UNIVERSIDAD
DE ATACAMA

FACULTAD DE INGENIERÍA
DEPTO. DE ING. INFORMÁTICA Y CIENCIAS DE LA COMPUTACIÓN

APLICACIÓN DE REGLAS DE ASOCIACIÓN PARA EL DESCUBRIMIENTO DE PATRONES ENTRE ENFERMEDADES RESPIRATORIAS Y VARIABLES MEDIOAMBIENTALES EN LA COMUNA DE COPIAPÓ.

Tesina presentada como parte de los requisitos para obtener el título de Ingeniero Civil Informática y las Ciencias de la Computación

Profesor Guía: Mg. Andrés Alfaro Avalos

Carlo Andrés Troncoso Araya

Copiapó, Chile 2024.

Agradecimientos

Agradezco a mi madre y a mi hermana por apoyarme en todo momento a lo largo de mi carrera profesional, por creer en mí y ayudarme a superar cada obstáculo al que me enfrente a través del curso de los años.

A mi profesor guía Andres Alfaro Avalos, Magister en Ingeniería Informática, cuya paciencia, comprensión y constancia facilitaron la realización de este trabajo de titulación, sus consejos fueron útiles y oportunos en el momento que más los necesite para seguir adelante.

A los docentes de esta carrera que acompañaron mi proceso de formación a lo largo de los años.

También a mis compañeros de carrera con quienes compartí bellos momentos e interminables horas de estudio que atesoro desde ya, especialmente a mi amigo Luis Espejo, quien siempre me apoyó y orientó en todo lo qué pudo.

Tabla de Contenidos

Índice de Figuras

Índice de Tablas

Resumen

Las enfermedades respiratorias son una de las principales causas de muerte a nivel mundial, con un gran impacto en la salud de niños y adultos mayores. Estas ocurren por diversos factores tales como infecciones, consumo de sustancias tóxicas e inhalación de humo.

En la actualidad, la gran cantidad de datos generados por el sector salud gracias a la revolución digital representa una oportunidad para implementar tecnologías que permitan aprovechar esta información de forma eficiente e innovadora.

El presente trabajo de titulación se centra en el análisis de las enfermedades respiratorias en la comuna de Copiapó, Chile, con el objetivo de descubrir patrones entre estas enfermedades y las variables medioambientales. Para ello, se utiliza un modelo de minería de datos y reglas de asociación, empleando datos recopilados entre enero de 2017 y diciembre de 2021, los cuales fueron convertidos a un sistema de categorías compatible con los algoritmos de extracción de reglas.

El estudio confirma que las enfermedades respiratorias constituyen un problema de salud pública en Copiapó. Las infecciones virales, la contaminación ambiental, y los factores sociodemográficos se identifican como los principales factores asociados a su desarrollo. Por ejemplo, un hallazgo importante del estudio indica que a finales de invierno y comienzo de primavera, ciertas concentraciones de MP10 según los datos analizados, apuntan a un posible aumento del riesgo de padecer otras enfermedades respiratorias, como la bronquitis, la rinitis, la faringitis y la amigdalitis.

En general, las reglas de asociación se presentan como una herramienta valiosa para el análisis del comportamiento de las enfermedades respiratorias, permitiendo descubrir patrones y relaciones entre variables que pueden ser de gran utilidad para mejorar la prevención y el tratamiento de estas enfermedades.

Capítulo I Introducción

Las enfermedades respiratorias son una de las principales causas de muerte a nivel mundial (Rivera et al., 2016), en Chile estas vienen a ser la tercera causa de muerte con un 9.5% del total, antecedida solo por enfermedades del sistema circulatorio y tumores (INE , 2016), aquejando principalmente a niños y adultos mayores. El deseo de reducir el número de afectados ha sido siempre un desafío difícil de afrontar. Sin embargo, como expone La Estrategia Nacional de Salud 2011-2020, cuyo objetivo *"es mejorar el acceso a la atención en salud oportuna, segura y de buena calidad, considerando las expectativas de la población, en un marco de respeto de los derechos de las personas en salud"*, también manifiesta que los esfuerzos realizados hasta la fecha no son suficientes y proponen, como una de las alternativas, aumentar la inversión en sistemas de información que ayuden a tomar decisiones que conduzcan a mejorar estos resultados.

Actualmente contamos con un gran número de datos en el sector de la salud, consecuencia de la gran cantidad de dispositivos electrónicos producto de la transformación digital, proceso que consiste en reorientar una organización hacia la aplicación y uso de tecnologías emergentes, como el Big Data, que se puede definir como el conjunto de estrategias que posibilitan recopilar y analizar un gran número de datos, en los cuales se detectan patrones ocultos que hacen visible información relevante y que no puede analizarse de forma convencional (Puyol Moreno, 2014), asimismo la Minería de Datos, que es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados. Empleando una amplia variedad de técnicas y ciencias (estadística, informática, matemáticas, ingeniería, entre otras) con el objetivo de extraer información no trivial de grandes volúmenes de datos, utilizando el análisis matemático para deducir los patrones y tendencias que existen (Siebes 2000), y por último del Aprendizaje de máquina, que en términos simples, se refiere a sistemas o máquinas que imitan la inteligencia humana para realizar tareas a través del reconocimiento de patrones (Vega et al., 2020). Con todo lo anterior se puede llevar a cabo modelos descriptivos que nos ayuden a entender mejor esta problemática.

Bajo este contexto el presente trabajo de apoyo a la investigación busca utilizar estas tecnologías para encontrar soluciones y mejoras en temas de salud. En particular el objetivo es la búsqueda de asociaciones significativas entre variables ambientales, de contaminación con la ocurrencia de enfermedades respiratorias, describiendo así su comportamiento, es decir detectar cuáles aspectos medioambientales (climatológicos, contaminantes) afectan o influyen en la reacción de este tipo de enfermedades en la comuna de Copiapó. Dicho en otras palabras, conocer cómo se relacionan o asocian este tipo de variables (relación de tipo inferencial, frecuencia, entre otras). Para ello se propone una metodología basada en KDD (*Knowledge Discovery in Databases*) y CRISP-DM (*Cross Industry Standard Process for Data Mining*) que consta de 8 etapas (1. Selección del subconjunto de datos, 2. Pre-procesamiento de los datos, 3. Revisión del conocimiento de dominio. 4. Creación del dataset, 5. Diseño e implementación del modelo, 6. Aplicación del modelo, 7. Validación del modelo, 8. Análisis y evaluación de los resultados). Con ello se pretende entregar como resultado un modelo que describa los patrones encontrados entre las variables de interés previamente mencionadas en esta sección, utilizando técnicas de asociación de minería de datos.

En el capítulo 2 se presenta el marco teórico que abarca los aspectos relacionados a las enfermedades respiratorias, las variables ambientales y contaminantes, así como las técnicas más utilizadas en minería de datos y los algoritmos de interés, en capítulo 3 se da a conocer un resumen de algunos trabajos relacionados con esta investigación, el capítulo 4 muestra detalladamente la metodología utilizada, para que finalmente en el capítulo 5 se da a conocer las conclusiones de esta investigación.

1.1. Objetivos

A continuación se detallan tanto el objetivo general como los objetivos específicos de esta investigación:

1.1.1. Objetivo general

Aplicación de reglas de asociación para el descubrimiento de patrones entre enfermedades respiratorias y variables medioambientales en la comuna de Copiapó.

1.1.2. Objetivos específicos

- Utilizar técnicas de minería de datos que permitan describir la relación entre las variables seleccionadas.

- Implementar técnicas de minería de datos seleccionadas para descubrir la relación entre las variables seleccionadas.

- Obtener los patrones de comportamiento de las enfermedades respiratorias a partir de los resultados alcanzados.

1.2. Alcance

Para esta investigación se consideró el material particulado de 10 micras (MP10) y el material particulado fino de menos de 2.5 micras (MP2.5) como variables contaminantes únicamente, el motivo detrás de esta decisión se debe a que el repositorio en donde se encuentra esta variable no cuenta con registros validados de las otras variables en el periodo de tiempo estudiado (2017-2021), por ello se decide prescindir de ellas, además las enfermedades respiratorias corresponden a los registros del hospital regional de Copiapó San José del Carmen, es decir, este estudio no considera los diagnósticos de otras instituciones médicas.

Capítulo II Marco teorico

En este capítulo se abordarán los conceptos clave necesarios para comprender los aspectos teóricos relacionados con las enfermedades respiratorias y su impacto en la salud de las personas. También se explorarán las técnicas de minería de datos utilizadas en el análisis de las variables empleadas en este proceso.

2.1. Enfermedades Respiratorias

Cuando se habla de enfermedades respiratorias, se hace referencias a la definición dada por la Organización Mundial de la Salud (OMS) como aquellas enfermedades que "*afectan a las vías respiratorias, incluidas las vías nasales, los bronquios y los pulmones. Incluyen desde infecciones agudas como la neumonía y la bronquitis a enfermedades crónicas como el asma y la enfermedad pulmonar obstructiva crónic*a" (W.H Organization, OMS, 2015), estas además se encuentran entre las principales causas de muerte a nivel mundial y con elevados gastos de hospitalización según expone A. M. Rivera et al., (2016).

Actualmente, en Chile existen iniciativas para combatir o más bien mitigar las consecuencias de estas enfermedades tales como el AUGE/GES, campañas de prevención y vacunación, programas de vigilancia epidemiológica, entre otros.

Para la clasificación de las enfermedades respiratorias, existe el sistema denominado clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud, conocido mundialmente como CIE-10 (CIE-10 ES). Este sistema clasifica a través de códigos los diferentes tipos de diagnósticos, organizando así las enfermedades. El capítulo X del CIE-10 clasifica las enfermedades respiratorias o "Enfermedades del aparato respiratorio", ordenándolas desde la J00 - J99 (Ver Tabla 2.1). Para este trabajo se consideran las enfermedades respiratorias presentes en este capítulo, específicamente aquellas que corresponden los rangos agrupados de J00 a J06, J09 a J18, J20 a J22, J40 a J47, J30 a J39 y J60 a J98, además del código U07.1 que se encuentra en el capítulo XXII llamado "Códigos para uso de emergencia" correspondiente a "Enfermedad respiratoria aguda debido al nuevo coronavirus SARS-CoV-2" más conocido como

COVID-19. Para un mayor acercamiento al detalle de estos códigos de enfermedades puede ver el Anexo C, correspondiente al capítulo X Enfermedades del aparato respiratorio del CIE-10.

Tabla 2.1: Categorización de enfermedades respiratorias

Código	**Categoría/Grupo**
J00–J06	Infecciones agudas de las vías respiratorias superiores
J10–J18	Influenza [gripe] y neumonía
J20–J22	Otras infecciones agudas de las vías respiratorias inferiores
J30–J39	Otras enfermedades de las vías respiratorias superiores
J40–J47	Enfermedades crónicas de las vías respiratorias inferiores
J60–J70	Enfermedades del pulmón debidas a agentes externos
J80–J84	Otras enfermedades respiratorias que afectan principalmente el intersticio
J85–J86	Afecciones supurativas y necróticas de las vías respiratorias inferiores
J90–J94	Otras enfermedades de la pleura
J95–J99	Otras enfermedades del sistema respiratorio

2.2. Variables Ambientales

Las variables ambientales o meteorológicas son aquellas que miden el estado de la atmósfera en un momento y lugar determinado (IDEAM, 2019). Estas variables son temperatura, presión, viento, humedad y precipitación. (Rodríguez Jiménez et al., 2004).

Las variables ambientales consideradas para esta investigación son temperatura, rocío,

humedad, temperatura mínima y temperatura máxima. Las cuales son de interés para este proyecto registradas desde el 01 enero 2017 al 31 de diciembre 2021. Estos datos son medidas con estaciones climáticas las cuales se valen de sensores para su registro, dichos registros fueron obtenidos de la Dirección General de Aeronáutica Civil (DGAC).

2.3. Variables Contaminantes

Las variables contaminantes miden los niveles de contaminación existentes en el aire, la OMS (WHO, 2021) define que entre las principales directrices se encuentran el material particulado (MP), el ozono (O3), el dióxido de nitrógeno (NO2) y el dióxido de azufre (SO2).

2.3.1 Material particulado (MP)

Se define al material particulado como un *"conjunto de partículas sólidas y/o líquidas (a excepción del agua pura) presentes en suspensión en la atmósfera"* (Mészáros, 1999, citado en Viana, 2003, p.1), por regla general el material particulado se compone por la mezcla de metales, sustancias salinas, materiales carbonosos, compuestos volátiles y las endotoxinas que suelen formar otros compuestos (Soukup & Becker, 2001; Alfaro et al., 2002; Schlesinger et al., 2006; Billet et al., 2007).

El material particulado se clasifica de acuerdo a su tamaño aerodinámico, es un parámetro importante para medir el efecto en la salud y su medio de transporte (Zheng et al., 2002), por consiguiente su clasificación por el tamaño aerodinámico del material particulado vendría ser la siguiente: Partículas cuyo diámetro es menor a 10 μm (MP10) se les conoce también como fracción respirable o inhalable y se considera un indicador confiable de exposición a las enfermedades respiratorias, también son denominadas como partículas gruesas, finalmente están las partículas con diámetro aerodinámico menor a 2,5 μm (MP2.5) que representa un indicador a la salud debido a que pueden penetrar en el sistema respiratorio y llegar hasta los conductos más bajos del pulmón (alvéolos), se considera también partículas finas. (Secretaría de Medio Ambiente y Recursos Naturales [SEMARNAT], 2011, p.15).

Ambos tipos de material particulado MP10 y MP2,5 son variables de interés en esta investigación y fueron obtenidos de las bases de datos publicadas por el Sistema de Información Nacional de Calidad del Aire (SINCA).

2.4. Análisis de Datos

La revolución digital ha hecho posible que la información digitalizada sea fácilmente capturada, almacenada, procesada y distribuida, en la medida en que la tecnología informática avanza y su adopción en los diferentes aspectos de la vida es cada vez más común, debido a esto se han generado una gran cantidad de datos que luego son almacenados en repositorios. Descubrir conocimiento en este enorme volumen de datos es un reto en sí mismo tal como expone José C. Riquelme, Roberto Ruiz, Karina Gilbert (Minería de Datos: Conceptos y Tendencias, 2006).

En este contexto el enfoque de esta investigación es el área de la salud en donde actualmente se genera una gran cantidad de datos, principalmente debido a lo mencionado en el párrafo anterior y la implementación de herramientas electrónicas y digitales (Bellinger, M. Jabbar, O.Zaïane, A.Osornio-Vargas, 2017).

Actualmente el uso de tecnologías y herramientas para extraer, transformar y obtener información de esta nube de datos está tomando cada vez más importancia. Sin embargo a nivel nacional su uso no se ha extendido (Estrategia nacional de salud, 2010).

A continuación se describen algunas tecnologías utilizadas en este trabajo.

2.4.1. Minería de datos

La minería de datos (*DM* por sus siglas en inglés) es un proceso iterativo y automatizado que consiste en la extracción de información útil y patrones de comportamiento que describen grandes volúmenes de datos (Linoff et al., 1997). Este proceso recibe su nombre debido a la analogía que se hace con la explotación de metales y piedras preciosas de la tierra. La comparación se debe a que, al igual que en la minería tradicional, en la minería de datos se busca extraer algo valioso y no trivial de una gran

cantidad de material. En la minería de datos, se utilizan técnicas de análisis estadístico y aprendizaje automático para identificar patrones y relaciones en los datos. A diferencia de la minería tradicional, la minería de datos se realiza de manera automatizada y se enfoca en la extracción de información útil en lugar de materiales físicos (MC Beatriz Beltrán Martínez, Minería De Datos, 2001).

Para llevar a cabo este propósito la minería de datos se vale de técnicas de varias disciplinas como son la estadística, inteligencia artificial (IA) o el aprendizaje automático, también de representaciones gráficas de información que permiten apoyar la toma de decisiones. Es importante recalcar que la minería de datos es una herramienta exploratoria y no explicativa, ya que explora los datos para sugerir hipótesis, pero es un error aceptar dichas hipótesis como explicaciones de tipo causa-efecto y viene a ser necesario validar los resultados con nuevos datos para tener una confirmación, debido a esto es que no existe una metodología estándar que resuelva todos los problemas pues estos deben ser adecuados a cada caso particular.

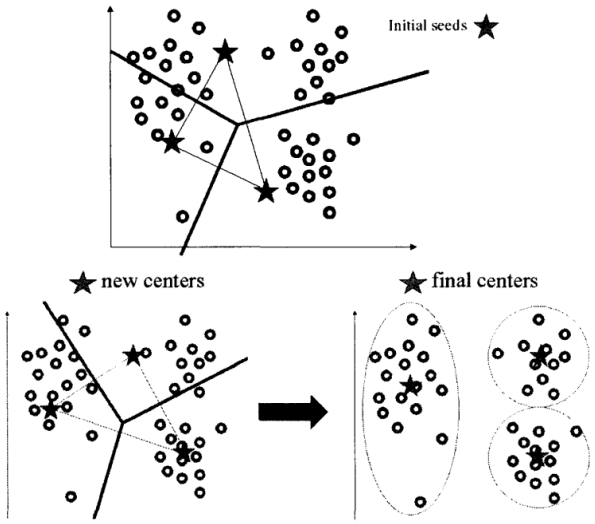
2.4.2. Tecnicas de Mineria de Datos

Las técnicas usadas en minería de datos se dividen principalmente en 2 grandes grupos, técnicas predictivas y técnicas descriptivas, un modelo predictivo pretende responder preguntas sobre datos futuros, para lo cual se vale generalmente del uso de algoritmos de aprendizaje supervisado que son capaces de predecir un atributo (etiqueta) de un conjunto de datos si se tienen los atributos de otro conjunto conocido denominado conjunto de entrenamiento, con esto se pretende obtener las etiquetas del conjunto desconocido (MC Beatriz Beltrán Martínez, Minería De Datos, 2001). Por el lado los modelos descriptivos no asignan un rol prioritario ni etiquetan las variables, ya que no suponen la existencia de un modelo previo a los datos, si no por el contrario, estos modelos se crean automáticamente partiendo del reconocimiento de patrones (C. P. López, 2017), algunas de las técnicas más utilizadas son:

-***Clustering:*** La técnica de agrupamiento o *Clustering* implica un análisis de uno o más atributos para identificar datos que son similares entre sí para comprender las diferencias

y similitudes entre el conjunto de datos (Abdullahi, 2019). A veces es llamado también como segmentación porque "segmenta" los datos en categorías para identificar un grupo de resultados correlacionados (Beltrán 2001). Uno de los algoritmos más utilizados y conocidos es el algoritmo K-Means ver Figura 2.1.

Figura 2.1: Funcionamiento del algoritmo K-Means



Fuente: Zaki, M. J., & Wong, L. (2004). Data Mining Techniques

-Clasificación: Esta técnica de clasificación es utilizada para clasificar una colección de datos en diferentes grupos o clases con el fin de obtener predicciones y análisis precisos en un gran conjunto de datos (Abdullahi, 2019). para identificar una clase en particular. Por ejemplo, puede categorizar fácilmente los edificios en diferentes tipos (según la ocupación o según el tipo de construcción) mediante la identificación de diferentes atributos (estructura, altura o unidad).

-**Asociación:** La asociación es una técnica de minería de datos que descubre patrones basados en la relación entre variables en una misma transacción. También se conoce como técnica de relación porque usa la relación entre elementos y descubre la ocurrencia frecuente de diferentes elementos que aparecen con las frecuencias más altas dentro del conjunto de datos. Estas técnicas tienen múltiples aplicaciones y se usan ampliamente para ayudar a descubrir correlaciones de ventas en bases de datos transaccionales en el ámbito del *retail* (Abdullahi Sidow Osman, Data Mining Techniques: Review, 2019).

Así como la minería de asociación es un área que se centra en encontrar patrones de asociación entre elementos en un conjunto de datos. Las reglas de asociación son una forma de representar estos patrones, en la sección a continuación estas reglas se describen en más detalle.

2.4.3. Reglas de Asociación

Las reglas de asociación corresponden a un tipo de análisis que extrae información buscando coincidencias, para ello realiza una búsqueda de correlaciones o coocurrencias en los sucesos de la base de datos y formaliza la obtención de reglas de tipo:

{si} -> {entonces} (también representado con el ejemplo: {pan}->{leche, huevos})

Transformándose así en un apoyo a la hora de descubrir conocimiento a partir de la información analizada. Un ejemplo típico de aplicación de las reglas de asociación es el análisis de la canasta de compra de las clientes de tiendas tipo *retail,* el descubrimiento de tales reglas ayuda a desarrollar estrategias de mercadeo que se traducen en un aumento de las ventas nos cuenta (Moya Amaris et al,2003).

En minería de datos los datos son típicamente representados en una tabla en donde cada fila es una muestra de los datos y cada columna es un atributo, este atributo representa una caracteristica de la muestra como por ejemplo: nombre, edad, sexo, estado civil, entre otras. En una base de datos transaccional cada atributo es conocido como "*ítem*" y el conjunto de *items* presente en la muestra (o transacción) se conoce como "*itemset*". Matemáticamente se puede representar la búsqueda de una regla de asociación de la forma:

Sea $D = \{d_1, d_2, d_3, \ldots, d_m\}$ un conjunto de *ítems* y $T = \{t_1, t_2, t_3, \ldots t_n\}$ un conjunto de transacciones, donde cada transacción t_i es un conjunto de *ítems* tal que

$$t_i \subseteq D \; donde \; 1 \leq i \leq n$$ (Malberti Riveros et al., 2015)

La implicancia de $X \Rightarrow Y$ es una regla de Asociación donde $X \subset D$, $Y \subset D$, $X \cap Y \neq \phi$, y $X \cup Y \subseteq t_i$, los conjuntos X e Y son mutuamente excluyentes, t_i es el itemset formado por el antecedente o el consecuente de la regla de Asociación, $X \cup Y$ debe estar contenido en alguna de las transacciones de T (Malberti Riveros, María Alejandra, & Elida Beguerí, Graciela, Reglas de Asociación con los datos de una biblioteca universitaria, 2015).

Las métricas más utilizadas en las reglas de Asociación son soporte, confianza y *lift.* A continuación se detallarán cada una de ellas.

