UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DISEÑO DE UN MÓDULO DE GENERACIÓN DE DATOS SINTÉTICOS PARA LA APLICACIÓN DE MODELOS DE MACHINE LEARNING EN PROYECTOS INTERDISCIPLINARIOS ASOCIADOS A SALUD

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

FERNANDO MARSHALL BOEHMWALD

PROFESOR GUÍA:
JUAN DOMINGO VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
ROCÍO BELÉN RUIZ MORENO
VÍCTOR ALEJANDRO HERNÁNDEZ MARTÍNEZ

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: **FERNANDO MARSHALL BOEHMWALD**
FECHA: 2022
PROF. GUÍA: JUAN DOMINGO VELÁSQUEZ SILVA

Diseño de un módulo de generación de datos sintéticos para la aplicación de modelos de machine learning en proyectos interdisciplinarios asociados a salud

Generar conocimiento en base a la evidencia en investigaciones clínicas es muchas veces un proceso lento, costoso y complejo. Dentro de los problemas a los que se enfrentan investigadoras e investigadores del rubro, está el bajo número de participantes en los experimentos, dada la dificultad de encontrar pacientes y el alto costo monetario y temporal de generar un nuevo registro. Esta escasez de información complejiza el trabajo estadístico, evitando la generalización de los resultados que se obtienen, dificultando la obtención de conclusiones aceptables que puedan ser aplicadas a la población. Actualmente es posible solucionar parcialmente la escasez de información utilizando datos de libre acceso, modelos más sencillos o aplicar distintas transformaciones a las fuentes de información. Sin embargo, ninguna de estas soluciones les permite a los investigadores e investigadoras utilizar todo el potencial de los datos que manejan.

Con el fin de entregar recomendaciones para resolver los problemas asociados a la escasez de datos en proyectos de aprendizaje de máquinas asociados a salud, en el presente trabajo de título se realizó un estudio de los algoritmos generadores de datos sintéticos más utilizados en la literatura para datos tabulares, basándose en los registros del proyecto Alzheimer Depression Diagnostic with Artificial Intelligence del Web Intelligence Centre. Se aplicaron tres algoritmos generativos en esta oportunidad, Generative Adversarial Networks, Variational Autoencoders y Gaussian Copula, siendo los dos primeros algoritmos de redes neuronales y el tercero un algoritmo estadístico.

Ningún algoritmo obtuvo mejores resultados al entrenar un modelo de clasificación en comparación con los datos reales, sin embargo, los mejores resultados provienen del algoritmo Gaussian Copula, presentando una diferencia de -9% y -5% para las métricas Recall y ROC AUC respectivamente al sólo utilizar datos sintéticos para el entrenamiento y otra de -18,5% y -13,5% en Recall y ROC AUC al unir los datos sintéticos y reales, todos estos resultados fueron obtenidos testeando dichos modelos con la información real de los pacientes. No fue posible probar distintos tipos de bases de datos, ya que todas poseían las mismas características; una variable binaria y varias variables numéricas. No obstante, los algoritmos que utilizan redes neuronales presentaron mejores resultados cuando las bases tenían una mayor cantidad de variables.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

Capítulo 1

Introducción

1.1 Antecedentes generales

1.1.1 Características de la organización

El presente trabajo de título se desarrollará bajo la tutela del Web Intelligence Centre, conocido por sus siglas como WIC, el cual es un centro de investigación perteneciente al Departamento de Ingeniería Civil Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. Este centro fue fundado en 2008 y busca entregar un enfoque en la investigación aplicada a la resolución de problemas que requieran utilización de tecnologías de la información y comunicaciones, en conjunto con la ciencia de datos e inteligencia artificial.

Gran parte de los proyectos que se realizan dentro del WIC son relacionados con al área de la salud, utilizando sus capacidades en ámbitos como Data Analytics, Business Intelligence, Data Science, Artificial Intelligence y Data Architecture para resolver problemas dentro del sector de la salud en Chile. La misión y visión del centro apuntan a utilizar estas poderosas herramientas tecnológicas y ponerlas a disposición de las personas que lo requieran, se detallan ambas a continuación:

Misión: Poner a disposición de la sociedad soluciones basadas en TICs, Ciencia de Datos e Inteligencia Artificial para entregar respuestas a problemas de la vida real.

Visión: Ser un centro de referencia en investigación, desarrollo y transferencia de conocimiento en soluciones basadas en TICs, DS e IA para Chile y el mundo

La estructura organizacional del centro se observa en la figura 1.1, este organigrama muestra las áreas funcionales de la organización en conjunto con los cargos responsables de estas.

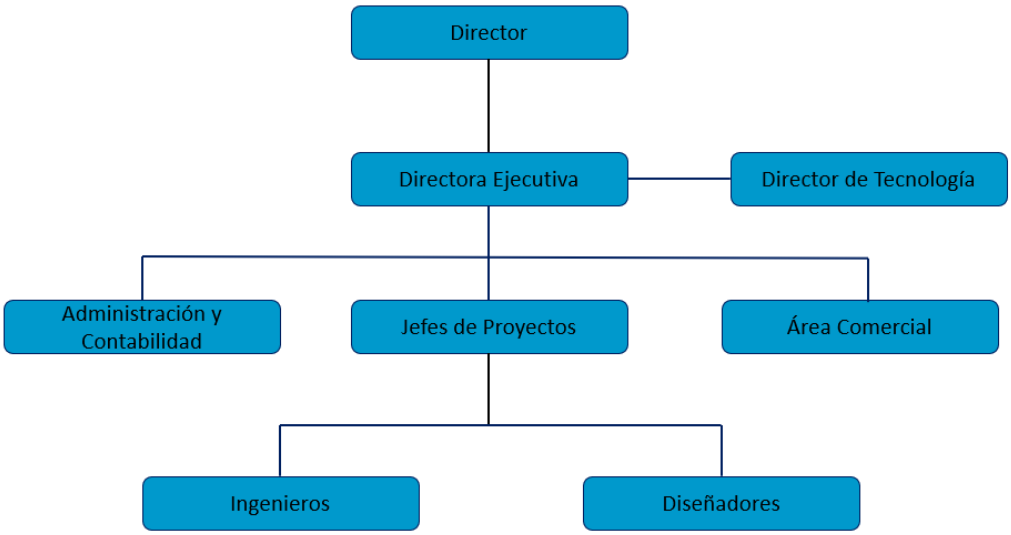

Figura 1.1: Organigrama del Web Intelligence Centre. Elaboración Propia.

Los productos y servicios que provee el WIC se catalogan en proyectos, investigaciones y asesorías, siendo las dos primeras las más relevantes en la organización. Tanto los proyectos como las investigaciones, pueden ser desarrolladas a partir de un fondo concursable adjudicado, sin embargo, los proyectos también pueden ser ejecutados a partir de una solicitud realizada por una institución para que el centro se haga cargo de una problemática mediante el uso de tecnologías de información. En cuanto a las asesorías, estas corresponden a servicios de consultorías que son realizadas a otras instituciones, abordando el ámbito de las tecnologías de la información.

Dentro del WIC se desarrollan tres tipos de servicios principales destacados en su página web [1]:

- **Salud Digital:** Corresponde a la Integración de diferentes aplicaciones y herramientas de aprendizaje de máquinas con el servicio de la salud. Sus principales objetivos son apoyar la toma de decisiones, mejorar la calidad de atención para los pacientes, reducir costos a lo largo de la cadena de atención médica y mejorar la experiencia de pacientes, médicos y otros colaboradores dentro del área de la salud.

- **Consultoría e Investigación:** Apoyo a empresas en la gestión de datos, tanto en su almacenamiento, cuidado, arquitectura y visualización de la información. Este

servicio ayuda en la toma de decisiones a sus clientes y les permite utilizar el potencial de sus datos de negocio.

- **Área Tecnológica:** Creación de plataformas tecnológicas utilizando algoritmos y herramientas de data science para apoyar la toma de decisiones y procesos en organizaciones.

El trabajo dentro del centro de investigación se desarrolla en base a proyectos. Estos proyectos están en su mayoría enfocados al área de la salud y se destacan algunos a continuación.

- **Proyecto Diagnóstico Diferencial Entre Depresión y Alzheimer (ADDAI)**: Desarrollo de un software de clasificación y scoring que integre señales de electroencefalogramas, eye trackers y test de navegación para diagnosticar de manera inicial a personas con depresión o alzheimer.

- **KEFURI**: Desarrollo de una aplicación móvil para dar aviso de posibles donantes de órganos a la Unidad de Procuramiento desde Urgencia y UCI.

- **SONAMA**: Plataforma informática de análisis de redes sociales para estudiar la prevalencia del consumo de marihuana y alcohol en Chile.

- **DOCODE**: Plataforma de detección de plagios que automatiza el proceso de análisis de documentos digitales, permitiendo encontrar coincidencias que podrían ser consideradas como plagio.

Es relevante mencionar que los proyectos e investigaciones desarrollados en el WIC pueden ser sustentados económicamente por fondos concursables adjudicados por el centro. Dentro de este contexto, se pueden determinar dos tipos de clientes principales para el Web Intelligence Centre: Entidades que fomentan la investigación y organizaciones que soliciten el servicio del centro.

Partiendo por las entidades que fomentan la investigación, estas organizaciones permiten financiar proyectos y entregan apoyo en torno a la investigación en Chile, en esta categoría, SERCOTEC y SENCE son clientes del WIC. Existen también agencias que potencian la investigación nacional a través de soporte en diferentes ámbitos a proyectos innovadores, dentro de este rubro se encuentra la Agencia Nacional de Investigación y Desarrollo (ANID) y la Corporación de Fomento de la Producción (CORFO). Por otro lado,

son clientes también las organizaciones que soliciten los servicios de consultoría o investigación. Dentro de esta categoría se encuentran organizaciones como la Cámara de Innovación Farmacéutica (CIF) y Amicar. Finalmente, se encuentran los usuarios del WIC, los cuales corresponden a personas beneficiadas por los diversos proyectos realizados dentro del centro de investigación.

El Web Intelligence Centre es uno de los más de 50 centros de investigación que posee la Universidad de Chile y a través de la obtención de fondos concursables, proyectos y consultorías a empresas públicas y privadas, maneja una cantidad de ingresos de $200 millones de pesos.

1.1.2 Marco institucional

Dentro del marco institucional en el que se desenvuelve el Web Intelligence Centre, existen tres actores principales encargados de regular el funcionamiento del centro y los cuales deben aprobar cada requerimiento que realiza el WIC. Estos actores corresponden a la Universidad de Chile, la Facultad de Ciencias Físicas y Matemáticas y el Departamento de Ingeniería Industrial. Sumándose a estos actores, se encuentran organizaciones externas que entregan financiamiento para proyectos privados o públicos, como ANID y CORFO, o SERCOTEC y SENCE. Los requerimientos para poder acceder a estos fondos dependen de la participación de concursos, presentaciones de proyectos y licitaciones, lo que implica competir por poder adjudicarse un proyecto contra el último actor dentro del marco en el que se desenvuelve el WIC, y corresponde a otros centros de investigación de la misma universidad u otra, consultoras y centros de investigación particulares, los cuales actúan como competencia directa al momento de buscar financiamiento pese a no pertenecer al rubro de la salud.

El proyecto en el cual se realizará una experiencia demostrativa durante esta memoria y será explicado con más detalle en la siguiente sección, corresponde al proyecto de diagnóstico diferencial entre el deterioro cognitivo causado por depresión o alzheimer, el cual logró adjudicarse fondos públicos a través del concurso IDeA I+D 2020 llevado a cabo por la Agencia Nacional de Investigación y Desarrollo. Considerando como sistema mayor este concurso, con un total de fondos a repartir de $7.991.711 millones, y que el WIC logró adjudicarse $199.940 millones destinados a este proyecto, la participación del centro de investigación dentro de este mercado corresponde a aproximadamente un 2,5%. [2]

Con respecto al marco regulatorio relevante para este trabajo de tesis, cabe destacar que el Web Intelligence Centre pertenece a la Universidad de Chile, por lo que debe regirse bajo los lineamientos de esta universidad pública para la adquisición de bienes y servicios, postulaciones a licitaciones, rendiciones de gastos y pagos de sueldos. Esto quiere decir que todo este proceso debe ser llevado a cabo a través del sistema de

mercado público, sujeto a la Ley N° 20285 Sobre Acceso a la Información Pública, o más conocida como Ley de Transparencia [3].

Junto con esto, y al estar trabajando con datos biométricos asociados a pacientes, existen dos leyes relevantes que procuran el cuidado de la información de las personas y que estos datos no terminen en manos de terceros:

Ley N° 19.628 Sobre Protección de la Vida Privada [4]: Esta ley explica los mecanismos para cuidar la vida privada y los datos de carácter personal de las personas, enfocándose en la recolección de datos a través de encuestas, estudios de mercado y obtenidos a través de mecanismos de recolección de información personal. La ley establece que se debe informar a las personas del propósito de la utilización de sus datos y cuales variables o preguntas son de carácter obligatorio. Se estipula además que la utilización y tratamiento de los datos sólo puede llevarse a cabo cuando la ley lo disponga o cuando exista una autorización escrita del titular de esta información.

Ley N° 19.223 Tipifica Figuras Penales Relativas a la Informática [5]: Esta ley rige las actividades informáticas relativas a delitos contra el apoderamiento, uso o difusión indebido de datos alojados en sistemas de información, también regula la destrucción, inutilización o alteramiento de estos sistemas de información de carácter malicioso.

1.2 Descripción del proyecto y justificación

1.2.1 Información del área de la organización

La escasez de datos en proyectos estadísticos o aplicaciones de aprendizaje de máquinas no es un problema reciente, es un problema que se ha discutido en distintas instancias, foros y publicaciones científicas por al menos 20 años. Como se puede apreciar en el siguiente estudio del 2001 [6], la discusión en torno a cómo afecta el p-valor a probar una hipótesis nula, y cómo este p-valor se ve afectado por el poder estadístico, es decir, la probabilidad de aceptar una hipótesis cuando es verdadera, el cual depende de la cantidad de datos que se manejan, es de gran relevancia el día de hoy para poder ejecutar modelos de aprendizaje de máquinas más certeros.

La significancia estadística o p-valor, el cual corresponde a la probabilidad de que una relación entre dos o más variables en un análisis no sea pura coincidencia, es fijada comúnmente en p = 0,05 en la mayoría de los estudios. Según [7], la probabilidad de tener un resultado contradictorio, fijando un p-valor de 0,05 corresponde a 29%. Este resultado se explica por la diferencia de poder estadístico que poseen los estudios, y cómo este poder se ve afectado a su vez por la falta de datos, lo que impide la generalización a la población de los experimentos que se realizan en dichos estudios.

Los modelos estadísticos, y por lo tanto los algoritmos de aprendizajes de máquinas que los utilizan, buscan poder explicar el comportamiento de la población a través de una muestra, en otras palabras, buscan poder generalizar los resultados que se obtienen en base a la información que poseen. Este gran salto de generalizar los resultados es lo que complica el trabajo de investigadores en diferentes áreas en donde acceder a una gran cantidad de datos es complejo, ya que, al tener una baja cantidad de muestras, y por lo tanto un bajo poder estadístico y escasa variabilidad de la muestra, no es posible encontrar un resultado y extrapolarlo a una población [8].

En el área de la salud, en particular en donde se realizan estudios a un grupo acotado de pacientes, la restricción de generalización (COG o Constraints on Generality en inglés) [9], puede ser un efecto bastante relevante, ya que impide extrapolar los resultados obtenidos a un público más general al no poder determinar de dónde provienen la población de estudio. Este efecto se puede apreciar cuando se realizan estudios clínicos con diferentes cantidades de pacientes, en donde se pueden obtener resultados distintos dependiendo de la cantidad de participantes. El problema al que se enfrentan los investigadores sobre todo en este tipo de investigaciones yace en que la participación por parte de los pacientes puede tener efectos adversos, como efectos secundarios de drogas que se están probando, o que participar implique no recibir el tratamiento respectivo por pertenecer a un grupo placebo.

Es por estas razones que la falta de datos tiende a ser un problema relevante en el ecosistema de investigación, genera resultados falsos o inciertos, provoca que estudios en donde los pacientes tienen efectos adversos no tengan resultados positivos y dificulta la generalización de los resultados de las investigaciones que si tienen resultados favorables.

En el contexto del proyecto ADDAI que se utilizará como piloto para esta memoria, las herramientas psico-geriátricas más utilizadas para diagnosticar, evaluar y controlar el progreso de la enfermedad son poco efectivas. Es por esto que es relevante diseñar nuevas maneras de detectar este tipo de enfermedades mentales. Por otro lado, existen métodos que permiten diferenciar con buenos resultados ambas enfermedades, los cuales estipulan largos protocolos y requieren de varios especialistas y doctores, por lo que su aplicación es bastante costosa y demanda de una gran cantidad de tiempo por parte de los pacientes. Es por estas razones que el objetivo principal de este proyecto es generar una herramienta precisa, objetiva, de bajo costo y ágil utilizando estrategias tecnológicas para el diagnóstico diferencial entre el deterioro cognitivo generado por depresión y el generado por la enfermedad de Alzheimer.

Si bien el Web Intelligence Centre no posee un área de proyectos, sino jefes encargados de cada proyecto, este trabajo de título será desarrollado en una primera instancia bajo el contexto del proyecto anteriormente mencionado. Este proyecto es de carácter

interdisciplinario, y cuenta con 20 profesionales de múltiples áreas, como médicos, ingenieros, diseñadores y enfermeras. Los usuarios son potenciales pacientes sin diagnóstico de depresión o alzheimer en nuestro país.

Las regulaciones que afectan al proyecto piloto son similares a las establecidas con anterioridad en este informe, pero el cuidado con los datos de las personas, en particular datos biométricos recaudados desde exámenes clínicos, requieren un cuidado adicional en el manejo y transformación de la información. Existe una ética detrás del manejo de la data de pacientes, la que resguarda la privacidad y seguridad de las personas al momento de tener que ser examinadas y evitar así que sus datos sean utilizados para otros fines.

Los actores dentro de este proyecto corresponden al Web Intelligence Centre, facultades de Odontología, Medicina y de Ciencias Físicas y Matemáticas de la Universidad de Chile, y el Laboratorio de Neurosistemas.

El solicitante de este trabajo de título es el WIC, en búsqueda de poder ejecutar este y otros proyectos relacionados a aprendizaje de máquinas que requieran una gran cantidad de datos y sea escasa la data real. El conflicto que busca resolver con esta memoria es encontrar diferentes formas de generar datos sintéticos, las cuales varían según el tipo de data que se tenga, para poder aumentar la cantidad de información en los modelos y permitirles utilizar herramientas de data mining con un mejor rendimiento.

1.2.2 Identificar el problema u oportunidad y su relevancia, con sus efectos y posibles causas.

Al momento de desarrollar una investigación, es muy difícil poder acceder a datos de toda la población, sobre todo cuando los datos que se buscan son muy específicos. Es por esto que los investigadores utilizan modelos estadísticos para poder generalizar fenómenos, ya que no pueden observar a la población en su totalidad, necesitan estimar el fenómeno que buscan en base a las observaciones que manejan. Es aquí cuando la falta de datos perjudica a los resultados de las investigaciones, cuando la información que se tiene no es suficiente para poder generalizar a la población y entender el fenómeno real.

Es por esta razón que el problema por abordar en este trabajo de título corresponde al Small Dataset Problem (SDP), o problema asociado a tener pocos datos en ámbitos de investigación y cómo estos escasos datos no permiten a estudiantes e investigadores a entrenar modelos, ya sean estadísticos o de aprendizaje de máquinas, que permitan agregar valor en las áreas en las que se desempeñan.

La escasez en la cantidad de datos proviene de tres factores que serán explicados con mayor detalle en la siguiente sección, pero que corresponden al tiempo que toman los exámenes en ser realizados, el costo asociado a generar un examen y el costo que implica participar para los pacientes. Esta falta de datos provoca dificultades al momento de ejecutar modelos de aprendizaje de máquinas, ya que al tener muy pocos datos, los modelos son muy sensibles a variaciones pequeñas, pierden el poder de generalización y aplicabilidad para el resto de los pacientes y pueden sufrir de overfitting, en otras palabras, de replicar exactamente los resultados provenientes de las bases de testeo. Estos problemas asociados a tener pequeñas muestras han sido explorados en [10].

La relevancia de solucionar esta falta de datos, en particular en el área de la salud en donde generar información es costoso y requiere de un manejo ético, recae en el objetivo no sólo de esta experiencia demostrativa en el proyecto ADDAI en particular, sino de los posibles proyectos del WIC que puedan tener problemas similares. Esto ya que es posible aplicar esta solución en otras instancias y así evitar problemas al entrenar modelos de aprendizaje de máquinas. Se debe entender entonces que poder crear una metodología para generar datos virtuales habilita a investigadores y estudiantes a seguir desarrollando soluciones que tengan un resultado e impacto positivos.

Dentro del marco del proyecto de depresión y Alzheimer, que será utilizado como experiencia demostrativa, se realiza un examen clínico a personas sanas, las cuales pertenecen a grupos de control, y a personas con un diagnóstico médico confirmando. Este examen obtiene información desde tres fuentes; Un electroencefalograma (EEG), una cámara que permite realizar seguimiento ocular del paciente, y un test de navegación, el cual es ejecutado a través de un juego virtual ejecutado en un computador,

lo que sumerge al paciente en un ambiente sintético para someterlo a pruebas que midan su memoria espacial y tiempos de reacción. Estos exámenes se aplican al mismo tiempo, buscando encontrar patrones que permitan diferenciar de manera sencilla y a través de los datos a los pacientes. La figura 1.2 ilustra de manera sencilla los exámenes realizados en paralelo y la información obtenida desde estos instrumentos de medición.

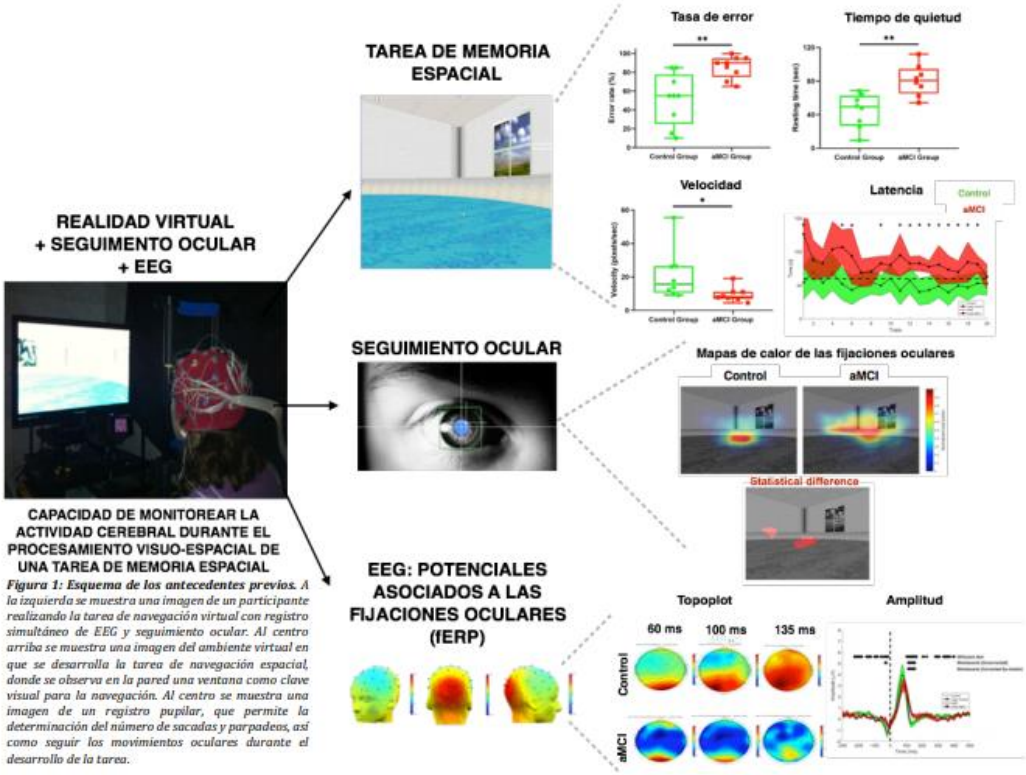

Figura 1: Esquema de los antecedentes previos. A la izquierda se muestra una imagen de un participante realizando la tarea de navegación virtual con registro simultáneo de EEG y seguimiento ocular. Al centro arriba se muestra una imagen del ambiente virtual en que se desarrolla la tarea de navegación espacial, donde se observa en la pared una ventana como clave visual para la navegación. Al centro se muestra una imagen de un registro pupilar, que permite la determinación del número de sacadas y parpadeos, así como seguir los movimientos oculares durante el desarrollo de la tarea.

Figura 1.2: Esquema del examen clínico. Fuente: Postulación del proyecto al concurso IDeA I+D 2020.

Como consecuencia de los extensos exámenes de entre una a una hora y media, los altos costos de las herramientas utilizadas y la necesidad de búsqueda de pacientes, es que actualmente existen menos de 40 registros de pacientes para alimentar un modelo de aprendizaje de máquinas. Este proyecto cuenta entonces con un problema de falta de datos y será utilizado como un proyecto piloto para poder construir una solución que permita a los investigadores poder superar esta restricción.

En este proyecto en particular, existe una necesidad de ejecutar modelos de clasificación para pacientes por parte del WIC con los datos obtenidos de los exámenes con el fin de poder realizar un prediagnóstico de los sujetos en base a sus datos y facilitar el diagnóstico temprano de las personas que padecen estas enfermedades.

1.2.3 Identificar hipótesis y posibles alternativas de solución para resolver el problema u oportunidad

Para poder solucionar la escasez de datos en esta área, una primera aproximación podría ser simplemente realizar una mayor cantidad de exámenes a pacientes, pero existen restricciones de tiempo, costos y búsqueda de pacientes, además de una necesidad de personal capacitado para ejecutar dichos procedimientos, lo que impide a esta opción de ser viable.

En particular para este proyecto, y según la postulación del proyecto al concurso IDeA I+D 2020 realizada por el equipo médico, el tiempo de la realización del Test de Navegación es de 60 minutos en una etapa inicial, y puede llegar a reducirse a 30-40 minutos tras la ejecución del proyecto, mientras que los costos asociados van desde los $40 mil pesos a los $80 mil pesos por examen. Por otro lado, el requerimiento de personal técnico capacitado para la ejecución de este proceso, y la búsqueda de pacientes que tengan un diagnóstico clínico de depresión y enfermedad de Alzheimer, hacen que los costos relacionados a esta medida sean muy elevados.

Existe otra arista importante, la cual corresponde a la utilización de data sensible, que está sujeta al código sanitario [11] en su artículo 101, en donde se explicita que las recetas médicas y su contenido, los análisis y exámenes de laboratorios clínicos y los servicios prestados relacionados con la salud serán reservados y considerados datos sensibles sujetándose a lo establecido en la ley N 19.628. Esto implica que, para tener acceso a dicha información, se requiere elevar una solicitud que puede o no ser aceptada, y que, en la mayoría de los casos, esa solicitud tarda mucho tiempo en ser respondida o es denegada.

Como se ha observado en [12], un problema clave a ser explorado es cómo la data sintética puede ser utilizada sin transgredir la privacidad de los pacientes, lo que se refiere a poder generar modelos sin entregar información de aquellos que estén presentes dentro del modelo. Para los efectos de poder medir qué tanta información entrega los modelos sobre estos pacientes, existen métricas que cuantifican la pérdida de privacidad y que también miden cuanta información real se entrega al presentar un resultado, con esto es posible detectar posibles fugas de información sensible y evitar su propagación.

Otra opción corresponde a la realización de variados exámenes para descartar otras enfermedades, sin embargo, esta opción sigue sujeta a restricciones de tiempo y costos, mientras que su escalabilidad en el sistema de salud de Chile se presenta como un desafío mayor. También existe la posibilidad de entrenar modelos de aprendizaje de máquinas con datos de otros países utilizando bases de datos públicas que contengan registros de pacientes. El problema dentro de este contexto es que en la mayoría de los casos los datos y las condiciones de los pacientes que se buscan son muy específicas y no están presentes dentro de estas bases de datos. Además, se pierde la localidad de

los datos, ya que se busca entrenar modelos utilizando información local, con el fin de ocuparlos para apoyar al diagnóstico de personas de nuestro país que pueden presentar diferencias con personas de otras partes del mundo.

Es por esto que se ha decidido apoyar al Web Intelligence Centre a buscar una metodología que permita generar datos de manera sintética, ya que permite evitar o disminuir las restricciones de tiempo y de costos mencionadas, y si los datos se generan de manera adecuada y responsable, se puede asegurar la anonimidad de la información. Esto permite su divulgación y facilidad de acceso por parte de investigadores y estudiantes que lo requieran, haciendo crecer al ecosistema de investigación y descubrimientos en torno a temáticas tan relevantes y que seguirán siéndolo en el futuro como lo son la utilización de herramientas de aprendizaje de máquinas en el área de la salud. Todo esto busca aportar a poder finalmente cubrir la necesidad de los pacientes de ser diagnosticados de manera temprana, la necesidad de los equipos médicos de poder ejecutar sus diagnósticos de manera ágil y la necesidad de los investigadores de poder acceder y trabajar con esta información.

Esta alternativa de solución está sujeta a tres hipótesis principales, las cuales son explicadas a continuación.

Hipótesis 1: Existe una cantidad suficiente y variada de datos, la que permite a los modelos generadores de datos crear datos sintéticos lo suficientemente cercanos a la realidad para ser utilizados.

Se considera que los datos reales provenientes de exámenes realizados a pacientes son suficientemente variados entre sí y serán una cantidad tal que permita utilizar modelos de generación de datos sintéticos y obtener resultados cercanos a los reales.

Esta es la hipótesis más relevante para el desarrollo de esta memoria ya que la cantidad de datos necesaria para ejecutar estos modelos de manera certera es incierta, y depende de la información que proporcionen las relaciones entre las variables involucradas y la distribución de los datos.

Hipótesis 2: Los datos sintéticos no comprometerán la privacidad de los pacientes, permitiendo la utilización de esta información para fines de investigación y desarrollo.

Es de gran importancia poder asegurar la anonimidad de la información generada, ya que de esta manera no se compromete la privacidad de los pacientes. Existen casos en donde pese a utilizar datos sintéticos, un posible ataque permite re identificar a pacientes reales en set de testeo.

Para poder medir la privacidad presente en la data sintética, se utilizará la distancia de Hamming como en [13]. Esta métrica permite identificar si es que un paciente real está presente en la base de testeo, calculando si este se encuentra a cierta distancia de al menos un dato generado sintéticamente.

Hipótesis 3: Los investigadores y equipos de ingenieros de datos estarán dispuestos a utilizar datos generados virtualmente.

Existen numerosas dudas sobre la generación de data sintética desde los principales usuarios de esta información, quienes son investigadores. Estas dudas corresponden a la confiabilidad de los datos y si es que realmente son una fiel representación de la realidad.

En una primera instancia, se debe asegurar la calidad de los datos utilizados. Es sabido ya que los modelos de aprendizaje de máquinas y sobre todo los algoritmos generadores replican la lógica que hay detrás de la información utilizando distribuciones y correlaciones entre variables, y que, si los datos en los que se entrenan estos algoritmos no son de calidad, o no representan fielmente los fenómenos que buscan explicar, entonces los modelos replicarán esta lógica y generarán datos que no aporten valor.

Para evitar dudas sobre la viabilidad de los datos y demostrar la utilidad de la generación de datos, se buscará responder a las siguientes preguntas durante la implementación de esta memoria:

- ¿Cuál es la diferencia en el rendimiento de entrenamiento y testeo de modelos de aprendizaje de máquinas utilizando data sintética versus la data real? Esto con el fin de poder asegurar la funcionalidad y similitud de los datos generados. Existe un dato extra que se obtiene al conocer este factor, y es que, si la diferencia es muy pequeña, los datos sintéticos serán idénticos a los reales, y se estará frente a un problema de overfitting.

- La segunda pregunta corresponde a conocer cuál es la varianza en las diferentes métricas de rendimiento entre modelos entrenados en data sintética o real. Para efectos de este trabajo de título, se utilizarán las métricas Accuracy, Recall, Precision, F1-Score y la curva ROC AUC. Estas métricas serán explicadas con más detalle en el apartado 1.4.5, correspondiente al marco conceptual.

