INSTITUTO POLITÉCNICO NACIONAL

Centro de Investigación en Computación

T E S I S

Minería de reglas de asociación mediante optimización multiobjetivo

PARA OBTENER EL GRADO DE:

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

Ing. Santiago Sinisterra Sierra

Directores de tesis:

Dr. Salvador Godoy Calderón

Dra. Miriam Pescador Rojas

Ciudad de México

Junio 2022

SIP-13
REP 2017

INSTITUTO POLITÉCNICO NACIONAL
SECRETARIA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REGISTRO DE TEMA DE TESIS
Y DESIGNACIÓN DE DIRECTOR DE TESIS

Ciudad de México, a 02 de marzo del 2021

El Colegio de Profesores de Posgrado del **Centro de Investigación en Computación** en su Sesión
(Unidad Académica)

Ordinaria No 12 celebrada el día 21 del mes diciembre de 2020, conoció la solicitud presentada por el (la) alumno (a):

Apellido Paterno:	SINISTERRA	Apellido Materno:	SIERRA	Nombre (s):	SANTIAGO

Número de registro: B 2 0 0 4 3 8

del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Referente al registro de su tema de tesis; acordando lo siguiente:

1.- Se designa al aspirante el tema de tesis titulado:

"Minería de reglas de asociación mediante optimización multiobjetivo"

Objetivo general del trabajo de tesis:

Aplicar métodos de optimización multiobjetivo sobre metaheurísticas bioinspiradas para llevar a cabo procesos de minería de reglas de asociación sobre grandes conjuntos de datos.

2.- Se designa como Directores de Tesis a los profesores:

Director: **Dr. Salvador Godoy Calderón** 2ª. Directora: **Dra. Miriam Pescador Rojas**

No aplica: ☐

3.- El Trabajo de investigación base para el desarrollo de la tesis será elaborado por el alumno en:

Centro de Investigación en Computación

que cuenta con los recursos e infraestructura necesarios.

4.- El interesado deberá asistir a los seminarios desarrollados en el área de adscripción del trabajo desde la fecha en que se suscribe la presente, hasta la aprobación de la versión completa de la tesis por parte de la Comisión Revisora correspondiente.

Director de Tesis

Dr. Salvador Godoy Calderón

2ª. Directora de Tesis

Dra. Miriam Pescador Rojas

Aspirante

Santiago Sinisterra Sierra

Presidente del Colegio

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION
EN COMPUTACION
DIRECCION

Dr. Marco Antonio Moreno Ibarra

SIP-14
REP 2017

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 12:00 horas del día 13 del mes de junio del 2022 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Posgrado de: Centro de Investigación en Computación para examinar la tesis titulada:

"Minería de reglas de asociación mediante optimización multiobjetivo" del (la) alumno (a):

Apellido Paterno:	SINISTERRA	Apellido Materno:	SIERRA	Nombre (s):	SANTIAGO

Número de registro: B 2 0 0 4 3 8

Aspirante del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Una vez que se realizó un análisis de similitud de texto, utilizando el software antiplagio, se encontró que el trabajo de tesis tiene 0 % de similitud. **Se adjunta reporte de software utilizado**.

Después que esta Comisión revisó exhaustivamente el contenido, estructura, intención y ubicación de los textos de la tesis identificados como coincidentes con otros documentos, concluyó que en el presente trabajo **SI** ☐ **NO** ☒ **SE CONSTITUYE UN POSIBLE PLAGIO.**

JUSTIFICACIÓN DE LA CONCLUSIÓN: *(Por ejemplo, el % de similitud se localiza en metodologías adecuadamente referidas a fuente original)*

El reporte turnitin indica 0% de similitud general.

****Es responsabilidad del alumno como autor de la tesis la verificación antiplagio, y del Director o Directores de tesis el análisis del % de similitud para establecer el riesgo o la existencia de un posible plagio.**

Finalmente y posterior a la lectura, revisión individual, así como el análisis e intercambio de opiniones, los miembros de la Comisión manifestaron **APROBAR** ☒ **SUSPENDER** ☐ **NO APROBAR** ☐ la tesis por **UNANIMIDAD** ☒ o **MAYORÍA** ☐ en virtud de los motivos siguientes:

Cumple con todos los requerimientos.

COMISIÓN REVISORA DE TESIS

Dr. Salvador Godoy Calderón
Director de Tesis

Dr. René Luna García

M. en C. Germán Téllez Castillo

Dra. Miriam Pescador Rojas
2° Director de Tesis

Dr. Francisco Hiram Calvo Castro

Dr. Gilberto Lorenzo Martínez Luna

Dr. Francisco Hiram Calvo Castro
PRESIDENTE DEL COLEGIO DE PROFESORES

INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
DIRECCIÓN
IPN-CIC

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA DE AUTORIZACIÓN DE USO DE OBRA PARA DIFUSIÓN

En la Ciudad de México el día 13 del mes de Junio del año 2022, el (la) que suscribe Santiago Sinisterra Sierra alumno(a) del programa Maestría en Ciencias de la Computación con número de registro B200438, adscrito(a) a Centro de Investigación en Computación manifiesta que es autor(a) intelectual del presente trabajo de tesis bajo la dirección de Dr. Salvador Godoy Calderón y Dra. Miriam Pescador Rojas y cede los derechos del trabajo intitulado Minería de reglas de asociación mediante optimización multiobjetivo, al Instituto Politécnico Nacional, para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expresado del autor y/o director(es). Este puede ser obtenido escribiendo a las siguiente(s) dirección(es) de correo sinisterra@protonmail.com. Si el permiso se otorga, al usuario deberá dar agradecimiento correspondiente y citar la fuente de este.

Santiago Sinisterra Sierra

Nombre completo y firma autografá del (de la) estudiante

Resumen

La minería de reglas de asociación es un conjunto de técnicas de análisis de datos que buscan descubrir relaciones interesantes entre los atributos de un conjunto de datos, articulándolas como expresiones lógicas.

Los algoritmos tradicionales de minería de reglas de asociación requieren de tiempo y recursos de cómputo considerables, generan grandes volúmenes de reglas que dependen de umbrales arbitrarios, los cuales pueden omitir inadvertidamente reglas interesantes.

Para resolver los retos mencionados, se propone un algoritmo evolutivo multi-objetivo, el cual genera, recombina y evalúa patrones, enfocándose en recuperar reglas de alta calidad que describan las relaciones existentes entre dos grupos de atributos, evaluando reglas con múltiples medidas de evaluación, incluyendo medidas tanto clásicas como de estimación causal. Además, se propuso un algoritmo de descubrimiento de grupos de atributos, con la intención de resaltar escenarios prometedores para la minería de reglas.

Las contribuciones propuestas fueron probadas usando un conjunto de datos compuesto por 15.5 millones de registros con información que describe la pandemia de COVID-19 en México. La estrategia de minería de datos generó reglas accionables que describen relaciones de causa y efecto entre los atributos.

Abstract

Association rule mining is a set of data analysis techniques aiming to discover interesting relationships among the attributes of a data set, articulating them as logical expressions.

Traditional association rule mining algorithms require considerable computation time and resources, generate large volumes of rules and depend on arbitrary thresholds, which might inadvertently omit interesting rules.

In order to solve such challenges, an evolutionary multi-objective algorithm is proposed that generates, recombines and evaluates patterns, focusing on recovering high-quality rules that describe existing relationships among two groups of attributes, evaluating rules in terms of multiple evaluation measures, including both classical and causal estimation measures. Furthermore, an attribute group discovery algorithm is presented, aiming to highlight promising rule mining scenarios.

The proposed contributions were tested with a dataset composed of 15.5 million records with official data describing the COVID-19 pandemic in Mexico. The algorithm recovered rules with actionable cause-effect relationships among attributes.

Agradecimientos

Mi más profundo agradecimiento al Centro de Investigación en Computación, al Instituto Politécnico Nacional y al Consejo Nacional de Ciencia y Tecnología por las facilidades otorgadas para desarrollar este trabajo.

A mis padres, mi hermana, así como a mi familia y amigos, por su permanente e incondicional apoyo.

A mis directores de tesis, por su invaluable experiencia y guía.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En este capítulo, se describe el contexto de la propuesta de tesis, los objetivos general y específicos; así como las contribuciones realizadas. Finalmente, se señala la estructura de este documento.

1.1 Contexto del trabajo

La minería de reglas de asociación es un conjunto de técnicas de análisis de datos con el propósito de descubrir relaciones interesantes entre los atributos de un conjunto de datos, articulándolas por medio de expresiones lógicas.

El nivel de interés de una regla de asociación puede ser evaluado tanto de forma objetiva como de forma subjetiva [1]. Por un lado, el interés objetivo muestra la relevancia de una regla estudiando sus propiedades estadísticas, a través de medidas como el soporte, confianza y ascenso. Por otro lado, evaluar el grado de interés de una regla desde punto de vista subjetivo depende del conocimiento y experiencia de quienes utilizan los resultados de la minería. Una regla de asociación tendrá un valor agregado de interés si coincide con el conocimiento sobre el fenómeno en particular sobre el cual se está realizando la minería,

revelando reglas que coincidan con las expectativas, intuición y experiencia de los usuarios. Para recuperar reglas de asociación interesantes desde una perspectiva subjetiva es necesario integrar conocimiento del dominio del problema para realizar una minería más efectiva.

Asimismo, una regla de asociación puede ser considerada como interesante si es accionable; es decir, si sugiere una relación de causa y efecto entre los atributos relacionados la cual ayude a comprender el entorno, realizar predicciones y sugerir posibles estrategias de acción para manipular el entorno. Es necesario considerar en la minería de reglas de asociación una clase de medidas que evalúen la aptitud de las reglas como una relación de causa y efecto para así encontrar reglas accionables.

Los algoritmos clásicos, al utilizar medidas clásicas de evaluación para las reglas, son aptas para recuperar reglas de asociación interesantes desde un punto de vista objetivo. Sin embargo, como se señaló previamente, es crucial la integración de conocimiento de dominio para recuperar reglas de asociación interesantes que sean consistentes con la intuición y experiencia de los usuarios, así como para identificar relaciones con alto potencial en su accionabilidad, evaluándolas desde una perspectiva causal.

Desde el paradigma de minería clásica, una regla de asociación es interesante si supera los umbrales mínimos de soporte y confianza definidos por los usuarios antes de realizar la minería, cuya definición es un proceso de ensayo y error. Este paradigma presenta varios inconvenientes derivados de la selección de los umbrales.

Por un lado, si se selecciona un umbral alto las reglas tienden a ser evidentes, excluyendo inadvertidamente reglas de soporte y confianza bajas que son interesantes desde un punto de vista subjetivo, esto por la interpretación de las variables siendo relacionadas. Por otro lado, si se selecciona un umbral muy bajo, se genera un gran volumen de reglas, dificultando identificar las reglas más valiosas, exigiendo considerable tiempo y recursos de cómputo tanto para la minería de patrones frecuentes, la enumeración de todas las reglas y de un proceso secundario de minería entre las reglas encontradas. Un algoritmo de minería

de reglas de asociación debe evitar estos inconvenientes eliminando la definición forzosa de umbrales mínimos.

En resumen, los algoritmos clásicos de minería de reglas de asociación únicamente consideran nociones de interés objetivo, evitando integrar conocimiento de dominio a la minería, son sensibles a los umbrales definidos para soporte y confianza, y exigen de considerables recursos y tiempo de cómputo.

La propuesta presentada en este trabajo tiene la motivación de resolver algunas de las problemáticas identificadas en la minería tradicional de reglas de asociación.

En primer lugar, se propone integrar conocimiento de dominio de los usuarios por medio de la definición de grupos de atributos, para los cuales se diseñó un algoritmo capaz de realizar propuestas preeliminares de grupos. A través de la definición de los grupos de atributos, los usuarios hacen explícito su conocimiento e intuición sobre las relaciones que existen entre atributos.

Asimismo, para recuperar reglas de asociación accionables se exploran medidas de estimación causal, las cuales evalúan las reglas de asociación en un contexto de causa y efecto. La técnica propuesta también admite el uso de medidas clásicas de evaluación.

Finalmente, al ser un algoritmo evolutivo multiobjetivo, el algoritmo explora las regiones del espacio de búsqueda más prometedoras, buscando las reglas mejor evaluadas en más de una medida de evaluación sin necesidad de definir forzosamente de antemano umbrales de soporte y confianza. Por medio de la configuración de parámetros del algoritmo evolutivo como el criterio de paro, el tamaño de la población, entre otros, los usuarios pueden controlar la profundidad de la exploración del espacio de búsqueda.

El caso de estudio seleccionado para el trabajo es la minería de reglas de asociación de un conjunto de datos de 15.5 millones de entradas con información relacionada con el COVID-19 en México. Realizar minería de datos sobre este conjunto de datos en particular es de importancia para describir este fenómeno, utilizando las reglas de asociación recuperadas

para describir relaciones interesantes entre atributos médicos, epidemiológicos y demográficos que sean potencialmente accionables.

Las pruebas realizadas con el algoritmo cubrieron una variedad amplia de experimentos, relacionando escenarios compuestos por diferentes grupos de atributos con 10 problemas mono-objetivo y 3 problemas multiobjetivo.

Una vez planteado el contexto, se procede a enunciar los objetivos definidos para el presente trabajo.

1.2 Objetivos

El trabajo tiene los siguientes objetivos:

Objetivo general

El objetivo general del presente trabajo es proponer un algoritmo evolutivo multiobjetivo para realizar minería de reglas de asociación.

Objetivos específicos

Como objetivos específicos se tiene:

- Implementar algoritmos evolutivos de minería de reglas de asociación.
- Estudiar diferentes paradigmas para evaluar reglas de asociación, tanto el paradigma clásico como desde una perspectiva causal.
- Sugerir estrategias de agrupamiento de atributos para guiar la minería de reglas de asociación.
- Realizar minería de reglas de asociación en grandes volúmenes de datos.

- Probar las contribuciones propuestas tomando como caso de estudio información relacionada con la pandemia de COVID-19 en México.

1.3 Contribuciones realizadas

Las contribuciones principales realizadas por este trabajo son:

- Diseño, implementación y pruebas de un algoritmo evolutivo para minería de reglas de asociación con la capacidad de integrar conocimiento de dominio para resolver problemas de minería de reglas de asociación tanto mono-objetivo como multiobjetivo.
- Desarrollo de un algoritmo de agrupamiento de atributos basado en cubrimientos de un grafo.
- Integración de los algoritmos propuestos en un nuevo flujo de minería de reglas de asociación.

1.4 Estructura de la tesis

El trabajo comprende seis capítulos. En el capítulo 1 se presentó una breve introducción al trabajo con el contexto, objetivos y contribuciones realizadas.

El capítulo 2 plantea el marco teórico y conceptual que fundamenta la propuesta realizada. En el capítulo 3 se discuten las propuestas del área que son comparables con el trabajo realizado. En el capítulo 4 se describe a detalle los diferentes elementos que componen las contribuciones realizadas, demostrando los resultados obtenidos en el capítulo 5. Finalmente, en el capítulo 6 se manifiestan las conclusiones alcanzadas y trabajo a futuro. Se incluyen al final del trabajo las referencias al material bibliográfico utilizado.

Capítulo 2

Marco teórico y conceptual

En este capítulo se plantean los elementos más importantes de este trabajo, los cuales constituyen el marco teórico y conceptual.

2.1 Minería de datos

De acuerdo con Dasu y Johnson [2], la minería de datos exploratoria es el proceso preeliminar de descubrimiento de estructura de un conjunto de datos. Se destaca que un buen método de minería de datos exploratoria debe cumplir con las siguientes características:

- **Amplio espectro de aplicación**, realizando pocas suposiciones sobre el proceso estadístico que genera los datos antes de ejecutar el método;
- **Tiempo de respuesta rápido**, ya que durante la fase de exploración es frecuente realizar múltiples análisis de forma simultánea, volviendo inaceptables a los métodos de análisis que tomen mucho tiempo;
- **Fácil de actualizar**, evitando la recalibración de los modelos desde cero cuando se recibe nueva información;
- **Generar resultados adecuados para su uso posterior**, reduciendo el volumen

de los datos a un formato menor, permitiendo realizar inferencias intuitivas sobre las asociaciones y patrones presentes en el conjunto de datos;

- **Fácil de usar e interpretar**, donde los resultados del método de minería de datos exploratoria no exijan conocimiento experto sobre los modelos utilizados para entender los mismos, así como que el funcionamiento interno del modelo sea transparente y confiable; a diferencia, por ejemplo, del uso de redes neuronales.

Se señalan dos estrategias principales: el análisis paramétrico, en el cual busca definir los parámetros que describen las distribuciones de los atributos; y el análisis no-paramétrico, el cual es guiado por los datos, sin realizar suposiciones previas sobre la distribución y relaciones entre atributos.

Entre los métodos no paramétricos, Dasu y Johnson identifican la construcción de tablas de frecuencias, sobre las cuales se obtienen medidas como la probabilidad marginal, probabilidad condicional, y probabilidad conjunta; las cuales se utilizan para describir la relación que comparten dos atributos. Las reglas de asociación son evaluadas con medidas de esta clase, describiendo así relaciones que existen entre conjuntos de atributos.

2.2 Reglas de asociación

La minería de reglas de asociación es un conjunto de técnicas de análisis de datos que buscan descubrir relaciones interesantes entre los atributos de un conjunto de datos, articulándolas como expresiones lógicas.

De acuerdo a Fürnkranz et al. [3], una regla de asociación se define de la siguiente forma: dado un conjunto de registros, donde cada registro es un conjunto de ítems, una regla de asociación es una expresión de la forma $A \rightarrow C$. $A \rightarrow C$ se interpreta como "Si ocurre A, entonces ocurre C", indicando que los registros de un conjunto de datos que contienen A tienden a contener C también.

Es posible utilizar un lenguaje de atributo-valor para describir las reglas de asociación. En dicho contexto, la regla de asociación toma la forma siguiente:

$$A_1 \wedge A_2 \wedge \ldots A_m \rightarrow C_1 \wedge C_2 \wedge \ldots \wedge C_n$$

Donde el antecedente (A) y el consecuente (C) son cláusulas conjuntivas de términos llamados selectores. Cada término A_i y C_i es un selector, el cual es un predicado delimitado por corchetes que relaciona un atributo del conjunto de datos con un valor o conjunto de valores en su dominio. En esta notación, el operador de conjunción se omite, al estar implícito en la expresión lógica.

A continuación se muestra una de las reglas de asociación con la notación utilizada a largo del trabajo:

`[Neumonía = Sí][COVID-19 = Confirmado] → [Intubado = Sí][Defunción = Sí]`,

la cual se interpreta de la siguiente forma: "Si un paciente tiene neumonía y es un caso confirmado de COVID-19, entonces fue intubado y falleció". La regla es cierta con una probabilidad dada, ya que existirán observaciones para las cuales la regla no se cumpla.

2.3 Causalidad

Uno de los beneficios del estudio de la causalidad es que poseer conocimiento sobre las relaciones de causa y efecto ayuda a comprender el entorno, realizar predicciones y actuar de formas diferentes [4].

En la presente sección se discuten conceptos importantes en el contexto de este trabajo con respecto al estudio de la causalidad, específicamente desde una perspectiva que habilita la agencia o manipulación, característica de gran peso en la disciplina de inteligencia artificial.

2.3.1 Causalidad como manipulación

Una de las corrientes de estudio de la causalidad considera su relación con la capacidad de manipulación, definida como teoría de agencia o manipulabilidad [5]. Dentro de este dominio, dada una relación de causa y efecto $C \rightarrow E$, donde C es la causa y E el efecto; actuar o manipular C permite modificar qué tanto se observa E.

Esta noción de causalidad es importante para la disciplina de inteligencia artificial, ya que la identificación de relaciones de causa y efecto es una habilidad crucial. Pearl [6] señala que un sistema autónomo inteligente que intenta construir un modelo de su entorno no puede basarse exclusivamente en conocimiento causal pre-programado; en lugar de esto, debe ser capaz de traducir observaciones directas a relaciones de causa y efecto.

2.3.2 Modelo de causalidad de Pearl

El modelo de causalidad de Judea Pearl, conocido también como la "escalera de causalidad"[7], describe una jerarquía de tres habilidades cognitivas asociadas a un agente de inferencia causal. Éstas habilidades, ordenadas de forma ascendente en complejidad, son *observar*, *hacer* y *entender*.

Al primer nivel, la *asociación*, le corresponde la habilidad de *observar*. El objetivo es describir relaciones entre elementos del modelo de dominio. Para realizar esta tarea, un agente artificial se vale de herramientas de probabilidad y estadística; por ejemplo, la probabilidad condicional y la correlación. Las medidas clásicas de evaluación de reglas de asociación tales como soporte, confianza y ascenso, entre otras, pertenecen, como su nombre lo indica, al nivel de asociación.

La *intervención*, el siguiente nivel de la jerarquía, está relacionado con la habilidad de *hacer*. Un agente en este nivel, de acuerdo con Pearl, "predice el efecto de alteraciones deliberadas para producir un efecto deseado" [7]. Si bien es infactible tanto para este caso de

estudio como para muchas otras instancias de minería de reglas de asociación que un agente artificial pueda experimentar del modo planteado por Pearl, es viable estimar el efecto de una intervención a través de un *estudio de caso-control*, comparando la incidencia del efecto en dos intervenciones. En éste trabajo se propone utilizar el *efecto causal* como la medida para evaluar una regla de asociación como una intervención.

Finalmente, en el nivel más alto se encuentra el *contrafactual*, relacionado con la habilidad de *entender*. Un agente habilitado para lidiar con conocimiento contrafactual es capaz de concebir un "mundo que no puede ser observado, porque contradice lo que se observa"[7].

El entendimiento proviene de contrastar el mundo factual, que representa lo observado; con un mundo contrafactual, que como su nombre lo indica, no es observable directamente, pero puede ser estimado a partir de lo observado. Para comparar ambos mundos es necesario partir de un *modelo causal* que, en palabras de Pearl, "habilite predecir qué ocurriría en situaciones aún no visualizadas".

Un modelo causal, al describir las relaciones de causa y efecto entre las variables consideradas, permite entonces estimar una nueva configuración del mundo como efecto de una intervención específica sobre el modelo causal. Una regla de asociación puede ser concebida como un modelo causal mínimo con dos variables en el que se afirma que el antecedente causa el consecuente.

2.3.3 Necesidad y suficiencia

La noción de condiciones suficientes y necesarias ayuda a comprender profundamente el comportamiento de una relación de causa y efecto entre dos elementos [4].

Dada una implicación $A \rightarrow C$, A es una *condición suficiente* para C, esto significa que la ocurrencia de A garantiza la ocurrencia de C; es decir, es imposible tener A sin C, indicando que si A está presente, entonces C también debe estar presente. La implicación

$A \rightarrow C$ se interpreta como "A es una condición suficiente para C".

Por otro lado, afirmar que A es una *condición necesaria* para C señala que la ocurrencia de A es necesaria para la ocurrencia de C; es decir, si A no ocurriera, entonces C no ocuriría tampoco. En este caso, es la implicación de la regla recíproca $C \rightarrow A$ la que describe a A como una condición necesaria para C.

Las condiciones de necesidad y suficiencia pueden combinarse; es posible que una condición sea necesaria sin ser suficiente, que sea no necesaria y suficiente, que no cumpla ninguno de los dos criterios, o que cumpla con ambos, determinando una condición necesaria y suficiente.

Una condición necesaria y suficiente está dada por la expresión lógica $A \leftrightarrow C$, una expresión bicondicional. Esta expresión es interpretada como "A es una condición necesaria y suficiente para C".

La noción de necesidad y suficiencia que existe en la lógica es aplicable también a la causalidad. Tian y Pearl [8] afirman que la causalidad tiene dos caras, la *necesaria* y la *suficiente*, para las cuales se puede estimar su probabilidad a partir de datos estadísticos.

2.4 Medidas de evaluación

En ésta sección se discuten dos clases de medidas que se utilizan para evaluar una regla de asociación: las *medidas clásicas*, que describen la asociación que existe entre antecedente y consecuente y *medidas de estimación causal*, a través de las cuales se determina si antecedente y consecuente comparten una relación de causa-efecto se estiman el impacto causal de dicha relación.

2.4.1 Medidas clásicas

- **Soporte** (*supp*, por su nombre en inglés *support*). El soporte indica el porcentaje de reglas de todo el conjunto de datos que cumple con la regla de asociación. Es la probabilidad de la intersección de A y C, como se muestra en la ecuación 2.1. Una regla con soporte alto indica que la regla se cumple frecuentemente en el conjunto de datos.

$$supp(A \to C) = \frac{|A \cap C|}{|U|} = P(A \cap C) = P(A, C) \tag{2.1}$$

- **Confianza** ($conf$, por su nombre en inglés *confidence*). La confianza indica la probabilidad de observar consecuente si se conoce el antecedente. Es la probabilidad condicional de C dado A, como se señala en la ecuación 2.2. Una regla con un valor alto de confianza indica que el antecedente es un buen predictor del consecuente.

$$conf(A \to C) = \frac{supp(A \to C)}{supp(C)} = \frac{P(A \cap C)}{P(C)} = P(C|A) \tag{2.2}$$

- **Ascenso** ($lift$, por su nombre en inglés *lift*). El ascenso compara la probabilidad observada para la regla con la probabilidad que se observaría si el antecedente y consecuente fueran independientes. Una regla con un valor alto de ascenso indica que antecedente y consecuente presentan un alto grado de asociación, descartando la alternativa de que sean independientes. Su cómputo se realiza con la ecuación 2.3.

$$lift(A \to C) = \frac{supp(A \to C)}{supp(A) \cdot supp(C)} = \frac{conf(A \to C)}{supp(C)} \tag{2.3}$$

2.4.2 Medidas de estimación causal

De acuerdo a Rosenbaum [9], un *estudio observacional* es una investigación empírica de los efectos de un tratamiento. Las observaciones que participan en el estudio se asignan a dos grupos balanceados: el grupo experimental, el cual presenta el tratamiento siendo estudiado; y el grupo de control, que no lo presenta. Un estudio observacional es necesario cuando la asignación al grupo experimental o el grupo de control no son factibles; a diferencia de una prueba controlada aleatorizada, en la que la asignación a los grupos se realiza por el azar.

En el contexto de una regla de asociación $A \rightarrow C$, el antecedente, A, es el tratamiento o causa; mientras que el consecuente, C, es el efecto. Las observaciones que presentan A pertenecen al grupo experimental; mientras que las que son cubiertas por $\neg A$ pertenecen al grupo de control. La relación entre A y C es representada por una matriz de confusión, como se indica en la Tabla 2.1.

Tabla 2.1: Composición de la matriz de confusión.

	Observación Positiva (C)	**Observación Negativa ($\neg C$)**
Predicción Positiva (A)	Verdadero Positivo (TP)	Falso Negativo (FN)
Predicción Negativa ($\neg A$)	Falso Positivo (FP)	Verdadero Negativo (TN)

Para determinar el efecto del tratamiento, se comparan la incidencia del efecto en ambos grupos, calculando la probabilidad condicional de C, de la siguiente forma:

- Incidencia en el grupo experimental (EER, en inglés, *Experimental Event Rate*). La incidencia en el grupo experimental es la confianza de $A \rightarrow C$ en una muestra balanceada, como se indica en la Ecuación 2.4.

$$EER = conf(A \rightarrow C) = P(C|A) \tag{2.4}$$

- Incidencia en el grupo de control (CER, en inglés, *Control Event Rate*). La incidencia en el grupo de control es la confianza de $\neg A \rightarrow C$ en una muestra balanceada, como se indica en la Ecuación 2.5.

$$CER = conf(\neg A \rightarrow C) = P(C|\neg A) \tag{2.5}$$

El efecto causal compara la *incidencia en el grupo experimental* (EER, *Experimental Event Rate*) con la *incidencia en el grupo de control* (CER, *Control Event Rate*). Entre más alto sea el efecto causal, mayores son los indicios de que el antecedente tiene un efecto sobre el consecuente.

