1469**	1464	1387
std	305.403	301.331	**307.767**	378.355
iqr	148.000	**148.000**	**148.000**	12.000
iqr_normal	109.713	**109.713**	**109.713**	8.896
mad	206.755	**203.168**	210.423	297.308
mad_normal	259.128	**254.634**	263.726	372.621
coef_var	0.208	0.205	**0.210**	0.272
range	1489	**1507**	1468	1303
max	1693	1693	1693	1693
min	204.000	**186.302**	224.863	390.000
skew	-2.019	**-2.024**	-1.982	-1.488
kurtosis	5.892	**5.896**	5.701	3.615
jarque_bera	22679	28453	**26433**	10613
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	1545	**1545**	**1545**	1541
mode_freq	0.493	0.516	**0.478**	0.058
median	1545	**1545**	**1545**	1542
0.1 %	450.696	510.097	**487.283**	410.000
1.0 %	531.000	536.173	**535.383**	443.000
5.0 %	628.900	**635.058**	622.694	523.000
25.0 %	1545	**1545**	**1545**	1537
75.0 %	1693	**1693**	**1693**	1549
95.0 %	1693	**1693**	**1693**	1692
99.0 %	1693	1693	1693	1693
99.9 %	1693	1693	1693	1693

Tabla A.28: Propiedades estadisticas de variable rooms, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[3. 2. 4. 1. 5.]	[3. 2. 4. 1. 5.]	[3. 2. 4. 1. 5.]	[3. 2. 4. 6. 5.]
top5_freq	[6355 4614 4168 2671 2232]	[8132 5808 5457 3244 2803]	[8221 5648 5434 3309 2991]	[9656 4931 3578 1807 1745]
top5_prob	[0.28809103 0.20916633 0.18894782 0.12108436 0.10118319]	[0.2949155 0.21063321 0.19790382 0.11764706 0.10165373]	[0.29814318 0.20483064 0.1970697 0.12000435 0.10847175]	[0.35018496 0.17882788 0.12975992 0.06553275 0.06328425]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	3.446	3.311	**3.316**	4.723
std_err	0.026	**0.013**	0.012	0.046
upper_ci	3.497	3.337	**3.340**	4.813
lower_ci	3.395	3.286	**3.292**	4.632
std	3.881	**2.168**	2.018	7.680
iqr	2.000	2.000	2.000	2.000
iqr_normal	1.483	1.483	1.483	1.483
mad	1.454	1.280	**1.284**	2.768
mad_normal	1.822	1.604	**1.609**	3.469
coef_var	1.126	**0.655**	0.608	1.626
range	399.000	182.000	49.000	**399.000**
max	400.000	183.000	50.000	**400.000**
min	1.000	1.000	1.000	1.000
skew	57.785	23.115	5.213	**30.247**
kurtosis	5331	**1737**	71	1386
jarque_bera	2.61061e+10	**3.45846e+09**	5.44919e+06	2.20157e+09
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	3.000	3.000	3.000	3.000
mode_freq	0.288	**0.295**	0.298	0.350
median	3.000	3.000	3.000	3.000
0.1 %	1.000	1.000	1.000	1.000
1.0 %	1.000	1.000	1.000	1.000
5.0 %	1.000	1.000	1.000	1.000
25.0 %	2.000	**2.000**	**2.000**	3.000
75.0 %	4.000	**4.000**	**4.000**	5.000
95.0 %	6.000	**6.000**	**6.000**	13.000
99.0 %	12.000	**10.000**	**10.000**	24.000
99.9 %	25.000	21.000	**25.000**	57.000

Tabla A.29: Propiedades estadisticas de variable property_type, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Departamento' 'Casa' 'Oficina o Casa Oficina' 'Parcela o Chacra' 'Local o Casa comercial']	['Departamento' 'Casa' 'Oficina o Casa Oficina' 'Parcela o Chacra' 'Local o Casa comercial']	['Departamento' 'Casa' 'Oficina o Casa Oficina' 'Parcela o Chacra' 'Departamento Amoblado']	['Casa' 'Departamento' 'Oficina o Casa Oficina' 'Parcela o Chacra' 'Local o Casa comercial']
top5_freq	[10592 8911 1553 413 255]	[13493 11433 1820 402 187]	[13619 11733 1781 237 110]	[17141 4844 2358 2003 580]
top5_prob	[0.48016683 0.4039621 0.0704021 0.01872252 0.01155991]	[0.48933778 0.41462972 0.06600421 0.01457895 0.00678175]	[0.4939073 0.42550954 0.06458983 0.00859505 0.00398927]	[0.62163632 0.17567274 0.08551534 0.07264089 0.02103431]
nobs	22059	27574	27574	27574
missing	22059	0	0	0