Si responder estas dos preguntas es posible durante el transcurso de la memoria, se podrá respaldar la validez de los datos y evaluar cuán exitosa fue esta etapa de generación, presentando información útil para diferenciar entre las técnicas generativas que se aplicarán en esta memoria.

1.2.4 Propuesta de valor de las posibles soluciones o impacto del cambio propuesto

El valor de este trabajo de título se basará principalmente en crear una metodología de generación de datos en base a la experiencia y los resultados experimentales obtenidos durante la implementación de distintos algoritmos generadores de datos sintéticos. Este proceso se repetirá con bases de datos de diferentes características, y será evaluado con diferentes métricas que permiten medir la similitud entre los datos reales y los generados virtualmente. Esto permitirá conocer cuáles son los algoritmos que generen los mejores resultados dependiendo de las características intrínsecas de los datos.

Esto permitirá al WIC poder discernir entre la utilización de esta clase de algoritmos para diferentes proyectos que requieran tener una mayor cantidad de datos para ejecutar modelos de aprendizaje de máquinas, lo cual es un problema habitual dentro del centro dado que su foco principal es la integración de nuevas tecnologías en el área de salud. Es en esta área en particular, en donde la generación de información tiene un alto costo, ya sea en el capital humano requerido para realizar los exámenes o los implementos utilizados para medir señales e impulsos, como el electroencefalograma. Junto con esto, existe una ética en el manejo de la información privada de los pacientes, la cual es respetada y asegurada al momento de utilizar esta información para alimentar modelos generadores de datos.

Esta metodología permitiría a investigadores y estudiantes poder acceder a información a la cual no tienen acceso actualmente, evitando problemas asociados a la privacidad de los pacientes y evitando también restricciones relacionadas a los costos y tiempo asociados a la necesidad de tener una gran cantidad de información para utilizar modelos sofisticados de aprendizaje de máquinas.

Tan sólo en el proyecto de diagnóstico diferencial entre depresión y Alzheimer, y en particular en la etapa de exámenes para diagnóstico de depresión, se estima según datos de la propuesta realizada por el equipo médico y el WIC para el concurso IDeA I+D 2020 de ANID, que el valor de estos exámenes ronda los $40 mil pesos - $80 mil pesos. Por simplicidad, se asume que los exámenes de depresión cuestan $60 mil pesos, por lo que, con tan sólo generar 100 exámenes de manera virtual dentro del contexto de este proyecto, se podrán ahorrar recursos del orden de $6 millones de pesos, sin contar el valor de los exámenes para la enfermedad de Alzheimer que tienden a ser más extensos y costosos.

Es relevante destacar que el valor de esta memoria no está sujeto a esta aplicación en particular, si existiese la posibilidad de aplicar esta metodología en otros proyectos, el posible valor podría aumentar, dependiendo de en qué tipo de exámenes se utilice y con qué fines.

1.3 Objetivos

A continuación, se detallan cuáles serán los objetivos del trabajo de título.

1.3.1 Objetivo General

Entregar recomendaciones para resolver los problemas asociados a la escasez de datos en proyectos de aprendizaje de máquinas asociados a salud, a través de pruebas y evaluaciones de diferentes métodos de generación de datos sintéticos en base a la experiencia demostrativa del proyecto ADDAI.

1.3.2 Objetivos Específicos

1. Investigar el estado del arte, con el fin de contextualizar el problema a tratar, conocer qué técnicas y métodos se utilizan para la resolución de dicho tipo de problemas, y que se ha realizado en casos similares.

2. Simular datos reales utilizando distintos métodos y algoritmos de generación de data sintética, utilizando como experiencia demostrativa el proyecto ADDAI, para poder aumentar la cantidad de registros con los que se cuenta actualmente y ampliar el espectro de modelos de aprendizaje de máquinas que puedan ser aplicados.

3. Medir el rendimiento de la simulación de datos a través de distintas métricas, definiendo casos de éxito y fracaso, con el fin de poder distinguir la calidad de los datos sintéticos generados.

4. Realizar una propuesta de producto o servicio, con el fin de mostrar el aporte que implica la generación de datos en proyectos similares.

5. Realizar un análisis de impacto económico, que pueda dar cuenta de cuán positiva puede ser la creación de datos sintéticos y a quienes puede afectar.

1.4 Marco conceptual

A continuación, se definirán y explicarán ciertos conceptos, técnicas, algoritmos y herramientas que se utilizarán en el futuro para el trabajo de título.

1.4.1 Bases de Datos

Una base de datos es una colección organizada de información [15]. La información almacenada en bases de datos no se limita a números, esta puede corresponder a imágenes, música o links. Esta colección se ordena generalmente a través de una tabla con columnas o variables y filas u observaciones. Existe una estructura que permite almacenar información en más de una tabla, la cual se denomina base de datos relacional, la cual permite establecer relaciones entre tablas a través de variables en común. El factor que distingue a las bases de datos de otras fuentes de información es su organización. Al coleccionar datos con cierta estructura, facilita el acceso y entendimiento de esta.

Hoy en día, existe una gran variedad de programas que permiten interactuar y manipular bases de datos, los cuales son denominados Sistemas de Administración de Bases de Datos (Database Management System, DBMS). Los softwares de DBMS además de interactuar, permiten llegar a grandes niveles de abstracción de la información, pudiendo aislar elementos y segmentar datos para facilitar el análisis.

1.4.2 Python

Este lenguaje de programación es uno de los más utilizados dentro del mundo de la ciencia de datos y en específico para aplicaciones de aprendizaje de máquinas. Python funciona utilizando librerías, las cuales aportan funcionalidades específicas dependiendo del problema que se está abordando. Este lenguaje de programación está basado en objetos, o sea, permite a los usuarios utilizar las librerías y bibliotecas importadas a través de la creación de objetos.

Python no posee integradas las funciones para aplicar algoritmos de aprendizaje de máquinas ni para el manejo de datos estandarizados y aquí es en donde se importan las bibliotecas que permiten realizar y trabajar con estos modelos, un ejemplo de biblioteca es Scikit-Learn, la cual es un módulo que integra una variada gama de algoritmos de ML supervisados y no supervisados [16]. La especialidad de esta biblioteca está en su facilidad de uso y eficiencia de sus algoritmos, lo que permite a los usuarios no especializados utilizar estos algoritmos de manera sencilla.

1.4.3 Algoritmos de aprendizaje de máquinas

El aprendizaje de máquinas es un método de análisis de datos el cual automatiza la construcción analítica de los modelos, en su nivel más básico, se refiere a programas computacionales capaces de aprender de los datos sin ser programados específicamente para un tipo de problema [85]. Estos métodos permiten a los investigadores y cientistas de datos combinar la gran cantidad de datos que se generan día a día en las distintas organizaciones de amplios rubros, en conjunto con el aumento constante de la capacidad de computación, para crear programas más complejos, los cuales son capaces de predecir fenómenos, clasificar objetos u optimizar factores.

Existen distintos tipos de algoritmos de aprendizaje de máquinas, según el tipo de entrenamiento que se lleve a cabo según Taiwo Ayodele [17] en su libro “Types of Machine Learning Algorithms”, estos tipos corresponden a:

1. **Aprendizaje supervisado:** Algoritmo que debe ser capacitado o entrenado para poder encontrar patrones etiquetados dentro de la data. Los algoritmos más característicos corresponden a modelos de clasificación.

2. **Aprendizaje no supervisado:** Algoritmo que no debe ser entrenado, ya que analiza patrones no etiquetados dentro de los datos.

3. **Aprendizaje semi supervisado:** Combina ambos tipos de algoritmos mencionados anteriormente.

4. **Aprendizaje por refuerzo:** Tipo de algoritmo que aprende cómo actuar dada una observación de los datos. Cada cambio que realice el modelo tiene un impacto en su ambiente y este ambiente sirve de guía para el proceso de aprendizaje del algoritmo.

1.4.4 Algoritmos Generadores de datos

Los algoritmos generadores de datos corresponden a modelos que permiten crear datos sintéticos a partir de datos reales, buscando replicar la información que existe detrás de la data en sí. Estos modelos utilizan la distribución, relaciones entre variables y tipos de variables para poder crear datos que, si bien no son iguales a los reales, buscan recrear fielmente la causalidad de la información que están buscando replicar.

Los algoritmos que se utilizarán en esta memoria corresponden a los siguientes:

Gaussian Copula (GC):

Una cópula es un modelo estadístico utilizado para entender las dependencias estructurales que existen entre diferentes distribuciones. Este tipo de modelos ha sido utilizado para generar datos anteriormente [18]. En términos matemáticos, una cópula es una función de distribución de probabilidad construida desde una distribución normal multivariable sobre el espacio de muestreo y es representada como un cubo unitario de dimensiones entre 0 y 1, uno de los ejes de este cubo corresponde a la función de probabilidad, mientras que el resto corresponden a las distribuciones de las variables. Esta función permite describir la distribución conjunta de múltiples variables aleatorias analizando la dependencia que existe entre las distribuciones marginales de las variables en la base de datos [19]. Es importante entender que Gaussian Copula es un generador estadístico, el cuál aprende los datos a través de simular distribuciones conocidas y posteriormente puede generar esta información obteniendo muestras desde estas distribuciones aprendidas.

Generative Adversarial Networks (GAN):

Las Generative Adversarial Networks [20] son un modelo de aprendizaje de máquinas para la generación de datos basado en la Teoría de Juegos. El fin de las GANs es entrenar dos redes neuronales, un algoritmo generador, que produzca muestras desde la distribución de la data, y otro discriminador, que sea capaz de diferencias las muestras reales de las sintéticas. El algoritmo generador es entrenado para engañar al discriminador y forzarlo a aceptar sus muestras como reales, mientras que el discriminador aprende a separar estas muestras. Estas redes neuronales se entrenan a partir de su participación en un juego continuo, en el cual el algoritmo generador aprende a engañar al discriminador mejorando la calidad de sus muestras, mientras que el generador aprende a no ser engañado y diferenciar los datos reales de los sintéticos, este proceso se realiza de manera no supervisada. La figura 1.3 muestra cómo funciona esta etapa de entrenamiento entre los algoritmos generadores (G) y los discriminadores (D).

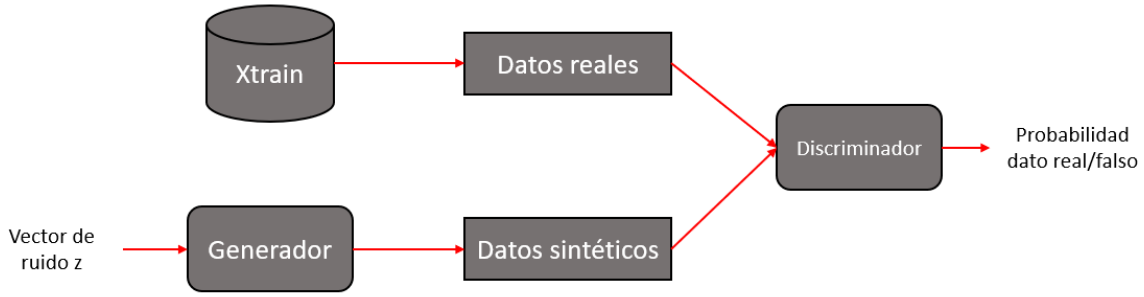

Figura 1.3: Arquitectura del algoritmo GAN. Basado en [21].

Más formalmente, y como es explicado en [21], para aprender la distribución de los datos (x) que entrega el generador, se define una variable de ruido $\boldsymbol{p_z(z)}$ y se representa también el espacio de la data como $\boldsymbol{G(z;\theta_g)}$, donde G es una red neuronal con parámetros $\boldsymbol{\theta_g}$. También se define el discriminador $\boldsymbol{D(x;\theta_d)}$ el cual entrega la probabilidad $\boldsymbol{D(x) \in [0,1]}$ de que el dato (x) haya sido obtenido desde la base de testeo en vez de provenir del generador. El algoritmo discriminador es entrenado para maximizar la probabilidad de asignar la etiqueta correcta a las muestras reales y a las sintéticas, generadas por $\boldsymbol{G}$. Se entrena simultáneamente al generador para minimizar $\boldsymbol{\log(1 - D(G(z)))}$. El problema de optimización final resuelto por las dos redes está definido como un juego de min-max de la siguiente manera:

$$\boldsymbol{\min_G \max_D} \mathbb{E}_{\mathbf{x}\sim p_{\text{data(x)}}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

Desde su creación, las GANs han sido muy populares en el mundo de la generación de datos, ya que permiten recrear múltiples tipos de variables y datos, lo cual las hace muy versátiles. Sin embargo, este proceso tiende a ser inestable y usualmente resulta en el colapso del modelo evitando obtener resultados positivos [22]. Esto ocurre cuando el generador encuentra algún grupo de datos con los que puede engañar al discriminador más fácilmente, por lo que comienza a favorecer estos tipos de datos en la generación y no se obtienen resultados fieles a la realidad. Es por esto que nacen iniciativas para utilizar variantes de las GANs, que les permiten adecuarse mejor a los modelos dependiendo del caso. Se distinguen dos grandes tipos de variantes de GANs: El primero responde a la inestabilidad del proceso de aprendizaje de manera más general, mientras que el otro es generar soluciones para problemas específicos, usualmente con arquitecturas de modelo diferentes a las GANs. Algunos ejemplos del primer tipo de

variantes son el Least Squares Generative Adversarial Network (LSGAN) [23] y el Wasserstein Generative Adversarial Network (WGAN) [24], mientras que para ejemplificar el mejoramiento de la arquitectura de los algoritmos existe Adversarial Variational Autoencoders (VAEGAN) [25] y Auxiliary Classifier Generative Adversarial Networks (ACGAN) [26]

Variational Autoencoders (VAE):

Antes de hablar del algoritmo VAE, es relevante precisar que es un algoritmo autocodificador. Este tipo de algoritmos utilizan redes neuronales artificiales de manera no supervisada para aprender a comprimir y codificar los datos, en algunos casos reduciendo la dimensionalidad de la muestra. Posterior a este proceso, se procede a descomprimir y decodificar la información, con el fin de volver a una representación de los datos lo más similar a la original. Al entrenar una red de compresión y otra de descompresión, el algoritmo puede inferir información de las distribuciones de los datos y las relaciones entre las variables, pudiendo ser utilizado con el fin de generar datos sintéticos. La siguiente figura muestra la arquitectura básica de los algoritmos autocodificadores.

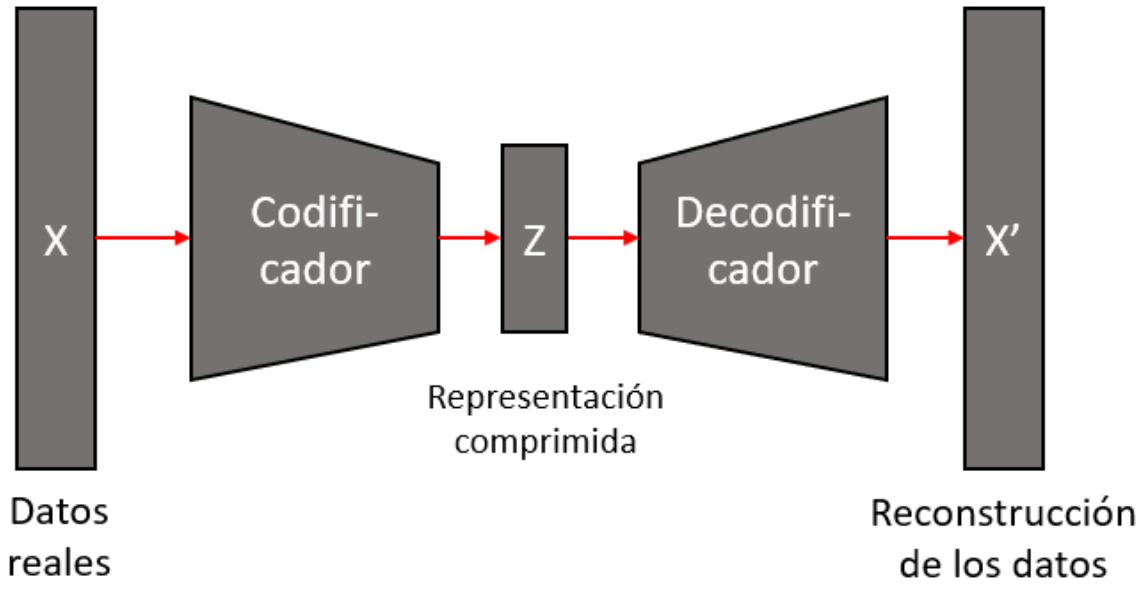

Figura 1.4: Arquitectura de un algoritmo autocodificador. Basado en [27].

Según lo visto en [28], la codificación típica utilizada en VAEs ha sido diseñada para retornar una distribución sobre una representación de data comprimida en vez de datos discretos. VAE es un algoritmo generador de datos, el cual busca estimar la función de

densidad de probabilidad de los datos reales. Esto lo logra produciendo dos vectores, uno de promedios (μ), y otro con las desviaciones estándar (σ) desde la data real. El algoritmo intenta aprender de las distribuciones de las variables latentes o inferidas, basándose en el promedio y la varianza por medio de compresiones y descompresiones de la información, utilizando un autocodificador. La figura 1.5 muestra un esquema de la arquitectura del algoritmo.

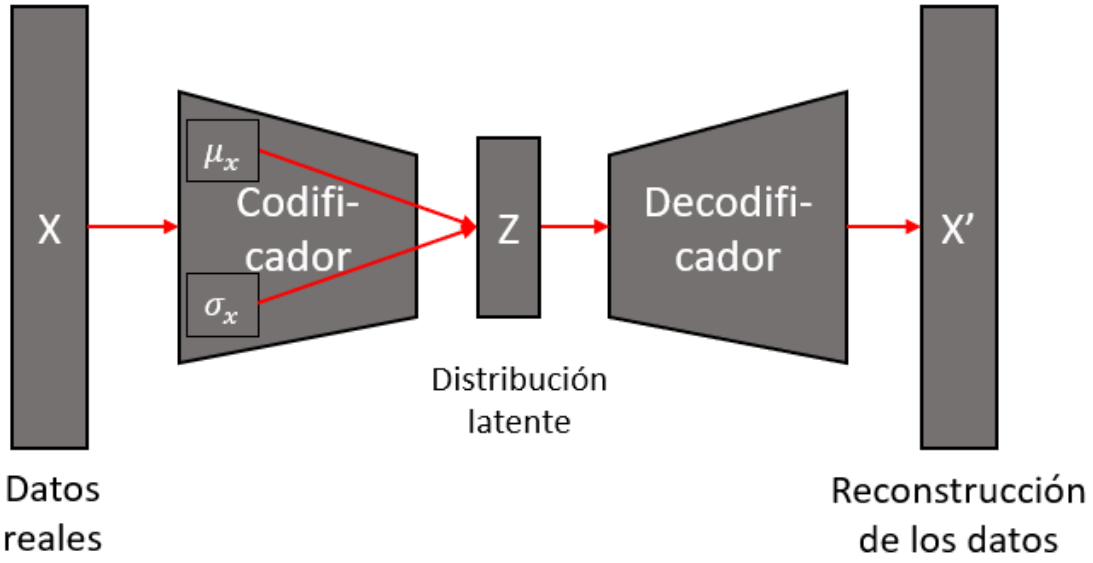

Figura 1.5: Arquitectura del algoritmo VAE. Basado en [27].

1.4.5 Evaluación de modelos generadores

Para poder medir el desempeño de los modelos y evaluar la calidad de los datos generados, es necesario poder comparar los modelos generadores de datos sintéticos entre sí. También existen métricas de evaluación que entregan información relevante que permite diferenciar el rendimiento de las variantes de cada algoritmo. Es por esto que para evaluar los resultados obtenidos durante el proceso de generación de data se utilizarán estas dos distinciones:

Evaluación de resultados entre modelos: La evaluación entre modelos permite comparar el rendimiento de los algoritmos evitando sesgos provocados por el funcionamiento de los algoritmos. Estas métricas se crean utilizando sólo los resultados obtenidos de los modelos. Algunos ejemplos vistos en [29] corresponden a los siguientes:

1. **Clasificación utilizando datos sintéticos:** Este método consiste en entrenar a un clasificador utilizando sólo data sintética para posteriormente clasificar datos reales, y comparar los resultados contra un clasificador que sólo ha sido entrenado utilizando datos reales. Mantener los resultados o tener una pequeña disminución

en los rendimientos en comparación con los datos reales, implica que los datos sintéticos son capaces de replicar la lógica y la información que existe detrás de los fenómenos reales.

2. **Clasificación mixta de data real y sintética:** Método similar a la clasificación utilizando datos sintéticos, diferenciándose en que los datos con los que se alimenta al clasificador son una combinación entre datos reales y sintéticos.

3. **Pruebas con algoritmos discriminadores:** Para evaluar la aplicabilidad de la data sintética generada y cómo puede ser utilizada en un escenario real, se combinan los datos sintéticos y reales en una única base de datos, la cual es etiquetada como real o sintética y posteriormente alimentada a un algoritmo clasificador, el cual ocupará el rol de discriminador y será entrenado con esta base. La meta del algoritmo es poder diferenciar la data real de la sintética, por lo que para que la generación de datos sea efectiva, el algoritmo no puede tener resultados muy malos, ya que los datos serán idénticos a los reales o muy buenos, ya que los datos serán totalmente diferentes a los originales. Este método busca imitar la arquitectura que utilizan los GANs.

Para evaluar el rendimiento de estos modelos de clasificación existen distintas métricas que permiten medir el ajuste de estos y la capacidad de realizar predicciones precisas y exactas. Para entender de mejor manera las métricas que serán utilizadas en este trabajo de tesis, es necesario conocer los conceptos de verdadero positivo, falso positivo, verdadero negativo y falso negativo, los cuales serán explicados a continuación. Para estos efectos, se utilizará el problema de clasificación binaria, el cual predice si un elemento pertenece o no a una clase.

- **Verdadero positivo** (VP): Predecir la clase A cuando realmente el elemento pertenece a esta clase.
- **Falso Positivo** (FP): Predecir la clase B cuando realmente el elemento pertenece a la clase A.
- **Verdadero negativo** (VN): Predecir la clase B cuando realmente el elemento pertenece a esta clase.
- **Falso negativo** (FN): Predecir la clase A cuando realmente el elemento pertenece a la clase B.

Estos valores se representan en una matriz llamada matriz de confusión, la que permite mostrar los niveles de ajuste del modelo. Las métricas que se explicarán a continuación utilizan los valores de estos indicadores.

Accuracy:

Representa la cantidad de clasificaciones correctas del total de clasificaciones o la probabilidad de predecir correctamente [69]. Formalmente se utiliza la siguiente fórmula para calcular el Accuracy del modelo.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{\#\ Predicciones\ correctas}}{\boldsymbol{Total\ de\ predicciones}} = \frac{\boldsymbol{VP + VN}}{\boldsymbol{VP + FP + VN + FN}}$$

Recall:
Corresponde a la tasa de verdaderos positivos, muestra la probabilidad del modelo de detectar casos realmente positivos [69]. Esta métrica responde a la pregunta de ¿Qué proporción de los datos en realidad positivos se predijo correctamente? Y formalmente se utiliza la siguiente fórmula.

$$\boldsymbol{Recall} = \frac{\boldsymbol{\#\ Predicciones\ positivas\ correctas}}{\boldsymbol{Total\ de\ reales\ positivos}} = \frac{\boldsymbol{VP}}{\boldsymbol{VP + FN}}$$

Precision:

Es la tasa de valores positivos predichos correctamente sobre el total de los valores positivos predichos [69]. Responde a la pregunta de ¿Qué proporción de los datos predichos positivamente son realmente positivos? Formalmente se utiliza la siguiente fórmula.

$$\boldsymbol{Precision} = \frac{\boldsymbol{\#\ Predicciones\ positivas\ correctas}}{\boldsymbol{Total\ de\ predicciones\ positivas}} = \frac{\boldsymbol{VP}}{\boldsymbol{VP + FP}}$$

F1 Score:

Corresponde a un promedio armónico entre Precision y Recall, este permite mostrar un desempeño general del modelo [70]. Se utiliza la siguiente fórmula para calcularlo.

$$\boldsymbol{F1\ Score} = \frac{\boldsymbol{2 * Recall * Precision}}{\boldsymbol{Recall + Precision}}$$

Curva ROC y AUC:

Una curva ROC (Receiver Operating Characteristic) es un método gráfico para evaluar el desempeño de un modelo de clasificación binaria. Esta curva representa el intercambio entre las siguientes tasas [83].

- Tasa de verdaderos positivos (TVP), el cual utiliza la misma fórmula que Recall.

$$TVP = \frac{VP}{VP + FN}$$

- Tasa de falsos positivos (TFP).

$$TFP = \frac{FP}{FP + TN}$$

Graficar esta curva permite entender el comportamiento del modelo, a medida que el TVP es mayor, el modelo tiene un mejor rendimiento, por el contrario, mientras mayor sea el TFP, el modelo tendrá un peor desempeño. Este comportamiento lo toma en cuenta el cálculo de AUC (Area Under Curve), que representa el área bajo la curva. Mientras más pequeño sea el TFP y más grande el TVP, la curva tenderá a posicionarse en el lado superior izquierdo del gráfico y esto hará que el área bajo la curva sea mayor. Este indicador toma valores entre 0 y 1 al igual que el resto de las métricas mencionadas anteriormente.

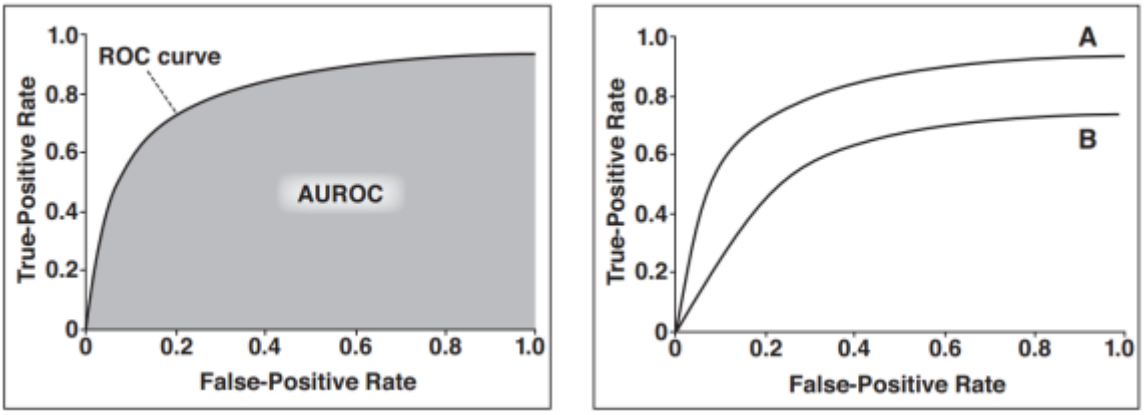

Figura 1.6: (a) Gráfico que muestra la curva ROC. (b) Mismo gráfico que muestra el área bajo la curva AUC. Fuente: Handelman et. al., 2019 [69].

Evaluación de resultados entre variantes de algoritmos: La evaluación entre las variantes de algoritmos permiten discernir y buscan explicar qué variante entrega mejores resultados dado un modelo. Algunos ejemplos vistos en [30] pueden ser vistos a continuación:

1. **Inception Score (IS):** Desarrollado por uno de los creadores del algoritmo GAN en [31], se utiliza para obtener información sobre la calidad del generador entrenado. Para calcular el IS, se entrena un clasificador en la data de testeo y es utilizado para determinar la entropía del modelo. Esto permite comparar diferentes variantes de GANs.

2. **Euclidean Distance:** Se utiliza esta distancia para evaluar cuán similar es la data generada a la data de entrenamiento. Esto se logra calculando dichas distancias con respecto a los datos reales y comparándolas. Al revisar los resultados de estas

distancias se puede obtener información sobre si el generador está imitando los datos desde la base de testeo o si está generando entradas nuevas.

3. **Sliced Wesserstein Distance (SWD):** Esta distancia es similar a la Wesserstein Distance, pero es más fácil de computar al reducir la dimensionalidad del problema en slices o pedazos de sólo una dimensión. Según [32], una métrica con un bajo valor de SWD indica que las dos distribuciones son similares en su apariencia y la variación de sus muestras.

1.4.6 Interpretabilidad versus funcionalidad de los modelos

Al trabajar con modelos de aprendizaje automatizado, un concepto clave es la interpretabilidad de los resultados que se obtienen [87]. Por un lado, el resguardo sobre la calidad de los datos utilizados debe ser prioridad al momento de entrenar estos modelos, pero esto no asegura que los resultados puedan ser explicados de una manera que permita obtener valor de ellos. Este concepto ha sido debatido en el último tiempo, generando investigaciones como en [86], puesto a que existe un intercambio entre el rendimiento de un modelo y su interpretabilidad. Como regla general, un modelo más sofisticado y con mejor rendimiento, tiende a poseer una menor explicabilidad que aquellos modelos más sencillos.

Este trabajo de título no es ajeno a este debate, sobre todo al momento de trabajar con datos de estudios clínicos, en donde si bien se busca conocer los fenómenos de la manera más certera posible, la interpretabilidad de los resultados es aún más importante y un factor limitante al momento de generar investigación en este entorno. Como se menciona en [86], los avances de los últimos años apuntan a evitar modelos denominados “Caja Negra”, en donde sólo se conocen los inputs y outputs del modelo, pero se desconoce el procedimiento para llegar a estos resultados. Con el fin de revisar estas diferencias entre algoritmos con distintas interpretabilidades es que se escogieron dos algoritmos de redes neuronales y uno estadístico para trabajar en esta memoria.

1.5. Metodología

El principal aporte dentro de este trabajo de título es poder crear una metodología en base a la experiencia obtenida en el proyecto piloto ADDAI, que permita a estudiantes e investigadores generar datos sintéticos para utilizar modelos de aprendizaje de máquinas en casos en los que no tienen acceso a la información o cuando no existen suficientes datos para entrenar estos modelos. Es interesante detallar este proceso de creación de datos falsos probando la mayor cantidad de algoritmos, para poder discernir qué herramienta utilizar dependiendo de las características de la base de datos con la que se trabaje.

Es por esta razón que la metodología de esta memoria será compuesta de cinco partes principales: conocimiento del proceso, generación de datos, evaluación de los modelos, impacto económico y propuesta de servicio. Por otro lado, independiente de si los resultados de este trabajo de título son positivos o no, aportarán conocimiento sobre qué camino seguir en el futuro para utilizar datos sintéticos.

- **Conocimiento del proceso:** Se busca conocer el proceso real y el manejo de datos que hay detrás de la toma de los diferentes exámenes del proyecto de diagnóstico de depresión o Alzheimer, es por esto que no es suficiente buscar información sobre este proceso, sino que es necesario entrevistar a las personas involucradas, tanto a los funcionarios que realizan los exámenes como los investigadores quienes serán los futuros usuarios de estos datos. Es por estas razones que el primer paso dentro de la metodología será la realización de entrevistas a ambos actores mencionados anteriormente.

- **Generación de datos:** La segunda parte de la metodología corresponde a la generación de datos, lo que significa poner en práctica todos los algoritmos y variantes posibles con respecto a múltiples bases de datos, para poder determinar cuál es el método óptimo para generar data sintética. Se utilizarán los algoritmos mencionados anteriormente (GAN, VAE y GC), posiblemente junto a algunas variantes de estos a través del software Python. Esta etapa sólo se centrará en la generación de datos. En este proceso es muy relevante describir las características propias de las bases de datos con las que se trabajará, con el fin de poder entender qué algoritmo es óptimo dependiendo de las condiciones iniciales del problema.

- **Evaluación de los modelos:** Posterior a este proceso, la tercera parte de la metodología de esta memoria corresponde a la evaluación de la generación de datos. Si bien algunos algoritmos como los GANs poseen un discriminador de resultados integrados dentro de su lógica, es interesante poder medir a todos los

algoritmos con métricas similares. Dentro de esta etapa, se entrenarán modelos de clasificación, que permitan medir el rendimiento de la generación de datos. Se aplicarán también test estadísticos que permitan diferenciar las muestras sintéticas de las reales, y qué tan bien pueden distinguir estas muestras dependiendo de los modelos generadores de donde provengan.