El efecto causal puede ser reportado tanto en la escala aditiva como en la escala multiplicativa[10], siendo conocido como *efecto absoluto* (AR, en inglés *Absolute Risk*) y *efecto relativo* (RR, en inglés, *Relative Risk*), respectivamente. La *razón de momios* (OR, en inglés, *Odds Ratio*), otra medida de estimación del efecto causal, asiste en la determinación de la significancia estadística del efecto [11].

El *efecto absoluto* (AR), como se muestra en la Ecuación 2.6, es la diferencia en porcentaje entre la incidencia en el grupo experimental y la incidencia en el grupo de control.

$$AR = EER - CER \tag{2.6}$$

El rango del efecto absoluto va de -1 a 1, donde un valor mayor a cero indica que el antecedente tiene un efecto causal sobre el consecuente.

El *efecto relativo* (RR, *relative risk*), expresado en la Ecuación 2.7, refleja, al ser un multiplicador, cuántas veces es mayor la incidencia en el grupo experimental que en el grupo de control.

$$RR = \frac{EER}{CER} \tag{2.7}$$

El *efecto relativo* tiende a ser muy grande cuando la incidencia en el grupo de control es baja, revelando la rareza del efecto.

La *razón de momios* (OR, *odds ratio*) estima el efecto causal entre antecedente y consecuente sin asumir la dirección del efecto; es decir, la razón de momios tendrá el mismo valor tanto para $A \rightarrow C$ como para $C \rightarrow A$. La razón de momios es entonces una *medida simétrica* [12] del efecto causal, contrastando con el efecto absoluto y el efecto relativo, las cuales son *medidas asimétricas* ya que asumen que la dirección del efecto es del antecedente actuando sobre el consecuente, tomando un valor diferente cuando se evalúa el efecto causal en $C \rightarrow A$.

La importancia de la razón de momios recae en la determinación de la significancia estadística del efecto causal. El cómputo de la razón de momios se describe en la Ecuación 2.8, donde se introducen los valores provenientes de la matriz de confusión con la estructura indicada en la Tabla 2.1.

$$OR = \frac{TP/FP}{TN/FN} \tag{2.8}$$

El rango de la razón de momios va de cero a infinito, donde un valor mayor o igual a uno indica que el antecedente y consecuente comparten un efecto causal, aunque sin asumir su dirección.

La *significancia estadística* de la razón de momios se determina calculando el intervalo de confianza al 95%. Se calcula en primera instancia el error estándar (ω), dado por la Ecuación 2.9.

$$\omega = SE\{\ln(OR)\} = \sqrt{\frac{1}{TP} + \frac{1}{TN} + \frac{1}{FP} + \frac{1}{FN}} \tag{2.9}$$

Posteriormente, se sustituye ω en la Ecuación 2.10 para evaluar las cotas inferior y superior del intervalo de confianza. La razón de momios es estadísticamente significativa si la cota inferior del intervalo de confianza es mayor a uno.

$$CI(OR) = (exp(ln(OR) - 1.96\omega), exp(ln(OR) + 1.96\omega)) \tag{2.10}$$

La *suceptibilidad* o *probabilidad de suficiencia* (PS) está dada por la ecuación 2.11. Pearl indica que ésta medida representa la "capacidad de x [el antecedente, A] para producir y [el consecuente, C]" [6] [13]. La expresión que describe esta medida de evaluación está dada por la Ecuación 2.11.

$$PS = \frac{P(C|A) - P(C|\neg A)}{1 - P(C|\neg A)} = \frac{AR}{1 - CER} \tag{2.11}$$

La *fracción atribuible en la población* o *impacto en la población* (AF_p) es una medida de evaluación utilizada para estudiar el impacto de la exposición a cierta variable en la población [14]. En el contexto de minería de datos, la población se refiere al total de registros que presentan C, el efecto siendo estudiado.

La fórmula para calcular el impacto en la población, propuesta por Miettinen [15], está dada en la Ecuación 2.12. La medida involucra el soporte de C y el riesgo relativo.

$$AF_p = supp(C) \cdot (1 - \frac{1}{RR}) \tag{2.12}$$

La medida de impacto en la población tiene una interpretación causal, la cual indica la fracción estimada de todas las observaciones del consecuente que no hubieran ocurrido si no se hubiera observado el antecedente [16].

2.5 Cómputo evolutivo

Poli et al. [17] definen a la programación genética como "un método sistemático e independiente de dominio para resolver problemas de forma automática a partir de declaraciones de alto nivel sobre el problema a resolver".

La programación genética evoluciona soluciones representadas por medio de árboles sintácticos, transformándolos de forma estocástica para generar soluciones mejores mediante operadores genéticos.

Dos de los operadores más utilizados son la recombinación y la mutación. El operador de recombinación produce nuevas estructuras a partir de elementos seleccionados de forma aleatoria entre dos soluciones, mientras que el operador de mutación altera una parte del mismo.

Tras la aplicación de los operadores genéticos, se evalúa la calidad de una solución mediante una función de aptitud. De acuerdo a la aptitud se seleccionan las mejores soluciones para la siguiente generación, aplicando de nuevo los operadores genéticos iterativamente hasta alcanzar el criterio de paro, usualmente, el número de generaciones.

2.6 Optimización multi-objetivo

La optimización, de acuerdo a la definición dada por Mirjalili y Dong [18], es el proceso de encontrar un conjunto de soluciones denominadas como óptimas entre el conjunto de todas las posibles soluciones para un problema dado. Para comparar soluciones entre sí y evaluar su calidad, se utilizan una o varias funciones objetivo. La cantidad de funciones objetivo describen al problema: denominando al problema como mono-objetivo si tiene una única función objetivo, mientras que es multi-objetivo para dos o más funciones objetivo.

Formalmente, un problema multi-objetivo se define como se indica en la Ecuación

2.13:

$$\begin{aligned} \text{maximizar } F(\vec{x}) &= \{f_1(\vec{x}), f_2(\vec{x}), \ldots, f_o(\vec{x})\} \\ \text{sujeto a } \; g_i(\vec{x}) &\geq 0, i = 1, 2, \ldots, m \\ h_i(\vec{x}) &= 0, i = 1, 2, \ldots, p \\ lb_i \leq x_i &\leq ub_i, i = 1, 2, \ldots, n \end{aligned} \tag{2.13}$$

donde $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ es el vector de variables de decisión, n es el número de variables, m es el número de condiciones de desigualdad, p es el número de condiciones de igualdad y lb_i y ub_i son las cotas inferior y superior, respectivamente de las variables de decisión i-ésima. Para esta definición, f_i son las funciones objetivo, g_i las condiciones de desigualdad, h_i las condiciones de igualdad y lb_i y ub_i las cotas de las variables de decisión.

En un problema de optimización multi-objetivo, cuando se optimizan múltiples objetivos de forma simultánea, no existe una solución óptima. En su lugar, existe un conjunto de soluciones óptimas, cada una considerando un balance o *trade-off* entre los objetivos. Un algoritmo de optimización multi-objetivo produce finalmente un conjunto de soluciones, delegando al tomador de decisión la selección de aquella solución que mejor satisface su problema [19] [20].

El frente de Pareto es el conjunto de soluciones óptimas que satisfacen el problema de optimización multi-objetivo planteado. Para identificar el frente de Pareto, es necesario realizar un ordenamiento no-dominado de las soluciones disponibles, comparándolas entre sí para determinar si una solución domina a otra.

De acuerdo a Deb [20] [21], una solución S_1 domina a una solución S_2 si y sólo si se satisfacen los siguientes criterios:

- La solución S_1 no es peor que la solución S_2 en alguno de los objetivos;

- La solución S_1 es estrictamente mejor que la solución S_2 en al menos uno de los objetivos.

Una vez aplicado el ordenamiento no-dominado, se puede identificar el frente de Pareto, el cual está compuesto por todas las soluciones no dominadas.

2.7 Técnicas de programación y manejadores de bases de datos

Memoización. La memoización es una técnica de programación que persiste la salida de una función para una entrada dada, reutilizando la salida de la función para llamadas posteriores con la misma entrada. Una función es apta para ser memoizable si es una función pura; es decir, para la misma entrada, la función debe siempre tener la misma salida. La función de evaluación de una reglas de asociación puede ser descrita como una función pura, siendo entonces candidata para la memoización. La función se ejecuta la primera vez que se llama con cierta entrada, momento en el que se persiste en memoria la salida de la función.

Base de datos en memoria. Las ventajas de la memoización expiran cuando el algoritmo se termina de ejecutar. Para preservar las evaluaciones de reglas a través del tiempo es necesario persistirlas en una base de datos. Éste caso de uso es apto para una base de datos en memoria, la cual es ágil para recuperar el valor de la evaluación asociado a la regla de asociación. Para ofrecer el mejor desempeño, el comportamiento de la función de evaluación buscará primero en la base de datos en memoria la evaluación de la regla, memoizando el resultado de la consulta ahorrando así consultas a la base de datos.

Base de datos columnar. Una base de datos en formato columnar está optimizada para responder a consultas de agregación, que son el tipo de consultas a hacer para evaluar una regla de asociación. Al cargar la fuente de datos a una base de datos columnar, es posible obtener respuestas en milisegundos a consultas a bases de datos con millones o cientos de millones de registros.

Capítulo 3

Estado del arte

En éste capítulo se describen los trabajos más relevantes con respecto a la minería de reglas de asociación; comenzando por las técnicas que tienen una perspectiva denominada como *minería clásica*, basada en la búsqueda de *ítemsets* frecuentes; pasando después a las técnicas de *minería evolutiva*, basadas en metaheurísticas bioinspiradas. Finalmente, se discuten técnicas de minería de reglas de asociación causales.

3.1 Minería clásica

Los algoritmos clásicos de minería de reglas de asociación recuperan todas las reglas de asociación que superen los umbrales mínimos de soporte y confianza.

APRIORI, propuesto por Agrawal y Srikant [22], es uno de los algoritmos pioneros de minería de reglas de asociación. Este algoritmo recorre la retícula del conjunto potencia de los atributos, generando todas las reglas de asociación candidatas posibles para cada conjunto de atributos frecuente. Un conjunto de atributos se le denomina como frecuente si supera el umbral mínimo de soporte.

El algoritmo *APRIORI* realiza una poda de los conjuntos de atributos valiéndose del

comportamiento anti-monotónico del soporte; el cual señala que si un conjunto de atributos es infrecuente, todos sus superconjuntos también serán infrecuentes, por lo que no es necesario generar reglas candidatas a partir de ellos. Una vez que se termina de recorrer la retícula y se generan todas las reglas de asociación candidatas, se evalúa su confianza, depurando las reglas que estén debajo del umbral mínimo de confianza, terminando entonces la minería.

APRIORI, al realizar una búsqueda exhaustiva, es capaz de recuperar todas las reglas de asociación. Sin embargo, únicamente es efectivo para un tamaño reducido de reglas candidatas, ya que el número de reglas candidatas crece exponencialmente con el número de atributos.

Han et al. [23] presentan FP-Growth, un algoritmo de minería de reglas de asociación basado en un árbol de prefijos, el cual organiza el espacio de conjuntos frecuentes.

Al presentar su solución, los autores expresamente señalan situaciones en las que el rendimiento de APRIORI disminuye, como en conjuntos que generan muchas reglas de asociación, patrones largos o umbrales bajos de soporte, los cuales reducen la capacidad de poda sobre la cual se basa APRIORI.

El algoritmo *FP-Growth*, para organizar la generación de soluciones candidatas, se genera un árbol de prefijos denominado como *FP-tree*, a partir del cual se pueden extraer los sub-árboles necesarios para generar los conjuntos frecuentes para cada atributo.

Si bien el costo de la generación de soluciones candidatas se reduce con el árbol de prefijos, *FP-Growth* también se ve limitado en su desempeño como APRIORI porque únicamente considera el soporte, sigue siendo sensible a umbrales bajos de soporte y exige la generación de todas las reglas candidatas que cumplan con las restricciones de soporte y confianza.

3.2 Minería evolutiva

Una de las problemáticas que enfrentan los algoritmos clásicos de minería de reglas de asociación es que realizan una búsqueda exhaustiva de todas las reglas de asociación en el espacio de búsqueda que superen un umbral dado de soporte y confianza, siendo esta tarea de considerable dificultad para volúmenes grandes tanto de transacciones como de atributos.

A diferencia de la minería clásica, la minería evolutiva de reglas de asociación utiliza algoritmos bio-inspirados para realizar la búsqueda de las reglas de asociación mejor evaluadas.

Telikani et al. [24] proponen una taxonomía para clasificar los algoritmos evolutivos de minería de reglas de asociación, en la que identifican cuatro categorías principales, siendo la más dominante la de los algoritmos basados en la evolución.

Dentro de los algoritmos basados en la evolución se destaca la clase de métodos basados en la genética, en el cual se señalan dos grupos principales, los cuales tienen cercana relación con el presente trabajo: algoritmos genéticos y programación genética. A continuación se discuten los trabajos más importantes para cada uno de los grupos señalados previamente.

Alatas y Akin [25] propusieron un algoritmo genético para realizar minería de reglas de asociación sin necesidad de generar conjuntos de atributos frecuentes y sin depender de umbrales mínimos de soporte y confianza. Para representar las reglas de asociación utilizan una codificación posicional, donde a cada atributo del conjunto de datos le corresponde un gen, en el cual se indica si el atributo aparece en el antecedente, el consecuente o no aparece. Como operadores genéticos consideran la recombinación y mutación. La función de aptitud es una combinación del soporte, confianza y número de atributos que aparecen en la regla. El criterio de paro utilizado fue el número de generaciones. El algoritmo fue probado en conjuntos de datos tanto sintéticos como reales, encontrando reglas de asociación con alto soporte y confianza, sin necesidad de encontrar conjuntos de ítems frecuentes o definir

umbrales mínimos de soporte y confianza.

Yan et al. [26] proponen ARMGA, un algoritmo genético para recuperar reglas de asociación sin especificar un soporte mínimo, fortaleciendo la noción de que los algoritmos evolutivos pueden realizar una búsqueda global y automatizada, característica importante para la minería de datos exploratoria. La representación elegida utiliza una codificación con longitud variable, donde el primer elemento del individuo indica el punto de corte entre antecedente y consecuente. El algoritmo implementa tres operadores genéticos: selección, recombinación multipunto y mutación. La función de aptitud es la confianza, sobre la cual se impone la restricción de que debe ser mayor que el soporte de la regla. El criterio de paro utilizado es el número de generaciones, limitando el tamaño de la población a 100 individuos.

Kabir et al. [27] identifican la importancia de una población inicial adecuada para un resultado efectivo de la minería evolutiva. Para atender esta problemática, proponen una técnica para subdividir el dominio de los atributos del conjunto de datos sobre el cual se realiza la minería. Sobre estas divisiones, denominadas como *m-dominios*, se generan individuos de tal forma que todos los atributos participen de forma balanceada en las reglas de asociación recuperadas, en lugar de una generación de individuos completamente aleatoria. La subdivisión del espacio de búsqueda preserva la diversidad de atributos en las reglas de asociación recuperadas.

Telikani et al. [24] señalan dos diferencias principales entre los algoritmos genéticos y la programación genética. La primera diferencia recae en la representación, donde los algoritmos genéticos tienden a usar una representación en forma de cadena, mientras que la programación genética lo hace en forma de árbol, listas dinámicas o mediante gramáticas. Una segunda diferencia es la longitud de la representación: en los algoritmos genéticos típicamente la longitud es fija, donde a cada cromosoma corresponde un atributo; mientras que en la programación genética la longitud siempre es variable.

Uno de los algoritmos de programación genética aplicados a reglas de asociación más

representativos es *G3PARM* (*Grammar-guided Genetic Programming Association Rule Mining*). Desarrollado por Luna et al. [28]; *G3PARM* resuelve mediante una gramática libre de contexto (*GLC*) uno de los problemas principales de los algoritmos de programación genética aplicados a reglas de asociación: la generación de individuos inválidos. Una gramática libre de contexto, compuesta por símbolos y reglas de producción, define las restricciones que debe tener un individuo, por lo que siempre se generan reglas de asociación válidas.

G3PARM utiliza como representación un árbol sintáctico dado por la gramática libre de contexto, sobre el cual se aplican un operador de recombinación, en el cual se intercambian sub-árboles compatibles con la *GLC*; así como un operador de mutación, el cual reemplaza un símbolo aleatorio del árbol por un nuevo símbolo del mismo tipo.

Las funciones objetivo utilizadas por *G3PARM* son soporte y confianza. Por un lado, durante el proceso evolutivo, *G3PARM* utiliza como función de aptitud el soporte. Al terminar cada generación se actualiza una población auxiliar con la población del algoritmo, la cual se ordena de acuerdo a la confianza y depura las reglas duplicadas, así como de reglas de bajo soporte y confianza. El criterio de paro seleccionado es por número de generaciones, con un tamaño de población de 70. *G3PARM*, un algoritmo mono-objetivo, se compara contra algoritmos multi-objetivo como SPEA2 y NSGA-II con medidas como el soporte promedio, confianza promedio y la proporción de registros cubiertos por el conjunto de reglas correspondientes a la población final.

Luna et al. [29] presentan una versión modificada de su algoritmo *G3PARM* para realizar minería de reglas de asociación en una base de datos relacional. Al momento de realizar la evaluación de una regla de asociación esta se convierte a una consulta SQL, la cual realiza una agregación de los datos en la base para así calcular el soporte de la regla de asociación. En los experimentos realizados con esta variante de *G3PARM*, utiliza un tamaño de población de 50 y como criterio de paro 100 generaciones. La función de aptitud combina las medidas de soporte, confianza y ascenso.

Martín et al. [30] proponen un algoritmo genético multi-objetivo de minería de reglas de asociación basado en NSGA-II. El algoritmo maximiza tres objetivos: interés, comprensibilidad y el producto de soporte y confianza. La representación elegida es de longitud fija, sobre la cual se aplican tres operadores: recombinación en un solo punto, mutación y reparación. El criterio de paro seleccionado es el número de evaluaciones, deteniendo el algoritmo al alcanzar 50,000 evaluaciones utilizando un tamaño de población de 100. Los resultados obtenidos se comparan estudiando los valores promedio de la población final de soporte, confianza y ascenso.

3.3 Minería de reglas causales

Li et al. [31] introducen una técnica de minería de reglas de asociación causales. Para el paradigma causal, los atributos del antecedente (A) se presentan como combinaciones de causas, las cuales tienen un efecto, representado por el consecuente. La identificación de estas reglas se realiza simulando un estudio observacional en el que se realiza un muestreo balanceado, donde el soporte de A y $\neg A$ sea el mismo; es decir, 50%. Posteriormente, se construye una tabla de contingencia para A y C, calculando la razón de momios u *odds ratio* (OR). Se afirma entonces que A causa C si la razón de momios es mayor a uno y es estadísticamente significativa.

El algoritmo de minería de reglas de asociación realiza una búsqueda por niveles hasta una profundidad máxima siendo su comportamiento similar al de algoritmos clásicos de minería de reglas de asociación. Los autores señalan en los resultados obtenidos que el conjunto de reglas de asociación causales es pequeño comparado con el conjunto de reglas de asociación, lo cual se debe a la restricción de significancia estadística de la razón de momios. Si bien utilizan la razón de momios para depurar reglas de asociación irrelevantes, la evaluación de las reglas se hace aún mediante soporte y confianza.

Yadav et al. [32] realizan minería de reglas de asociación usando medidas de evaluación

causal. En su trabajo, los autores prueban cinco técnicas para estimar el efecto causal de una regla de asociación; entre los cuales se encuentra el efecto absoluto, el cual está dado por la diferencia en confianza de $A \Rightarrow C$ y $\neg A \Rightarrow C$. La minería de reglas de asociación se realiza aplicando el algoritmo APRIORI.

A modo de conclusión, en la revisión del estado del arte se comenzó con una revisión de los algoritmos clásicos de minería de reglas de evaluación, describiendo los elementos más importantes de la disciplina como las medidas de evaluación más comunes, así como enunciando las limitaciones de los algoritmos clásicos debido al tamaño del espacio de búsqueda y la dependencia de umbrales arbitrarios de soporte y confianza.

Los algoritmos de minería evolutiva realizan aportes para resolver las problemáticas de los algoritmos clásicos al eliminar la necesidad de realizar minería de patrones frecuentes y la definición de umbrales mínimos de soporte y confianza, los cuales son difíciles de definir en un contexto de minería exploratoria de datos. Además, los algoritmos evolutivos pueden realizar la búsqueda de las mejores reglas de evaluación tanto en un contexto mono-objetivo como multi-objetivo.

Finalmente se estudiaron trabajos en los que se experimenta con la minería de reglas de asociación causales, las cuales describen relaciones de causa y efecto entre los atributos de un conjunto de datos, criterio de especial importancia para recuperar reglas de asociación que puedan ser de interés en el problema de minería de datos.

Estos trabajos fundamentan la propuesta realizada: un algoritmo evolutivo multi-objetivo capaz de recuperar reglas de asociación tanto causales como no-causales.

Capítulo 4

Diseño de la propuesta

En este capítulo se describe el flujo propuesto para realizar minería de reglas de asociación así como dos de los algoritmos diseñados: el algoritmo de descubrimiento de grupos de atributos, el cual provee las relaciones más prometedoras entre variables para realizar minería de datos y el algoritmo evolutivo para la búsqueda eficiente de reglas de asociación.

En la figura 4.1 se muestra un diagrama de bloques del flujo, el cual incluye cinco fases, las cuales serán descritas en las siguientes secciones.

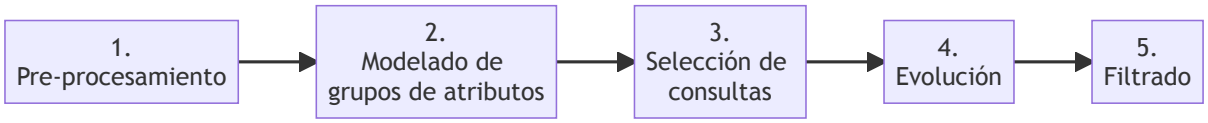

Figura 4.1: Diagrama a bloques del flujo de minería de reglas de asociación

4.1 Fase de pre-procesamiento de datos

En esta sección se discuten las técnicas de pre-procesamiento de datos para la propuesta de tesis, se comparan las técnicas de discretización exploradas y se analizan las ventajas

que ofrecen para su acoplamiento a la minería de reglas de asociación.

Por lo general los conjuntos de datos sobre los cuales se realiza minería tienen atributos tanto categóricos como numéricos, haciendo necesaria una discretización de los datos numéricos para transformarlos en atributos categóricos. Por lo tanto, es necesario aplicar una técnica de discretización, transformando un atributo con valores continuos a n valores discretos.

Se probaron dos técnicas de discretización: la discretización basada en clases (*bins*) y la discretización basada en cuantiles. La discretización basada en clases divide un atributo numérico en n rangos de la misma amplitud; mientras que la discretización basada en cuantiles divide al atributo numérico en n rangos con amplitud variable, buscando que cada rango cubra aproximadamente la misma cantidad de instancias.

Entre las dos técnicas se seleccionó la discretización basada en cuantiles [33], eligiendo dividir los atributos numéricos en cinco cuantiles, conocidos también como quintiles. Esta decisión se justifica por las siguientes razones:

- **Soporte uniforme**. Cada valor del atributo discretizado tiene aproximadamente el mismo soporte, el cual es del 20%. En una discretización basada en clases ésta condición no se cumple.
- **Menor impacto de valores atípicos**. Una discretización basada en clases puede producir uno o varios atributos con soporte muy bajo, generando valores atípicos. En contraste, una discretización basada en cuantiles acumula valores atípicos en dos rangos, asignando valores muy bajos al primer quintil y valores muy altos al último quintil.
- **Principio de Pareto o regla del 80-20**. Este principio empírico afirma que el 80% de la incidencia de un factor es atribuible al 20% de los observaciones [34]. Dividir en quintiles arroja atributos cuyos valores tienen un 20% de soporte cada uno.

En la Tabla 4.1 se muestra una comparativa entre ambas técnicas de discretización para el atributo de edad. Como se puede apreciar, la discretización basada en cuantiles, a

diferencia de la discretización basada en clases, genera atributos con soporte uniforme y contiene a los valores atípicos en un único rango.

Tabla 4.1: Comparativa entre métodos de discretización.

	Clases			**Quintiles**		
	Rango	Registros	Soporte (%)	Rango	Registros	Soporte (%)
1	[0, 24]	3,196,583	20.51	[0, 24]	3,196,583	20.51
2	[25, 48]	8,143,960	52.26	[25, 32]	3,157,761	20.26
3	[49, 72]	3,685,629	23.65	[33, 42]	3,263,396	20.94
4	[73, 96]	544,522	3.49	[43, 53]	2,986,363	19.16
5	[96, inf]	11,808	0.05	[54, inf]	2,978,399	19.11

También, al analizar los datos, se reporta como ejemplo de la regla del 80-20 para el caso de estudio, ya que, al estudiar la incidencia de las defunciones en cada rango de edad, se destaca que el 74% de las defunciones ocurrieron en personas con 54 o más años, las cuales representan solo un 19% de los registros.

4.2 Fase de modelado de grupos de atributos

Los atributos del conjunto de datos se distribuyen en grupos, los cuales son conjuntos mutuamente excluyentes.

Esta fase puede realizarse de forma manual, especificando directamente los grupos de atributos. La determinación de grupos de atributos puede ser arbitraria, agrupandolos de forma semántica. Un ejemplo puede ser reunir los atributos relacionados con padecimientos, datos demográficos, tratamientos médicos, entre otros.

Sin embargo, la determinación manual de grupos de atributos puede ser infactible en algunas situaciones. Esta tarea plantea dificultades cuando se aplica el algoritmo de minería para realizar exploración de datos, fase en la cual se tiene poco conocimiento experto sobre los atributos y su contenido; cuando el volumen de atributos es grande, o bien, si se desea mantener ocultos los nombres de los atributos, por ejemplo, cuando se tienen datos sensibles.

Para resolver dicha problemática se propone un algoritmo de descubrimiento de grupos de atributos, el cual realiza un agrupamiento mediante la construcción de un grafo a partir de una medida de asociación estadística entre pares de atributos.

Utilizar el algoritmo de descubrimiento propuesto es de utilidad incluso cuando la determinación de los grupos se realiza de forma manual, ya que permite delegar al algoritmo la generación de una propuesta de grupos para ser ajustada posteriormente.

Otro aporte que hace el algoritmo de descubrimiento de grupos de atributos es proponer sugerencias de consultas a realizar, identificando pares de grupos que estén relacionados a través de los atributos que los componen.

En conclusión, la determinación de grupos semánticos puede efectuarse de forma manual, diseñada por los usuarios de la minería; de forma automatizada, a través de un algoritmo de descubrimiento de grupos de atributos; o de forma híbrida, ajustando manualmente a partir de conocimiento experto una propuesta inicial generada por el algoritmo. De igual forma, otra de las capacidades del algoritmo es que puede identificar escenarios potenciales de grupos sobre los cuales realizar minería.

4.3 Algoritmo de descubrimiento de grupos de atributos

Un algoritmo de descubrimiento de grupos de atributos sirve como auxiliar en la minería de datos, cuando los usuarios tienen dificultades para determinar los grupos de atributos a relacionar. En esta sección se describen las fases que componen al algoritmo de descubrimiento de grupos de atributos.

En la Figura 4.2 se muestra el flujo del algoritmo de descubrimiento de grupos de atributos, el cual consta de cinco pasos:

1. **Construcción del grafo de asociación.** Se construye un grafo de asociación en el cual los vértices son los atributos del conjunto de datos, conectándolos en el grafo si

están asociados estadísticamente.

2. **Depuración de aristas.** Se ordenan las aristas de forma ascendente, eliminándolas una a una del grafo de asociación hasta encontrar una arista de corte, momento en el que se detiene el proceso.

