Desarrollo y validación de un algoritmo para realizar matching inteligente, en estudios epidemiológicos analíticos de gran escala

Autor:
Rivera Palma Alejandra Sofia

Tutor:
Cinthya Urquidi

Fecha de Defensa:
2022-04-29 00:00:00

Proyecto de grado

Desarrollo y validación de un algoritmo para realizar *matching* inteligente, en estudios epidemiológicos analíticos de gran escala

Fecha: 17.01.2022

RESUMEN

Relevancia del tema: en la era actual, la generación de datos en salud ha ido incrementando en forma exponencial en el transcurso de los últimos años, junto con la transformación digital de los registros en salud. Por consiguiente, el desarrollo de estudios epidemiológicos observacionales poblacionales basados en grandes volúmenes de datos es una realidad, y se ha posicionado como una opción ante la alternativa de realizar estudios poblacionales a un costo accesible.

Planteamiento del problema: el control de sesgos al momento de planificar y analizar grandes volúmenes de datos se ha transformado en un desafío para los investigadores. *Matching* (emparejamiento), es una alternativa que contribuye al control del sesgo de confusión, al crear grupos comparables de estudios, el cual ha sido ampliamente demostrado en diversos estudios observacionales. Por otro lado, las herramientas y métodos actuales para realizar *matching* exigen bases de datos perfectamente estructuradas y con un lenguaje de codificación unificado, escenario que no es una realidad en las bases de datos en salud. Este punto dificulta el control de los sesgos en estudios a gran escala.

Estado del arte: el presente proyecto desarrollará un algoritmo inteligente capaz de realizar *matching* en grandes volúmenes de datos parametrizados codificados y no codificados, basado en Inteligencia Artificial (IA)/*Machine Learning* (ML), cuya finalidad es contribuir al control del sesgo de confusión. Dentro de las herramientas actuales para realizar *matching*, se encuentran los softwares R y STATA®. Estos contienen diversos métodos y paquetes para realizar *matching*. Pero, estas herramientas actuales no permiten realizar emparejamiento en bases de datos parametrizadas codificadas y no codificadas, además de necesitar un número limitado de observaciones y datos.

Supuesto: un algoritmo inteligente basado en *Machine Learning* es capaz de leer datos parametrizados codificados y no codificado para realizar *matching* automático en estudios epidemiológicos de gran escala; de manera de lograr grupos de estudios comparables en las variables emparejadas. **Objetivo general:** Desarrollar un algoritmo de *matching* inteligente (**Epimatch**) por medio de *Machine Learning*, para emparejar unidades de observación en base de datos parametrizados codificados y no codificados, en estudios epidemiológicos analíticos de gran escala. **Diseño metodológico del estudio**: Inteligencia Artificial basado en Machine Learning con aprendizaje supervisado y no supervisado, utilizando métodos de similitud y semejanza. **Aplicabilidad:** en estudios observacionales epidemiológicos a gran escala, en bases de datos parametrizados codificados y no codificados. Permitirá crear grupos de estudios comparables y contribuir al control del sesgo de confusión, con el consiguiente aumento de la validez interna de los resultados.

Comentado [1]:

- **RELEVANCIA DEL TEMA Y CARACTERIZACIÓN DEL PROBLEMA**

- **RELEVANCIA DEL TEMA**

En la actualidad, se están generando estudios epidemiológicos observacionales a partir de grandes volúmenes de datos para distintas poblaciones, gracias a la disposición de diversas bases de datos en salud. De esta manera se convierten en una alternativa de menor costo frente a la opción de realizar estudios convencionales, como es el caso de los estudios primarios que requieren del reclutamiento de un elevado número de personas para levantar la información. En este contexto, FONIS incentiva a través del financiamiento el desarrollo de la investigación aplicada, con la finalidad de aumentar, tanto el conocimiento científico como la resolución de situaciones en salud que se ajusten a los objetivos estratégicos del país. Así mismo, es relevante la validez interna de estos estudios pues permiten obtener resultados importantes para la toma de decisiones en materia de salud pública.

La ejecución de estudios epidemiológicos, sobre todo aquellos que se ejecutan a nivel poblacional, requieren de un arduo trabajo al momento de plantear su estructura, por ejemplo, diseñar el estudio adecuado, seleccionar las variables, calcular una muestra representativa de la población por estudiar, definir el *outcome* de interés y el método de análisis, entre otros. A su vez, los diseños analíticos observacionales son una buena alternativa para estudiar factores de riesgo para el desarrollo de diversas enfermedades, el efecto de intervenciones en condiciones reales o cuando no es posible realizar estudios experimentales. Sin embargo, el gran desafío que deben enfrentar los investigadores es controlar las distintas fuentes de sesgo, especialmente el de confusión. Cuando no se controlan en el estudio, disminuye su validez interna, con lo cual se obtienen resultados que no son representativos de la realidad de la población.

Con el fin de controlar el sesgo de confusión, existen varias técnicas metodológicas, y una de ellas es el *matching* (emparejamiento). Es una técnica ampliamente utilizada en el ámbito de los estudios observacionales, permite controlar el sesgo de confusión y aumentar el grado de validez interna en este tipo de estudios. Además, es fácil de comprender y ejecutar cuando las variables por emparejar son mínimas, están perfectamente codificadas y el tamaño muestral es acotado. Si el emparejamiento es adecuado, el riesgo del sesgo de confusión es el mínimo. No obstante, en grandes volúmenes de datos y tamaños muestrales, o si las variables no se encuentran codificadas, el *matching* se vuelve un desafío para los investigadores (1).

De acuerdo con lo anterior, este proyecto propone desarrollar y validar una nueva herramienta tecnológica e inteligente basada en *Machine Learning (ML)*, la cual permitirá facilitar el *matching* en estudios observacionales cuando se disponen de grandes volúmenes de datos o grandes tamaños muestrales, con variables de distintas parametrizaciones. En este sentido, esta innovadora herramienta se ajusta con FONIS, ya que dará respuesta a la necesidad de controlar el sesgo de confusión en estudios de grandes volúmenes de datos, que le otorgará una mayor validez interna a sus resultados. Además, esta herramienta tendrá las características de ser sencilla, comprensible, de fácil acceso y analizar datos que no estén parametrizados y no codificados o codificados. Todos estos atributos, permitirán a los investigadores que no tienen experiencia en *Data Science*, trabajar en este tipo de bases de datos y obtener resultados confiables.

