



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

DATA SINTÉTICA PRIVADA, EJECUCIÓN Y EVALUACIONES DE MODELOS

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

GERARDO JORGE VILLARROEL GONZÁLEZ

PROFESOR GUÍA:
ANDRES ABELIUK

MIEMBROS DE LA COMISIÓN:

—
—
—

SANTIAGO DE CHILE
2023

Resumen

Nulla ac nisl. Nullam urna nulla, ullamcorper in, interdum sit amet, gravida ut, risus. Aenean ac enim. In luctus. Phasellus eu quam vitae turpis viverra pellentesque. Duis feugiat felis ut enim. Phasellus pharetra, sem id porttitor sodales, magna nunc aliquet nibh, nec blandit nisl mauris at pede. Suspendisse risus risus, lobortis eget, semper at, imperdiet sit amet, quam. Quisque scelerisque dapibus nibh. Nam enim. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc ut metus. Ut metus justo, auctor at, ultrices eu, sagittis ut, purus. Aliquam aliquam.

Etiam pede massa, dapibus vitae, rhoncus in, placerat posuere, odio. Vestibulum luctus commodo lacus. Morbi lacus dui, tempor sed, euismod eget, condimentum at, tortor. Phasellus aliquet odio ac lacus tempor faucibus. Praesent sed sem. Praesent iaculis. Cras rhoncus tellus sed justo ullamcorper sagittis. Donec quis orci. Sed ut tortor quis tellus euismod tincidunt. Suspendisse congue nisl eu elit. Aliquam tortor diam, tempus id, tristique eget, sodales vel, nulla. Praesent tellus mi, condimentum sed, viverra at, consectetur quis, lectus. In auctor vehicula orci. Sed pede sapien, euismod in, suscipit in, pharetra placerat, metus. Vivamus commodo dui non odio. Donec et felis.

Etiam suscipit aliquam arcu. Aliquam sit amet est ac purus bibendum congue. Sed in eros. Morbi non orci. Pellentesque mattis lacinia elit. Fusce molestie velit in ligula. Nullam et orci vitae nibh vulputate auctor. Aliquam eget purus. Nulla auctor wisi sed ipsum. Morbi porttitor tellus ac enim. Fusce ornare. Proin ipsum enim, tincidunt in, ornare venenatis, molestie a, augue. Donec vel pede in lacus sagittis porta. Sed hendrerit ipsum quis nisl. Suspendisse quis massa ac nibh pretium cursus. Sed sodales. Nam eu neque quis pede dignissim ornare. Maecenas eu purus ac urna tincidunt congue.

*A todos los lectores no organicos, que para cuando interiorisen estas palabras espero hayamos
aprendido a ser buenos padres*

Agradecimientos

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Tabla de Contenido

1. Introducción	1
1.1. Estructura del documento	1
1.2. Equifax: contexto y limitaciones	3
1.3. Contexto Temporal/tecnológico	5
1.4. Objetivo	6
2. Revisión Bibliográfica	7
2.1. Tipos de Datos	7
2.2. Privacidad de Datos	8
2.2.1. Tipo de datos a ser protegidos	8
2.2.2. Tipos de riesgos de divulgación	9
2.2.3. Regulación de datos sintéticos	10
2.2.4. Protección de Privacidad	10
2.3. Generación de Datos Sintéticos	10
2.3.1. Generación de datos tabulares	11
2.3.2. Generación de texto en base de datos tabulares	12
3. Desarrollo	15
3.1. Recursos disponibles	15
3.1.1. Datasets	15
3.1.2. Computación y Software	15
3.2. Arquitectura de software	16

3.3. Dependencias	16
3.4. Modelos	16
3.4.1. Datos tabulares	16
3.4.2. Datos de texto	16
4. Resultados	17
4.1. Datos tabulares	17
4.1.1. Metricas	17
4.1.2. Ejemplos	17
4.2. Datos de texto	17
4.2.1. Metricas	17
4.2.2. Ejemplos	17
5. Conclusiones	19
6. Discusión	21
Bibliografía	23
Apéndice A. Anexo	24

Índice de Tablas

2.1. Tipos de datos estructurados	7
2.2. Niveles de revelación y ejemplos	8
2.3. Tipos de Riesgos de Divulgación y sus Descripciones	9
2.4. Estado del arte en generación de datos tabulares	11
2.5. Estado del arte en generación de textos en base a datos	12
2.6. Ejemplo de tabla de entrada	13
3.1. Computador Usado	15
5.1. Tabla 1	20