3. **Cubrimiento del grafo de asociación.** Un cubrimiento en *cliques* o *grafos completos* del grafo de asociación se obtiene realizando una coloración de su grafo complemento.

4. **Construcción del grafo de grupos.** Se crea un grafo de grupos, en el que los nodos son los grupos y las aristas relacionan grupos que comparten atributos asignados a diferentes grupos.

5. **Identificación de escenarios potenciales.** Se seleccionan las aristas más importantes en el grafo de grupos, reportándolas como escenarios potenciales a ejecutar.

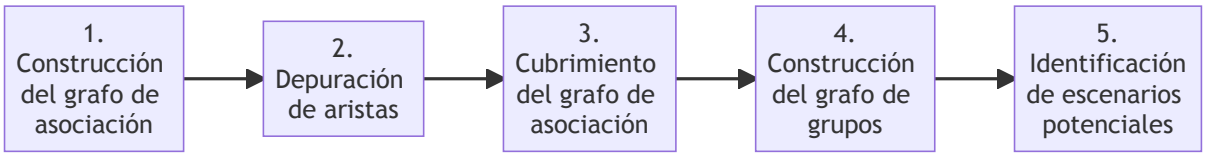

Figura 4.2: Diagrama a bloques del algoritmo de descubrimiento de grupos de atributos

4.3.1 Construcción del grafo de asociación

En la primera fase para la construcción del grafo de asociación, se seleccionó como medida de asociación la V de Crámer (V), la cual es una medida de asociación entre dos atributos categóricos derivada a partir de la tabla de contingencia utilizada para la prueba χ^2 de Pearson. La fórmula para calcular la V de Crámer se muestra en la Ecuación 4.1:

$$V = \sqrt{\frac{\chi^2/n}{min(k-1, r-1)}}, \tag{4.1}$$

donde n es el total de registros, k el total de columnas de la tabla de contingencia y r el total de filas.

La V de Crámer tiene un rango de valores interpretables en sus extremos, el cual va de cero a uno; donde cero indica que los atributos no están asociados, y uno indica que existe asociación total. Es una diferencia importante con χ^2, la cual no está acotada a un rango específico y crece conforme aumentan los grados de libertad de la tabla de contingencia. Asimismo, al ser derivada de χ^2, comparten el mismo p-valor al evaluar su significancia estadística, criterio necesario para incluir la arista en el grafo de asociación.

4.3.2 Depuración de aristas

El grafo de asociación de aristas usualmente se encuentra fuertemente conectado, resultando en pocos grupos al buscar un cubrimiento en grafos completos en la siguiente fase. Debido a lo anterior, será necesario eliminar aristas de asociaciones débiles para generar más grupos.

Una opción para realizar esta tarea es seleccionar un umbral de corte; eligiendo depurar todas las aristas cuyo peso sea menor o igual a cierto valor de referencia que indique una asociación débil.

Esta estrategia presenta dos inconvenientes. El primero es que los valores de referencia dados en la literatura para la V de Crámer indicando que una asociación es débil son inconsistentes y dependen del dominio de los datos. Además, en circunstancias en las que se utilice el algoritmo para la exploración de datos, fase en la cual no se posee de suficiente conocimiento experto, es infactible determinar qué umbral utilizar sin un proceso de ensayo y error.

Un umbral seleccionado de forma arbitraria puede generar más de una componente conectada en el grafo de asociación. Tener más de una componente conectada resulta en grupos aislados, lo que impide a dichos grupos participar en fases posteriores del algoritmo

ya que no comparten conexiones con otros grupos.

La solución para la selección del umbral es eliminar en orden ascendente cada una las aristas, hasta encontrar una arista de corte. Una arista de corte, conocida también como *puente*, es una arista que al ser eliminada generaría más de un componente conectado en el grafo de asociación. Esta estrategia conserva el grafo conectado y evita la selección de un umbral de forma arbitraria en caso de que no se tenga suficiente información para tomar dicha decisión.

4.3.3 Cubrimiento del grafo de asociación

Con las aristas débiles depuradas, se procede a encontrar un cubrimiento en *cliques* o grafos completos del grafo de asociación. Un *clique* es un grafo tal que cada par de nodos que lo componen son adyacentes. Para encontrar un cubrimiento en cliques se propone utilizar una coloración del grafo complemento [35].

Este método se compone de los siguientes pasos:

1. Obtener el complemento del grafo de asociación
2. Efectuar una coloración del grafo complemento, donde cada color conforma un *clique*.
3. Asignar los *cliques* de acuerdo a la coloración del grafo complemento.

Al construirse los grupos a partir de la coloración del grafo, el número cromático del grafo complemento indica el mínimo de grupos que pueden ser generados con esta técnica de cubrimiento.

Para realizar la coloración de los nodos, se seleccionó el algoritmo DSATUR[36]. Esta propuesta implementa un algoritmo voraz que prioriza los nodos a colorear con el *grado de saturación* del nodo, el cual está dado por la cantidad de colores adyacentes a un nodo aún sin colorear. La implementación utilizada es la que se encuentra integrada en la biblioteca *NetworkX*, disponible para Python.

Para ejemplificar el método de cubrimiento se generó un grafo aleatorio, al cual se le aplica el método de cubrimiento. En la Figura 4.3 se muestra a la izquierda el grafo generado aleatoriamente, mientras que a la derecha, con las aristas punteadas y en color rojo, se muestra su grafo complemento. Como se puede apreciar, en el grafo complemento dos nodos están conectados si el mismo par de nodos está desconectado en el grafo original.

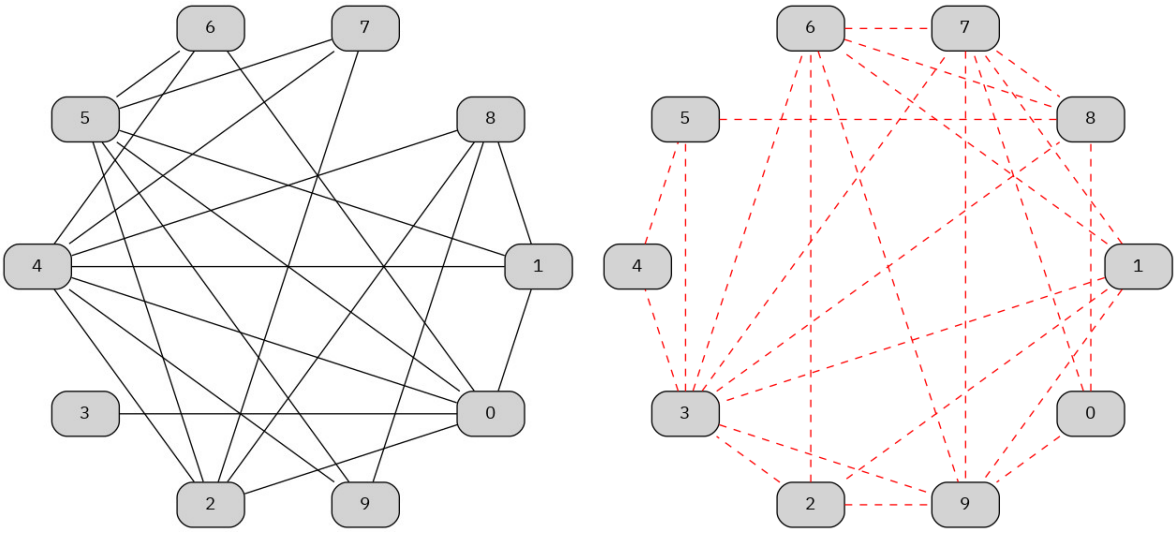

Figura 4.3: Grafo de ejemplo y su complemento

Posteriormente, se procede a realizar una coloración del grafo complemento. La coloración mostrada es una coloración válida ya que no existe un par de nodos adyacentes con el mismo color. La coloración del grafo complemento se muestra en la Figura 4.4, requiriendo de cinco colores.

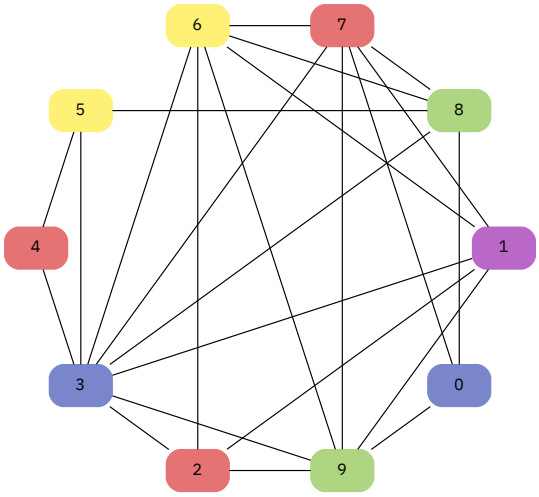

Figura 4.4: Coloración del grafo complemento del grafo de ejemplo

Finalmente, se asignan los colores dados a cada nodo en el grafo complemento al grafo original. Los nodos que comparten un mismo color conforman un *clique*. El cubrimiento del grafo de asociación se muestra en la Figura 4.5.

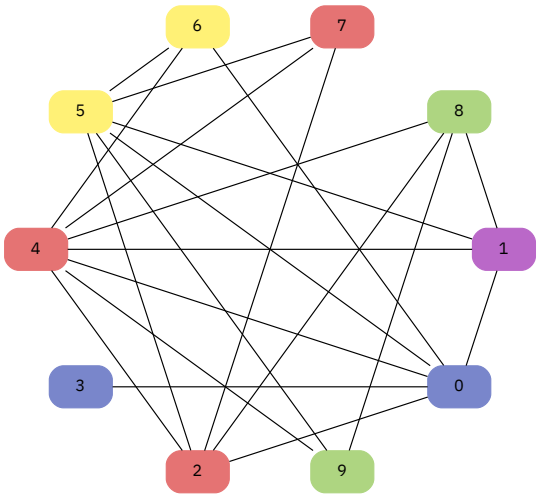

Figura 4.5: Cubrimiento en cliques del grafo de ejemplo

4.3.4 Construcción del grafo de grupos

Una vez que se tiene un cubrimiento en cliques del grafo de asociación, se procede a construir el grafo de grupos, el cual es un grafo cuyos nodos son los grupos identificados en la fase de cubrimiento.

Dos grupos, g_1 y g_2, están conectados por una arista en el grafo de grupos si existe al menos una arista en el grafo de atributos tal que conecte un atributo que pertenezca a g_1 con un atributo en g_2. El peso de cada arista en el grafo de grupos es la cantidad de aristas en el grafo de grupos que satisfacen la condición anterior.

4.3.5 Identificación de escenarios potenciales

Un escenario es un par de grupos de atributos sobre los cuales se realiza la minería de reglas de asociación, asignando uno al antecedente y otro al consecuente.

Si bien no hay impedimento para realizar una minería exhaustiva relacionando todas las permutaciones de grupos de atributos, no se recomienda ya que la cantidad de escenarios puede crecer significativamente y no superaría en desempeño a un algoritmo clásico de minería de reglas de asociación.

En su lugar es necesaria una técnica de identificación de escenarios potenciales a relacionar, para evitar así una minería exhaustiva, privilegiando los escenarios que puedan generar las reglas de asociación más prometedoras. Otro de los aportes es que la identificación asiste al caso de uso de exploración, ofreciendo sugerencias de escenarios sobre los cuales realizar una minería preliminar, dado el caso en el que no se conozca de antemano qué grupos deberían relacionarse.

La identificación de escenarios potenciales puede realizarse a partir del grafo de asociación. En primera instancia se obtiene el árbol de expansión máxima del grafo de asociación, para después seleccionar como escenario potencial aquellos pares de grupos (g_i, g_j) tales que

exista al menos una arista en en el árbol de expansión máxima que enlace a un atributo perteneciente a g_i con otro atributo en g_j. Utilizar el árbol de expansión máxima asegura que se enlacen los grupos por medio de atributos que compartan asociaciones fuertes.

4.4 Fase de construcción de consultas

Concluido el modelado de grupos de atributos, la siguiente fase consiste en la selección de las medidas que serán utilizadas para evaluar las reglas de asociación que serán minadas. Una consulta es la definición de un problema de optimización a resolver para un escenario en particular.

Cada problema de optimización está compuesto por un conjunto de funciones a optimizar así como de restricciones, tanto en el espacio de variables de decisión como en el espacio de los objetivos.

Para este caso de estudio se plantean tres consultas a realizar: una consulta de minería clásica, en la que se maximizan soporte, confianza y ascenso; una consulta de minería de reglas causales, identificando causas necesarias y/o suficientes; y finalmente una consulta de las reglas de asociación con altos valores de susceptibilidad y en la fracción atribuible en la población.

4.4.1 Restricciones por omisión

Por omisión se definen tres restricciones de caja para cada problema de optimización:

Soporte mayor a cero. La regla de asociación debe cumplirse para al menos un registro. En la definición formal de los problemas de optimización se plantea como $supp(A \rightarrow C) > 0$.

Significancia estadística. La razón de momios (OR, ver ecuación 2.8) debe ser

estadísticamente significativa; es decir, la cota inferior de su intervalo de confianza al 95% (CI^{inf}_{OR}, ver ecuación 2.10) debe ser mayor o igual a uno. Formalmente, en los problemas de optimización se plantea como $CI^{inf}_{OR}(A \rightarrow C) \geq 1$.

Efecto absoluto positivo. Una regla de asociación con un efecto absoluto (AR, ver ecuación 2.6) positivo, con valor mayor a cero, indica que observar el antecedente incrementa la probabilidad de observar el consecuente, descartando así las reglas en las que el antecedente inhibe al consecuente. Formalmente, en los problemas de optimización se plantea como $AR(A \rightarrow C) > 0$.

4.4.2 Problemas mono-objetivo

En esta sección se presentan diez problemas mono-objetivo, en los cuales se prueba maximizar las medidas de evaluación presentadas en el marco teórico y conceptual, así como combinaciones de las mismas a través de la media geométrica ya que esta admite la combinación de valores en diferentes escalas. Para todos los problemas mono-objetivos planteados se aplican las tres restricciones presentadas en la sección 4.4.1.

Los problemas planteados en las ecuaciones 4.2, 4.3 y 4.4 se maximizan respectivamente el soporte, confianza y ascenso, las medidas clásicas de evaluación de reglas de asociación. En la Ecuación 4.5 se maximiza la media geométrica de soporte, confianza y ascenso. En este problema se aprecia la importancia de utilizar la media geométrica, ya que el ascenso, al tener un rango de 0 a infinito, tomaría un peso desproporcionado en la media aritmética.

$$\begin{aligned} &\text{maximizar } supp(A \rightarrow C) \\ &\text{sujeto a } supp(A \rightarrow C) > 0, \\ &AR(A \rightarrow C) > 0, \\ &CI^{inf}_{OR}(A \rightarrow C) \geq 1. \end{aligned} \tag{4.2}$$

$$\begin{aligned} &\text{maximizar } conf(A \to C) \\ &\text{sujeto a } supp(A \to C) > 0, \\ &AR(A \to C) > 0, \\ &CI_{OR}^{inf}(A \to C) \geq 1. \end{aligned} \tag{4.3}$$

$$\begin{aligned} &\text{maximizar } lift(A \to C) \\ &\text{sujeto a } supp(A \to C) > 0, \\ &AR(A \to C) > 0, \\ &CI_{OR}^{inf}(A \to C) \geq 1. \end{aligned} \tag{4.4}$$

$$\begin{aligned} &\text{maximizar } \sqrt[3]{supp(A \to C) \cdot conf(A \to C) \cdot lift(A \to C)} \\ &\text{sujeto a } supp(A \to C) > 0, \\ &AR(A \to C) > 0, \\ &CI_{OR}^{inf}(A \to C) \geq 1. \end{aligned} \tag{4.5}$$

El problema de la ecuación 4.6 busca las reglas que maximicen el efecto causal de $A \to C$, buscando identificar causas suficientes; mientras que el problema de la ecuación 4.7 es de maximización del efecto causal de $C \to A$, la regla recíproca, la cual identifica a A como causa necesaria de C. En el problema 4.8 se buscan reglas bicondicionales del tipo $A \leftrightarrow C$, procurando encontrar reglas con valores altos de efecto causal y de efecto causal en la regla recíproca, combinándolos también con la media geométrica.

$$
\begin{aligned}
\text{maximizar} \quad & AR(A \to C) \\
\text{sujeto a} \quad & supp(A \to C) > 0, \\
& AR(A \to C) > 0, \\
& AR(C \to A) > 0, \\
& CI_{OR}^{inf}(A \to C) \geq 1, \\
& CI_{OR}^{inf}(C \to A) \geq 1.
\end{aligned}
\tag{4.6}
$$

$$
\begin{aligned}
\text{maximizar} \quad & AR(C \to A) \\
\text{sujeto a} \quad & supp(A \to C) > 0, \\
& AR(A \to C) > 0, \\
& AR(C \to A) > 0, \\
& CI_{OR}^{inf}(A \to C) \geq 1, \\
& CI_{OR}^{inf}(C \to A) \geq 1.
\end{aligned}
\tag{4.7}
$$

$$
\begin{aligned}
\text{maximizar} \quad & \sqrt[2]{AR(A \to C) \cdot AR(C \to A)} \\
\text{sujeto a} \quad & supp(A \to C) > 0, \\
& AR(A \to C) > 0, \\
& AR(C \to A) > 0, \\
& CI_{OR}^{inf}(A \to C) \geq 1, \\
& CI_{OR}^{inf}(C \to A) \geq 1.
\end{aligned}
\tag{4.8}
$$

Finalmente, en el problema 4.9 se realiza la minería de reglas de asociación con alta

susceptibilidad, señalando reglas en las que el antecedente tenga una alta probabilidad de producir el consecuente; mientras que en el problema 4.10 se buscan reglas con alto valor en la medida de impacto en la población, señalando causas a las cuales se les pueda atribuir la mayor proporción posible de los efectos. En el problema definido en la Ecuación 4.11 se buscan reglas con altos valores tanto de susceptibilidad e impacto en la población, combinando las medidas con la media geométrica.

$$\begin{aligned} \text{maximizar } & PS(A \to C) \\ \text{sujeto a } & supp(A \to C) > 0, \\ & AR(A \to C) > 0, \\ & CI^{inf}_{OR}(A \to C) \geq 1. \end{aligned} \tag{4.9}$$

$$\begin{aligned} \text{maximizar } & AF_p(A \to C) \\ \text{sujeto a } & supp(A \to C) > 0, \\ & AR(A \to C) > 0, \\ & CI^{inf}_{OR}(A \to C) \geq 1. \end{aligned} \tag{4.10}$$

$$\begin{aligned} \text{maximizar } & \sqrt[2]{PS(A \to C) \cdot AF_p(A \to C)} \\ \text{sujeto a } & supp(A \to C) > 0, \\ & AR(A \to C) > 0, \\ & CI^{inf}_{OR}(A \to C) \geq 1. \end{aligned} \tag{4.11}$$

4.4.3 Problemas multi-objetivo

Problema 1: Soporte, confianza y ascenso

La minería clásica de reglas de asociación busca obtener reglas que tengan el mayor soporte, confianza y ascenso posible. Sin embargo, la relación entre las medidas es no-monótona; es decir, un incremento o decremento sostenido en una no garantiza un comportamiento en la misma dirección en las otras dos.

La definición formal del problema de optimización de esta consulta se describe en la ecuación 4.12:

$$\begin{aligned} &\text{maximizar } supp(A \to C), \\ &\qquad conf(A \to C), \\ &\qquad lift(A \to C) \\ &\text{sujeto a } supp(A \to C) > 0, \\ &\qquad AR(A \to C) > 0, \\ &\qquad CI_{OR}^{inf}(A \to C) \geq 1. \end{aligned} \tag{4.12}$$

Problema 2: Minería de bicondicionales a través del efecto causal

Una expresión bicondicional ($A \Leftrightarrow C$), desde una perspectiva lógica, puede interpretarse como "A si y sólo si C" o "A es una condición necesaria y suficiente para C". Su valor de verdad equivale al de la expresión $(A \to C) \wedge (C \to A)$.

La condición de suficiencia recae sobre la regla de asociación $A \to C$, interpretada como "A es una condición suficiente para C". Se considera que la condición de suficiencia se satisface si el efecto causal de $A \to C$ es considerable. Por otro lado, para satisfacer la condición de necesidad de la expresión bicondicional, es necesario que el efecto causal de

$C \rightarrow A$ sea considerable también.

El problema de optimización a resolver para esta consulta se describe formalmente en la ecuación 4.13:

$$\begin{aligned} \text{maximizar} \quad & AR(A \rightarrow C), \\ & AR(C \rightarrow A) \\ \text{sujeto a} \quad & supp(A \rightarrow C) > 0, \\ & AR(A \rightarrow C) > 0, \\ & AR(C \rightarrow A) > 0, \\ & CI^{inf}_{OR}(A \rightarrow C) \geq 1. \end{aligned} \tag{4.13}$$

Los efecto causales de $A \rightarrow C$ y $C \rightarrow A$ se encuentran en oposición por la falacia de la afirmación del consecuente, la cual señala que si A causa C no es necesariamente cierto que C implique A, razón por la cual es necesario evaluar la regla tanto como su recíproca.

Problema 3: Susceptibilidad e impacto en la población

En este escenario se busca encontrar las reglas que maximicen la *susceptibilidad*, medida que cuantifica la capacidad del antecedente para producir el consecuente; y el *impacto en la población* o *fracción atribuible en la población*, medida que indica la proporción de observaciones del consecuente que fueron causadas por el antecedente.

El problema de optimización para esta consulta se describe formalmente en la ecuación 4.14.

$$
\begin{aligned}
&\text{maximizar } PS(A \to C),\\
&AF_p(A \to C)\\
&\text{sujeto a } \ supp(A \to C) > 0,\\
&AR(A \to C) > 0,\\
&CI^{inf}_{OR}(A \to C) \geq 1.
\end{aligned}
\tag{4.14}
$$

4.5 Fase de Evolución

En la fase de evolución, se utiliza un algoritmo evolutivo en el cual se utilizan estrategias de recombinación y mutación para generar reglas de asociación que serán evaluadas para satisfacer las consulta indicada. La Figura 4.6 incluye el diagrama de flujo que describe esta fase.

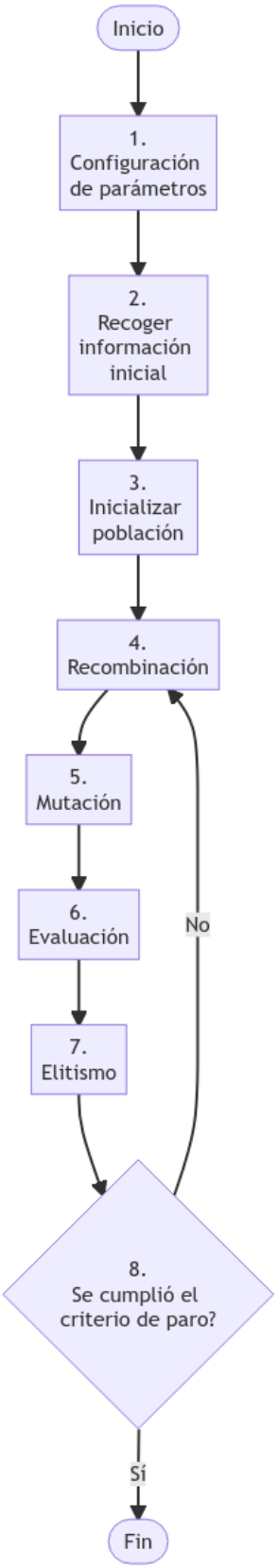

Figura 4.6: Diagrama de flujo de la fase de evolución

4.5.1 Descripción general

El algoritmo evolutivo consta de ocho etapas:

1. **Configuración de parámetros**. Se determinan los parámetros como el tamaño de población, la condición de paro, los grupos a relacionar, el problema de optimización a resolver, restricciones al espacio de variables de decisión y restricciones al espacio de los objetivos.
2. **Análisis del conjunto de datos**. Se realiza un análisis del conjunto de datos para determinar el dominio de cada atributo, paso necesario para generar selectores válidos.
3. **Inicializar población**. Se crea un conjunto inicial de reglas que fueron generadas en forma aleatoria.
4. **Recombinación**. Las reglas se asocian en pares, recombinando cada par a través de cuatro operadores: recombinación por intercambio, recombinación por unión, recombinación por intersección y recombinación por diferencia simétrica.
5. **Mutación**. Se introduce variabilidad a las reglas con tres operadores de mutación: extensión, contracción y reemplazo.
6. **Evaluación**. Se evalúan las reglas de asociación descendientes generadas a partir de los operadores de recombinación y mutación.
7. **Elitismo**. Se seleccionan las reglas más aptas para sobrevivir a la siguiente generación a través de dos operadores de elitismo: ordenamiento no dominado y preservación de diversidad.
8. **Comprobación del criterio de paro**. Terminado el paso de elitismo, se comprueba si se ha alcanzado el criterio de paro. Si se cumple el criterio de paro, el algoritmo termina; de otro modo, inicia una nueva iteración de recombinación, mutación, evaluación y elitismo. El criterio de paro difiere entre problemas mono-objetivo y multi-objetivo.

4.5.2 Representación seleccionada

Para ser manipulada por medio de un algoritmo evolutivo, una regla de asociación debe transformarse de su representación como expresión lógica a un árbol sintáctico.

En la Figura 4.7 se muestra el árbol sintáctico correspondiente a la regla de asociación `[Neumonía = Sí][Hipertensión = Sí] →[Hospitalización = Sí][Defunción = Sí]`.

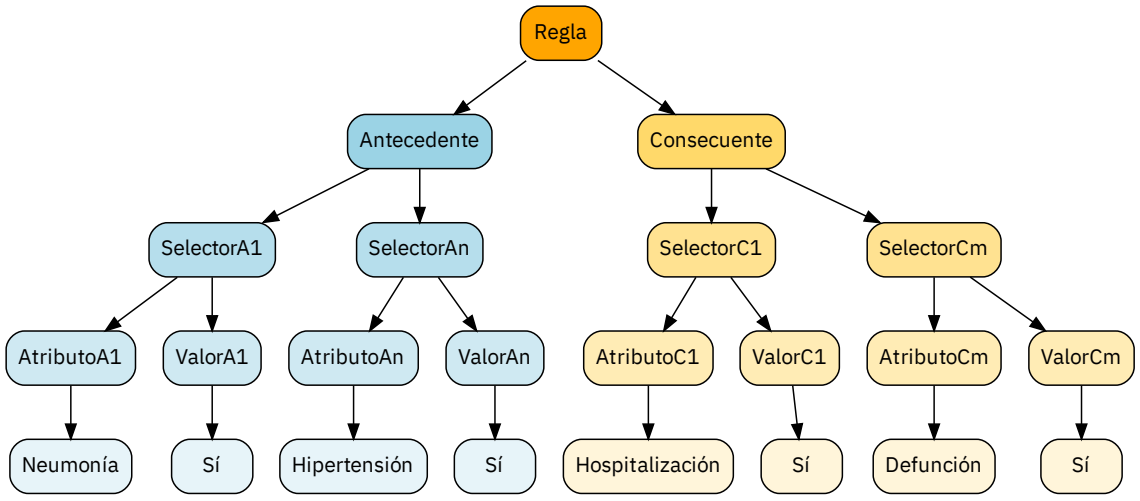

Figura 4.7: Árbol sintáctico para [Neumonía = Sí][Hipertensión = Sí]→[Hospitalización = Sí][Defunción = Sí]

En la raíz se encuentra el nodo `Regla`, compuesto de un `Antecedente` y un `Consecuente`; cada uno con uno o más nodos de tipo `Selector`. Bajando un nivel, cada nodo tipo `Selector` se compone de un nodo de tipo `Atributo`, que contiene el nombre de la columna, y un nodo `Valor`, con uno de los valores que puede tomar el atributo. Los nodos de tipo `Antecedente` y `Consecuente` crecen a lo ancho en lugar de crecer en profundidad, ya que no existe una precedencia entre los selectores al estar todos en conjunción.