Por último, entender la importancia del control de confusores que puedan sesgar y disminuir la validez interna es altamente relevante, sobre todo cuando se analizan grandes volúmenes de datos y que, a partir de sus resultados, se toman importantes decisiones en salud pública.

- **PLANTEAMIENTO DEL PROBLEMA**

El área de la salud es un escenario rico en datos e información que se han incrementado notablemente en los últimos años. Esto se ha generado por la transformación digital, lo que ha significado una mejora sustancial en los registros clínicos, en el traspaso del sistema en papel al electrónico y al mayor almacenamiento de datos. Por ejemplo, en un día habitual de un servicio de medicina intensiva en España, se generan, en promedio, 1.400 nuevas unidades de información por cada paciente ingresado; esto se traduce en 10 millones de unidades de información al año (2). En Chile, entre los años 2017 al 2020, se han registrado 6.304.409 altas hospitalarias a nivel nacional (3). En el ámbito epidemiológico, se pueden citar las múltiples encuestas nacionales que reclutan más de 6 mil encuestados, lo que se traduce en más de 7 millones de datos por encuesta (4). A su vez, en salud pública, se requieren estudios analíticos observacionales con base poblacional que contribuyan a la resolución de problemas sanitarios. Estos pueden ser primarios, aunque muy costosos o, por el contrario, secundarios que emplean el volumen de datos ya existente. Sin embargo, los investigadores de diversas áreas se deben enfrentar al desafío de controlar el sesgo de confusión que puede influir en la validez interna de los resultados, característico de estos diseños.

La técnica *matching* es un método de control de confusores ampliamente utilizado tanto estudios epidemiológicos observacionales, como experimentales en situaciones específicas. Así mismo, esta técnica permite emparejar grupos de sujetos o unidades de estudio lo más parecido posible en cuanto a sus características, en el que un grupo estará expuesto al factor de riesgo o enfermedad y otro grupo, no. Por ejemplo, al estudiar una enfermedad y su posible factor de riesgo, los investigadores identifican un tercer factor, que tiene las características de ser una variable confusora. Ellos pueden controlar esta tercera variable y procurar que esta sea lo más similar posible entre ambos grupos; de esta manera, la validez interna del estudio no se ve comprometida (5).

De igual modo, existen múltiples estudios en salud que usan la técnica *matching*, de diferentes diseños, escalas y poblaciones. Como ejemplo se pueden citar tres estudios: en primer lugar, un estudio español poblacional, estimó la incidencia de infarto agudo al miocardio (IAM) en pacientes con diabetes mellitus tipo 2 (DM2), con el uso de datos secundarios. Se realizó *matching* en 109.759 hombres y 44.589 mujeres con y sin DM2 ≥ 40 años. Las variables usadas para el *matching* fueron edad, sexo, código IAM y año de hospitalización (6).

En segundo lugar, en el área clínica, ante la dificultad de realizar un ensayo clínico aleatorizado (ECA) para evaluar la efectividad del Programa OncoGenomic (POG), se comparó a un grupo de pacientes expuestos al programa con otro que no lo estuvo. Con este propósito, se utilizó la técnica de *genetic matching*1en un total de 203 pacientes; el *matching* se realizó en las variables sexo, área geográfica de residencia, presencia de LHA y sitio del cáncer primario. De esta manera, se evidenció que el riesgo de mortalidad disminuyó en los

pacientes sometidos a POG. Este estudio resultó ser una buena alternativa ante la ausencia de un ECA, En la toma de decisiones para el área oncológica (7).

En tercer lugar, en un tema de interés mundial, Israel realizó dos estudios importantes para evaluar la efectividad de la vacuna *ARNm BNT162b2* contra el COVID-19, desde una gran base de datos secundaria parametrizada del país. El *matching* se realizó en las características demográficas y clínicas de los sujetos. Primero evidenciaron la efectividad de la vacuna con dos dosis, con la cifra no menor de 1.163.534 vacunados y 596.618 controles. En el segundo estudio, comprobaron la efectividad de la tercera dosis de la misma vacuna para prevenir las formas graves de COVID-19, en 1.158.269 individuos vacunados con tercera dosis, y el grupo control (dos dosis) de 728.321 individuos. En cada estudio, se analizaron más de 10 millones de unidades de datos y demostraron la utilidad del *matching* en grandes volúmenes de datos. Estos resultados fueron valiosos para la toma de decisiones en salud pública (8,9).

En conclusión, el *matching* es una técnica versátil y útil al momento de ser aplicada en distintos diseños epidemiológicos. No obstante, realizar *matching* en estudios primarios que requieren de grandes tamaños de muestras o cuando los datos se encuentran parametrizados codificados o no codificados, puede ser muy complejo y afectar la validez interna del estudio. También, puede requerir personal altamente capacitado en *Data Science* para un correcto *matching*. A raíz de este problema, nace la necesidad de crear una herramienta de *matching* automatizada por medio de ML, que permita realizar emparejamiento de datos de forma rápida, sencilla y confiable.

- **ANÁLISIS DEL ESTADO DEL ARTE**

En la actualidad, existen pocas herramientas disponibles y de libre acceso para realizar *matching* en estudios epidemiológicos. En la búsqueda, se encontraron dos paquetes en *softwares* estadísticos: R y STATA®.

En R, el paquete “MATCHIT” (10) contiene variados métodos de *matching* y los más importantes se describen a continuación:

Nearest Neighbor Matching (method_nearest): este tipo de *matching* es ampliamente utilizado, ya que calcula la distancia entre cada unidad tratada o caso y cada unidad control, una por una. A cada unidad tratada, se le asigna una unidad control semejante.

Optimal Pair Matching (method_optimal): esta técnica realiza un *matching* óptimo, por medio de la suma de las distancias absolutas de los pares de la muestra emparejada, y esta distancia es la más reducida entre los casos y controles. Con el fin de lograr este tipo de emparejamiento, se requiere de las especificaciones de la medida de la distancia entre las unidades de caso y control; con esto se logra formar muestras emparejadas muy similares.

Exact Matching (method_exact): esta técnica realiza un emparejamiento exacto, que utiliza el cruce completo de las covariables, tanto de las unidades tratadas o expuestas como de las unidades control, con lo que se forman subclases para cada combinación de las covariables. En este sentido, para cada subclase que no contenga el mismo número de unidades tratadas y de control, se descarta y solo se dejan subclases que contengan unidades de tratamiento y controles semejantes.