Índice de Ilustraciones

5.1. Logo de la Facultad	19
------------------------------------	----

Capítulo 1

Introducción

Cuando se revise esta tesis, estará desactualizada. Desde AlexNet [19] en 2012, el liderazgo en el problema de clasificación de imágenes ha cambiado al menos 15 veces [4]. En el campo de texto a imágenes, modelos como DALL-E 2 [1], Google Imagen [2] y Stable Diffusion [5] fueron presentados en 2022, mientras que para el 2023 se pronostica el inicio de una carrera de inteligencia artificial en el campo de los chatbots entre Google y Microsoft [20, 3]. En definitiva, es un campo actualmente en crecimiento y que seguirá sorprendiendo con nuevas técnicas y productos, en variedad y calidad.

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

Según el libro *Practical synthetic data generation: balancing privacy and the broad availability of data* [14] los datos sintéticos ofrecen dos beneficios principales:

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

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

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

1.1. Estructura del documento

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

partes originales utilizando deep learning.

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

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

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

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

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

1.2. Equifax: contexto y limitaciones

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

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

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

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

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

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

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

1.3. Contexto Temporal/tecnológico

Usando ChatGPT en el febrero 2023 [21].

Prompt

Introducción a la relevancia de la generación de datos sintéticos.

ChatGPT

La generación de datos sintéticos se ha vuelto fundamental en diversos campos de investigación y aplicación, como el aprendizaje automático, la inteligencia artificial y la simulación. Debido a la creciente necesidad de entrenar modelos de aprendizaje automático con conjuntos de datos grandes y de alta calidad, se ha visto la generación de datos sintéticos como una forma de mejorar la calidad de los modelos y ampliar los conjuntos de datos disponibles.

La generación de datos sintéticos también es de gran utilidad en situaciones donde la recopilación de datos reales es costosa o impracticable, como en la simulación de escenarios extremos o peligrosos. En estos casos, la generación de datos sintéticos puede proporcionar una alternativa segura y controlada para la experimentación.

Además, la generación de datos sintéticos puede ayudar a proteger la privacidad y la confidencialidad de los datos personales. Los datos reales pueden ser sensibles o privados y no se pueden compartir fácilmente. Al generar datos sintéticos, se pueden mantener características importantes de los datos originales mientras se preserva la privacidad y la confidencialidad.

En resumen, la generación de datos sintéticos es una técnica valiosa y cada vez más popular que puede mejorar la calidad de los modelos de aprendizaje automático, proporcionar alternativas seguras y controladas para la experimentación, y proteger la privacidad y la confidencialidad de los datos personales.

1.4. Objetivo

Objetivo General:

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

Objetivos Específicos:

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

Capítulo 2

Revisión Bibliográfica

2.1. Tipos de Datos

Los tipos de datos tienen varias implicancias en su generación, como su representación, almacenamiento y procesamiento. Los datos estructurados están presentados en Tabla 2.1. En el trabajo de

En el 2012 IDC estimó que para el 2020 más del 95 % sería data no estructurada [15]. En seguimiento en el análisis de Kiran Adnan y Rehan Akbar [8] el texto es el tipo de dato no estructurado con mayor crecimiento en las publicaciones, seguidos en orden por imagen, video y finalmente el audio.

La Tabla 2.1 Tipos de datos estructurados se resume la lista que figura 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
	Categorico	Datos que pueden tomar solo un conjunto específico de valores que representan un conjunto de categorías posibles.	-
	Binario	Un caso especial de datos categóricos con solo dos categorías de valores	0/1, verdadero/falso
	Ordinal	Datos categóricos que tienen un ordenamiento explícito.	pequeña/ mediana/ grande

2.2. Privacidad de Datos

La privacidad de datos es un tema crítico en la generación de datos sintéticos. Si los datos pertenecen a recetas o automóviles, entonces la privacidad no es relevante. Sin embargo, la síntesis de datos sobre individuos lo es [11]. Es por ello que para Equifax es un tema relevante. Muchos de los conjuntos de datos tienen información personal.

2.2.1. Tipo de datos a ser protegidos

Para determinar qué campos de datos son relevantes en términos de privacidad, se puede utilizar la definición resumida en la Tabla 2.2 de *Data privacy: Definitions and techniques* [13].