En la implementación en Python, una regla de asociación se representa mediante tuplas anidadas. Por ejemplo, la regla de asociación `[Neumonía = Sí][Hipertensión = Sí] →[Hospitalización = Sí][Defunción = Sí]` representada mediante tuplas anidadas

toma la siguiente forma:

```
regla_asociacion = (
    (("NEUMONÍA", "Sí"), ("HIPERTENSIÓN", "Sí")), # antecedente
    (("HOSPITALIZACIÓN", "Sí"), ("DEFUNCIÓN", "Sí")) #consecuente
)
```

Para los atributos numéricos, el valor que toma el selector corresponde al rango determinado por la discretización en cuantiles:

```
regla_asociacion = (
    (("EDAD", "[24, 32]"), ("SEXO", "MUJER")), # antecedente
    (("INTERVALO_SINTOMAS_ENFERMEDAD", "[0, 3]"), ("DEFUNCIÓN", "No")) #consecuente
)
```

4.5.3 Configuración de parámetros

En el paso de configuración de parámetros se definen las condiciones de operación del algoritmo: la descripción del experimento, que incluye los grupos a relacionar, el problema de optimización a resolver y sus restricciones, tanto en el espacio de variables de decisión como restricciones sobre el espacio de las funciones objetivo; así como el criterio de paro y el tamaño de población.

Un grupo debe ser asignado al antecedente, al cual se le denomina como G_A y otro al consecuente, conocido como G_C. Cada parte de la regla, antecedente y consecuente, únicamente podrá tener selectores provenientes del grupo respectivo que le fue asignado en la configuración de parámetros.

Un problema de optimización indica la función o funciones a maximizar o minimizar, así como restricciones. El algoritmo admite la definición de dos tipos de restricciones:

restricciones sobre el espacio de variables de decisión y restricciones sobre el espacio de las funciones objetivo.

Un tipo de restricción sobre el espacio de variables de decisión regula los valores que pueden tomar los selectores permitiendo a los usuarios definir expresamente los valores que pueden tomar ciertos atributos, ya sea a través de listas en las que se enumeran los valores permitidos o excluidos para cierto conjunto de atributos.

Un segundo tipo de restricción a definir sobre el espacio de variables de decisión es la cardinalidad máxima tanto del antecedente como del consecuente; por omisión, cada parte de la regla de asociación puede incluir todos los selectores del grupo asignado. Cabe destacar que las asignaciones de grupos también son restricciones sobre el espacio de decisión.

Las restricciones definen los valores permitidos que pueden tomar las medidas de evaluación de las reglas de asociación durante el proceso evolutivo. Como se señaló en la sección 4.4.1, por omisión siempre se aplican tres restricciones: soporte mayor a cero, efecto absoluto positivo y significancia estadística. Sin embargo, el algoritmo admite la definición de más restricciones por los usuarios en caso de que se desee agregarlas.

El tamaño de la población se calcula a partir de los grupos a relacionar. En primera instancia es necesario definir un tamaño mínimo de población, cuyo valor por omisión es de 50. Para determinar del tamaño de población, se calcula para la asignación de grupos dada el total de selectores válidos, y se aumenta dicho número en un 50%. Se conserva como tamaño de población el mayor entre el tamaño mínimo y el tamaño calculado.

El criterio de paro está dado por un número máximo de generaciones sin variación en el indicador de paro después de que el algoritmo se haya ejecutado por 50 generaciones, siendo el indicador de paro de una generación es la aptitud máxima. Se recogen los últimos 30 valores del indicador de paro, para calcular después la desviación estándar. Si la desviación estándar es de cero, señalando una falta de variación en el indicador de paro, el algoritmo se detiene.

En la Tabla 4.2 se muestran los valores por omisión que tienen algunos parámetros del algoritmo genético.

Tabla 4.2: Parámetros y valores por omisión del algoritmo genético

Parámetro	**Valor por omisión**
Tamaño mínimo de población (P)	50
Máximo de generaciones sin variación	30
Mínimo de generaciones	50
Cardinalidad máxima del antecedente	Total de atributos en G_A
Cardinalidad máxima del consecuente	Total de atributos en G_C

4.6 Análisis del conjunto de datos

En este paso se realiza un análisis del conjunto de datos para determinar el dominio de cada atributo; es decir, los valores posibles que puede tomar un atributo. Éstos son los símbolos que participarán como valores de selector en las reglas de asociación.

Una vez obtenido el dominio de cada atributo, se aplican las restricciones indicadas por las listas de valores permitidos y excluidos, definidas en el paso de configuración de parámetros.

4.7 Inicialización de la población

La inicialización de la población crea la primer generación de reglas que formarán parte del proceso evolutivo.

Se crean $3P$ reglas aleatorias de cardinalidad (1, 1); es decir, reglas con un selector en el antecedente y un selector en el consecuente, siendo P el tamaño de población definido en

la configuración de parámetros. Se crean $3P$ reglas en lugar de P para tener un excedente de reglas válidas y así intentar garantizar tener un tamaño suficiente. Una vez creadas las reglas se evalúan, descartando primero aquellas reglas inválidas por no satisfacer las restricciones para después conservar las P mejores reglas y así conformar la población inicial.

En dado caso que no existieran suficientes reglas de cardinalidad (1, 1), se van generando reglas aleatorias una por una, agregándolas a la población inicial sólo si son válidas. Una vez alcanzado el tamaño de población deseado se tiene la primera generación de reglas, dando por concluido el proceso.

Si después de 5,000 intentos aún no fuera posible generar una población inicial de reglas válidas con el tamaño deseado, el flujo de minería se detiene prematuramente.

En primer lugar, el algoritmo inicia con reglas de cardinalidad (1, 1) debido a la propiedad planteada por Li [11] que enuncia que si una regla es estadísticamente significativa, entonces todas sus versiones más específicas también lo serán. El algoritmo de minería puede ser visto entonces como un proceso iterativo de generalización y especialización de reglas de asociación.

4.7.1 Recombinación

El paso de recombinación genera nuevas reglas a partir de la combinación de reglas existentes con el propósito de generar descendientes que compartan las mejores características de sus padres.

Los operadores de recombinación empleados en esta propuesta requieren una selección de pares de reglas (R1, R2), los cuales se generan de forma aleatoria. A cada par de reglas se le aplicarán los cuatro operadores de recombinación propuestos: recombinación por intercambio, recombinación por unión, recombinación por intersección y recombinación por diferencia simétrica.

Para ejemplificar las reglas resultantes de cada operador de recombinación se utilizarán como la regla R1 `[Neumonía = Sí][Hipertensión = Sí] → [Defunción = Sí][UCI = Sí][Intubado = Sí]` y como la regla R2 `[Diabetes = Sí][Hipertensión = Sí] → [Defunción = No][UCI = Sí]`.

Recombinación por intercambio. En la recombinación por intercambio, para cada par de reglas se producen dos reglas al intercambiar sus consecuentes. Los descendientes de la recombinación por intercambio son entonces:

`[Neumonía = Sí][Hipertensión = Sí] → [Defunción = No][UCI = Sí]` y

`[Diabetes = Sí][Hipertensión = Sí] → [Defunción = Sí][UCI = Sí][Intubado = Sí]`.

Recombinación por unión. La recombinación por unión genera una regla descendiente que acumula los atributos de sus padres. Si existe más de un selector para un atributo, se elige aleatoriamente uno de los dos selectores para preservar.

En la recombinación por unión de R1 y R2 colisionan los atributos `Hipertensión` y `Defunción`. Para `Hipertensión` el selector elegido es invariante, mientras que para `Defunción` se eligió aleatoriamente el selector con el valor de `No`. La regla resultante es entonces `[Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí] → [Defunción = No][UCI = Sí][Intubado = Sí]`.

Recombinación por intersección. La recombinación por intersección genera una regla descendiente a partir de los atributos que comparten las reglas progenitoras. Si para los atributos de intersección existen selectores con valores diferentes en las reglas, se elige un selector aleatoriamente.

En la recombinación por intersección de R1 y R2 colisionan los atributos de `Hipertensión` y `Defunción`. Así como en la recombinación por unión, el valor que toma el selector de `Hipertensión` es `Sí`, mientras que para el selector de defunción queda seleccionado en esta ocasión el valor de `No`. La regla descendiente es entonces `[Hipertensión =`

Sí] → [Defunción = Sí][UCI = Sí].

Recombinación por diferencia simétrica. La recombinación por diferencia simétrica genera una regla descendiente a partir de los atributos que están exactamente en una de las reglas progenitoras.

Es importante reiterar que la diferencia simétrica, así como la unión e intersección, se determinan a partir de los atributos y no de los selectores. Es por eso que en la recombinación por diferencia simétrica de R1 y R2 no se considera que exista una colisión, razón por la cual los selectores de Defunción quedan excluídos de la regla descendiente. El resultado de este operador de recombinación es entonces [Neumonía = Sí][Diabetes = Sí] → [Intubado = Sí].

4.7.2 Mutación

El paso de mutación busca introducir variabilidad a las reglas de asociación. Para cumplir con este objetivo se diseñaron tres operadores: mutación por extensión, mutación por contracción y mutación por reemplazo. Los operadores de mutación se aplican a las reglas resultantes de los operadores de recombinación.

Mutación por extensión. En la mutación por extensión, la regla mutada se extiende con un selector aleatorio cuyo atributo aún no esté presente en la regla de asociación, generando una regla más específica que la regla original.

La mutación por extensión de R1 extiende a la regla con el selector [EPOC = Sí], resultando entonces en la regla [EPOC = Sí][Neumonía = Sí][Hipertensión = Sí] → [Defunción = Sí][UCI = Sí][Intubado = Sí]

Mutación por contracción. En la mutación por contracción, se elimina aleatoriamente un selector de la regla, generando una regla más general que la regla original. Para evitar generar reglas con antecedentes o consecuentes vacíos, las reglas de cardinalidad (1,

1) quedan exentas de este operador; mientras que para las reglas de cardinalidad $(m, 1)$ y $(1, n)$ donde m y n son mayores a uno, el selector a eliminar se elige entre las partes de cardinalidad mayor a uno.

Como ejemplo, en la mutación por contracción de R1 se elige eliminar al selector `[UCI = Sí]`, produciendo entonces la regla descendiente `[Neumonía = Sí][Hipertensión = Sí] → [Defunción = Sí][Intubado = Sí]`.

Mutación por reemplazo En la mutación por reemplazo, se elige aleatoriamente un selector de la regla, reemplazándolo por un selector con el mismo atributo, pero de diferente valor.

Como ejemplo, en la mutación por reemplazo de R2, queda elegido el selector `[Defunción = Sí]`. El dominio del atributo `Hipertensión` es {`Sí`, `No`, `Se desconoce`}. Se selecciona un valor aleatorio entre el dominio excluyendo al valor `Sí`, quedando elegido el valor `No`. El resultado de la mutación por extensión de R2 pasa a ser `[Diabetes = Sí][Hipertensión = No] → [Defunción = Sí][UCI = Sí][Intubado = Sí]`

4.7.3 Evaluación

En la fase de evaluación se obtienen los valores de las medidas de evaluación descritas en el marco teórico y conceptual para cada regla generada por los operadores de recombinación y mutación. La secuencia de operaciones realizadas para evaluar una regla de asociación se describe en la Figura 4.8.

En la fase de evaluación recae el mayor uso de recursos de cómputo, exigiendo el uso de técnicas de programación como la memoización; así como de artefactos tales como una base de datos en memoria para persistir las evaluaciones y reutilizarlas en ejecuciones posteriores, y una base de datos columnar, capaz de responder ágilmente a las consultas de agregación necesarias para evaluar una regla de asociación.

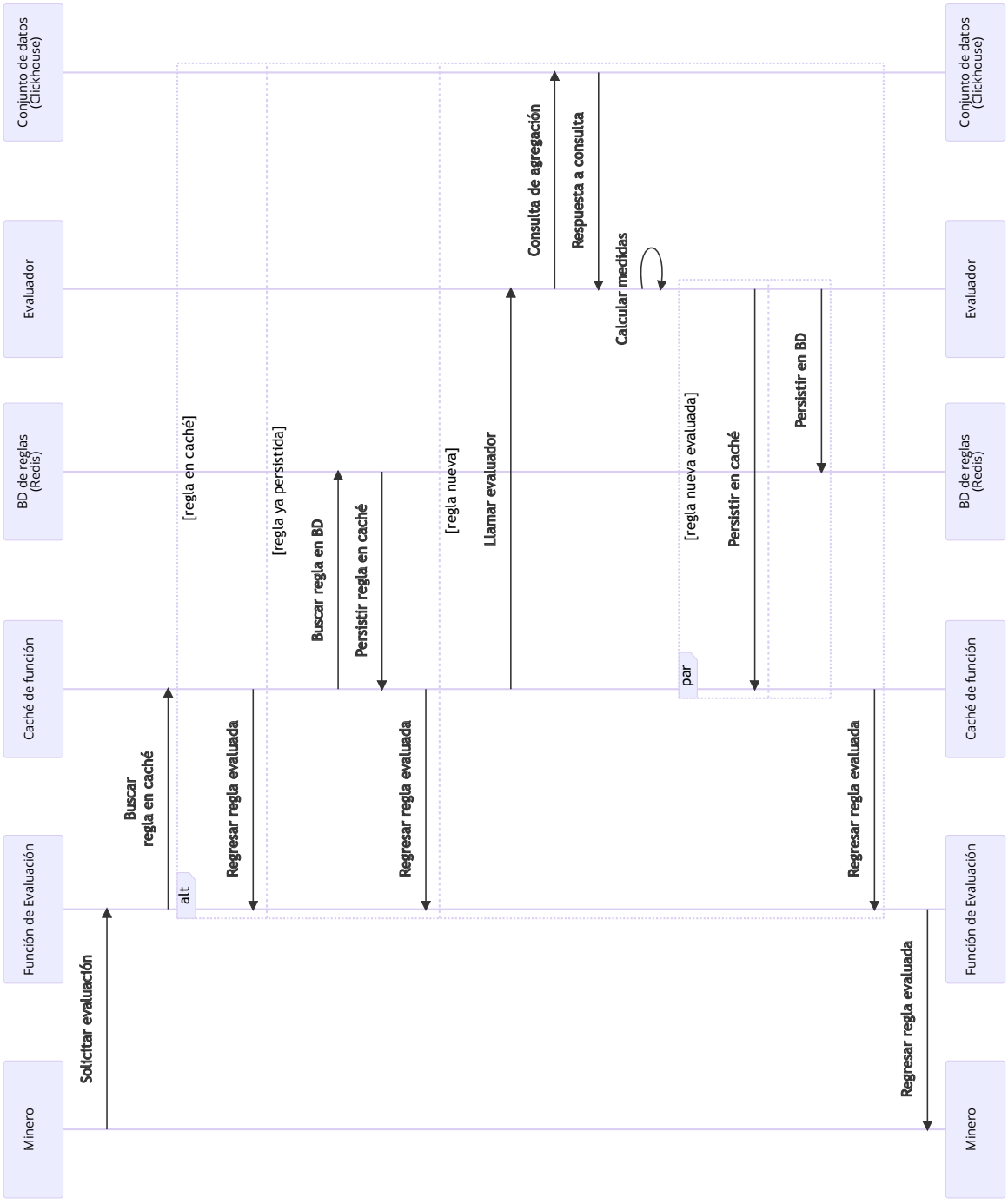

Figura 4.8: Diagrama de secuencia para evaluación

En el diagrama de secuencia se identifican tres alternativas de flujo al momento de evaluar una regla de asociación:

1. **Regla en caché.** Se activa este flujo si la regla solicitada ya ha sido evaluada previamente durante esta ejecución. La caché de la función de evaluación recupera el resultado y lo regresa.
2. **Regla persistida previamente.** Se activa este flujo si la regla que se solicita evaluar no ha sido evaluada previamente en esta ejecución, pero sí ha sido evaluada en ejecuciones anteriores del algoritmo de minería. Se recupera el resultado de la base de datos de reglas, se agrega a la caché de la función para finalmente regresar el resultado.
3. **Regla nueva.** Se activa este flujo si la regla no está registrada ni por la función de evaluación ni por la base de datos de reglas. En este paso se va directamente a la fuente de datos haciendo una consulta de agregación para recoger la información necesaria para hacer los cálculos de evaluación. Una vez terminados, se despachan paralelamente peticiones para persistir la evaluación de la regla tanto en la base de datos de reglas como en la caché de la función de evaluación.

4.7.4 Elitismo

En la etapa de elitismo se seleccionan las reglas más aptas para conformar la siguiente generación del proceso evolutivo. En una nueva iteración del proceso evolutivo se les aplicarán los pasos de recombinación, mutación, evaluación y elitismo hasta alcanzar el criterio de paro.

En primera instancia, se eliminan aquellas reglas que no cumplan con las restricciones de caja definidas en la configuración de parámetros. Ya que se han excluido las reglas inválidas, se aplican dos operadores para obtener las reglas élite: primero, se ordenan las reglas de acuerdo a su aptitud; en el caso multi-objetivo se aplica un ordenamiento no dominado para encontrar las reglas más aptas en el espacio de los objetivos para después aplicar un operador de diversidad, el cual preserva la variedad en el espacio de variables de decisión al

conservar las mejores reglas en las que participa cada selector.

Ordenamiento no-dominado. Iterativamente se va aplicando un algoritmo de ordenamiento no dominado, a través del cual se identifican a las reglas que ofrecen el mejor balance (*tradeoff*) en el espacio de los objetivos. Las reglas que se extraen en la primera iteración del ordenamiento no dominado conforman el conjunto de Pareto. El proceso se repite hasta que todas las reglas tienen asignado un nivel en la jerarquía de no dominancia.

Operador de diversidad. Además de conservar las mejores reglas de asociación en el espacio de los objetivos, es importante también que las reglas resultantes de la minería sean diversas en el espacio de variables de decisión, cubriendo la mayor cantidad de selectores diferentes posibles.

Para preservar dicha característica en la población se busca en la población actual, por cada selector, la regla de asociación en el nivel más alto del ordenamiento no dominado. Si llegara a existir más de una regla en el mismo nivel, se selecciona la que tenga mayor soporte. Este operador ayuda a conservar la diversidad de selectores que componen las reglas, garantizando que haya al menos una regla de cada selector.

Ya habiendo marcado el nivel en la jerarquía de no dominancia y su selección por el operador de diversidad, la población se reordena considerando los siguientes aspectos:

1. **Conjunto de Pareto**
2. **Diversidad**.
3. **Reglas válidas**.

La población de reglas ya evaluada se trunca, para seleccionar las P mejores de acuerdo al ordenamiento descrito previamente.

4.7.5 Comprobación del criterio de paro

Terminada la fase de elitismo, se comprueba si se ha alcanzado el criterio de paro. Si ya se ha alcanzado el número de generaciones definido en los parámetros de inicio, el algoritmo se detiene. De otro modo, se repiten los pasos de recombinación, mutación, evaluación y elitismo en una nueva iteración.

4.8 Fase de filtrado

El resultado del algoritmo de minería son las reglas que conforman la población final al momento en el que se alcanza el criterio de paro. Sin embargo, es posible que sea necesario aplicar un filtrado posterior a la población para destacar las reglas más importantes en caso de que el tamaño de población sea relativamente grande.

En el filtrado de la población final se destacan dos tipos de reglas: las que conforman el conjunto de Pareto, ofreciendo el mejor balance entre las medidas de evaluación que componen la consulta; y las reglas seleccionadas a través del operador de diversidad, revelando las reglas de mayor aptitud en las que participa cada selector. Cada tipo de regla se obtiene aplicando los operadores de ordenamiento no dominado y de diversidad descritos en el paso de elitismo de la fase de evolución.

Es entonces que se da por concluido el flujo de minería de reglas de asociación.

Capítulo 5

Resultados experimentales

En el presente capítulo se describe el diseño experimental y los resultados obtenidos, tanto del algoritmo de descubrimiento de grupos de atributos como del algoritmo de minería de reglas de asociación aplicado a problemas mono-objetivo y multiobjetivo.

En la sección 5.1 se describe el conjunto de datos sobre el cual se realizó la minería así como las transformaciones realizadas para obtenerlo. Posteriormente, se reportan en la sección 5.2 los resultados de cada una de las fases del algoritmo de descubrimiento de grupos de atributos.

La sección 5.3 describe el diseño experimental utilizado para probar el algoritmo de minería de reglas de asociación, dedicando las secciones 5.4 y 5.5 para describir los resultados de la experimentación para la clase de problemas mono-objetivo y la interpretación de las reglas obtenidas; mientras que las secciones 5.6 y 5.7 para los resultados e interpretación, respectivamente, de los experimentos multiobjetivo.

5.1 Descripción del conjunto de datos

El conjunto de datos a utilizar se titula "Información relacionada con COVID-19 en México", recuperado el 1° de Abril de 2022 del Portal de Datos Abiertos del gobierno de México [37]. El conjunto de datos está compuesto de 15,578,792 registros, cada uno con 37 atributos. Entre los atributos, 33 son categóricos y 4 numéricos. El periodo que comprenden los datos va del 1° de enero de 2020 al 1° de abril de 2022.

A continuación se reportan diferentes distribuciones de edad y sexo, con el propósito de caracterizar el conjunto de datos, así como algunas sub-poblaciones de interés, tales como los casos confirmados de COVID-19, casos que requirieron de hospitalización y defunciones.

En todo el conjunto de datos, el cual tiene 15.5 millones de registros, la distribución de la edad tiene una media de 38.55 años, la mediana es de 37 años y desviación estándar de 17.03. El número de mujeres es de 8.3 millones (53 %), mientras que 7.2 millones (47 %) son hombres.

Con respecto al contagio de COVID-19, 5.34 millones (34.3 %) de las observaciones son de pacientes para los cuales se confirmó que padecían COVID-19 por medio de una prueba. La distribución de edad para esta población tiene una media de 40.3 y una desviación estándar de 16.8, con una mediana de 38; siendo 2.78 millones (52%) de casos positivos en mujeres y 2.56 millones (48%) en hombres.

El conjunto de datos indica que 1.2 millones (7.7%) de los pacientes fueron hospitalizados. La distribución de la edad en los pacientes hospitalizados tiene una media de 52.6 años, con desviación estándar de 22 y mediana de 56. La cantidad de hombres hospitalizados fueron 668,979 (55.7%) son hombres y 532,307 (44.3%) son mujeres.

Se reportan 415,855 decesos en el conjuntos de datos, los cuales comprenden 2.66% de los registros. Para esta sub-población, la distribución de edad tiene una media de 62.9 años, desviación estándar de 15.8 y mediana de 64; donde 254,160 (61.1%) son hombres y

161,695 (38.9%) son mujeres.

Como se menciona en la metodología, en la sección 4.1, los cuatro atributos numéricos fueron sometidos a una discretización basada en quintiles, esto con el propósito de que cada intervalo tenga un soporte de aproximadamente el 20%.

La discretización del atributo `Edad` se muestra en la Tabla 5.1, mientras que la discretización de los atributos de fecha; es decir, `Fecha de ingreso`, `Fecha de síntomas` y `Fecha de defunción`, se muestran en la Tabla 5.2. Las semanas comprendidas por cada intervalo varían debido a que se busca tener una distribución uniforme en el número de registros que comprende cada intervalo, el cual es uno de los efectos del método de discretización elegido.

Tabla 5.1: Discretización en quintiles del atributo `Edad`

Rango	Soporte	Soporte (%)
[0, 24]	3,196,583	20.51
[34 42]	3,263,396	20.94
[24, 32]	3,157,761	20.26
[43, 53]	2,986,363	19.16
[54, 266]	2,978,399	19.11

De acuerdo al número 05-22 del Informe integral de COVID-19 en México, publicado por la Secretaría de Salud del gobierno de México [38], se identifican cuatro olas de contagio en el periodo comprendido por los datos del caso de estudio. Las olas de contagio se presentan en la Tabla 5.3, reportando sus fechas de inicio y fin así como el total de registros en cada ola.

Tabla 5.2: Discretización de los atributos de fecha

Atributo	Rango	Inicio	Término	Semanas	Registros	Soporte (%)
Fecha de defunción	[0, 210]	2020-01-01	2020-07-30	30	83,939	0.54
	[211, 337]	2020-07-31	2020-12-04	18	82,622	0.53
	[338, 404]	2020-12-05	2021-02-08	9	83,709	0.54
	[405, 593]	2021-02-09	2021-08-16	27	82,704	0.53
	[594, 820]	2021-08-17	2022-04-01	32	82,881	0.53
Fecha de ingreso	[0, 342]	2020-01-01	2020-12-08	49	3,137,397	20.13
	[342, 455]	2020-12-09	2021-03-31	16	3,108,784	19.95
	[455, 590]	2021-04-01	2021-08-13	19	3,114,946	19.99
	[590, 726]	2021-08-14	2021-12-27	19	3,118,845	20.02
	[726, 820]	2021-12-28	2022-04-01	13	3,102,530	19.91
Fecha de síntoma	[0, 338]	2020-01-01	2020-12-04	48	3,107,094	20
	[338, 453]	2020-12-05	2021-03-29	16	3,129,137	20.08
	[454, 588]	2021-03-30	2021-08-11	19	3,134,419	20.11
	[589, 723]	2021-08-12	2021-12-24	19	3,095,423	19.86
	[724, 820]	2021-12-25	2022-04-01	13	3,106,429	19.94

Tabla 5.3: Descripción de las olas de contagio

Ola	Fecha de inicio	Fecha de fin	Registros
1	2020-02-16	2020-09-26	1,955,291
2	2020-09-27	2021-04-17	4,604,490
3	2021-06-06	2021-10-23	4,220,735
4	2021-12-19	2022-03-05	3,027,248

5.2 Algoritmo de descubrimiento de grupos de atributos

En la presente sección se describen los resultados obtenidos para cada una de las cinco fases del algoritmo de descubrimiento de grupos de atributos.

5.2.1 Fase de construcción del grafo de asociación

En la fase de construcción del grafo de asociación se aplicó una estrategia de agrupamiento que requiere la evaluación de un indicador conocido como la V de Cramér [39] para todos los pares posibles de los 37 atributos, resultando en 666 evaluaciones. Se creó un vértice por cada atributo del grafo de asociación, agregando una arista para cada par de nodos evaluados.

La distribución de las evaluaciones de la V de Cramér tuvo una media de 0.156, una desviación estándar de 0.21 y la mediana fue de 0.071. Un histograma de la distribución se muestra en la Figura 5.1. Observar la distribución de la V de Crámer señala que en este conjunto de datos predominan las asociaciones de bajo peso.