2.4.3.1. Soporte

Se define el soporte de la regla de Asociación de tipo $X \Rightarrow Y$ en el *itemset D* como el porcentaje de transacciones en D que contienen a $X \cup Y$, es decir, la frecuencia con la que aparece un itemset en la base de datos y se calcula como la proporción de transacciones que contienen el itemset, de la forma:

$$soporte(X \Rightarrow Y) = soporte(X \cup Y)$$ (Malberti Riveros et al., 2015).

2.4.3.2. Confianza

Se define la confianza de una regla de Asociación de tipo $X \Rightarrow Y$ en el *itemset D* como el porcentaje de transacciones en D que sabiendo que contienen X además contienen Y (Zaki, M. J., & Wong, L., 2004) como sigue:

$$confianza(X \Rightarrow Y) = \frac{soporte(X \cup Y)}{soporte(X)}$$

La confianza puede ser considerada como una probabilidad condicional pero como nos indica (Romero Morales, 2003) esta no es capaz de mostrar la independencia estadística entre el antecedente y consecuente, para ello se emplea el *Lift.*

2.4.3.3. *Lift*

El *lift* también traducido como "empuje" es una medida que relaciona los conjuntos de ítems que conforman al antecedente y consecuente sean estadísticamente independientes de la forma (Silverstein et al., 1998):

$$lift(X \Rightarrow Y) = \frac{confianza(X \Rightarrow Y)}{soporte(Y)} = \frac{soporte(X \cup Y)}{soporte(X) \times soporte(Y)}$$

El *lift* es simétrico, por lo que $lift(X \Rightarrow Y) = lift(Y \Rightarrow X)$, un valor de lift mayor a 1 indica una mayor asociación entre sus ítems, mientras que valores menores a 1 pueden indicar su independencia.

2.5. Algoritmos principales de Reglas de Asociación.

A continuación se describen algunos de los algoritmos principales utilizados para la extracción de las reglas de asociación.

2.5.1. Algoritmo Apriori

El algoritmo Apriori se utiliza en minería de datos, sobre bases de datos transaccionales, este permite encontrar de forma eficiente "conjuntos de ítems frecuentes", los cuales sirven para generar reglas de asociación. Debe su nombre a que utiliza conocimiento para generar a priori los *itemsets* frecuentes. Su funcionamiento se resume en 2 pasos:
Primero genera todos los *itemset* que contienen un solo elemento, luego reutiliza estos conjuntos y realiza las combinaciones con 2 elementos y así sucesivamente, tomando todos los pares cuyos *ítems* cumplan con las métricas mínimas de soporte, esto permite eliminar a los que no cumplen y quedarse con los demás candidatos, los que no cumplan no entran al análisis. En segundo lugar se generan las reglas con aquellas que cumplan con el mínimo de confianza. Si un conjunto cumple la condición de soporte y confianza sus subconjuntos también la cumplen, y por el caso contrario, si un conjunto no lo cumple sus superconjuntos tampoco y pueden descartarse (Moya Amaris, M. E., y Rodríguez Rodríguez, J. E., La contribución de las reglas de asociación a la minería de datos, 2003).

2.5.2. Algoritmo FP-Growth

El algoritmo FP-Growth es un método para obtener todos los *itemsets* frecuentes sin la necesidad de generar candidatos (Said et al., 2009). Utiliza una estructura llamada *FP-Tree* (*Frequent Pattern Tree*). En el árbol *FP*, los nodos representan un elemento con su recuento actual y cada rama representa una asociación diferente. El algoritmo construye el árbol leyendo el conjunto que conforma una transacción a la vez y asignando cada transacción a una ruta del árbol de *FP* (Han, Kamber y Pei, 2011). Si varias transacciones tienen elementos en común, estas rutas pueden superponerse, lo que permite obtener una mayor comprensión de la estructura del árbol *FP*. Si el tamaño del árbol generado es tan pequeño que puede permanecer cargado en memoria sin almacenar nada, es posible extraer el conjunto de elementos frecuentes directamente sin tener que hacer pasos extras. Para trabajar sobre el árbol, se comienza con los patrones de longitud uno, se construye su patrón basado en el condicional (set de rutas de prefijo que concurre con el patrón sufijo) y luego se procede a la construcción del árbol *FP* condicional. Se aplica minería de forma recursiva en el árbol (Han et al., 2011) para aumentar el tamaño del patrón concatenando el patrón sufijo con los frecuentes generados a partir de un árbol *FP* condicional (Pérez-Gómez, R. ,2020).

2.5.2. Algoritmo Eclat

El algoritmo Eclat, también conocido como *Equivalence Class Transformation*, emplea un enfoque basado en intersecciones y utiliza una estructura de base de datos vertical para calcular el soporte (Goethals 2003). Para ello, transforma la base de datos transaccional en una matriz llamada "matriz de incidencia", en la que cada fila representa un elemento y cada columna una transacción. Cada elemento debe indicar su presencia o ausencia en la transacción, es decir, debe ser binario o booleano. Eclat busca conjuntos frecuentes de manera recursiva, comenzando por considerar cada elemento como un conjunto frecuente y luego explorando las combinaciones de elementos para encontrar conjuntos de mayor dimensión.

Este algoritmo es especialmente útil en bases de datos grandes, donde hay una mayor densidad de conjuntos frecuentes, debido a su eficiencia en la búsqueda de los conjuntos frecuentes más relevantes a gran velocidad.

Finalizando con los algoritmos principales para la extracción de reglas de asociación el siguiente apartado aborda conceptos del análisis de componentes principales, una técnica ampliamente utilizada en análisis de datos.

2.6. Análisis de Componentes Principales (PCA)

El análisis de componentes principales PCA (*Principal Components Analysis*) es una técnica que se utiliza básicamente para hallar una serie de vectores ortogonales que expliquen de forma más eficiente las varianzas de las variables observadas, el objetivo es realizar una reducción de la dimensión del conjunto de datos conservando la variación en los mismos (Chávez Chong, Sánchez García, 2015). Para lograr esta reducción de dimensión se realizan las combinaciones lineales de las variables originales proyectando la máxima varianza de los datos, esto permite conservar la mayor cantidad de información en una menor cantidad de componentes, facilitando así su observación y análisis (Sánchez Mangas, A. 2012).

En resumen, las enfermedades respiratorias son un problema de salud pública mundial. Las infecciones virales son la causa principal de las infecciones respiratorias agudas en niños menores de 5 años. Además, la contaminación ambiental y sanitaria también puede contribuir a las afecciones respiratorias.

Este capítulo explora las técnicas de minería de datos utilizadas para analizar los datos de enfermedades respiratorias. El algoritmo apriori es el que genera un mayor número de reglas y es el más conocido.

Las variables de interés corresponden al número de casos o de diagnosticados con enfermedades respiratorias en la comuna de Copiapó, las variables ambientales son humedad, rocío, temperatura, temperatura máxima y mínima, finalmente las variables contaminantes son MP10 y MP2.5.

Capítulo III Estado del Arte

En este capítulo se da a conocer el estado del arte de esta investigación:

1. En el trabajo realizado por Luis Gonzalo Espejo Tapia (2022) denominado "Análisis del comportamiento de las enfermedades respiratorias en la comuna de Copiapó, utilizando algoritmos de clustering." donde demuestra que efectivamente se puede analizar el comportamiento de las enfermedades respiratorias con respecto a las variables ambientales y de contaminación, centrado en el año 2019 en la ciudad de Copiapó, utilizando minería de datos y técnicas de agrupamiento.

2. En la investigación hecha por Görkem Sarıyer, Ceren Öcal Taşar en (2020) titulado "*Highlighting the rules between diagnosis types and laboratory diagnostic tests for patients of an emergency department*: *Use of association rule mining*" en donde utilizan reglas de asociación en pacientes diagnosticados con diferentes enfermedades para determinar el uso óptimo de las pruebas de diagnóstico hechas, por motivo de costo y tiempo de estas, generando un trabajo que servirá como apoyo a las decisiones en el ámbito médico.

3. El estudio titulado "*A systematic review of data mining and machine learning for air pollution epidemiology*" y realizado por Bellinger, Colin Mohomed Jabbar, Mohomed Shazan Zaïane, Osmar Osornio-Vargas, Alvaro, si bien no estudia directamente enfermedades respiratorias, utiliza los mismos conceptos tecnológicos (minería de datos) para identificar factores de riesgo epidemiológicos pulmonares que afectan a embarazadas en la ciudad de Manizales, Colombia.

4. El trabajo realizado por Rojas, Gutiérrez, Erika Andrea Rojas Juan Sebastián Aguilar, denominado "Minería de Datos para el Descubrimiento de Patrones en Enfermedades Respiratorias en Bogotá, Colombia", propone descubrir patrones de enfermedades respiratoria en la ciudad de Bogotá, utilizando técnicas de

minería de datos. De esta manera se pudo agrupar (*Clúster*) e identificar, por ejemplo, que en el año 2014 hubo un aumento en los diagnosticados con ASMA, siendo el género masculino entre 4 y 5 años el más afectado.

5. Y por último, a nivel nacional encontramos el proyecto FONDEF de investigación y desarrollo denominado "Desarrollo y Evaluación de algoritmos de Data Mining para la predicción del Riesgo de Crisis en Pacientes Ambulatorios de un Hospital Pediátrico". Código del proyecto: CA13I10300, y desarrollado por Sebastián Ríos Pérez. Esta investigación es una de las pocas iniciativas nacionales que utiliza estas nuevas tecnologías, tiene como objetivo desarrollar un paquete de medidas tecnologías (algoritmos y sistemas de predicción de riesgos) que permitirán detectar el peligro de crisis respiratorias

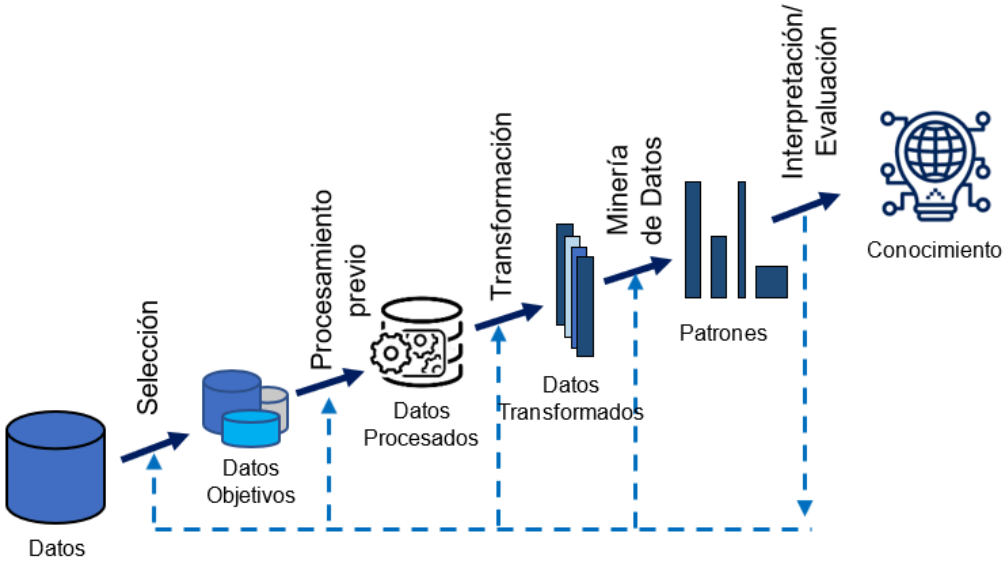
Capítulo IV Metodología

Las metodologías de minería de datos son una guía para llevar a cabo el proceso de descubrimiento de conocimiento de forma sistemática y no trivial. Estas son una ayuda para entender el proceso, la planificación y ejecución de proyectos.

El modelo KDD (*Knowledge Discovery in Databases*) es un modelo de proceso de minería de datos que establece las etapas principales de un proyecto de explotación de información. En la actualidad, el término KDD y minería de datos se utilizan indistintamente para referirse al proceso completo de descubrimiento de conocimiento, a partir del año 2000 surgen nuevas metodologías que plantean un proceso algo más sistémico en donde destaca CRISP-DM (*Cross Industry Standard Process for Data Mining*) llegando a ser una de las más utilizadas (Moine, Haedo & Gordillo, 2011), a continuación se ofrece una pequeña descripción de cada una:

-Knowledge Discovery in Databases: Por sus siglas KDD consiste en utilizar algoritmos de minería de datos para extraer conocimiento de acuerdo con ciertos parámetros especificados sobre una base de datos, el objetivo consiste en encontrar patrones en los datos y finalmente validar la información obtenida (Mejía, J. C. G., 2019), ver figura 4.1.

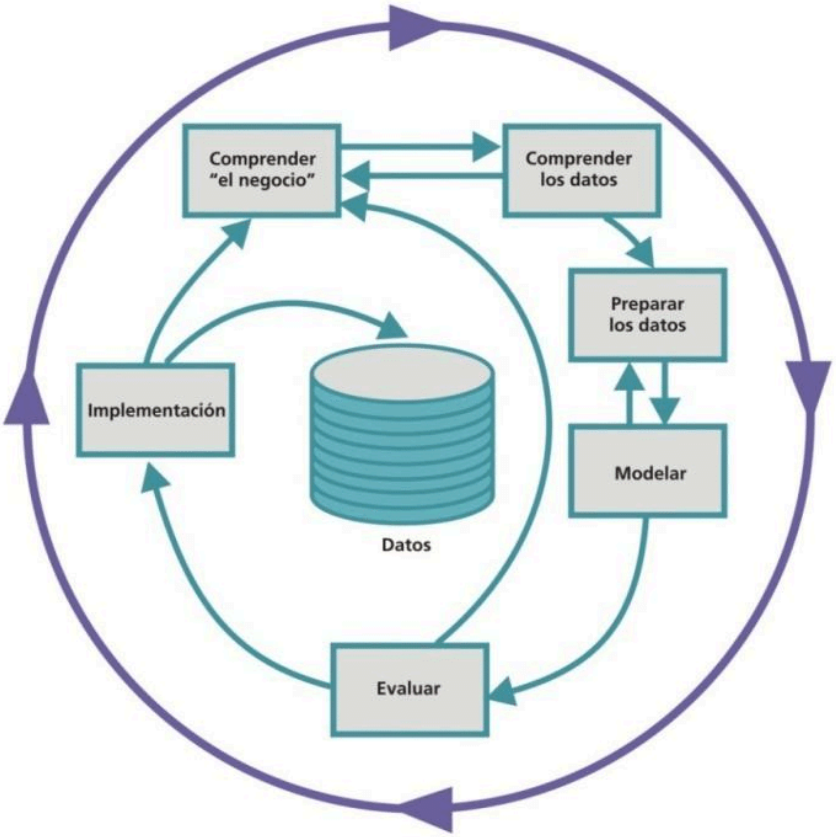
Figura 4.1: Proceso KDD



Fuente: Zambrano, J.-C., Quiroz-Palma, P., Santamaría-Philco, A., & Zamora, W. (2022). "Covid-19 en Ecuador: Aplicación de minería de datos. Informática Y Sistemas"

-Cross Industry Standard Process for Data Mining: Por sus siglas CRISP-DM es un marco de trabajo compuesto de un modelo y una guía, está organizado en seis etapas distintas. Algunas de estas etapas permiten una retroalimentación bidireccional, lo que significa que se puede regresar a una fase anterior para realizar revisiones. La secuencia de las fases no sigue un orden lineal necesariamente desde la primera hasta la última (Cortina, V. G. 2015). Las fases se encuentran en la figura 4.2 a continuación:

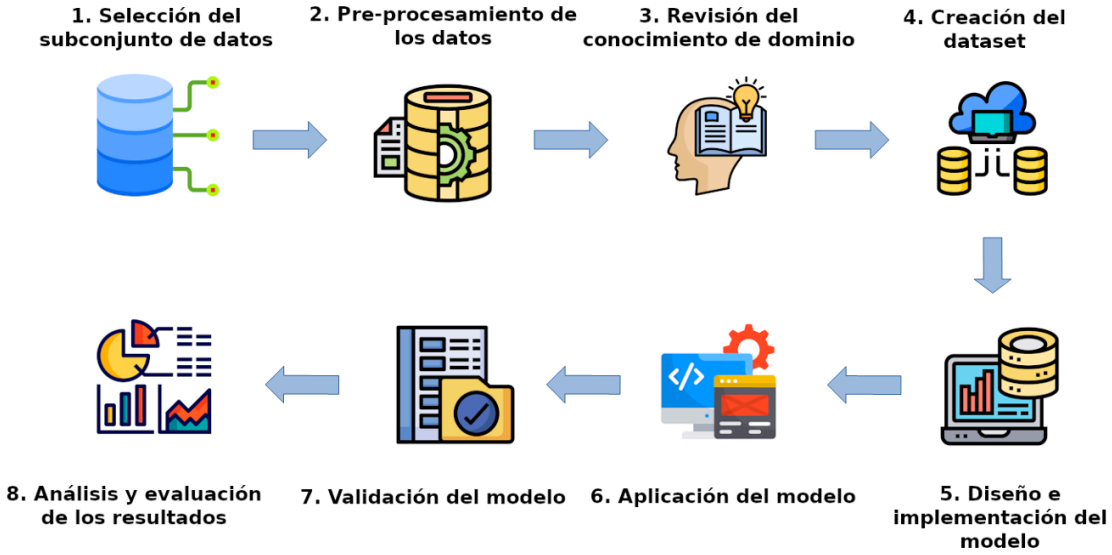
Figura 4.2: Proceso CRISP-DM



Fuente: Yanina Bellini S. Martín V. Santiago B. Romina M. (2014).

Basado en CRISP-DM y KDD se propone una metodología de 8 pasos o etapas:: 1. Selección del subconjunto de datos, 2. Pre-procesamiento de los datos, 3. Revisión del conocimiento de dominio. 4. Creación del *dataset*, 5. Diseño e implementación del modelo, 6. Aplicación del modelo, 7. Validación del modelo, 8. Análisis y evaluación de los resultados. En la figura 4.3 se puede apreciar cada una de las etapas que componen el proceso.

Figura 4.3: Metodología propuesta



Fuente: Elaboración propia.

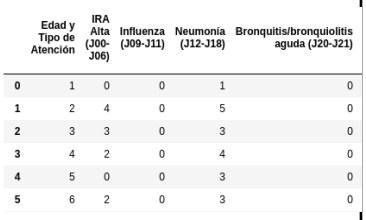
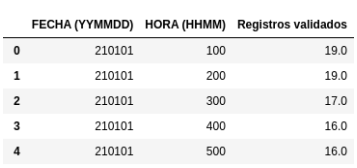
A continuación se detalla cada una de las etapas y el proceso que se llevó a cabo en ellas.

4.1. Selección del subconjunto de datos

La primera etapa consiste en buscar y almacenar los datos necesarios para un análisis y entendimiento previo, que permita a posterior realizar los tratamientos deseados (Kimball, R. & Caserta, J. 2004).

En la tabla 4.1 se puede apreciar la fuente de donde se obtuvieron los datos y las variables de interés para cada fuente para luego indicar cuáles fueron las variables de interés seleccionadas .

Tabla 4.1: Fuentes de Datos

Fuente de datos	Variables de interés	Representación
Hospital Regional de Copiapó	Enfermedades respiratorias	
Dirección General de Aeronáutica Civil (DGAC)	Variables ambientales	
Sistema de Información Nacional de Calidad del Aire (SINCA)	Variables contaminantes	

Enfermedades respiratorias: El estudio utiliza 9 variables para analizar las enfermedades respiratorias, que corresponden a la cantidad de personas diagnosticadas por semana en el hospital Regional de Copiapó. Estas variables son:

- Infección respiratoria aguda alta (IRA_Alta)
- Influenza
- Neumonía
- Bronquitis Bronquiolitis (Bronquitis_bronquiolitis)
- Crisis obstructiva bronquial (Crisis_obstructiva_bronquial)
- Otra causa del sistema respiratorio (Otra_causa_respiratoria)
- Causas del sistema respiratorio (CAUSAS_SISTEMA_RESPIRATORIO)
- COVID-19 confirmado en urgencia (COVID19_confirmado_u)
- COVID-19 confirmado hospitalizado (COVID19_confirmado_h)

Los datos se encuentran agrupados por año y rango de edades. Los años correspondientes son 2017-2021 y los rangos de edades son:

- Todas las edades
- Menores de 1 año
- Niños de 1 a 4 años
- Niños de 5 a 14 años
- Adultos de 15 a 64 años
- Adultos mayores de 65 años

Variables medioambientales: Las variables medioambientales se obtienen de los repositorios online de la Dirección General de Aeronáutica Civil DGAC. Son 5 en total:

- Humedad
- Rocío
- Temperatura
- Temperatura mínima
- Temperatura máxima

Las variables contaminantes: se obtienen de los repositorios online del Sistema de Información Nacional de Calidad del Aire SINCA. Son 2 en total:

- Material particulado grueso de 10 micras (MP10)
- Material particulado fino de 2.5 micras (MP2.5)

4.2. Pre-procesamiento de los datos.

Una vez seleccionadas y almacenadas correctamente los datos recopilados de las distintas fuentes, se procede a realizar un análisis exploratorio de estos para comprobar su integridad, como es habitual en minería de datos comenzaremos haciendo una revisión de manera visual de cada una de las filas que componen el almacén en busca de datos faltantes, ruido o datos atípicos (Shafique, U., & Qaiser, H. 2014), a continuación

se presentan algunas acciones que podemos tomar en caso de encontrar estos datos faltantes o perdidos.

4.2.1. Datos perdidos

En las grandes bases de datos es posible encontrar datos perdidos o faltantes, esto puede ocurrir por varios motivos, algunos habituales son: cuando surgen errores durante la recolección, por problemas técnicos, u omisión de respuestas emitidas durante en entrevistas o encuestas. Lidiar con este tipo de problemas es algo habitual en análisis de datos y se debe recurrir a diversas estrategias, entre las más utilizadas se encuentran:

1. Ignorar o eliminar: en ocasiones se tiene un pequeño número de datos faltantes en comparación con el tamaño de la base y prescindir de ellos no tiene gran influencia en los resultados, entonces estos se pueden despreciar eliminando esos registros (Dagnino, J. 2014).
2. Método de la media: Para reemplazar los valores faltantes, el método de la media reemplaza cada valor faltante por la media de los valores disponibles (Rosas et al., 2009).
3. El método del vecino más cercano: Este método reemplaza los valores faltantes por el valor de la variable más cercana en una variable auxiliar. En el caso de que haya varios valores equidistantes, el método elige uno de ellos al azar (Rosas et al., 2009).
4. Imputación de los datos: La imputación de los datos consiste en generar modelos matemáticos con el propósito de predecir o estimar los datos faltantes que completarán los registros (Allison, P. D. 2001).