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 de *Practical synthetic data generation* [11] se resumen 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 datos compartidos, utilizando técnicas de filtrado para llegar a una única opción.
Divulgación de nueva información	Este riesgo asume el riesgo de Divulgación de Identidad y, además, implica una ganancia de información adicional sobre el individuo a partir de los datos compartidos.
Divulgación de Atributos	Este riesgo se presenta cuando, a pesar de no poder identificar a un individuo, se puede identificar un atributo común en varios registros, lo que permite obtener información sensible sobre un grupo de individuos.
Divulgación Inferencial	Este riesgo se refiere a la posibilidad de inferir información sensible a partir de los datos compartidos, mediante la utilización de técnicas de análisis estadístico o de aprendizaje automático. Por ejemplo si filtrando todos los registros. 80 % de los registros con las mismas características tienen cancer, se podría inferir que el individuo buscado puede tener cancer.

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

1. En términos prácticos, normalmente la data sintética busca tener cierta permeabilidad con respecto a la **Divulgación Inferencial**, ya que queremos que estadísticamente sea similar. adicionalmente se busca proteger la identidad de los individuos, pero no es la unica condición, tambien se busca proteger esos atributos que pueden ser sensibles, por ejemplo enfermedades. A todo este conjunto se le denomina **Revelación de indentidad significativa**. Particularmente riesgoso por discriminación en ciertos grupos que cumplan los atributos criterios.
2. Los mismos atributos pueden tener más relevancia para ciertos grupos de la población que para otro. El ejemplo que indica [14] es que debido a que el numero de hijo igual a 2, es menos frecuente en una etnia que otra, una 40% y la segunda un 10%, ese dato es más relevante en la segunda. Ya que es un factor que filtra de mejor manera y por ello puede conocerse mejor ese grupo específico. Esto lo denomina **Definición de información ganada**

2.2.3. Regulación de datos sintéticos

Debido a que los datos sintéticos son basados en datos reales, pueden ser afectos a las regulaciones de sobre protección de datos [11]. Los nuevos datos podrían ser afectos por:

1. Regulation (EU) 2016/679 of the European Parliament and of the Council [24]
2. The California consumer privacy act: Towards a European-style privacy regime in the United States [22]
3. Health insurance portability and accountability act of 1996 [7]

2.2.4. Protección de Privacidad

TBD

2.3. Generación de Datos Sintéticos

Los datos sintéticos no son datos reales, pero se intenta que conserven algunas propiedades de los datos reales. El grado en que los datos sintéticos pueden servir como proxy de los datos reales es la medida de su utilidad [11]. Dependiendo del uso de los datos reales, se pueden separar en tres tipos: los que utilizan datos reales, los que no lo hacen y los híbridos.

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

Datos no basados en datos reales: utilizan el conocimiento del mundo. Por ejemplo, un conjunto de nombres al azar con un apellido al azar para formar un nombre completo.

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

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

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

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

2.3.1. Generación de datos tabulares

En la Tabla 2.4 Estado del arte en generación de datos tabulares, se resumen las últimas publicaciones sobre generación de datos tabulares, indicando la fecha de publicación y si se puede acceder al código fuente o no, a febrero de 2023.

Tabla 2.4: Estado del arte en generación de datos tabulares

Nombre	Fecha ↓	Código
REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers [25]	2023-02-04	Github
PreFair: Privately Generating Justifiably Fair Synthetic Data [23]	2022-12-20	
GenSyn: A Multi-stage Framework for Generating Synthetic Microdata using Macro Data Sources [6]	2022-12-08	Github
TabDDPM: Modelling Tabular Data with Diffusion Models [18]	2022-10-30	Github
Language models are realistic tabular data generators [10]	2022-10-12	Github
Ctab-gan+: Enhancing tabular data synthesis [28]	2022-04-01	Github
Ctab-gan: Effective table data synthesizing [27]	2021-05-31	Github
Modeling Tabular data using Conditional GAN [26]	2019-10-28	Github
SMOTE: synthetic minority over-sampling technique [12]	2002-06-02	Github

2.3.2. Generación de texto en base de datos tabulares

En la Tabla 2.5 Estado del arte en generación de textos en base a datos, se listan las publicaciones en la generación de texto a partir de datos estructurados.

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

Nombre	Fecha ↓	Modelo Base
Table-To-Text generation and pre-training with TABT5 [9]	2022-10-17	T5
Text-to-text pre-training for data-to-text tasks [17]	2021-07-09	T5
TaPas: Weakly supervised table parsing via pre-training [16]	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.