- **Impacto económico:** El cuarto paso dentro de la metodología consiste en medir el impacto económico de la generación de datos, conocer el valor del proceso de búsqueda de pacientes, toma de exámenes y evaluación de resultados para poder estimar el aporte económico que pudiese tener la creación de data sintética dentro del mundo de la investigación. Es importante destacar en este punto, que existe un valor no percibido o desperdiciado al no poder crear modelos de aprendizaje de máquinas en situaciones que se requieran. Se buscará poder estimar este costo también.

- **Propuesta de producto o servicio:** Finalmente, la última etapa de esta memoria corresponde a realizar una propuesta de cómo esta generación de datos podría ser un servicio o producto, especificando el aporte de valor que pueden tener, las situaciones en las que son útiles y los recursos necesarios para su aplicación.

Estos cinco puntos mencionados anteriormente convergen para dar pie a una metodología de generación de datos virtuales, la cual permita a estudiantes e investigadores acceder a la información que hay detrás de la data, respetando la privacidad de los pacientes y evitando restricciones de costos y tiempo para poder ejecutar modelos que requieran una gran cantidad de datos.

1.6 Alcances

Las aplicaciones dentro de este trabajo de título pueden ser muchas, existen muchos tipos de bases de datos, dentro de la literatura existe una gran cantidad de algoritmos generadores de datos y sus variantes, y diversas métricas de evaluación de desempeño. Es relevante entonces definir el alcance de esta memoria a continuación.

Los datos con los que se trabajarán provienen del proyecto de diagnóstico diferencial entre depresión y enfermedad de Alzheimer, los cuales fueron realizados a sujetos voluntarios por el Laboratorio de Neurosistemas. Con el fin de poder generar una metodología que aborde varios tipos de datos, se ha decidido intentar generar datos en los resultados de los tres exámenes con los que se trabaja dentro de este proyecto, los cuales corresponden a exámenes de movimiento ocular, electroencefalograma en al menos un canal y en los datos relacionados al test de navegación.

Con respecto a los algoritmos generadores de datos que se utilizarán para la creación de data sintética, se buscará abordar al menos tres algoritmos, los cuales corresponden a generación de datos a través de Gaussian Copula, Generative Adversarial Networks y Variational Autoencoders. Se escoge GC ya que permite una fácil interpretación al ser un modelo estadístico y será utilizado como la base para evaluar el resto de los modelos más complejos. La idea es poder cuantificar la mejoría de los algoritmos GAN y VAE en la generación con respecto a la pérdida de explicabilidad del modelo, la cual es característica en la utilización de redes neuronales, en otras palabras, qué tan superior son los datos sintéticos de estos modelos complejos de explicar, versus la generación de un modelo más simple. Es posible incorporar alguna variante de GAN en reemplazo de su algoritmo en estado puro, esto depende de qué tan inestable sea el entrenamiento de esta herramienta.

Con respecto a la evaluación de los modelos de generación, se buscará poder explicar las diferencias en los resultados obtenidos, por lo cual se espera contar con la mayor cantidad de información respecto a las características de los datos. Por otra parte, es más relevante poder medir la diferencia entre los modelos que entre algunas variantes de algoritmos, por lo que para esta etapa, se contará con tres métricas de evaluación; la clasificación utilizando datos sintéticos, clasificación mixta con data real y sintética y testeos con algoritmos discriminadores.

Al no poder testear o comparar diferentes variantes de los algoritmos GAN y VAE, queda fuera de análisis la integración de estas herramientas. Se decide no evaluar en una primera instancia estas variantes ya que tienden a ser muy dependientes de los modelos de datos, dependiendo de la particularidad de cada caso y complejizando la generalización de su aplicación.

Capítulo 2

Estado del Arte

Uno de los problemas usuales a los que se enfrentan investigadores y estudiantes, sobre todo al momento de realizar investigaciones en torno al área de la salud, es la falta de datos, ya sea por el difícil acceso a esta información, o por la escasez de exámenes específicos realizados a pacientes como para entrenar algún modelo de aprendizaje de máquinas. A continuación, se especificarán algunos problemas asociados a la falta de datos, cómo se utilizan los modelos de aprendizaje de máquinas en el área de la salud y finalmente cómo han solucionado en la literatura este problema.

2.1 Problemas asociados a tener escasez de datos

Cada día es más sencillo utilizar técnicas complejas de aprendizaje de máquinas, como aplicar redes neuronales para clasificar elementos o generar imágenes. Estos algoritmos, cada vez más complejos, requieren a su vez para un correcto funcionamiento una gran cantidad de datos, ya que de esta forma son capaces de detectar comportamientos diferentes entre grupos estudiados y generalizar las soluciones que encuentren.

Este problema consiste en cómo afecta la carencia de datos en áreas donde no necesariamente se puede generar u obtener esta información, dadas restricciones de costo, tiempo o espacio para poder crear datos. Las áreas típicamente afectadas por este fenómeno corresponden al área de la psicología, la medicina y la geología. Dentro de este trabajo de título, se enfocarán los esfuerzos para mostrar cómo impacta al área de la medicina, y cómo la costosa generación de exámenes tanto en tiempo como en recursos, fuerza a investigadores a trabajar con bases con pocos registros, lo que a su vez compromete los resultados de las investigaciones.

Los problemas asociados a tener pocos datos han sido nombrados recurrentemente en la literatura, y corresponden principalmente a problemas de generalización de resultados y de desbalanceo de clases al momento de tener que separar la base de datos en base de entrenamiento y de prueba, herramienta muy utilizada al tener que entrenar modelos de aprendizaje de máquinas.

Algunas publicaciones apuntan a cómo el poder estadístico se ve afectado al momento de tener pocos datos [33], y cómo esto reduce la posibilidad de encontrar un efecto real. El poder estadístico es la probabilidad de que un test de hipótesis encuentre un efecto cuando realmente hay un efecto que encontrar. Este permite mostrar que mientras más pequeño sea el efecto buscado, mayor será el número de muestras que se necesitan para poder encontrar y demostrar realmente el impacto que tiene el efecto estudiado.

Por otro lado, en [34] se ha mencionado que los modelos entrenados con pocos datos tienden a ser más imprecisos y con peor rendimiento, complicando la extracción de información relevante desde bases de datos con pequeñas cantidades de datos. Esto se explica con mayor detalle en [35], donde se muestran que las brechas o la distancia entre los datos observados hacen que la información que se extrae de estos sea incompleta y no represente a la población general, lo cual a su vez dificulta entrenar herramientas predictivas. La siguiente figura ilustra de manera más sencilla el concepto de brecha de información dada por la falta de datos.

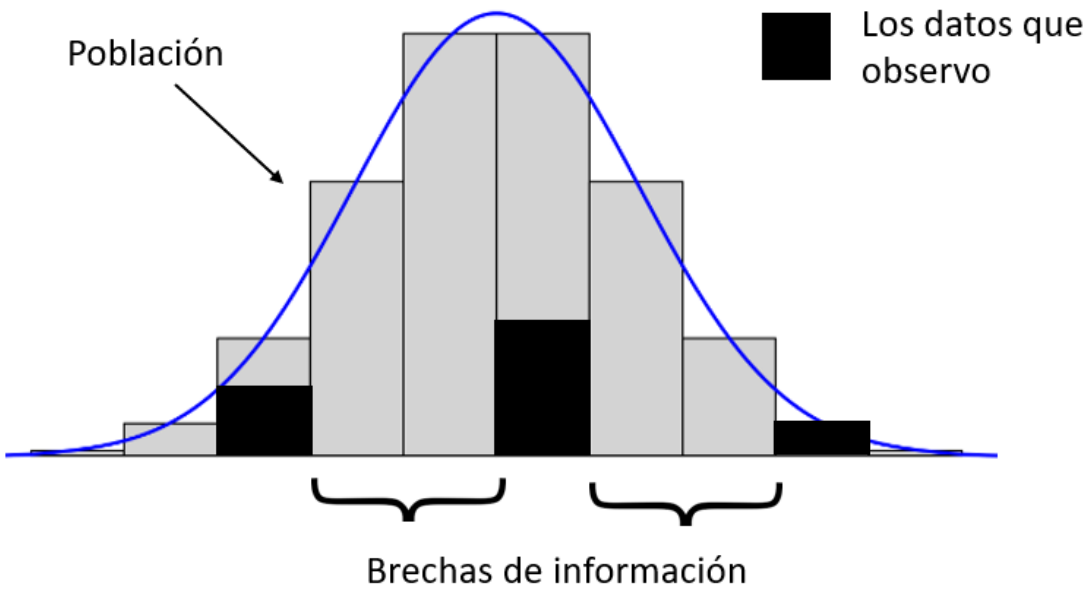

Figura 2.1: Distribución de una base de datos con pocos datos relativa a su población. Basado en [35].

Es relevante mencionar, que dentro de la literatura también se habla de cómo la incapacidad de obtener un resultado dada la falta de datos dentro de una investigación afecta negativamente a las personas que la financian [50]. Esto tiene una gran importancia al momento de manejar proyectos financiados por fondos públicos, y cómo se pueden desperdiciar estos recursos, generando una pérdida social.

2.2 Utilización de herramientas de aprendizaje de máquinas en el área de la salud

El uso o la discusión sobre aprendizaje de máquinas o inteligencia artificial en el área médica tiene sus raíces en la etapa inicial de estas aplicaciones [36], y por antigua que sea esta discusión, sigue siendo un tema relevante para muchos investigadores en la actualidad.

Las herramientas de aprendizaje de máquinas han tenido un gran reconocimiento en el área de la salud, visto en varias aplicaciones mencionadas a lo largo de esta memoria, pero aún existen impedimentos para poder utilizar el máximo potencial de estas herramientas. La interpretabilidad de las soluciones que entregan los algoritmos es de gran importancia al momento de tener que utilizar esta información para generar un diagnóstico [37], características que algunos algoritmos más sofisticados carecen, como los que utilizan redes neuronales, transformándose en una “caja negra” o algo de lo que no se tiene explicación o un por qué.

Existe un gran debate frente al intercambio que sufren los modelos entre explicabilidad y rendimiento, ya que los modelos más simples y fáciles de interpretar tienden a tener un peor desempeño frente a otros más complejos. Es por esta razón que investigadores y científicos han tenido que encontrar un balance que les permita explicar fenómenos de manera certera a través de modelos adecuados. Este debate es aún más relevante al momento de utilizar estas herramientas para apoyar diagnósticos, ya que la consecuencia de que un modelo se equivoque es potencialmente muy alta.

Es relevante también poder conocer el uso de técnicas de aprendizaje de máquinas en particular en el área de salud mental, y cómo han ido evolucionado con el tiempo. Como se ha visto en [38], de la mayoría de las publicaciones científicas relativas a aprendizaje de máquinas en esta área, el 63% corresponde a modelos que apoyan el diagnóstico y la detección, categoría a la cual el proyecto ADDAI pertenece. Dentro de esta categoría, la mayoría de los estudios utilizan datos provenientes de una resonancia magnética, electroencefalograma o tomografía por emisión de positrones.

Los modelos de aprendizaje de máquinas utilizados dentro de esta área han mostrado resultados prometedores, sin embargo, un gran problema es la inconsistencia de las pruebas realizadas en un laboratorio, en contraste del comportamiento real que pueda tener la persona fuera de ese espacio. Se destaca que en ningún punto los resultados de un modelo pueden reemplazar un diagnóstico, sobre todo en casos como Alzheimer o depresión, ya que las pruebas para diagnosticar a estos sujetos corresponden a pruebas didácticas, en las que el especialista realiza cuestionarios y entabla conversaciones no sólo con los pacientes, sino que con la familia también. Es por esta razón es que se hace un hincapié en enfocar los esfuerzos en estudiar y entrenar modelos que utilicen

información más sencilla y que pueda ser accesible para todos, ya que, como se mencionó anteriormente, la mayoría de las investigaciones se realizan en neuroimagen, la cual depende de una resonancia magnética, y esta herramienta no está a disposición de todo el público.

2.3 Generación de datos en la literatura

Actualmente existen diferentes métodos para poder evitar el problema de tener pocos datos, algunos corresponden a escoger modelos más sencillos, a eliminar valores extremos u outliers y otros a entrenar modelos utilizando transferencia de aprendizaje, entrenando modelos con datos de libre acceso y luego adaptando los resultados para utilizarlos con datos locales. Ninguna de estas soluciones es tan eficiente como la generación de datos sintética, ya que permite a los investigadores trabajar con información específica y local, evita eliminar datos cuando ya son escasos y les permite aplicar los modelos más complejos, los cuales tienden a tener mejores resultados en términos de precisión.

Es por estas razones que se ha decidido estudiar la opción de extender las bases de datos que se poseen generando datos sintéticos para este trabajo de título, algunos ejemplos de cómo se realiza este proceso será visto a continuación.

Existen métodos que generan perturbaciones en los datos agregando ruido a la base de datos original, como rotaciones, eliminación de datos o aplicación de transformaciones lineales [47]. Para producir bases de datos más diversas, que busquen entrenar modelos de clasificación más generalizables, se han utilizado métodos que agregan ruido desde alguna distribución conocida, este mecanismo ha sido utilizado también para anonimizar datos, como se ha visto en PrivBayes [48].

Por otra parte, existen modelos más complejos que capturan las relaciones y distribuciones de las variables, estos modelos pueden ser construidos específicamente para cada base de datos o inferido desde los datos a través de modelos como redes Bayesianas o redes neuronales [12]. Generative Adversarial Networks o GAN [20] han sido de las redes neuronales más utilizadas para estos fines, siendo el algoritmo más popular dentro de la literatura, sobre todo para poder replicar imágenes ya que es un algoritmo versátil, al poder cambiar su función objetivo, y la construcción de su modelo es más robusta ya que consta de dos partes, un generador y un discriminador.

Existen también modelos estadísticos que permiten la generación de datos sintéticos, como Synthetic Minority Oversampling Technique (SMOTE) el cual fue diseñado para evitar el problema de tener clases desbalanceadas al momento de entrenar modelos de clasificación insertando datos generados virtualmente [49].

En resumen, dentro de la literatura respectiva a la generación de datos se encuentran modelos que utilizan redes neuronales, modelos estadísticos, modelos que agregan perturbaciones, y modelos que anonimizan los datos. Todos estos modelos buscan solucionar problemas relacionados a la falta de datos, como lo son sesgos, desbalanceo de clases y problemas al entrenar modelos complejos. El desafío más grande dentro de estas investigaciones [12], es el problema de la caja negra, el cual se refiere a la explicabildiad de los modelos utilizados, cómo interactúan las variables dentro estos modelos y cómo identificar cuando un modelo está entendiendo los datos de manera correcta, sobre todo cuando se trabaja con parámetros complejos. Algunas aproximaciones para entender de mejor manera estos modelos es a través de la aplicación de modelos de árboles de decisión [51] o modelos que muestren gráficamente las distribuciones probabilísticas.

2.4 Arquitectura de los algoritmos

Como ha sido mencionado en el marco conceptual del presente informe, se ha decidido trabajar con tres algoritmos, GAN, VAE y GC. Esto con el fin de poder medir el rendimiento de los primeros dos algoritmos, los cuales utilizan redes neuronales, en contra de un algoritmo estadístico con mayor poder de interpretación, como lo es GC. En este apartado se mostrará la arquitectura de los algoritmos utilizados.

2.1.1 Conditional Tabular GAN (CTGAN):

Antes de explicar cómo funcionan estos algoritmos, es relevante mencionar el por qué se genera una distinción al llamarse tabular con respecto a otros algoritmos GANS. Los datos tabulares son aquellos datos dispuestos en el formato de una tabla, donde cada celda representa la intersección de una columna y una fila y esta estructura ordenada permite almacenar y encontrar información de manera más sencilla. Como se menciona en [14], los algoritmos GANs han sido utilizados ampliamente en la literatura para replicar imágenes o texto, pero en los últimos años se ha avanzado para utilizar estas herramientas en datos tabulares con el fin de habilitar investigación en campos donde este tipo de datos escasean.

Los problemas asociados a generar datos tabulares mediante este tipo de algoritmos se exploran también en Bourou, S. et al. [14] dentro de los cuales se menciona la dificultad de utilizar diferentes tipos de datos, desbalanceo de clases y trabajar con distribuciones que no son Gaussianas, las que provocan que el proceso de aprendizaje se vea comprometido al utilizar redes neuronales.

Es por estas razones que se decide utilizar CTGAN [52] que es un modelo basado en Generative Adversarial Networks, el cual permite crear entradas sintéticas desde una distribución utilizando redes neuronales. El diseño condicional de este algoritmo

generador, en conjunto con su entrenamiento a través de muestreo, le permiten evitar errores al trabajar con distribuciones no Gaussianas y multimodales, en conjunto con poder generar datos desde columnas discretas desbalanceadas.

El generador de los algoritmos GANs es entrenado tradicionalmente utilizando un vector de muestreo proveniente de una distribución normal multivariable. Al entrenar este en conjunto con un algoritmo discriminador, se obtiene un mapa determinístico que une la distribución de los datos con esta distribución normal multivariable. Este método de entrenamiento no cuenta con el desbalanceo que puede existir en las columnas categóricas dentro de la base, por lo que estos datos tenderán a estar poco representados en la base sintética final.

La solución propuesta por Lei Xu et al. [52] al problema del desbalanceo de las variables categóricas corresponde a un cambio en la arquitectura del algoritmo GAN denominado CTGAN, como se muestra en la Figura 4.1, entregándole un vector condicional con las variables categóricas dentro de la base al generador, el cual aprende las diferencias que existen entre las categorías y cómo se comportan el resto de las variables respecto a ellas. Con este cambio, CTGAN puede explorar las relaciones que existen entre las categorías y generar datos más certeros.

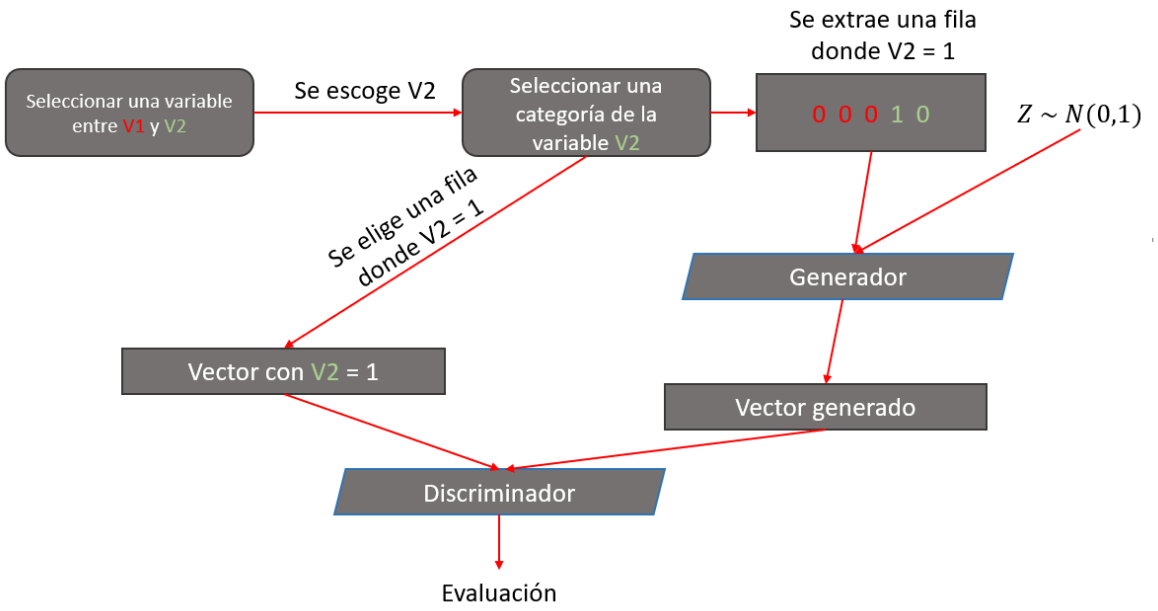

Figura 4.1: Arquitectura del algoritmo CTGAN. Esta construcción le permite generar muestras para cada categoría en las variables. Adaptado desde [52].

2.1.2 Tabular VAE (TVAE):

Variational autoencoder o VAE son otro modelo generativo que utiliza redes neuronales para comprimir y descomprimir los datos, cambiando de dimensionalidad. Mediante la

repetición de este proceso, el modelo aprende las relaciones que tienen las variables latentes. En [52] se adapta este modelo para funcionar con datos tabulares cambiando la función de pérdida del modelo. TVAE utiliza dos redes neuronales, una para comprimir y otra para descomprimir los datos, las cuales se entrenan utilizando la función de pérdida ELBO (evidence lower-bound), de igual manera que D. Kingma en [28] cuando describe por primera vez el algoritmo VAE.

2.1.3 Gaussian Copula (GC)

Existen diferentes métodos para generar datos de manera tabular, sin embargo, la mayoría de estos utilizan el mismo principio el cual corresponde a representar todos los datos numéricos como distribuciones y covarianzas. Estos conceptos describen los valores dentro de una columna y la dependencia entre columnas, respectivamente. Al unir esta información, se obtiene un modelo descriptivo de toda la tabla.

Un modelo generativo necesita conocer la forma de las distribuciones para cada columna numérica dentro de la base de datos, lo cual es complejo de calcular para bases extensas. Es por esta razón que se utilizan aproximaciones para estimar estas distribuciones, asumiendo que sus valores pertenecen a distribuciones Gaussianas uniformes, Beta o Exponenciales. Junto con esto, este modelo debe también conocer la covarianza de los datos, pero en algunos casos, la forma de la distribución puede afectar las estimaciones de covarianza.

Es por estas razones que en [53] se propone la utilización de Gaussian Copula, ya que evita el sesgo producido por la forma de la distribución, ya que convierte todas las distribuciones en normales y las estandariza en un rango entre 0 y 1.

Al poder generar una matriz de covarianza y conocer las distribuciones de cada columna, el modelo es capaz de adquirir toda la información de la tabla de manera compacta, y es posible utilizarla para generar muestras sintéticas. La Figura 4.2 ejemplifica de manera sencilla la arquitectura de GC.

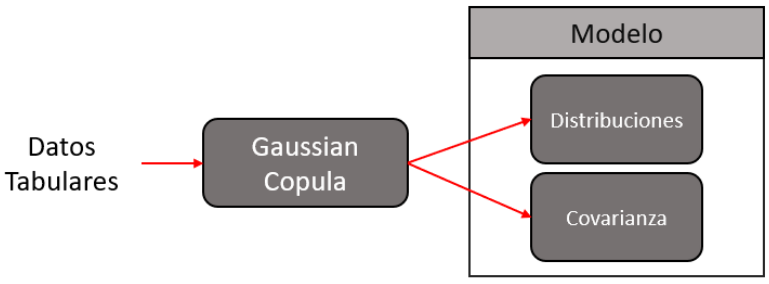

Figura 4.2: Arquitectura simplificada del algoritmo Gaussian Copula. Basado en [53].

Capítulo 3

Generación y manipulación de datos reales

El proyecto ADDAI es un proyecto financiado por el Fondo de Fomento al Desarrollo Científico y Tecnológico (FONDEF) que está siendo desarrollado por el Web Intelligence Centre y la Facultad de Medicina de la Universidad de Chile. Este proyecto busca poder generar una herramienta multimodal, es decir, una herramienta que a través de varias fuentes de información recaudadas de los pacientes de manera simultánea, les permita a investigadores determinar si un sujeto padece de Enfermedad de Alzheimer o Depresión. Esto puesto a que el diagnóstico realizado por médicos especialistas como neurólogos o psiquiatras tiende a confundir ambas enfermedades en sus etapas iniciales, ya que el deterioro cognitivo y sus síntomas tienden a ser similares. [39]

Los exámenes realizados dentro del contexto del proyecto ADDAI son explicados en detalle a continuación:

- **Test de Navegación:** Consiste en un juego ejecutado en un computador, en donde el sujeto es invitado a moverse virtualmente utilizando el teclado para encontrar una plataforma. Dentro de este programa se registran los siguientes datos:

    1. Tasa de error: Frecuencia de errores ocurridos.
    2. Latencia a la plataforma: Tiempo empleado para encontrar la plataforma.
    3. Tiempo de reacción: Tiempo de retraso inicial antes de que los sujetos comienzan a escanear el grupo para llegar a la plataforma.
    4. Tiempo de descanso: Tiempo dedicado por los pacientes para encontrar la plataforma.
    5. Velocidad: Indicador que muestra la rapidez con la que se mueve el sujeto.
    6. Entropía: El grado en que la búsqueda se centra en la ubicación de la plataforma.
    7. Distancia recorrida: Distancia recorrida por los sujetos mientras navegan dentro del ambiente virtual.
    8. Correlación entre las tasas de latencia: La prueba se repite, por lo que los sujetos que tengan una menor tasa de latencia demostrarán un proceso de aprendizaje motor.

- **Electroencefalograma (EEG):** Durante el test de navegación, y para medir la amplitud y propagación de la señal eléctrica en el cerebro del paciente, se utiliza una gorra con electrodos que permite recoger estas señales. Utilizando este equipo sumado a un software que permite traducir estas señales en el rendimiento

del sujeto durante las pruebas, se generará un informe automático del desempeño del paciente.

Se utilizará un sistema de electroencefalografía de 32 + 8 canales (32 canales EEG, más 8 canales externos). La señal analógica adquirida se filtrará entre 0 (CC real) y 1000 Hz antes de su conversión, y se generarán muestras a 2048 Hz, lo que implica una cantidad de datos de 2048 entradas por cada segundo de duración del test.

- **Eye Tracking o Registro de Movimiento Ocular:** Utilizando una cámara que permite grabar los movimientos oculares, se determinan el número de movimientos y fijaciones, que son un grupo de puntos de mirada que están muy cerca entre ellos e indican que la persona se concentra y procesa lo que está observando. Con esto se construye un mapa de calor con los movimientos de los sujetos mientras ejecutan el test de navegación.

  El seguimiento ocular se logrará empleando un sistema Eyelink® 1000, que digitaliza y almacena datos de seguimiento ocular en un archivo binario convertible a texto, desde el cual se obtendrá la posición bidimensional de las pupilas a una frecuencia de 500 Hz, o sea, 500 entradas por cada segundo de duración del test.

Como se ha mencionado con anterioridad, estos exámenes tienden a ser costosos y longevos, de aproximadamente una hora por paciente, lo que implica que tener una gran cantidad de estos exámenes es una tarea compleja. Se ha decidido poder estimar de manera precisa el proceso, tanto en costos monetarios como temporales, para poder conocer el posible impacto de la generación de datos en este proyecto en particular.

Para esto, se ha decidido entrevistar a las personas encargadas de realizar los exámenes, y también a encargados de procesar y manejar los datos para construir modelos de aprendizaje de máquinas, con el fin de entender los tiempos y costos que existen detrás de estas etapas.

3.1 Proceso de Búsqueda de Pacientes

Encontrar a una persona que padezca de una enfermedad, que posea un diagnóstico clínico, pertenezca al grupo etario al cual apunta el proyecto y que además esté dispuesta y en condiciones de participar en los exámenes antes descritos es complejo, dado que es un grupo muy reducido de personas. Es por esto que previo al proceso de examinado, existe un proceso de búsqueda de pacientes, el cual consiste en invitar e incentivar a pacientes a participar en los exámenes.

Dentro del sistema del Hospital Clínico de la Universidad de Chile (HCUCH), cuando un neurólogo o psiquiatra en su consulta detecta posibles síntomas de depresión o Enfermedad de Alzheimer en su consulta, invita al paciente a participar de estos exámenes. Si el paciente acepta, se realiza un contrato de consentimiento informado con el paciente y se coordina el día y la hora de la realización de los exámenes. Es importante destacar en esta etapa que se buscan a pacientes con deterioro cognitivo leve, depresión y pacientes de control. Actualmente este proceso de búsqueda de pacientes es llevado a cabo por sólo una neuróloga, pero esta pronosticado aumentar este número a 3 especialistas en el hospital clínico.

De la totalidad de los pacientes que participan de los exámenes relacionados al proyecto ADDAI, el 70% de ellos accede a través de este proceso. El 30% restante, son pacientes que asisten a estos exámenes ya que conocen el proyecto mediante afiches en el mismo hospital o a través de contactos familiares.

Para asegurar la participación de los sujetos en los exámenes, se cuenta con un incentivo que permite tener una mayor cantidad de pacientes para la muestra. Este incentivo corresponde a una consulta neurológica o psiquiátrica, escogida por el paciente y de manera totalmente gratis, en conjunto con tener un trato más directo con su clínico tratante, como poder contactar telefónicamente al especialista ante eventualidades, aunque no tengan relevancia dentro del estudio. Esta consulta según la página web del HCUCH [40], está avaluada entre $36.000 y $47.000 pesos.

3.2 Realización de los exámenes

Posterior al proceso de búsqueda de pacientes, y ya coordinando una consulta neurológica o psicológica, es que comienza la etapa de generación de exámenes reales, la cual consta de tres partes: La consulta con un especialista para poder diagnosticar al paciente y asegurar su participación en el estudio, el examen utilizando la herramienta multimodal, y finalmente una evaluación neuropsicológica, que permita confirmar el diagnóstico en caso de ser confuso. Como se ha mencionado anteriormente, los síntomas entre depresión y Enfermedad de Alzheimer son similares en sus etapas iniciales, lo cual lleva a especialistas a confundir el diagnóstico y es por esta razón que la confirmación del diagnóstico es necesaria.

Durante la consulta con un especialista, se realizan test para evaluar la memoria del paciente, se obtienen datos generales del paciente, como su situación socioeconómica, edad, historial clínico. Junto con esto se revisa que el paciente haya padecido de diabetes, hipertensión, accidentes graves o cirugías previas, ya que todos estos factores pueden producir un deterioro cognitivo sin tener necesariamente depresión o Alzheimer.

Posterior a la consulta, y con un diagnóstico clínico para cada paciente que asegure su participación en el grupo de estudio o de control, es que se realiza el examen utilizando

la herramienta multimodal. El funcionamiento de la herramienta depende de dos Técnicos de Enfermería de Nivel Superior (TENS) previamente capacitados, pero es posible utilizarla con sólo uno. Esta herramienta corresponde a tres pruebas en simultáneo, en donde a los sujetos se les sitúa en un ambiente virtual a través de un juego de computador, cuyo objetivo es encontrar una plataforma escondida. Mientras el paciente realiza la prueba, se mide la actividad electroencefalográfica a través de una gorra y el movimiento ocular con una cámara especial.

En esta etapa es importante destacar que el posicionamiento de la gorra en los pacientes es relevante, por lo que existe un proceso de calibración de en promedio 20 minutos, esta gorra tiene que ser instalada con un gel que permite recibir las señales desde el cráneo. Junto con esto, la cámara de movimiento ocular también requiere ser calibrada, cuyo proceso no tarda más de 1 o 2 minutos, pero este proceso tiene que ser repetido al menos cuatro veces por cada paciente para mantener la fidelidad de las imágenes obtenidas durante la prueba.

El o los técnicos no entregan mucha información sobre la meta del juego de la plataforma ya que es autocontenido, pero han existido casos en los que pacientes requieren ayuda para poder lograr el objetivo o entender cómo utilizar el programa.
La duración de la prueba de navegación dura aproximadamente una hora por paciente, oscilando entre 45 minutos y una hora y media el rango de tiempo que tardan las personas en ejecutar las tareas dentro del juego.

Terminando el examen, los pacientes pueden optar a un lavado de pelo si es que lo desean, este proceso es llevado a cabo por un enfermero externo al equipo médico encargado de la investigación. El lavado de pelo se le realiza a alrededor del 50% de los sujetos que pasan por el examen.

Para terminar el proceso de prueba de los pacientes, se les cita a una tercera evaluación, esta vez con un neuropsicólogo, el cual es un especialista en las aristas cognitivas de los pacientes. Se realizan mayoritariamente test de lápiz y papel que buscan medir el grado de memoria y aprendizaje que manejan los sujetos, estos test no cuentan como un diagnóstico, pero si permiten confirmar y diferenciar a pacientes que previamente padezcan de Enfermedad de Alzheimer o depresión. El costo asociado a esta prueba va desde los $70.000 a los $80.000 pesos chilenos a nivel comercial.