Al eliminar aquellas subclases que no son coincidentes, se reduce drásticamente la precisión del estudio, lo que repercute en la posibilidad de extrapolar estos resultados a la población objetivo.

Optimal Full Matching (method_full): esta técnica realiza *matching* que forma subclasificaciones de todas las unidades, tratadas y controles, es decir, a la muestra completa. Además, asigna una subclase y cada una debe recibir, por lo menos, una coincidencia de emparejados. Cuando realiza *matching* entre las unidades, la distancia absoluta entre ellas es lo más pequeña posible.

En STATA®, los comandos (.dos) para realizar *matching*, se describen a continuación:

Psmatch2: es un comando ampliamente utilizado en aquellos estudios en los cuales no es posible realizar un ECA. Su forma de operar es por medio de la coincidencia completa de *mahalanobis*2y *propensity score*3, para ajustar las diferencias observables antes del tratamiento, entre un grupo de tratados y otro no tratado (11).

Nnmatch: este comando implementa estimaciones de *noarest neighbor*. De esta manera, estima el efecto promedio del tratamiento para los tratados y controles o ambos a la vez, además de sus errores estándar (12).

Teffects: es el comando más completo de STATA® y es utilizado para estimar el efecto de tratamientos en estudios observacionales. Considera los siguientes promedios: del resultado potencial (POM), de los efectos del tratamiento y de los efectos del tratamiento en los tratados. Así mismo, proporciona estimadores de ajuste de regresión, ponderación de probabilidad inversa y *matching*, métodos robustos que combinan el ajuste de regresión y la ponderación de probabilidad inversa (13).

Las alternativas expuestas anteriormente, si bien son aptas para realizar *matching* en grandes volúmenes de datos, necesariamente sus variables deben estar parametrizadas y correctamente codificadas, para que el conjunto de datos sea susceptible de ser emparejado, lo cual no es una realidad en las bases de datos en salud. A su vez, ambos paquetes requieren de conocimiento para su uso y el manejo de un nivel intermedio o avanzado en *software* estadísticos. Adicionalmente, STATA® requiere de una licencia. En el caso de R, no tiene algoritmos unificados, ya que cada uno de ellos se almacena en un paquete distinto, y estos deben cambiar en función de cada necesidad, lo que significa un mayor tiempo de trabajo.

Comentado [2]:

Comentado [3]:

Considerando los factores antes mencionados, el presente proyecto desarrollará un algoritmo inteligente capaz de realizar *matching* en grandes volúmenes de datos parametrizados codificados y no codificados, el cual se basará en Inteligencia Artificial (IA)/*Machine Learning* (ML). Aunado a esto, el análisis en datos por medio de ML permite no hacer inferencias sobre la distribución de los datos, en este caso datos no gaussianos. En este sentido, el ML evitará modelar mal la distribución subyacente de los datos. Así mismo, la no codificación de los datos hace referencia cuando estos no se encuentran estructurados o no siguen un mismo lenguaje de unificación como lo es el texto libre, escenario que es muy común en las bases de datos de salud.

El ML consiste en el aprendizaje de una máquina (computador) que puede aprender de forma automática e iterativa, por medio de datos, para realizar tareas específicas como lo es el *matching*, utilizando conocimientos y técnicas estadísticas, para la consiguiente creación de diversos algoritmos, capaces de identificar patrones, tomar decisiones de forma automática y procesar grandes volúmenes de datos con la mínima intervención humana posible.

Una de las características importantes del ML, que le confiere atributos únicos, es su aspecto iterativo. La iteración de los modelos o algoritmos en los nuevos datos permite que los mismos puedan aprender cálculos previos para la toma de decisiones y producir resultados confiables y repetibles.

En el ML existen diferentes métodos para el entrenamiento de algoritmos por desarrollar y su descripción es necesaria para entender el desarrollo del algoritmo por crear en el presente proyecto; estos métodos se describen a continuación (14):

1. Aprendizaje supervisado: permite desarrollar algoritmos utilizando datos etiquetados esto quiere decir que, desde el procesamiento de los datos, se conocen ya los resultados deseados.

Este tipo de aprendizaje es utilizado en métodos de clasificación, regresión, predicción y aumento de gradiente, para la predicción de eventos futuros probables, desde el procesamiento de datos históricos.

2. Aprendizaje no supervisado: sin duda es el método utilizado en ML, ya que permite procesar principalmente datos que no se encuentran etiquetados, como aquellos que sí lo están. Esta característica le confiere versatilidad en su uso. Al trabajar con datos que no se encuentran etiquetados, el modelo no entrega una respuesta esperada al sistema, por lo que debe explorar los datos para hallar los resultados.

Este método explora las características de los datos con el fin de detectar su estructura interna. Una vez que se identifica esta estructura, se pueden establecer los atributos similares que existen entre los datos.

3. Aprendizaje semisupervisado: utiliza técnicas de aprendizaje supervisado, pero emplea datos etiquetados y otros que no lo están para el entrenamiento de algoritmos. Es una alternativa cuando existe un alto costo en el procesamiento de datos etiquetados con aprendizaje supervisado.

Finalmente, utilizar las actuales tecnologías que puedan facilitar el análisis de datos a gran escala con fines epidemiológicos, independiente de si los datos se encuentran parametrizados codificados y no codificados, permitirá el desarrollo de estudios observacionales poblacionales a gran escala. Así mismo, el desarrollo de esta nueva herramienta inteligente capaz de realizar *matching* facilitará la obtención de resultados con una mayor validez interna y extrapolable a la población de estudio.

Bibliografía

- Pallás JA, Villa JJ. Estudios de casos y controles. En: Métodos de investigación clínica y epidemiológica. La Villa y Corte de Madrid, España: Elsevier España; 2019. p. 81–2.
- Núñez Reiz A, Armengol de la Hoz MA, Sánchez García M. Big Data Analysis and Machine Learning in intensive care units. Med Intensiva (Engl Ed). 2019;43(7):416–26.
- MINSAL. Estadísticas de egresos hospitalarios a nivel país, según diagnóstico principal de hospitalización, sexo, grupo etario y previsión. Por año y nacionalidad [Internet]. https://informesdeis.minsal.cl/. 2021 [citado el 4 de noviembre de 2021].
- MINSAL. ENCUESTA NACIONAL DE SALUD 2016-2017 Primeros resultados. 2017.
- Celentano DD, Szklo M. Estudios observacionales. En: Gordis Epidemiología. La Villa y Corte de Madrid, España: Elsevier España; 2019. p. 167–8.
- Lopez-de-Andres A, Jimenez-Garcia R, Hernández-Barrera V, de Miguel-Yanes JM, Albaladejo-Vicente R, Villanueva-Orbaiz R, ¿et al. Are there sex differences in the effect of type 2 diabetes in the incidence and outcomes of myocardial infarction? A matched-pair analysis using hospital discharge data. Cardiovasc Diabetol [Internet]. 2021;20(1). Disponible en: http://dx.doi.org/10.1186/s12933-021-01273-y.
- Weymann D, Laskin J, Jones SJM, Lim H, Renouf DJ, Roscoe R, et al. Matching methods in precision oncology: An introduction and illustrative example. Mol Genet Genomic Med. 2021;9(1): e1554.