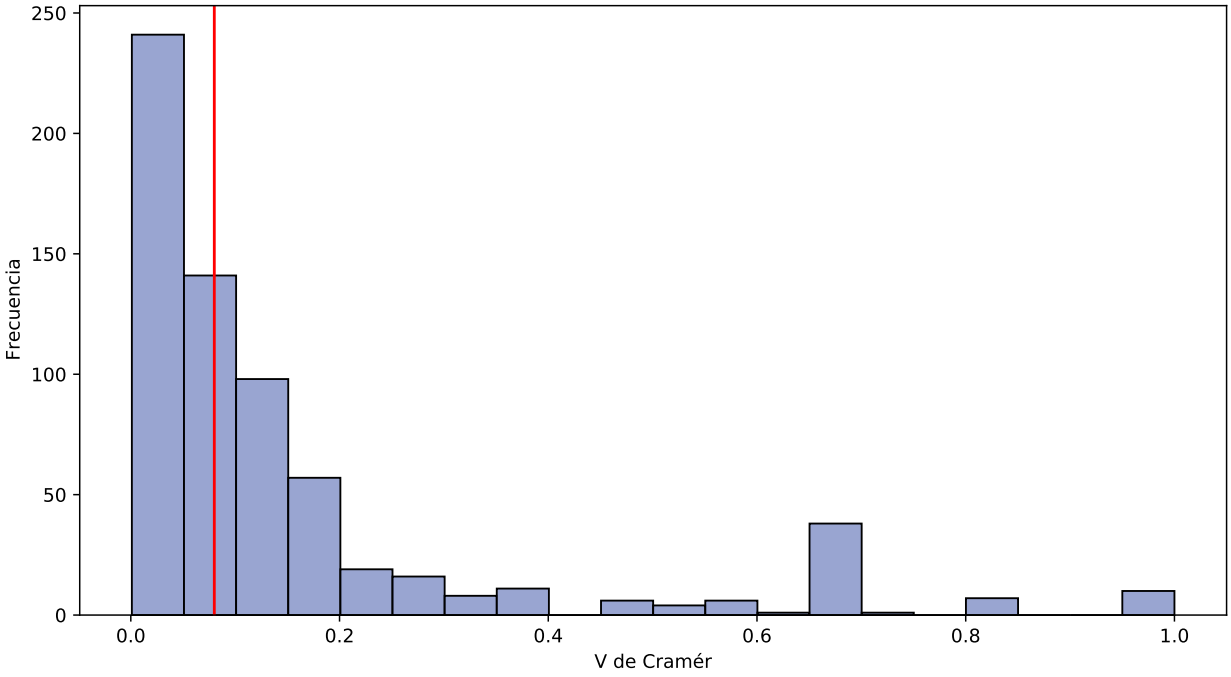

Figura 5.1: V de Cramér

5.2.2 Fase de depuración de aristas

Después se procedió con la depuración de aristas, eliminando primero las aristas sin significancia estadística y después depurando las aristas de bajo peso hasta encontrar una arista de corte o puente. Dicha arista determina el umbral de corte.

En primera instancia se eliminaron 2 aristas debido a que el valor de la V de Crámer no era estadísticamente significativo, ya que tenían un p-valor menor de 0.001. Posteriormente, en la depuración de aristas de bajo peso se eliminaron 348 aristas (52% de las aristas del grafo original), resultando en un grafo con 316 aristas. Las aristas de corte identificadas fueron `(SEXO, EMBARAZO)` y `(SEXO, TABAQUISMO)`, ambas con un valor de 0.08; fijando en dicho punto el umbral de corte, el cual se señala con una línea roja en la Figura 5.1. Los resultados de la depuración se muestran en la Tabla 5.4.

Tabla 5.4: Resultados de la fase de depuración de aristas.

Propiedad	Valor
Vértices	37
Aristas evaluadas	666
Aristas significativas	664 de 666 (99.7%)
Aristas significativas depuradas por bajo peso	348 de 664 (52.4%)
Total de aristas depuradas	350 de 664 (52.2%)
Aristas en el grafo final de asociación	316 de 666 (47.8%)

5.2.3 Fase de cubrimiento del grafo de asociación

Para obtener un cubrimiento en cliques del grafo de asociación, es necesario construir el grafo complemento del grafo de asociación para después buscar una coloración de sus vértices. Cada conjunto de vértices que comparte un color en el grafo complemento forma un clique en el grafo de asociación.

La coloración del grafo complemento mediante el algoritmo DSATUR [36] utilizó siete colores, particionando entonces el grafo de asociación en siete cliques. Se verificó de forma independiente el número cromático del grafo complemento, el cual fue de siete. Esto muestra que el algoritmo DSATUR utilizado para la coloración del grafo complemento está produciendo el número mínimo de cliques posible para el grafo de asociación indicado.

En la Tabla 5.6 se muestran los grupos resultantes de la fase de cubrimiento, indicando el identificador de cada grupo, su tamaño y los atributos que lo componen. El grafo resultante del cubrimiento se muestra en la Figura 5.2.

Tabla 5.5: Atributos que componen los escenarios utilizados en la experimentación.

ID	Atributos
Enfermedades	`Asma`, `Cardiovascular`, `COVID-19`, `Diabetes`, `EPOC`, `Hipertensión`, `Inmusupr.`, `Neumonía`, `Obesidad`, `Renal crónica`, `Tabaquismo` , `Otro caso`, `Otra complicación`
Edad y sexo	`Edad`, `Sexo`
Ubicación	`Entidad de atención médica`, `Atn. misma entidad`, `Sector`
Atención Médica	`Intubado`, `UCI`, `Defunción`, `Hospitalización`
G_0	`Asma`, `Inmunosupresores`, `Tabaquismo`, `Obesidad`, `Renal crónica`, `EPOC`, `Cardiovascular`, `Diabetes`, `Hipertensión`, `COVID-19`
G_1	`Edad`, `Fecha defunción`, `Intubado`, `UCI`, `Toma muestra antígeno`, `Toma muestra lab.`, `Defunción`, `Hospitalización`
G_3	`Origen`, `Resultado_antígeno`, `Fecha de ingreso`, `Fecha de síntomas`, `Resultado laboratorio`

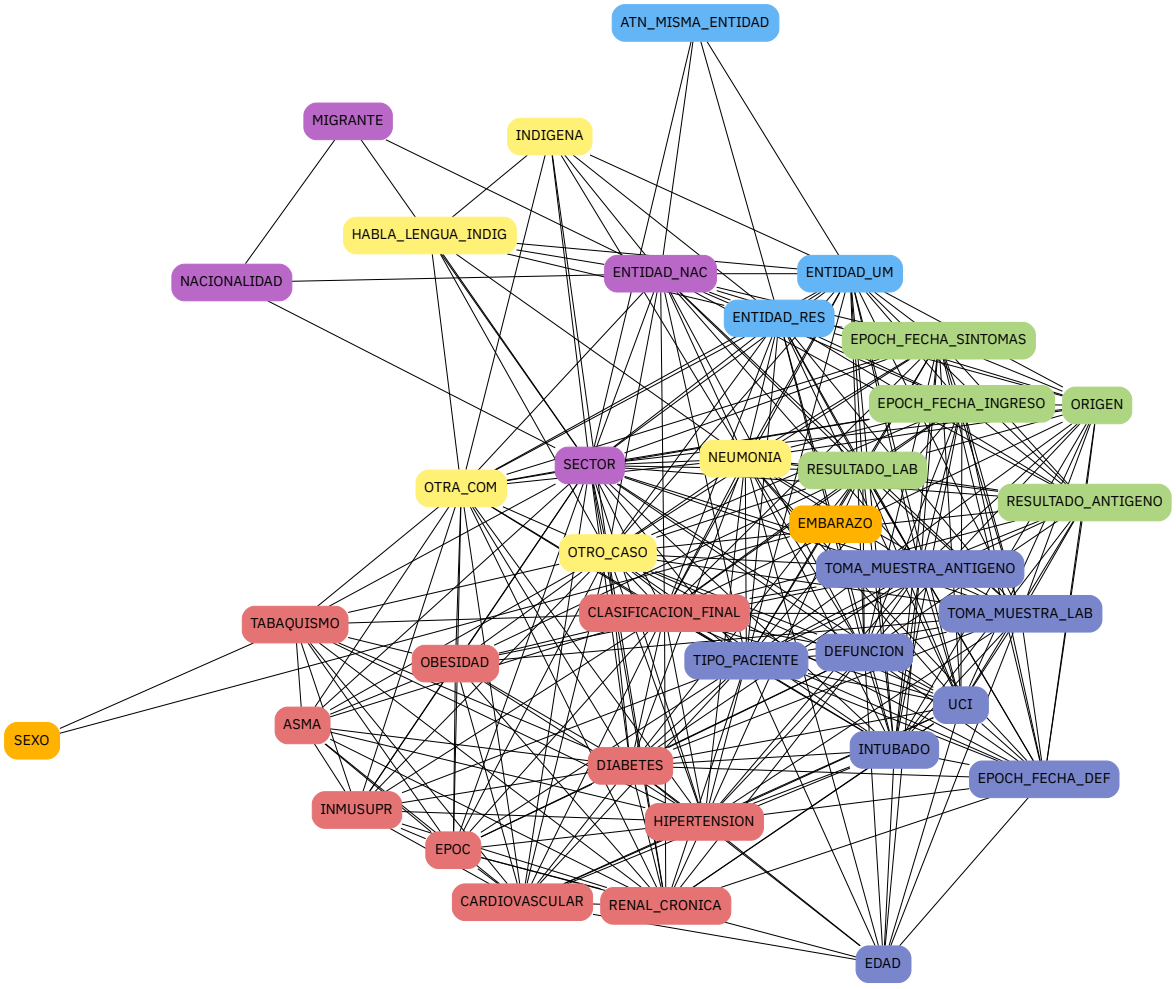

Figura 5.2: Grafo de asociación con aristas depuradas.

5.2.4 Fases de construcción del grafo de grupos e identificación de escenarios potenciales

Para realizar la identificación de escenarios potenciales se obtiene en primera instancia el árbol de máximo alcance del grafo de asociación, el cual puede ser visualizado en la Figura 5.3.

Figura 5.3: Árbol de máximo alcance en el grafo de asociación.

Un par de grupos (g_1, g_2) conforman un escenario potencial si existe en el árbol de máximo alcance del grafo de asociación al menos una arista que conecte un vértice de g_1 con un vértice de g_2. En la Figura 5.4 se conectan los escenarios potenciales identificados por el árbol de máximo alcance, resultando en veinte permutaciones de grupos de atributos. Mediante la identificación de escenarios potenciales se sugieren 20 de las 42 permutaciones posibles de grupos.

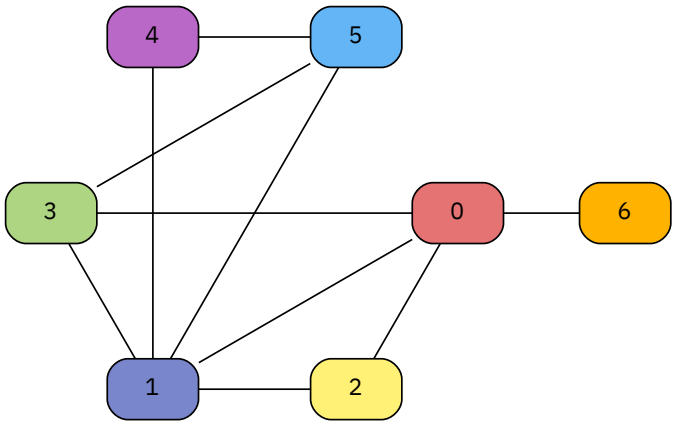

Figura 5.4: Grafo de escenarios potenciales

Gráficamente, se puede observar en la Figura 5.5 cada uno de los cliques que particionan al grafo de asociación. Las aristas del árbol de expansión que conectan atributos de diferentes grupos se muestran punteadas y en color rojo.

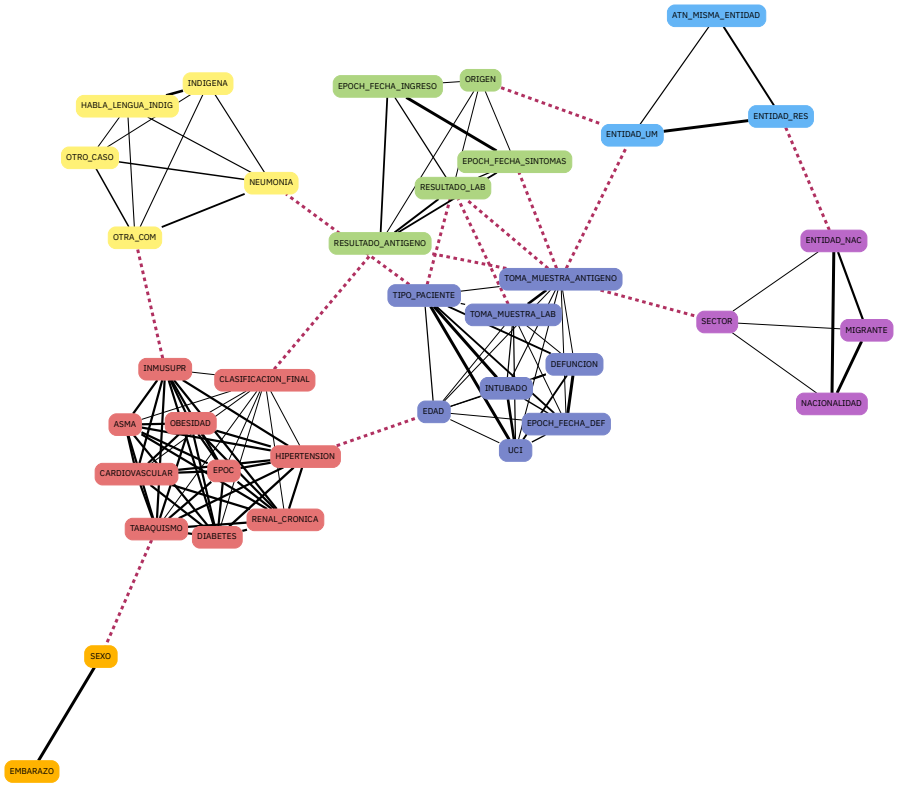

Figura 5.5: Árbol de máximo alcance y cliques del grafo de asociación.

5.3 Diseño de los experimentos

Se realizaron dos clases de experimentos para aplicar la técnica de minería de reglas de asociación: una clase de experimentos mono-objetivo y una clase multiobjetivo.

Retomando la terminología utilizada a través de este trabajo, un **escenario** es una tupla ordenada de grupos de atributos, en la que los atributos del primer grupo conforman los selectores de antecedente, y los atributos del segundo grupo conforman los selectores de consecuente; un **experimento** está descrito por un *problema de optimización* y un *escenario*; y una **ejecución** es el resultado del algoritmo evolutivo configurado de acuerdo a las indicaciones de un experimento dado.

Para componer los escenarios se seleccionaron siete grupos de atributos. Cuatro de

estos grupos fueron construidos de forma semántica, mientras que tres son producto del algoritmo de descubrimiento de grupos de atributos. Los atributos que componen cada grupo se muestran en la Tabla 5.6.

Tabla 5.6: Atributos que componen los escenarios utilizados en la experimentación.

ID	Atributos
Enfermedades	`Asma`, `Cardiovascular`, `COVID-19`, `Diabetes`, `EPOC`, `Hipertensión`, `Inmusupr.`, `Neumonía`, `Obesidad`, `Renal crónica`, `Tabaquismo` , `Otro caso`, `Otra complicación`
Edad y sexo	`Edad`, `Sexo`
Ubicación	`Entidad de atención médica`, `Atn. misma entidad`, `Sector`
Atención Médica	`Intubado`, `UCI`, `Defunción`, `Hospitalización`
G_0	`Asma`, `Inmunosupresores`, `Tabaquismo`, `Obesidad`, `Renal crónica`, `EPOC`, `Cardiovascular`, `Diabetes`, `Hipertensión`, `COVID-19`
G_1	`Edad`, `Fecha defunción`, `Intubado`, `UCI`, `Toma muestra antígeno`, `Toma muestra lab.`, `Defunción`, `Hospitalización`
G_3	`Origen`, `Resultado_antígeno`, `Fecha de ingreso`, `Fecha de síntomas`, `Resultado laboratorio`

En la Tabla 5.7 se muestra para cada grupo, el total de atributos, el total de selectores que pertenecen al grupo, así como el total de combinaciones válidas de selectores para cada grupo. El total de combinaciones se calcula multiplicando sucesivamente la cantidad de selectores de cada atributo más uno. Un atributo con m selectores tiene $m + 1$ opciones, ya que es necesario considerar la opción de vacío; es decir, que el atributo no esté presente en la combinación. Se multiplican todas las opciones para cada atributo, restando uno al producto final, excluyendo así la opción en la que todos los atributos están vacíos.

Tabla 5.7: Atributos que componen los escenarios

Grupo	N° atributos	N° selectores	Combinaciones
Enfermedades	13	43	134,217,727
Ubicación	3	66	3,266
Atención médica	4	12	224
Edad y sexo	2	7	17
G_0	10	34	2,097,151
G_1	8	27	85,049
G_3	5	20	2,591

En la Tabla 5.8 se muestran los seis escenarios planteados sobre los cuales se realizaron los experimentos. Tres escenarios, A, B y C, reflejan relaciones de interés para el caso de estudio, relacionando grupos de atributos construidos de forma semántica; mientras que los siguientes tres, D, E y F; relacionan grupos de atributos obtenidos por el algoritmo de descubrimiento de grupos de atributos, identificados en el grafo de grupos como escenarios potenciales. El tamaño del espacio de búsqueda está dado por el producto de las combinaciones para cada par de grupos en el espacio de búsqueda.

Tabla 5.8: Descripción de los escenarios.

Escenario	Antecedente	Consecuente	Espacio de búsqueda
A	Edad y sexo	Enfermedades	2,281,701,359
B	Enfermedades	Atención médica	30,064,770,848
C	Ubicación	Enfermedades	438,355,096,382
D	G_0	G_1	178,360,595,399
E	G_0	G_3	5,433,718,241
F	G_1	G_3	220,361,959

En las secciones 5.4 y 5.5 se describen los resultados de la clase de experimentos mono-objetivo; mientras que en las secciones 5.6 y 5.7 se describen los resultados de la clase de experimentos multiobjetivo.

5.4 Experimentación mono-objetivo

Para realizar la experimentación mono-objetivo se utilizaron los seis escenarios descritos en la Tabla 5.8; así como diez problemas de optimización mono-objetivo, resultando en 60 experimentos al ser un experimento el producto cartesiano de un escenario y un problema de optimización. Para cada experimento se realizaron 5 ejecuciones independientes de cada experimento, variando cada vez la semilla para generación de aleatorios, acumulando un total de 300 ejecuciones para toda la clase de experimentos. El flujo de experimentación se describe en la Figura 5.6.

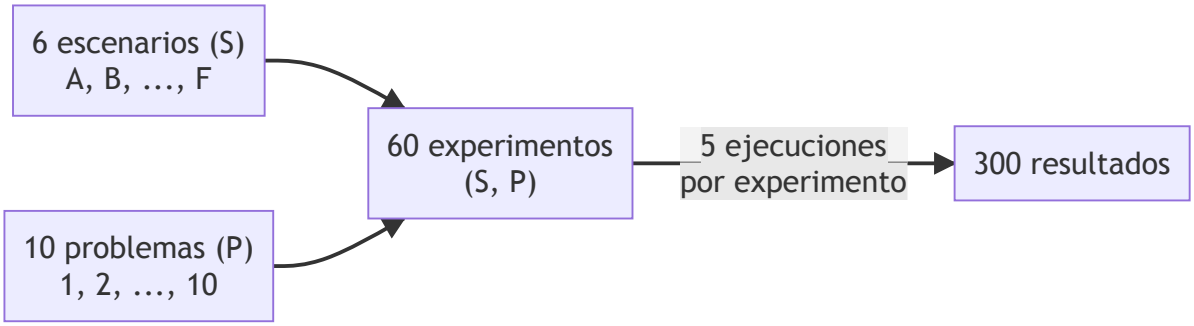

Figura 5.6: Descripción de la experimentación mono-objetivo

Los diez problemas de optimización mono-objetivo consideran funciones objetivo que corresponden a medidas de evaluación tanto clásicas como causales. Entre las medidas clásicas se seleccionó el soporte, confianza y ascenso; así como medidas de evaluación causales, como el efecto causal, la susceptibilidad e impacto en la población. También se incluyeron tres problemas de optimización en los que se combinan dos y tres medidas de evaluación, donde la combinación se realiza a través de la media geométrica. Los problemas de optimización seleccionados se describen en la Tabla 5.9 .

Tabla 5.9: Problemas mono-objetivo

Problema	Medida a maximizar	Ecuación
1	Soporte	4.2
2	Confianza	4.3
3	Ascenso	4.4
4	Soporte, confianza y ascenso (media geométrica)	4.5
5	Efecto causal	4.6
6	Efecto causal en la regla recíproca	4.7
7	Efectos causales (media geométrica)	4.8
8	Susceptibilidad	4.9
9	Impacto en la población	4.10
10	Susceptibilidad, impacto (media geométrica)	4.11

Para la ejecución de cada experimento se seleccionó la regla de asociación con la mayor aptitud. Posteriormente se calcularon las medidas de tendencia central para el conjunto de reglas de asociación seleccionadas. Se reporta la media (μ) y desviación estándar (σ) de las

cinco ejecuciones de cada experimento en la Tabla 5.10.

Tabla 5.10: Resultados de las funciones mono-objetivo (1-10). A-F son los grupos a relacionar, μ el promedio y σ la desviación estándar

Prob.	A		B		C		D		E		F	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
1	0.322	0	0.588	0	0.356	0.025	0.588	0	0.465	0	0.752	0
2	0.655	0	0.993	0	0.993	0.01	0.993	0	0.788	0.002	1	0
3	7.696	0	54.137	0	100.96	40.79	167.046	0	232.383	0	763.247	84.159
4	0.583	0	0.84	0	0.652	0	0.84	0	0.789	0	0.842	0
5	0.272	0	0.869	0	0.873	0.06	0.841	0	0.663	0	0.881	0.042
6	0.739	0	0.891	0	0.296	0.001	0.834	0.132	0.963	0	0.979	0
7	0.347	0	0.624	0	0.217	0	0.839	0	0.702	0	0.771	0
8	0.293	0	0.941	0	0.973	0	0.986	0	0.702	0	1	0
9	0.913	0	0.935	0	0.447	0	0.912	0	1	0	1	0
10	0.375	0	0.627	0	0.222	0	0.842	0	0.784	0	0.828	0

Con respecto a las reglas encontradas, 88 reglas diferentes fueron las de mayor aptitud a través de las 300 ejecuciones, indicando alta consistencia en la regla de mayor aptitud. Dicha consistencia está reflejada por el valor de cero en la la desviación estándar para 48 de los 60 experimentos realizados.

5.5 Interpretación de las reglas mono-objetivo

A continuación se discuten algunas de las reglas de asociación encontradas en la experimentación mono-objetivo. Debido a que todos los problemas de optimización considerados buscan maximizar la función objetivo, se seleccionaron las reglas de asociación con el mayor valor de aptitud para cada problema, independientemente del escenario relacionado con el experimento. Muchas de las reglas obtenidas en los problemas mono-objetivo destacan en el valor numérico que toman, llegando en muchos casos a los valores extremos para las funciones objetivo siendo optimizadas. Para cada regla se muestra el valor de la función objetivo así como el soporte de la regla.

La regla de mayor soporte es `[Defunción = No] → [Origen = Fuera de USMER]`, con un soporte del 75% (11.7 millones de registros). Esta regla señala que si un paciente no

falleció, recibió atención en una unidad de salud diferente a las Unidades de salud monitoras de enfermedades respiratorias (USMER).

La regla de mayor confianza fue `[Defunción = Sí][Fecha de Defunción = [2020-01-01, 2020-07-30]] → [Fecha Ingreso = [2020-01-01, 2020-12-08]]`, con confianza del 100% y soporte del 0.5% (83,939) registros. Esta regla señala que todos los decesos comprendidos en el primer intervalo de fechas de defunción corresponden al primer intervalo de las fechas de ingreso.

La regla de mayor ascenso fue `[Defunción = Sí][Edad = [54, inf]][Fecha defunción = [2020-12-05, 2021-02-08]][Tipo paciente = Hospitalizado][Toma muestra antígeno = Sí][Toma muestra lab. = Sí][UCI = Sí] → [Fecha síntomas = [2020-12-05, 20201-03-29]][Origen = Fuera de USMER][Resultado antígeno = Sí][Resultado laboratorio = No adecuado]`, con un ascenso de 800.8 y soporte de 27 registros.

La regla con la mayor media geométrica de soporte, confianza y ascenso es `[Hospitalización = No] → [Origen = Fuera de USMER]` con soporte del 73% (11,370,521 registros), confianza del 79% y ascenso de 1.0344, cuya media geométrica es de 0.842. Esta regla alcanza valores altos de confianza y soporte, pero ascenso muy bajo.

La regla de mayor efecto causal es `[Atención en la misma entidad = Sí][Entidad = Yucatán][Sector = Privado] → [Neumonía = Sí]`, con un valor de efecto causal de 90% y soporte de 1,005 registros. Esta regla indica que la probabilidad de observar neumonía en presencia del antecedente aumenta considerablemente. Debido al alto valor del efecto causal, es posible afirmar que el antecedente es una causa suficiente para observar el consecuente.

La regla de mayor efecto causal en su regla recíproca es `[Toma muestra antígeno = Sí][Toma muestra lab. = Sí] → [Resultado antígeno = Negativo][Resultado lab. = Negativo]`, con un valor de 98% y soporte del 3% (473,020 registros). El alto valor del

efecto causal en la regla recíproca señala que el antecedente es una causa necesaria para observar el consecuente. Este resultado es consistente con la interpretación semántica que se le da a la regla, señalando que para tener resultados negativos en las pruebas de COVID-19 es necesario que se tomen muestras para cada tipo de prueba.

La regla con el mayor valor en la media geométrica del efecto causal de la regla y su recíproca es la regla `[COVID-19 = Sospechoso] → [Defunción = No][Toma muestra antígeno = No][Toma muestra lab. = No]`, con una media geométrica de 0.84; efecto causal del 80% y efecto causal en la regla recíproca de 88%. El soporte de la regla es del 3%, cubriendo 465,592 registros. Esta regla identifica que ser un caso sospechoso de COVID-19 es una causa suficiente y necesaria para observar el consecuente. Es una causa suficiente debido al alto efecto causal, así como necesaria, por el alto efecto causal en la regla recíproca. La interpretación de esta regla es consistente con la descripción dada en el diccionario de datos, la cual señala que un caso sospechoso de COVID-19 es un caso para el cual no existe ni prueba de antígeno ni prueba de laboratorio para confirmar el contagio de COVID-19. Esta regla es también la de mayor media geométrica de susceptibilidad e impacto en la población, con susceptiblidad del 80% e impacto en la población del 89%.

La regla con mayor susceptibilidad es la regla `[Fecha de defunción = [2020-01-01, 2020-07-30]] → [Fecha de síntomas = [2020-01-01, 2020-12-04]]`, con susceptibilidad de 100% y soporte de 83,939. Esta regla indica que todos los casos en los que se presentó defunción en los primeros 210 días fueron de contagios en los primeros 338 días para los cuales se tienen datos. Otra regla con alta susceptibilidad es `[Entidad = Nayarit][Sector = Cruz Roja] → [COVID-19 = Confirmado]`, con valor del 97%. La interpretación de este valor es que la probabilidad de que se presente un caso confirmado de COVID-19 en presencia del antecedente es del 97%. La regla tiene un soporte de 56 registros.

La regla con mayor valor de impacto en la población (AF_p) fue `[Toma muestra de`

`antígeno = Sí] → [Resultado Antígeno = Negativo]`, con valor del 100% y soporte del 45% (6,993,425 registros). Esta regla indica que el 100% de los casos de resultados negativos en pruebas de antígeno requirieron de la toma de una prueba de este tipo, una dependencia necesaria. Otra regla de alto impacto en la población es `[Neumonía = Sí] → [Defunción = Sí][Intubado = Sí][Hospitalizado = Sí][UCI = Sí]`, con impacto en la población del 93% y soporte del 1.7% (26,057 registros). Este valor tan alto de impacto en la población señala que la neumonía tuvo un efecto causal para el 93% de los decesos que recibieron la mayor atención médica.

5.6 Experimentación multi-objetivo

Para realizar la experimentación multi-objetivo se utilizaron los seis escenarios descritos en la Tabla 5.8, resolviendo los tres problemas de optimización descritos en la Tabla 5.11, presentados originalmente en el Capítulo 4.

Tabla 5.11: Problemas multi-objetivo

Problema	**Medidas a maximizar**	**Ecuación**
1	Soporte, confianza, ascenso	4.12
2	Efectos causales de $A \Rightarrow C$ y $C \Rightarrow A$	4.13
3	Susceptibilidad, impacto en la población	4.14

Cada experimento se ejecutó diez veces, variando cada vez la semilla de generación de aleatorios, acumulando un total de 180 ejecuciones. El criterio de paro seleccionado fue detener la ejecución al llegar a 100 generaciones y que no haya mejora en las funciones objetivo. El flujo de experimentación se describe en la Figura 5.7.