Para el proyecto ADDAI es relevante esta última etapa, ya que es el estándar contra el cual se pueden medir los resultados y la capacidad de diferenciar pacientes con Enfermedad de Alzheimer o depresión, ya que, frente a diagnósticos dudosos de especialistas clínicos, un neuropsicólogo realiza pruebas para poder apoyar al diagnóstico del neurólogo o psiquiatra.

3.3 Procesamiento de los datos

Una vez los exámenes de la herramienta multimodal son realizados, estos son almacenados en un servidor seguro ubicado en el HCUCH. A esta información sólo pueden acceder los especialistas a cargo de los pacientes y los investigadores del WIC con un permiso especial a través de un protocolo SFTP o Secure File Transfer Protocol, el cual requiere conexión a un servidor a través de un usuario y una contraseña.

Los datos del electroencefalograma provienen de un aparato de marca BIOSEMI, el cual es un instrumento destinado a la investigación en el área médica y que permite obtener y traducir los impulsos eléctricos y el movimiento de los ojos en información útil para los investigadores. Según su página web [39], los productos BIOSEMI no son dispositivos médicos para utilizar en el diagnóstico, ya que son libres de configurarlos dependiendo de las necesidades de investigación, utilizando también software de código abierto u open source. Los datos provenientes del EEG son series de tiempo muestreadas a 2048 Hz, lo que corresponde a 2048 datos o entradas por segundo de la prueba. Para poder interpretar esta información, es que se utilizan indicadores como los vistos en [41], dependiendo de la frecuencia de las ondas percibidas, los cuales serán descritos a continuación:

- **Ondas Gamma:** Son las ondas de mayor frecuencia, por sobre los 32 Hz, estas se originan en el tálamo, cerca de la corteza cerebral. Estas ondas han sido tratadas como ruido en varios estudios, pero otros afirman que son las ondas predominantes en estados de meditación. Algunos estudios vinculan la actividad irregular de las ondas gamma con la Enfermedad de Alzheimer y epilepsia [42].

- **Ondas Beta:** Corresponden al espectro de ondas de entre 12 y 32 Hz y provienen de varias ubicaciones de la corteza cerebral. Esta onda se divide en tres sub espectros denominados Lo beta (12-15 Hz) relacionado con la capacidad de enfoque y concentración introvertida, Beta 2 (15-22 Hz) relacionado con un alto nivel de energía y ansiedad y Hi Beta (22-32 Hz) relacionado con estrés y alta excitación. [43]

- **Ondas Alpha:** Son parte del rango de frecuencias entre 8 y 13 Hz y provienen de la zona occipital del cerebro. Representan la calma y coordinación mental, un estado de relajación en el cual la persona está descansando, pero no durmiendo. Estas ondas son predominantes cuando los ojos están cerrados y se atenúan cuando se abren. [44]

- **Ondas Theta:** Equivalen al rango de frecuencias entre 4 y 8 Hz y provienen del hipocampo. Estas ondas están presentes en adultos al momento de despertar, y

puede que no se presenten en algunas personas. Han sido conectadas con procesos de memoria y navegación. [45]

- **Ondas Delta:** Son las ondas de mayor amplitud y menor frecuencia, por bajo los 4 Hz. Estas ondas son asociadas con estados de sueño profundo y su presencia en adultos despiertos puede indicar un problema en las funciones cerebrales. Se ha detectado que actividades como caminar dormido o hablar ocurren cuando hay una alta actividad de ondas delta. [46]

Estos datos requieren ser filtrados, mediante la inspección de un neurofisiólogo clínico capacitado para evaluar la calidad de las señales adquiridas y eliminar registros anormales, como ritmos basales anormales o actividad epileptiforme. Los componentes de ruido asociado a los parpadeos, sacadas y algún otro movimiento ocular son detectados de manera semiautomática para cada paciente y son eliminados los registros.

Los datos de la cámara de movimiento ocular provienen de un hardware llamado Eyelink 1000, el cual digitaliza y almacena datos del movimiento de los ojos de los sujetos en un archivo binario en formato Eye Data Format (EDF) convertible a texto. Dentro de los datos que recauda este aparato está la posición bidimensional de las pupilas, el tamaño de estas, el tiempo, parpadeos, fijaciones y sacadas oculares. Esta información es relevante y permite el procesamiento y análisis de las señales del electroencefalograma cuando es necesario filtrar y analizar las señales asociadas a la fijación ocular durante la prueba de navegación virtual. Se estudian las variables relacionadas al comportamiento y movimiento ocular, movimientos sacádicos y parpadeos, en cantidad, duración y frecuencia durante el examen.

Estos datos corresponden a series de tiempo, las cuales se muestrean a 1000 Hz, lo que corresponde a mil datos por segundo que transcurre la prueba. Para poder interpretar esta información, es que las series de tiempo pasan por funciones, las que transforman estas series en indicadores, los cuales permiten interpretar los datos de manera más sencilla. Algunos de los indicadores corresponden al tiempo en que las personas fijan el ojo en cada cuadrante, la desviación estándar del movimiento en los ejes X e Y, y el promedio de dilatación de la pupila durante el examen. Son con estos indicadores que los científicos de datos ejecutan un modelo de clasificación asistida para poder determinar las diferencias entre el grupo de control y el de prueba.

Por otro lado, los datos provenientes del test de navegación son registrados por el mismo computador en el cual es ejecutado el juego, y estos corresponden a los mencionados anteriormente, como lo son el tiempo de latencia, de reacción, aprendizaje y demora hacia la plataforma. A diferencia de los datos anteriores, estos no se dividen de manera temporal, sino que se marcan las etapas en las cuales la persona realiza ciertas acciones

y también se dividen en cuatro, que corresponden a las veces que se repite la misma prueba.

3.4 Generalización del proceso

Es relevante entender que como el proyecto ADDAI sólo se utilizará de experiencia demostrativa dentro de esta memoria, la generalización de este proceso de generación de datos reales es de gran importancia, para poder entender los costos y tiempos asociados a este proceso. Es por esto que se ha separado el proceso en cuatro etapas principales, las cuales pueden corresponder a distintos proyectos dentro del área de salud que requieran la utilización de datos y tengan problemas con la cantidad de información que manejan.

1. **Búsqueda de pacientes:** Dada la falta de sujetos que participen en los estudios, se genera un proceso de búsqueda activa de pacientes, la cual se lleva a cabo dentro de la misma institución clínica utilizando las citas médicas con especialistas para promover la participación, con afiches o volantes informativos y a través de contactos familiares o cercanos.

2. **Confirmación del diagnóstico:** Esta etapa corresponde al correcto etiquetado de los pacientes y sus diagnósticos. En algunos casos una cita médica no es suficiente para poder confirmar el estado de los pacientes, debido a la necesidad de realizar exámenes complejos o dificultades dada la naturaleza de la enfermedad para llegar a un diagnóstico correcto.

3. **Generación de muestras:** La etapa de generación de muestras es donde los sujetos previamente etiquetados con un diagnóstico y pertenecientes a un grupo definido entre control y estudio, participan en las pruebas o exámenes clínicos necesarios para la investigación. Este proceso puede requerir la participación de los sujetos más de una vez, someterlos a múltiples pruebas y tener efectos secundarios para las personas involucradas. Los costos y tiempos asociados a esta tarea dependen de que tan extensa sea, los actores involucrados en esta y la cantidad de sujetos participando en el estudio.

4. **Procesamiento de los datos:** Una vez los datos son generados, estos deben ser procesados para extraer la información real que contienen. En esta etapa se crean variables de interés, se condensa la información mediante agregaciones de datos y se puede eliminar información que no sea relevante para los modelos que se busca entrenar. Los costos y tiempos asociados a esta etapa dependen de la naturaleza de los datos con los que se trabaja, ya que se requiere un mínimo de

conocimiento para poder tomar decisiones sobre la importancia de los datos con los que se trabaja.

Capítulo 4

Generación de datos sintéticos

Como se ha mencionado anteriormente, en el presente trabajo de título se intentará resolver los problemas asociados a la falta de datos a través de la generación de datos sintéticos. En la siguiente sección, se mostrará dicho proceso, junto con los modelos utilizados y las decisiones respecto a la arquitectura de estos. Durante este proceso se trabajará con las tres bases de datos utilizadas dentro del proyecto ADDAI, que corresponden a la información recibida desde el electroencefalograma, la cámara de movimiento ocular y el test de navegación. Se entrenarán tres modelos por cada base de datos, lo que permitirá generar en una primera instancia 9 bases de datos con datos sintéticos.

4.1 Características de las bases de datos

Dentro del marco del proyecto ADDAI utilizado como experiencia demostrativa en esta memoria, se almacenan datos desde tres fuentes principales, un electroencefalograma, una cámara de movimiento ocular y una prueba de navegación a través de un programa.

Los datos provenientes del electroencefalograma corresponden a series de tiempo, que se reciben a una frecuencia de 2048 Hz, lo que implica que se obtienen 2048 datos por cada segundo de duración de la prueba por cada paciente. Existe por lo tanto una excesiva cantidad de información por cada sujeto que rinde estos exámenes y es necesario condensar estos datos, para estos efectos, se implementan funciones que agrupan los datos por cada canal del electroencefalograma y por cada tipo de onda, las cuales se dividen por frecuencias. Estas agrupaciones corresponden a calcular promedios, varianzas y medidas estadísticas de posición de los resultados obtenidos durante la duración del test de navegación. Posterior a este proceso es que se obtiene una base de datos con sólo 38 entradas, una por cada paciente, y con 362 columnas, las cuales muestran los resultados calculados anteriormente mencionados.

Los datos provenientes de la cámara ocular también corresponden a series de tiempo, las cuales son recibidas con una frecuencia de 1000 Hz, lo que implica 1000 datos por cada segundo por cada paciente que realiza el examen. Junto con estos datos, se registran mapas de calor en forma de imágenes que muestran las posiciones más concurridas por los ojos de los sujetos, sin embargo, para efectos de este proyecto, sólo se utilizan los datos tabulares. Estos datos son construidos de igual manera que los provenientes del electroencefalograma, a través de funciones que condensan la información a través de transformaciones como promedios, varianzas y cálculos de distancias entre los movimientos oculares. Tras realizar la agregación de los datos, se obtiene una base de datos con 34 entradas, una por cada paciente y 11 columnas, las

cuales hacen referencia a la posición del ojo en un plano XY, la dilatación de la pupila, los tiempos en cada cuadrante, y la distancia recorrida por la pupila.

La tercera fuente de datos corresponde al test de navegación, el cual se ejecuta en un computador y es el mismo programa el que registra la información y la almacena. Estos datos también corresponden a series de tiempo, y de igual manera que los exámenes anteriores, son condensados a través de funciones lineales, agregando los datos para su utilización. Posterior a este proceso es que se obtiene una base de datos con 38 entradas, una por cada paciente y 7 columnas, las cuales muestran la distancia recorrida, latencia o demora hacia la plataforma y velocidad de navegación entre otros.

Es relevante destacar que, dentro de todas las bases de datos mencionadas anteriormente, existe una columna que indica el número del sujeto y si el sujeto es parte del grupo de control o el grupo de investigación. Actualmente el grupo de investigación cuenta solamente con pacientes con un diagnóstico de enfermedad de Alzheimer, ya que se espera que las diferencias en los datos sean más notorias y permitan el entrenamiento de los modelos con una mayor facilidad. Finalmente, se ha confeccionado una tabla que resume la información con la que se trabajará, es importante destacar que la mayoría de las variables dentro de las bases de datos corresponden a variables numéricas, excepto la variable que enumera al sujeto y señala a que grupo de estudio pertenece.

Tabla 4.1: Características de las bases de datos para cada tipo de examen. Elaboración propia.

	Electroencefalograma	**Cámara Mov. Ocular**	**Test de Navegación**
Cantidad de columnas	362	182	7
Cantidad de filas	36	34	38
Cantidad de variables categóricas	1	1	1

4.2 Implementación de los algoritmos generativos

Antes de ejecutar modelos de aprendizaje de máquinas, es necesario pasar por una etapa de transformación de los datos, esto con el fin de poder evitar sesgos y utilizar sólo la información relevante, para asegurar la calidad de los resultados al final del proceso. En esta etapa es posible crear nuevas variables y eliminar otras, ya que se busca trabajar

con la menor cantidad de información repetida posible para poder entrenar modelos más certeros. Dentro de este contexto, es natural cuestionar en qué etapa de este proceso de limpieza y preprocesamiento de la información debe entrar la generación de datos, ya que es posible generar datos en las bases de datos en todas las etapas dentro del preprocesamiento.

Según la pauta para generación de datos sintéticos de Fida Dankar [57], la mayoría de las investigaciones que utilizan datos sintéticos los generan con la data sin procesar, y después de las pruebas conducidas en su artículo, no hay beneficios asociados a preprocesar los datos antes de la generación de datos. Por esta razón es que se ha decidido trabajar en esta primera etapa con los datos en su estado puro en forma de base de datos, antes de ser procesados por los investigadores.

La implementación de estos algoritmos fue realizada a través de Python, mediante la utilización de librerías específicas que apoyan la utilización de redes neuronales, en conjunto con la librería pandas para el manejo de bases de datos. Las librerías serán detalladas a continuación.

Para la confección del modelo GAN, se utilizó la librería CTGAN [54], la cual permite entrenar la red generadora y discriminadora de manera sencilla, pero teniendo control sobre la cantidad de épocas, el tamaño del grupo de datos asignado a cada etapa, la velocidad de aprendizaje y si es que se utiliza la log-frecuencia de las variables categóricas. El modelo es entrenado utilizando los datos reales utilizando de hiperparámetros 500 épocas, un batch size de 20 filas y una arquitectura de cuatro capas tanto para el generador como el discriminador, junto con esto, se decidió realizar tres actualizaciones del discriminador por cada una del generador. Esto es una práctica común que se puede ver en la arquitectura de WGAN, en donde en ese caso en particular se utilizan 5 actualizaciones [24]. Posterior al proceso de aprendizaje que toma entre 4 y 20 minutos dependiendo de la cantidad de variables, es posible obtener datos sintéticos creando muestras desde el generador.

El modelo TVAE fue implementado a través de la librería SDV (Synthetic Data Vault) [55], esta permite de igual manera que la librería CTGAN, entrenar dos redes neuronales, teniendo control sobre la cantidad de épocas, el tamaño de los grupos y la velocidad de aprendizaje. Para entrenar estas redes, se decidió utilizar 500 épocas, un batch size de 20 filas y una arquitectura de 2 capas para las redes de compresión y de descompresión. Los tiempos de demora van entre los 2 a 8 minutos, dependiendo de la cantidad de variables de la base de datos. Posterior al entrenamiento, es posible obtener muestras desde este algoritmo generador.

Es importante destacar que las decisiones sobre los hiperparámetros de épocas y batch size son escogidas puesto a que se realizó un análisis de sensibilidad que será explicado más adelante en la etapa de evaluación de los modelos en 5.3, mientras que las

arquitecturas se definieron mediante pruebas de ensayo y error que apuntaron a buscar el mejor rendimiento de los modelos.

Por último, el modelo que utiliza el algoritmo Gaussian Copula, fue entrenado utilizando nuevamente la librería SDV, en particular sus funciones relacionadas con este algoritmo [56]. De la misma manera que en los casos anteriores, se entrena a este modelo con los datos reales crudos, y posterior a este proceso que toma entre 1 y 4 minutos dependiendo de la base de datos, se generan muestras de datos sintéticos. Al terminar este proceso, es que se obtienen 9 bases de datos, una por cada algoritmo generador y cada base de datos utilizada.

Posterior a esta primera generación de datos, y frente a la necesidad de entrenar el modelo de clasificación que se utiliza en el proyecto piloto ADDAI, es que se decidió generar datos nuevamente utilizando los datos ya procesados y en la forma de una base unificada. Esto debido a que al generar tres bases por separado se pierde la unicidad de cada paciente, haciendo que no concuerden los sujetos de control y de estudio en las bases de datos.

El procesamiento de los datos corresponde a la verificación de que cada sujeto tenga todos los datos de los exámenes que le pertenezcan, junto con la selección de las variables más relevantes para explicar la varianza dentro de cada examen, esta selección se realizó a través de una selección de características conducida en el proyecto ADDAI y se utilizarán las mismas variables obtenidas para trabajar en esta memoria. En las siguientes tablas se adjuntan las variables utilizadas.

Tabla 4.2: Variables seleccionadas provenientes del Test de Navegación. Elaboración propia.

Variable	**Descripción**
Avg_pathlengthRat	Promedio del ratio entre el largo del camino recorrido y el esperado
Avg_avgNavigVeloc	Velocidad promedio de navegación
Avg_plat_LatencRat	Promedio del ratio entre la latencia y el largo del camino esperado
Change_avgNavigVeloc	Cambio entre el primer y último intento en la velocidad promedio
Change_pathLengthRat	Cambio entre el primer y último intento en el ratio del largo del camino
change_plat_LatencRat	Cambio entre el primer y último intento en la latencia hacia la plataforma

Tabla 4.3: Variables seleccionadas provenientes de la cámara de movimiento ocular. Elaboración propia.

Variable	Descripción
Xstd_grouped	Desviación estándar del movimiento en el eje X
Ystd_grouped	Desviación estándar del movimiento en el eje Y
Cuad1	Movimiento en el primer cuadrante de la cámara
Cuad2	Movimiento en el segundo cuadrante de la cámara
Cuad3	Movimiento en el tercer cuadrante de la cámara
Cuad4	Movimiento en el cuarto cuadrante de la cámara
Dilmean_Group	Dilatación pupilar promedio
Dist_euc	Distancia euclidiana de los puntos

Tabla 4.4: Variables seleccionadas provenientes del electroencefalograma. Elaboración propia.

Variable	Descripción
AF3_Theta	Ondas Theta detectadas por el electrodo AF3
CP6_Beta	Ondas Beta detectadas por el electrodo CP6
Mean_F4	Promedio de los impulsos detectados por el electrodo F4
P8_Gamma	Ondas Gamma detectadas por el electrodo P8
PO4_Theta	Ondas Theta detectadas por el electrodo PO4
T7_Beta	Ondas Beta detectadas por el electrodo T7
Variance_PO4	Varianza de los impulsos detectados por el electrodo F4

Para entender de donde provienen las variables del electroencefalograma, se adjunta la siguiente figura que muestra la posición de los electrodos en el cráneo, que es de donde provienen las señales que se indican en la tabla 4.4.

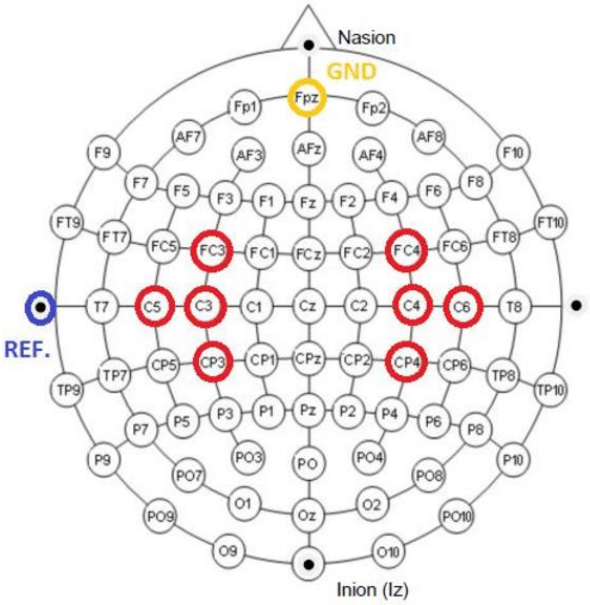

Figura 4.3: Posiciones de los electrodos al realizar un examen de EEG. Fuente: Vourvopolous et. al., 2015 [60].

Después de este proceso se obtiene una base de datos con 38 filas, las cuales representan a los sujetos, una variable categórica que indica la pertenencia al grupo de control o de estudio y las variables numéricas mencionadas anteriormente, resultando en 22 columnas en total. Es con esta base de datos que se procede a generar muestras sintéticas utilizando los mismos algoritmos generativos utilizados con anterioridad; CTGAN, TVAE y Gaussian Copula.

Se obtienen finalmente 6 bases de datos utilizando este método, tres con la misma cantidad de filas que la base original, siendo una por cada algoritmo generativo, y tres bases con 150 filas con el fin de poder aumentar la cantidad de sujetos. Al terminar todos los procesos generativos, se obtienen 15 bases de datos sintéticos a través de los algoritmos mencionados a lo largo de este trabajo de título, se esquematiza el proceso en la siguiente figura.

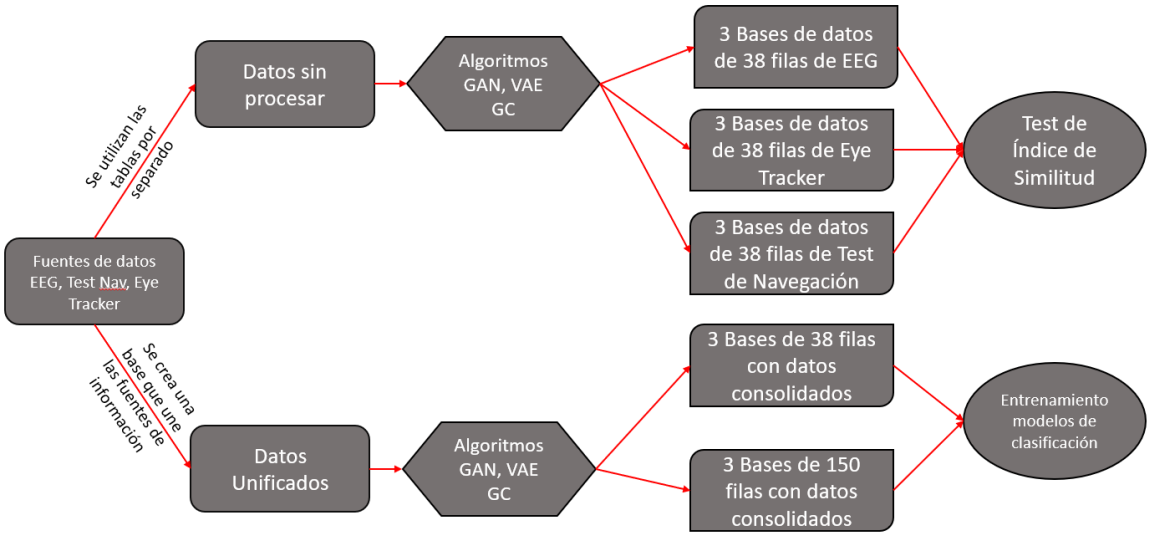

Figura 4.4: Esquema del proceso de implementación y su posterior evaluación. Elaboración Propia.

4.3 Indicadores de la generación de datos

Para ejemplificar el rendimiento de este proceso de generación se utilizarán dos métodos. Para las primeras nueve bases de datos generadas, se utilizará un histograma en una variable, el cual permite ver la distribución de esta y revisar las diferencias entre datos reales y sintéticos. Para esta etapa se escogió una variable al azar de las tablas 4.2, 4.3 y 4.4 por cada tipo de examen. Por otro lado, para las seis bases de datos restantes, que son las que se utilizarán más adelante para entrenar modelos de clasificación, se utilizará un gráfico de líneas que muestra el promedio y desviación estándar de seis variables, el que permite comparar los datos dependiendo de su origen.

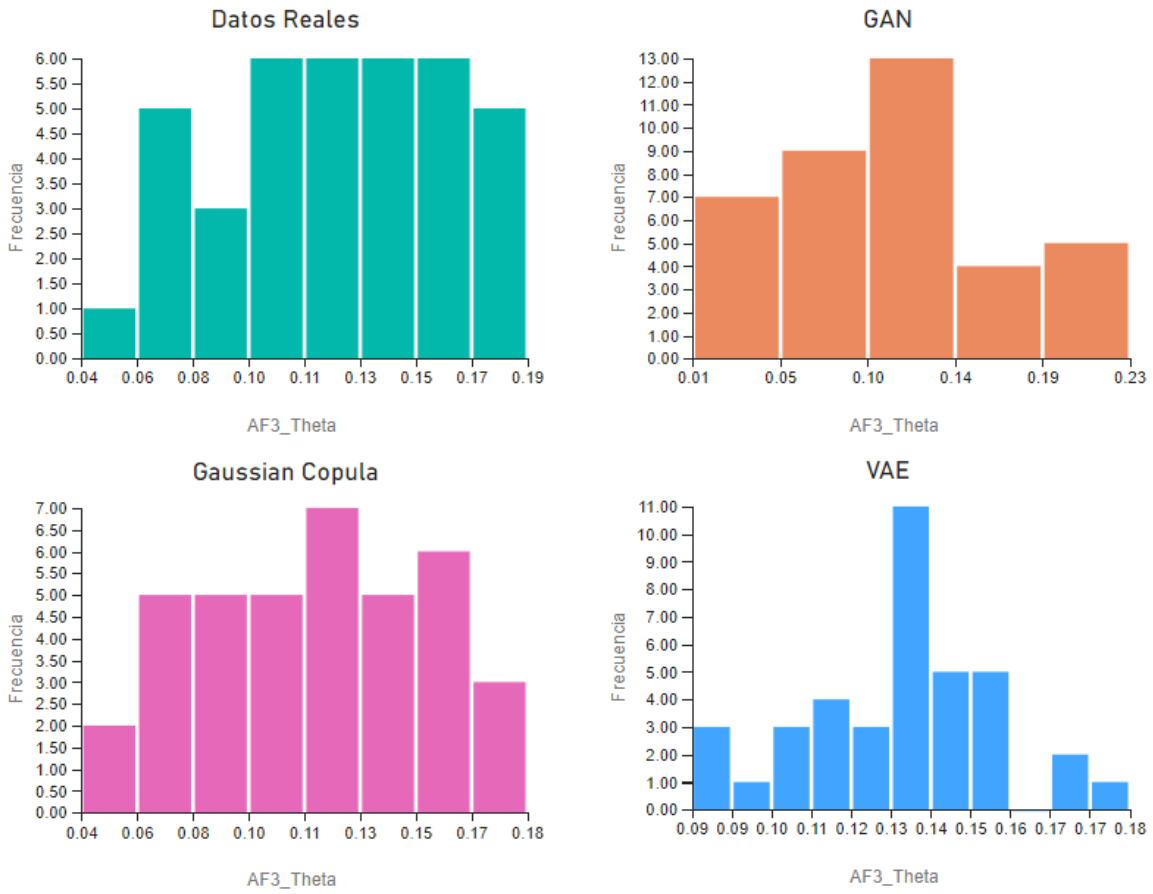

Figura 4.5: Distribución de la variable AF3_Theta (EEG) para los datos reales y los generados sintéticamente.

En la Figura 4.4 se puede ver que el algoritmo que mantiene una distribución más similar a la original corresponde a Gaussian Copula, mientras que los algoritmos VAE y GAN presentan grandes diferencias con respecto a los datos reales. Si bien la distribución de GAN es más cercana a la distribución original en comparación a VAE, también presenta valores que están fuera del rango de la variable real y esto hace que la calidad de los datos generados por este algoritmo se vea comprometida.

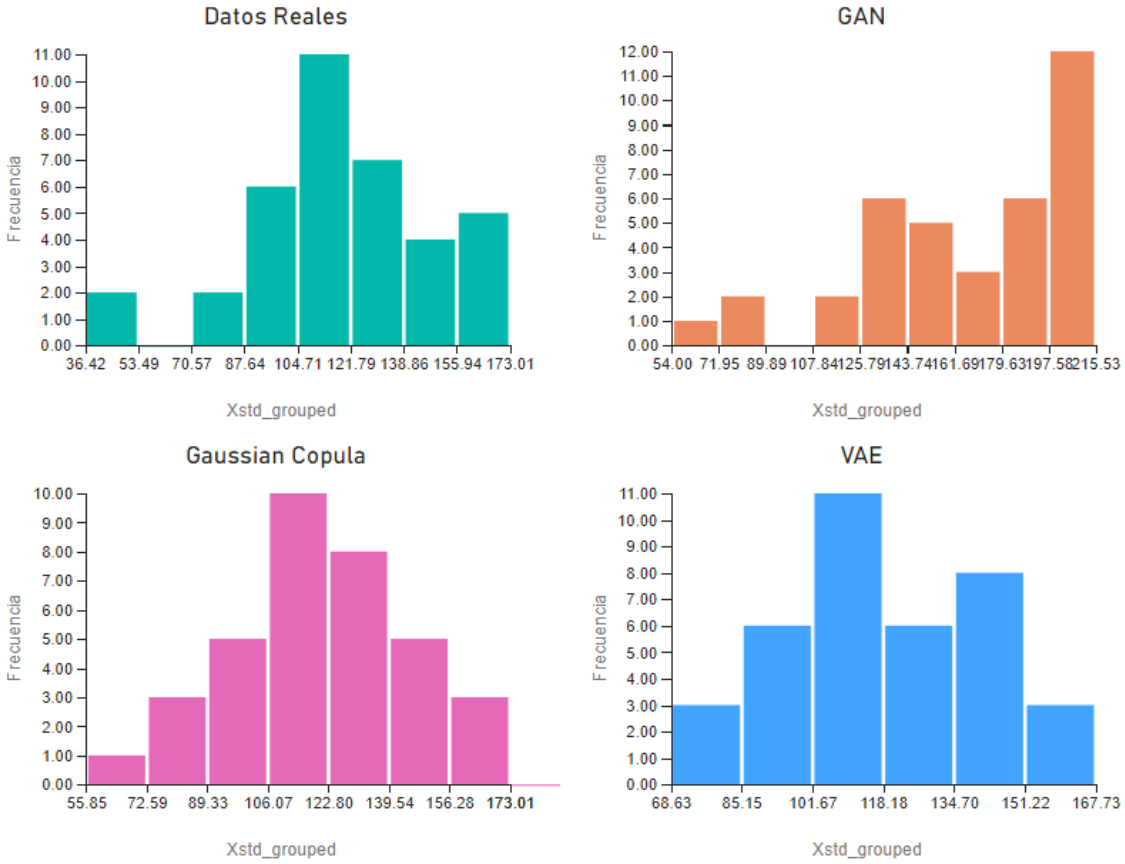

Figura 4.6: Distribución de la variable Xstd_grouped (Cámara de movimiento ocular) para los datos reales y los generados sintéticamente.

Son interesantes los resultados de la Figura 4.5, ya que permiten mostrar algunas cualidades de cada uno de los algoritmos generativos. Por un lado, el algoritmo con peores resultados nuevamente es GAN, pero al mirar el resto de las distribuciones, se puede notar que es el único capaz de mantener y generar valores atípicos o outliers para estos resultados. Por otro lado, Gaussian Copula al ser un algoritmo estadístico, se basa en encontrar la distribución de origen de los datos reales para generar nuevas entradas, en este caso, la distribución que logró encontrar es muy similar a una distribución normal y se ve reflejada en su gráfico. Por último, el resultado más cercano en esta etapa lo tiene el algoritmo VAE, que si bien no mantiene el mismo rango que la variable real, su distribución es muy similar a esta.