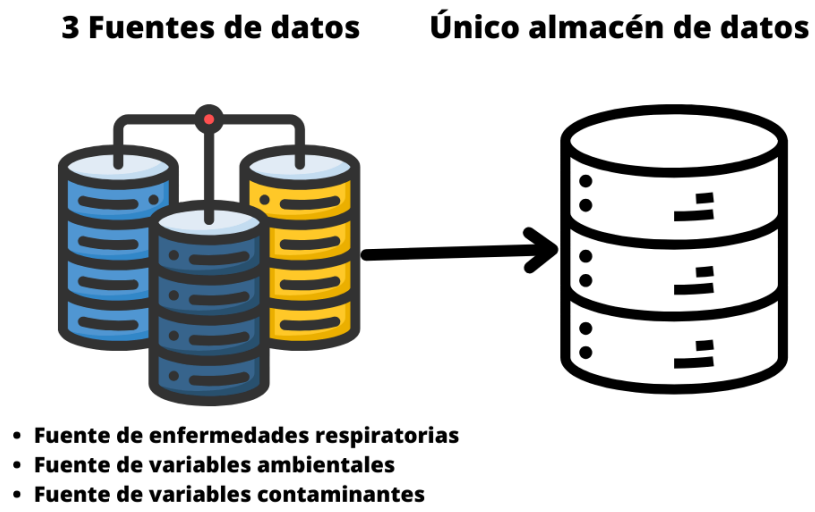
En algunas de las fuentes de datos utilizadas se detectaron datos faltantes, específicamente en las variables medioambientales los cuales debieron ser completados para asegurar la integridad y mejorar los análisis, para ello se utilizó el método de la media en conjunto con de los vecinos más cercanos, es decir, se calculó la media de los registros aledaños para completar los registros faltantes, teniendo así un mínimo impacto en el cálculo de la varianza, esto debido a que los registros por su naturaleza no aleatoria

(estos provienen de los registros validados de una estación meteorológica) (Allison, P. D. 2001), posterior a esto se debe crear una estructura de datos que reúna todas las variables de interés en un solo lugar para así facilitar su almacenamiento. En la próxima sección se detalla este proceso.

4.2.2.Creación de la Estructura de datos

Un *DataFrame* es una estructura de datos bidimensional flexible que se compone de filas y columnas etiquetadas (Petersohn et al., 2020). Es una herramienta eficaz para representar y manipular datos tabulares, ya que permite realizar operaciones aritméticas de forma automática. Se puede pensar en un DataFrame como un diccionario de datos, donde cada fila representa un registro y cada columna representa un atributo.
Para realizar un análisis más eficaz, se creó un DataFrame que reúne a las 3 fuentes de datos en una sola estructura, ver Figura 4.4.

Figura 4.4: Estructura de datos



Fuente: Creación propia.

Para crear un DataFrame con sentido, fue necesario limpiar las fuentes de datos y alinear la granularidad de las mismas. La granularidad de las fuentes de datos se define como el nivel de detalle de los datos (Cravero et al., 2013). En este caso, la fuente de datos con la menor granularidad era la del hospital regional, que representa los datos como "casos

por semana". Por lo tanto, fue necesario ajustar las otras fuentes de datos al promedio por semana para que todas las fuentes tuvieran la misma granularidad.

Finalmente, se consiguió una estructura de 16 columnas que corresponden a 9 tipos de enfermedades respiratorias, 5 variables ambientales y 2 variables contaminantes. Las 261 filas representan las semanas en las que se realiza el registro, desde la primera semana de enero de 2017 hasta la última de diciembre de 2021, para un total de 5 años seguidos.

A continuación podemos apreciar la información del DataFrame con el método info() de la librería Pandas de Python que proporciona una interfaz intuitiva para trabajar con *DataFrames*, lo que facilita el análisis y la manipulación de grandes conjuntos de datos (Wu, Y., 2020)., ver Figura 4.5.

Figura 4.5: Información Dataframe

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 261 entries, 0 to 260
Data columns (total 16 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   IRA_Alta                      261 non-null    float64
 1   Influenza                     261 non-null    float64
 2   Neumonia                      261 non-null    float64
 3   Bronquitis_bronquiolitis      261 non-null    float64
 4   Crisis_obstructiva_bronquial  261 non-null    float64
 5   Otra_causa_respiratoria       261 non-null    float64
 6   CAUSAS_SISTEMA_RESPIRATORIO   261 non-null    float64
 7   COVID19_Confirmado_u          261 non-null    float64
 8   COVID19_Confirmado_h          261 non-null    float64
 9   Humedad                       261 non-null    float64
 10  Rocio                         261 non-null    float64
 11  Temperatura                   261 non-null    float64
 12  Temperatura_min               261 non-null    float64
 13  Temperatura_max               261 non-null    float64
 14  MP10                          261 non-null    float64
 15  MP2.5                         261 non-null    float64
dtypes: float64(16)
memory usage: 32.8 KB
```

Fuente: Creación propia.

Con el almacén de datos ya creado, podemos realizar operaciones con mayor facilidad y dar inicio a la siguiente sub etapa: el análisis estadístico descriptivo.

4.2.3. Análisis estadístico descriptivo.

Para comprender mejor las características de las variables de este estudio, se utilizaron técnicas de estadística descriptiva, como la media y la varianza (Aroca et al., 2009). Estas técnicas permiten calcular medidas de tendencia central y dispersión, que ayuda a describir la distribución de los datos, ver figura 4.6.

Figura 4.6: Media y Varianza de cada variable