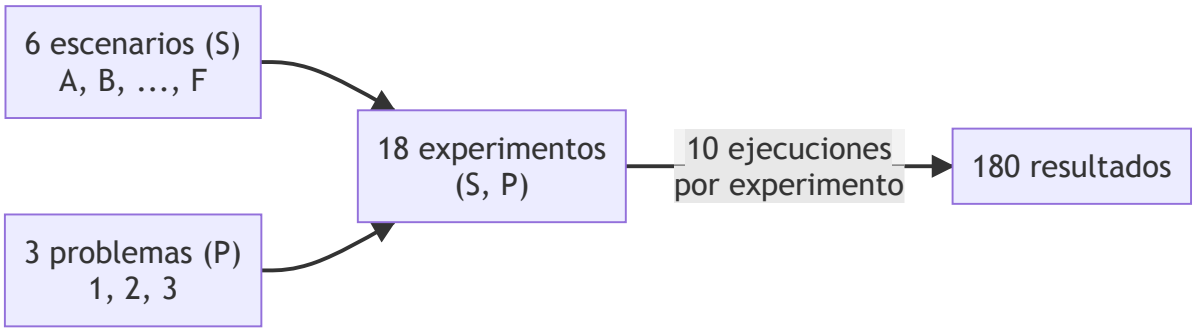

Figura 5.7: Descripción de la experimentación multi-objetivo

Por cada ejecución se calculó el tamaño del frente de Pareto en la población final. En la Tabla 5.12 se reportan la media (μ) y desviación estándar (σ) del tamaño del frente de Pareto para los dieciocho experimentos realizados. En 7 de los 18 experimentos, el algoritmo de minería de reglas de asociación llega en todas sus ejecuciones a frentes de Pareto con la misma longitud.

Tabla 5.12: Medidas de tendencia central de las longitudes del frente de Pareto

Escenario	Problema 1		Problema 2		Problema 3	
	μ	σ	μ	σ	μ	σ
A	27	0	9	0	11	0
B	39	0	20	0	21	0
C	16	0.707	8.8	1.643	11.2	1.095
D	38.8	1.304	31.2	1.789	29	0.707
E	32.8	0.837	2	0	3.8	0.447
F	34.2	0.837	11	0	6.8	0.447

Para describir las características de los frentes de Pareto, se recogieron por cada experimento las reglas con los valores más altos para cada una de las funciones objetivo del problema, así como la regla con el mayor valor en la media geométrica de las funciones objetivo. La media geométrica se reporta con el objetivo de comparar los resultados obtenidos en los experimentos mono-objetivo utilizando a la media geométrica como función de aptitud. Posteriormente se obtuvo la media (μ) y desviación estándar (σ) para dichos valores.

Los resultados de los diferentes experimentos para el problema 1 se reportan en la Tabla 5.13, los del problema 2 se reportan en la Tabla 5.14 y finalmente los del problema 3 en la Tabla 5.15.

Tabla 5.13: Máximos obtenidos para el problema 1. Los escenarios A-F son los grupos a relacionar, μ la media y σ la desviación estándar.

Escenario	Soporte		Confianza		Ascenso		Media geométrica	
	μ	σ	μ	σ	μ	σ	μ	σ
A	0.322	0	0.655	0	7.696	0	0.583	0
B	0.588	0	0.993	0	33.967	0	0.84	0
C	0.365	0.021	0.88	0.119	41.11	31.003	0.64	0.027
D	0.588	0	0.993	0	63.897	9.575	0.84	0
E	0.465	0	0.789	0	27.524	0.489	0.789	0
F	0.752	0	1	0	445.549	246.863	0.842	0

Tabla 5.14: Máximos obtenidos para el problema 2. Los escenarios A-F son los grupos a relacionar, μ la media y σ la desviación estándar.

Escenario	Efecto causal		Efecto causal (r.)		Media geométrica	
	μ	σ	μ	σ	μ	σ
A	0.272	0	0.739	0	0.347	0
B	0.869	0	0.891	0	0.624	0
C	0.823	0.094	0.245	0.019	0.182	0
D	0.683	0	0.597	0.002	0.347	0
E	0.663	0	0.825	0	0.702	0
F	0.806	0	0.979	0	0.771	0

Tabla 5.15: Máximos obtenidos para el problema 3. Los escenarios A-F son los grupos a relacionar, μ la media y σ la desviación estándar.

Escenario	Susceptibilidad		AF_p		Media geométrica	
	μ	σ	μ	σ	μ	σ
A	0.293	0	0.913	0	0.375	0
B	0.941	0	0.935	0	0.627	0
C	0.951	0.019	0.377	0.031	0.222	0
D	0.841	0	0.908	0.001	0.357	0
E	0.702	0	1	0	0.784	0
F	1	0	1	0	0.828	0

Como ejemplo de los frentes de Pareto encontrados, se reportan los frentes de Pareto para los tres problemas en el escenario B, el cual relaciona el grupo de atributos de enfermedades con el de atención médica.

Para el problema 1, el cual maximiza las medidas clásicas de soporte, confianza y ascenso. La Tabla 5.16, muestra una selección de las reglas

pertenecientes al frente de Pareto que estuvieron entre las primeras 10 en aptitud para al menos una de las medidas de evaluación.

El problema 2, que maximiza los efectos causales, tiene su frente de Pareto representado de forma gráfica en la Figura 5.8. Las reglas específicas encontradas se muestran en la Tabla 5.17.

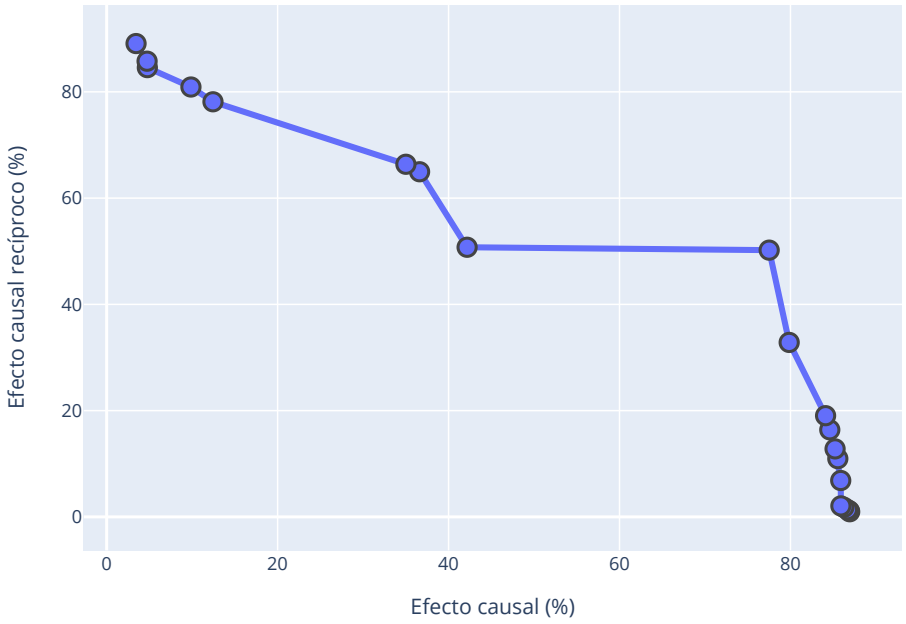

Figura 5.8: Representación gráfica del frente de Pareto para el experimento multi-objetivo B2

Finalmente, el problema 3, que maximiza las medidas de susceptibilidad e impacto en la población, tiene el frente de Pareto representado gráficamente en la Figura 5.9. Las reglas correspondientes se indican en la Tabla 5.18.

Tabla 5.16: Selección de reglas en el frente de Pareto para el experimento multi-objetivo B1.

Regla	Registros	Soporte	Confianza	Ascenso
[COVID-19 = Negativo] → [Defunción = No]	9,158,220	58.8%	99.1%	1.02
[COVID-19 = Negativo] → [Hospitalizado = No]	8,774,148	56.3%	95.0%	1.03
[COVID-19 = Confirmado] → [Hospitalizado = Sí]	635,567	4.1%	11.9%	1.54
[Neumonía = Sí] → [Hospitalizado = Sí]	614,838	4.0%	81.5%	10.57
[COVID-19 = Confirmado][Neumonía = Sí] → [Hospitalizado = Sí]	400,302	2.6%	85.1%	11.05
[Hipertensión = Sí][Neumonía = Sí] → [Hospitalizado = Sí]	230,899	1.5%	90.4%	11.73
[COVID-19 = Confirmado][Neumonía = Sí] → [Defunción = Sí]	213,573	1.4%	45.4%	17.02
[Diabetes = Sí][Neumonía = Sí] → [Hospitalizado = Sí]	198,472	1.3%	91.1%	11.82
[Asma = Sí][COVID-19 = Negativo] → [Defunción = No]	193,117	1.2%	99.3%	1.02
[COVID-19 = Confirmado][Hipertensión = Sí][Neumonía = Sí][Obesidad = Sí] → [Hospitalizado = Sí]	41,956	0.3%	91.7%	11.89
[Neumonía = Sí][Renal Crónica = Sí] → [Hospitalizado = Sí]	39,608	0.3%	92.6%	12.01
[Hipertensión = Sí][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalizado = Sí]	29,769	0.2%	93.1%	12.08
[Diabetes = Sí][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalizado = Sí]	25,029	0.2%	93.5%	12.12
[COVID-19 = Confirmado][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalizado = Sí]	21,683	0.1%	93.8%	12.17
[COVID-19 = Confirmado][Hipertensión = Sí][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalizado = Sí]	16,439	0.1%	94.3%	12.23
[COVID-19 = Confirmado][Neumonía = Sí][Obesidad = Sí] → [Defunción = Sí][Intubado = Sí]	14,584	0.1%	15.0%	26.58
[COVID-19 = Confirmado][Diabetes = Sí][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalizado = Sí]	13,929	0.1%	94.5%	12.26
[COVID-19 = Confirmado][Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalizado = Sí]	11,655	0.1%	94.6%	12.27
[COVID-19 = Confirmado][Hipertensión = Sí][Neumonía = Sí][Obesidad = Sí] → [Intubado = Sí][Hospitalizado = Sí]	8,664	0.1%	18.9%	25.61
[COVID-19 = Confirmado][Hipertensión = Sí][Neumonía = Sí][Obesidad = Sí] → [Defunción = Sí][Intubado = Sí]	7,498	0.1%	16.4%	29.01
[COVID-19 = Confirmado][Neumonía = Sí][Obesidad = Sí] → [Defunción = Sí][Intubado = Sí][UCI = Sí]	5,102	0.0%	5.3%	29.45
[COVID-19 = Confirmado][Diabetes = Sí][Neumonía = Sí][Obesidad = Sí] → [Defunción = Sí][Hospitalizado = Sí][UCI = Sí]	2,727	0.0%	7.5%	29.56
[COVID-19 = Confirmado][Hipertensión = Sí][Neumonía = Sí][Obesidad = Sí] → [Defunción = Sí][Intubado = Sí][UCI = Sí]	2,422	0.0%	5.3%	29.69
[COVID-19 = Confirmado][Hipertensión = Sí][Neumonía = Sí][Obesidad = Sí][Tabaquismo = Sí] → [Intubado = Sí][Hospitalizado = Sí]	954	0.0%	20.5%	27.74
[COVID-19 = Confirmado][Hipertensión = Sí][Neumonía = Sí][Obesidad = Sí][Tabaquismo = Sí] → [Defunción = Sí][Intubado = Sí]	821	0.0%	17.7%	31.25
[COVID-19 = Confirmado][Diabetes = Sí][Inmunosupr. = Sí][Neumonía = Sí][Obesidad = Sí] → [Intubado = Sí]	241	0.0%	21.4%	28.99
[COVID-19 = Confirmado][Diabetes = Sí][Hipertensión = Sí][Inmunosupr. = Sí][Neumonía = Sí][Obesidad = Sí] → [Defunción = Sí][Intubado = Sí][Hospitalizado = Sí]	151	0.0%	19.2%	33.97

Tabla 5.17: Frente de Pareto del experimento multi-objetivo B2

Regla	$AR(A \Rightarrow C)$	$AR(C \Rightarrow A)$
[COVID-19 = Confirmado] [Diabetes = Sí] [Hipertensión = Sí] [Neumonía = Sí] [Renal Crónica = Sí] → [Hospitalizado = Sí]	86.9%	1.0%
[COVID-19 = Confirmado] [Diabetes = Sí] [Neumonía = Sí] [Renal Crónica = Sí] → [Hospitalizado = Sí]	86.9%	1.2%
[COVID-19 = Confirmado] [Hipertensión = Sí] [Neumonía = Sí] [Renal Crónica = Sí] → [Hospitalizado = Sí]	86.7%	1.4%
[COVID-19 = Confirmado] [Neumonía = Sí] [Renal Crónica = Sí] → [Hospitalizado = Sí]	86.3%	1.8%
[Diabetes = Sí] [Neumonía = Sí] [Renal Crónica = Sí] → [Hospitalizado = Sí]	85.9%	2.1%
[COVID-19 = Confirmado] [Diabetes = Sí] [Hipertensión = Sí] [Neumonía = Sí] → [Hospitalizado = Sí]	85.9%	6.9%
[COVID-19 = Confirmado] [Diabetes = Sí] [Neumonía = Sí] → [Hospitalizado = Sí]	85.5%	11.0%
[COVID-19 = Confirmado] [Hipertensión = Sí] [Neumonía = Sí] → [Hospitalizado = Sí]	85.2%	12.8%
[Diabetes = Sí] [Neumonía = Sí] → [Hospitalizado = Sí]	84.6%	16.4%
[Hipertensión = Sí] [Neumonía = Sí] → [Hospitalizado = Sí]	84.1%	19.1%
[COVID-19 = Confirmado] [Neumonía = Sí] → [Hospitalizado = Sí]	79.9%	32.8%
[Neumonía = Sí] → [Hospitalizado = Sí]	77.5%	50.2%
[COVID-19 = Confirmado] [Neumonía = Sí] → [Defunción = Sí] [Hospitalizado = Sí]	42.2%	50.8%
[Neumonía = Sí] → [Defunción = Sí]	36.6%	65.0%
[Neumonía = Sí] → [Defunción = Sí] [Hospitalizado = Sí]	35.0%	66.4%
[Neumonía = Sí] → [Intubado = Sí] [Hospitalizado = Sí]	12.4%	78.1%
[Neumonía = Sí] → [Defunción = Sí] [Intubado = Sí]	9.9%	80.9%
[Neumonía = Sí] → [Intubado = Sí] [Hospitalizado = Sí] [UCI = Sí]	4.8%	84.6%
[Neumonía = Sí] → [Defunción = Sí] [UCI = Sí]	4.7%	85.8%
[Neumonía = Sí] → [Defunción = Sí] [Intubado = Sí] [UCI = Sí]	3.4%	89.1%

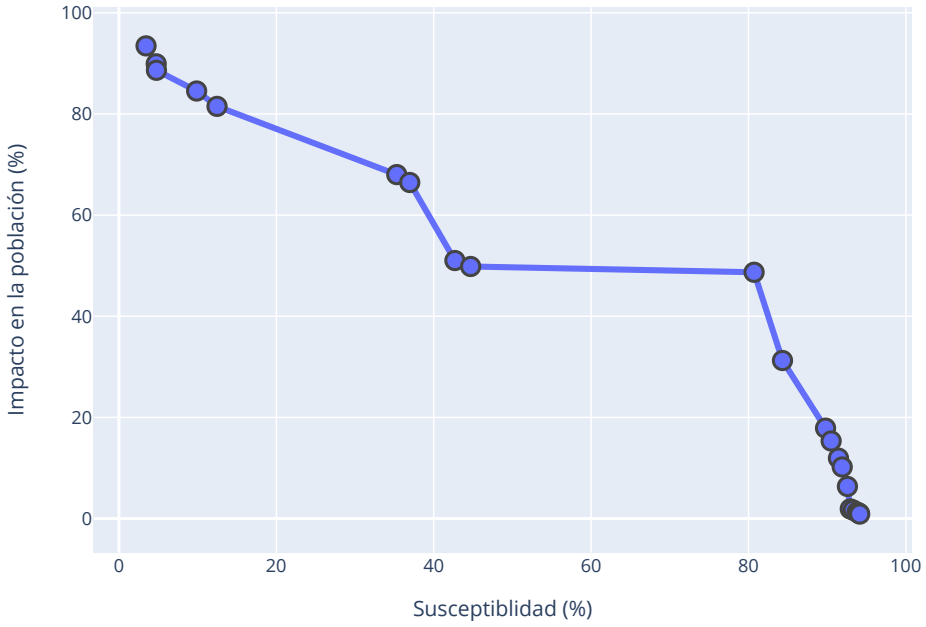

Figura 5.9: Representación gráfica del frente de Pareto para el experimento multi-objetivo B3

Tabla 5.18: Reglas que conforman el frente de Pareto en el experimento multi-objetivo B3

Regla	Susceptibilidad	AF_p
[COVID-19 = Confirmado][Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalización = Sí]	94.1%	0.9%
[COVID-19 = Confirmado][Diabetes = Sí][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalización = Sí]	94.1%	1.1%
[COVID-19 = Confirmado][Hipertensión = Sí][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalización = Sí]	93.8%	1.3%
[COVID-19 = Confirmado][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalización = Sí]	93.3%	1.7%
[Diabetes = Sí][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalización = Sí]	92.9%	1.9%
[COVID-19 = Confirmado][Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	92.6%	6.4%
[COVID-19 = Confirmado][Diabetes = Sí][Neumonía = Sí] → [Hospitalización = Sí]	91.9%	10.2%
[COVID-19 = Confirmado][Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	91.4%	11.9%
[Diabetes = Sí][Neumonía = Sí] → [Hospitalización = Sí]	90.5%	15.3%
[Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	89.8%	17.9%
[COVID-19 = Confirmado][Neumonía = Sí] → [Hospitalización = Sí]	84.3%	31.3%
[Neumonía = Sí] → [Hospitalización = Sí]	80.7%	48.7%
[COVID-19 = Confirmado][Neumonía = Sí] → [Defunción = Sí]	44.7%	49.8%
[COVID-19 = Confirmado][Neumonía = Sí] → [Defunción = Sí][Hospitalización = Sí]	42.7%	51.0%
[Neumonía = Sí] → [Defunción = Sí]	37.0%	66.5%
[Neumonía = Sí] → [Defunción = Sí][Hospitalización = Sí]	35.3%	68.0%
[Neumonía = Sí] → [Intubado = Sí]	12.5%	81.5%
[Neumonía = Sí] → [Defunción = Sí][Intubado = Sí][Hospitalización = Sí]	9.9%	84.5%
[Neumonía = Sí] → [Intubado = Sí][UCI = Sí]	4.8%	88.6%
[Neumonía = Sí] → [Defunción = Sí][Hospitalización = Sí][UCI = Sí]	4.7%	89.9%
[Neumonía = Sí] → [Defunción = Sí][Intubado = Sí][Hospitalización = Sí][UCI = Sí]	3.4%	93.5%

5.7 Interpretación de las reglas multi-objetivo

A continuación se hace una revisión e interpretación de las reglas de asociación obtenidas en los experimentos multi-objetivo para los escenarios A (Edad y Sexo ⇒ Enfermedades), B (Enfermedades ⇒ Atención médica) y C (Ubicación ⇒ Enfermedades). Se seleccionaron estos escenarios en particular por su importancia para describir el comportamiento de la pandemia de COVID-19 en México.

Para cada experimento se seleccionaron las reglas que se encuentran en los extremos del frente de Pareto así como la que toma el mayor valor en la media geométrica de las funciones objetivo que componen el problema, realizando una interpretación de las medidas siendo optimizadas.

5.7.1 Escenario A: Edad y sexo, enfermedades

Problema 1: Soporte, confianza y ascenso

Para este experimento, la regla de asociación con mayor soporte es `[Sexo = Femenino] → [COVID-19 = Negativo]`, cubriendo un 32% de los registros (5.01 millones), con una confianza del 60% y ascenso de 1.02. Esta es la regla que tiene el valor más alto en la media geométrica de soporte, confianza y ascenso, con un valor de 0.58.

La regla de mayor confianza fue `[Edad = [0, 24]][Sexo = Femenino] → [COVID-19 = Negativo]`, con un valor de confianza de 66%, soporte de 0.07 (1.07 millones) y ascenso de 1.11. Esta regla, contrastándola con la regla de mayor soporte, indica que la confianza de no tener COVID-19 aumenta en mujeres en el primer quintil de edad; es decir, entre 0 y 24 años.

La regla de mayor ascenso fue `[Edad = [54, inf]][Sexo = Masculino] → [COVID-19 = Confirmado][EPOC = Sí][Neumonía = Sí][Tabaquismo = Sí]`, con un ascenso de 7.70, soporte de 0.02% (2,422) y confianza de 0.2%. El valor del ascenso de 7.7 indica que existe una asociación entre la población de hombres en el quintil de mayor edad con la COVID-19 así como con el padecimiento simultáneo de neumonía, tabaquismo y obstrucción pulmonar (EPOC).

El resultado de este experimento es un buen ejemplo de la importancia de utilizar la media geométrica en lugar de la media aritmética para promediar valores de diferentes rangos, ya que el uso de la media aritmética en este experimento da peso desproporcionado al ascenso.

Por un lado, la regla de mayor media geométrica, R_g, es `[Sexo = Femenino] → [COVID-19 = Negativo]`, con un valor de 0.58 y un valor en su media aritmética de 0.66. Por otro lado, la regla de mayor media aritmética, R_a, `[Edad = [54, inf]][Sexo = Masculino] → [COVID-19 = Confirmado][EPOC = Sí][Neumonía = Sí][Tabaquismo =`

Sí] tiene una media aritmética de 2.56 y una media geométrica de 0.01. La media aritmética de R_a es 4 veces mayor que la de R_g, a pesar de ser R_g la regla de mayor soporte y la segunda en confianza del frente de Pareto, disminuyendo considerablemente el peso de los valores de soporte y confianza.

Los valores obtenidos de la media geométrica refuerzan la importancia de tratar el problema de minería de reglas de asociación como un problema multi-objetivo, ya que, para el caso particular de este experimento, la combinación de las funciones objetivo en una sola función de aptitud maximiza únicamente soporte y confianza, evitando explorar regiones del espacio de búsqueda con valores altos de ascenso.

Problema 2. Efectos causales

La regla con el mayor efecto causal en este experimento es `[Edad = [54, inf]] → [Hipertensión = Sí]`, con un efecto causal de 27% y un efecto causal en la regla recíproca de 44%. Esta regla maximiza también la media geométrica de ambos efectos causales.

Por otro lado, la regla con el mayor efecto causal en su recíproca es `[Edad = [54, inf]] → [COVID-19 = Negativo][EPOC = Sí][Hipertensión = Sí][Neumonía = Sí]`, con un efecto causal de 0.19% y un efecto causal en la recíproca de 74%. La interpretación de esta regla es que tener 54 o más años es una causa necesaria, pero no suficiente, para presentar las enfermedades que componen el consecuente. Es una causa necesaria por el alto efecto causal en la recíproca, pero no es una causa suficiente, por el bajo efecto causal del antecedente sobre el consecuente.

Problema 3. Susceptibilidad e impacto en la población

La regla con mayor susceptibilidad es `[Edad = [54, inf]][Sexo = Femenino] → [Hipertensión = Sí]`, con susceptibilidad de 29% y AF_p de 25%. La susceptibilidad indica

que las personas con 54 años o más tienen un 29.2% de probabilidad de presentar hipertensión. El valor de impacto en la población señala que pertenecer al rango de edad de 54 años o más y ser mujer tuvo un efecto causal en el 25% de los casos de hipertensión.

Por otro lado, la regla con mayor valor en AF_p es `[Edad = [54, inf]] → [COVID-19 = Negativo][EPOC = Sí][Hipertensión = Sí][Neumonía = Sí]`, con un valor de susceptibilidad de 0.19% y un impacto en la población de 91%. La susceptibilidad indica que la probabilidad de que el consecuente dependa únicamente de la edad es baja; sin embargo, el alto impacto en la población indica que la edad es un factor importante, ya que la edad tuvo un efecto causal en el 91% de las observaciones del consecuente.

La regla con la mayor media geométrica de susceptibilidad e impacto en la población es `[Edad = [54, inf]] → [Hipertensión = Sí]`. La media geométrica es de 0.37, promediando los valores de susceptibilidad de 28.7% y AF_p de 49%. Si se contrasta con la regla de mayor susceptibilidad, se puede observar que generalizar la regla a hombres y mujeres decrementa ligeramente la susceptibilidad, en un 0.5%. El impacto en la población es del 49%, señalando que pertenecer al quintil de mayor edad tuvo un efecto causal en el 49% de los casos de hipertensión.

5.7.2 Escenario B: Enfermedades y atención médica.

Problema 1. Soporte, confianza y ascenso

La regla `[COVID-19 = Negativo] → [Defunción = No]` es la de mayor soporte, cubriendo al 59% (9.15 millones) de los registros, con un valor de confianza de 99.1% y ascenso de 1.0187. Ésta regla es también la regla con la mayor media geométrica.

La regla de mayor confianza es `[Asma = Sí][COVID-19 = Negativo] → [Defunción = No]`, con un soporte del 1% (193,117), confianza del 99.3% y ascenso de 1.01.

La regla de mayor ascenso es `[COVID-19 = Confirmado][Diabetes = Sí][Hipertensión`

= Sí][Inmunosupresores = Sí][Neumonía = Sí][Obesidad = Sí] → [Defunción = Sí][Intubado = 1][Hospitalización = Sí]. Ésta regla tiene un soporte de 151 registros, con una confianza de 19.2% y un ascenso de 33.97. El soporte esperado cuando antecedente y consecuente son independientes es de 4.44, señalando que existe una asociación entre antecedente y consecuente. Esta suposición es razonable dadas las seis comorbilidades incluidas en el antecedente, exigiendo una atención médica que requiere hospitalizar e intubar al paciente.

Problema 2. Efectos causales

La regla con mayor efecto causal es [COVID-19 = Confirmado][Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí][Renal crónica = Sí] → [Hospitalización = Sí], con un efecto causal del 87% y un efecto causal en su recíproca de 0.9%, indicando que el antecedente es una causa suficiente y no necesaria para la hospitalización. El efecto causal alto indica que el antecedente es una causa suficiente para presentar hospitalización. Sin embargo, el antecedente no es una causa necesaria, ya que la hospitalización tiene un efecto causal en menos del 1% de las observaciones del antecedente.

La regla con mayor efecto causal en la regla recíproca es [Neumonía = Sí] → [Defunción = Sí][Intubado = Sí][UCI = Sí][Hospitalización = Sí], con un efecto causal del 3% y un efecto causal en la regla recíproca del 89%. Los valores que toman estas medidas describen a la neumonía como una causa necesaria, pero no suficiente, para el fallecimiento de un paciente que fue hospitalizado, intubado y que recibió cuidados intensivos. La neumonía es una causa necesaria para observar el consecuente, ya que el efecto causal del consecuente sobre el antecedente es alto (89%). Por otro lado, la neumonía no es una causa suficiente por su bajo efecto causal sobre el consecuente, incrementando la probabilidad de observar el consecuente solo en un 3%.

La regla con el mayor valor en la media geométrica de los efectos causales es

[Neumonía = Sí] → [Hospitalización = Sí], con una media geométrica de 0.62, promediando un efecto causal del 78% y un efecto causal en la regla recíproca de 50%. Puede afirmarse que la neumonía es una causa necesaria y suficiente para la hospitalización ya que los efectos causales tanto de la regla como de su recíproca son altos, indicando que existe un incremento considerable en el riesgo de hospitalización cuando se padece neumonía y viceversa. Esta regla toma también el mayor valor en la media geométrica de susceptibilidad e impacto en la población.