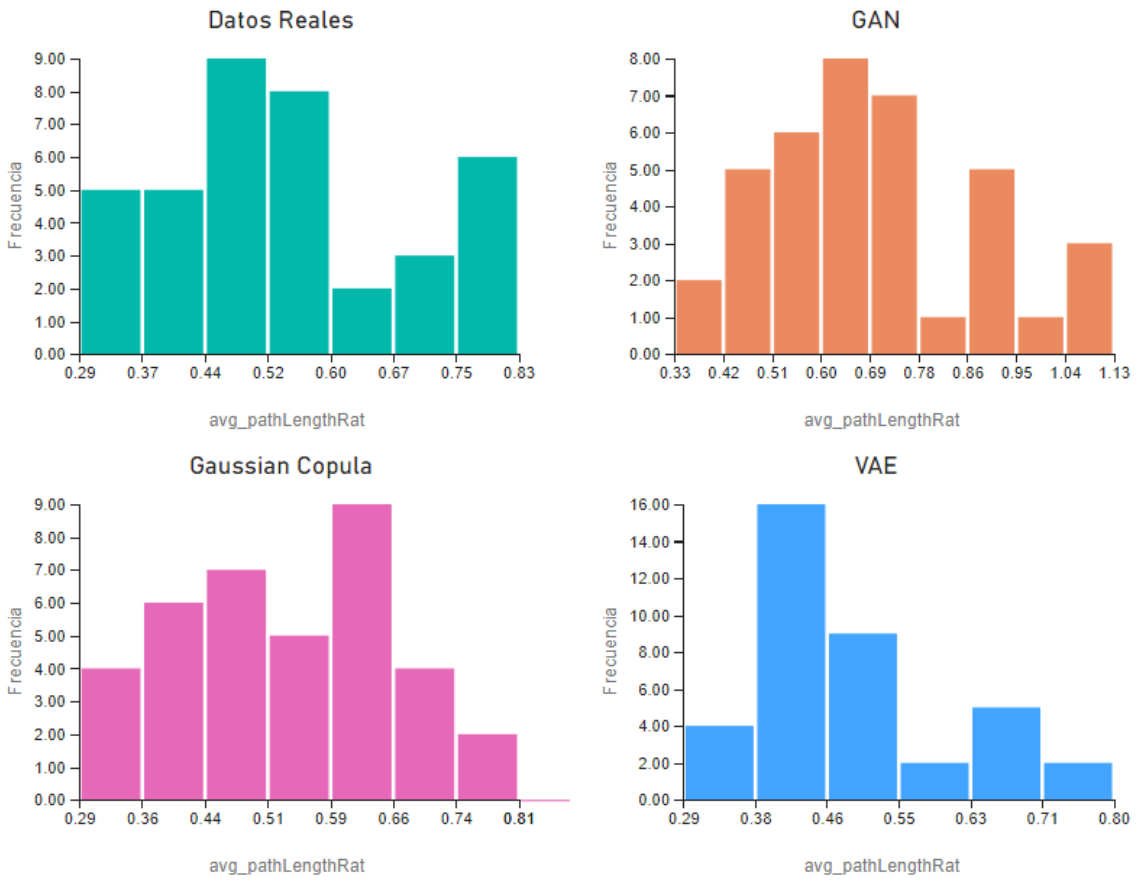

Figura 4.7: Distribución de la variable avg_pathLengthRat (Test de navegación) para los datos reales y los generados sintéticamente.

En el caso de la Figura 4.6 que representa datos provenientes del Test de Navegación, las tres distribuciones entregadas por los algoritmos son bastante diferentes entre sí. Por un lado, el algoritmo GAN tiene un rango distinto y su distribución no es muy similar a la de los datos reales. Por otro lado los algoritmos VAE y Gaussian Copula respetan el rango en el que se mueve la variable, pero las distribuciones no se asemejan de gran manera a los datos reales. El algoritmo VAE le da un gran peso a los valores pertenecientes al intervalo [0.38,0,46], con 16 individuos en esta categoría. Al ver la distribución de los datos reales, podemos notar una gran diferencia en esta categoría, con sólo 5 individuos. Es por esto que el mejor resultado en este caso corresponde a Gaussian Copula.

Para poder evaluar la generación en las bases de datos unificadas, que tienen variables desde las tres fuentes de datos distintas, es que en una primera etapa se calculan estadísticos descriptivos de las variables para medir la similitud entre los algoritmos generadores y los datos reales. Estos resultados pueden ser revisados en el apartado A de los Anexos, pero al ser tan extensos, es más sencillo entender el rendimiento a través de gráficos que resuman estas diferencias. Es por esta razón que en esta etapa se decide utilizar gráficos de líneas, que muestren el promedio o desviación

estándar de las variables, y así evidenciar las diferencias para cada algoritmo. Las figuras 4.7 a 4.10 muestran este proceso, seleccionando dos variables por cada fuente de datos dentro de todas las bases, se selección, haciendo distinción cuando se utilizan la misma cantidad de filas que la base original o 150 filas. Cabe destacar que se escogieron variables en base al rango de estas, seleccionando dos que pertenezcan a un rango entre 0 y 1.5, esto para poder mostrarlas en el mismo gráfico y notar las diferencias. Se destaca que para este proceso, se crean dos variables llamadas Xstd/100 y Ystd/100, que simplemente corresponden a las variables Xstd_grouped y Ystd_grouped provenientes del Eye Tracker divididas por 100. Esto se hace ya que estas variables y sus diferencias están en una escala de magnitud mayor que el resto, por lo que utilizar esta transformación ayuda a la lectura y entendimiento de los gráficos que se muestran a continuación.

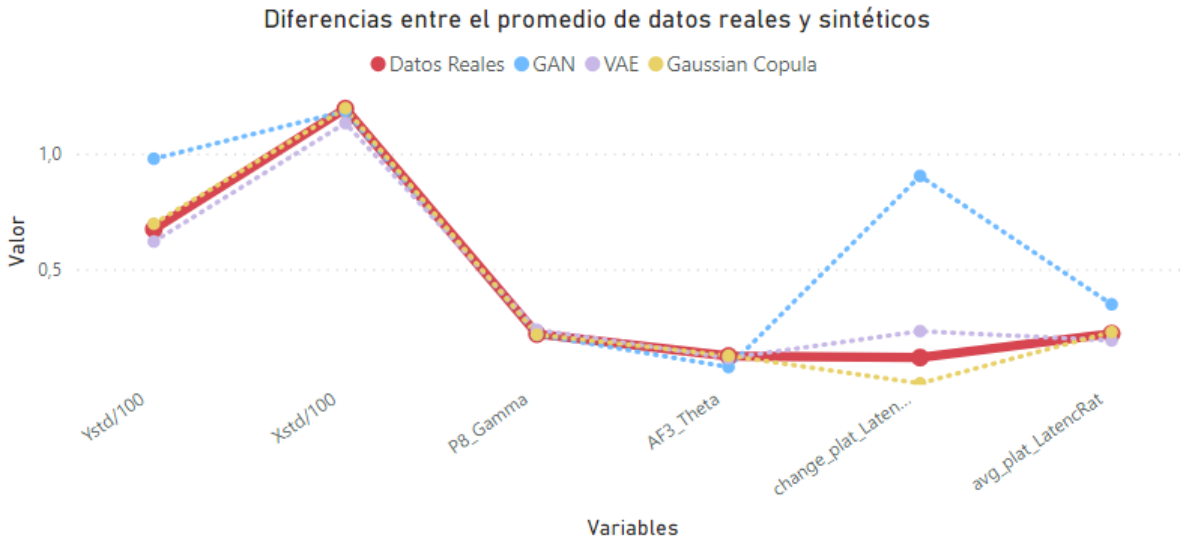

Figura 4.8: Diferencias entre el promedio de los datos reales y sintéticos para las variables Ystd/100, Xstd/100 (Cámara de movimiento ocular), P8_Gamma, AF3_Theta (EEG), change_plat_LatencRat y avg_plat_latencRat (Test de Navegación).

De acuerdo con la Figura 4.7, los mejores resultados corresponden a las variables provenientes del EEG para todos los algoritmos, recordando que es la fuente de datos con mayor cantidad de variables y por lo tanto, es posible especular una conexión entre este factor y la calidad de los datos generados. Con respecto a la comparación entre algoritmos, las distancias entre los promedios de las variables generadas sintéticamente y las reales en la mayoría de los casos son pequeñas, lo que es una buena señal sobre el proceso de generación de datos. Sin embargo, en las variables Ystd/100 y change_plat_LatencRat, las diferencias son notorias y relevantes, en particular para el algoritmo GAN, es por esta razón que en esta etapa es el peor evaluado dentro de los tres algoritmos. Por otro lado, es difícil evidenciar las diferencias entre VAE y Gaussian Copula, por lo que para estos efectos se ha decidido utilizar la desviación estándar en conjunto con el promedio para continuar con la evaluación. Estos resultados son presentados en la Figura 4.8 a continuación.

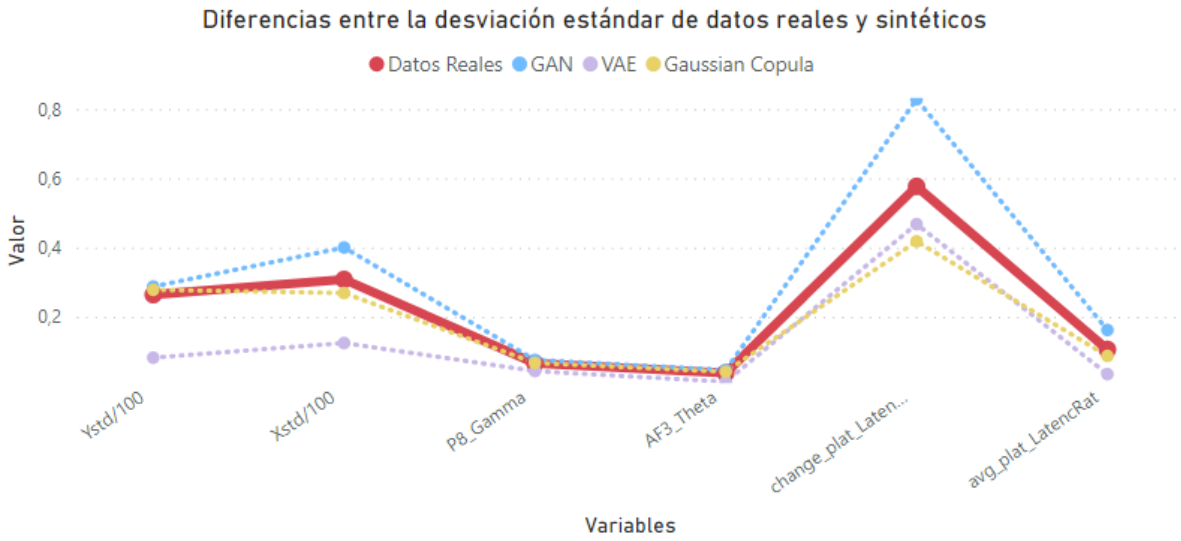

Figura 4.9: Diferencias entre la desviación estándar de los datos reales y sintéticos para las variables Ystd/100, Xstd/100 (Cámara de movimiento ocular), P8_Gamma, AF3_Theta (EEG), change_plat_LatencRat y avg_plat_latencRat (Test de Navegación).

Al contar con la información que entrega la Figura 4.8, es posible notar que las desviaciones estándar más similares a las reales para la mayoría de las variables corresponden a las generadas por el algoritmo Gaussian Copula, por lo tanto es el que tiene resultados más fieles a los reales al compararlo con el resto de los algoritmos en esta etapa. Este proceso se repite una vez más, pero utilizando las bases de datos sintéticas con 150 entradas, los resultados se muestran en las Figuras 4.9 y 4.10.

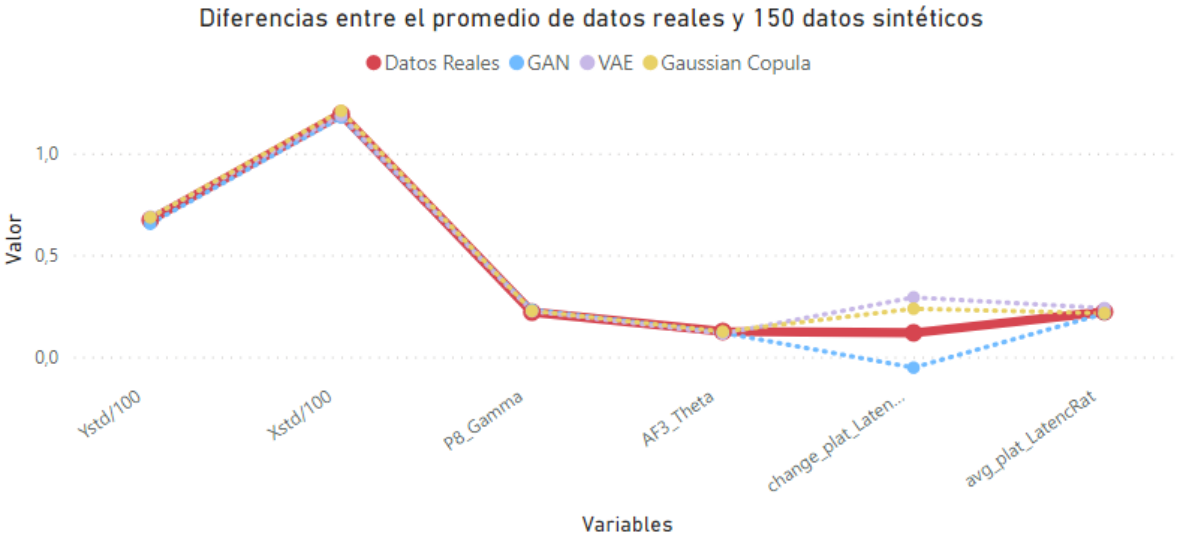

Figura 4.10: Diferencias entre el promedio de los datos reales y sintéticos de 150 filas para las variables Ystd/100, Xstd/100 (Cámara de movimiento ocular), P8_Gamma, AF3_Theta (EEG), change_plat_LatencRat y avg_plat_latencRat (Test de Navegación).

En la Figura 4.9 están los promedios de todos los algoritmos, los cuales son similares para todas las variables excepto en la variable changle_plat_LatencRat del Test de Navegación. Como la mayoría de los resultados son similares a los reales, es que es necesario analizar las diferencias en las desviaciones estándar para poder concluir sobre esta etapa. Pese a no tener un resultado claro aún, este gráfico permite evidenciar la capacidad de los tres algoritmos de replicar el promedio de los datos reales al generar datos sintéticos.

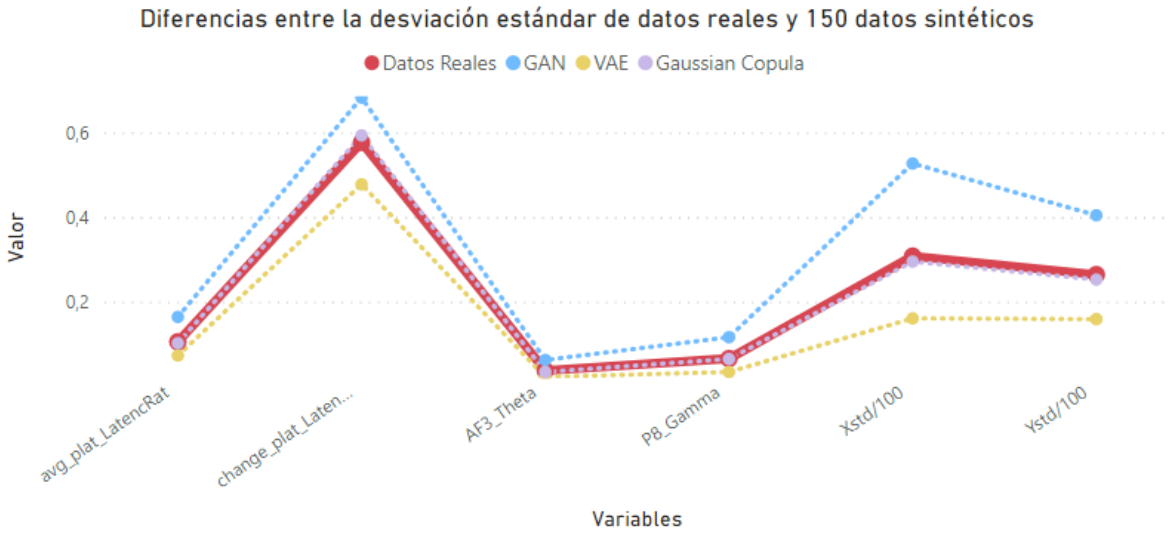

Figura 4.11: Diferencias entre la desviación estándar de los datos reales y sintéticos de 150 filas para las variables Ystd/100, Xstd/100 (Cámara de movimiento ocular), P8_Gamma, AF3_Theta (EEG), change_plat_LatencRat y avg_plat_latencRat (Test de Navegación).

Finalmente, la Figura 4.10 permite distinguir al algoritmo con mejor desempeño en la generación de 150 datos sintéticos, el cual corresponde a Gaussian Copula. Este algoritmo es el que simula de mejor manera las distribuciones estándares para todas las variables estudiadas, teniendo resultados casi idénticos a los datos reales. Por otro lado, el algoritmo peor evaluado de esta etapa es GAN. Es interesante poder expandir en la evaluación de los algoritmos, probando con ejemplos y aplicaciones reales los datos generados. En el próximo capítulo se presentarán evaluaciones más exhaustivas de los algoritmos, que permitirán reconocer sus diferencias y decidir cuál es el algoritmo con mejor rendimiento en esta generación de datos.

Capítulo 5

Evaluación de algoritmos generadores

Parte importante de este trabajo de título es poder no sólo generar datos sintéticos a través de múltiples herramientas, sino que evaluar el rendimiento de estos algoritmos para poder entregar una indicación de cuáles son los más efectivos en base a la evidencia. Es por esta razón que el siguiente capítulo se encargará de revisar diferentes formas de evaluar los rendimientos de la generación de datos, entregando información relevante para comparar los resultados obtenidos.

5.1 Índice de similitud

Este índice es construido en base a una colección de métricas estadísticas, las cuales son estandarizadas en el intervalo [0,1] y posteriormente son promediadas.
Este indicador compara las similitudes que existen entre la base de datos real y la base de datos sintética de manera estadística, comparando las distribuciones y correlaciones que existen entre bases.

Es relevante poder aclarar cómo se construye este índice, para estos fines se presenta la siguiente tabla, en la cual se encuentran las métricas utilizadas para calcular el índice de similitud. Como se mencionó anteriormente, estas métricas se calculan individualmente, son normalizadas dentro de un rango entre 0 y 1, y posteriormente promediadas.

Tabla 5.1: Métricas utilizadas para evaluar el rendimiento de los algoritmos. Extraído de [58].

Métrica	**Prueba que utiliza**
CSTest	Chi-Cuadrado
KSTest	Prueba de Kolmogorov-Smirnov
KSTestExtended	Prueba de Kolmorogov-Smirnov
LogisticDetection	Regresión Logística
SVCDetection	Clasificador de vectores de soporte
BNLikelihood	Redes Bayesianas
BNLogLikelihood	Redes Bayesianas
LogisticParentChildDetection	Regresión logística
SVCParentChildDetection	Clasificador de vectores de soporte

A continuación, se explican las diferentes métricas que componen este índice:

1. **Chi-Cuadrado:** Como se explica en [61] esta prueba tiene dos propósitos: monitorear la independencia entre dos variables, y testear que tan bien se ajusta una distribución a otra a través de un test de hipótesis.

2. **Prueba de Kolmogorov-Smirnov:** Es otra prueba estadística no-paramétrica que muestra la bondad de ajuste entre distribuciones. Comúnmente se utiliza para saber si dos muestras vienen de distribuciones diferentes [62].

3. **Regresión logística:** Es un análisis estadístico que permite conocer la relación entre las variables independientes y las dependientes [63]. Estima la probabilidad de que un evento ocurra y en este caso, de que una muestra pertenezca a una base de datos real o sintética.

4. **Clasificador de vectores de soporte:** Es la aplicación de clasificación que tiene el algoritmo Support Vector Machines (SVM). Así como se explica en [64], este aprende a utilizando los datos de prueba para encontrar un hiperplano óptimo dentro de un espacio hipotético, el cual es una colección de hiperplanos. Estos hiperplanos corresponden a divisiones que separan una clase de otra. En el caso lineal, el hiperplano se encuentra en el espacio muestral de los datos y le permite encontrar una solución. Este método se utiliza para clasificar y detectar posibles diferencias entre los datos reales y sintéticos.

5. **Redes Bayesianas:** Son un modelo probabilístico que representa un grupo de variables y sus dependencias condicionales [65]. Estas redes sirven para determinar la probabilidad de que alguna causa conocida haya afectado a un evento que ya ocurrió. En este caso, se utilizan para determinar la posibilidad de que, dado un grupo de datos, estos pertenezcan a una base de datos real o una sintética.

6. **Relaciones entre bases (Parent-Child detection):** Permite detectar las relaciones de jerarquía entre las variables de las bases reales y sintéticas [59].

La implementación de este índice de similitud fue realizada a través de la librería SDV en Python [58] y los resultados de la generación de datos han sido adjuntados en la sección A de los anexos, en donde se muestran las descripciones estadísticas tanto de las bases de datos reales como de las bases de datos generadas. En la siguiente tabla, se presentan los valores del índice de similitud calculado para los tres algoritmos, en las tres bases de datos pertenecientes al proyecto piloto ADDAI.

Tabla 5.2: Test de similitud entre los datos reales y los generados sintéticamente. Elaboración propia.

	EEG	Eye Tracker	Test de navegación
CTGAN	0.591	0.575	0.562
TVAE	0.585	**0.611**	0.516
GC	**0.704**	0.602	**0.649**

Como este índice de similitud es un promedio de varias métricas, permite mostrar de manera general cuán parecidas son las bases de datos sintéticos en comparación con los datos reales. Mientras más cerca del uno se encuentre el indicador, más similar será la base de datos de la base real. Es importante destacar que similar no se refiere a igualdad en este caso, por lo que tener un resultado muy cercano a uno no implica tener dos bases de datos iguales, sino que los datos que las componen provienen de una distribución similar o igual. Para este trabajo de título, se decidió que un mal resultado o una generación de datos no exitosa, será cualquier resultado que esté entre 0 y 0.4, mientras que un resultado entre 0.41 y 0.75 se considera un resultado positivo y cualquier resultado mayor a 0.75 será considerado un buen resultado.

En esta primera aproximación a poder evaluar el rendimiento de los algoritmos, es posible notar pequeñas diferencias entre el desempeño de los algoritmos que utilizan redes neuronales, como CTGAN y VAE y el algoritmo estadístico Gaussian Copula. Este último presenta un mejor desempeño en dos de las tres bases de datos, siendo muy cercano su resultado con respecto al primer lugar en la base del Test de Navegación. Si bien los resultados en esta etapa no son excelentes, si permiten continuar trabajando con los datos generados, ya que todos se posicionan en el rango en que la generación de datos fue exitosa. Estos indicadores no permiten diferenciar de manera clara los resultados obtenidos con estos algoritmos, por lo que es necesario profundizar esta investigación utilizando otras formas de evaluación utilizadas en la literatura las cuales serán exploradas en el próximo apartado.

5.2 Evaluación utilizando un modelo de clasificación

Para darle profundidad al análisis, se utilizarán las métricas mencionadas en 1.4.5, las cuales permiten evaluar a los algoritmos entrenando un modelo de clasificación con datos sintéticos y con la mezcla de datos sintéticos y datos reales, comparando estos resultados al caso base de sólo utilizar datos reales. Para esto se utilizarán las métricas de Accuracy, Recall, Precision, F1 Score y ROC AUC mencionadas en 1.4.5 para tener un gran espectro de evaluación, siendo Recall la métrica más relevante dentro del análisis en el área de la salud [66], puesto a que muestra la capacidad del modelo de identificar realmente a los pacientes que padezcan una enfermedad. Junto con esto, la métrica ROC AUC también se considera importante para el análisis, ya que toma en cuenta a Recall dentro de su construcción y muestra la capacidad del modelo de distinguir entre pacientes control y estudio. El modelo de clasificación corresponde a una regresión logística y existen dos tipos de modelos: Uno que calcula la probabilidad por tipo de dato o examen, agrupando los datos pertenecientes a cada tipo de examen y posteriormente calculando la probabilidad para cada uno, mientras que el otro modelo, de probabilidad global, utiliza todas las variables mencionadas en 4.2 al mismo tiempo. Se escoge este modelo ya que es el que utilizan en el proyecto ADDAI para clasificar pacientes, por lo que es lo más cercano a la aplicación real que podrían tener estos datos.

Estos algoritmos operan a través de una función creada con la librería Scikit-Learn [67] de Python, la cual permite definir el modelo, en este caso de regresión logística. Esta función es alimentada con cuatro conjuntos de variables provenientes de la base de datos inicial: *X_train, X_test, Y_train e Y_test*, siendo X el conjunto de variables independientes e Y el vector de variables dependientes que se busca predecir, repartiendo el 80% de los datos en la base de entrenamiento y el 20% restante en la de testeo, las bases de *train* son las bases que utiliza el modelo para el entrenamiento y las bases de *test* son las que se utilizan posteriormente para generar una predicción y comparar los resultados obtenidos. De esta manera, el modelo es capaz de aprender sobre el fenómeno y es posible evaluarlo con datos que desconoce para medir su rendimiento real.

Se comienza este proceso de evaluación utilizando los datos reales para entrenar ambos modelos, obteniendo una base de comparación para evaluar el desempeño de los algoritmos generativos. Se presentan a continuación los resultados de las métricas en las bases de *test* y *train* con el fin de indicar que el modelo está funcionando correctamente.

Tabla 5.3: Métricas asociadas al rendimiento de los modelos entrenados con datos reales. Elaboración propia.

	Accuracy		Recall		Precision		F1 Score		ROC AUC	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Probabilidad por tipo de dato	0,96	1	1	1	0,93	1	0,96	1	0,59	0,9
Probabilidad Global	0,9	1	1	1	0,81	1	0,90	1	1	1

Los resultados tanto en la base de *train* como en la de *test* permiten mostrar que ambos modelos funcionan adecuadamente. Siguiendo con las características de los modelos, estos se adecuan bien a los datos pese a tener una escasa cantidad y dado el Recall, pueden finalmente clasificar de manera correcta a un paciente con enfermedad de Alzheimer de un paciente control en la base de *test*.

5.2.1 Datos sintéticos

Ya teniendo una primera aproximación de cómo funciona el modelo en los datos reales, es posible comparar estos resultados a los que se obtendrán utilizando los datos sintéticos al entrenar el mismo modelo, pudiendo así poder medir las diferencias entre estos rendimientos. En la siguiente figura se muestra un esquema que entrega más detalles sobre el proceso de evaluación.

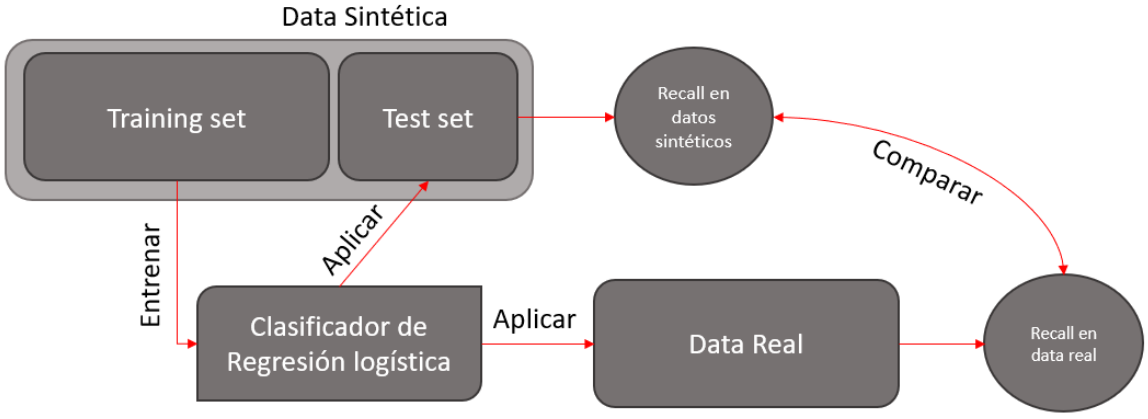

Figura 5.1: Pasos para poder evaluar el rendimiento de los datos sintéticos utilizando un algoritmo de clasificación y los datos reales. Basado en [13].

Siguiendo la lógica de este esquema es que se calculan los rendimientos en las métricas mencionadas anteriormente, junto con las diferencias en comparación al modelo entrenado con datos reales. En esta ocasión, sólo se mostrarán los resultados de la evaluación realizada a toda la base de datos reales, o sea, se evaluará el rendimiento del

modelo entrenado en datos sintéticos utilizando como base de *test* los datos reales, simulando lo que sería una aplicación real de esta herramienta.

Utilizando las bases de datos sintéticas con 38 muestras, la misma cantidad que la base de datos real, es que se entrenan los modelos de probabilidad por tipo de dato y probabilidad global del proyecto piloto ADDAI. Los resultados se presentan con color rojo si son negativos, negro si no hubo cambios y verde si es que hubo una mejora en el rendimiento, estos se muestran en las tablas 5.4, 5.5 y 5.6 a continuación.

Tabla 5.4: Métricas obtenidas entrenando un clasificador con datos sintéticos y testeándolo en datos reales en el modelo de probabilidad por tipo de dato. Elaboración propia.

	Accuracy	**Recall**	**Precision**	**F1 Score**	**ROC AUC**
	Test	**Test**	**Test**	**Test**	**Test**
GAN	0,68 (-0,32)	0,88 (-0,12)	0,58 (-0,42)	0,7 (-0,3)	0,62 (-0,28)
VAE	0,63 (-0,37)	0,5 (-0,5)	0,57 (-0,43)	0,53 (-0,47)	0,66 (-0,24)
GC	0,82 (-0,18)	0,94 (-0,06)	0,71 (-0,29)	0,81 (-0,19)	0,85 (-0,05)
Control	1	1	1	1	0,9

Tabla 5.5: Métricas obtenidas entrenando un clasificador con datos sintéticos y testeándolo en datos reales en el modelo de probabilidad global. Elaboración propia.

	Accuracy	**Recall**	**Precision**	**F1 Score**	**ROC AUC**
	Test	**Test**	**Test**	**Test**	**Test**
GAN	0,66 (-0,34)	0,88 (-0,12)	0,56 (-0,44)	0,68 (-0,32)	0,45 (-0,45)
VAE	0,84 (-0,16)	1 (0)	0,72 (-0,28)	0,84 (-0,16)	1 (0)
GC	0,82 (-0,18)	0,88 (-0,12)	0,74 (-0,26)	0,8 (-0,2)	0,95 (-0,05)
Control	1	1	1	1	1

Tabla 5.6: Variación promedio de las métricas de evaluación entre ambos modelos al utilizar datos sintéticos. Elaboración Propia.

	Accuracy	**Recall**	**Precision**	**F1 Score**	**ROC AUC**
GAN	-0,33	-0,12	-0,43	-0,31	-0,365
VAE	-0,265	-0,25	-0,355	-0,315	-0,12
GC	**-0,18**	**-0,09**	**-0,275**	**-0,195**	**-0,05**

Los resultados obtenidos en esta etapa son similares a los obtenidos en 5.1, pero permiten ver con más detalle las diferencias entre los algoritmos. Si bien ambas redes neuronales siguen funcionando de peor manera en comparación con el algoritmo estadístico, VAE presenta resultados bastante cercanos al modelo original con datos reales en el modelo de probabilidad global, sobre todo en las métricas Recall y ROC AUC, mientras que el algoritmo GAN no presenta resultados lo suficientemente robustos para poder entrenar ninguno de modelos de clasificación. Sin dudas, el algoritmo mejor

evaluado en esta etapa corresponde al Gaussian Copula, el cual tiene resultados más cercanos que el resto a los datos reales en la mayoría de las métricas, esto se puede ver al comparar las variaciones promedio, las cuales son menores en todas las métricas para este algoritmo.

No se puede dejar de lado una de las características importantes de los datos sintéticos, y es que permiten generar una cantidad de datos mucho mayor en comparación a la fuente original, es entonces relevante poder probar esta capacidad y ver cómo se comportan los modelos al incorporar una mayor cantidad de datos sintéticos en la etapa de entrenamiento. Las tablas 5.7, 5.8 y 5.9 representan los resultados obtenidos de este proceso, el cual es similar en estructura a la evaluación anterior, pero entrenando estos modelos con bases sintéticas de 150 datos.

Tabla 5.7: Métricas obtenidas entrenando un clasificador con 150 datos sintéticos y testeándolo en datos reales en el modelo de probabilidad por tipo de dato. Elaboración propia.

	Accuracy	**Recall**	**Precision**	**F1 Score**	**ROC AUC**
	Test	**Test**	**Test**	**Test**	**Test**
GAN	0,63 (-0,37)	0,63 (-0,37)	0,56 (-0,44)	0,59 (-0,41)	0,6 (-0,30)
VAE	0,81 (-0,19)	0,69 (-0,31)	0,85 (-0,15)	0,76 (-0,24)	0,63 (-0,27)
GC	0,95 (-0,05)	1 (0)	0,88 (-0,12)	0,94 (-0,06)	0,5 (-0,40)
Control	1	1	1	1	0,9

Tabla 5.8: Métricas obtenidas entrenando un clasificador con 150 datos sintéticos y testeándolo en datos reales en el modelo de probabilidad global. Elaboración propia.