```
----------------------
Media de cada variable
----------------------
IRA_Alta                        95.268199
Influenza                        7.747126
Neumonia                         8.812261
Bronquitis_bronquiolitis        37.486590
Crisis_obstructiva_bronquial    20.724138
Otra_causa_respiratoria         12.134100
CAUSAS_SISTEMA_RESPIRATORIO      9.965517
COVID19_Confirmado_u             4.996169
COVID19_Confirmado_h             2.735632
Humedad                         66.502299
Rocio                            9.253640
Temperatura                     16.304981
Temperatura_min                 11.306513
Temperatura_max                 24.459770
MP10                            39.676628
MP2.5                           12.868199
dtype: float64
```

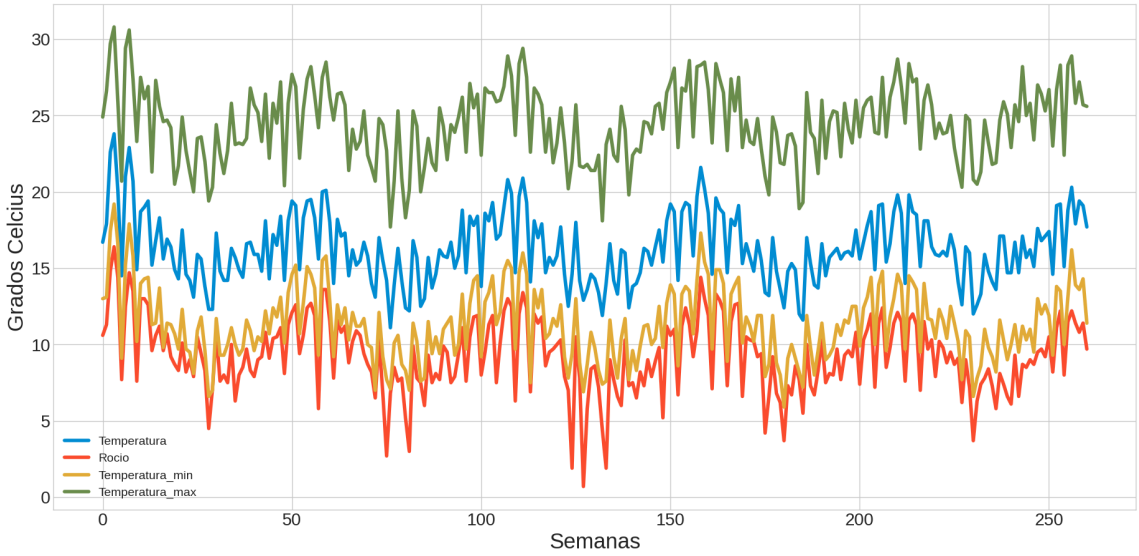
```
------------------------
Varianza de cada variable
------------------------
IRA_Alta                        3009.889331
Influenza                        189.735809
Neumonia                          41.253080
Bronquitis_bronquiolitis        1412.327704
Crisis_obstructiva_bronquial     212.308223
Otra_causa_respiratoria           32.870410
CAUSAS_SISTEMA_RESPIRATORIO       35.733422
COVID19_Confirmado_u             115.719216
COVID19_Confirmado_h              35.733687
Humedad                           15.946149
Rocio                              5.934573
Temperatura                        5.171244
Temperatura_min                    5.704073
Temperatura_max                    6.193260
MP10                             187.217106
MP2.5                             21.322331
dtype: float64
```

Fuente: Creación propia.

Este análisis abarca un periodo de cinco años, desde 2017 hasta 2021. Con la información recopilada, se facilita la comparación y contrastación con otras fuentes externas, con el propósito de verificar que la dirección tomada es positiva. Por ejemplo, en este análisis la media de la temperatura es de aproximadamente 16,3 °C. Según la Biblioteca del Congreso Nacional, en la sección "Clima y Vegetación Región de Atacama" (https://www.bcn.cl/siit/nuestropais/region3/clima.htm), la temperatura media de Copiapó es de 15°C, clasificando la zona como de clima desértico marginal. Además según (Infante Amunátegui et al. 2015), la temperatura media anual de Copiapó es de 16°C, cifra aún más cercana. En cualquier caso, el valor obtenido sigue estando dentro del rango esperado. Además el trabajo realizado por (Gómez Sarria et al. 2014) expone que, Copiapó como ciudad-oasis, incluye islas de calor y áreas de fresco urbano, y estos fenómenos afectan las temperaturas, ya sea aumentándolas o disminuyéndolas. Esto se

refleja en las temperaturas observadas en el gráfico de la figura 4.7, con una máxima de 30.8 °C y una mínima de 5.9 °C.

Figura 4.7: Gráfico lineal Temperatura, Rocio, Temperatura mínima y Temperatura máxima



Fuente: Creación propia.

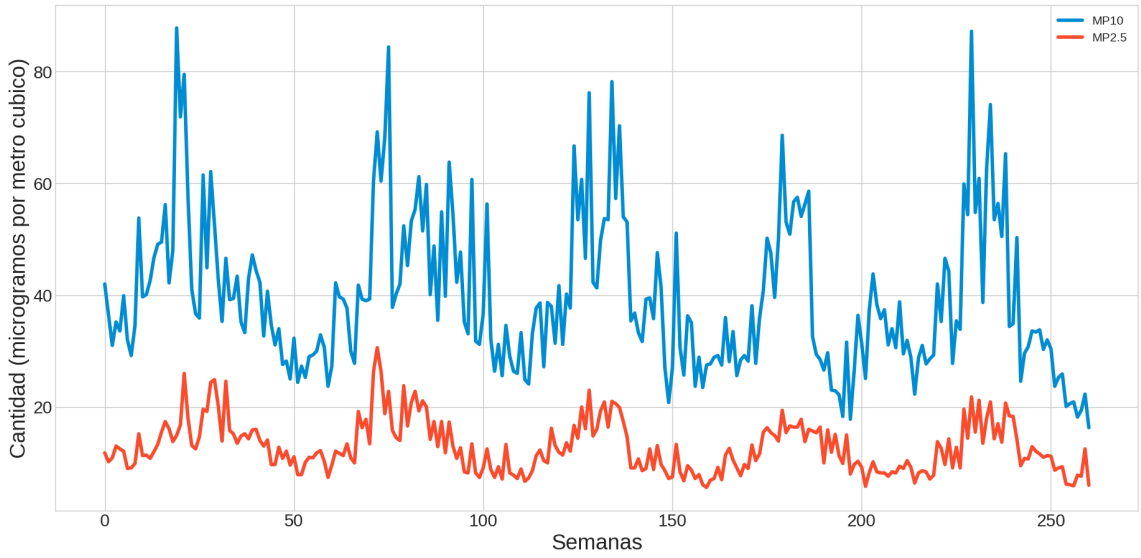
En el caso de la variable contaminantes MP 2.5, el Decreto de Ley N°12 de Chile define los niveles que determinan situaciones de emergencia, en él se establece la concentración de MP 2.5 en un intervalo de 24 horas, mientras que el Decreto de Ley 59 de igual manera establece los niveles respirables para el material contaminante de 10 micras MP 10, los rangos están definidos en la tabla 4.2. En este caso, tanto para la variable MP 2.5 como para MP 10, no se sobrepasa el nivel de preemergencia o el nivel 1°, respectivamente. Esto indica que se mantiene un nivel respirable bueno o aceptable, de acuerdo con lo observado, en la Figura 4.8.

Tabla 4.2: Niveles de contaminación por decreto ley

Variable contaminante	**Nivel**	**Rango**
MP 2.5	Alerta	80-109 (µg/m3)
	Preemergencia	110-169 (µg/m3)
	Emergencia	170 (µg/m3) o superior

Variable contaminante	Nivel	Rango
MP 10	Nivel 1°	195-239 (µg/m3)
	Nivel 2°	240-329 (µg/m3)
	Nivel 3°	330 (µg/m3) o superior

Figura 4.8: Gráfico lineal Variables contaminantes



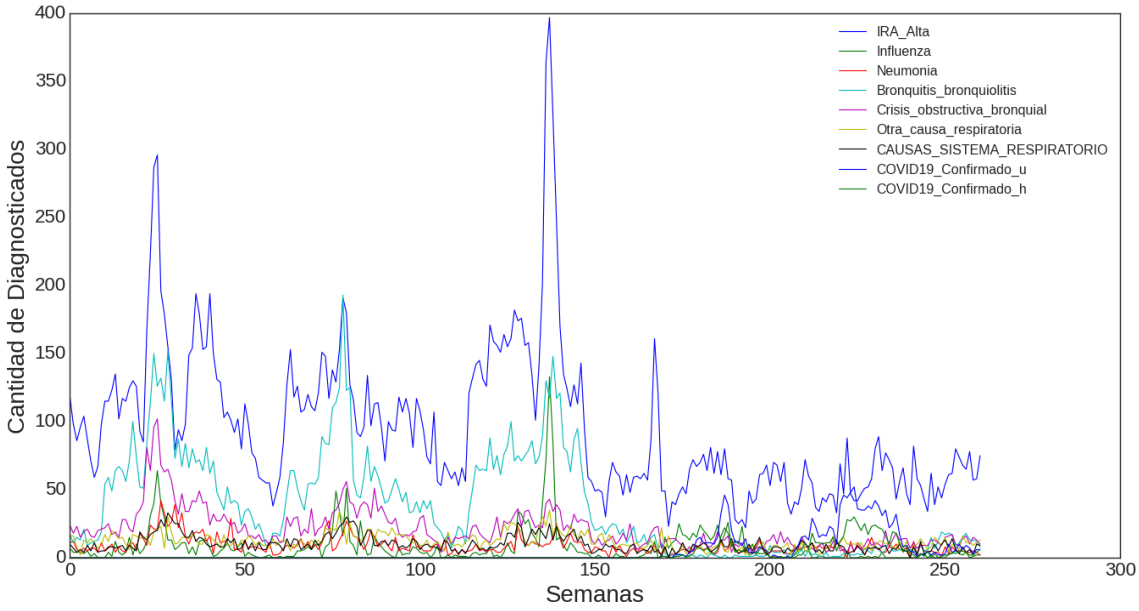
Fuente: Creación propia.

Respecto a las enfermedades respiratorias, la Figura 4.9 destaca un notable aumento en los diagnósticos de IRA Alta durante la semana 137 (12 de agosto de 2019). La variable, representada en color azul, registra un total de 397 casos, siendo la más destacada del gráfico.

El informe del "Boletín Epidemiológico Trimestral: Influenza, SE 1 – 39, Año 2019" proporciona información detallada sobre el aumento de casos de infecciones respiratorias en Copiapó durante el período de 2019. El informe describe el aumento de la transmisibilidad de la influenza, la notificación de Enfermedad Tipo Influenza (ETI) y la vigilancia de casos de Infección Respiratoria Aguda Grave (IRAG) en varios hospitales centinelas. El informe destaca el aumento de la transmisibilidad en grupos escolares, la detección de influenza A y B, y la importancia de mantener la vigilancia de

la influenza a través del monitoreo semanal. Estos hallazgos corroboran el aumento de casos de infección respiratoria aguda observado en la figura 4.9.

Figura 4.9: Gráfico lineal Enfermedades respiratorias



Fuente: Creación propia.

El análisis estadístico de este primer estudio muestra una coincidencia significativa con los informes y fuentes citados. Por lo tanto, se considera que los datos utilizados en esta investigación son fiables. La siguiente sub-etapa es el análisis de componentes principales.

4.2.4. Análisis de componentes principales

Y por último como parte del Pre-procesamiento de los datos, el análisis de componentes principales es una técnica utilizada para proporcionar una explicación más eficiente de las variaciones presentes en las variables (Karamizadeh et al., 2013), con el propósito de reducir su dimensión y obtener una visión preliminar de su comportamiento en el conjunto de datos, permitiendo identificar patrones y tendencias en los datos, facilitando evaluar visualmente similitudes y diferencias entre muestras y determinar si las muestras pueden ser agrupadas (Ringnér 2008),

Para realizar un análisis de componentes principales, el primer paso consiste en normalizar los datos. Este proceso se realiza mediante el método "fit_transform" de la biblioteca Sklearn de Python, el cual convierte los registros en valores normalizados. Es decir, se establece una media de 0 y una desviación estándar de 1. Al utilizar el método "*describe*", podemos corroborar esta transformación, como se muestra en la figura 4.10. En dicha figura, se pueden observar valores promedio muy cercanos a cero en el registro "*mean*" (media), así como desviaciones estándar cercanas a uno en el registro de "*std*" (*standard deviation*).

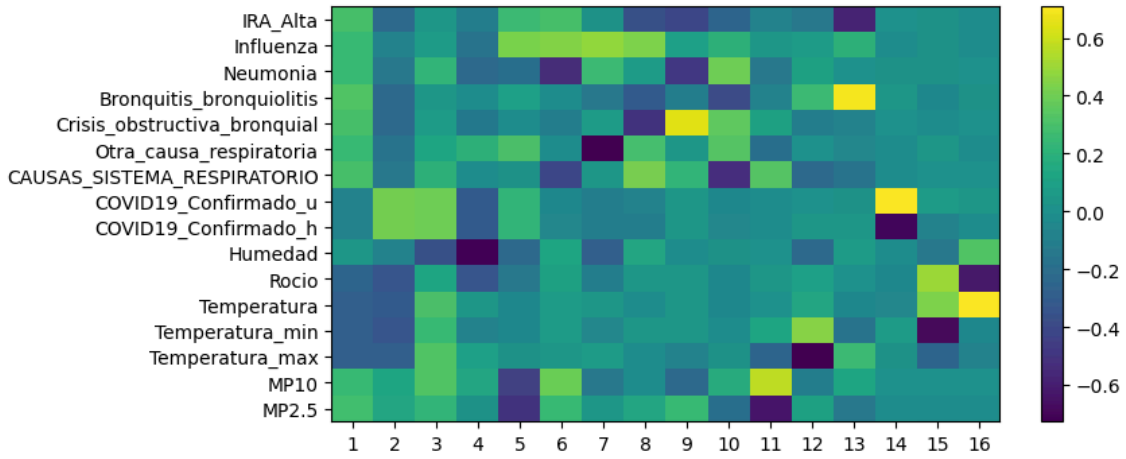
Figura 4.10: Descripción Dataframe normalizado

	IRA_Alta	Influenza	Neumonia	Bronquitis_bronquiolitis	Crisis_obstructiva_bronquial	Otra_causa_respiratoria
count	2.610000e+02	2.610000e+02	2.610000e+02	2.610000e+02	2.610000e+02	2.610000e+02
mean	5.444772e-17	3.062684e-17	5.444772e-17	-5.444772e-17	-1.088954e-16	-1.088954e-16
std	1.001921e+00	1.001921e+00	1.001921e+00	1.001921e+00	1.001921e+00	1.001921e+00
min	-1.338055e+00	-5.635073e-01	-1.374652e+00	-9.994056e-01	-1.218752e+00	-1.770990e+00
25%	-6.988699e-01	-4.907697e-01	-5.946863e-01	-8.394435e-01	-6.686537e-01	-5.477014e-01
50%	-2.788343e-01	-3.452946e-01	-2.827001e-01	-4.128779e-01	-3.248424e-01	-1.981902e-01
75%	4.699249e-01	1.638686e-01	3.412723e-01	7.068568e-01	4.315425e-01	5.008320e-01
max	5.510353e+00	9.110593e+00	5.177058e+00	4.146042e+00	5.588712e+00	3.995943e+00

Fuente: Creación propia.

El siguiente paso es, crear las componentes principales para todas las variables involucradas y cuyos coeficientes se representan a través de un mapa de calor que facilita su observación, ver Figura 4.11.

Figura 4.11: Mapa de calor de todas las componentes principales

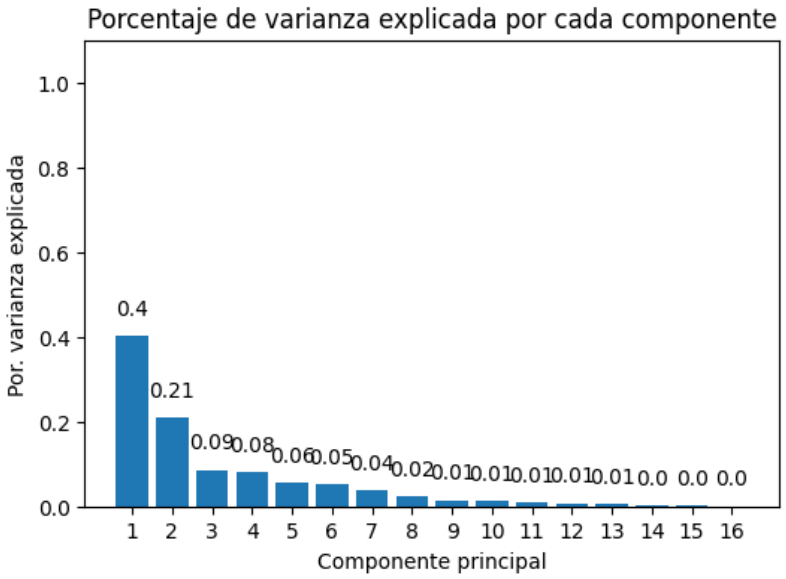


Fuente: creación propia.

Este mapa nos permite observar el nivel de representación que tiene cada variable en cada una de las componentes generadas, donde los colores más cercanos al amarillo indican una mayor representacion que tiene una variable en la componente que pertenece (las cuales comianzan de izquierda a derecha representadas por los numeros del 1 al 16) y los mas cercanos al azul indican lo contrario, es decir, una menor representacion, por ejemplo la variable humedad en la cuarta componente en color azul con una representacion en torno al -0.6.

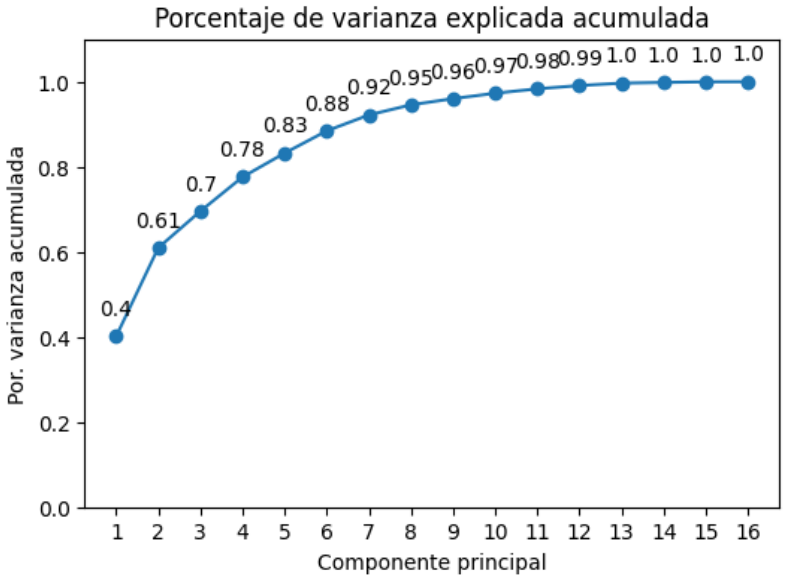
Con estas componentes previamente generadas podemos determinar el número de componentes principales que se ajustan mejor a el propósito de la investigación, para ello calculamos la varianza acumulada de cada componente y utilizamos la cantidad de componentes que expliquen el máximo de varianza acumulada y que al mismo tiempo sean las mínimas posibles, con esto en mente si se observa la figura 4.13 podemos notar que con las primeras 4 componentes se tiene una varianza total acumulada que explica el 78% de la colección de datos, lo cual es aceptable para el propósito de esta investigación.

Figura 4.12: Varianza explicada de cada componente



Fuente: Creación propia.

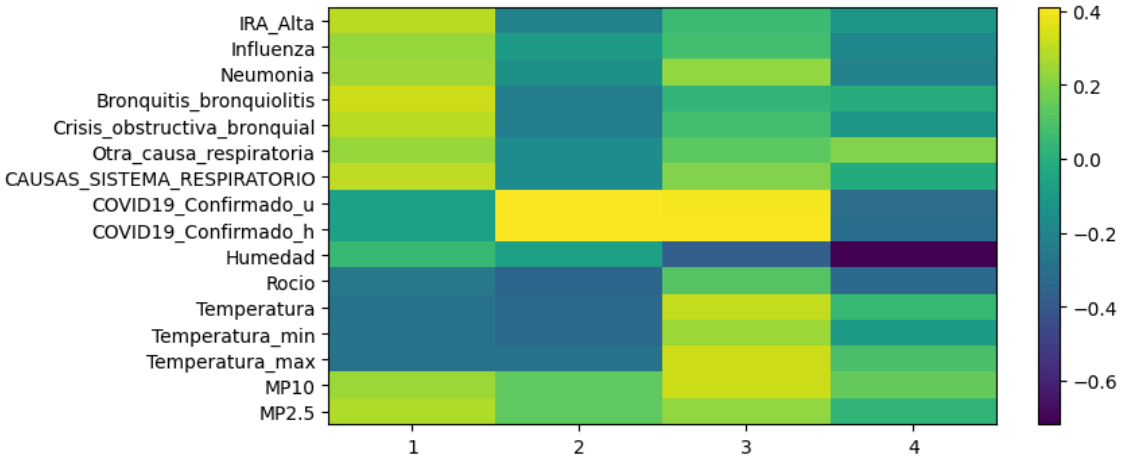
Figura 4.13: Varianza explicada acumulada de cada componente



Fuente: Creación propia.

Por otra parte , en la Figura 4.14 se observa un mapa de calor correspondiente a 4 componentes con una varianza acumulada que explica el 78% del total de los datos. Este mapa permite ver con mayor claridad la representación que tienen las variables en cada componente. Por ejemplo, las enfermedades tienen un nivel de representación en la primera componente, a excepción de las 2 de Covid (COVID19_Confirmado_u y COVID19_Confirmado_h). Sin embargo, estas últimas 2 variables tienen un comportamiento semejante en todas las componentes (casi el mismo color en el mapa). Lo mismo ocurre con las 2 variables contaminantes (MP10 y MP2.5), que comparten una representación parecida en todas las 4 componentes. Este comportamiento es acorde a lo esperado.

Figura 4.14: Mapa de calor primeras 4 componentes



Fuente: Creación propia.

Los resultados del análisis confirman la confiabilidad de las fuentes de datos. Esto basado en la comparación satisfactoria de los resultados con las fuentes citadas, así como en el comportamiento observado de las variables en el mapa de calor. Es importante destacar que los registros de variables contaminantes se encuentran por debajo de los niveles mínimos de emergencia. En cuanto a las enfermedades respiratorias, la IRA alta y la bronquitis presentan la mayor varianza, lo que las convierte en las más volátiles. La siguiente etapa de la metodología se centrará en la revisión del conocimiento de dominio.

4.3. Revisión de conocimiento de dominio

Esta etapa tiene como objetivo comprender el contexto y las particularidades del problema a resolver. Por ello, se realiza un estudio comparativo de metodologías y procesos empleados en proyectos de investigación similares. El objetivo es buscar soluciones basadas en experiencias previas, además de identificar el algoritmo que mejor se ajusta a las características específicas de esta investigación.

Para conocer más sobre la extracción de reglas de asociación se reviso el trabajo realizado por (Kang et al., 2014), donde se extraen reglas de asociación mediante un proceso iterativo de ejecución y análisis, a partir de una base de datos de tipo transaccional, en donde los items corresponden a características de interés de los

pacientes, tales como sexo, edad, tipo de diagnóstico. Dichas características se utilizaron como antecedentes al determinar la probabilidad de padecer la afección estudiada. Se utilizó un soporte del 10% y una confianza mínima del 50% para identificar las reglas más relevantes. Posteriormente, se compararon las reglas con datos demográficos para identificar asociaciones significativas. Para ello, se utilizaron medidas de tendencia central para clasificar los datos en diferentes categorías.

A diferencia del trabajo mencionado en el párrafo anterior, que utiliza una base de datos de tipo transaccional, la colección de datos de este trabajo contiene atributos de tipo float. Por lo tanto, es necesario convertir primero los datos a un formato compatible con los algoritmos de extracción de reglas de asociación.

Del mismo modo, para el proceso de discretización, necesario para obtener las reglas de asociación se analizó el trabajo realizado por Ramon Sangüesa, que expone el convertir variables a través de un proceso de discretización el cual define como *"establecer un criterio por medio del cual se puedan dividir los valores de un atributo en dos o más conjuntos disjuntos"* (Ramon Sangüesa, 2019, p.23), que puede ser de dos tipos: Supervisada y no supervisada, a continuación, se ofrece una definición de cada uno de estos tipos:

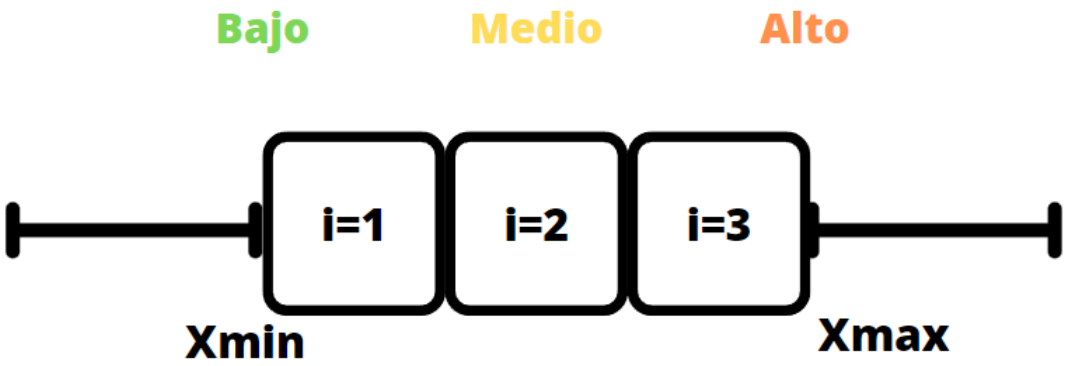
> **Discretización supervisada:** Es un método que utiliza las etiquetas de clase para encontrar los límites de los intervalos de discretización. La discretización supervisada es una de las formas en que se pueden clasificar los métodos de discretización (Dougherty, et al., 1995).
>
> **Discretización no supervisada:** La discretización no supervisada es un método de discretización que no utiliza las etiquetas de clase para encontrar los límites de los intervalos de discretización. Algunos ejemplos de métodos de discretización no supervisados son la discretización por igual frecuencia y la discretización por igual anchura (Dougherty, et al., 1995).

Discretización por intervalos de igual ancho: Es un método no supervisado que se utiliza para producir valores nominales a partir de variables continuas. Consiste en ordenar los valores observados de una característica continua y dividir el rango de valores observados en *k* intervalos de igual tamaño, donde *k* es un parámetro proporcionado por el usuario (Dougherty, et al., 1995). Si una variable X tiene valores limitados por Xmín y Xmax, este método construye lo *k* intervalos como sigue:

$$\delta_i = \frac{Xmax - Xmin}{k}$$

En este método, cada intervalo se calcula sumando k veces a Xmin hasta completar el ancho total del intervalo (ver Figura 4.15).

Figura 4.15: Método de intervalos de igual ancho



Fuente: Creación propia.

El método se aplica a cada característica continua de forma independiente. Como no utiliza información de clase, es un método de discretización no supervisado. Esto significa que las categorías se asignan a los datos sin tener en cuenta la clase a la que pertenecen.

Por ejemplo, si se tiene una temperatura que oscila entre 10°C y 25°C, la diferencia es 15°, al dividir ese dato en 3 se tiene 5°, entonces los valores intermedios se ubican en las categorías de temperatura: bajo, para aquellos entre 10° y 15°, medio para aquellos entre 15° y 20° y alto para los que oscilan entre 20° y 25°.

En este contexto, es posible transformar la colección de datos a un sistema de categorías utilizando la discretización por intervalos de igual ancho. Esto permite convertir la colección a un formato similar al de una "cesta de compra", sin embargo, un inconveniente de este método es que algunas características de los datos tienen una gran dispersión. Esto provoca que en la conversión existan intervalos que acaparan la mayor cantidad de registros. Por ejemplo, en la Figura 4.9, el valor máximo de la variable IRA_Alta es muy alto en comparación al resto de registros, generando intervalos cuyos máximos no son aplicables a los registros en todos los periodos de tiempo.

Para evitar este inconveniente se separó la colección de datos en períodos de tiempo de similar estación. Los datos se agruparon por meses: Enero-Febrero, Marzo-Abril, Mayo-Junio, Julio-Agosto, Septiembre-Octubre y Noviembre-Diciembre. Cada período se analizó durante 5 años (2017-2021). De esta forma, se obtuvieron 6 colecciones de datos distintos.

Finalmente para extraer las reglas de asociación se estudia el uso de un algoritmo apropiado para ello, en la tabla 4.3 se observa una comparación de las ventajas y desventajas de los más populares algoritmos para la extracción de reglas de asociación creada a partir de otros estudios comparativos, particularmente Patil et al., (2022) y Heaton, J. (2016).

Tabla 4.3: Comparativa ventajas y desventajas algoritmos

	Apriori	**Eclat**	**FP-Growth**
Fácil de implementar y entender	✅	⬜	⬜
Escanea los datos de manera eficiente	⬜	✅	⬜
No necesita generar candidatos	⬜	⬜	✅
Eficiente y escalable	⬜	✅	✅
Soporte para datos dispersos	✅	⬜	✅

	Apriori	Eclat	FP-Growth
Tiempo de ejecución	**Moderado**	**Rápido**	**Rápido**
Cumple con el criterio :✅ **No cumple con el criterio:🔲**			

Considerando su facilidad de uso, el soporte para datos dispersos y por el dominio del conocimiento del problema, una buena opción es utilizar el algoritmo Apriori para la extracción de reglas de asociación, considerando además que su mayor tiempo de ejecución no representa un inconveniente para la cantidad de datos con los que contamos, ya que no son demasiados registros.
Con esto ya podemos continuar con la próxima etapa de la metodología que consiste en la creación del dataset que será la entrada del algoritmo de extracción de reglas.

4.4. Creación del dataset

Los datasets son una herramienta esencial para este tipo de investigación científica, ya que permiten identificar patrones, realizar investigaciones y tomar decisiones basadas en datos (Guía para la Gestión de Datos de Investigación del Ministerio de Ciencia, Tecnología e Innovación, Minciencias). Los datasets se pueden utilizar para entrenar modelos de aprendizaje automático y evaluar su rendimiento , así como para validar y evaluar modelos (Huang, Z., 1997), además, los datasets pueden presentar diferentes tipos y formatos, adaptándose a las necesidades específicas de cada proyecto.

A partir de la definición y la importancia de esta colección de datos , se crea un dataset utilizando el método de discretización por intervalos de igual ancho descrito en la sección 4.3, a continuación se observa el cambio de cada variable o característica por 3 nuevas de tipo boolean que llevan la leyenda bajo, medio y alto al final, para facilitar su lectura y posterior interpretación de los resultados, la tabla 4.4 muestra el detalle de este cambio.

Tabla 4.4: Conversión a variables categóricas

Variable original	Formato original	Variable nueva	Formato nuevo
Dato_Variable	Float	Dato_Variable_Bajo	Boolean
		Dato_Variable_Medio	Boolean
		Dato_Variable_Alto	Boolean

Además, en esa misma sección se menciona la necesidad de dividir la colección de datos original en periodos de 2 meses para facilitar el análisis y evitar la dispersión de datos, teniendo así 6 colecciones de datos.

Con esto en mente se lleva a cabo la transformación de los datos obteniendo como resultado un conjunto de datos que puede ser usado como *input* en el algoritmo de extracción de reglas de asociación, el cual a partir de ahora será denominado **"dataset model"**, puesto qué es esencial para el modelo propuesto en este proyecto, ver figura 4.16 para un ejemplo real.

Figura 4.16: Ejemplo Dataset model

	IRA_Alta_Bajo	IRA_Alta_Medio	IRA_Alta_Alto	Influenza_Bajo	Influenza_Medio	Influenza_Alto	Neumonia_Bajo	Neumonia_Medio	Neumonia_Alto
0	False	True	False	True	False	False	True	False	False
1	True	False	False	True	False	False	False	False	True
2	False	True	False	True	False	False	True	False	False
3	False	True	False	True	False	False	True	False	False
4	False	True	False	True	False	False	True	False	False
5	False	True	False	False	False	True	True	False	False
6	True	False	False	True	False	False	False	False	True
7	False	True	False	True	False	False	False	True	False
8	False	True	False	True	False	False	True	False	False
9	False	True	False	True	False	False	True	False	False
10	False	True	False	True	False	False	True	False	False
11	False	True	False	True	False	False	True	False	False
12	False	False	True	True	False	False	True	False	False
13	False	True	False	True	False	False	True	False	False
14	False	False	True	True	False	False	True	False	False
15	False	True	False	True	False	False	False	True	False
16	False	True	False	True	False	False	True	False	False
17	False	True	False	True	False	False	False	True	False

Fuente: Creación propia.

Con lo anterior previamente realizado se está en condiciones de pasar a la siguiente etapa.

4.5. Diseño e implementación del modelo

Con los pasos anteriores ya completados sólo resta el diseño e implementación del modelo, para ello se implementó un *script* en *Python* con las librerías: *Pandas, Sklearn, Mlxtend, Matplotlib* para realizar las tareas pertinentes descritas en las secciones 4.1. y 4.2. A continuación se ofrece una breve descripción de cada una de las librerías utilizadas :

Pandas: Es una librería de software escrita para el lenguaje de programación *Python* que se utiliza principalmente para la manipulación y análisis de datos. Ofrece estructuras de datos y operaciones para manipular tablas numéricas y series de tiempo (Stančin, et al., 2019).

Sklearn: Es una librería de aprendizaje automático de software libre para el lenguaje de programación *Python*. Ofrece varios algoritmos de clasificación, regresión y agrupamiento, incluyendo máquinas de vectores de soporte, bosques aleatorios, aumento de gradiente (Stančin, et al., 2019).

Mlxtend: Es una librería de herramientas útiles para las tareas diarias de ciencia de datos y aprendizaje automático en *Python*. MLxtend ofrece funcionalidades adicionales y puede ser una valiosa adición a su kit de herramientas de ciencia de datos. La librería tiene muchas funciones interesantes para el análisis de datos y tareas de aprendizaje automático, la creación de gráficos de correlación PCA, la descomposición de sesgo-varianza, la creación de regiones de decisión para modelos de clasificación, la creación de matrices de gráficos de dispersión y el remuestreo, entre otros (Stančin, et al., 2019).

Matplotlib: Matplotlib es una librería de *Python* de código abierto que permite crear visualizaciones de datos estáticas, animadas e interactivas con poco código. Útil para aquellos que trabajan con NumPy y se utiliza en servidores de aplicaciones web, shells y scripts de *Python*. Permite crear trazados, histogramas, diagramas de barra y otros tipos de gráficas. Algunas de las ventajas de

Matplotlib son que es fácil de aprender, permite el control de cada elemento en una figura, tiene una salida de alta calidad en muchos formatos y es muy personalizable. Matplotlib es una herramienta muy completa que permite generar visualizaciones de datos muy detalladas y es ampliamente utilizada en la comunidad científica y de análisis de datos (Ari, et al,. 2014).

De este modo el modelo realiza la limpieza y transformaciones de los datos originales utilizando la librería Pandas principalmente, posteriormente muestra un análisis estadístico sencillo, para eso incorpora la librería SKlearn y Matplotlib, con ayuda de funciones incorporadas en ellas realiza el análisis de componentes principales y permite desplegar gráficos respectivamente, finalmente para extraer las reglas de asociación utilizamos el algoritmo apriori que se encuentra en la librería Mlxtend. La siguiente etapa trata con más detalle la ejecución del algoritmo Apriori.

4.6. Aplicación del modelo

Esta etapa se centra en la extracción de las reglas de asociación más interesante según el caso de estudio, para ello se aplica el algoritmo apriori sobre el dataset model descrito en el apartado 4.4, es decir, posterior a la discretización de variables y la separación de la colección original por periodo de tiempo, se realiza además una separación de acuerdo a los rangos de edad (menores de 1 año, niños de 1 a 4 años, niños de 5 a 14 años, adultos de 15 a 64 años, adultos mayores de 65 años y todas las edades), ver ejemplo figura 4.17.

Figura 4.17: Ejemplo reglas de asociación

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
187	(Temperatura_Bajo, CAUSAS_SISTEMA_RESPIRATORIO...	(Neumonia_Bajo, Temperatura_min_Bajo)	0.200	0.225	0.200	1.0	4.444444
189	(CAUSAS_SISTEMA_RESPIRATORIO_Bajo, Temperatura...	(Temperatura_Bajo, Neumonia_Bajo)	0.200	0.225	0.200	1.0	4.444444
2	(Rocio_Bajo)	(Temperatura_min_Bajo)	0.200	0.250	0.200	1.0	4.000000
20	(Temperatura_Bajo, Neumonia_Bajo)	(Temperatura_min_Bajo)	0.225	0.250	0.225	1.0	4.000000
48	(Temperatura_Bajo, CAUSAS_SISTEMA_RESPIRATORIO...	(Temperatura_min_Bajo)	0.200	0.250	0.200	1.0	4.000000
...	...	...	...	...	...	...	...
1807	(Otra_causa_respiratoria_Bajo, Temperatura_Med...	(Crisis_obstructiva_bronquial_Bajo, Rocio_Medi...	0.500	0.200	0.200	0.4	2.000000
1823	(Rocio_Medio)	(Otra_causa_respiratoria_Bajo, Influenza_Bajo,...	0.500	0.200	0.200	0.4	2.000000
1844	(Otra_causa_respiratoria_Bajo, Temperatura_Med...	(Rocio_Medio, Influenza_Bajo, CAUSAS_SISTEMA_R...	0.500	0.200	0.200	0.4	2.000000
1943	(Rocio_Medio)	(Otra_causa_respiratoria_Bajo, Temperatura_min...	0.500	0.200	0.200	0.4	2.000000
2032	(Crisis_obstructiva_bronquial_Bajo, Temperatur...	(Neumonia_Bajo, Otra_causa_respiratoria_Bajo, ...	0.500	0.200	0.200	0.4	2.000000

2034 rows × 9 columns

Fuente: Creación propia.

Para evaluar la cantidad y la calidad de las reglas, se analizaron aquellas con soportes entre el 15% y el 30%, ordenándolas por confianza de mayor a menor. Posteriormente, se estableció un soporte mínimo del 20% y un lift de 2 para garantizar que las reglas sean frecuentes y con asociaciones fuertes. Esta decisión se debe básicamente a que el soporte mínimo depende de los objetivos del proyecto y las consideraciones más relevantes (Piatetsky-Shapiro et al., 2000). Por último, se ordenaron las reglas por confianza de mayor a menor, estableciendo un umbral de confianza del 50%. Esto garantiza que las reglas ocurran con una frecuencia bastante alta.

En la siguiente etapa se presenta un método para realizar la validación de estas reglas de asociación entregadas por el modelo.

4.7. Validación del modelo

La validación del modelo se hizo de acuerdo al método del juicio de expertos o también conocido como criterio experto que consiste en obtener la opinión informada de personas con trayectoria en un tema específico, que son reconocidas por otros como expertos cualificados en este, y que pueden dar información, evidencia, juicios y valoraciones. Este método se utiliza para verificar la fiabilidad de una investigación y para obtener una amplia y detallada información sobre el objeto de estudio y la calidad de las respuestas por parte de los jueces (Garrote, P. R., & del Carmen Rojas, M. 2015).

Para llevar a cabo el proceso de validación por parte de los expertos, se realizó un primer filtrado del conjunto de reglas de asociación, obteniendo 211 reglas divididas por períodos de análisis. Posteriormente, un equipo de dos expertas en la materia, la Mg María Paola Vieytes Carrizo y la licenciada en medicina Sonia Ibaceta Lorca, evaluó las reglas filtradas. Con base en su conocimiento y experiencia, determinaron que 27 reglas eran representativas del contexto de la investigación. Es importante mencionar que el segundo filtrado se realizó a ciegas, es decir, las métricas no fueron proporcionadas a las expertas, esto se hizo para que su selección no dependiera de indicadores, sino del contexto específico del estudio.

La siguiente sección muestra los resultados finales y un pequeño análisis de estos.

4.8. Análisis y validación de los resultados

La última etapa del modelo propuesto consiste en el análisis y validación de los resultados.

Tras la selección realizada por los expertos en la sección anterior (segundo filtro, que dio como resultado 27 reglas), se realizó un tercer filtrado. A partir de aquí, se seleccionaron únicamente las reglas con una confianza superior al 80%. Como resultado, se obtuvieron 6 reglas (ver tabla 4.5), divididas en los diferentes períodos de análisis.

La tabla 4.5 proporciona la siguiente información:

- **#:** Número que identifica la regla.
- **Antecedentes, Consecuentes:** Conjuntos de reglas de asociación de variables contaminantes, ambientales y enfermedades descritas en la sección 4.1, posterior a la discretización y categorización.
- **Periodo:** Periodo de tiempo en meses al que pertenece la reglas, en donde la equivalencia es:
    - Ene-Feb: Periodo entre Enero y Febrero.
    - Mar-Abr: Periodo entre Marzo y Abril.
    - May-Jun: Periodo entre Mayo y Junio.
    - Jul-Ago: Periodo entre Julio y Agosto.
    - Sep-Oct: Periodo entre Septiembre y Octubre.
    - Nov-Dic: Periodo entre Noviembre y Diciembre.
- **Rango:** Corresponde al rango de edades:
    - Todos: Para todas las edades.
    - < 1: Para los niños menores de 1 año.
    - 1 a 4: Para los niños entre 1 a 4 años.
    - 5 a 14: Para los niños entre 5 a 14 años.
    - 15 a 64: Para los adultos entre 15 a 64 años.
    - 65 y más: Para los adultos entre mayores de 65 años.

Tabla 4.5: Resultados reglas de asociación

#	Antecedentes		Consecuentes	Periodo	Rango
1	{'Influenza URGENCIA ', 'Bronquitis URGENCIA'}	**->**	{'Neumonía URGENCIA', 'IRA Alta URGENCIA'}	May-Jun	5 a 14
2	{'ROCIO', 'Influenza URGENCIA', 'HUMEDAD', 'Temperatura mínima'}	**->**	{'IRA Alta URGENCIA', 'TEMPERATURA'}	May-Jun	15 a 64
3	{'Otra causa respiratoria URGENCIA', 'Temperatura mínima'}	**->**	{ 'IRA Alta URGENCIA'}	May-Jun	15 a 64
4	{'TEMPERATURA', 'Temperatura mínima'}	**->**	{'Causas sistema respiratorio HOSPITALIZACIÓN'}	May-Jun	65 y más
5	{'Neumonía URGENCIA', 'MP10'}	**->**	{'Otra causa respiratoria URGENCIA'}	Sep-Oct	65 y más
6	{'MP2.5', 'TEMPERATURA'}	**->**	{'ROCIO', 'Neumonía URGENCIA', 'Temperatura máxima'}	Nov-Dic	Todos

-Regla 1: La **influenza** y la **bronquitis** son enfermedades respiratorias que pueden derivar en neumonía e infecciones **respiratorias agudas (IRA)**, especialmente en niños, como lo indica Borrell et al. (2016) y Cifuentes Martínez et al. (2020). Tanto la **influenza** como la **neumonía** pueden tener un origen viral y afectar las vías respiratorias inferiores. Los síntomas de ambas, al igual que la bronquiolitis, incluyen tos seca, dificultad respiratoria, fiebre y sibilancias.

Es importante destacar que la mayoría de las infecciones respiratorias presentan un patrón estacional, siendo más frecuentes durante los meses fríos del año. Esta primera regla evidencia la relación entre estas enfermedades, mientras que el período y rango de edad coinciden con lo expuesto por la fuente citada.

-Regla 2: La combinación de ciertas variables climáticas, como una **temperatura** de **rocío** entre 7 y 11 °C, una **humedad** relativa del 60 al 70 % y **temperaturas bajas** entre 15 y 18 °C, puede aumentar el riesgo de **infecciones respiratorias agudas (IRA)** en adultos durante los meses de mayo y junio (Dünner et al., 2020). La influenza también puede contribuir a este aumento de riesgo.
Estos virus pueden ser más activos en épocas de mayor humedad ambiental. Por lo tanto, es importante tomar medidas para reducir la exposición a estos virus, como lavarse las

manos con frecuencia, mantener una dieta saludable y controlar el estrés ambiental (De León et al., 1997; Pino et al., 2015).

-Regla 3: Presenta una conexión entre las enfermedades vinculadas a **otras causas del sistema respiratorio** y las **temperaturas mínimas** entre 6 y 9 °C que pueden debilitar el sistema inmunológico, lo que hace que las personas sean más susceptibles a las **infecciones respiratorias agudas (IRA)**, especialmente si tienen otras enfermedades respiratorias subyacentes (Ardusso et al., 2019; De León et al., 2020).

-Regla 4: Expone como **temperaturas bajas** de entre 5 y 13 °C pueden aumentar el riesgo de **enfermedades respiratorias generales** (Pino et al,. 2015; Ardusso et al., 2019; Dünner et al., 2020). Esto puede deberse a que las fluctuaciones de temperatura pueden provocar alteraciones en la exposición a pólenes, es decir, cuando las temperaturas son más bajas, el polen se vuelve más pesado y es menos probable que se disperse en el aire. Esto puede provocar una acumulación de polen en el aire, lo que puede aumentar la exposición a las personas alérgicas.

-Regla 5: A finales de invierno y comienzo de primavera, las concentraciones de **material particulado de 10 micras (MP10)** entre 17 y 34 microgramos por metro cúbico (µg/m³) pueden aumentar el riesgo de **otras enfermedades respiratorias**, como la **bronquitis**, la **rinitis**, la **faringitis** y la **amigdalitis**. El **MP10** es un tipo de contaminación del aire que puede causar inflamación de las vías respiratorias, lo que dificulta la respiración y aumenta el riesgo de infección (Dünner et al., 2020; Pino et al,. 2015; Ardusso et al., 2019).

Además, las concentraciones elevadas de **MP10** pueden aumentar el riesgo de **enfermedades pulmonares** crónicas, como la neumoconiosis y la silicosis. Estas enfermedades se producen cuando las partículas de contaminación se acumulan en los pulmones y causan daño tisular.

-Regla 6: A finales de año, muestra una conexión entre las concentraciones de **material particulado de 2.5 micras (MP2.5)** entre 6 y 9 microgramos por metro cúbico (µg/m³)

y las **temperaturas elevadas** entre 18 y 29 °C pueden aumentar el riesgo de **neumonía**. Las temperaturas elevadas pueden debilitar el sistema inmunológico, de acuerdo con Dünner et al. (2020), Ardusso et al. (2019) y Cifuentes Martínez et al. (2020), donde además se indica que factores contaminantes como MP2.5 junto con episodios de calor o frío extremo, crean condiciones ideales para la propagación de virus responsables de la **neumonía**. lo que hace que las personas sean más susceptibles a las infecciones.

Además, la combinación de contaminación y temperaturas elevadas puede crear condiciones ideales para la propagación de virus responsables de la **neumonía**. Por ejemplo, un estudio realizado en China encontró que las personas que vivían en áreas con altos niveles de **MP2.5** con altas temperaturas tenían un riesgo mayor de desarrollar **neumonía** (Tian et al., 2019).

4.8.1 Conclusiones de las reglas seleccionadas

Los resultados muestran que durante el mes de Otoño, la Infección Respiratoria Aguda (IRA) es común en Copiapó, predominantemente asociada con influenza, bronquitis y otras afecciones respiratorias. En cambio, en primavera, las condiciones climáticas y la contaminación son los principales factores relacionados con enfermedades como bronquitis, rinitis, faringitis y amigdalitis, junto con enfermedades pulmonares debido a la exposición a contaminantes externos. El análisis de reglas de asociación por período y edad revela tendencias específicas en diferentes épocas del año y grupos de edad, lo que ayuda a comprender mejor estas relaciones. Según datos del Sistema de Información Nacional de Calidad del Aire (SINCA) entre 2017 y 2021, los niveles de MP10 y MP2.5 se mantienen dentro de los límites establecidos, pero siguen siendo desencadenantes de enfermedades respiratorias.

La investigación encontró que existe una relación entre los niveles de contaminación ambiental por MP2.5 y MP10 con el número de consultas respiratorias, independientemente de la edad de los usuarios. En promedio, las personas de todas las edades consultaron más a los servicios de salud durante episodios de contaminación, incluso cuando estos no alcanzaron niveles de emergencia. Este hallazgo es importante

porque sugiere que la contaminación ambiental puede tener un impacto negativo en la salud respiratoria de las personas de todas las edades, incluso en condiciones que no se consideran peligrosas.

Capítulo V Discusión

En este estudio, se aplicaron reglas de asociación a un conjunto de datos de enfermedades respiratorias y variables ambientales en Copiapó. Los resultados confirmaron que las enfermedades respiratorias son un problema de salud pública en la ciudad, y que la exposición a la contaminación ambiental, especialmente a material particulado (MP10 y MP2.5), es un factor de riesgo importante para su desarrollo.

Los resultados de la investigación se alinean con la literatura previa sobre la relación entre la contaminación del aire y las enfermedades respiratorias. Diversos estudios (Dünner et al., 2020; Pino et al., 2015; Tian et al., 2019) han demostrado que la exposición a MP10 y MP2.5 incrementa el riesgo de IRA, bronquitis, neumonía y otras enfermedades del sistema respiratorio.

En particular, el trabajo de Cifuentes Martínez et al. (2020) encontró una relación significativa entre MP2.5 y enfermedades respiratorias, coincidiendo con algunas de las limitaciones del presente estudio. Entre estas, se destaca la falta de estaciones de monitoreo y repositorios de datos confiables. Adicionalmente, la falta de diferenciación por sexo y la amplitud de los rangos de edad dificultan la interpretación de los resultados. Sin embargo, a diferencia del estudio de Cifuentes Martínez et al. (2020) que se enfocó en dos ciudades y seis establecimientos de salud, este estudio se limita a una ciudad, pero analiza otras variables medioambientales. Esta diferencia permite una comparación más precisa del impacto de la contaminación del aire en la salud respiratoria dentro de un contexto específico.

Los patrones identificados podrían utilizarse para desarrollar sistemas de alerta temprana que informen a la población sobre condiciones ambientales que pueden aumentar el riesgo de enfermedades respiratorias. Estos sistemas permitirían a las autoridades de salud tomar medidas proactivas para reducir la exposición a la contaminación ambiental y proteger la salud pública.

En definitiva la minería de datos puede ser una herramienta útil para monitorear la situación de las enfermedades respiratorias en Copiapó y para evaluar la eficacia de las medidas de intervención implementadas y este estudio proporciona un punto de partida para futuras investigaciones sobre la relación entre la contaminación del aire y las enfermedades respiratorias en Copiapó. En general, este estudio proporciona evidencia de que la contaminación del aire es un problema de salud pública en Copiapó y que la minería de datos puede ser una herramienta útil para identificar patrones y asociaciones entre variables ambientales y enfermedades respiratorias.

5.1 Limitaciones

A pesar de los esfuerzos por garantizar la confiabilidad del estudio, es importante reconocer algunas limitaciones que podrían haber impactado en los resultados y su interpretación.

En primer lugar, la falta de datos sobre otras variables ambientales limita la comprensión de las asociaciones entre los contaminantes y las enfermedades respiratorias. No se puede determinar con precisión el impacto individual de cada variable, ni cómo interactúan entre sí para afectar la salud. En segundo lugar, la ausencia de información sobre el sexo de los pacientes y de categorías diagnósticas detalladas para las enfermedades respiratorias limita la generalización de los hallazgos. El sexo y la enfermedad específica pueden influir en la susceptibilidad a la contaminación del aire, por lo que no se pueden sacar conclusiones definitivas para toda la población.

Finalmente, el análisis de rangos etarios amplios (15 a 64 años) no permite distinguir entre adolescentes, adultos jóvenes, adultos y adultos mayores. Esto limita la identificación de grupos poblacionales especialmente vulnerables a la contaminación del aire.

Capítulo VI Conclusiones

Este trabajo de apoyo a la investigación presenta un método novedoso para identificar patrones entre las enfermedades respiratorias y las variables medioambientales en la ciudad de Copiapó. El modelo utiliza reglas de asociación, un tipo de aprendizaje que ha demostrado ser eficaz en la identificación de patrones complejos. El trabajo también utiliza herramientas de programación, principalmente en el lenguaje Python, para facilitar la implementación del modelo.

La metodología del estudio no permite afirmar que la contaminación ambiental causa enfermedades respiratorias, sino sólo que existe una asociación entre ambas variables. Por lo tanto, es posible que otros factores, como la genética o el estado de salud individual, también influyen en el riesgo de contraer una enfermedad respiratoria. La evidencia sugiere que los niños y adultos mayores son los grupos más susceptibles a los efectos de la contaminación atmosférica, incluso si no son la mayoría de los diagnósticos.

Se resalta la necesidad de ampliar la investigación sobre la relación entre la contaminación ambiental y las enfermedades respiratorias a otras ciudades del país. Una limitación importante es la falta de estaciones de monitoreo de calidad del aire que permitan mejorar el acceso a los datos y contrastar la información.

Los resultados del estudio ofrecen una base sólida para futuras investigaciones que exploren la relación entre las variables medioambientales y las enfermedades respiratorias. Estas investigaciones podrían ayudar a comprender mejor los mecanismos que subyacen a esta relación, lo que podría conducir al desarrollo de estrategias de prevención y tratamiento más efectivas.

Para finalizar se propone como trabajo futuro utilizar series temporales para generar reglas de predicción de enfermedades respiratorias. También se propone estudiar el

impacto del COVID-19 por separado, ya que su aparición fue un evento inesperado que no se puede considerar dentro de las condiciones estudiadas.

Referencias

Asociación Latinoamericana de Tórax, “Foro de las Sociedades Respiratorias Internacionales El impacto mundial de la Enfermedad Respiratoria,” México, 2017.

Instituto Nacional de Estadísticas (INE), “Estadistísticas Vitales Informe Anual 2016,” Chile, 2016.

W. H. Organization, “OMS | Infecciones del tracto respiratorio,” OMS, 2015. [Online]. Available: https://www.who.int/topics/respiratory_tract_diseases/es/. [Accessed: 24-Mar-2020].

A. M. Rivera, C. A., Díaz, R. A., Céspedes, P. F., & Kalergis, “Virus Respiratorio Sincicial: un desafío para la salud pública a nivel mundial | Revista de la Sociedad Española de Bioquímica y Biología Molecular | SEEBM,” *EN EL CONO SUR*, vol. 26, 2016.

K. Homero Puppo, K. Rodrigo Torres-Castro, and K. Javiera Rosales-Fuentes, “REHABILITACIÓN RESPIRATORIA EN NIÑOS,” Rev. Médica Clínica Las Condes, vol. 28, no. 1, pp. 131–142, Jan. 2017.

Puyol Moreno, J. (2014). Una aproximación a Big Data= An approach to Big Data.

Siebes, A. (2000). Data Mining and Statistics: A Systems Point of View. In Computational Intelligence in Data Mining (pp. 1-38). Vienna: Springer Vienna.

Vega, M. Á., Mora, L. M. Q., & Badilla, M. V. C. (2020). Inteligencia artificial y aprendizaje automático en medicina. Revista médica sinergia, 5(8), e557-e557.

López, C. P. (2017). Minería de datos: técnicas y herramientas. España

Mamani Rodríguez, Z., Del Pino Rodríguez, L., & Cortez Vasquez, A. (2017). Minería de datos distribuida usando clustering k-means en la predictibilidad del proceso petitorio en una organización pública. Industrial Data, 20(2), 123–130. https://doi.org/10.15381/idata.v20i2.13949

J. Rodríguez-Núñez, Iván & Torres, Gerardo & Zenteno, Daniel & Tapia, Ximena & Medina, Kimberly & Tapia, "Programa de Rehabilitación Respiratoria Infantil en un Hospital Público de Chile | Request PDF," Arch. Argent. Pediatr., vol. 117, 2019.

R. Sepúlveda M., "Las enfermedades respiratorias del adulto mayor en Chile: un desafío a corto plazo," Rev. Chil. enfermedades Respir., vol. 33, no. 4, pp. 303–307, Dec. 2017.

R. González, R. Pinto, and J. P. Álvarez, "LAS ENFERMEDADES RESPIRATORIAS EN CHILE: UN REFLEJO DE NUESTRA HISTORIA," Rev. Médica Clínica Las Condes, vol. 28, no. 1, pp. 152–154, Jan. 2017.

M. Barros Monge, "Sociedad Chilena de Enfermedades Respiratorias: 75 años de historia," Rev. Chil. enfermedades Respir., vol. 21, no. 1, Jan. 2005.

G. de Chile, ESTRATEGIA NACIONAL DE SALUD Para el cumplimiento de los Objetivos Sanitarios de la Década. 2010. https://extranet.who.int/nutrition/gina/sites/default/filesstore/CHL_2011%20Estrategia%20Nacional%20de%20Salud.pdf

CIE-10 ES. Clasificación Internacional de Enfermedades - 10ª Revisión. Modificación Clínica. 2ª Edición-Enero 2018. Tomo I: Diagnósticos.

Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) (Agosto de 2019). Glosario Meteorológico. Colombia. p. 286.

Rodríguez Jiménez, Rosa María; Benito Capa, Águeda; Portela Lozano, Adelaida (2004). Meteorología y Climatología. Fundación Española para la Ciencia y la Tecnología (FECYT). p. 12 a 33.

Keneth Denner Chagua Namuche "ASOCIACIÓN DEL MATERIAL PARTICULADO (PM10-PM2.5) CON LAS ENFERMEDADES RESPIRATORIAS EN JESÚS MARÍA" http://repositorio.unjfsc.edu.pe/bitstream/handle/20.500.14067/6471/KENETH%20DENNER%20CHAGUA%20NAMUCHE_compressed%20%281%29.pdf?sequence=1&isAllowed=y

José C. Riquelme, Roberto Ruiz, Karina Gilbert “Minería de Datos: Conceptos y Tendencias”. 2006
https://idus.us.es/bitstream/handle/11441/43290/Miner%c3%ada%20de%20datos.pdf?sequence=1&isAllowed=y

Bellinger, M. Jabbar, O. Zaïane, A. Osornio-Vargas (2017).

Linoff, G., & Berry, M. J. (1997). Data Mining Techniques: For Marketing, Sales, and Customer Support. Hoboken.

MC Beatriz Beltrán Martínez, “MINERÍA DE DATOS”, 2001
https://scholar.google.com/scholar?hl=es&as_sdt=0%2C5&q=%22mineria+de+datos%22&btnG=

Luis Gonzalo Espejo Tapia (“Análisis del comportamiento de las enfermedades respiratorias en la comuna de Copiapó, utilizando algoritmos de clustering.”, 2022)

Zaki, M. J., & Wong, L. (2004). DATA MINING TECHNIQUES. Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, 125–163.

Abdullahi Sidow Osman “Data Mining Techniques: Review”, (2019)
http://ojs.mediu.edu.my/index.php/IJDSR/article/view/1841/717

Moya Amaris, M. E., y Rodríguez Rodríguez, J. E. (2003). La contribución de las reglas de asociación a la minería de datos. Tecnura, 7(13), 94–109. https://doi.org/10.14483/22487638.6175

Malberti Riveros, María Alejandra, & Elida Beguerí, Graciela. (2015). Reglas de Asociación con los datos de una biblioteca universitaria. Revista Cubana de Ciencias Informáticas
http://scielo.sld.cu/pdf/rcci/v9n4/rcci03415.pdf

Silverstein, C., Brin, S., & Motwani, R. (1998). Beyond market baskets: Generalizing association rules to dependence rules. Data mining and knowledge discovery, 2, 39-68.

ROMERO, C. Aplicación de técnicas de adquisición de conocimiento para la mejora de cursos hipermedia
adaptativos basados en Web. Tesis Doctoral.. Universidad de Granada. E.T.S. Ingeniería Informática. 2003.

Said, A. M., Dominic, P. D. D., & Abdullah, A. B. (2009). A comparative study of fp-growth variations. International journal of computer science and network security, 9(5), 266-272.

Han, J., Kamber, M., y Pei, J. (2011). Data mining : Concepts and techniques. doi:10.1016/C2009-0-61819-5

Pérez-Gómez, R. (2020). Generación de reglas de asociación para productos de retail utilizando el algoritmo FP-Growth paralelo. Actas Del Congreso Internacional De Ingeniería De Sistemas, 231-250. https://doi.org/10.26439/ciis2019.5349

Bart Goethals Survey on Frequent Pattern Mining. HIIT Basic Research Unit
Department of Computer Science, University of Helsinki P.O. box 26, FIN-00014
http://adrem.uantwerpen.be/~goethals/publications/pubs/fpm_survey.pdf

Chávez Chong, C. O., Sánchez García, J. E., & DelaCerda, J. (2015). Análisis de componentes principales funcionales en series de tiempo económicas (Analysis of principal functional components in economic time series). GECONTEC: Revista Internacional de Gestión del Conocimiento y la Tecnología, 3(2).

Sánchez Mangas, A. (2012). Análisis de componentes principales: versiones dispersas y robustas al ruido impulsivo (Master's thesis).
https://e-archivo.uc3m.es/bitstream/handle/10016/15618/PFC_Andres%20Sanchez%20Mangas.pdf?sequence=1&isAllowed=y

Sarıyer G, Öcal Taşar C. Highlighting the rules between diagnosis types and laboratory diagnostic tests for patients of an emergency department: Use of association rule mining. Health Informatics Journal. 2020;26(2):1177-1193. doi:10.1177/1460458219871135

Moine, J. M., Haedo, A. S., & Gordillo, S. E. (2011). Estudio comparativo de metodologías para minería de datos. In XIII Workshop de Investigadores en Ciencias de la Computación. http://sedici.unlp.edu.ar/handle/10915/20034

Zambrano, J.-C., Quiroz-Palma, P., Santamaría-Philco, A., & Zamora, W. (2022). Covid-19 en Ecuador: Aplicación de minería de datos. Informática Y Sistemas: Revista De Tecnologías De La Informática Y Las Comunicaciones, 6(1), 35–52. https://doi.org/10.33936/isrtic.v6i1.4366

Mejía, J. C. G. (2019). Aplicación de la técnica regresión logística de la minería de datos en el proceso de descubrimiento de conocimiento (KDD) en bases de datos operativas o transaccionales. Perspectiv@ s, 14(13), 51-55.

Cortina, V. G. (2015). Aplicación de la metodología crisp-dm a un proyecto de minería de datos en el entorno universitario. Universidad Carlos III de Madrid.

Yanina Bellini Saibene, Martín Volpaccio, Santiago Banchero, Romina Mezher, (2014), "Desarrollo y uso de herramientas libres para la explotación de datos de los radares meteorológicos del INTA" - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Fases-del-proceso-de-CRISP-DM-Adaptado-de-10_fig2_306959832 [accessed 14 Jun, 2023]

Kimball, R., & Caserta, J. (2004). The data warehouse ETL toolkit. John Wiley & Sons.

BOLETÍN EPIDEMIOLÓGICO TRIMESTRAL: INFLUENZA, SE 1 – 39, AÑO 2019

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), 217-222.

Dagnino, J. (2014). Datos faltantes (missing values). Rev Chil Anest, 43, 332-4.

Rosas, J. F. M., & Verdejo, E. Á. (2009). Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus. Revista de Métodos Cuantitativos para la Economía y la Empresa, 7, 3-30.

Allison, P. D. (2001). Missing data. Sage publications.

Petersohn, D., Macke, S., Xin, D., Ma, W., Lee, D., Mo, X., ... & Parameswaran, A. (2020). Towards scalable dataframe systems. arXiv preprint arXiv:2001.00888.

Wu, Y. (2020). Is a dataframe just a table?. In 10th Workshop on Evaluation and Usability of Programming Languages and Tools (PLATEAU 2019). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Cravero, A. L., Sepúlveda, S. E., Mazón, J. N., & Trujillo, J. C. (2013). Un enfoque de ingeniería de requerimientos basada en el alineamiento de almacenes de datos y la estrategia del negocio. Ingeniare. Revista chilena de ingeniería, 21(3), 314-327.

Aroca, P. R., García, C. L., & López, J. J. G. (2009). Estadística descriptiva e inferencial. Revista el auge de la estadística en el siglo XX, 22, 165-176.

Infante Amunátegui, A. (2015). Hoya n° 302 Copiapó. Catastro de pozos al 31 de mayo de 1971.

Gómez Sarria, N. (2014).Climatología urbana de Copiapó como ciudad localizada en un medio ambiente árido. Disponible en https://repositorio.uchile.cl/handle/2250/130424

Sepúlveda, B., Santana, R., & Soto, J. (2016). Producción masiva forzada de árboles silvestres del desierto de Atacama, Copiapó (Chile). Idesia (Arica), 34(3), 7-16.

Ramon Sangüesa i Solé (2019). Preparación de datos

Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., & Hooman, A. (2013). An overview of principal component analysis. Journal of Signal and Information Processing, 4(3B), 173.

Ringnér, M. (2008). What is principal component analysis?. Nature biotechnology, 26(3), 303-304.

Kang, S. M., & Wagacha, P. W. (2014). Extracting diagnosis patterns in electronic medical records using association rule mining. International Journal of Computer Applications, 108(15).

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In Machine learning proceedings 1995 (pp. 194-202). Morgan Kaufmann.

Patil, M., & Patil, T. (2022). Apriori Algorithm against Fp Growth Algorithm: A Comparative Study of Data Mining Algorithms. Available at SSRN 4113695.

Heaton, J. (2016, March). Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms. In SoutheastCon 2016 (pp. 1-7). IEEE.

Guía para la Gestión de Datos de Investigación del Ministerio de Ciencia, Tecnología e Innovación, Ministerio de Ciencia, Tecnología e Innovación (Minciencias) de Colombia

Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. Dmkd, 3(8), 34-39.

Cevallos-Macías , J., Solórzano-Cadena , R., Palma-Menéndez , S., & Verduga-Urdánigo , F. (2020). APLICACIÓN DE REGLAS DE ASOCIACIÓN SOBRE MICROSERVICIOS EN LAS MICROEMPRESAS: ARTÍCULO DE INVESTIGACIÓN. REVISTA CIENTÍFICA MULTIDISCIPLINARIA ARBITRADA YACHASUN - ISSN

Stančin, I., & Jović, A. (2019, May). An overview and comparison of free Python libraries for data mining and big data analysis. In 2019 42nd International convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 977-982). IEEE.

Ari, N., & Ustazhanov, M. (2014, September). Matplotlib in python. In *2014 11th International Conference on Electronics, Computer and Computation (ICECCO)* (pp. 1-6). IEEE.

Piatetsky-Shapiro, G., & Steingold, S. (2000). Measuring lift quality in database marketing. ACM SIGKDD Explorations Newsletter, 2(2), 76-80.

Garrote, P. R., & del Carmen Rojas, M. (2015). La validación por juicio de expertos: dos investigaciones cualitativas en Lingüística aplicada. Revista Nebrija de lingüística aplicada a la enseñanza de lenguas, (18), 124-139.

Galicia Alarcón, Liliana Aidé, Balderrama Trápaga, Jorge Arturo, & Edel Navarro, Rubén. (2017). Validez de contenido por juicio de expertos: propuesta de una herramienta virtual. Apertura (Guadalajara, Jal.), 9(2), 42-53. https://doi.org/10.32870/ap.v9n2.993

Borrell, L. S., & Segura, M. C. (2016). Neumonía y neumonía recurrente. Pediatrá integral, 20(1), 38-42.

Cifuentes Martínez, Paula, Rodríguez-Fernández, Alejandra, Luengo M., Carolina, & Tapia O., Leonardo. (2020). Relación entre contaminación atmosférica y consultas por enfermedades respiratorias en atención primaria de urgencia. Revista chilena de enfermedades respiratorias, 36(4), 260-267. https://dx.doi.org/10.4067/S0717-73482020000400260

De León, J. M., Acosta, D., Lorduy, F. A., De la Cruz Pinzón, C., Arrieta, J. M. E., Jaramillo, C., ... & Pinzón, H. (1997). Infección respiratoria aguda. Instituto de Seguros Sociales.

Dünner, M. A., Puentes, R., Vaquero, A., & Díaz, J. (2020). Enfermedades no transmisibles y clima en Chile: un resumen de evidencia para el período 1990-2019. Revista del Instituto de Salud Pública de Chile, 4(2).

Pino, P., Iglesias, V., Garreaud, R., Cortés, S., Canals, M., Folch, W., ... & Steenland, K. (2015). Chile confronts its environmental health future after 25 years of accelerated growth. Annals of global health, 81(3), 354-367.

Ardusso, L. R., Neffen, H. E., Fernández-Caldas, E., Saranz, R. J., Parisi, C. A., Tolcachier, A., ... & Marino, D. (2019). Intervención ambiental en las enfermedades respiratorias. MEDICINA (Buenos Aires), 79(2), 123-136.

Tian, Y., Liu, H., Wu, Y., Si, Y., Li, M., Wu, Y., ... & Hu, Y. (2019). Ambient particulate matter pollution and adult hospital admissions for pneumonia in urban China: a national time series analysis for 2014 through 2017. PLoS medicine, 16(12), e1003010.

Anexo A Diccionario de datos

Diccionario de datos				
Enfermedades respiratorias				
#	Nombre	Tipo	Definición breve	Código CIE-10
0	IRA_Alta	Flotante	Infecciones agudas de las vías respiratorias superiores.	(J00-J06)-U
1	Influenza	Flotante	Enfermedad respiratoria contagiosa provocada por virus de influenza.	(J09-J11)-U
2	Neumonia	Flotante	Gripe y Neumonía.	(J12-J18)-U
3	Bronquitis_bronquiolitis	Flotante	Otras infecciones agudas de las vías respiratorias inferiores.	(J20-J21)-U
4	Crisis_obstructiva_bronquial	Flotante	Enfermedades crónicas de las vías respiratorias inferiores.	(J40-J46)-U
5	Otra_causa_respiratoria	Flotante	Otras enfermedades respiratorias.	(J22,J30,J39,J47,J60-J98)-U
6	CAUSAS_SISTEMA_RESPIRATORIO	Flotante	Enfermedades respiratorias generales.	-H
7	COVID19_confirmado_u	Flotante	Confirmado de COVID-19, virus identificado (urgencias).	(U07.1)-U
8	COVID19_confirmado_h	Flotante	Confirmado de COVID-19, virus identificado (hospitalización).	(U07.1)-H
Variables ambientales				
#	Nombre	Tipo	Definición breve	Unidad medida
9	Humedad	Flotante	Agua de la que está impregnado un cuerpo.	%
10	Rocio	Flotante	Temperatura a la cual se debe enfriar el aire para que el vapor de agua se condense.	[C°]
11	Temperatura	Flotante	Magnitud fisica que expresa el grado de calor del ambiente.	[C°]
12	Temperatura_min	Flotante	Temperatura mínima que podría existir en un periodo de tiempo.	[C°]
13	Temperatura_max	Flotante	Temperatura máxima que podría existir en un periodo de tiempo.	[C°]
Variables contaminantes				
#	Nombre	Tipo	Definición breve	Unidad medida
14	MP10	Flotante	Material particulado con diámetro aerodinámico menor o igual a 10 micras.	$[\mu g/m^3]$
15	MP2.5	Flotante	Material particulado con diámetro aerodinámico menor o igual a 2.5 micras.	$[\mu g/m^3]$

Anexo B Análisis estadístico por periodos de tiempo

Periodo Enero-Febrero (2017-2021) para todas las edades

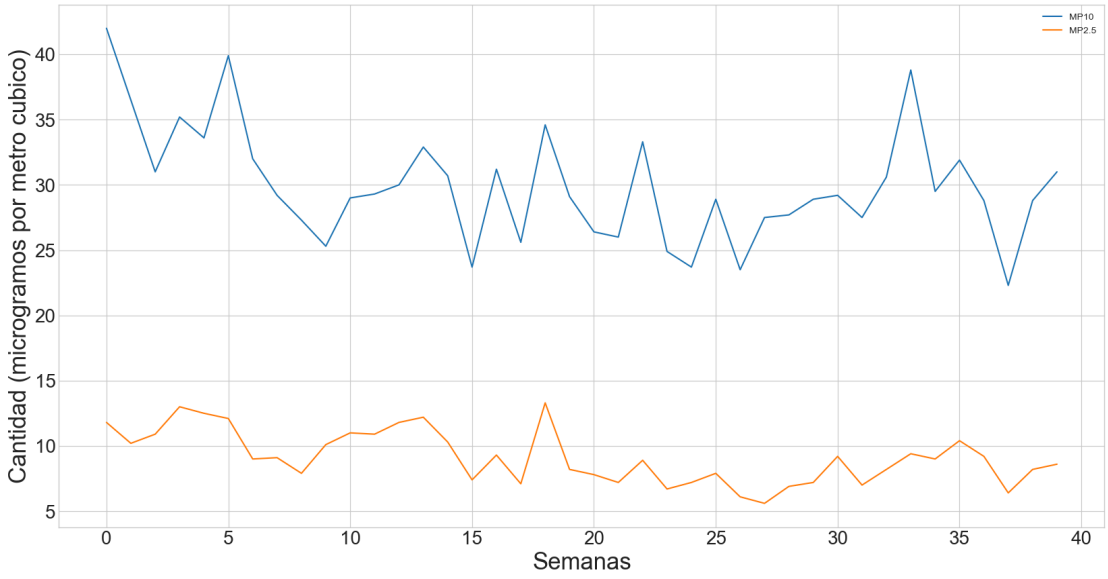
Media cada variable