Problema 3: Susceptibilidad e impacto en la población

La regla con mayor susceptibilidad, tomando un valor de 94% es [COVID-19 = Confirmado][Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí][Renal Crónica = Sí] → [Hospitalización = Sí], con valor de AF_p de 0.8%. La regla señala que existe una susceptibilidad muy alta de hospitalización si se presentan simultáneamente las enfermedades descritas por el antecedente; aunque dicha combinación de enfermedades causó menos del 1% de los casos de hospitalización.

La regla con mayor valor de impacto en la población es [Neumonía = Sí] → [Defunción = Sí][Intubado = Sí][UCI = Sí][Hospitalización = Sí], con susceptibilidad del 3% e impacto en la población de 93%. Los valores que toma esta regla reflejan que la probabilidad de que la neumonía sea suficiente para requerir la máxima atención médica y fallecer es baja. La regla indica también que es importante considerar que la neumonía causó el 93% de los casos más graves.

Una de las reglas de mayor valor en la media geométrica es [COVID-19 = Confirmado][Neumonía = Sí] → [Hospitalización = Sí], con susceptibilidad del 84% y un valor de impacto en la población del 31%, con una media geométrica de 0.51. Los pacientes con COVID-19 que presentan neumonía tienen una probabilidad del 84% de ser hospitalizados. Además, el valor de impacto en la población señala que el 31% de las

hospitalizaciones fueron causadas por la combinación de ambas enfermedades.

5.7.3 Escenario C. Ubicación y enfermedades.

Problema 1: Soporte, confianza y ascenso

La regla de mayor soporte es `[Sector = SSA] → [COVID-19 = Confirmado]`, la cual cubre el 37% de las instancias (5.82 millones), tiene un valor de confianza de 66.3% y un ascenso de 1.11. Es también la regla con la mayor media geométrica, con un valor de 0.65.

La regla de mayor confianza es `[Entidad = Guanajuato][Sector = Privado] → [COVID-19 = Confirmado]`, con un soporte de 0.18% (27,297), confianza del 98% y ascenso de 2.85.

Finalmente, la regla de mayor ascenso encontrada en este experimento fue `[Entidad = Yucatán][Sector = Privado] → [Cardiovascular = Sí][COVID-19 = Confirmado][Hipertensión = Sí][Neumonía = Sí]`, con soporte de 55 registros, confianza del 4.5% y ascenso de 54.5.

Problema 2. Efectos causales

La regla con el mayor efecto causal es `[Entidad = Yucatán][Sector = Privado] → [Neumonía = Sí]`, con un valor del 89% de efecto causal y un valor de 0.02% de efecto causal en la regla recíproca. Estas medidas indican que el antecedente es una causa suficiente, pero no necesaria, pera presentar neumonía.

La regla `[Sector = IMSS] → [COVID-19 = Confirmado][Hipertensión = Sí][Inmunosupr. = Sí][Renal Crónica = Sí]` es la que tiene el mayor valor de efecto causal en la regla recíproca. El efecto causal de la regla recíproca, con un valor de 27%, indica que, entre los que presentan los padecimientos indicados por el consecuente, la

probabilidad de atenderse en el Instituto Mexicano del Seguro Social se incrementa en un 27%. Esta regla tiene un efecto causal de 0.02%, el cual, al ser tan próximo a cero señala que no existe una relación causal entre recibir atención en dicha institución de salud con los padecimientos observados.

La regla con el mayor valor en la media geométrica de los efectos causales es `[Sector = IMSS] → [COVID-19 = Confirmado]`, tomando un valor de 0.18, donde el efecto causal y el efecto causal en la regla recíproca es del 18%.

Problema 3. Susceptibilidad e impacto en la población

La regla de mayor susceptibilidad es `[Entidad = Nayarit][Sector = Cruz Roja] → [COVID-19 = Confirmado]`, indicando que los pacientes que recibieron atención médica en la Cruz Roja en dicha entidad fueron predominantemente casos confirmados de COVID-19, con una susceptibilidad del 97%. El valor de impacto en la población es muy próximo a cero, señalando que el antecedente tuvo un efecto causal en una fracción muy pequeña de los casos confirmados de COVID-19.

La regla con el mayor valor de impacto en la población es `[Sector = IMSS] → [Hipertensión = Sí][Inmunosupr. = Sí][Neumonía = Sí][Renal Crónica = Sí]`, tomando un valor de impacto en la población de 40% y una susceptibilidad de 0.02%. Estos valores señalan que el recibir atención en el IMSS tuvo un efecto causal en el 40% de los casos en los que se observa el consecuente. Sin embargo, la susceptibilidad tan cercana a cero indica que los pacientes del IMSS tienen poca probabilidad de terminar presentando el consecuente.

La regla con el mayor valor en la media geométrica es `[Sector = SSA] → [COVID-19 = Negativo]`, con un valor de 0.22. Esta regla indica una susceptibilidad del 32% en pacientes que recibieron atención por parte de la SSA, señalando además, a través del valor del impacto en la población, que la atención en dicho sector tuvo un efecto causal en el 15% de los casos

negativos de COVID-19.

5.8 Comparativa entre los resultados mono-objetivo y multi-objetivo

El constraste de los resultados obtenidos en experimentos mono-objetivo con los multi-objetivo justifica la importancia de tratar la minería de reglas de asociación como un problema multi-objetivo, ya que las medidas de evaluación con las cuales se experimentó presentan intercambios.

En los experimentos multi-objetivo, las reglas que pertenecen al frente de Pareto ofrecen al tomador de decisiones una variedad de opciones así como más información para comprender el comportamiento de las medidas de evaluación seleccionadas, ofreciendo reglas con valores a través de todo el rango de cada medida. Las reglas con valores altos para cada medida así como aquellos intermedios pudieron ser recuperadas en los experimentos multi-objetivo.

Este comportamiento no se observa en los experimentos mono-objetivo donde las reglas con los valores más altos en una medida usualmente tenían valores muy bajos para las otras medidas de evaluación del problema. Por otro lado, el combinar las diferentes medidas en una función de aptitud recupera las reglas que quedan en un punto intermedio del frente de Pareto; sin embargo, esto reduce la posibilidad de llegar a los extremos, ya que al combinar las medidas, la aptitud general de la regla se reduce.

5.9 Experimentación por olas de contagio

Los resultados reportados en las secciones 5.4 y 5.6 se obtuvieron del conjunto de datos en su totalidad, por lo que reportan un promedio a través de la temporalidad de datos disponible. En la presente sección se presentan los resultados de realizar la minería en subconjuntos de datos correspondientes a las cuatro olas de contagio.

5.9.1 Diseño de los experimentos

Para esta clase de experimentos, un experimento es la combinación entre uno de los tres escenarios definidos manualmente, uno de los tres problemas multiobjetivo y una de las cuatro olas de contagio. Los escenarios corresponden a los planteados en la Tabla 5.8, los problemas a los indicados en la tabla 5.11 y las olas de contagio son las señaladas en la Tabla 5.3.

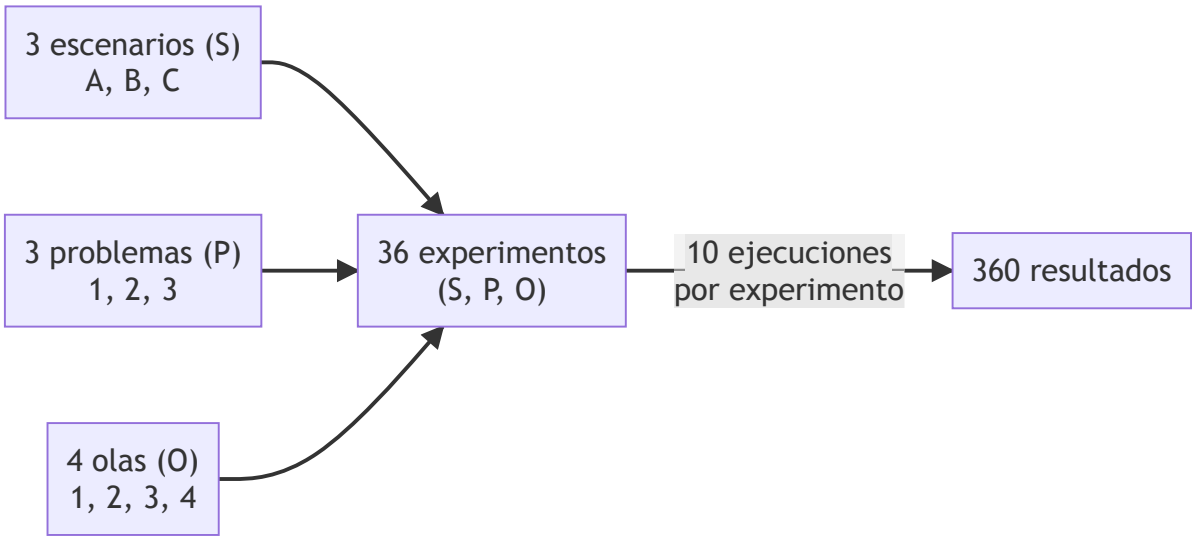

Figura 5.10: Descripción de la experimentación por olas de contagio

La configuración de parámetros tales como el tamaño de la población y criterio de paro son consistentes con las planteadas para la experimentación multi-objetivo.

5.9.2 Soluciones no dominadas por ola de contagio

En la Tabla 5.19 se muestran la media (μ) y desviación estándar (σ) del total de soluciones no dominadas en cada ejecución de los experimentos realizados.

Tabla 5.19: Resultados del número total de soluciones no dominadas por ola de contagio

Experimento	Ola 1		Ola 2		Ola 3		Ola 4		Todos los datos	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
A1	28	0	28	0	29	0	26	0	27	0
A2	9	0	28	0	14	0	15	0	9	0
A3	13	0	15	0	13	0	14	0	11	0
B1	38	0	40	0	39	0	40	0	39	0
B2	16	0	21	0	24	0	28	0	20	0
B3	23	0	22	0	23	0	28	2.345	21	0
C1	16.4	1.14	16.8	1.789	18.2	1.095	10.2	0.837	16	0.707
C2	9.6	1.517	13.4	1.14	10.8	1.643	3	0.707	8.8	1.643
C3	10.6	1.14	12.6	2.408	11.4	0.548	8.2	1.304	11.2	1.095

El escenario C tiende a tener más variación en la cantidad de soluciones no dominadas en los experimentos realizados, esto debido a que tiene una cantidad de selectores que participan en la construcción de reglas que es considerablemente mayor a la de otros escenarios.

5.9.3 Reglas recuperadas en todas las olas de contagio

Es importante identificar aquellas reglas que se mantienen a través de diferentes intervalos de tiempo. Para esto se identificó la cantidad de veces en las que una regla fue recuperada, tanto en los experimentos realizados en las diferentes olas de contagio como en el conjunto de datos en su totalidad. Si una regla se repite cinco veces corresponde a que dicha regla fue recuperada en cada ola de contagio como en todo el conjunto de datos. Los resultados correspondientes se muestran en la Tabla 5.20.

Tabla 5.20: Número de repeticiones de las reglas no dominadas obtenidas en los diferentes conjuntos de datos.

Experimento	1	2	3	4	5	Total	Media
A1	19	6	4	9	12	50	2.78
A2	31	3	2	3	4	43	1.74
A3	12	2	1	3	7	25	2.64
B1	68	16	8	11	9	112	1.9
B2	23	7	4	9	5	48	2.29
B3	25	6	6	5	10	52	2.4
C1	66	12	9	6	0	93	1.52
C2	39	11	7	3	0	60	1.57
C3	54	7	4	6	2	73	1.56

El experimento con la mayor cantidad de reglas recuperadas en todos los conjuntos de datos es el experimento A1, con 12 reglas. Por otro lado, en los experimentos C1 y C2 no existieron reglas tales que se repitieran en los cinco conjuntos de datos, lo cual se puede atribuir al que el espacio de búsqueda inducido por el escenario C es mayor.

El experimento en el que en promedio existen más repeticiones es en el escenario A1, con una media de 2.78 repeticiones; seguido por el experimento A3, en el que una regla se repitió 2.64 veces en promedio. En los experimentos en los que hay menos repeticiones en promedio son los experimentos C1, C2 y C3, todos con aproximadamente 1.5 repeticiones por regla.

5.9.4 Variación en los efectos causales por ola de contagio

Para ilustrar la variación en las medidas de evaluación por ola de contagio, se seleccionaron las reglas que fueron recuperadas en los subconjuntos de datos por olas así como

en el conjunto de datos completo.

En la Tabla 5.21 se muestran los valores que toman las 9 reglas que ocurren en todos los conjuntos de datos seleccionados. Las columnas numeradas del 1 al 4 representan una ola diferente respectivamente, mientras que la columna marcada como T representa el total del conjunto de datos.

Tabla 5.21: Medidas de evaluación para reglas del problema 2

Exp.	**Regla**	**Efecto causal**				
		1	**2**	**3**	**4**	**T**
A2	[Edad = [54, inf]] → [Diabetes = Sí][Hipertensión = Sí]	0.164	0.137	0.129	0.109	0.135
A2	[Edad = [54, inf]] → [Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí]	0.069	0.04	0.025	0.016	0.035
A2	[Edad = [54, inf]] → [Hipertensión = Sí]	0.311	0.28	0.259	0.227	0.272
A2	[Edad = [54, inf]] → [Hipertensión = Sí][Neumonía = Sí]	0.123	0.072	0.044	0.027	0.063
B2	[Diabetes = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.764	0.846	0.876	0.848	0.846
B2	[Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.76	0.843	0.872	0.841	0.841
B2	[Neumonía = Sí] → [Defunción = Sí]	0.372	0.409	0.349	0.24	0.366
B2	[Neumonía = Sí] → [Defunción = Sí][Hospitalización = Sí]	0.349	0.391	0.344	0.23	0.35
B2	[Neumonía = Sí] → [Hospitalización = Sí]	0.754	0.796	0.782	0.643	0.775
Exp.	**Regla**	**Efecto causal recíproco**				
		1	**2**	**3**	**4**	**T**
A2	[Edad = [54, inf]] → [Diabetes = Sí][Hipertensión = Sí]	0.504	0.53	0.54	0.521	0.534
A2	[Edad = [54, inf]] → [Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí]	0.571	0.618	0.644	0.678	0.63
A2	[Edad = [54, inf]] → [Hipertensión = Sí]	0.428	0.447	0.438	0.414	0.442
A2	[Edad = [54, inf]] → [Hipertensión = Sí][Neumonía = Sí]	0.553	0.597	0.61	0.65	0.605
B2	[Diabetes = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.195	0.185	0.137	0.115	0.164
B2	[Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.219	0.221	0.155	0.14	0.19
B2	[Neumonía = Sí] → [Defunción = Sí]	0.641	0.657	0.641	0.534	0.65
B2	[Neumonía = Sí] → [Defunción = Sí][Hospitalización = Sí]	0.655	0.67	0.655	0.548	0.664
B2	[Neumonía = Sí] → [Hospitalización = Sí]	0.576	0.557	0.455	0.339	0.502

Para las reglas seleccionadas correspondientes al escenario A, en el que se relacionan las variables de edad y sexo con las enfermedades, se observa que el efecto causal va disminuyendo conforme fueron ocurriendo nuevas olas de contagio. También se aprecia que la ola que tiene valores más próximos al valor obtenido para todo el conjunto de datos es la ola 2.

Con respecto al efecto causal recíproco, este se incrementa en la mayoría de reglas conforme avanzaron las olas de contagio, señalando que la relación del antecedente como causa necesaria del consecuente se fue fortaleciendo. En este escenario, la ola con mayor similaridad entre los valores obtenidos al realizar minería sobre todo el conjunto de datos es con los valores de la ola 2.

En el escenario B, en el que se relacionan las enfermedades con la atención médica recibida, se observa que es en las olas 2 y 3 cuando las reglas alcanzan el mayor efecto causal, decrementando en las fases posteriores. Así como en el escenario A, los valores que más se aproximan al comportamiento general son los de la ola 2. Analizando el efecto causal recíproco, se aprecia también que se reduce para las reglas del escenario B conforme ocurrieron nuevas olas de contagio. Para muchas reglas, el comportamiento del conjunto de datos en su totalidad es similar al de la ola 2.

Cabe recalcar que para el problema 2 no existieron reglas que ocurrieran en todos los conjuntos de datos, tanto en las olas de contagio como en el conjunto de datos completo.

Para el problema 3 se maximizan simultáneamente las medidas de susceptibilidad e impacto en la población (AF_p).

En el escenario A, el cual relaciona la edad y el sexo con las enfermedades, los valores de susceptibilidad de las reglas obtenidas se reduce conforme ocurren nuevas olas de contagio, siendo la ola 2 la que se aproxima mejor al comportamiento general observado en todo el conjunto de datos. La regla con el valor más alto de susceptibilidad en este escenario es `[Edad = [54, inf]] → [Hipertensión = Sí]` en la ola 1, con un 33.9% de susceptibilidad. Con respecto al impacto en la población, la regla de mayor impacto en la población para una ola en particular es `[Edad = [54, inf]] → [EPOC = Sí][Hipertensión = Sí][Neumonía = Sí]`, con un valor de 93.5% en la ola 4.

Para el escenario B, donde se relacionan las enfermedades con la atención médica recibida, se observa en términos generales un incremento de la ola 1 a la ola 2, reduciéndose tanto la susceptibilidad como el impacto en la población en olas subsecuentes. La regla con el valor más alto de susceptibilidad en alguna de las olas es `[COVID-19 = Confirmado][Diabetes = Sí][Neumonía = Sí] → [Hospitalización = Sí]` en la ola 3, con una susceptibilidad del 94.2% y un impacto en la población del 9.6%. Por otro lado, la regla con el valor más alto de impacto en la población en alguna de las olas es `[Neumonía = Sí] → [Defunción`

= Sí][Intubado = Sí][Hospitalización = Sí][UCI = Sí] en la ola 2, con un 94.9% de impacto en la población y una susceptibilidad del 3.4%.

Finalmente, en el escenario C, el cual relaciona variables como la ubicación y la institución de atención médica con las enfermedades, se observa que la regla con mayor susceptibilidad es [Sector = SSA] → [COVID-19 = Negativo] en la ola 4, con una susceptibilidad de 36.2% y un impacto en la población de 27.7%. Por otro lado, la regla con mayor impacto en la población es [Entidad = Ciudad de México] → [COVID-19 = Negativo][Tabaquismo = Sí] en la ola 2, con un impacto en la población de 32.3% y susceptibilidad del 3.8%.

Tabla 5.22: Medidas de evaluación para reglas del problema 3

Exp.	**Regla**	**Susceptibilidad**				
		1	**2**	**3**	**4**	**T**
A3	[Edad = [54, inf]] → [Diabetes = Sí][Hipertensión = Sí]	0.168	0.14	0.131	0.111	0.138
A3	[Edad = [54, inf]] → [Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí]	0.069	0.04	0.025	0.016	0.036
A3	[Edad = [54, inf]] → [EPOC = Sí][Hipertensión = Sí]	0.022	0.014	0.014	0.011	0.015
A3	[Edad = [54, inf]] → [EPOC = Sí][Hipertensión = Sí][Neumonía = Sí]	0.01	0.005	0.004	0.003	0.005
A3	[Edad = [54, inf]] → [EPOC = Sí][Neumonía = Sí]	0.018	0.008	0.006	0.004	0.008
A3	[Edad = [54, inf]] → [Hipertensión = Sí]	0.339	0.299	0.271	0.238	0.288
A3	[Edad = [54, inf]] → [Hipertensión = Sí][Neumonía = Sí]	0.124	0.072	0.044	0.027	0.063
B3	[COVID-19 = Conf.][Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.902	0.93	0.95	0.909	0.926
B3	[COVID-19 = Conf.][Diabetes = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.898	0.924	0.942	0.887	0.919
B3	[COVID-19 = Conf.][Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.891	0.921	0.938	0.878	0.914
B3	[Diabetes = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.89	0.912	0.921	0.876	0.905
B3	[Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.881	0.906	0.916	0.869	0.898
B3	[Neumonía = Sí] → [Defunción = Sí]	0.38	0.414	0.351	0.241	0.37
B3	[Neumonía = Sí] → [Defunción = Sí][Intubado = Sí][Hospitalización = Sí][UCI = Sí]	0.04	0.034	0.035	0.019	0.034
B3	[Neumonía = Sí] → [Defunción = Sí][Hospitalización = Sí]	0.356	0.396	0.346	0.231	0.353
B3	[Neumonía = Sí] → [Defunción = Sí][Hospitalización = Sí][UCI = Sí]	0.053	0.048	0.049	0.029	0.047
B3	[Neumonía = Sí] → [Hospitalización = Sí]	0.816	0.829	0.807	0.659	0.807
C3	[Entidad = CIUDAD DE MÉXICO] → [COVID-19 = Negativo][Tabaquismo = Sí]	0.044	0.038	0.025	0.016	0.031
C3	[SECTOR = SSA] → [COVID-19 = Negativo]	0.265	0.35	0.284	0.362	0.322
Exp.	**Regla**	AF_p				
		1	**2**	**3**	**4**	**T**
A3	[Edad = [54, inf]] → [Diabetes = Sí][Hipertensión = Sí]	0.62	0.65	0.618	0.609	0.633
A3	[Edad = [54, inf]] → [Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí]	0.734	0.788	0.757	0.814	0.772
A3	[Edad = [54, inf]] → [EPOC = Sí][Hipertensión = Sí]	0.809	0.837	0.833	0.819	0.833
A3	[Edad = [54, inf]] → [EPOC = Sí][Hipertensión = Sí][Neumonía = Sí]	0.882	0.913	0.912	0.935	0.907
A3	[Edad = [54, inf]] → [EPOC = Sí][Neumonía = Sí]	0.839	0.887	0.873	0.908	0.872
A3	[Edad = [54, inf]] → [Hipertensión = Sí]	0.474	0.504	0.475	0.457	0.488
A3	[Edad = [54, inf]] → [Hipertensión = Sí][Neumonía = Sí]	0.696	0.753	0.715	0.779	0.736
B3	[COVID-19 = Conf.][Diabetes = Sí][Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.064	0.074	0.058	0.055	0.064
B3	[COVID-19 = Conf.][Diabetes = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.107	0.116	0.096	0.079	0.102
B3	[COVID-19 = Conf.][Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.121	0.14	0.108	0.097	0.119
B3	[Diabetes = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.168	0.172	0.131	0.111	0.153
B3	[Hipertensión = Sí][Neumonía = Sí] → [Hospitalización = Sí]	0.19	0.206	0.148	0.135	0.179
B3	[Neumonía = Sí] → [Defunción = Sí]	0.681	0.674	0.651	0.54	0.664
B3	[Neumonía = Sí] → [Defunción = Sí][Intubado = Sí][Hospitalización = Sí][UCI = Sí]	0.93	0.949	0.931	0.912	0.935
B3	[Neumonía = Sí] → [Defunción = Sí][Hospitalización = Sí]	0.701	0.689	0.666	0.554	0.68
B3	[Neumonía = Sí] → [Defunción = Sí][Hospitalización = Sí][UCI = Sí]	0.902	0.9	0.905	0.864	0.899
B3	[Neumonía = Sí] → [Hospitalización = Sí]	0.545	0.54	0.443	0.332	0.487
C3	[Entidad = Ciudad de México] → [COVID-19 = Negativo][Tabaquismo = Sí]	0.234	0.323	0.236	0.28	0.276
C3	[SECTOR = SSA] → [COVID-19 = Negativo]	0.215	0.181	0.098	0.277	0.153

5.9.5 Frentes de Pareto por olas de contagio

Para comparar los frentes de Pareto obtenidos a través de las diferentes olas de contagio se acumularon, por experimento, los frentes de Pareto de todas las ejecuciones de un experimento para una ola de contagio. En este acumulado se realizó un nuevo ordenamiento no dominado para identificar un nuevo frente de Pareto compuesto por reglas de diferentes ejecuciones. En la Tabla 5.23 se muestra el total de soluciones no dominadas tras aplicar la técnica descrita previamente. Las columnas corresponden a cada una de las 4 olas de contagio así como al conjunto de datos en su totalidad.

Tabla 5.23: Total de soluciones no dominadas en los acumulados por experimento.

Experimento	**Ola 1**	**Ola 2**	**Ola 3**	**Ola 4**	**Todos los datos**
A1	28	28	29	27	27
A2	9	5	14	15	9
A3	13	15	13	14	11
B1	38	40	39	40	44
B2	16	21	24	28	20
B3	23	22	23	31	21
C1	21	26	26	17	19
C2	12	16	16	4	12
C3	13	17	14	12	13

Posteriormente se procedió a graficar los frentes de Pareto identificados en cada experimento. Para los experimentos que involucran el problema 1 se utilizó una gráfica de coordenadas paralelas debido a que tiene más de dos objetivos. La gráfica de coordenadas paralelas permite mostrar a través de poli-líneas las soluciones no dominadas del problema multiobjetivo, cada solución es representada por un color diferente y su nivel correspondiente por cada función objetivo. Finalmente, para los experimentos de los problemas 2 y 3 se usó

una gráfica de dispersión al ser problemas de dos objetivos.