En los siguientes ejemplos, se utilizará la Tabla 2.6 Ejemplo de tabla de entrada para ilustrar cómo se puede utilizar para generar texto utilizando los modelos de *fine-tuning* mencionados anteriormente. Esta tabla representa información sobre películas, incluyendo el nombre de la película, el director, el año de lanzamiento y el género, y se utilizará para generar preguntas y respuestas a partir de la información proporcionada.

Tabla 2.6: 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 [9, 16]. 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* 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

Capítulo 3

Desarrollo

3.1. Recursos disponibles

3.1.1. Datasets

3.1.2. Computación y Software

Para llevar a cabo los experimentos, se utilizó un computador con las siguientes especificaciones técnicas, como se muestra en la Tabla 3.1 Computador Usado. El procesador empleado fue un AMD Ryzen 9 7950X 16-Core Processor, con cuatro módulos de 32 GB para una memoria total de 128 GB DDR5. La tarjeta gráfica empleada fue una NVIDIA GeForce RTX 4090, y se contó con dos discos duros de 500 GB SSD. La utilización de un equipo con estas características permitió una ejecución eficiente de los modelos de generación de datos, asegurando la viabilidad de los experimentos. Es importante destacar que la elección de los componentes del computador fue cuidadosamente considerada para asegurar que los resultados obtenidos no se vieran limitados por un hardware insuficiente.

En cuanto al software utilizado, se utilizó el sistema operativo Ubuntu 22.04.2 LTS. Los modelos de generación de datos fueron desarrollados utilizando el lenguaje de programación Python, haciendo uso principalmente de las bibliotecas SDV y PyTorch. Se detalla una lista en ?? ?. Cabe resaltar que se optó por el uso de estas herramientas debido a TabDDPM.

Tabla 3.1: 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

El código fuente de los modelos de generación de datos, así como los scripts de análisis y

visualización de los resultados, se encuentra disponible en un repositorio público de Github.

3.2. Arquitectura de software

3.3. Dependencias

3.4. Modelos

3.4.1. Datos tabulares

3.4.2. Datos de texto

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Definición 3.1 (ver [?]) *Definición definitiva*

$$\frac{d}{dx} \int_a^x f(y) dy = f(x).$$

Capítulo 4

Resultados

4.1. Datos tabulares

4.1.1. Metricas

4.1.2. Ejemplos

4.2. Datos de texto

4.2.1. Metricas

4.2.2. Ejemplos

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Definición 4.1 (ver [?]) *Definición definitiva*

$$\frac{d}{dx} \int_a^x f(y) dy = f(x).$$

Capítulo 5

Conclusiones

Mauris ac ipsum. Duis ultrices erat ac felis. Donec dignissim luctus orci. Fusce pede odio, feugiat sit amet, aliquam eu, viverra eleifend, ipsum. Fusce arcu massa, posuere id, nonummy eu, pulvinar ut, wisi. Sed dui. Vestibulum nunc nisl, rutrum quis, pharetra eget, congue sed, dui. Donec justo neque, euismod eget, nonummy adipiscing, iaculis eu, leo. Duis lectus. Morbi pellentesque nonummy dui.

Aenean sem dolor, fermentum nec, gravida hendrerit, mattis eget, felis. Nullam non diam vitae mi lacinia consectetur. Fusce non massa eget quam luctus posuere. Aenean vulputate velit. Quisque et dolor. Donec ipsum tortor, rutrum quis, mollis eu, mollis a, pede. Donec nulla. Duis molestie. Duis lobortis commodo purus. Pellentesque vel quam. Ut congue congue risus. Sed ligula. Aenean dictum pede vitae felis. Donec sit amet nibh. Maecenas eu orci. Quisque gravida quam sed massa.

Nunc euismod, mauris luctus adipiscing pellentesque, augue ligula pellentesque lectus, vitae posuere purus velit a pede. Phasellus leo mi, egestas imperdiet, blandit non, sollicitudin pharetra, enim. Nullam faucibus tellus non enim. Sed egestas nunc eu eros. Nunc euismod venenatis urna. Phasellus ullamcorper. Vivamus varius est ac lorem. In id pede eleifend nibh consectetur faucibus. Phasellus accumsan euismod elit. Etiam vitae elit. Integer imperdiet nibh. Morbi imperdiet orci euismod mi.

Donec tincidunt tempor metus. Aenean egestas cursus nulla. Fusce ac metus at enim viverra lacinia. Vestibulum in magna non eros varius suscipit. Nullam cursus nibh. Mauris neque. In nunc quam, convallis vitae, posuere in, consequat sed, wisi. Phasellus bibendum consectetur massa. Curabitur quis urna. Pellentesque a justo.