Tabla A.30: Propiedades estadisticas de variable transaction_type, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	['Venta' 'Arriendo' 'Busco arriendo' 'Compro']	['Venta' 'Arriendo' 'Busco arriendo' 'Compro']	['Venta' 'Arriendo']	['Venta' 'Arriendo' 'Compro' 'Busco arriendo']
top5_freq	[17540 4517 1 1]	[22110 5460 3 1]	[22067 5507]	[19148 6575 1831 20]
top5_prob	[7.95140306e-01 2.04769029e-01 4.53329707e-05 4.53329707e-05]	[8.01842315e-01 1.98012621e-01 1.08798143e-04 3.62660477e-05]	[0.80028288 0.19971712]	[0.69442228 0.23844926 0.06640313 0.00072532]
nobs	22059	27574	27574	27574
missing	22059	0	0	0

Tabla A.31: Propiedades estadisticas de variable m_size, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[5000. 50. 60. 200. 70.]	[5000. 50. 200. 70. 60.]	[5000. 30. 45. 60. 50.]	[0. 211628.46 159330.27 220573.96 198682.57]
top5_freq	[601 342 321 285 281]	[767 425 366 330 317]	[167 76 72 72 67]	[3068 2 2 2 2]
top5_prob	[0.02724512 0.01550388 0.01455188 0.0129199 0.01273856]	[0.02781606 0.01541307 0.01327337 0.0119678 0.01149634]	[0.00605643 0.00275622 0.00261116 0.00261116 0.00242983]	[1.11264234e-01 7.25320955e-05 7.25320955e-05 7.25320955e-05 7.25320955e-05]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	146269	818272	163524	**161700**
std_err	105454	817548	**87196**	550
upper_ci	352956	2420637	**334425**	162777
lower_ci	-60417	-784093	**-7378**	160622
std	15662334	135757324	**14479306**	91298
iqr	340.500	**322.785**	368.787	134708.178
iqr_normal	252.413	**239.281**	273.382	99859.321
mad	290635	1635037	**325380**	75918
mad_normal	364257	2049215	**407804**	95149
coef_var	107.079	165.907	**88.546**	0.565
range	2.24100e+09	2.25431e+10	**1.80341e+09**	3.75531e+05
max	2.24100e+09	2.25431e+10	**1.80341e+09**	3.75531e+05
min	0.000	0.015	1.000	**0.000**
skew	134.762	166.045	**108.046**	-0.352
kurtosis	19053	27572	**12365**	2
jarque_bera	3.33616e+11	8.73361e+11	**1.75623e+11**	1.38898e+03
jarque_bera_pval	0.000	**0.000**	**0.000**	0.000
mode	5000	**5000**	**5000**	0
mode_freq	0.027	**0.028**	0.006	0.111
median	145.000	**143.698**	149.347	175767.490
0.1 %	2.000	**2.796**	14.702	0.000
1.0 %	22.000	**23.716**	26.000	0.000
5.0 %	35.000	**35.000**	35.010	0.000
25.0 %	66.000	**67.000**	67.283	98010.742
75.0 %	406.500	**389.785**	436.070	232718.920
95.0 %	5000	**5000**	4257	293055
99.0 %	10200	7662	**8430**	323719
99.9 %	70000	33073	**43532**	350745

Tabla A.32: Propiedades estadisticas de variable bathrooms, Economicos (A-3)