Problema 1. Soporte, confianza y ascenso

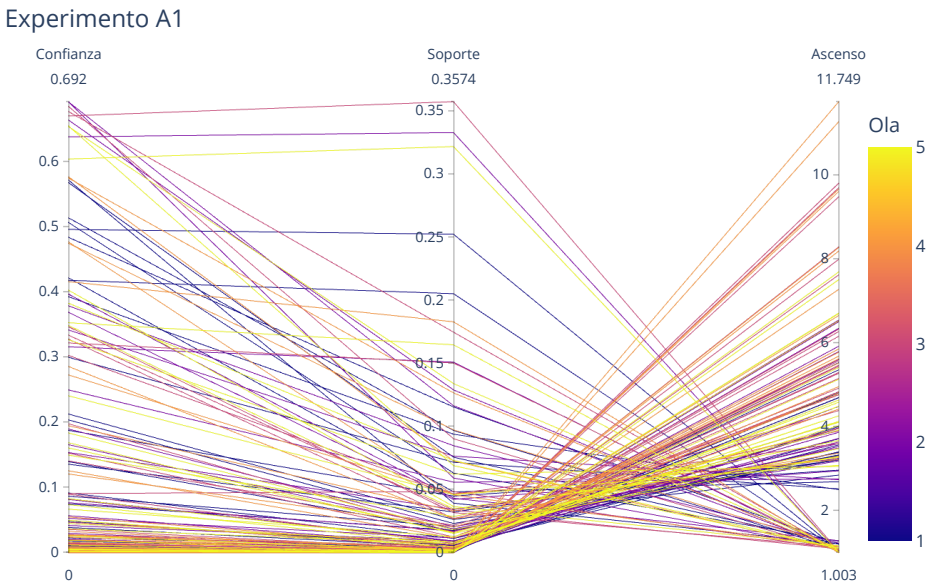

Figura 5.11: Representación gráfica de los frentes de Pareto para cada ola (experimento A1)

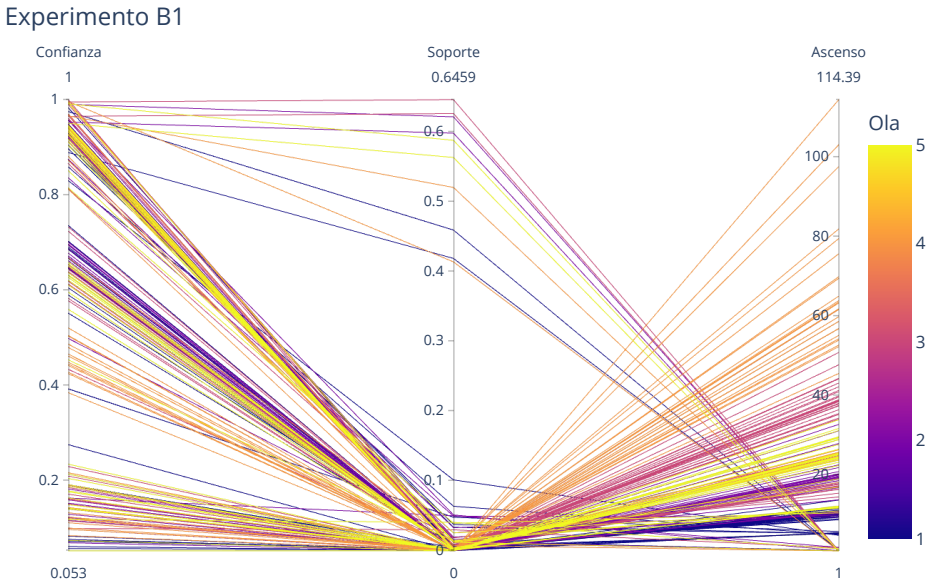

Figura 5.12: Representación gráfica de los frentes de Pareto para cada ola (experimento B1)

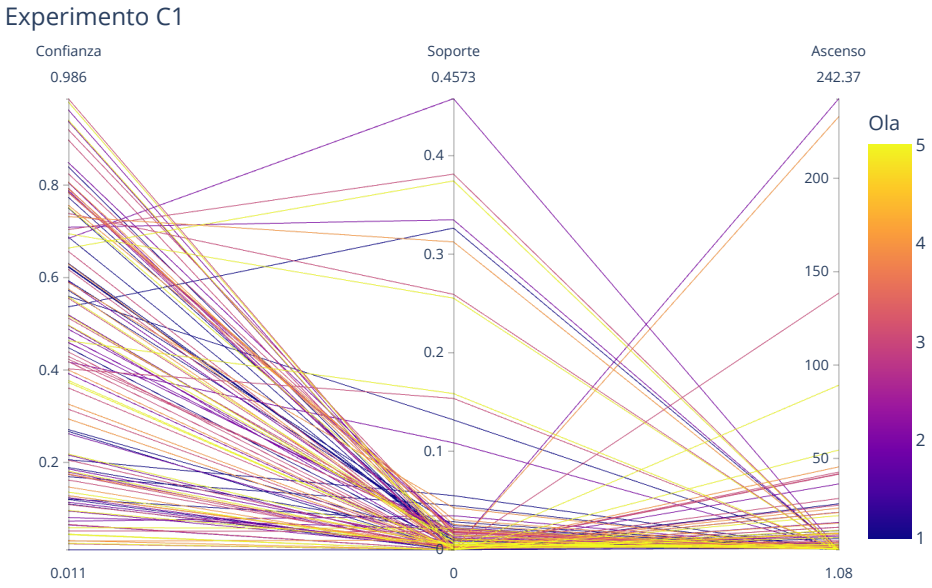

Figura 5.13: Representación gráfica de los frentes de Pareto para cada ola (experimento C1)

Problema 2. Efectos causales

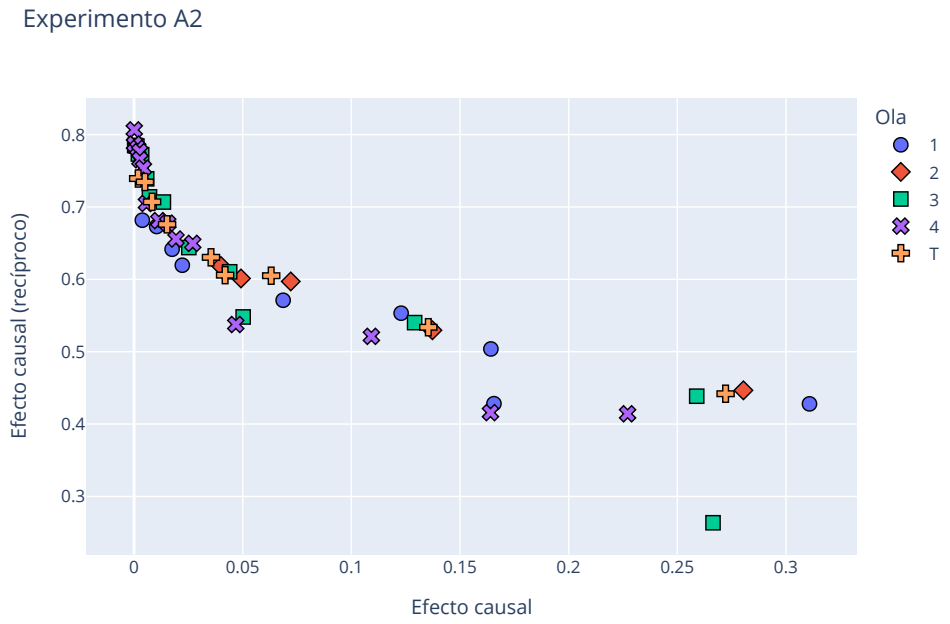

Figura 5.14: Representación gráfica de los frentes de Pareto para todas las olas (experimento A2)

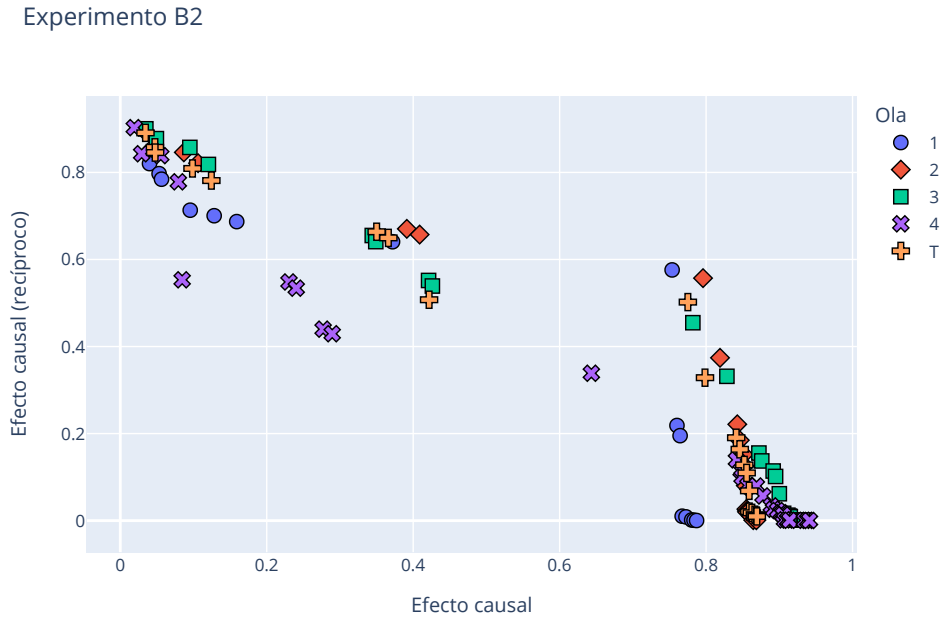

Figura 5.15: Representación gráfica de los frentes de Pareto para todas las olas (experimento B2)

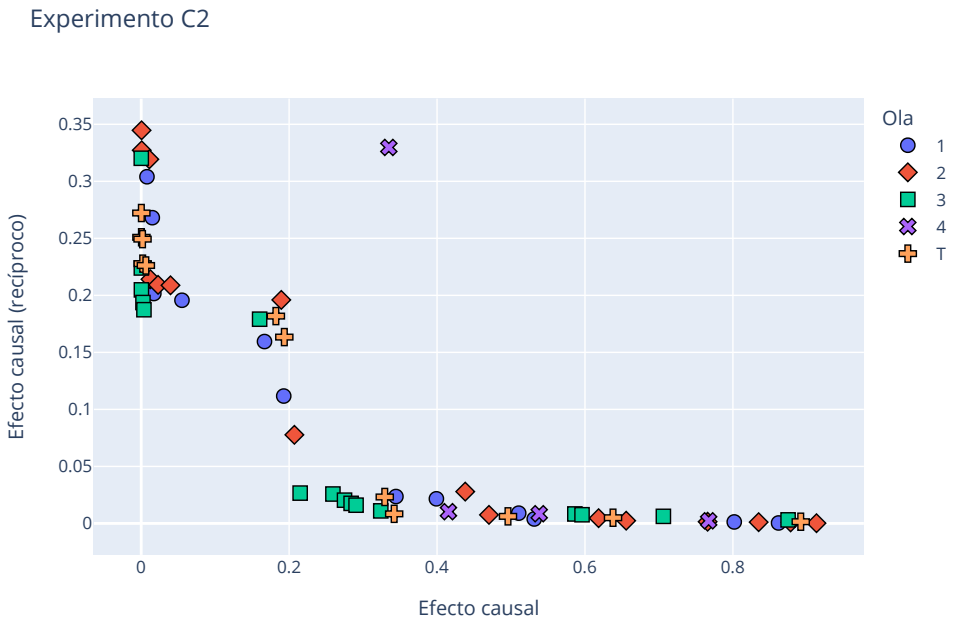

Figura 5.16: Representación gráfica de los frentes de Pareto para todas las olas (experimento C2)

Problema 3. Susceptibilidad, impacto en la población

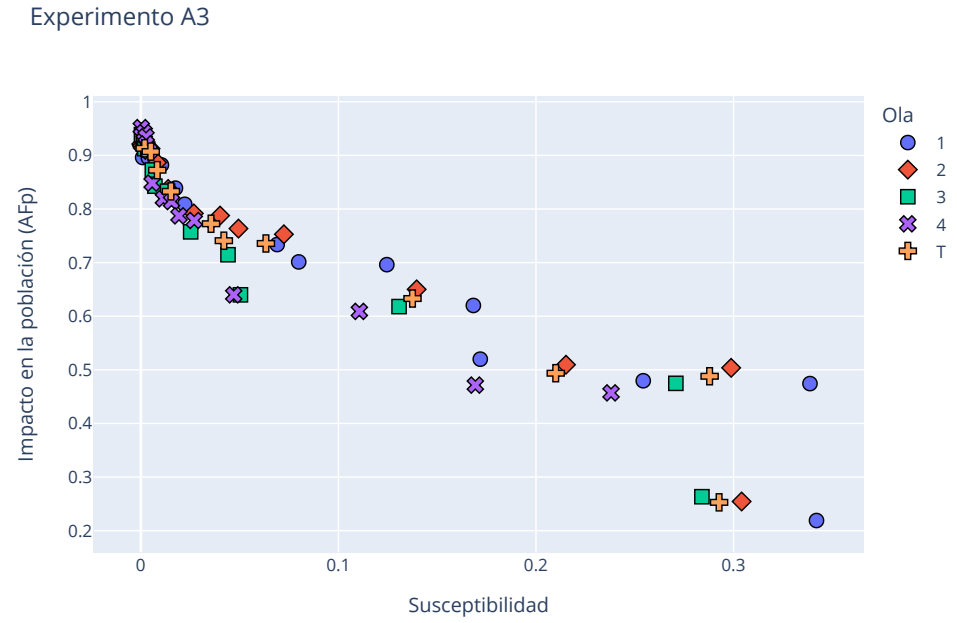

Figura 5.17: Representación gráfica de los frentes de Pareto para todas las olas (experimento A3)

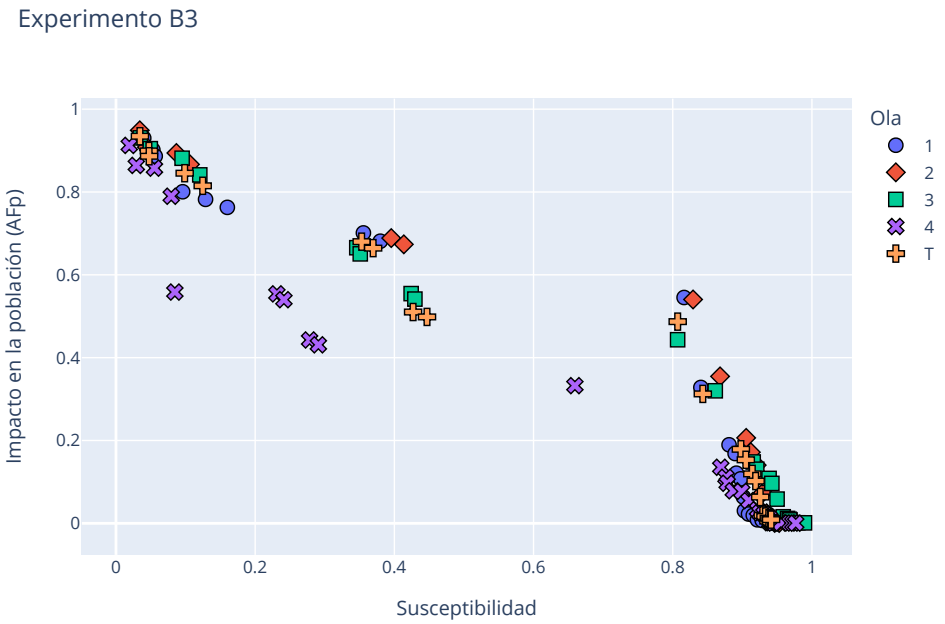

Figura 5.18: Representación gráfica de los frentes de Pareto para todas las olas (experimento B3)

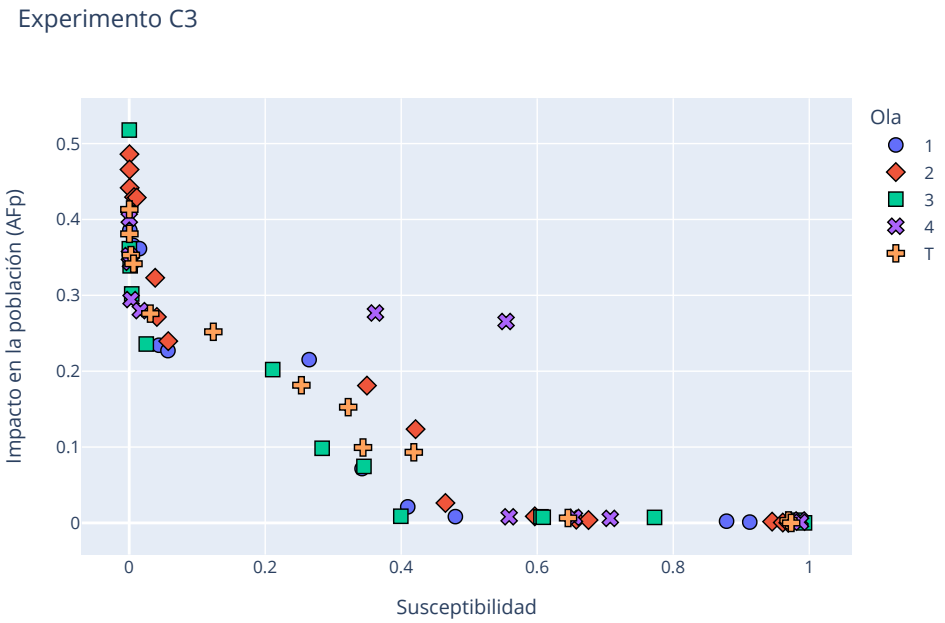

Figura 5.19: Representación gráfica de los frentes de Pareto para todas las olas (experimento C3)

Capítulo 6

Conclusiones y trabajo a futuro

En el trabajo de tesis se presentó una técnica de minería de minería de reglas de asociación basada en un algoritmo evolutivo multi-objetivo.

Al ser un algoritmo evolutivo, la técnica genera y combina patrones para después realizar la minería de las reglas de asociación mejor evaluadas por las funciones objetivo dadas. Esta característica evita tanto la minería de patrones frecuentes como la evaluación exhaustiva de reglas, estrategias utilizadas por los algoritmos clásicos de minería de reglas de asociación, las cuales exigen de considerable tiempo y recursos de cómputo.

La definición de grupos de atributos permite a la técnica propuesta el responder a consultas específicas sobre la relación entre dos conjuntos de atributos en particular, asignando uno al antecedente y otro al consecuente. Además, se propuso un algoritmo de descubrimiento de grupos de atributos, el cual asiste en la construcción de grupos de atributos así como en identificar pares de grupos de interés sobre los cuales realizar minería.

A diferencia de los algoritmos clásicos, la técnica propuesta es capaz de realizar minería de reglas de asociación sin requerir forzosamente de la definición de umbrales mínimos de soporte y confianza, cuya selección usualmente es arbitraria, depende de ensayo y error y puede realizar inadvertidamente omisiones de reglas de asociación interesantes.

Al integrar un algoritmo multi-objetivo, la técnica propuesta es capaz de recuperar las reglas de asociación mejor evaluadas para múltiples medidas de evaluación. Estas medidas de evaluación pueden ser clásicas, como soporte, confianza y ascenso; así como medidas de estimación causal, tales como el efecto causal, susceptibilidad e impacto en la población.

Los resultados obtenidos en el trabajo confirman la importancia de tratar la minería de reglas de asociación como un problema multi-objetivo. En los experimentos multi-objetivo, los conjuntos de reglas resultantes tienen tanto valores altos como valores bajos en las medidas de evaluación, ofreciendo una perspectiva amplia al momento de realizar la interpretación de las reglas; a diferencia de los experimentos mono-objetivo, los cuales usualmente se quedan en los extremos, dependiendo de la medida de evaluación seleccionada.

El uso de medidas de estimación causal para evaluar reglas de asociación fue probada en los experimentos multi-objetivo, en particular el uso de esta clase de medidas de evaluación para identificar y describir relaciones de causa y efecto, realizando un aporte al estudio de la causalidad desde una perspectiva computacional.

El reto de realizar minería sobre un conjunto de datos con un volumen de 15.5 millones de datos exigió la integración de tecnologías como bases de datos columnares y en memoria, reduciendo el tiempo de cómputo considerablemente en comparación con la minería clásica.

Con respecto al caso de estudio, las reglas recuperadas del conjunto de datos son útiles para describir, en términos generales, las relaciones que comparten las variables demográficas, epidemiológicas y médicas sobre la pandemia de COVID-19 en México, considerando principalmente la información reportada por fuentes oficiales. Es importante observar también que, para el caso de estudio, el espacio de la evolución artificial resulta mucho mayor que el espacio de los datos observados, razón por la cual sería de especial valor el enriquecimiento de los datos de COVID-19, esto con el propósito de recuperar reglas que describan a mayor detalle el fenómeno siendo estudiado.

Dados los planteamientos anteriores, se afirma que se cumplieron los objetivos, tanto

generales como específicos planteados para este trabajo.

Como trabajo futuro se propone extender el algoritmo con la inclusión de operadores lógicos además de la conjunción. El presente trabajo, el cual ya ha sido probado con la representación conjuntiva, serviría entonces como marco de referencia para comparar las reglas conjuntivas con reglas que estén compuestas por expresiones con diversos operadores lógicos.

Se sugiere también explorar mejoras al algoritmo de descubrimiento de grupos de atributos para que pueda admitir restricciones definidas por los usuarios, en las que se especifiquen conjuntos de atributos que deban siempre pertenecer al mismo grupo o atributos que deban estar siempre separados.

Es de especial valor considerar como trabajo futuro el realizar la minería de datos de reglas causales en diferentes periodos de tiempo, así como proponer una técnica para comparar de forma sistemática las reglas obtenidas y diferencias entre las mismas en cada periodo. Asimismo, debe considerarse como trabajo futuro el analizar las funciones objetivo para determinar si generan reglas consideradas como óptimos locales.

Finalmente, se propone probar la técnica propuesta para que realice minería sobre conjuntos de datos de dominios diferentes al seleccionado para el caso de estudio.

Referencias

[1] R. Sethi y B. Shekar, «Subjective Interestingness in Association Rule Mining: A Theoretical Analysis», en *Digital Business*, vol. 21, X.-S. Yang, S. Patnaik, M. Tavana, F. Popentiu-Vlădicescu, y F. Qiao, Eds. Cham: Springer International Publishing, 2019, pp. 375-389. doi: 10.1007/978-3-319-93940-7_15.

[2] T. Dasu y T. Johnson, *Exploratory data mining and data cleaning*. New York: Wiley-Interscience, 2003.

[3] J. Fürnkranz, D. Gamberger, y N. Lavrač, *Foundations of Rule Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-540-75197-7.

[4] J. Y. F. Lau, *An introduction to critical thinking and creativity: think more, think better*. Hoboken, N.J: Wiley, 2011.

[5] P. Menzies y H. Price, «Causation as a Secondary Quality», *The British Journal for the Philosophy of Science*, vol. 44, n.º 2, pp. 187-203, 1993, Accedido: 1 de junio de 2022. [En línea]. Disponible en: https://www.jstor.org/stable/687643

[6] J. Pearl, *Causality: models, reasoning, and inference*. Cambridge, U.K. ; New York: Cambridge University Press, 2000.

[7] J. Pearl y D. Mackenzie, *The book of why: the new science of cause and effect*. New York: Basic Books, 2018.

[8] J. Tian y J. Pearl, «Probabilities of causation: Bounds and identification», *Annals of Mathematics and Artificial Intelligence*, vol. 28, n.º 1/4, pp. 287-313, 2000, doi: 10.1023/A:1018912507879.

[9] P. R. Rosenbaum, *Design of observational studies*, Second edition. Cham, Switzerland: Springer, 2020.

[10] M. A. Hernán y J. M. Robins, *Causal Inference: What If*, 1.ª ed. Boca Raton, FL: Chapman & Hall/CRC, 2020.

[11] J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin, y B. Sun, «Mining Causal Association Rules», en *2013 IEEE 13th International Conference on Data Mining Workshops*, dic. 2013, pp. 114-123. doi: 10.1109/ICDMW.2013.88.

[12] P. Cummings, «The Relative Merits of Risk Ratios and Odds Ratios», *Archives of Pediatrics & Adolescent Medicine*, vol. 163, n.º 5, p. 438, 2009, doi: 10.1001/archpediatrics.2009.31.

[13] P. W. Cheng, «From covariation to causation: A causal power theory.», *Psychological Review*, vol. 104, n.º 2, pp. 367-405, 1997, doi: 10.1037/0033-295X.104.2.367.

[14] M. A. Mansournia y D. G. Altman, «Population attributable fraction», *BMJ*, p. k757, feb. 2018, doi: 10.1136/bmj.k757.

[15] O. S. Miettinen, «PROPORTION OF DISEASE CAUSED OR PREVENTED BY A GIVEN EXPOSURE, TRAIT OR INTERVENTION», *American Journal of Epidemiology*, vol. 99, n.º 5, pp. 325-332, 1974, doi: 10.1093/oxfordjournals.aje.a121617.

[16] B. Rockhill, B. Newman, y C. Weinberg, «Use and misuse of population attributable fractions.», *American Journal of Public Health*, vol. 88, n.º 1, pp. 15-19, ene. 1998, doi: 10.2105/AJPH.88.1.15.

[17] R. Poli, W. Langdon, y N. Mcphee, *A Field Guide to Genetic Programming*. 2008.

[18] S. Mirjalili y J. S. Dong, *Multi-Objective Optimization using Artificial Intelligence Techniques*. 2020. Accedido: 25 de noviembre de 2020. [En línea]. Disponible en: https://link.springer.com/10.1007/978-3-030-24835-2

[19] C. A. Coello Coello, «A Comprehensive Survey of Evolutionary-Based Multiobjective Optimization Techniques», *Knowledge and Information Systems*, vol. 1, n.º 3, pp. 269-308, ago. 1999, doi: 10.1007/BF03325101.

[20] G. Pappa y A. Freitas, «Evolving rule induction algorithms with multi-objective grammar-based genetic programming», *Knowl. Inf. Syst.*, vol. 19, pp. 283-309, jun. 2009, doi: 10.1007/s10115-008-0171-1.

[21] K. Deb, *Multi-objective optimization using evolutionary algorithms*, 1st ed. Chichester ; New York: John Wiley & Sons, 2001.

[22] R. Agrawal, T. Imieliński, y A. Swami, «Mining association rules between sets of items in large databases», en *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*, 1993, pp. 207-216. doi: 10.1145/170035.170072.

[23] J. Han, J. Pei, y Y. Yin, «Mining frequent patterns without candidate generation», *ACM SIGMOD Record*, vol. 29, n.º 2, pp. 1-12, jun. 2000, doi: 10.1145/335191.335372.

[24] A. Telikani, A. H. Gandomi, y A. Shahbahrami, «A survey of evolutionary computation for association rule mining», *Information Sciences*, vol. 524, pp. 318-352, jul. 2020, doi: 10.1016/j.ins.2020.02.073.

[25] B. Alataş y E. Akin, «An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules», *Soft Computing*, vol. 10, n.º 3, pp. 230-237, feb. 2006, doi: 10.1007/s00500-005-0476-x.

[26] X. Yan, C. Zhang, y S. Zhang, «Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support», *Expert Systems with Applications*, vol. 36, n.º 2, pp. 3066-3076, mar. 2009, doi: 10.1016/j.eswa.2008.01.028.

[27] M. Md. J. Kabir, S. Xu, B. H. Kang, y Z. Zhao, «A new multiple seeds based genetic algorithm for discovering a set of interesting Boolean association rules», *Expert Systems with Applications*, vol. 74, pp. 55-69, 2017, doi: 10.1016/j.eswa.2017.01.001.

[28] J. M. Luna, J. R. Romero, y S. Ventura, «G3PARM: A Grammar Guided Genetic Programming algorithm for mining association rules», en *IEEE Congress on Evolutionary Computation*, jul. 2010, pp. 1-8. doi: 10.1109/CEC.2010.5586504.

[29] J. M. Luna, A. Cano, y S. Ventura, «Genetic Programming for Mining Association Rules in Relational Database Environments», p. 20, 2015.

[30] D. Martín, A. Rosete, J. Alcalá-Fdez, y F. Herrera, «A multi-objective evolutionary algorithm for mining quantitative association rules», en *2011 11th International Conference on Intelligent Systems Design and Applications*, nov. 2011, pp. 1397-1402. doi: 10.1109/ISDA.2011.6121855.

[31] J. Li *et al.*, «From Observational Studies to Causal Rule Mining», *ACM Transactions on Intelligent Systems and Technology*, vol. 7, n.º 2, pp. 14:1-14:27, nov. 2015, doi: 10.1145/2746410.

[32] P. Yadav, P. J. Caraballo, M. Steinbach, V. Kumar, M. R. Castro, y G. Simon, «Frequent Causal Pattern Mining: A Computationally Efficient Framework For Estimating Bias-Corrected Effects», *Proceedings : ... IEEE International Conference on Big Data. IEEE International Conference on Big Data*, vol. 2019, pp. 1981-1990, dic. 2019, doi: 10.1109/bigdata47090.2019.9005977.

[33] H. Elhilbawi, S. Eldawlatly, y H. Mahdi, «The Importance of Discretization Methods in Machine Learning Applications: A Case Study of Predicting ICU Mortality», en *Advanced Machine Learning Technologies and Applications*, vol. 1339, K.-C. Chang, A.-E. Hassanien, y T. Mincong, Eds. Cham: Springer International Publishing, 2021, pp. 214-224. doi: 10.1007/978-3-030-69717-4_23.

[34] K. Tanabe, «Pareto's 80/20 rule and the Gaussian distribution», *Physica A: Statistical Mechanics and its Applications*, vol. 510, pp. 635-640, nov. 2018, doi: 10.1016/j.physa.2018.07.023.

[35] B. Kouakou, «On-line algorithm for the minimal b-clique cover problem in interval graphs». 2006. Accedido: 8 de abril de 2022. [En línea]. Disponible en: https://halshs.archives-ouvertes.fr/halshs-00123607

[36] D. Brélaz, «New methods to color the vertices of a graph», *Communications of the ACM*, vol. 22, n.º 4, pp. 251-256, abr. 1979, doi: 10.1145/359094.359101.

[37] S. de Salud, «Información referente a casos COVID-19 en México - datos.gob.mx/busca». 2022. Accedido: 18 de marzo de 2022. [En línea]. Disponible en: https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico

[38] S. de Salud, «Informe integral de COVID-19 en México». abril de 2022. Accedido: 6 de junio de 2022. [En línea]. Disponible en: https://coronavirus.gob.mx/wp-content/uploads/2022/05/Info-05-22-Int_COVID-19_6abr_26abr22OK.pdf

[39] W. Bergsma, «A bias-correction for Cramér's and Tschuprow's», *Journal of the Korean Statistical Society*, vol. 42, n.º 3, pp. 323-328, sep. 2013, doi: 10.1016/j.jkss.2012.10.002.