- Dagan N, Barda N, Kepten E, Miron O, Perchik S, Katz MA, et al. BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting. N Engl J Med. 2021;384(15):1412–23.
- Barda N, Dagan N, Cohen C, Hernán MA, Lipsitch M, Kohane IS. Effectiveness of a third dose of the BNT162b2 mRNA COVID-19 vaccine for preventing severe outcomes in Israel: an observational study. The Lancet Journal. el 29 de octubre de 2021.
- Ho D, Imai K, King G, Stuart E, King G, Whitworth A, et al. Nonparametric Preprocessing for Parametric Causal Inference. CRAN; 2021.
- Help for psmatch2. STATA®.
- Abadie A, Drukker D, Herr JL, Imbens GW. Implementing matching estimators for average treatment effects in Stata. The Stata Journal. 2004;290–311.
- STATA. teffects — Treatment-effects estimation for observational data [Internet]. Disponible en: https://www.stata.com/manuals/teteffects.pdf.
- Deo RC. Machine learning in medicine. Circulation 2015; 132:1920–30. https://doi.org/10.1161/circulationaha.115.001593.

- **<u>SOLUCIÓN E INVESTIGACIÓN</u>**
- **SOLUCIÓN PROPUESTA Y ESCENARIOS DE APLICABILIDAD**

En la actualidad, el uso de la IA/ML se ha expandido de tal manera que son variadas las áreas que se han beneficiado de esta revolucionaria tecnología y una de ellas es la epidemiología. La IA permite el análisis de grandes volúmenes de datos, por medio de algoritmos inteligentes de predicción o discriminación en estudios analíticos poblacionales, usualmente observacionales. Así mismo, es importante tener en consideración el manejo de aquellos factores que puedan inducir confusión y posterior sesgo en los resultados, lo que disminuye la validez interna del estudio. En este sentido, la técnica *matching* es útil para el manejo de estos confusores, tema ya mencionado anteriormente.

En relación con lo anterior, existen pocas alternativas tradicionales para realizar *matching*, que debe ser aplicada en bases de datos correctamente codificadas y parametrizadas. En este

sentido, no es la realidad del área de la salud, ya que las bases de datos existentes no cuentan con un lenguaje estructurado ni unificado. Por ejemplo, si se tiene el diagnóstico de hipertensión, este se puede digitar como HTA, HAS, HAM, Hta, hipert, hipertensión etc. en un registro clínico. En este escenario, se debe limpiar, filtrar y codificar la base, para luego realizar el análisis; en consecuencia, el tiempo invertido es mayor, más si se trata de un estudio a gran escala. Adicionalmente, es habitual que los investigadores realicen *matching* de forma manual en este tipo de casos.

El presente proyecto propone desarrollar una herramienta tecnológica basada en ML, algoritmo denominado **Epimatch**. Esta herramienta responde a la necesidad de realizar *matching* en múltiples variables parametrizadas codificadas y no codificadas, en grandes volúmenes de datos o grandes tamaños muestrales, con el fin de controlar el sesgo de confusión en estudios epidemiológicos de gran escala. **Epimatch**, pretende ser un algoritmo sencillo, completo y entendible para el usuario, a un costo accesible.

Comentado [4]:

Los mayores beneficiarios de este nuevo algoritmo serán aquellos investigadores *seniors*, *juniors* y en formación, además de la comunidad científica y académica, que no necesariamente tiene un manejo avanzado en *Data Science*, estadística y *software* estadístico, que realicen estudios observacionales a nivel poblacional con grandes volúmenes de datos.

Así mismo, este proyecto se ajusta con la nueva Política Nacional en Inteligencia Artificial, cuyo propósito es empoderar a las personas en el uso y desarrollo de herramientas de IA, posicionarse por sobre el promedio OCDE y ser el país más avanzado en América Latina y el caribe para el año 2031 en esta área.

Finalmente, este proyecto se ajusta a los lineamientos de FONIS, ya que busca incentivar y desarrollar las capacidades investigativas de las personas en salud. Con esta innovadora herramienta se podrán desarrollar estudios epidemiológicos con grandes volúmenes de datos, fundamentados en bases de datos parametrizadas codificadas y no codificadas. La obtención de estos resultados tendrá una mayor validez interna, lo que permitirá tomar decisiones en salud con resultados confiables y extrapolables a la realidad de la población, como también orientar futuras políticas públicas.

- **PREGUNTA DE INVESTIGACIÓN**

¿Un algoritmo inteligente basado en *Machine Learning* es capaz de realizar *matching* en datos parametrizados codificados o no codificados en estudios epidemiológicos analíticos de gran escala?

- **HIPÓTESIS O SUPUESTOS DE INVESTIGACIÓN**

Supuesto:

Un algoritmo inteligente basado en *Machine Learning* es capaz de leer datos parametrizados codificados y no codificado para realizar *matching* (emparejamiento) automático en estudios epidemiológicos de gran escala; de manera de lograr grupos de estudios comparables en las variables emparejadas.

- **OBJETIVOS**

- **OBJETIVO GENERAL**

Desarrollar un algoritmo de *matching* inteligente (**Epimatch**) por medio de *Machine Learning*, para emparejar unidades de observación en base de datos parametrizados codificados y no codificados, en estudios epidemiológicos analíticos de gran escala.

- **OBJETIVOS ESPECÍFICOS**

1. Entrenar un algoritmo de *matching* (**Epimatch**) en un conjunto de datos de salud parametrizados codificados y no codificados, a través de métodos de similitud y semejanza en *Machine Learning*.
2. Validar el algoritmo de *matching* (**Epimatch**) desarrollado en un estudio epidemiológico analítico, basado en datos clínicos de una institución de salud.
3. Desarrollar una interfaz que sea capaz de reproducir el algoritmo **Epimatch**.