	Accuracy	**Recall**	**Precision**	**F1 Score**	**ROC AUC**
	Test	**Test**	**Test**	**Test**	**Test**
GAN	0,66 (-0,34)	0,63 (-0,37)	0,59 (-0,41)	0,61 (-0,39)	0,55 (-0,35)
VAE	0,73 (-0,27)	0,75 (-0,25)	0,67 (-0,33)	0,71 (-0,29)	0,8 (-0,10)
GC	0,97 (-0,03)	1 (0)	0,94 (-0,06)	0,97 (-0,03)	0,98 (-0,02)
Control	1	1	1	1	1

Tabla 5.9: Variación promedio de las métricas de evaluación entre ambos modelos al utilizar 150 datos sintéticos. Elaboración Propia.

	Accuracy	**Recall**	**Precision**	**F1 Score**	**ROC AUC**
GAN	-0,355	-0,37	-0,425	-0,4	-0,325
VAE	-0,23	-0,28	-0,24	-0,265	-0,185
GC	**-0,04**	**0**	**-0,09**	**-0,045**	**-0,16**

Los resultados obtenidos en esta etapa cuentan una historia similar a la anterior, pero permiten discernir sobre que algoritmo posee mejor capacidad para replicar la información contenida en los datos reales. El algoritmo GAN presenta resultados que no permiten entrenar ambos modelos correctamente, disminuyendo sus resultados en comparación con la evaluación anterior en todas las métricas. Esto se explica ya que, al generar una mayor cantidad de datos, las diferencias en las distribuciones desde donde provienen estos datos son aún más evidentes, por lo que el modelo ya no es capaz de etiquetar correctamente los datos reales, puesto a que fue entrenado con una base de entrenamiento muy diferente.

Por otro lado, el algoritmo VAE presenta mejores resultados en comparación a la evaluación anterior. Si bien es cierto que el Recall disminuye en esta etapa, esta disminución es mucho menor que la del algoritmo GAN y si es posible notar un incremento en las métricas Accuracy y Precision. Este resultado indica que los datos generados a través del algoritmo Variational Autoencoders son de más cercanos a la realidad que los generados por GAN.

Finalmente, el algoritmo Gaussian Copula es nuevamente el mejor evaluado en esta etapa, teniendo resultados muy similares a los modelos originales. Al revisar las variaciones promedio, este algoritmo mejoró en casi todas sus métricas en esta etapa en comparación con la anterior. Esto demuestra su utilidad en esta aplicación en particular, en donde se trabaja con bases de datos numéricas con sólo una variable categórica.

Falta aún revisar otra forma de implementar datos sintéticos, la cual corresponde a la combinación entre datos sintéticos y reales para entrenar un modelo. Es por esto que en el siguiente apartado se revisará este proceso para poder concluir sobre la calidad de las muestras generadas por los algoritmos.

5.2.2 Datos mixtos

En esta última etapa de evaluación, se combinarán los datos reales con los generados sintéticamente por los algoritmos uno por uno. Se entrenarán los modelos de probabilidad por tipo de dato y probabilidad global y se harán pruebas sobre la base de *test*. La estructura de esta evaluación será distinta a la realizada anteriormente, ya que, al combinar los datos no tiene sentido probar los algoritmos sobre la base completa de datos reales. En la siguiente figura se muestra un esquema que entrega más detalles sobre el proceso de evaluación.

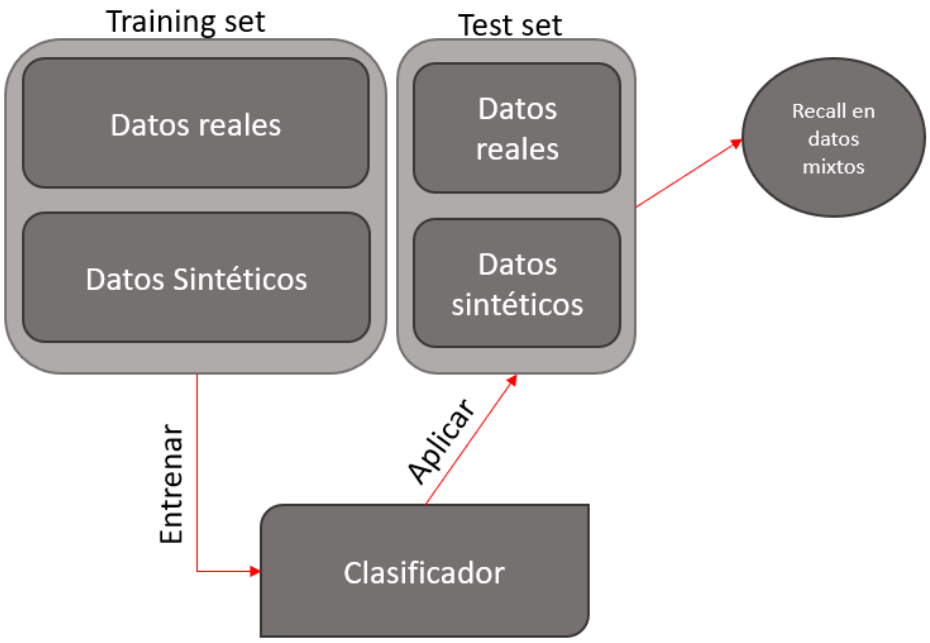

Figura 5.2: Pasos para poder evaluar el rendimiento de los datos mixtos utilizando un algoritmo de clasificación y los datos reales. Basado en [13].

Como se indica en la Figura 5.2, se utilizará la arquitectura clásica de evaluación de un modelo de clasificación. Esta se realiza separando la base en dos, una base de *training* o entrenamiento y una base de *test* o prueba, entrenando el algoritmo clasificador utilizando la base de entrenamiento y realizando las evaluaciones en la base de testeo. A través de este proceso se confeccionan las tablas 5.10, 5.11 y 5.12, las cuales muestran el rendimiento de los dos modelos utilizando las métricas mencionadas en 1.4.5.

Tabla 5.10: Métricas obtenidas entrenando un clasificador con datos mixtos y testeándolo en datos reales en el modelo de probabilidad por tipo de dato. Elaboración propia.

	Accuracy	**Recall**	**Precision**	**F1 Score**	**ROC AUC**
	Test	**Test**	**Test**	**Test**	**Test**
GAN	0,94 (-0,06)	0,86 (-0,14)	1 (0)	0,92 (-0,08)	0,55 (-0,35)
VAE	0,94 (-0,06)	0,83 (-0,17)	1 (0)	0,91 (-0,09)	0,5 (-0,4)
GC	0,88 (-0,12)	0,75 (-0,25)	1 (0)	0,86 (-0,14)	0,66 (-0,24)
Control	1	1	1	1	0,9

Tabla 5.11: Métricas obtenidas entrenando un clasificador con datos mixtos y testeándolo en datos reales en el modelo de probabilidad global. Elaboración propia.

	Accuracy	Recall	Precision	F1 Score	ROC AUC
	Test	Test	Test	Test	Test
GAN	0,94 (-0,06)	1 (0)	0,88 (-0,12)	0,83 (-0,17)	0,95 (0,05)
VAE	0,69 (-0,31)	0,50 (-0,50)	0,60 (-0,40)	0,68 (-0,32)	0,8 (-0,20)
GC	0,94 (-0,06)	0,88 (-0,12)	1 (0)	1 (0)	0,97 (-0,03)
Control	1	1	1	1	1

Tabla 5.12: Variación promedio de las métricas de evaluación entre ambos modelos al utilizar datos mixtos. Elaboración Propia.

	Accuracy	Recall	Precision	F1 Score	ROC AUC
GAN	**-0,06**	**-0,07**	-0,06	-0,125	-0,15
VAE	-0,185	-0,335	-0,2	-0,205	-0,3
GC	-0,09	-0,185	**0**	**-0,07**	**-0,135**

Los resultados obtenidos en esta instancia siguen mostrando la superioridad y estabilidad del algoritmo Gaussian Copula, que si bien no es el mejor evaluado en esta etapa por tener una diferencia de Recall más grande que el algoritmo GAN, mantiene buenos resultados para todas las evaluaciones que se han realizado. En esta instancia el algoritmo peor evaluado es VAE, el cual al presentar una diferencia de Recall de -0,34 y de ROC AUC de -0,30 indica que no ha podido entrenar los modelos de manera correcta en esta etapa. Si bien las métricas para el algoritmo GAN tienen un buen resultado en esta evaluación, no son capaces de compensar con el mal rendimiento que ha tenido a lo largo de este proceso, presentando resultados cercanos o peores que los que ha demostrado el algoritmo VAE, la otra red neuronal que se utilizó.

Para terminar la etapa de evaluación de los modelos, es posible concluir en base a la evidencia que el algoritmo con mejor rendimiento para las tres bases de datos con las que se trabajó corresponde al algoritmo estadístico Gaussian Copula. Este permitió entrenar a los modelos de manera muy cercana a utilizar datos reales en las tres evaluaciones que se realizaron. Su peor evaluación corresponde a la última aplicación, en la que se mezclaron datos sintéticos y datos reales, mientras que se desempeñó de mejor manera en la evaluación que utilizaba 150 datos sintéticos para entrenar los modelos.

Con respecto a los algoritmos que utilizan redes neuronales, es posible confirmar que el algoritmo Variational Autoencoders permite generar datos de manera exitosa, los cuales pueden ser utilizados para entrenar modelos en grandes cantidades, ya que su mejor evaluación correspondió a la etapa de 150 datos. Finalmente, se considera que, pese a que el algoritmo Generative Adversarial Networks fue exitoso al momento de generar

datos, las evaluaciones lo posicionaron como el peor algoritmo para entrenar ambos modelos utilizados en el proyecto ADDAI. Sin embargo, este algoritmo tuvo un gran éxito particular al momento de ser combinado con los datos reales para entrenar al modelo, presentando mejores resultados que el algoritmo Gaussian Copula, por lo que se cree que sus posibles aplicaciones deben seguir estas mismas directrices.

5.3 Análisis de sensibilidad

Al trabajar con redes neuronales, es importante poder entender cómo los argumentos o *inputs* que se le entregan a los algoritmos afectan a los resultados, esto se hace para poder entender con más detalle el impacto que tienen sobre el modelo estas variables. Es por esta razón que se decide realizar un análisis de sensibilidad sobre las variables *batch size* o tamaño del lote, que representa la cantidad de muestras que se le entregan en una misma etapa al algoritmo, y *epochs* o épocas, que representan cuando pasa la totalidad de los datos dentro de la base una vez por la red neuronal [68]. La arquitectura de ambos modelos como se mencionan en 4.2 corresponden a cuatro capas para las redes discriminadora y generadora del algoritmo GAN y dos capas para las redes compresora y descompresora del algoritmo VAE. Las siguientes tablas se construyeron utilizando el índice de similitud descrito en 5.1, el cual busca evaluar si los datos sintéticos provienen de distribuciones similares a las reales y si es que las variables mantienen las relaciones originales, se comienza evaluando el impacto del tamaño del lote en los resultados de las redes neuronales.

Tabla 5.13: Índice de similitud para datos sintéticos generados con el algoritmo GAN en las diferentes bases de datos utilizando como base 300 épocas. Elaboración propia.

		Batch = 10	**Batch = 20**	**Batch = 50**	**Batch = 100**
GAN	**EEG**	**0.6078**	0.4949	0.5221	0.5746
	Eye Tracker	0.5127	**0.5658**	0.5049	0.5192
	Test Nav	0.5497	0.5142	0.542	**0.6382**
	Promedio	0.5567	0.5249	0.523	**0.5773**

Tabla 5.14: Índice de similitud para datos sintéticos generados con el algoritmo VAE en las diferentes bases de datos utilizando como base 300 épocas. Elaboración propia.

		Batch = 10	**Batch = 20**	**Batch = 50**	**Batch = 100**
VAE	**EEG**	0.6133	0.5235	0.5728	**0.6163**
	Eye Tracker	0.5881	**0.6646**	0.6335	0.586
	Test Nav	0.3356	0.4467	0.5925	**0.6383**
	Promedio	0.5123	0.5449	0.5996	**0.6135**

Con la información de las tablas 5.13 y 5.14, es posible notar que los mejores promedios del índice de similitud para ambos algoritmos corresponden a utilizar un *batch size* de 100 unidades, sin embargo, no es posible olvidar que la base de datos contiene solamente 38 filas. Como este parámetro indica la cantidad de datos que se le entrega a la red neuronal en cada etapa, esto implica que en cada etapa de aprendizaje, se le entrega al modelo la base repetida 2.5 veces, recordando también la definición de época, la cual corresponde a la cantidad de veces que pasa la totalidad de la base de datos por la red neuronal, esto sugiere que por cada etapa se estaría avanzando 2.5 épocas, lo que reduciría el proceso de aprendizaje.

Como se expresa en [71], al utilizar un *batch size* grande es posible encontrarse con problemas de sobreajuste de los modelos, y frente a un *batch size* pequeño, puede presentar diferencias en las propiedades exploratorias de la red neuronal, generando dificultades para generalizar los resultados. Es por estas razones que pese a obtener mejores resultados para un *batch size* de 100 unidades, se decide utilizar uno de 20 unidades, puesto a que es un número menor a la cantidad de datos y en promedio funciona mejor que el de 10 unidades. Ya decidido el tamaño del lote, se fija este y se prosigue a medir el impacto que tienen las épocas en los resultados, este proceso se muestra a continuación en las tablas 5.15 y 5.16.

Tabla 5.15: Índice de similitud para datos sintéticos generados con el algoritmo GAN en las diferentes bases de datos utilizando como base un batch size de 20 muestras. Elaboración propia.

		Épocas = 50	**Épocas = 150**	**Épocas = 500**	**Épocas = 1000**
GAN	**EEG**	0.4032	0.4691	**0.5914**	0.5652
	Eye Tracker	0.5379	0.4456	**0.5752**	0.4544
	Test Nav	**0.578**	0.5496	0.5627	0.5089
	Promedio	0.5063	0.4881	**0.5763**	0.5095

Tabla 5.16: Índice de similitud para datos sintéticos generados con el algoritmo VAE en las diferentes bases de datos utilizando como base un batch size de 20 muestras. Elaboración propia.

		Épocas = 50	**Épocas = 150**	**Épocas = 500**	**Épocas = 1000**
VAE	**EEG**	0.6206	**0.6306**	0.5857	0.6167
	Eye Tracker	0.4891	0.5391	**0.6112**	0.535
	Test Nav	0.6697	**0.6773**	0.5166	0.5497
	Promedio	0.5931	**0.6156**	0.5711	0.5671

Los resultados indican que los algoritmos funcionan mejor en épocas distintas, siendo GAN más preciso en 500 épocas y VAE en 150. Si bien es posible utilizar una cantidad distinta para cada uno, se busca poder comparar los resultados obtenidos para cada algoritmo, por lo que mantener una igualdad de condiciones en esta etapa es importante. Es por esta razón que se decide utilizar 500 épocas para ambos, ya que los resultados para ambos son mejores en promedio que los resultados con 150 épocas. Concluyendo este análisis se decide utilizar un *batch size* de 20 unidades y una cantidad de 500 épocas para el entrenamiento de los modelos.

Capítulo 6

Evaluación del impacto económico

Al poder generar datos sintéticos y aumentar la cantidad de datos que se manejan en proyectos de investigación, se genera valor dentro de dichos proyectos, ya sea agilizando procesos que demandan horas hombre costosas o evitando costos al necesitar un menor número de sujetos para entrenar modelos. Para poder medir este impacto económico, es que primero se detallarán los costos del proceso de generación de exámenes reales de manera general, se estimarán los participantes del mercado de investigación utilizando herramientas de aprendizaje de máquinas en el área de la salud, y finalmente se calcularán los beneficios de utilizar estas herramientas dentro de los proyectos.

6.1 Costos de generar exámenes

Para realizar una estimación económica de cuánto puede costar la generación de un examen en un proyecto de salud similar, se utilizará la misma generalización vista en el apartado 3.4 y se estimarán los costos relativos a cada etapa del proyecto, utilizando como base la información sobre los cargos y funciones del proyecto ADDAI, la información brindada por el equipo médico y los datos de acceso público del HCUCH sobre costos de personal. Según la postulación al fondo IDeA I+D 2020 del proyecto ADDAI, se estimó la participación de 80 sujetos para ser examinados, y como esta etapa busca ser una generalización de aquel proceso, se estimará la participación de 100 pacientes para todas las etapas por simplicidad. A continuación, se adjuntan cuatro tablas que condensan esta información entregando un total para cada etapa.

1. **Búsqueda de pacientes:**

Tabla 6.1: Costos asociados a la etapa de búsqueda de pacientes. Elaboración Propia.

Detalle	Porcentaje del total de pacientes*	Costo	Costo promedio por paciente	Costo total del proceso
Cita médica grátis	100%	$40.000 - $60.000	$45.000	$4.500.000
Problemas de transporte o movilidad	20% a 30%	$10.000 – $15.000	$3.750	$375.000
Extras (colación, EPP, extras)	100%	$5.000	$5.000	$500.000
Total			$51.750	**$5.175.000**

*Costos y porcentajes estimados en base a entrevistas con equipo médico encargado de realizar exámenes en el proyecto ADDAI.

2. **Confirmación del diagnóstico:**

Tabla 6.2: Costos asociados a la etapa de confirmación del diagnóstico. Elaboración Propia.

Detalle	Porcentaje del total de pacientes*	Costo	Costo estimado por paciente	Costo total del proceso
Exámenes	85%	$20.000 - $120.000	$59.200	$5.920.000
Cita extra	15%	$60.000 - $80.000	$10.500	$1.050.000
Total			$69.700	**$6.970.000**

*Costos y porcentajes estimados en base a entrevistas con equipo médico encargado de realizar exámenes en el proyecto ADDAI.

3. **Generación de muestras:**

Tabla 6.3: Costos asociados a la etapa de generación de muestras. Elaboración Propia.

Detalle	Porcentaje del total de pacientes*	Costo	Costo estimado por paciente	Costo total del proceso
Pago 1 TENS	100%	$4.550/HH	$4.550	$455.000
Pago 1 Enfermera/o	100%	$11.000/HH	$11.000	$1.100.000
Total			$15.550	**$1.555.000**

*Costos y porcentajes estimados en base a entrevistas con equipo médico encargado de realizar exámenes en el proyecto ADDAI.

**Costo HH estimado en base a sueldos dispuestos en [78].

4. **Procesamiento de los datos:**

Tabla 6.4: Costos asociados a la etapa de procesamiento de datos. Elaboración Propia.

Detalle	% De dedicación*	Costo/hora	Costo total del proceso
Pago HH 2 investigadores responsables del análisis de datos	30%	$6.666	$7.839.000
Pago HH investigador encargado	20%	$8.888	$3.484.000
Pago HH directora del centro de investigación	30%	$8.333	$4.899.900
Pago HH jefe de área comercial del centro de investigación	30%	$5.555	$3.466.320
Total			**$19.688.320**

*Porcentaje de dedicación estimado en base a la postulación del proyecto ADDAI al concurso IDeA I+D 2020.

****Costo total estimado como HH al año * % de dedicación * Costo HH

6.2 Mercado de investigación

Ya teniendo un valor estimado para cada etapa de un proyecto generalizado de aprendizaje de máquinas asociado a salud, es momento de revisar el mercado existente para este tipo de proyectos. Para estos fines, se investigó a los principales actores en este mercado en Chile; Concurso FONDECYT, concurso IDeA I+D y concurso Fondo Nacional de Investigación en Salud (FONIS), los cuales entregan fondos para desarrollar investigación en torno a diferentes temáticas. En este caso, se buscaron sólo proyectos asociados a salud y más en particular aún, que utilicen datos dentro de este. Se construyeron dos tablas filtrando las bases de datos originales, una para los proyectos adjudicados del concurso IDeA I+D y otra para el concurso FONIS, con el nombre del proyecto y el monto asignado, ambas tablas están adjuntas en la sección B de los anexos. Por otro lado, no se encontró el mismo nivel de detalle para el concurso FONDECYT, pero sí una tabla que entrega información para la categoría de medicina. Si bien no es posible poder descartar proyectos médicos que no utilicen datos, se decidió utilizar una aproximación del 40%, es decir, se estimó que esa proporción de los proyectos dentro del concurso corresponden a la categoría de medicina y además utilizan datos. La información condensada por concurso se adjunta en la siguiente tabla.

Tabla 6.5: Cantidad de proyectos y sus montos asignados por cada concurso de investigación. Elaboración propia.

	Proyectos relacionados a salud que utilicen datos	**Monto total asignado (en millones de pesos)**
FONDECYT	17	$4064,57
IDeA I+D	12	$2925,76
FONIS	4	$237
Total	33	$7227,33

Con la información de la tabla 6.5 es posible notar que se estima un total de 33 proyectos dentro de este mercado, y el monto total es de $7.227 millones de pesos. Hay que recordar que estos concursos son realizados anualmente, por lo que el capital y la cantidad de proyectos pueden ir variando año a año y los montos deben ser percibidos como un flujo.

6.3 Beneficios de utilizar datos sintéticos

Para calcular los beneficios de utilizar datos sintéticos, un primer punto por considerar es que los proyectos si necesitan datos reales para llegar a conclusiones que aporten valor, por lo tanto, las etapas de búsqueda de pacientes, confirmación de diagnóstico y generación de muestras son independientes de que se apliquen datos sintéticos o no. Al final del día, se necesita generar esta información y si un proyecto requiere de una cantidad de pacientes para probar un fenómeno, esta cantidad debe ser alcanzada a través de las etapas anteriormente mencionadas independientemente de la cantidad de tiempo que se requiera. Esto sucede ya que si bien los datos sintéticos pueden replicar la información que existe detrás de los datos reales, no son capaces de reemplazarlos para estas aplicaciones de investigación en donde los riesgos de llegar a una conclusión equivocada a través de los datos pueden ser muy altos.

Por otra parte, y pensando que estos proyectos son avaluados anualmente, si es que se obtiene una cantidad de datos inferior a los datos requeridos por la dificultad de encontrar pacientes que participen en los ensayos clínicos o que los mismos exámenes tomen más tiempo del estipulado, se debe seguir pagando el sueldo de los investigadores y científicos de datos involucrados en el proyecto pese a no poder llegar a un resultado concreto. Estos atrasos terminan siendo costosos y nunca se debe olvidar que en la mayoría de los casos se utilizan recursos públicos para financiar estos proyectos. Es aquí donde los datos sintéticos pueden aportar valor, en la etapa de procesamiento de datos, ya que puede ayudar a investigadores a entrenar o mejorar la calidad de sus modelos sin necesitar una gran cantidad de información, haciendo que estos sean capaces de obtener resultados antes de que concluya la etapa de obtención de muestras reales. El alcance práctico que tienen los datos sintéticos en esta aplicación corresponde a agilizar el proceso de creación y entrenamiento de los modelos, facilitando la obtención de un modelo funcional que pueda ser utilizado con los datos reales para generar valor o descubrimientos.

Junto con esto, un problema recurrente de investigadores en estudios clínicos es no poder llegar a conclusiones dada la escasez de datos en sus proyectos. Según un estudio conducido por Oracle en 2018 [88], al preguntarle a investigadores sobre cuáles son los riesgos más grandes para una investigación de ensayos clínicos, un 49% declara que no tener todos los datos para determinar la eficacia de un tratamiento es uno de los problemas más relevantes en el rubro, siendo catalogado como el segundo más importante. Otro beneficio de utilizar datos sintéticos es poder llegar a conclusiones y aplicaciones en estas investigaciones que se estancan por no tener suficientes datos. Si bien este beneficio es relevante, se desconoce la información de las investigaciones que no llegan a conclusiones, sobre todo en nuestro país, ya que estas no son publicadas en la mayoría de los casos. Es por esta razón que no se generará una estimación monetaria de este beneficio en esta memoria, pero será declarado como trabajo futuro, esperando

que existan indicadores que permitan estimar el valor que se pierde al no poder concluir una investigación.

Utilizando la información sobre los costos de la etapa de procesamiento de datos dispuesta en 6.1, se obtiene un costo mensual de $1.640.690 pesos para esta etapa. Juntando esto con el tamaño del mercado de investigación en torno al área de la salud que ocupan datos explicada en 6.2, es posible estimar los beneficios económicos de utilizar datos sintéticos, planteando cuatro escenarios en donde gradualmente se incrementa la cantidad de proyectos que se atrasan por tener escasez de datos con diferentes periodos de atraso. Para estos efectos se construye la siguiente tabla que expresa los resultados de este procedimiento.

Tabla 6.6: Costos generados por el atraso de distintas cantidades de proyectos en diferentes tiempos. Elaboración propia.

	3 Meses	**6 Meses**	**9 Meses**	**12 Meses**
1 Proyecto	$ 4.922.080	$ 9.844.160	$ 14.766.240	$ 19.688.320
8 Proyectos (25% del total)	$ 39.376.640	$ 78.753.280	$118.129.920	$157.506.560
17 Proyectos (50% del total)	$ 83.675.360	$167.350.720	$251.026.080	$334.701.440
25 Proyectos (75% del total)	$123.052.000	$246.104.000	$369.156.000	$492.208.000
33 Proyectos (100% del total)	$162.428.640	$324.857.280	$487.285.920	$649.714.560

Como se expresa en la tabla 6.6, existen diferentes escenarios que permiten ver los efectos de que un proyecto se atrase por no tener suficientes datos o no contar con los sujetos necesarios para poder generar resultados. Teniendo en cuenta que los montos máximos que entregan los concursos varían, con el concurso FONIS entregando $60 millones de pesos [80], mientras que el concurso FONDECYT provee $57 millones de pesos [81] e IDeA I+D $200 millones de pesos [82], estos atrasos pueden ser un factor importante al momento de considerar el financiamiento de un proyecto de esta índole. Por ejemplo, si un proyecto financiado con el máximo cupo del concurso FONIS se atrasa 6 meses y debe seguir financiando a su equipo de procesamiento de datos, este perdería una suma de $9.844.160 o el equivalente al 16.4% de su presupuesto total. Este tipo de atrasos puede ser el factor clave que diferencia a proyectos capaces de generar

conocimiento en base a lo investigado frente a otros que no obtienen resultados concluyentes.

A un nivel más global, estas pérdidas pueden alcanzar los $640 millones de pesos en un escenario pesimista en donde todos los proyectos se atrasan, $334 millones en un escenario intermedio y $157 millones en un escenario optimista en donde sólo se atrasan 8 proyectos. Esta proyección muestra el impacto que pueden tener los datos sintéticos en esta etapa, permitiendo a los investigadores abaratar costos en los casos de atraso o frente a la insuficiencia de sujetos dentro de una investigación clínica. No es posible dejar de lado que se trabaja con recursos públicos en los tres concursos mencionados, por lo que lograr eficientizar este tipo de proyectos genera un beneficio social, aprovechando de mejor manera los recursos disponibles para estos fines.

Capítulo 7

Datos sintéticos como un producto o servicio

Es interesante poder revisar no sólo la generación y evaluación de datos sintéticos, sino cómo estos pueden ser integrados dentro de un modelo de negocios como un producto o servicio en una institución de salud real. En este capítulo se revisará el proceso de creación de un servicio o producto a través de la realización de un *benchmark* revisando aplicaciones similares y tomando en cuenta también las restricciones que existen en el área de salud.

7.1 Aplicaciones que utilizan datos sintéticos

7.1.1 M-Sense

M-Sense [72] es una aplicación alemana que permite a sus usuarios registrar sus dolores de cabezas o migrañas a través de una aplicación móvil. Esta app además entrega información útil para los usuarios, que les permite entender el impacto de su estilo de vida en los dolores de cabeza, reconocer síntomas, evaluar los patrones que existen en la frecuencia de estos y entrega métodos validados para evitar o aliviar los síntomas. Junto con esto, utilizan datos sintéticos para anonimizar los datos ingresados por los usuarios y entregar esta información a comunidades científicas, permitiéndoles investigar patrones sobre esta enfermedad [73].

7.1.2 Medkit-learn

Es un paquete de Python de uso público que permite a los usuarios tener un acceso fácil a data clínica de alta fidelidad [74]. Esta herramienta utiliza datos sintéticos para entrenar una variedad de modelos que permiten replicar situaciones médicas reales, además permite representar el comportamiento humano basado en decisiones [76]. Con estas características fomenta el ecosistema de investigación dentro de estas áreas, ya que garantiza el acceso a información sensible sin perjudicar la privacidad de los pacientes.

7.1.3 DECAF

DECAF es otra librería de Python que se preocupa de los sesgos que puedan tener los datos, buscando poder entrenar modelos más justos [75]. Existen distintos sesgos de género, etnia, sociales, etc. Que residen en los datos que se manejan actualmente, y estos son replicados naturalmente cuando se entrenan modelos con esta información. Es por esta razón que DECAF busca a través del uso de datos sintéticos evitar estos sesgos generando indicadores y estándares de justicia, evitando que se repliquen estas desigualdades, existe un enfoque también en que la calidad de los datos no disminuya a causa de este proceso.

De este proceso se rescatan características en común como el respeto a la privacidad de los usuarios, evitar sesgos que puedan contener los datos, facilidad de uso de las aplicaciones o soluciones, simulación de datos y eventos reales y por sobre todo un apoyo a investigadores en el área de la salud al poder brindar acceso a información anónima y de calidad.

7.2 Restricciones en el área de la salud

Dentro de las aplicaciones de aprendizaje de máquinas en el área de la salud, existen restricciones que impiden a los científicos de datos poder utilizar todas las herramientas existentes en el estado del arte, las que ya han sido exploradas con anterioridad en este trabajo de tesis. A continuación, se mencionan brevemente estas restricciones.

1. **Conectividad a servidores online o plataformas Cloud:** Establecer políticas de conectividad, uso de la información o plataformas Cloud se transforma en un proceso más complejo, ya que mantener la privacidad de los pacientes es prioridad, haciendo que los costos de implementación sean muy altos, además, una infracción a la seguridad puede tener consecuencias graves [84].

2. **Baja accesibilidad de la información:** Nuevamente la justificada preocupación por la seguridad de la información dificulta el acceso a los datos para un equipo de investigadores. Esta barrera frena posibles investigaciones que puedan aportar valor a la industria.

3. **Escaso número de sujetos:** Un bajo número de pacientes participando en un experimento hacen que sea difícil poder generalizar los resultados, u obtener resultados que permitan demostrar la hipótesis a estudiar.

7.3 Modelando datos sintéticos como un producto o servicio

Teniendo en cuenta las aplicaciones que utilizan datos sintéticos en sus funcionalidades, y las restricciones que existen en el área de la salud, es que se prosigue a modelar una alternativa de servicio que permita cumplir con los requerimientos y estándares de los usuarios dentro de este mercado.

La aplicación buscada debe conservar la privacidad de los pacientes, por lo que se descarta un servicio personalizado en donde los usuarios carguen y envíen sus datos para poder ser procesados por un tercero, además, deberá permitir a los usuarios generar datos sintéticos de manera sencilla y evaluar el nivel de similitud y privacidad que tienen los datos generados, entregando información sobre que significa tener buenos resultados en ambos ámbitos para su fácil entendimiento.

Al no poder acceder a los datos en tiempo real siendo un ente externo, y dada la complejidad de mantener un estándar de seguridad de información en la red de salud nacional como para instaurar un servicio en la nube, es que se escoge la opción para proveer este servicio de manera más sencilla; mediante un software como un servicio que deberá manipular el usuario final y no almacenará la información sobre los pacientes. Este producto se destinará para investigadores que manipulen datos y requieran aumentar o anonimizar la información que manejan.