In sit amet dui eget lacus rutrum accumsan. Phasellus ac metus sed massa varius auctor. Cu-



Figura 5.1: Logo de la Facultad

Tabla 5.1: Tabla 1

Campo 1	Campo 2	Num
Valor 1a	Valor 2a	3
Valor 1b	Valor 2b	3

rabitur velit elit, pellentesque eget, molestie nec, congue at, pede. Maecenas quis tellus non lorem vulputate ornare. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Etiam magna arcu, vulputate egestas, aliquet ut, facilisis ut, nisl. Donec vulputate wisi ac dolor. Aliquam feugiat nibh id tellus. Morbi eget massa sit amet purus accumsan dictum. Aenean a lorem. Fusce semper porta sapien.

Curabitur sit amet libero eget enim eleifend lacinia. Vivamus sagittis volutpat dui. Suspendisse potenti. Morbi a nibh eu augue fermentum posuere. Curabitur elit augue, porta quis, congue aliquam, rutrum non, massa. Integer mattis mollis ipsum. Sed tellus enim, mattis id, feugiat sed, eleifend in, elit. Phasellus non purus sed elit viverra rhoncus. Vestibulum id tellus vel sem imperdiet congue. Aenean in arcu. Nullam urna justo, imperdiet eget, volutpat vitae, semper eu, quam. Sed turpis dui, porttitor ut, egestas ac, condimentum non, wisi. Fusce iaculis turpis eget dui. Quisque pulvinar est pellentesque leo. Ut nulla elit, mattis vel, scelerisque vel, blandit ut, justo. Nulla feugiat risus in erat.

Capítulo 6

Discusión

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Definición 6.1 (ver [?]) *Definición definitiva*

$$\frac{d}{dx} \int_a^x f(y) dy = f(x).$$

Bibliografía

- [1] DALL·e 2.
- [2] Imagen: Text-to-image diffusion models.
- [3] 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.
- [4] Papers with code - ImageNet benchmark (image classification).
- [5] Stable diffusion public release.
- [6] Angeela Acharya, Siddhartha Sikdar, Sanmay Das, and Huzefa Rangwala. GenSyn: A multi-stage framework for generating synthetic microdata using macro data sources.
- [7] Accountability Act. Health insurance portability and accountability act of 1996. 104:191.
- [8] Kiran Adnan and Rehan Akbar. An analytical study of information extraction from unstructured and multidimensional big data. 6:1–38. Publisher: Springer.
- [9] Ewa Andrejczuk, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, and Yase-min Altun. Table-to-text generation and pre-training with TabT5.
- [10] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators.
- [11] Peter Bruce, Andrew Bruce, and Peter Gedeck. *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O’Reilly Media.
- [12] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. 16:321–357.
- [13] Sabrina De Capitani Di Vimercati, Sara Foresti, Giovanni Livraga, and Pierangela Samarati. Data privacy: Definitions and techniques. 20(6):793–817. Publisher: World Scientific.
- [14] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O’Reilly Media.
- [15] John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. 2007(2012):1–16.

- [16] Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. TaPas: Weakly supervised table parsing via pre-training.
- [17] Mihir Kale and Abhinav Rastogi. Text-to-text pre-training for data-to-text tasks.
- [18] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models.
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- [20] Dan Milmo and Dan Milmo Global technology editor. Google v microsoft: who will win the AI chatbot race?
- [21] OpenAI. ChatGPT: a large language model trained by OpenAI.
- [22] Stuart L Pardau. The california consumer privacy act: Towards a european-style privacy regime in the united states. 23:68. Publisher: HeinOnline.
- [23] David Pujol, Amir Gilad, and Ashwin Machanavajjhala. PreFair: Privately generating justifiably fair synthetic data.
- [24] Protection Regulation. Regulation (EU) 2016/679 of the european parliament and of the council. 679:2016.
- [25] Aivin V Solatorio and Olivier Dupriez. REaLTabFormer: Generating realistic relational and tabular data using transformers.
- [26] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. 32.
- [27] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. CTAB-GAN: Effective table data synthesizing. In Vineeth N. Balasubramanian and Ivor Tsang, editors, *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pages 97–112. PMLR.
- [28] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. CTAB-GAN+: Enhancing tabular data synthesis.

Apéndice A

Anexo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Definición A.1 (ver [?]) *Definición definitiva*

$$\frac{d}{dx} \int_a^x f(y) dy = f(x).$$