- **METODOLOGÍA, ÉTICA Y PLANIFICACIÓN**

- **METODOLOGÍA Y PROCEDIMIENTOS**

Considerando los factores antes mencionados, el presente proyecto desarrollará un algoritmo inteligente capaz de realizar *matching* en grandes volúmenes de datos parametrizados codificados o no codificados, usando la Inteligencia Artificial (IA) específicamente *Machine Learning* (ML).

Se decidió por el ML ya que consiste en el aprendizaje de una máquina (computador) que puede aprender de forma automática e iterativa, por medio de datos, para realizar tareas específicas como lo es el *matching*, utilizando conocimientos y técnicas estadísticas, para la consiguiente creación de diversos algoritmos, capaces de identificar patrones, tomar decisiones de forma automática y procesar grandes volúmenes de datos con la mínima intervención humana posible.

Para el entrenamiento de **Epimatch** se utilizarán dos tipos de ML, el aprendizaje supervisado y no supervisado, técnicas descritas anteriormente. De esta forma se podrá probar y desarrollar diferentes algoritmos para un mismo fin, y seleccionar el mejor método que pueda realizar *matching* en grandes volúmenes de datos, especialmente en aquellos datos parametrizados codificados y no codificados.

Dentro de lo anterior, cabe destacar que aquellos datos definidos como codificados son aquellos que se encuentran estandarizados dentro de una base de datos, como por ejemplo la variable sexo se puede definir con el número 1 para hombres y el número 0 para mujeres. De lo contrario, aquellos datos que no se encuentran codificados puedes ser digitados y registrados en texto libre, como por ejemplo mujer, femenino, femenina, hembra, hombre, masculino, macho, etc. Además, la parametrización hace referencia a la celda en la cual se encuentra contenido el dato de la observación, a diferencia de las imágenes que se consideran como no parametrizados, ya que no se encuentran contenidos en un espacio en específico de la base de datos.

Procedimientos para el desarrollo del algoritmo Epimatch

1. Bases de datos

Con el propósito de entrenar el algoritmo, se utilizarán bases de datos libres proporcionadas por la página *kaggle.com.* Esta plataforma es gratuita y pone a disposición diversas bases de datos, una de ellas en el área de la salud, para la solución de problemas en análisis predictivo, *Data Science* y ML. Además, es ampliamente conocida entre científicos de datos e ingenieros, ya que *Kaggle* imparte concursos para resolver desafíos en el análisis de datos.

De las bases de datos disponibles, los criterios de inclusión para la selección de la base de datos son los siguientes:

1. Bases de datos de origen en salud, epidemiológico o clínico.
2. Contener datos codificados y no codificados.
3. Con registro de datos, para el algoritmo basado en aprendizaje supervisado, mayor o igual a 10.000 observaciones.

Los criterios de exclusión para la selección de la base de datos en salud son los siguientes:

1. En formato de imagen.
2. Con pérdidas de datos >30%
3. No anonimizadas.

2. Tamaño muestral para el entrenamiento

Este es un punto relevante, ya que se debe obtener una base con grandes volúmenes de datos, con el objetivo de entrenar y validar el algoritmo. En este sentido, se debe seleccionar una base de datos con un número mayor a 10.000 observaciones (personas o sujetos), con un máximo de 10 variables a estudiar (columnas), obteniendo en total un número mayor a 100.000 datos.

Este criterio se establece para lograr una buena precisión del algoritmo a desarrollar.

3. Métodos para el desarrollo del algoritmo

Se utilizarán tres diferentes técnicas para la creación del algoritmo. Aquel que tenga mejor rendimiento será seleccionado para la posterior validación.

Comentado [5]:

1. *K-Nearest Neighbor (KNN):* definido como método de aprendizaje supervisado del ML. Su forma de operar es sencilla, ya que busca observaciones más cercanas al resultado que se intenta predecir, y clasifica los puntos de interés en función de los datos que lo rodean. Con este objetivo, almacena e identifica las distancias entre todos los datos, utilizando la función de distancia, a un *K de* distancia. Sus resultados serán los siguientes: la etiqueta más frecuente y los valores media de la *K neighbor* más cercanos4.

2. *K-Means Clustering (KMC)*: definido como técnica de aprendizaje no supervisado del tipo iterativo, que dividirá las observaciones (n) en *clúster* (k), en que cada observación se encuentra ubicada según la media más cercana al *clúster*. Principalmente, agrega una colección de puntos a los datos en función de sus similitudes. Esta técnica es frecuentemente usada en Big Data5

Comentado [6]:

Comentado [7]:

3. Método *A priori*: este modelo se comenzó a utilizar ampliamente en la minería de datos, específicamente en bases de datos de transacciones. Con esta técnica se extraerá conjuntos de elementos frecuentes, por medio de un método iterativo llamado búsqueda de capa por capa, en que se usan *k* conjuntos de elementos para explorar6.

4. Desarrollo del algoritmo *matching*

En el desarrollo del algoritmo se utilizará la multiplataforma de código abierto *Python*®, que es útil y versátil para el desarrollo de algoritmos y procesamiento de datos. Siguiendo las siguientes etapas secuencialmente:

1. Recolección de datos

Como se mencionó anteriormente, se utilizará una base de datos libre de la página web *Kaggle.com*, la cual debe ser seleccionada en función de los criterios de inclusión y exclusión establecidos.

2. Preparación de los datos

Una vez seleccionada la base de datos, se procederá a su limpieza. En este proceso, se utilizarán dos librerías de *Python® NumPy y Pandas*. El primero, se especializa en el cálculo numérico y análisis de grandes volúmenes de datos a través de la formación de *arrays,* que es una estructura de datos de un mismo tipo organizada en forma de tabla o cuadrícula de distintas dimensiones. El segundo, se especializa en el análisis y manejo de estructura de datos, basado en los *arrays* formadas por la librería *NumPay.* De esta manera, aquellos datos que sean reconocidos como texto libre serán reconocidos como numéricos para su posterior análisis.

Los datos seleccionados para la construcción del algoritmo de *matching* se dividen de la siguiente forma:

1. Un 80% de los datos se utilizarán para entrenar el algoritmo.

2. Un 10% se emplearán para validar el algoritmo en la base de datos de entrenamiento.
3. El 10% restante se utilizará para testear o someter a prueba el algoritmo; de esta forma, se podrá comprobar que el algoritmo funcione correctamente.