```
----------------------
Media de cada variable
----------------------
IRA_Alta                      63.8750
Influenza                      1.8500
Neumonia                       6.6000
Bronquitis_bronquiolitis      13.9250
Crisis_obstructiva_bronquial  13.1000
Otra_causa_respiratoria        9.5000
CAUSAS_SISTEMA_RESPIRATORIO    7.4000
COVID19_Confirmado_u           3.6000
COVID19_Confirmado_h           2.1000
Humedad                       65.5950
Rocio                         11.4150
Temperatura                   18.6725
Temperatura_min               13.8150
Temperatura_max               26.8125
MP10                          29.9325
MP2.5                          9.1300
dtype: float64
```

Varianza cada variable

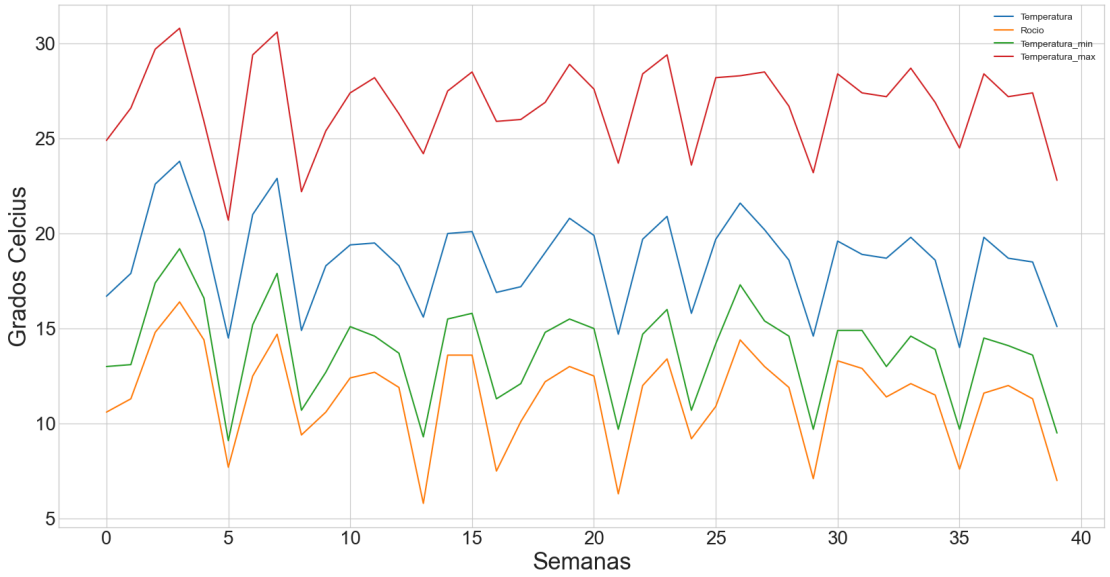
```
-------------------------
Varianza de cada variable
-------------------------
IRA_Alta                      387.394231
Influenza                       4.643590
Neumonia                        8.605128
Bronquitis_bronquiolitis       57.353205
Crisis_obstructiva_bronquial   27.938462
Otra_causa_respiratoria        13.384615
CAUSAS_SISTEMA_RESPIRATORIO     9.323077
COVID19_Confirmado_u           58.246154
COVID19_Confirmado_h           20.605128
Humedad                         9.093308
Rocio                           6.365410
Temperatura                     5.795891
Temperatura_min                 6.431564
Temperatura_max                 5.382660
MP10                           19.626353
MP2.5                           4.146769
dtype: float64
```

Variables contaminantes

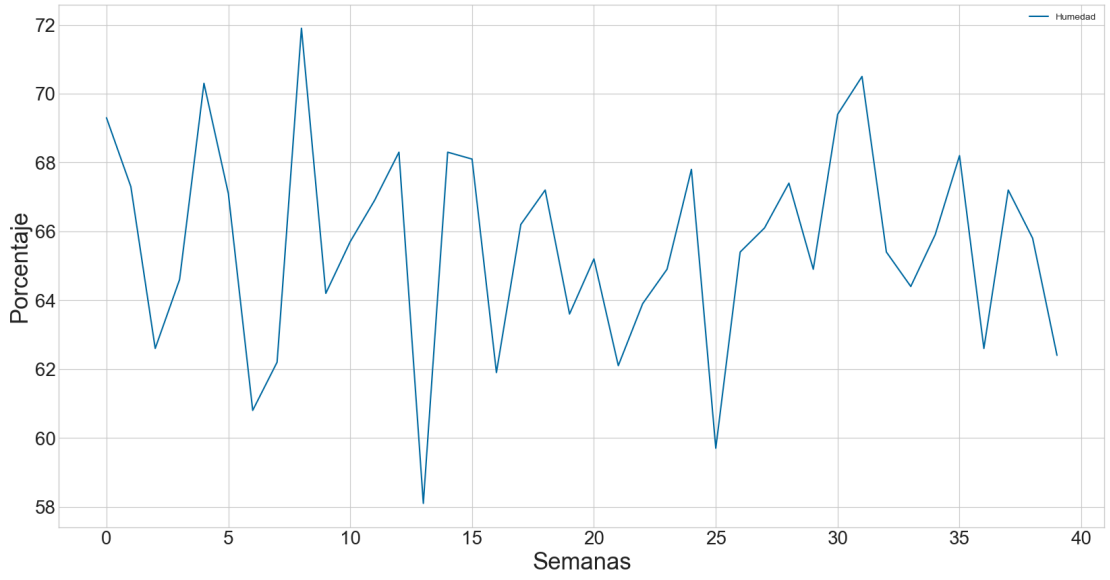


Periodo Enero-Febrero (2017-2021) para todas las edades

Variables ambientales

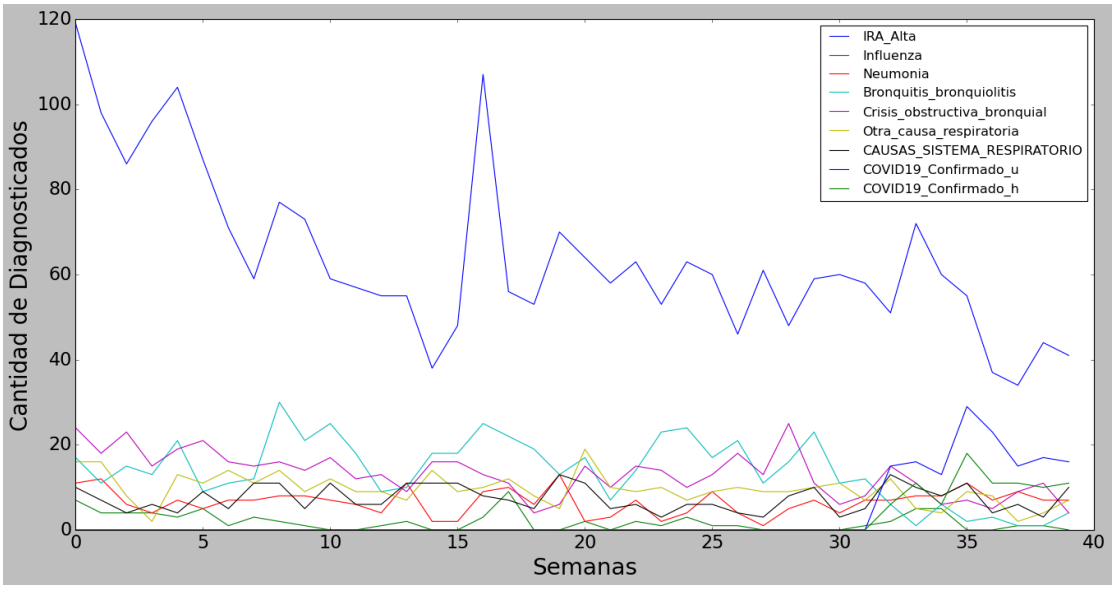


Humedad

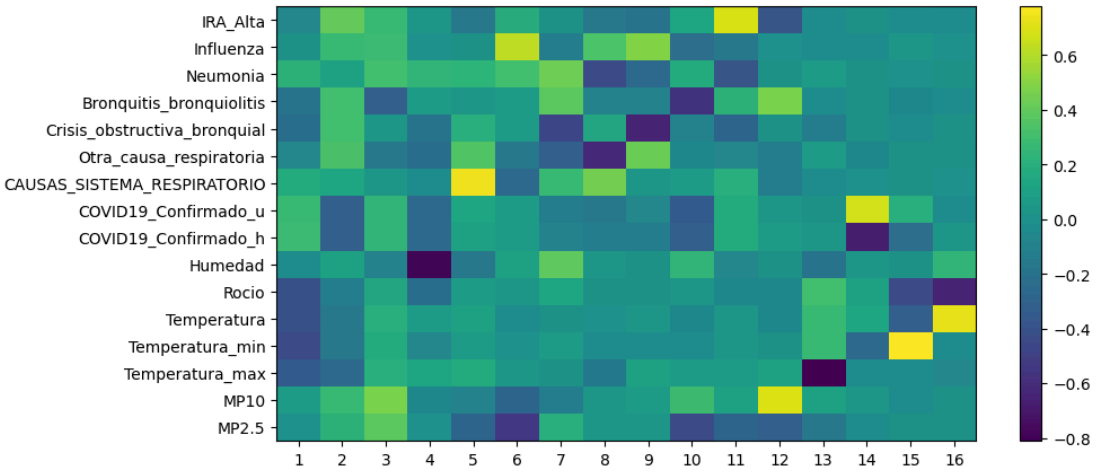


Periodo Enero-Febrero (2017-2021) para todas las edades

Variables enfermedades respiratorias

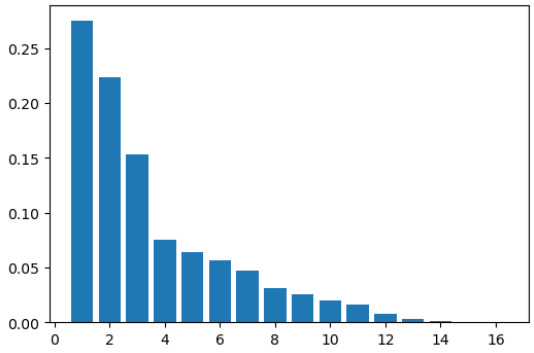


Mapa calor PCA 16 componentes

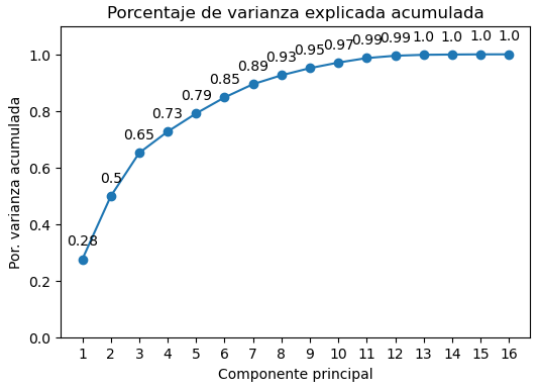


Periodo Enero-Febrero (2017-2021) para todas las edades

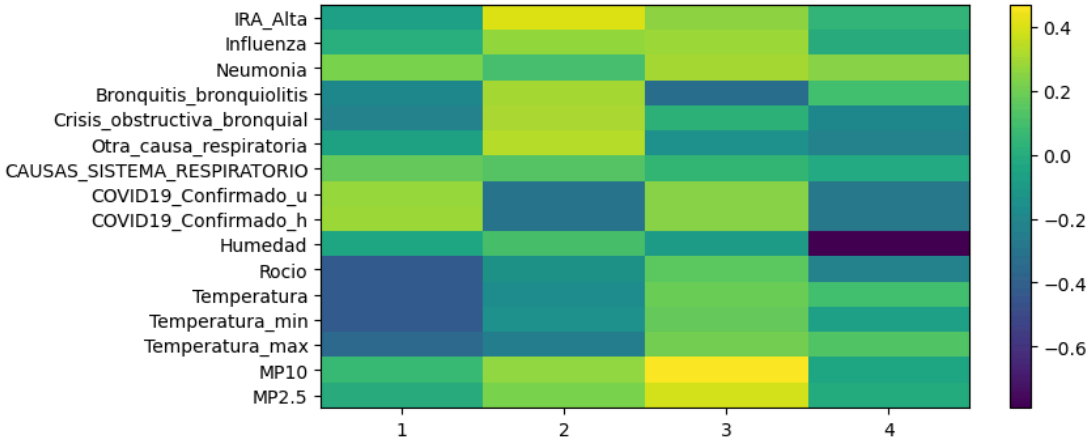
Varianza cada componente



Varianza explicada acumulada



Mapa calor PCA 4 primeras componentes



Periodo Marzo-Abril (2017-2021) para todas las edades

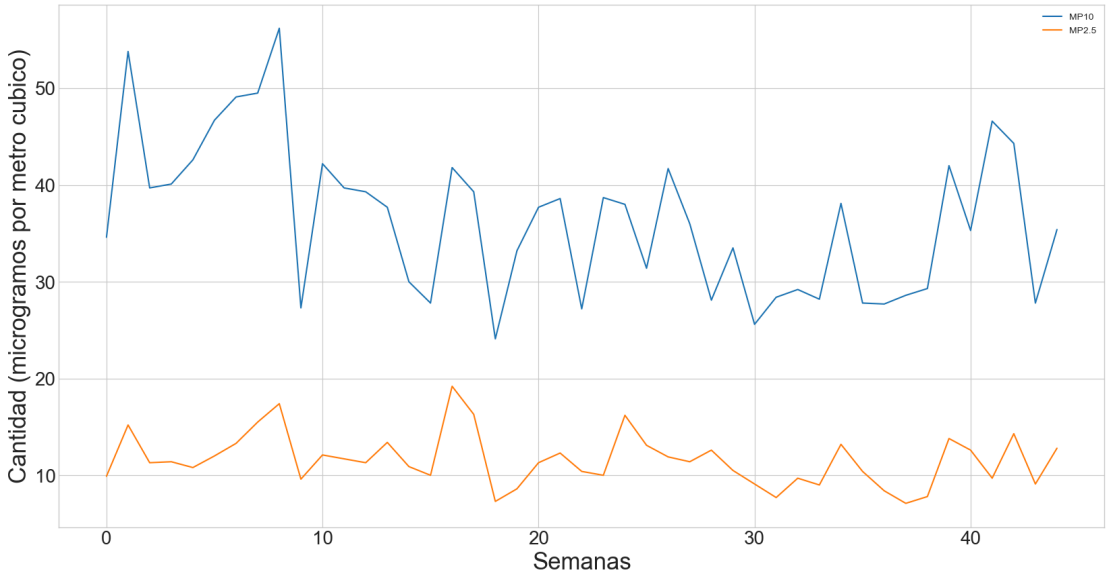
Media cada variable