Variable/Modelo	Real	tddpm_mlp	smote-enc	ctgan
top5	[2. 1. 3. 4. 5.]	[2. 1. 3. 4. 5.]	[2. 1. 3. 4. 5.]	[2. 1. 3. 5. 4.]
top5_freq	[7511 5440 4486 2665 1084]	[9567 6748 5742 3392 1316]	[9394 6785 5574 3366 1328]	[10737 6090 3978 2454 2009]
top5_prob	[0.34049594 0.24661136 0.20336371 0.12081237 0.04914094]	[0.34695728 0.24472329 0.20823965 0.12301443 0.04772612]	[0.34068325 0.24606513 0.20214695 0.12207152 0.04816131]	[0.38938855 0.22086023 0.14426634 0.08899688 0.07285849]
nobs	22059	27574	27574	27574
missing	0.000	0.000	0.000	0.000
mean	2.604	2.511	**2.587**	3.053
std_err	0.025	0.011	0.012	**0.035**
upper_ci	2.652	2.533	**2.610**	3.121
lower_ci	2.556	2.490	**2.563**	2.985
std	3.655	1.814	**2.023**	5.750
iqr	1.000	1.000	1.000	1.000
iqr_normal	0.741	0.741	0.741	0.741
mad	1.203	1.094	**1.180**	1.742
mad_normal	1.507	1.372	**1.479**	2.183
coef_var	1.404	0.723	0.782	**1.884**
range	435.000	179.000	179.000	**435.000**
max	436.000	180.000	180.000	**436.000**
min	1.000	1.000	1.000	1.000
skew	82.448	35.917	27.561	**51.076**
kurtosis	9252	3349	2189	**3614**
jarque_bera	7.86518e+10	1.28702e+10	5.49124e+09	**1.49972e+10**
jarque_bera_pval	0.000	0.000	0.000	0.000
mode	2.000	2.000	2.000	2.000
mode_freq	0.340	0.347	**0.341**	0.389
median	2.000	2.000	2.000	2.000
0.1 %	1.000	1.000	1.000	1.000
1.0 %	1.000	1.000	1.000	1.000
5.0 %	1.000	1.000	1.000	1.000
25.0 %	2.000	2.000	2.000	2.000
75.0 %	3.000	3.000	3.000	3.000
95.0 %	5.000	**5.000**	**5.000**	7.000
99.0 %	8.000	7.000	**8.000**	14.000
99.9 %	17.942	12.000	**18.000**	30.000

A.11. Estadísticos Económicos - Conjunto B

Tabla A.33: Propiedades estadisticas de variable state, Economicos (B-3)

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

Tabla A.34: Propiedades estadisticas de variable publication_date, Economicos (B-3)

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

Tabla A.35: Propiedades estadisticas de variable property_type, Economicos (B-3)

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

Tabla A.36: Propiedades estadisticas de variable transaction_type, Economicos (B-3)

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

Tabla A.37: Propiedades estadisticas de variable bathrooms, Economicos (B-3)

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

Tabla A.38: Propiedades estadisticas de variable rooms, Economicos (B-3)

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

Tabla A.39: Propiedades estadisticas de variable _price, Economicos (B-3)

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

Tabla A.40: Propiedades estadisticas de variable m_size, Economicos (B-3)

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

Tabla A.41: Propiedades estadisticas de variable m_built, Economicos (B-3)

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

Tabla A.42: Propiedades estadisticas de variable county, Economicos (B-3)

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

A.12. Ejemplos de código con fines de reproducibilidad

En el Código 5, se muestra cómo se calcula y se muestra el puntaje 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.

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

Código 5: Mostrando Puntajes Promedios Calculados

En el ejemplo presentado en el Código 6, se crea una instancia de la clase *Synthetic* utilizando 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 almacenarán los archivos temporales, se seleccionan los modelos a utilizar, se establece el número de registros sintéticos deseados y se define una columna objetiva para realizar pruebas con aprendizaje automático y estratificar los conjuntos parciales de datos que se utilizarán. De esta manera, se configura de forma flexible el proceso de generación de datos sintéticos según las necesidades específicas del usuario.

```
    syn = Synthetic(df_converted,
            id="url",
            category_columns=category_columns,
            text_columns=("description", "price", "title", "address", "owner",),
            exclude_columns=tuple(),
            synthetic_folder = "../datasets/economicos/synth",
            models=['copulagan', 'tvae', 'gaussiancopula', 'ctgan', 'smote-enc'],
            n_sample = df.shape[0],
            target_column="_price"
    )
```

Código 6: Instanciando clase Synthetic