Por último, una vez obtenida la base de datos limpia, se debe plantear la pregunta de investigación a trabajar. Este paso es importante, ya que se van a definir las variables con las cuales se realizará el análisis y posterior *matching*. Además, se deben definir las características de los grupos a comparar, identificando los casos o sujetos que presentan el evento de interés y, controles o sujetos que no presentan el evento de interés pero que se asemejan a los casos en sus características.

4. Medidas indicadoras de éxito

En el entrenamiento del algoritmo *KNN y KMC*, se utilizará la librería *Scikit-learn*, que es gratuita y se encuentra disponible en *Python*®. Esta librería puede ser utilizada en métodos de aprendizaje supervisado y no supervisado, y es aplicada en técnicas de clasificación, regresión, *clustering* y reducción de dimensionalidad.

K-Nearest Neighbor (KNN)

1. Primero se debe determinar el valor de *k*, número de vecinos, en el entrenamiento de la base de datos.
2. Calcular la distancia de los datos del set de entrenamiento, en ambos grupos (casos y controles), usando la distancia euclidiana.
3. Ordenar la distancia entre los datos del más pequeño al más grande según el número de *k*.
4. Seleccionar las *k* muestras conocidas más cercanas.
5. Se espera un *accuracy* o precisión del set de entrenamiento >90%, y en el set de prueba >85%.

Comentado [8]:

k-means clustering (KMC):

1. Primero se debe dividir el conjunto de datos en k-números de grupos. Este proceso debe ser itinerante, hasta que no se encuentren mejores grupos o centros.
2. Posteriormente se debe calcular la distancia al cuadrado entre, los puntos de datos y de todos los centroides formados.
3. Asignar a cada punto de datos el grupo más cercano.

4. Calcular el promedio de los centroides o grupos de todos los puntos de datos que pertenecen a cada grupo.

5. Se espera un *accuracy* o precisión del set de entrenamiento >90%, y en el set de prueba >85%.

Método *Apriori*:

Se utilizará la librería *Apyori* para el desarrollo del algoritmo. Esta se encuentra disponible y gratuito en *Python*®.

1. Primero se debe calcular el valor de soporte, esto quiere decir el número de ocurrencias para para cada variable del *dataset.*

2. Posteriormente se debe definir el umbral de apoyo. Este paso permite filtrar aquellos elementos que no son frecuentes, y su valor se define según el *dataset*, pero puede ser un 60%, 70%, 80%, etc.

3. Una vez establecido el umbral de apoyo, se deben extraer todos los subconjuntos que tengan un valor de soporte superior al umbral.

4. Por último, se deben seleccionar todos los subconjuntos con un valor de confianza superior al umbral mínimo, y se deben ordenar de forma descendente.

5. Se espera un *accuracy* o precisión del set de entrenamiento >90%, y en el set de prueba >85%.

6. Validación del algoritmo

Una vez que se desarrolle el algoritmo, este será validado en una base de datos clínica de una institución de salud. Así mismo, se elaborará un protocolo de investigación para cumplir con este punto, y será presentado al CEC del establecimiento para su aprobación.

La base de datos a solicitar debe cumplir las mismas características que fueron establecidas en los criterios de inclusión y exclusión. En tal sentido, se respetarán los cimientos utilizados para el entrenamiento del algoritmo, y de esta manera poder reproducir el algoritmo **Epimatch** en otras bases de salud.

La validación se basará en la capacidad del algoritmo de formar grupos comparables para dar respuesta a la pregunta de investigación que se pueda plantear. En sentido, el algoritmo debe ser capaz de agrupar casos y controles lo más similar posible, esto quiere decir que sus distancias deben ser lo más cercana entre cada una de ellas. Además, se medirá la rapidez del algoritmo en formar estos grupos, el cual se medirá en segundos.

7. Desarrollo de la interfaz

El algoritmo **Epimatch** será reproducido por medio de una interfaz, la cual será desarrollada por una empresa que preste servicios en informática y su principal característica debe ser accesible, fácil de usar y comprensible para el usuario. Por lo tanto, se va a privilegiar una plataforma sencilla y amigable.

Los pasos para el desarrollo de la interfaz son las siguientes:

1. Seleccionar una empresa que preste servicios en informática especializada en la creación de interfaces.
2. Crear un diseño o maqueta basado en los objetivos del algoritmo.
3. Especificar los iconos e imágenes a utilizar para los mensajes de la interfaz.
4. Minimizar los pasos para reproducir el algoritmo, esto le otorgará la característica de ser sencillo.
5. Diseñar un estilo atractivo para la interfaz.
6. Una vez creada la interfaz, será probada y testeada en distintas bases de datos de salud.

La validación de la interfaz será por los mismos potenciales usuarios como académicos, investigadores, estadísticos, entre otros. Los usuarios tendrán la oportunidad de utilizar la interfaz y ellos determinarán si es capaz de realizar *matching* en las bases de datos que estimen pertinente para sus estudios epidemiológicos.

Por último, la difusión del algoritmo e interfaz será por medio de cursos y seminarios que se impartirán en formato online y presencial por parte de la Universidad de los Andes.

En conclusión, el algoritmo **Epimatch** se desarrollará basado en ML y probado en lenguaje supervisado y no supervisado, utilizando tres técnicas distintas de similitud de datos. El algoritmo que presente el mejor rendimiento para realizar *matching* será seleccionado con el fin de validarlo en la base de datos que será solicitada a una institución de salud. Finalmente, una vez validado el algoritmo, expertos en el área desarrollarán una interfaz con el propósito de utilizarla en las distintas áreas académicas que lo requieran.

- **ANÁLISIS DE LAS IMPLICANCIAS ÉTICAS**
  - **ANÁLISIS DE RIESGO-BENEFICIO**

El presente proyecto no representa un daño directo ni indirecto a la salud de las personas, ya que se trabajará con datos en salud. No obstante, el hecho de trabajar solo con datos no evita que haya un daño, pues las personas pueden sentirse vulneradas al saber que su información privada se pueda utilizar con fines investigativos. Por este motivo, se trabajará con datos anonimizados, lo que resguardará la información sensible.

En cuanto al beneficio, este es mayor al riesgo, ya que se diseñará una herramienta inteligente basada en ML para realizar *matching* en grandes volúmenes de datos, lo que brindará la posibilidad de ejecutar estudios en salud a gran escala, cuyos resultados tendrán el menor sesgo posible y una mayor validez interna. Además, a partir de estos resultados se tomarán decisiones en salud pública con un impacto positivo en la población objetivo.