Para estos fines, este software automatizado deberá permitir a los investigadores diferenciar los tipos de datos que poseen, describiendo cuales son datos categóricos, numéricos o fechas y los rangos en los que se mueven las variables numéricas. Posterior a este proceso, y para utilizar diferentes algoritmos generadores, se proseguirá a ofrecer distintas opciones en la arquitectura de estos, por ejemplo, para los que utilizan redes neuronales, se ofrecerán varias arquitecturas con distintas épocas, capas, tamaño de lote y nivel de aprendizaje. Finalizado esta generación, se ofrecerán herramientas para medir el rendimiento y la anonimidad de los datos, evaluando todos los algoritmos utilizando estas técnicas para entregar una recomendación final. En esta etapa, deberá ser posible utilizar modelos básicos de clasificación y regresión para la evaluación, pero también se deberá dar la oportunidad de utilizar los propios modelos que el usuario desee cargar, buscando replicar la situación real en la que serán utilizados los datos para medir su rendimiento.

Esto permitirá a investigadores utilizar una herramienta flexible en casos en los que requieran generar datos para aumentar la cantidad de pacientes que existen dentro de las investigaciones con el fin de entrenar modelos de aprendizaje de máquinas, permitiendo además anonimizar los datos con el fin de facilitar la divulgación de estos. Todo esto, tomando en cuenta las restricciones que existen dentro del área de la salud

para la manipulación de información de pacientes y también basándose en las soluciones que existen actualmente en el mercado.

7.4 Requerimientos funcionales y no funcionales

Con el fin de formalizar una solución que permita cumplir con lo estipulado en el apartado anterior, se buscan definir los requerimientos funcionales y no funcionales de la aplicación que se busca crear. En base a [89] se definen los requerimientos funcionales como todo lo que debe proveer un servicio, que características tiene y que funciones cumple, por otro lado, los requerimientos no funcionales son aquellos que describen las propiedades generales del sistema. Estos se detallan a continuación:

- **Requerimientos funcionales:**

    1. **¿Qué datos deben ser ingresados al sistema?:** De los usuarios se esperan cinco inputs principales; La carga de datos manual desde un archivo csv o Excel descargado, posible registro manual de los tipos de datos, selección de arquitecturas de los algoritmos, selección de los modelos de evaluación y carga de modelos propios si es que lo requieren.

    2. **¿Qué mostrará el sistema?:** El sistema tendrá una pantalla de inicio que explique lo que hace la aplicación, otra que permita cargar los datos y seleccionar los tipos de variables, una pantalla de espera mientras se entrenan los modelos que muestre el estado del proceso, una que muestre las evaluaciones en un reporte y también le permita a los investigadores descargar los datos.

    3. **¿Cómo son los flujos de trabajo que realizará el sistema?:** El sistema utilizará los datos cargados y los tipos de datos ingresados para entrenar la cantidad de algoritmos que el usuario requiera, en base a las arquitecturas definidas, posterior a esto, generará datos sintéticos con estos y los evaluará con modelos predefinidos o alguno que el usuario decida cargar.

    4. **¿Qué reportes u outputs tendrá el sistema?:** El sistema tendrá dos diferentes entregas, por un lado, están los datos generados sintéticamente y por otro un reporte que muestre las métricas de los modelos de evaluación y gráficos de las variables que el usuario escoja.

    5. **¿Quién puede cargar datos a la aplicación?:** Sólo los usuarios a través de un archivo csv o Excel, para evitar el uso de credenciales y conexiones a data

warehouses, es posible en un futuro integrar estas conexiones si es que son requeridas, pero requieren de un trabajo personalizado con el usuario para evitar pérdidas de información.

6. **¿Cómo cumplirá el sistema la regulación pertinente?:** El sistema no almacenará datos en ninguna nube, será un ejecutable que le permita al usuario crear datos de manera remota a través de una aplicación, con esto se espera mantener el nivel de seguridad de los datos de los pacientes, evitando fugas de información.

- **Requerimientos no funcionales:**

1. **Seguridad:** El sistema no tendrá acceso a la web y será utilizado exclusivamente con archivos planos, con el fin de evitar fugas de información. Además, será utilizado exclusivamente por los científicos de datos que estén a cargo de un proyecto relacionado, pensando en que ya tienen acceso a estos datos sensibles. El sistema no almacenará de ninguna forma los datos generados.

2. **Usabilidad:** El sistema está pensado para ser utilizado por científicos de datos que estén familiarizado con Python o R, pero no se espera un nivel de sofisticación mayor a cargar archivos y seleccionar características que deseen. De todas maneras, se espera el conocimiento por parte del usuario sobre qué son los modelos y sus variables, para poder tener buenos resultados. Además, el sistema debe poseer un manual de usuario y explicar en cada etapa lo que se realiza en el backend, para transparentar la utilización de la información sensible. Es importante que esta aplicación sea de código abierto para que los investigadores puedan revisar si es que sus datos están en riesgo en todo momento. Los reportes y gráficos que entrega el sistema deben ser autocontenidos, todos los gráficos deben tener correctas sus etiquetas y título.

Con respecto a la arquitectura tecnológica de esta aplicación, será desarrollada en Python a través de las siguientes librerías.

- **Pandas [90]:** Librería que permite el manejo de bases de datos dentro de Python.
- **Messytables [91]:** Esta librería permite identificar el tipo de dato contenido en cada variable de manera automática, con el fin de clasificarlos y poder entregarles a los modelos generadores los datos etiquetados.
- **Pickle [96]:** Para permitir que el usuario cargue sus propios modelos, Pickle codifica y decodifica modelos pre entrenados y permite utilizarlos de una manera sencilla. No requiere que el modelo se entrene nuevamente o que se le entreguen

los parámetros, sólo se necesita cargar un archivo .pkl que contenga la información de este para poder generar predicciones o muestras.

- **SDV [55]:** Synthethic Data Vault es una librería que permite generar registros sintéticos con diferentes herramientas de manera rápida e intuitiva.
- **Scikit-Learn [92]:** Esta librería permitirá obtener predicciones y métricas de los modelos generados por los usuarios.
- **Seaborn [93]:** Librería que permite utilizar una mayor variedad de gráficos, siendo más agradables visualmente, lo que mejora la experiencia de usuario.
- **Streamlit [94]:** Streamlit permite montar una aplicación que llevará a los usuarios cargar sus datos, seleccionar variables, generar datos, cargar sus propios modelos y finalmente recibir un reporte y los datos que generaron en el proceso.
- **Docker [95]:** Esta librería permite generar un container o imagen del código con el que se trabaja, su principal función en este caso es mantener una versión estable de Python y todas las librerías mencionadas anteriormente, con el fin de poder ejecutar esta aplicación en cualquier máquina sin tener que preocuparse por tener las mismas versiones que el computador en el que se programó la aplicación.

A continuación, se presenta un esquema que muestra cómo funcionaría la aplicación, en conjunto con la arquitectura tecnológica que compone cada etapa.

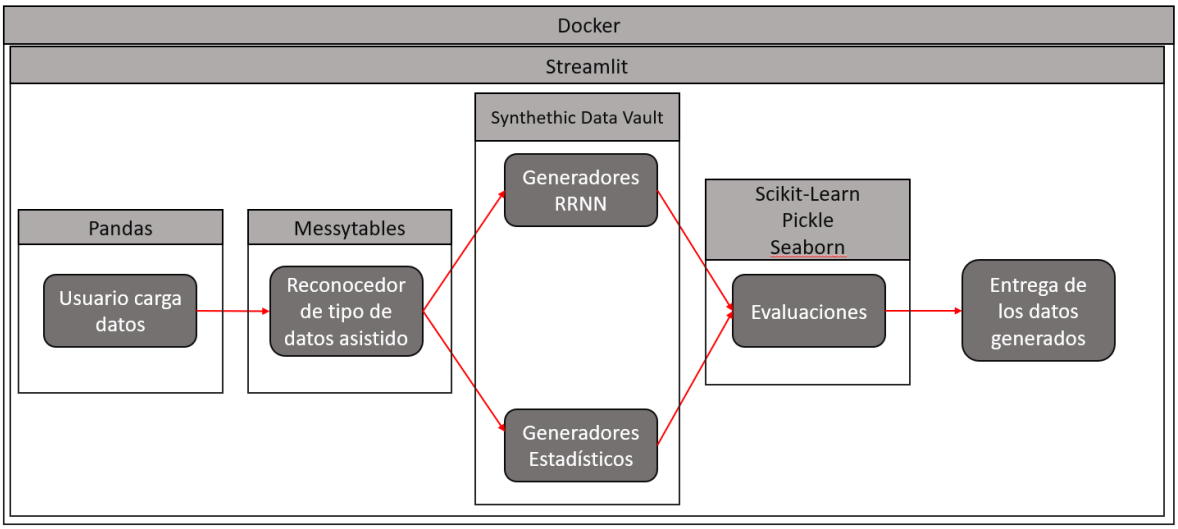

Figura 7.1: Diagrama del sistema de la aplicación y bloques tecnológicos que lo componen. Elaboración Propia.

La aplicación trabajará de la siguiente manera: Primero, el usuario cargará una base de datos al sistema y ésta se leerá utilizando Pandas, para clasificar los tipos de datos habrán dos opciones, de manera manual o automática utilizando messytables, posterior a este proceso se generarán datos utilizando la librería SDV con los algoritmos que el usuario seleccione, para posteriormente generar un reporte de evaluación utilizando un modelo de clasificación predeterminado o uno que cargue el usuario ocupando Pickle.

Las métricas de Accuracy, Recall o Precision se calcularán utilizando Scikit-Learn y se graficarán las variables a través de Seaborn. Para crear una aplicación ejecutable se utilizará Streamlit, que permite montar todas las librerías mencionadas anteriormente, y para poder generar una versión estable, se trabajará con Docker, creando una imagen de la aplicación que no dependerá de las versiones de las librerías que tenga el usuario.

Finalmente, y como se trabajará sin conexión a internet, es posible entregar una aplicación ejecutable que utilice los archivos en formato csv o Excel que manejen los investigadores y que utilice la información contenida ahí para generar datos sintéticos y reportes sobre la generación de estos. Con esto se busca entregar una herramienta de sencillo uso que permita a investigadores probar diferentes generadores, crear datos sintéticos y evaluar su desempeño, con el fin de expandir la utilización de estas herramientas de una manera más amigable para el usuario, entregándoles nuevas posibilidades a investigadores y buscando solucionar los problemas asociados a no tener una cantidad de datos suficiente.

Capítulo 8

Conclusiones y recomendaciones

En el siguiente capítulo se presentan las conclusiones más importantes de este trabajo de título, mostrando el conocimiento obtenido en base a los resultados y la experiencia desarrollada. Además, se proponen recomendaciones para avanzar en los aspectos que no fueron abordados en este trabajo por estar fuera de los alcances definidos previamente.

8.1 Conclusiones

En el capítulo 1 de este documento se plantearon 5 objetivos específicos para lograr cumplir con el objetivo general, el cual corresponde a entregar recomendaciones para resolver los problemas asociados a la escasez de datos en proyectos de aprendizaje de máquinas asociados a salud, en base a la experiencia demostrativa del proyecto ADDAI. Estos objetivos fueron cumplidos y a continuación se detalla brevemente los resultados obtenidos.

1. Del estudio y análisis del estado del arte se logró llegar a la conclusión de que el no contar con una cantidad de datos suficientes para poder investigar un fenómeno sí es un problema real que ha sido recurrente en los últimos años, sobre todo en las áreas de investigación clínica. Además, existe un área de la investigación enfocada en aplicar herramientas de aprendizaje de máquinas en el área de la salud, la cual crece año a año, por lo que el problema de tener pocos datos en el rubro de la investigación será cada vez más relevante con el paso del tiempo. Finalmente, existen diversas maneras de generar datos sintéticos en la literatura, algunas herramientas utilizan sistemas complejos, como redes neuronales, mientras que otros utilizan algoritmos estadísticos o transformaciones lineales. El desafío más relevante en este rubro es superar el complejo de la caja negra, en donde se pierde la comprensibilidad de la información a cambio de un mejor rendimiento de los modelos.

2. Con respecto a la generación de datos sintéticos mediante diferentes herramientas generativas utilizando como experiencia demostrativa al proyecto ADDAI, se concluye que esta generación fue exitosa para los algoritmos GAN, VAE y GC. Se afirma esto pues los resultados obtenidos son aceptables al momento de ser comparados con los datos reales. Al finalizar este proceso, se obtuvieron 15 bases de datos sintéticos, 9 de ellas provienen de utilizar un algoritmo en cada base de datos y el resto corresponden a los generados en la base de datos agrupados, una vez por cada algoritmo y una vez para una cantidad de 38 y 150 filas.

3. De la etapa de evaluación de los datos sintéticos, se destaca que el algoritmo con mejor rendimiento en la mayoría de las etapas, ya sea utilizando el índice de similitud o la evaluación con un modelo de clasificación, es Gaussian Cópula, seguido por TVAE. Por otra parte, se concluye que los mejores resultados obtenidos para los tres algoritmos se alcanzaron al combinar los datos sintéticos con los datos reales. Pese a esto, en ninguno de las evaluaciones se obtuvo un modelo más certero que el original entrenado con datos reales.

4. Se logró realizar una propuesta de servicios utilizando datos sintéticos como el eje principal. En esta propuesta se estudiaron diferentes aplicaciones de estos datos y se sumaron las restricciones del sistema de salud para proponer un software que permita a investigadores generar datos sintéticos a través de distintas fuentes, entregando evaluaciones y métricas que les permitan distinguir la calidad y privacidad de los resultados.

5. Finalmente, de la evaluación del impacto económico de la aplicación de datos sintéticos, que se basó únicamente en investigación pública, se estima que es posible ahorrar en proyectos similares al proyecto ADDAI evitando atrasos en un monto de hasta $334 millones de pesos al año en un escenario intermedio, siendo estos fondos de uso público. Se concluye que esta suma es relevante para el fomento de la investigación y el desarrollo en Chile, por lo que beneficiaría a una gran cantidad de instituciones y personas a futuro.

Se logra entonces cumplir los 5 objetivos específicos propuestos al inicio de este trabajo de título, por lo que es posible afirmar que se cumple el objetivo general a cabalidad. El único punto restante es poder entregar recomendaciones para aplicaciones futuras de los métodos utilizados en a lo largo de este trabajo, lo cual se procede a realizar en la siguiente sección.

8.2 Recomendaciones

A partir del trabajo realizado, se proponen recomendaciones que aportan continuidad del trabajo futuro para la incorporación de algoritmos generativos en proyectos de investigación.

1. **Preprocesamiento de los datos:** Al momento de generar datos sintéticos es natural cuestionarse cuando será el mejor momento para utilizar a los algoritmos generativos, en este contexto existen dos opciones claras, utilizarlos previamente al preprocesamiento de datos, o después. Si bien en la literatura no se presentan mejorías para los modelos que generen datos de manera previa o posterior, si es relevante entender que sucede al momento de trabajar con diferentes fuentes de datos. En estos casos, la generación de datos debe ser posterior al proceso de preprocesamiento, en donde se agrega la información proveniente de todas las

fuentes de información. Esto sucede ya que al generar más de una base de datos sintéticos, estos algoritmos son incapaces de mantener la unicidad de los pacientes individuales, en otras palabras, un paciente “n” puede pertenecer al grupo de control en una base de datos y en otra pertenecer al grupo de estudio.

2. **Elección de algoritmos generativos:** Escoger un algoritmo generativo puede ser una tarea abrumadora en algunos casos, dada la variedad y las variantes de cada algoritmo que existen actualmente. Si bien no existe una guía de que algoritmos utilizar dado el contexto de los datos, en este trabajo sólo se revisó un tipo de base de datos, que corresponde a aquellas con varias variables numéricas y una binaria, para estos casos, el algoritmo con mejor resultado es Gaussian Copula, mientras que los algoritmos de redes neuronales mejoraban sus resultados al tener una mayor cantidad de variables. Para casos más generales, se recomienda tener una aproximación más amplia, probando dos o tres algoritmos y evaluando los resultados, si es que no son suficientes estos resultados, se recomienda probar variantes del algoritmo con mejores resultados en la etapa anterior.

3. **Arquitectura de los algoritmos:** Actualmente existen bibliotecas de Python que facilitan la generación de datos a través de estos algoritmos de redes neuronales, pero una gran interrogante es qué arquitectura elegir para dichos algoritmos. Escoger la cantidad de capas de una red neuronal, la velocidad de aprendizaje e incluso las épocas y el tamaño del lote es una tarea que no es sencilla y puede tener grandes repercusiones en los resultados finales. Se recomienda en esta etapa tener una aproximación similar a la que se tuvo en el análisis de sensibilidad, donde se prueban muchos modelos ligeramente diferentes para revisar cómo cambian los resultados en torno a pequeñas modificaciones en la arquitectura, con el fin de encontrar la que mejor funcione. Además, se puede afirmar en base a la experiencia adquirida que definir el tamaño del lote a una cantidad más grande que la cantidad de filas que tiene la base de datos, sólo hace que las épocas avancen más rápido, lo que implica que la etapa de aprendizaje es más corta, esto puede perjudicar los resultados obtenidos. Junto con esto, y por lo visto en el apartado 5.3, se puede afirmar que no siempre una mayor cantidad de épocas entregan un mejor resultado para los algoritmos GAN y VAE.

4. **Formas de evaluación:** Existen varias maneras de evaluar la calidad de los datos generados, algunas involucran crear estadísticos que muestren la distancia entre los datos reales y sintéticos, mientras que otros utilizan modelos de clasificación para determinar si un dato proviene de la misma distribución que los datos reales o no. Si bien la diferencia entre ambos está en la profundidad del análisis que permiten realizar, las conclusiones tienden a ser similares. Por otra parte, es

posible evaluar de manera gráfica los resultados de la generación de datos como en el apartado 4.3, pero al contar con bases de datos con muchas variables, este análisis pierde relevancia frente a otros que permiten resumir los resultados de manera más compacta en sólo una tabla. La recomendación en esta etapa es ir incrementando la complejidad del análisis paulatinamente, es decir, utilizar una combinación entre análisis gráfico en una primera etapa, seguido de algún estadístico que permita comprender hacia donde se dirigen los resultados y terminar con un análisis más profundo utilizando modelos de clasificación. Todo esto con el fin de evitar realizar análisis muy complejos cuando los resultados iniciales no son los esperados, ahorrando tiempo a investigadores.

5. **Aplicaciones:** Con respecto a las aplicaciones que pueden tener los datos sintéticos en el rubro de la investigación, se destaca que los mejores resultados en este trabajo de títulos se obtuvieron al combinar datos reales y sintéticos, formando una base de datos de 76 filas. Se destaca esta información ya que estos resultados son mejores a los obtenidos con bases de datos sintéticos con 150 filas. Esto quiere decir que siempre debe privilegiarse la calidad de la información sobre la cantidad, pese a tener la posibilidad de generar una cantidad indefinida de datos. La recomendación en este punto corresponde a aplicar estas herramientas siempre que sea posible en combinación con los datos reales, para poder obtener los mejores resultados posibles, además, no siempre una mayor cantidad es mejor que una cantidad moderada de datos sintéticos.

6. **Trabajo futuro:** Dado el alcance de esta memoria, no se han podido abarcar algunos puntos relevantes para continuar la investigación y aplicación de datos sintéticos para resolver el problema de la escasez de datos en el área de la salud. Estos puntos se detallan a continuación.

    a. **Generación para datos no procesados:** Una línea de investigación que no se abordó en esta memoria corresponde a generar datos desde la fuente de origen, en este caso, directamente desde las series de tiempo que generan las herramientas de medición del proyecto ADDAI. Se sugiere explorar el rendimiento de los algoritmos aplicados en este trabajo de título a los datos no procesados, o sea, directamente a las series de tiempo y no a los datos tabulares ya procesados.

    b. **Variantes de algoritmos generativos:** Como se explicó en el marco teórico, existen muchas variantes para cada algoritmo generativo, y posiblemente en el futuro siga aumentando esta cantidad de variantes disponibles. Si bien estas son muy dependientes del modelo y los datos con los que se trabajan, y por esta misma razón no se utilizaron en esta memoria, se recomienda en una primera instancia realizar un análisis de

cómo funcionan estas variantes, para posteriormente poder generar datos sintéticos con ellas. Por lo que se revisó en el estado del arte, estas variantes permiten obtener mejores resultados para algunos tipos de datos, por lo que estudiarlas podría fomentar aún más la aplicación de estas herramientas para evitar la escasez de datos en proyectos de investigación.

c. **Evaluación de la privacidad de los datos:** Si bien durante este trabajo de título se utilizaron algoritmos que permiten anonimizar la información, no se estudió la pérdida o el cambio de privacidad de los datos por falta de tiempo y quedar fuera de los alcances. Actualmente existen métricas y evaluaciones que permiten demostrar cuanta información de los datos originales se traspasa al modelo, con el fin de evitar posibles fugas de información, las cuales tienen diversas consecuencias dependiendo de la industria. Se sugiere en este punto poder estudiar con una mayor profundidad cómo se evalúa la privacidad de un modelo, como se conserva y mediante qué herramientas se mide, con el fin de poder mantener un alto nivel de seguridad y resguardo al momento de compartir información sensible.

Bibliografía

[1] Web Intelligence Centre. Home. Recuperado 11 de julio de 2021. https://wic.uchile.cl/

[2] Agencia Nacional de Investigación y Desarrollo. (2020). Resultados concuros IDeA I+D 2020: https://s3.amazonaws.com/documentos.anid.cl/fondef/2020/ideaid/Res9413_IDeA2020.pdf

[3] Ley N° 20.285. Diario Oficial de la República de Chile, 20 de agosto de 2008. http://bcn.cl/2f8ep

[4] Ley N° 19.628. Diario Oficial de la República de Chile, 28 de agosto de 1999. http://bcn.cl/2f7cg

[5] Ley N° 19.223. Diario Oficial de la República de Chile, 07 de junio de 1993. http://bcn.cl/2gf9s

[6] Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration ofpValues for Testing Precise Null Hypotheses. The American Statistician, 55(1), 62–71. https://www.tandfonline.com/doi/abs/10.1198/000313001300339950

[7] Colquhoun D. 2014 An investigation of the false discovery rate and the misinterpretation ofp-values.R. Soc. open sci.1: 140216. http://dx.doi.org/10.1098/rsos.140216

[8] Science Direct. (2021). Sample Variability. [En Línea] https://www.sciencedirect.com/topics/mathematics/sample-variability

[9] Simons, Daniel J.; Shoda, Yuichi; Lindsay, D. Stephen (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. Perspectives on Psychological Science. https://pubmed.ncbi.nlm.nih.gov/28853993/

[10] Zhao, Wei. (2017). Research on the deep learning of the small sample data based on transfer learning. AIP Conference Proceedings. https://doi.org/10.1063/1.4992835

[11] Decreto 725. Código Sanitario. 31 de enero de 1968. https://www.bcn.cl/leychile/navegar?idNorma=5595

[12] Tucker, A., Wang, Z., Rotalinti, Y., & Myles, P. (2020). Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ digital medicine, 3(1), 147. https://doi.org/10.1038/s41746-020-00353-9

[13] Goncalves, A., Ray, P., Soper, B. et al. Generation and evaluation of synthetic patient data. BMC Med Res Methodol 20, 108 (2020). https://doi.org/10.1186/s12874-020-00977-1

[14] Bourou, S., El Saer, A., Velivassaki, T.-H., Voulkidis, A., Zahariadis, T. A Review of Tabular Data Synthesis Using GANs on an IDS Dataset. Information 2021, 12, 375. https://doi.org/10.3390/ info12090375

[15] Derclaye, Estelle. (2005). What is a Database?. The Journal of World Intellectual Property. http://dx.doi.org/10.1111/j.1747-1796.2002.tb00189.x

[16] Python Software Foundation. (2021). The Python Tutorial — Python 3.9.6 documentation. [En Línea] https://docs.python.org/3/tutorial/

[17] Ayodele, Taiwo. (2010). Types of Machine Learning Algorithms. https://doi.org/10.5772/9385

[18] Kao, S.-C., Kim, H. K., Liu, C., Cui, X., & Bhaduri, B. L. (2012). Dependence-Preserving Approach to Synthesizing Household Characteristics. Transportation Research Record, 2302(1), 192–200. https://doi.org/10.3141/2302-21

[19] Li, Z., Zhao, Y., & Fu, J. (2020). SynC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources. *2020 International Conference on Data Mining Workshops (ICDMW)*, 571-578. https://arxiv.org/abs/2009.09471

[20] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27. https://arxiv.org/pdf/1406.2661.pdf

[21] Hayes, J., Melis, L., Danezis, G., & De Cristofaro, E. (2017). Logan: Membership inference attacks against generative models. https://arxiv.org/abs/1705.07663

[22] Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. https://arxiv.org/abs/1701.00160

[23] Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2794-2802). https://doi.org/10.1109/ICCV.2017.304

[24] Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214-223). PMLR. https://arxiv.org/abs/1701.07875

[25] Larsen, A.B.L., Sønderby, S.K., Larochelle, H. & Winther, O.. (2016). Autoencoding beyond pixels using a learned similarity metric. Proceedings of The 33rd International Conference on Machine Learning, in Proceedings of Machine Learning Research http://proceedings.mlr.press/v48/larsen16.html

[26] Odena, A., Olah, C., & Shlens, J. (2017, July). Conditional image synthesis with auxiliary classifier gans. In International conference on machine learning (pp. 2642-2651). PMLR. https://arxiv.org/abs/1610.09585

[27] Elbattah, Mahmoud & Loughnane, Colm & Guerin, Jean-Luc & Carette, Romuald & Cilia, Federica & Dequen, Gilles. (2021). Variational Autoencoder for Image-Based Augmentation of Eye-Tracking Data. Journal of Imaging. https://doi.org/10.3390/jimaging7050083

[28] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. https://arxiv.org/abs/1312.6114

[29] Aznan, N. K. N., Atapour-Abarghouei, A., Bonner, S., Connolly, J. D., Al Moubayed, N., & Breckon, T. P. (2019, July). Simulating brain signals: Creating synthetic eeg data via neural-based generative models for improved ssvep classification. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE. https://arxiv.org/abs/1901.07429

[30] Hartmann, K. G., Schirrmeister, R. T., & Ball, T. (2018). EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. https://arxiv.org/abs/1806.01875

[31] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. Advances in neural information processing systems, 29, 2234-2242. https://arxiv.org/pdf/1606.03498.pdf

[32] Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., & Rohde, G. K. (2019). Generalized sliced Wasserstein distances. https://arxiv.org/abs/1902.00434

[33] Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability

of neuroscience. Nature Reviews Neuroscience, 14(5), 365–376. https://doi.org/10.1038/nrn3475

[34] Al-Janabi, Samaher & al-bakry, Abbas. (2010). Genetic Programming Data Construction Method to Handle Data Scarcity Problem. International Journal of Advancements in Computing Technology (IJACT). https://www.researchgate.net/publication/214699989_Genetic_Programming_Data_Construction_Method_to_Handle_Data_Scarcity_Problem

[35] Tsai, C.-H., & Li, D.-C. (2015). Improving Knowledge Acquisition Capability of M5' Model Tree on Small Datasets. 2015 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence. https://doi.org/10.1109/ACIT-CSI.2015.72

[36] TR Addis. Towards an" expert" diagnostic system. ICL Technical Journal, 1:79–105, 1956.

[37] McKelvey, T. & Ahmad, Muhammad & Teredesai, Ankur & Eckert, Carly. (2018). Interpretable Machine Learning in Healthcare. https://www.researchgate.net/publication/328416903_Interpretable_Machine_Learning_in_Healthcare

[38] Shatte ABR, Hutchinson DM, Teague SJ (2019). Machine Learning in mental health: a scoping review of methods and applications. Psychological Medicine 1–23. https://doi.org/10.1017/S0033291719000151

[39] Mayo Clinic. (2019). Alzheimer's or depression: Could it be both? https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers/art-20048362.

[40] Red Hospital Clínico Universidad de Chile. (2021). Aranceles HCUCH https://www.redclinica.cl/Portals/_default/Skins/Skin_HCUCH_17_03/images/arancel_2021_publicacion_3.pdf

[41] Kalra, Anubha & Anand, Gautam & Lowe, Andrew. (2020). Interpreting Electroencephalogram (EEG) – An Introductory Review of Assessment and Measurement Procedures. Modern Applied Science. 14. 47. https://doi.org/10.5539/mas.v14n6p47.

[42] Uhlhaas, P. J., & Singer, W. (2006). *Neural Synchrony in Brain Disorders: Relevance for Cognitive Dysfunctions and Pathophysiology. Neuron, 52(1), 155–168.* https://doi.org/10.1016/j.neuron.2006.09.020

[43] Abhang, Priyanka & Gawali, Bharti & Mehrotra, Suresh. (2016). Technical Aspects of Brain Rhythms and Speech Parameters. http://dx.doi.org/10.1016/B978-0-12-804490-2.00003-8

[44] Foster, Joshua & Sutterer, David & Serences, John & Vogel, Edward & Awh, Edward. (2017). Alpha-Band Oscillations Enable Spatially and Temporally Resolved Tracking of Covert Spatial Attention. Psychological Science. 28. http://dx.doi.org/10.1177/0956797617699167

[45] Lega, B. C., Jacobs, J., & Kahana, M. (2012). Human hippocampal theta oscillations and the formation of episodic memories. *Hippocampus*, *22*(4), 748–761. https://doi.org/10.1002/hipo.20937

[46] Pilon, M., Zadra, A., Joncas, S., & Montplaisir, J. (2006). Hypersynchronous delta waves and somnambulism: brain topography and effect of sleep deprivation. *Sleep*, *29*(1), 77–84. https://doi.org/10.1093/sleep/29.1.77

[47] Roth, H. R. Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE Trans. Med. Imaging 35, 1170–1181 (2016). https://arxiv.org/abs/1505.03046

[48] Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D. & Xiao, X. PrivBayes: private data release via Bayesian Networks. ACM Trans. Database Syst. 42, 25 (2017). https://doi.org/10.1145/3134428

[49] Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. J. AI Res. 16, 321–357 (2002). https://arxiv.org/abs/1106.1813

[50] Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. Dental Press Journal of Orthodontics, 19(4), 27–29. https://doi.org/10.1590/2176-9451.19.4.027-029.ebo

[51] Hapfelmeier, A., Hothorn, T., Ulm, K. & Strobl, C. A new variable importance measure for random forests with missing data. Stat. Comput. 24, 21–34 (2014). http://dx.doi.org/10.1007/s11222-012-9349-1

[52] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. https://arxiv.org/abs/1907.00503

[53] Neha Patki, Roy Wedge, Kalyan Veeramachaneni. The Synthetic Data Vault. IEEE DSAA 2016. http://dx.doi.org/10.1109/DSAA.2016.49

[54] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2021). CTGAN Library. https://github.com/sdv-dev/CTGAN

[55] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2021). SDV Library. https://github.com/sdv-dev/SDV

[56] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2021). Gaussian Copula Function https://sdv.dev/SDV/user_guides/single_table/gaussian_copula.html

[57] Dankar, Fida & Ibrahim, Mahmoud. (2021). Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. Applied Sciences. 11. 2158. https://doi.org/10.3390/app11052158

[58] Massachusetts Institute of Technology. (2022). Synthethic Data Vault Evaluation Framework. https://sdv.dev/SDV/user_guides/evaluation/evaluation_framework.html

[59] Massachusetts Institute of Technology. (2022) Synthethic Data Vault multi table evaluation metrics. (2022). https://sdv.dev/SDV/user_guides/evaluation/multi_table_metrics.html

[60] Vourvopoulos, Athanasios & Liarokapis, Fotis & Chen, Monchu. (2015). The Effect of Prior Gaming Experience in Motor Imagery Training for Brain-Computer Interfaces: A Pilot Study. http://dx.doi.org/10.1109/VS-GAMES.2015.7295789

[61] Singhal, Richa & Rana, Rakesh. (2015). Chi-square test and its application in hypothesis testing. Journal of the Practice of Cardiovascular Sciences. 1. http://dx.doi.org/10.4103/2395-5414.157577

[62] The Concise Encyclopedia of Statistics. (2008). Kolmogorov–Smirnov Test. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_214

[63] IBM. (2022). What is logistic regression? https://www.ibm.com/topics/logistic-regression

[64] Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. https://doi.org/10.1007/3-540-44673-7_12

[65] Friedman, Nir & Geiger, Dan & Goldszmidt, Moises. (1997). Bayesian Network Classifiers. Machine Learning. 29. 131-163. http://dx.doi.org/10.1023/A:1007465528199

[66] Hicks, S.A., Strümke, I., Thambawita, V.L., Hammou, M., Halvorsen, P., Riegler, M.A., & Parasa, S. (2021). On evaluation metrics for medical applications of artificial intelligence. *medRxiv*. https://doi.org/10.1101/2021.04.07.21254975

[67] Scikit-learn (2022). User guide. https://scikit-learn.org/stable/user_guide.html

[68] Sharma Sagar. (2017). Epoch vs Batch Size vs Iterations. Towards Data Science. https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9

[69] Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Asadi, H. (2019). Peering into the black box of artificial intelligence: evaluation metrics of Machine Learning methods. American Journal of Roentgenology, 212(1), 38-43. https://pubmed.ncbi.nlm.nih.gov/30332290/

[70] Scikit-learn developers. (2022). F1-score user guide. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[71] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. https://doi.org/10.48550/arXiv.1609.04836

[72] Raffaelli, B., Mecklenburg, J., Overeem, L. H., Scholler, S., Dahlem, M. A., Kurth, T., Oliveira Gonçalves, A. S., Reuter, U., & Neeb, L. (2021). Determining the Evolution of Headache Among Regular Users of a Daily Electronic Diary via a Smartphone App: Observational Study. *JMIR mHealth and uHealth*, *9*(7), e26401. https://doi.org/10.2196/26401

[73] Elise Devaux. (2020). Newsenselab able to guarantee medical data anonymity. Statice.ai. https://www.statice.ai/post/newsenselab-make-medical-data-available-for-research-while-guaranteeing-patients-anonymity

[74] Chan. A et al. (2021). The Medkit-Learn Environment. Github. https://github.com/XanderJC/medkit-learn

[75] van Breugel, B., Kyono, T., Berrevoets, J., & van der Schaar, M. (2021). DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks. Advances in Neural Information Processing Systems, 34. https://doi.org/10.48550/arXiv.2110.12884

[76] Chan, A. J., Bica, I., Huyuk, A., Jarrett, D., & van der Schaar, M. (2021). The Medkit-Learn (ing) Environment: Medical Decision Modelling through Simulation. https://doi.org/10.48550/arXiv.2106.04240

[77] Comisión Nacional de Investigación Científica y Tecnológica (CONICYT). (2020). Concurso nacional de proyectos de investigación y desarrollo en salud FONIS, Resultados 2020: https://s3.amazonaws.com/documentos.anid.cl/fonis/2020/sai20/Resolucion9414-FONISXVIII2020.pdf

[78] Universidad de Chile. (2020). Dotación a contrata. http://web.uchile.cl/transparencia/contrataene2020efg.html

[79] Agencia Nacional de Investigación y Desarrollo. (2020). Fondo Nacional de Desarrollo Científico y Tecnológico regular (FONDECYT), Resultados 2020: https://s3.amazonaws.com/documentos.anid.cl/regular/2020/fallo/Estadisticas_y_recursos_totales_asignados_por_Disciplina.pdf

[80] Universidad de Chile. (2021). INSTRUCTIVO PROCESO DE POSTULACIÓN AL CONCURSO FONIS 2021 https://www.uchile.cl/documentos/instructivo-procedimiento-interno_172548_2_3538.pdf

[81] Comisión Nacional de Investigación Científica y Tecnológica (CONICYT). (2020). Bases concurso nacional de proyectos FONDECYT regular 2020. https://www.conicyt.cl/fondecyt/files/2019/05/Bases-Concurso-FONDECYT-Regular-2020.pdf

[82] Chile Atiende. (2021). Concurso Fondef IDeA de Investigación y Desarrollo (concurso IDeA I+D) https://www.chileatiende.gob.cl/fichas/74998-concurso-fondef-i-de-a-de-investigacion-y-desarrollo-concurso-i-de-a-i-d#:~:text=El%20monto%20m%C3%A1ximo%20a%20solicitar,15%20de%20abril%20de%202021.