```
----------------------
Media de cada variable
----------------------
IRA_Alta                          93.155556
Influenza                          3.111111
Neumonia                           6.622222
Bronquitis_bronquiolitis          31.288889
Crisis_obstructiva_bronquial      16.311111
Otra_causa_respiratoria           11.266667
CAUSAS_SISTEMA_RESPIRATORIO        7.044444
COVID19_Confirmado_u               7.133333
COVID19_Confirmado_h               3.800000
Humedad                           68.904444
Rocio                             10.371111
Temperatura                       16.728889
Temperatura_min                   11.942222
Temperatura_max                   24.851111
MP10                              36.442222
MP2.5                             11.591111
dtype: float64
```

Varianza cada variable

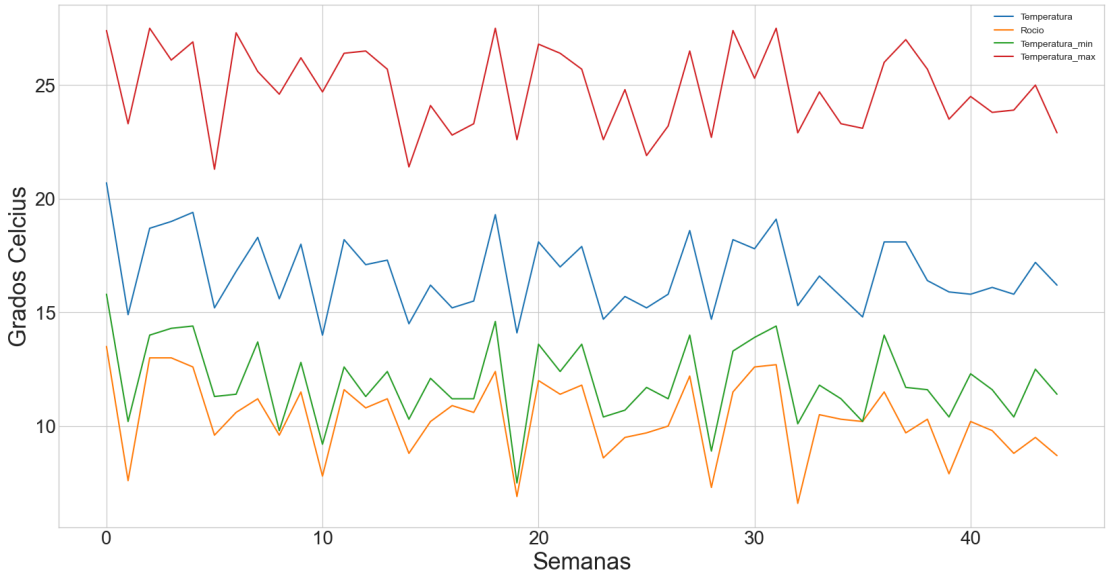
```
-------------------------
Varianza de cada variable
-------------------------
IRA_Alta                        1581.952525
Influenza                         14.328283
Neumonia                          10.467677
Bronquitis_bronquiolitis         706.391919
Crisis_obstructiva_bronquial      51.264646
Otra_causa_respiratoria           15.927273
CAUSAS_SISTEMA_RESPIRATORIO        9.497980
COVID19_Confirmado_u             242.254545
COVID19_Confirmado_h              71.845455
Humedad                           12.023162
Rocio                              2.968465
Temperatura                        2.680737
Temperatura_min                    3.057495
Temperatura_max                    3.363465
MP10                              61.783859
MP2.5                              7.318556
dtype: float64
```

Variables contaminantes

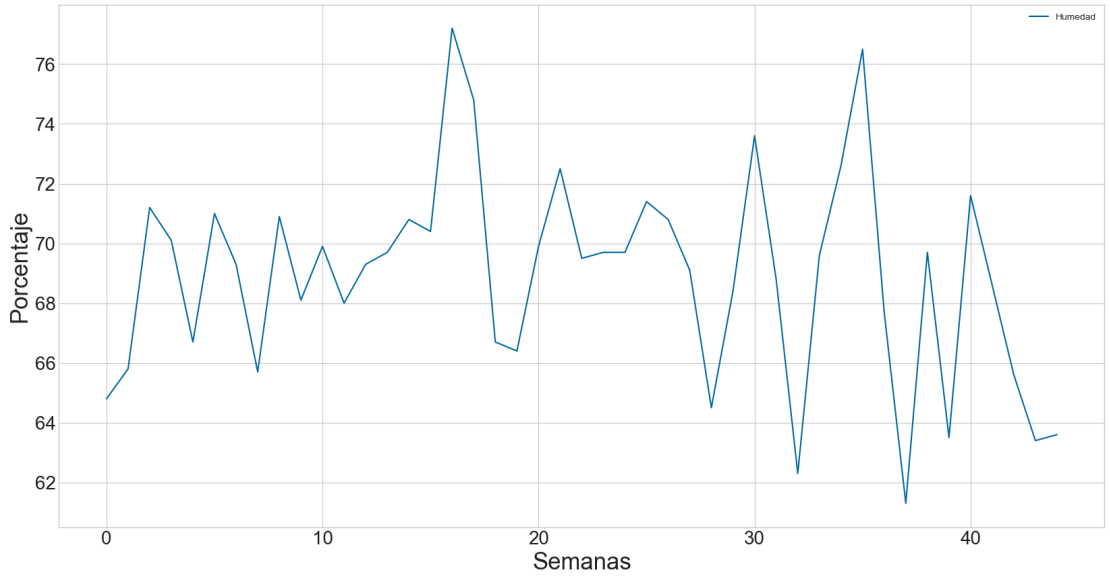


Periodo Marzo-Abril (2017-2021) para todas las edades

Variables ambientales

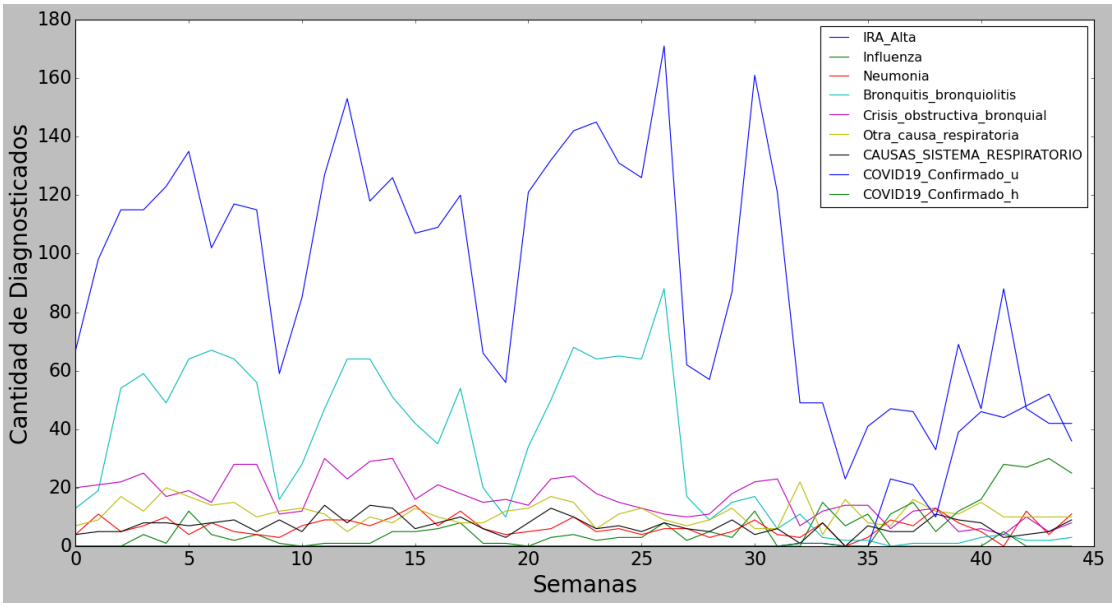


Humedad

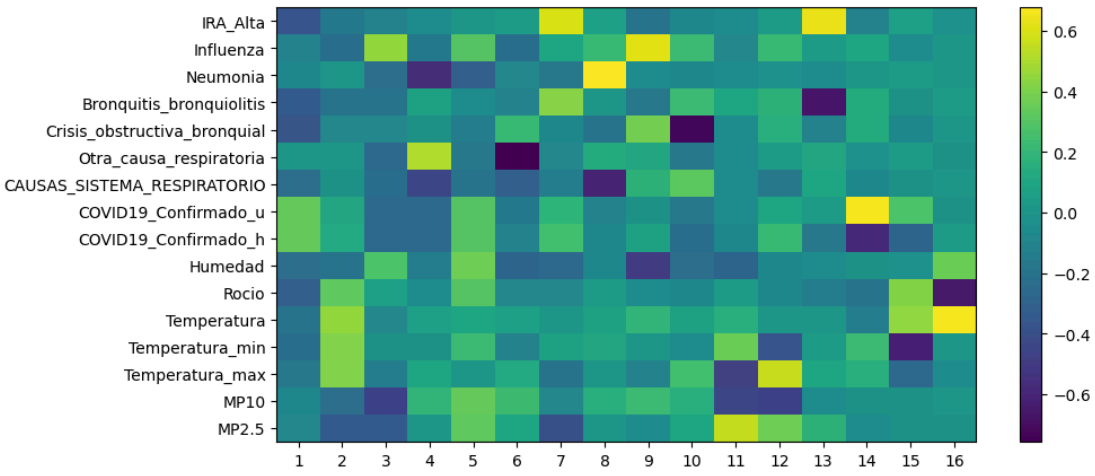


Periodo Marzo-Abril (2017-2021) para todas las edades

Variables enfermedades respiratorias

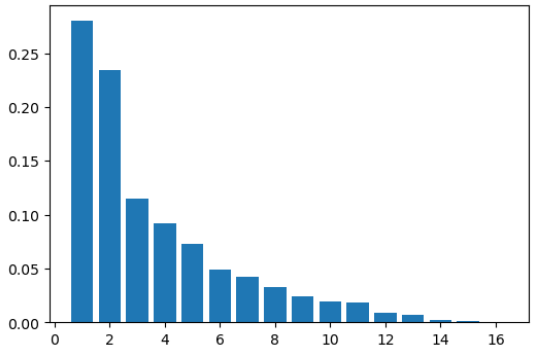


Mapa calor PCA 16 componentes

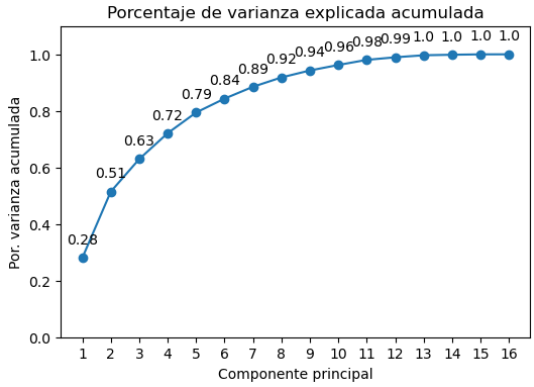


Periodo Marzo-Abril (2017-2021) para todas las edades

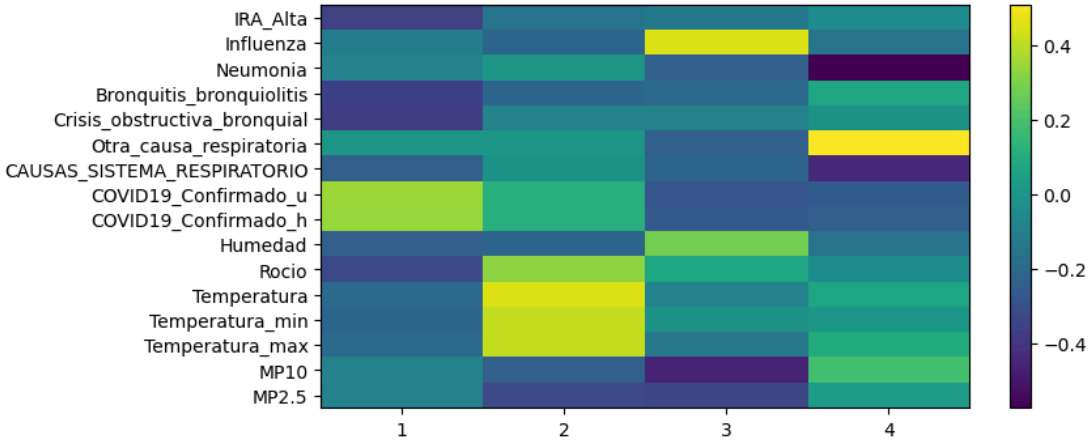
Varianza cada componente



Varianza explicada acumulada



Mapa calor PCA 4 primeras componentes



Periodo Mayo-Junio (2017-2021) para todas las edades

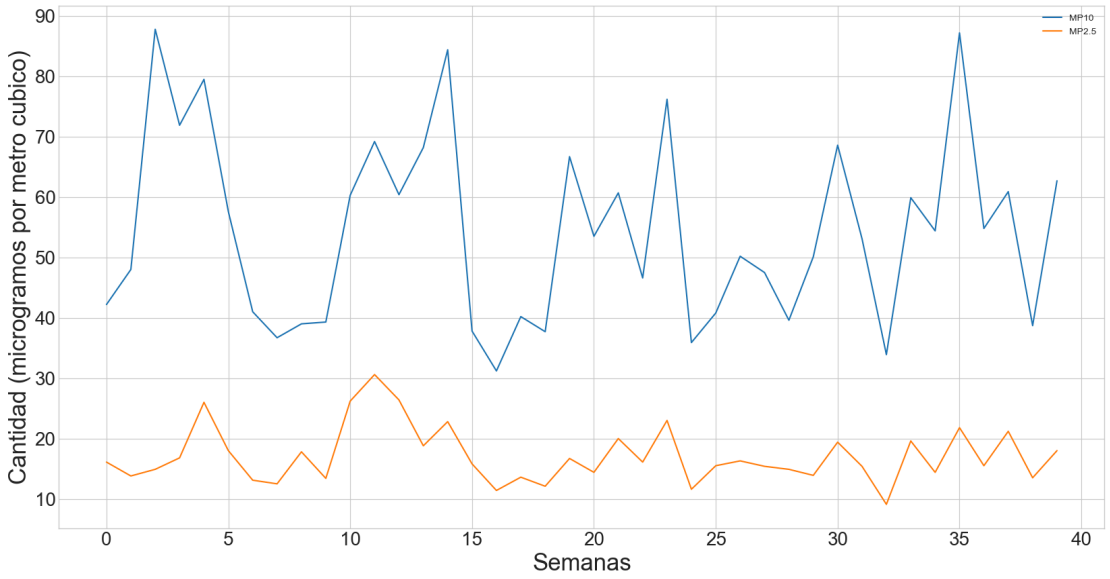
Media cada variable