- **RESGUARDO DE LA CONFIDENCIALIDAD**

Con respecto a este punto, la Inteligencia Artificial ha sido tema de interés y ha motivado la creación de políticas públicas en relación con el uso de grandes volúmenes de datos sensibles.

Inicialmente, en el presente proyecto, se utilizarán bases de datos libres, obtenidas de la página *Kaggle.com*, las cuales se encontrarán anonimizadas, por lo tanto, no se trabajará con datos que identifiquen a las personas.

En una segunda instancia, se empleará una base de datos clínica de salud para la validación del algoritmo. Además, se elaborará un protocolo que será presentado al Comité de Ética Científico (CEC) de la institución de salud. Una vez que el protocolo sea aprobado, se solicitará la base de datos, pero la información sensible será anonimizada, como por ejemplo Rut, nombre, apellidos, teléfono, etc.

- **CONSENTIMIENTO/ASENTIMIENTO INFORMADO**

El presente proyecto no requerirá de consentimiento informado. Sin embargo, para solicitar la base de datos clínica con el cual se validará el algoritmo, se realizará un protocolo del proyecto para ser presentado al CEC de la institución de salud. Asimismo, una vez obtenida la aprobación del CEC, se solicitará la dispensa de consentimiento informado. El investigador principal del proyecto será la persona responsable por velar, en todo momento, por la anonimización de los datos en conjunto con el equipo del proyecto.

- **AUTORIZACIONES INSTITUCIONALES REQUERIDAS**

Se solicitará autorización al CEC de la institución de salud para la obtención de la base de datos clínica. Además, se debe contar con la aprobación de la dirección médica del lugar con el propósito de iniciar la recolección de datos.

- **PLAN DE TRABAJO**

Objetivo Especifico	*Actividad*	AÑO 1												AÑO 2											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1. Entrenar un algoritmo de *matching* (**Epimatch**) en un conjunto de datos de salud parametrizados codificados y no codificados, a través de métodos de similitud y semejanza en *Machine Learning*.	1.Explorar métodos de similitud y semejanza basados en ML.																								
	2. Obtener una base de datos en salud libre, desde la página web kaggle.com																								
	3.Limpiar la base de datos seleccionada.																								
	4. Desarrollar el algoritmo de *matching*, basado en métodos del ML con lenguaje supervisado y no supervisado, utilizando técnicas en similitud y semejanza.																								
2. Validar el algoritmo de *matching* (**Epimatch**) desarrollado en un estudio	1. Crear un protocolo de proyecto en salud, para solicitar una base de datos clínica, en una institución en salud.																								

epidemiológico analítico, basado en datos clínicos de una institución de salud.	2. Presentar el protocolo al Comité de Ética Científico de la institución de salud.																								
	3.Obtener la autorización del CEC de la institución para obtener la base de datos																								
	4. Solicitar la base de datos al área de registros clínicos previa aprobación del CEC.																								
	5. Validar el algoritmo de *matching* en la base de datos clínica.																								
3. Desarrollar una interfaz que sea capaz de reproducir el algoritmo **Epimatch**.	1. Explorar softwares adecuados en el diseño de interfaces, para reproducir el algoritmo **Epimatch** en grandes volúmenes de datos.																								
	2. Crear una interfaz para reproducir el algoritmo																								

	Epimatch en función de los objetivos de este.																							
	3. Testear la interfaz en usuarios potenciales del algoritmo: universidades, departamentos de investigación, departamentos de análisis de datos, estadísticos, etc.																							
	4. Crear un manual de usuario de **Epimatch**																							
	5. Difundir el algoritmo en seminarios presenciales y online.																							
	6. Publicación científica del algoritmo **Epimatch**																							

1. RESULTADOS, IMPLEMENTACIÓN Y DIFUSIÓN

1. IMPLEMENTACIÓN DEL(LOS) PRODUCTO(S) ESPERADO(S)

1. RESULTADOS Y/O PRODUCTOS ESPERADOS

Nombre del resultado/producto	Ingresar una breve descripción del resultado/producto
Algoritmo Epimatch	Algoritmo inteligente de ML, capaz de realizar *matching* en grandes volúmenes de datos en salud, en bases parametrizados codificadas y no codificadas, para eliminar el sesgo de selección y el consiguiente sesgo de confusión. Epimatch será inscrita como marca registrada.
Interfaz Epimatch	Interfaz capaz de reproducir el algoritmo de forma sencilla y fácil.
Manual de uso del algoritmo Epimatch	Se creará un manual de usuario que estará disponible de forma gratuita en el portal de la Universidad de los Andes. Este manual contendrá los pasos tanto para su descarga como de su uso.
Presentación en congresos de informática “JIS Go Live”	Presentar el algoritmo en congresos del área informática en salud, con la intención de masificar su conocimiento y uso.
Presentación del algoritmo en la comunidad científica	Se presentará el algoritmo en universidades y centros científicos para su difusión.
Publicación científica	Se publicará el algoritmo en revistas científicas nacionales e internacionales.

2. IMPLEMENTACIÓN DE EL(LOS) PRODUCTO(S) ESPERADO(S)

Los principales beneficiaros del algoritmo **Epimatch** serán investigadores senior, junior y en formación como también académicos que realizan estudios epidemiológicos a grandes escalas, fundamentados en bases de grandes volúmenes de datos parametrizadas codificadas y no codificadas. Se espera que el algoritmo sea utilizado principalmente en universidades, departamentos de investigación y de manejo de datos, entre otras áreas especializadas en esta área.

En relación con lo anterior, la primera estrategia de implementación es crear alianzas con áreas académicas dedicadas a la investigación epidemiológica como también departamentos dedicados al análisis de datos y estadística, de esta manera se busca compartir y difundir las propiedades y beneficios del uso de **Epimatch**.

La segunda estrategia de implementación es realizar seminarios y charlas para difundir el algoritmo **Epimatch**. En estas instancias se expondrán los detalles del algoritmo y aplicabilidad en los estudios epidemiológicos a gran escala. Además, se presentará el manual de usuario de **Epimatch**, el cual se encontrará en formato online disponible en la página web institucional de la Universidad de los Andes.