[83] Fawcett, Tom. (2006). Introduction to ROC analysis. Pattern Recognition Letters. 27. 861-874. http://dx.doi.org/10.1016/j.patrec.2005.10.010

[84] Al-Issa, Y., Ottom, M. A., & Tamrawi, A. (2019). eHealth Cloud Security Challenges: A Survey. Journal of healthcare engineering, 2019, 7516035. https://doi.org/10.1155/2019/7516035

[85] Wehle, Hans-Dieter. (2017). Machine Learning, Deep Learning, and AI: What's the Difference?. https://www.researchgate.net/publication/318900216_Machine_Learning_Deep_Learning_and_AI_What's_the_Difference

[86] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1-85. https://arxiv.org/abs/2103.11251

[87] Wagstaff, K. (2012). Machine learning that matters. arXiv preprint arXiv:1206.4656. https://arxiv.org/ftp/arxiv/papers/1206/1206.4656.pdf

[88] Pharma Intelligence. Oracle. (2018). Challenges And Opportunities In Clinical Data Management. https://www.oracle.com/oce/dc/assets/CONTEEDBE293BDEF418C998B39AE60FCBA35/native/oracle-clinical-data-report-1809-final-26-sept.pdf?elqTrackId=a3c3795787d24ddb905a0872489fcbd8&elqaid=75274&elqat=2

[89] Altexsoft. (2021). Functional and Nonfunctional Requirements: Specification and Types. https://www.altexsoft.com/blog/business/functional-and-non-functional-requirements-specification-and-types/

[90] McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56). https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

[91] The Open Knowledge Foundation Ltd. (2013). Messytables: all your rows are belong to us. https://messytables.readthedocs.io/en/latest/

[92] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830. https://arxiv.org/abs/1201.0490

[93] Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open-Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021

[94] Streamlit Inc. (2022). Getting Started. https://docs.streamlit.io/library/get-started

[95] Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. Linux Journal, 2014(239), 2. https://dl.acm.org/doi/fullHtml/10.5555/2600239.2600241

[96] Van Rossum, G. (2020). *The Python Library Reference, release 3.8.2*. Python Software Foundation. https://docs.python.org/3/library/pickle.html

Anexos

A. Resultados de la generación de datos

A.1. Test de navegación

Tabla A.1.1: Descripción estadística de los datos reales relacionados al test de navegación. Elaboración Propia.

	avg_pathLengthRat	avg_avgNavigVeloc	avg_plat_LatencRat	change_avgNavigVeloc	change_pathLengthRat	change_plat_LatencRat
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000
mean	0.542695	20.094931	0.221199	0.042612	0.182680	0.117671
std	0.151445	17.154715	0.105410	0.363503	0.560755	0.576653
min	0.288898	3.819944	0.068483	-0.457699	-0.709266	-0.627124
25%	0.438224	7.911884	0.135005	-0.194805	-0.170017	-0.354746
50%	0.519024	12.683025	0.214311	0.032798	-0.005991	0.064749
75%	0.627486	28.365956	0.283057	0.230486	0.349652	0.406589
max	0.827259	73.820030	0.455746	1.525110	1.831605	1.711947

Tabla A.1.2: Descripción estadística de los datos generados utilizando GAN relacionados al test de navegación. Elaboración Propia.

	avg_plat_LatencRat	avg_pathLengthRat	avg_avgNavigVeloc	change_plat_LatencRat	change_pathLengthRat	change_avgNavigVeloc
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000
mean	0.214200	0.691494	16.095686	0.127929	1.100045	0.098086
std	0.165446	0.200617	19.248001	0.678340	0.900879	0.414015
min	-0.113633	0.334218	-6.992015	-0.893047	-0.659617	-0.708649
25%	0.113890	0.570124	4.890640	-0.355823	0.557116	-0.234598
50%	0.215040	0.650967	12.664454	-0.015506	0.908935	0.163838
75%	0.344618	0.803071	21.294274	0.335471	1.852782	0.406128
max	0.562911	1.128753	85.731757	1.654925	2.762073	0.836939

Tabla A.1.3: Descripción estadística de los datos generados utilizando VAE relacionados al test de navegación. Elaboración Propia.

	avg_plat_LatencRat	avg_pathLengthRat	avg_avgNavigVeloc	change_plat_LatencRat	change_pathLengthRat	change_avgNavigVeloc
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000
mean	0.206100	0.491667	14.980546	-0.088910	0.029783	0.089060
std	0.073741	0.114953	11.362436	0.434590	0.253275	0.230784
min	0.068483	0.293455	4.419169	-0.627124	-0.457440	-0.350705
25%	0.160432	0.411876	8.962573	-0.336113	-0.120179	-0.074413
50%	0.205115	0.455743	11.426972	-0.199954	0.037446	0.112782
75%	0.244853	0.535394	14.971671	0.029339	0.104757	0.231739
max	0.421600	0.797294	58.029378	1.246416	0.750553	0.518766

Tabla A.1.4: Descripción estadística de los datos generados utilizando Gaussian Copula relacionados al test de navegación. Elaboración Propia.

	avg_plat_LatencRat	avg_pathLengthRat	avg_avgNavigVeloc	change_plat_LatencRat	change_pathLengthRat	change_avgNavigVeloc
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000
mean	0.197246	0.527830	18.570077	0.091904	0.094771	-0.028827
std	0.088126	0.139203	17.714345	0.590785	0.468254	0.281345
min	0.080005	0.288898	3.819944	-0.624869	-0.709266	-0.457699
25%	0.110862	0.425532	8.591592	-0.404214	-0.309338	-0.203953
50%	0.190643	0.524633	12.704091	0.000752	0.169396	-0.057512
75%	0.258280	0.624426	20.034332	0.330420	0.412741	0.125241
max	0.391042	0.812264	73.820030	1.644868	0.896033	0.646691

A.2. Cámara de movimiento ocular

Tabla A.2.1: Descripción estadística de los datos reales relacionados al movimiento ocular. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	cuad1	cuad2	cuad3	cuad4	dilmean_Group	dist_euc
count	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000
mean	119.293343	67.035085	0.055375	0.016096	0.226063	0.702466	-106.637982	213.785516
std	31.192844	26.715248	0.076187	0.019500	0.092608	0.100945	101.901675	55.403399
min	36.415103	17.152944	0.000543	0.000000	0.001641	0.470701	-328.628915	59.092375
25%	100.543502	55.218111	0.010780	0.003365	0.182537	0.670232	-172.359648	177.324740
50%	116.025100	63.887829	0.025359	0.011497	0.221454	0.703968	-88.319380	224.957649
75%	135.382918	78.648669	0.065242	0.021805	0.267413	0.763669	-33.423069	257.939852
max	173.014826	155.698032	0.314687	0.081978	0.450015	0.891407	88.088373	308.852970

Tabla A.2.2: Descripción estadística de los datos generados utilizando GAN relacionados al movimiento ocular. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	dilmean_Group	dist_euc	cuad1	cuad2	cuad3	cuad4
count	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000
mean	166.621993	54.969095	59.550102	187.205381	0.121203	-0.000456	0.266698	0.539752
std	40.876750	25.187569	123.321675	74.277756	0.069868	0.014211	0.109473	0.118434
min	53.998443	7.487423	-220.092691	58.377790	-0.012928	-0.022016	0.075882	0.366119
25%	139.934593	36.889707	14.490844	125.330742	0.077347	-0.010136	0.190087	0.437073
50%	176.740159	57.285234	88.408880	185.943209	0.113374	-0.003576	0.246574	0.531077
75%	201.364357	74.563477	155.831731	229.396501	0.165287	0.005497	0.342536	0.620373
max	215.530326	102.504307	236.933279	365.186087	0.286622	0.029182	0.517290	0.798055

Tabla A.2.3: Descripción estadística de los datos generados utilizando VAE relacionados al movimiento ocular. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	dilmean_Group	dist_euc	cuad1	cuad2	cuad3	cuad4
count	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000
mean	119.018667	68.495157	-124.201566	218.641345	0.033835	0.012102	0.256817	0.698567
std	24.878017	15.828372	81.390404	42.564923	0.019219	0.007622	0.070438	0.075154
min	68.631817	33.212618	-328.628915	105.195775	0.000543	0.000000	0.094031	0.512830
25%	103.919435	56.440184	-165.969661	200.794435	0.024151	0.007110	0.210145	0.662303
50%	117.092245	71.090703	-121.819724	227.009264	0.033857	0.011040	0.250453	0.716102
75%	134.943547	78.512115	-76.111732	241.949932	0.042476	0.014810	0.300652	0.745006
max	167.733412	103.044064	33.710427	295.665699	0.078219	0.034711	0.388350	0.817337

Tabla A.2.4: Descripción estadística de los datos generados utilizando Gaussian Copula relacionados al movimiento ocular. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	dilmean_Group	dist_euc	cuad1	cuad2	cuad3	cuad4
count	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000	37.000000
mean	124.144222	72.788423	-80.003461	221.933915	0.046655	0.014951	0.243724	0.705104
std	27.438216	27.519266	100.396969	52.772555	0.050424	0.019870	0.097424	0.087544
min	55.853832	27.676193	-275.580469	130.433152	0.000547	0.000063	0.001641	0.470701
25%	108.840105	60.230756	-158.206509	181.442853	0.007395	0.002151	0.180909	0.678498
50%	122.673215	67.650656	-105.719108	220.329400	0.033238	0.007689	0.222966	0.709870
75%	140.855654	80.152093	2.076281	271.754679	0.063805	0.020194	0.295050	0.759984
max	173.014826	146.048205	78.556828	308.852970	0.217171	0.081978	0.450015	0.835586

A.3. Electroencefalograma

Tabla A.3.1: Descripción estadística de los datos filtrados reales relacionados al electroencefalograma. Elaboración Propia.

	AF3_Theta	CP6_Beta	Mean_F4	P8_Gamma	PO4_Theta	T7_Beta	Variance_PO4
count	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000	3.600000e+01
mean	0.123716	0.271072	0.002012	0.219862	0.118166	0.266935	1.163774e+06
std	0.038533	0.078326	0.013337	0.067460	0.029292	0.056747	4.619934e+06
min	0.044389	0.105873	-0.009042	0.073783	0.050391	0.097152	2.689677e+01
25%	0.096826	0.222295	-0.001764	0.164973	0.102923	0.253604	4.932837e+01
50%	0.125247	0.266694	-0.000332	0.228691	0.113822	0.272495	1.396452e+02
75%	0.154271	0.326231	0.002230	0.277965	0.135646	0.293670	6.729906e+02
max	0.185029	0.396599	0.074764	0.318719	0.195673	0.364665	2.405899e+07

Tabla A.3.2: Descripción estadística de los datos filtrados generados utilizando GAN relacionados al movimiento ocular. Elaboración Propia.

	AF3_Theta	CP6_Beta	Mean_F4	P8_Gamma	PO4_Theta	T7_Beta	Variance_PO4
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	3.800000e+01
mean	0.114774	0.316956	-0.170704	0.253861	0.103969	0.240580	9.935133e+05
std	0.057372	0.117879	0.732473	0.063997	0.037299	0.049106	1.870563e+06
min	0.009256	0.112253	-4.473525	0.040423	0.046053	0.139869	-1.987565e+06
25%	0.081381	0.222008	-0.173303	0.217547	0.069620	0.209048	4.472380e+04
50%	0.106358	0.349138	-0.068019	0.250569	0.098821	0.236034	8.309293e+05
75%	0.139280	0.392354	0.053614	0.298761	0.132592	0.263879	1.602385e+06
max	0.234978	0.509866	0.238726	0.347337	0.187166	0.368099	9.852425e+06

Tabla A.3.3: Descripción estadística de los datos filtrados generados utilizando VAE relacionados al movimiento ocular. Elaboración Propia.

	AF3_Theta	CP6_Beta	Mean_F4	P8_Gamma	PO4_Theta	T7_Beta	Variance_PO4
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000
mean	0.132310	0.255355	-0.004025	0.249788	0.112862	0.266546	78047.782815
std	0.022393	0.030831	0.014405	0.044725	0.010200	0.027531	62019.481611
min	0.085712	0.176289	-0.026027	0.155047	0.096449	0.162729	26.896772
25%	0.116261	0.235376	-0.015427	0.242094	0.105735	0.255331	23794.876260
50%	0.134362	0.256050	-0.007002	0.266213	0.110865	0.268817	76009.365937
75%	0.144599	0.276514	0.006280	0.278204	0.120216	0.283537	115358.742113
max	0.183556	0.312039	0.027919	0.298008	0.140419	0.309793	271267.387457

Tabla A.3.4: Descripción estadística de los datos filtrados generados utilizando Gaussian Copula relacionados al movimiento ocular. Elaboración Propia.

	AF3_Theta	CP6_Beta	Mean_F4	P8_Gamma	PO4_Theta	T7_Beta	Variance_PO4
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000
mean	0.118384	0.276442	-0.043008	0.229465	0.110466	0.281970	1132.941634
std	0.036839	0.065459	0.272699	0.055822	0.027056	0.043939	2836.543048
min	0.044389	0.136817	-1.673891	0.135629	0.050391	0.186589	26.897373
25%	0.090396	0.238064	-0.003266	0.189072	0.092637	0.256467	32.413338
50%	0.121802	0.261889	-0.000107	0.230948	0.112001	0.281398	107.013680
75%	0.148277	0.319872	0.004724	0.268854	0.130165	0.303725	748.017068
max	0.183968	0.396599	0.074764	0.318719	0.170916	0.364665	13851.903335

A.4. Base de datos ya procesados y unificados

Tabla A.4.1: Descripción estadística de los datos reales en la base de datos ya unificada. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	dist_euc	avg_plat_LatencRat	avg_avgNavigVeloc	change_plat_LatencRat	AF3_Theta	P8_Gamma	Variance_PO4
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	3.800000e+01
mean	119.407393	67.324308	214.218696	0.221199	20.094931	0.117671	0.124389	0.219409	1.102778e+06
std	30.776463	26.412002	54.714776	0.105410	17.154715	0.576653	0.037588	0.065640	4.500983e+06
min	36.415103	17.152944	59.092375	0.068483	3.819944	-0.627124	0.044389	0.073783	2.689677e+01
25%	101.609575	55.555858	177.524697	0.135005	7.911884	-0.354746	0.098672	0.167795	5.091414e+01
50%	116.651334	63.942533	225.418437	0.214311	12.683025	0.064749	0.128883	0.216871	1.476239e+02
75%	135.248978	78.492897	257.268982	0.283057	28.365956	0.406589	0.153172	0.277400	1.108635e+03
max	173.014826	155.698032	308.852970	0.455746	73.820030	1.711947	0.185029	0.318719	2.405899e+07

Tabla A.4.2: Descripción estadística de los datos generados utilizando el algoritmo GAN en la base de datos ya unificada. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	dist_euc	avg_plat_LatencRat	avg_avgNavigVeloc	change_plat_LatencRat	AF3_Theta	P8_Gamma	Variance_PO4
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	3.800000e+01
mean	118.337418	97.734838	196.114768	0.347690	7.651366	0.902986	0.076807	0.216431	1.247375e+06
std	40.010283	28.733916	69.027320	0.161869	19.003799	0.828329	0.046229	0.075255	8.696130e+05
min	3.655912	54.380664	65.314568	0.003706	-24.071428	-0.908338	0.000017	-0.000791	-8.653917e+05
25%	90.058199	81.108905	141.327527	0.245154	-2.718045	0.244404	0.047158	0.184884	6.877951e+05
50%	124.117305	91.386211	195.751580	0.399245	4.426025	0.996542	0.062592	0.218580	1.471360e+06
75%	142.636562	103.392648	249.627913	0.461643	13.681059	1.549018	0.108053	0.264411	1.960431e+06
max	215.970104	167.861901	335.454193	0.577304	56.167830	2.634786	0.160580	0.320625	2.342139e+06

Tabla A.4.3: Descripción estadística de los datos generados utilizando el algoritmo VAE en la base de datos ya unificada. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	dist_euc	avg_plat_LatencRat	avg_avgNavigVeloc	change_plat_LatencRat	AF3_Theta	P8_Gamma	Variance_PO4
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000
mean	120.069178	70.278239	209.457920	0.211365	12.908874	0.203768	0.126610	0.234318	7003.901045
std	9.727343	7.001471	25.360484	0.031091	7.012062	0.340877	0.007828	0.037788	18411.013520
min	104.326149	54.385587	167.335630	0.157877	7.375046	-0.419633	0.106801	0.144952	26.896772
25%	111.287589	65.781735	190.005506	0.184357	10.347134	-0.200200	0.120714	0.239706	26.896772
50%	120.528240	71.922712	203.857317	0.212764	11.248290	0.350661	0.127516	0.245622	26.896772
75%	126.775367	75.300015	225.466286	0.236614	11.794539	0.466690	0.131922	0.257388	26.896772
max	144.121394	82.583471	257.162826	0.274776	36.469903	0.576793	0.142311	0.271818	87970.807065

Tabla A.4.4: Descripción estadística de los datos generados utilizando Gaussian Copula en la base de datos ya unificada. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	dist_euc	avg_plat_LatencRat	avg_avgNavigVeloc	change_plat_LatencRat	AF3_Theta	P8_Gamma	Variance_PO4
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	3.800000e+01
mean	120.564952	72.792403	222.807149	0.233066	13.259872	0.131627	0.129712	0.222971	6.530328e+04
std	26.516304	20.970727	52.068309	0.107550	9.634806	0.562864	0.031512	0.057578	2.777995e+05
min	59.933676	39.548950	120.855526	0.073868	3.819944	-0.603905	0.052884	0.110802	2.689677e+01
25%	103.773997	63.715179	197.911354	0.160375	6.197911	-0.318198	0.108209	0.186433	3.126532e+01
50%	119.323422	69.936828	220.293815	0.191860	11.530444	-0.003505	0.134072	0.223353	3.166050e+02
75%	132.879479	77.948524	264.528113	0.311587	15.868010	0.537230	0.151469	0.270428	2.245209e+03
max	173.014826	129.461037	305.797813	0.443222	45.415099	1.273258	0.185029	0.318182	1.404287e+06

A.5. 150 Datos para base ya procesada y unificada

Tabla A.5.1: Descripción estadística de los datos reales en la base de datos ya unificada. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	dist_euc	avg_plat_LatencRat	avg_avgNavigVeloc	change_plat_LatencRat	AF3_Theta	P8_Gamma	Variance_PO4
count	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	38.000000	3.800000e+01
mean	119.407393	67.324308	214.218696	0.221199	20.094931	0.117671	0.124389	0.219409	1.102778e+06
std	30.776463	26.412002	54.714776	0.105410	17.154715	0.576653	0.037588	0.065640	4.500983e+06
min	36.415103	17.152944	59.092375	0.068483	3.819944	-0.627124	0.044389	0.073783	2.689677e+01
25%	101.609575	55.555858	177.524697	0.135005	7.911884	-0.354746	0.098672	0.167795	5.091414e+01
50%	116.651334	63.942533	225.418437	0.214311	12.683025	0.064749	0.128883	0.216871	1.476239e+02
75%	135.248978	78.492897	257.268982	0.283057	28.365956	0.406589	0.153172	0.277400	1.108635e+03
max	173.014826	155.698032	308.852970	0.455746	73.820030	1.711947	0.185029	0.318719	2.405899e+07

Tabla A.5.2: Descripción estadística de los datos generados utilizando el algoritmo GAN en la base de datos ya unificada.

	Xstd_grouped	Ystd_grouped	dist_euc	avg_plat_LatencRat	avg_avgNavigVeloc	change_plat_LatencRat	AF3_Theta	P8_Gamma	Variance_PO4
count	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	1.500000e+02
mean	117.740719	65.421739	229.841451	0.212518	16.436268	-0.053527	0.117977	0.231002	1.270219e+06
std	52.669404	40.430301	90.160361	0.164221	20.618429	0.682262	0.062434	0.116299	6.748028e+06
min	11.639081	-25.879946	29.735378	-0.148522	-24.281970	-1.395841	-0.001771	-0.024363	-6.974667e+06
25%	82.446014	36.919932	157.413635	0.083803	3.561649	-0.555479	0.066228	0.146486	-1.295124e+06
50%	120.989008	65.916814	250.072897	0.218187	12.277981	-0.144339	0.112275	0.237136	-1.373432e+05
75%	152.708268	94.091393	309.886309	0.335278	22.539147	0.445223	0.168682	0.320458	1.231365e+06
max	226.357986	183.618877	391.456922	0.558659	94.123231	1.704195	0.241912	0.442711	4.294187e+07

Tabla A.5.3: Descripción estadística de los datos generados utilizando el algoritmo VAE en la base de datos ya unificada. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	dist_euc	avg_plat_LatencRat	avg_avgNavigVeloc	change_plat_LatencRat	AF3_Theta	P8_Gamma	Variance_PO4
count	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000
mean	118.680761	68.938305	205.892629	0.239347	12.511044	0.292001	0.113974	0.223913	115861.366335
std	16.077557	15.896632	44.319295	0.073053	6.244405	0.477395	0.022380	0.034182	106889.367163
min	69.941957	17.152944	105.383660	0.068483	3.819944	-0.627124	0.053033	0.138357	26.896772
25%	108.953521	58.834192	179.221787	0.191753	8.832168	-0.035079	0.098863	0.200212	18335.824489
50%	118.762097	68.734059	207.536716	0.235725	11.837303	0.363518	0.114121	0.225264	83866.941383
75%	129.256483	78.467597	236.942191	0.290370	15.075679	0.615590	0.126625	0.247809	192363.621050
max	158.672464	110.090568	307.013620	0.453283	54.114903	1.550951	0.177962	0.318719	460709.998139

Tabla A.5.4: Descripción estadística de los datos generados utilizando Gaussian Copula en la base de datos ya unificada. Elaboración Propia.

	Xstd_grouped	Ystd_grouped	dist_euc	avg_plat_LatencRat	avg_avgNavigVeloc	change_plat_LatencRat	AF3_Theta	P8_Gamma	Variance_PO4
count	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	1.500000e+02
mean	120.971778	68.548316	215.263580	0.213937	16.547240	0.236428	0.123923	0.226118	2.969693e+05
std	29.491294	25.241643	52.808084	0.102169	15.488718	0.593041	0.034284	0.064748	2.483812e+06
min	45.612054	18.457806	87.810842	0.071280	3.819944	-0.622665	0.044389	0.080538	2.689677e+01
25%	99.158121	53.126438	177.125468	0.125106	6.416531	-0.278316	0.100682	0.180466	2.689677e+01
50%	117.385877	64.291820	217.787729	0.195540	12.157392	0.200392	0.127304	0.229562	1.273536e+02
75%	138.318836	77.825359	259.845157	0.294051	19.411674	0.630678	0.148628	0.286153	5.512410e+02
max	173.014826	155.698032	306.960196	0.451550	73.820030	1.536680	0.185029	0.318641	2.405899e+07

B. Mercado de investigación relacionada al área de la salud

Tabla B.1: Proyectos relacionados al área de la salud y el manejo de datos que se adjudicaron el fondo IDeA I+D 2020. Extraído de [2].

Código	TITULO	NOTA FINAL	INSTITUCION BENEFICIARIA PRINCIPAL	OTRAS BENEFICIARIAS	MONTO FONDEF (M$)
ID20i10192	Development of therapeutic human monoclonal antibodies to treat covid-19	4,48	Universidad De Concepción		200,00
ID20i10297	Sistema tecnológico para la evaluación de riesgos ergonómicos de trastornos musculoesqueléticos de acuerdo a normativa nacional y metodologías internacionales	4,45	Universidad De Concepción		199,94
ID20i10174	Plataforma informática basada en inteligencia artificial para la caracterización e identificación del grado de adherencia al tratamiento	4,36	Universidad De Chile		199,39
ID20i10279	Desarrollo de una membrana antibacteriana con capacidad regenerativa ósea para uso potencial en el tratamiento de la periimplantitis	4,09	Universidad De La Frontera	Universidad De Antofagasta	200,00

ID20i10332	Sistema de inteligencia artificial para el apoyo en el diagnóstico y priorización de exámenes mamográficos	4,04	Universidad De Valparaíso		200,00
ID20i10106	Desarrollo de un nuevo biofármaco con actividad anti-tumoral para el tratamiento de cáncer: anticuerpo monoclonal completamente humano dirigido contra el sitio de combinación de MICA a su receptor de activación	3,94	Universidad De Chile	Pontificia Universidad Católica De Valparaíso	200,00
ID20i10252	Una inmunoterapia efectiva para el tratamiento de la enfermedad de alzheimer: ensayo pre-clínico	3,84	Universidad De Chile		199,94
ID20i10056	Desarrollo de un prototipo mejorado de la prueba inmunocromatográfica rápida (loxo-test) para la detección temprana del loxoscelismo presente en Chile, Argentina, Brasil y Perú.	3,82	Universidad De Antofagasta	Instituto De Salud Pública De Chile	199,09
ID20i10082	Desarrollo de un microorganismo probiótico con capacidad antiinflamatoria para tratar enfermedades inflamatorias de causa autoinmune e infecciosa.	3,79	Pontificia Universidad Católica De Chile		150,52
ID20i10152	Machine learning para el diagnóstico precoz y perfil multimodal de la enfermedad de alzheimer basado en mirna de exosomas circulantes.	4,55	UNIVERSIDAD ADOLFO IBÁÑEZ	UNIVERSIDAD MAYOR UNIVERSIDAD DE CHILE	200,00
ID20i10004	Desarrollo de un nuevo paradigma de implantes de pene para el tratamiento de la disfunción eréctil severa	4,39	PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE		200,00
ID20i10040	Desarrollo de filamentos de impresión 3d en base a materiales compuestos para aplicaciones biomédicas.	4,34	UNIVERSIDAD DE CONCEPCIÓN		199,33

ID20i10001	Advanced data science methods for medication error prevention	4,15	UNIVERSIDAD DE CHILE		180,92
ID20i10234	Herramienta multimodal no invasiva para el diagnóstico diferencial entre depresión y enfermedad de alzheimer inicial	4,09	UNIVERSIDAD DE CHILE		199,99
ID20i10371	Test audiológico detecta en forma precoz la presencia de deterioro cognitivo en adultos mayores	4,05	UNIVERSIDAD DE CHILE		196,65

Tabla B.2: Proyectos relacionados al área de la salud y el manejo de datos que se adjudicaron el fondo FONIS 2020. Extraído de [77].

Código	TITULO	NOTA FINAL	INSTITUCION BENEFICIARIA PRINCIPAL	OTRAS BENEFICIARIAS	MONTO FONDEF (M$)
SA20i0078	Derivación y validación de una regla de predicción clínica de infecciones urinarias adquiridas en la comunidad causadas por bacterias resistentes en pacientes adultos hospitalizados	4,34	PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE		60
SA20i0002	Efecto de terapia basada en realidad virtual para disminución de síntomas negativos en pacientes con esquizofrenia usuarios de clozapina	4,3	Hospital clínico metropolitano la florida dra. Eloisa Diaz Insunza		60
SA20i0095	Validación de un screening de memoria basado en realidad virtual para el diagnóstico de trastornos neurocognitivos en adultos mayores chilenos	4,28	Universidad de Valparaiso		57

SA20i0031	Diseño, implementación y evaluación de la eficacia de un modelo colaborativo multidimensional para mejorar la resolución de la depresión en equipos de atención primaria de la región Del Maule	4,23	Universidad de Talca		60

Tabla B.3: Cantidad de proyectos aprobados y montos totales del concurso FONDECYT por especialidad. Extraído de [79].

	PROYECTOS		RECURSOS TOTALES (miles $)	
DISCIPLINA	**Concursados**	**Aprobados**	**Solic. Total**	**Aprob. Total**
TECNOLOGIAS				
INGENIERIA 1	69	21	12.213.294	3.184.237
INGENIERIA 2	114	38	16.516.708	4.422.158
INGENIERIA 3	91	27	17.732.353	4.414.180
MEDICINA G1 - CIENCIAS BIOMEDICAS	**79**	**22**	**19.006.554**	**5.346.729**
MEDICINA G2-G3 CIENCIAS CLINICAS Y CIENCIAS DE LA SALUD PUBLICA	**84**	**20**	**19.454.207**	**4.814.687**
AGRONOMIA	87	23	18.432.818	4.889.869
SALUD Y PRODUCCION ANIMAL	53	15	10.897.181	2.927.491
SUBTOTAL TECNOLOGIAS	**577**	**166**	**114.253.115**	**29.999.351**