```
----------------------
Media de cada variable
----------------------
IRA_Alta                        113.475
Influenza                        10.975
Neumonia                          9.450
Bronquitis_bronquiolitis         50.050
Crisis_obstructiva_bronquial     24.275
Otra_causa_respiratoria          13.025
CAUSAS_SISTEMA_RESPIRATORIO      10.825
COVID19_Confirmado_u              8.375
COVID19_Confirmado_h              4.775
Humedad                          65.915
Rocio                             7.385
Temperatura                      14.675
Temperatura_min                   9.550
Temperatura_max                  22.525
MP10                             54.360
MP2.5                            17.145
dtype: float64
```

Varianza cada variable

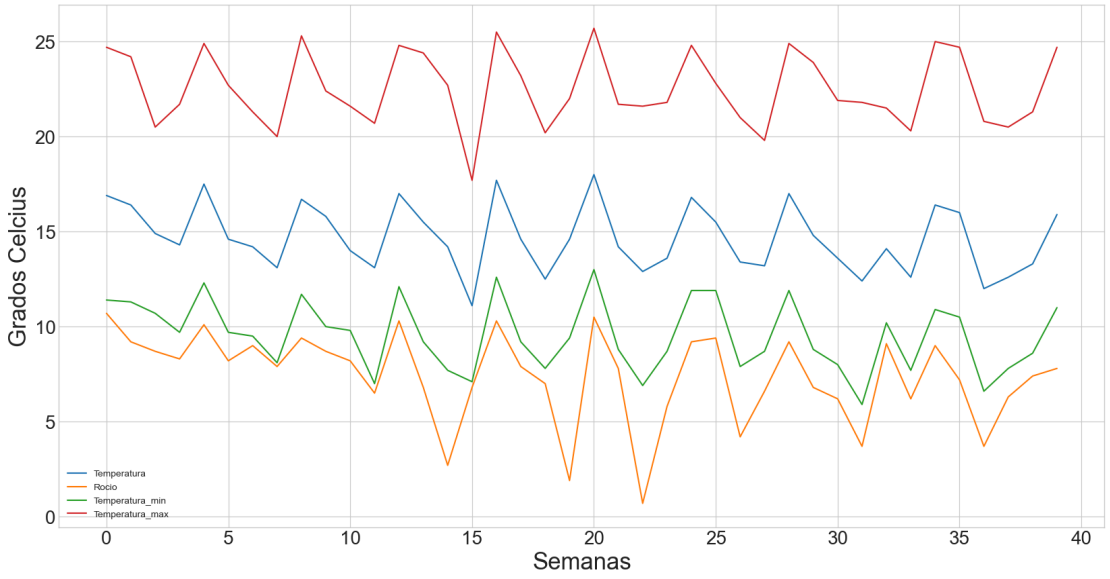
```
-------------------------
Varianza de cada variable
-------------------------
IRA_Alta                        3014.460897
Influenza                        139.153205
Neumonia                          41.023077
Bronquitis_bronquiolitis        1835.125641
Crisis_obstructiva_bronquial     380.409615
Otra_causa_respiratoria           40.486538
CAUSAS_SISTEMA_RESPIRATORIO       30.045513
COVID19_Confirmado_u             211.060897
COVID19_Confirmado_h              71.871154
Humedad                           30.225923
Rocio                              5.662846
Temperatura                        3.093718
Temperatura_min                    3.469744
Temperatura_max                    3.859359
MP10                             244.352718
MP2.5                             21.807667
dtype: float64
```

Variables contaminantes

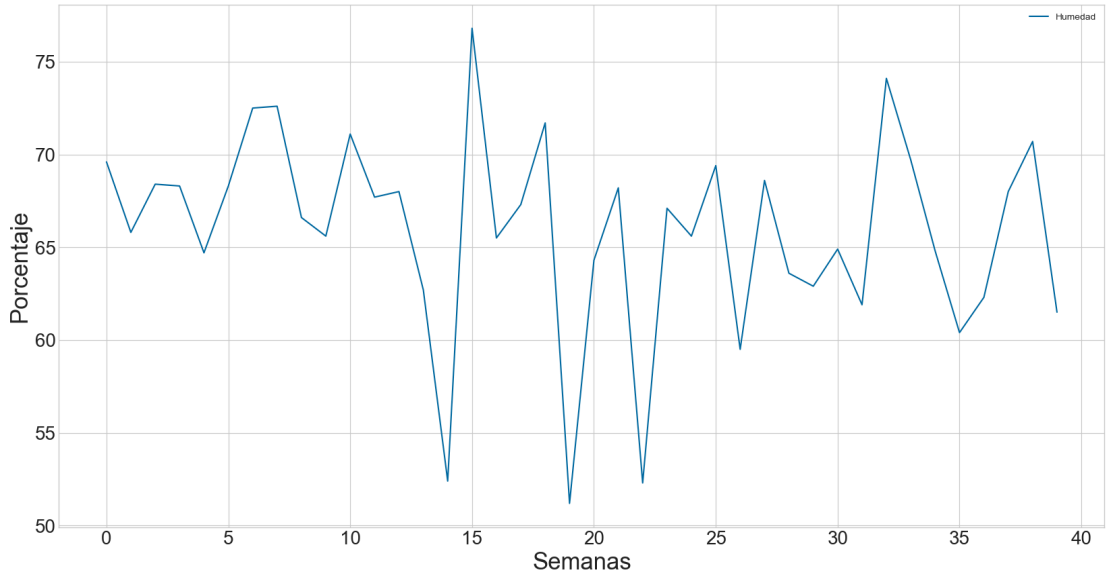


Periodo Mayo-Junio (2017-2021) para todas las edades

Variables ambientales

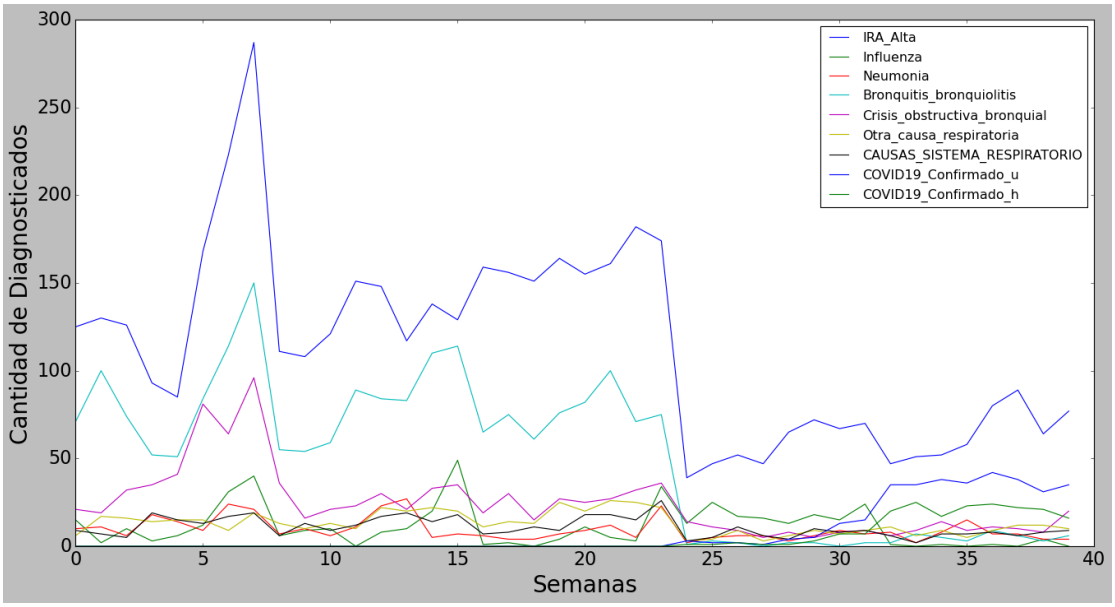


Humedad

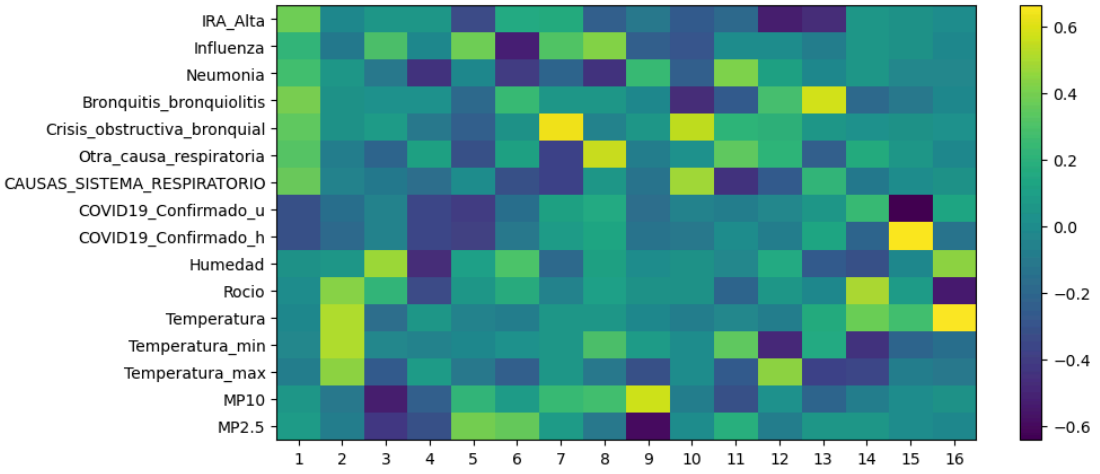


Periodo Mayo-Junio (2017-2021) para todas las edades

Variables enfermedades respiratorias

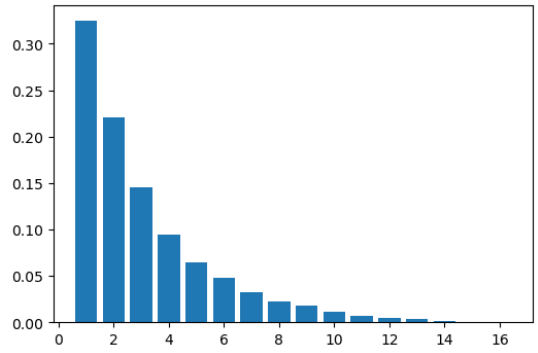


Mapa calor PCA 16 componentes

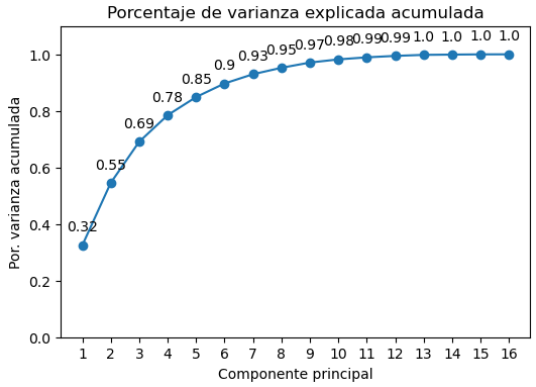


Periodo Mayo-Junio (2017-2021) para todas las edades

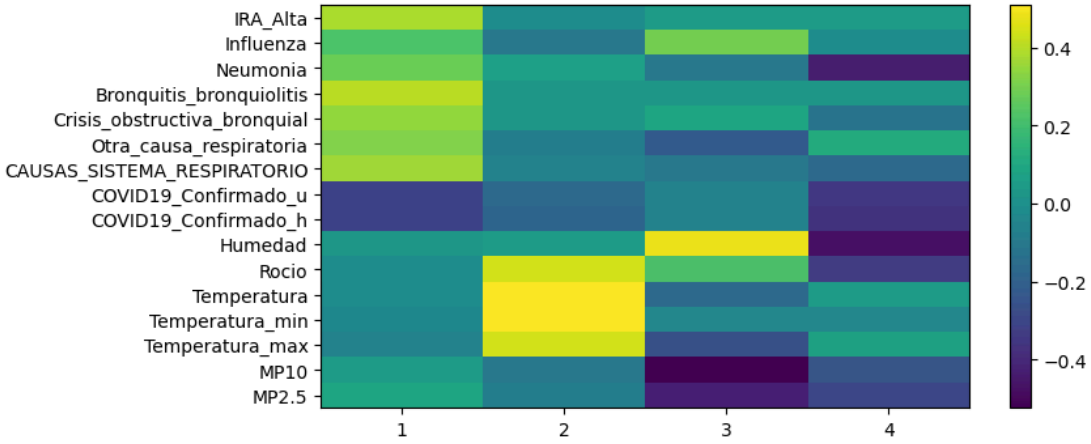
Varianza cada componente



Varianza explicada acumulada



Mapa calor PCA 4 primeras componentes



Periodo Julio-Agosto (2017-2021) para todas las edades

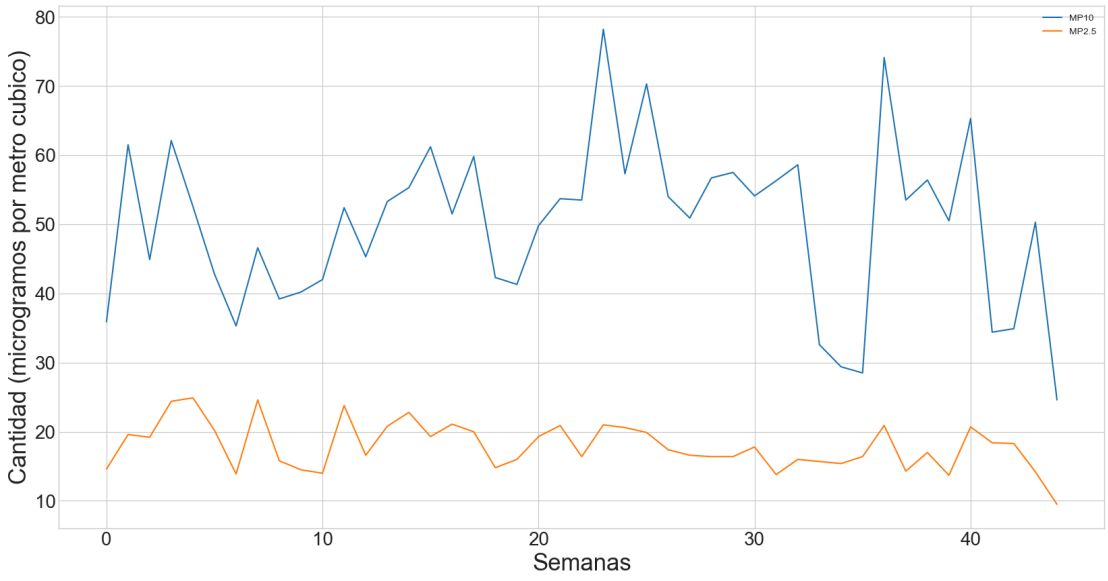
Media cada variable