Por último, se espera que junto con las actividades de difusión el algoritmo sea utilizado en diversos estudios epidemiológicos poblacionales a gran escala basado en datos parametrizados codificados y no codificados. También se espera que sea incorporado en los estudios como parte de su metodología para realizar *matching* como estrategia para contribuir a la disminución del sesgo de confusión.

2. ACTIVIDADES DE DIFUSIÓN

Comunidad científica

Se difundirá el algoritmo **Epimatch** en la Sociedad Chile de Epidemiología (SOCHEPI), Ministerio de Salud (MINSAL), Sociedad científicas, universidades y centros de investigación en salud y epidemiológicos del país.

Por lo tanto, para dar cumplimiento a este punto se solicitarán reuniones a directivos y personas pertinentes a cada área expuesta, con el fin de difundir y exponer las funciones y utilidades de **Epimatch** al emparejar unidades de observaciones en grandes volúmenes de datos parametrizados codificados y no codificados.

Congresos

Epimatch será presentado en congresos de informática en salud, *Data Science* como también aquellos dedicados al área de la epidemiología, sobre todo aquellos que tengan un enfoque poblacional. En estos escenarios se realizará una demostración del uso del algoritmo en grandes volúmenes de datos, y como el algoritmo puede realizar *matching* para contribuir al control del sesgo de confusión en estudios observacionales.

Talleres

Se realizarán talleres de aprendizaje enfocados en el modo de uso y aplicabilidad del algoritmo, los cuales se impartirán de forma remota y/o presencial. En cada taller que se imparta, se entregará un manual de uso de **Epimatch** digitalizado.

Prensa

Se realizará un punto de prensa en la Universidad de los Andes para presentar el nuevo algoritmo **Epimatch**, y como esta innovadora tecnología se ajusta con las políticas públicas del país, como lo es la nueva Política Nacional de Inteligencia Artificial.

Redes sociales

El departamento de comunicaciones de la Universidad de los Andes se encargará de difundir el algoritmo **Epimatch** en twitter, Instagram y Facebook, como también en la página web de la misma universidad.

2. CAPACIDAD DE GESTIÓN Y ASOCIATIVIDAD

1. CAPACIDAD DE GESTIÓN

COSTO TOTAL DEL PROYECTO

ITEM	VALOR
GASTOS EN PERSONAL	6.000.000
DESARROLLO INTERFAZ	15.000.000
EQUIPAMENTO	0
INFRAESTRUCTURA Y MOBILIARIO	0
PATENTE DEL PROYECTO	160.000
CAPACITACIÓN/TALLERES	200.000
PASAJES	100.000
CONGRESOS	100.000
MANUAL	300.000
DIFUSIÓN EPIMATCH	3.000.000
PUBLICACIÓN CIENTÍFICA	1.000.000
TOTAL	25.860.000.-

NOMBRE	INSTITUCIÓN	PROFESIÓN	CARGO EN EL PROYECTO	Funciones y Capacidades Críticas que aportará al proyecto	% dedicación Mensual (calculado en base a 180hrs. mensuales)	$/HH
EQUIPO DE INVESTIGACIÓN						
Investigador 1 (director)	Clínica Dávila	Enfermera	Investigadora principal	Coordinación y dirección del proyecto.	22.2% (40hrs)	10.000
Investigador 2 (director alterno)	Universidad de los Andes	Epidemióloga	Investigadora secundaria	Soporte en la coordinación y dirección del proyecto	11.1%(20hrs)	10.000

Investigador 3	Profesor *Data Science*	*Data Science*	Desarrollador	Desarrollo del algoritmo	22.2% (40hrs)	15.000
Investigador 4	Clínica Dávila	Médico – gestión bases de datos.	Gestor bases de datos	Encargado de obtener las bases de datos en salud	4.4%(8hrs)	10.000
PERSONAL TÉCNICO DE APOYO						
Informático I	Empresa de informática	Ingeniero informático	Desarrollador de interfaz	Encargado del desarrollo de la interfaz del algoritmo	11.1%(20hrs)	10.000
Informático II	Empresa de informática	Ingeniero informático	Desarrollador de interfaz	Encargado del desarrollo de la interfaz del algoritmo	11.1%(20hrs)	10.000
PERSONAL ADMINISTRATIVO						
Secretaria	Clínica Dávila	Secretaria	Secretaria	Encargada de la recepción de documentos, recibir llamadas y organizar reuniones	4.4%(8hrs)	5.000

Porcentaje de Dedicación mensual en otros Proyectos				
CARGO EN EL PROYECTO	**NOMBRE**	**2022**	**2023**	**2024**
Director	XXX	25% de Dedicación	25% de Dedicación	25% de Dedicación
Director Alterno	XXX	25% de Dedicación	25% de Dedicación	25% de Dedicación
Investigador 3	XXX	20% de Dedicación	20% de Dedicación	20% de Dedicación
Investigador 4	XXX	20% de Dedicación	20% de Dedicación	20% de Dedicación

2. ANTECEDENTES CURRICULARES DEL EQUIPO DE INVESTIGACIÓN

Director: profesional de la salud con experiencia en investigaciones epidemiológicas observacionales poblacionales retrospectivas, como también en investigaciones biomédicas. También cuenta con experiencia en coordinación de protocolos y proyectos de investigación, y en manejo de base de datos.

Director alterno: profesional de la salud y académica, con una amplia trayectoria en el desarrollo de proyectos y protocolos en salud e investigación. Se especializa en el área de la epidemiología, específicamente en estudios poblacionales a gran escala, como también contar con estudios estadísticos especializados.

Investigador 3: docente en diversos cursos de Inteligencia Artificial y Machine Learning. Se ha especializado como desarrollador de algoritmos para el análisis de datos y fundar su propia empresa en informática.

Investigador 4: profesional de la salud con experiencia en el manejo de bases de datos en el ambiente clínico. Cuenta con formación y especialización en el extranjero, y actualmente desempeña sus funciones en una institución clínica de salud.

3. PARTICIPACION DE INVESTIGADORES EN FORMACIÓN

Investigador en formación: se invitará a participar a un estudiante de post grado en formación en *Data Science*. Esta persona participará de todo el proceso del proyecto y junto con el equipo de informática, podrá desarrollar el algoritmo que propone el presente proyecto.

4. ASOCIATIVIDAD

Universidad de los Andes: esta institución de educación superior aportará con personal, recursos e infraestructura para el desarrollo del proyecto.

Clínica Dávila: esta institución de salud aportará con personas y bases de datos en salud para la validación del algoritmo, previa autorización del Comité Ético Científico (CEC).