```
----------------------
Media de cada variable
----------------------
IRA_Alta                        121.488889
Influenza                        19.555556
Neumonia                         13.400000
Bronquitis_bronquiolitis         59.088889
Crisis_obstructiva_bronquial     30.422222
Otra_causa_respiratoria          14.222222
CAUSAS_SISTEMA_RESPIRATORIO      15.600000
COVID19_Confirmado_u              7.088889
COVID19_Confirmado_h              3.688889
Humedad                          66.402222
Rocio                             7.342222
Temperatura                      14.486667
Temperatura_min                   9.391111
Temperatura_max                  22.582222
MP10                             50.020000
MP2.5                            17.953333
dtype: float64
```

Varianza cada variable

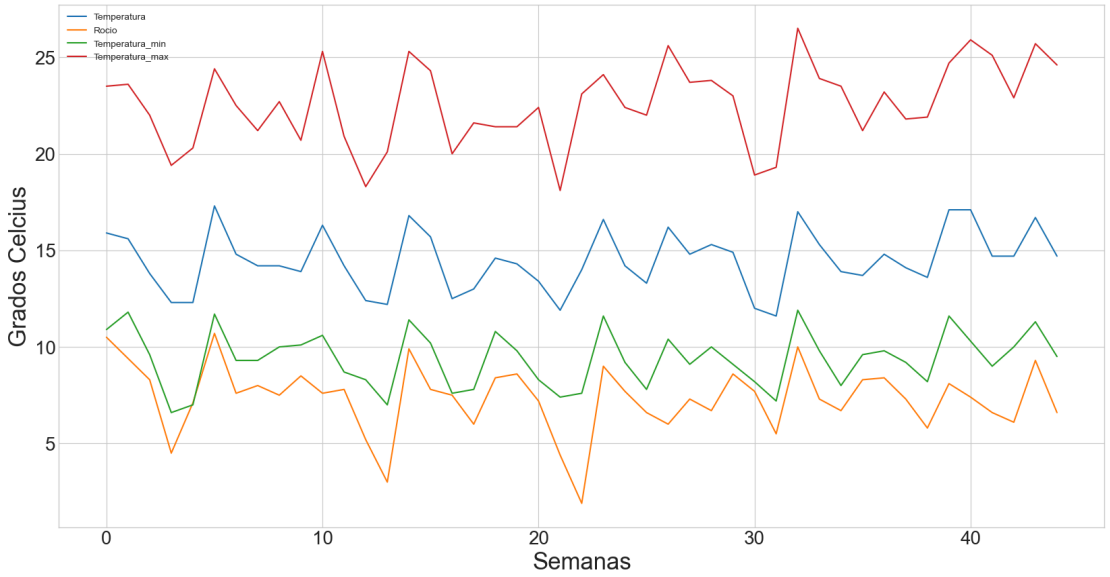
```
-------------------------
Varianza de cada variable
-------------------------
IRA_Alta                        6081.937374
Influenza                        577.570707
Neumonia                          93.927273
Bronquitis_bronquiolitis        2845.537374
Crisis_obstructiva_bronquial     419.749495
Otra_causa_respiratoria           68.858586
CAUSAS_SISTEMA_RESPIRATORIO       77.700000
COVID19_Confirmado_u             140.037374
COVID19_Confirmado_h              38.310101
Humedad                           21.714768
Rocio                              3.167949
Temperatura                        2.452091
Temperatura_min                    2.109465
Temperatura_max                    4.505586
MP10                             140.938455
MP2.5                             11.589364
dtype: float64
```

Variables contaminantes

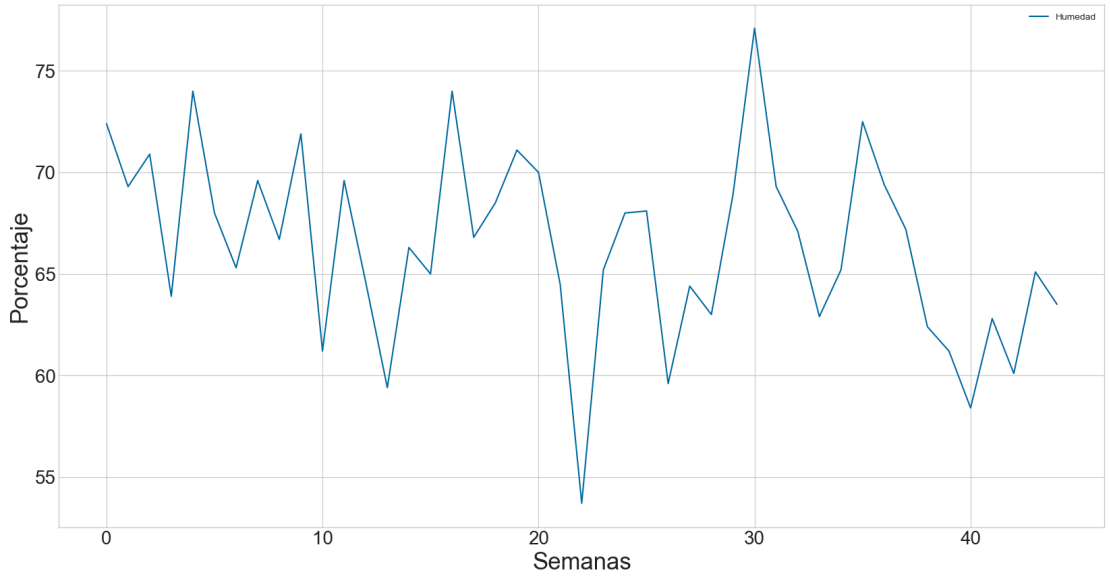


Periodo Julio-Agosto (2017-2021) para todas las edades

Variables ambientales

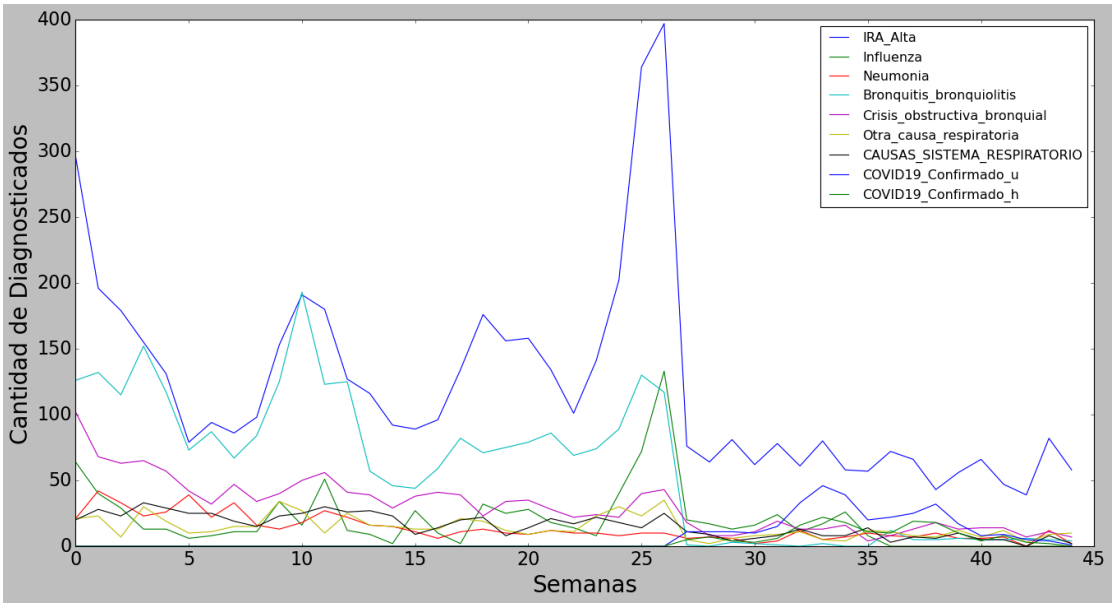


Humedad

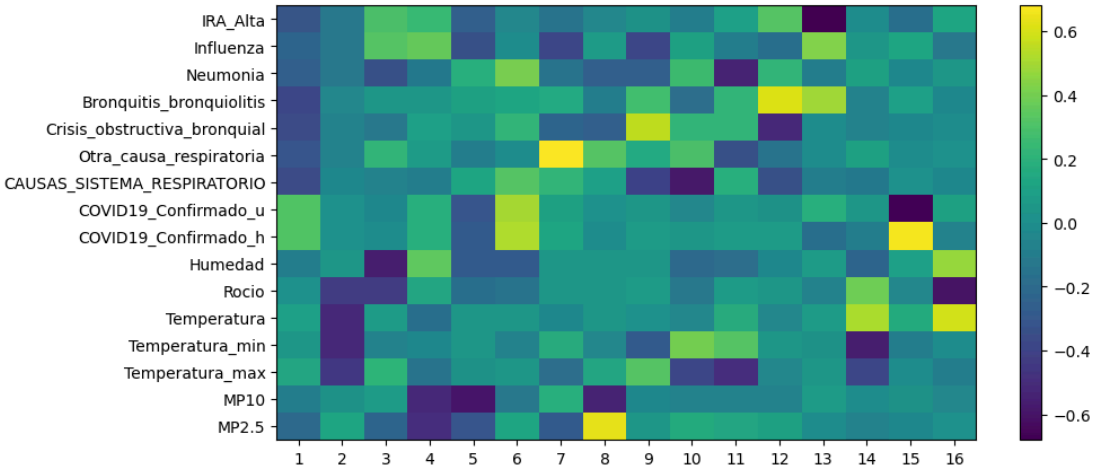


Periodo Julio-Agosto (2017-2021) para todas las edades

Variables enfermedades respiratorias

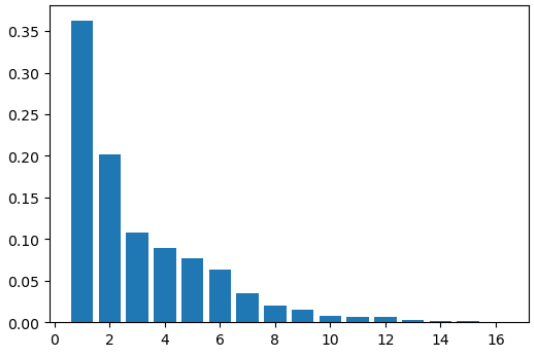


Mapa calor PCA 16 componentes

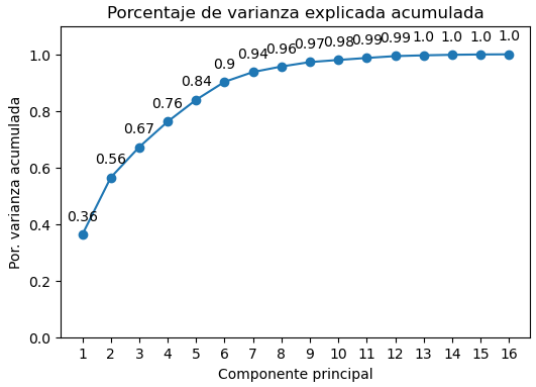


Periodo Julio-Agosto (2017-2021) para todas las edades

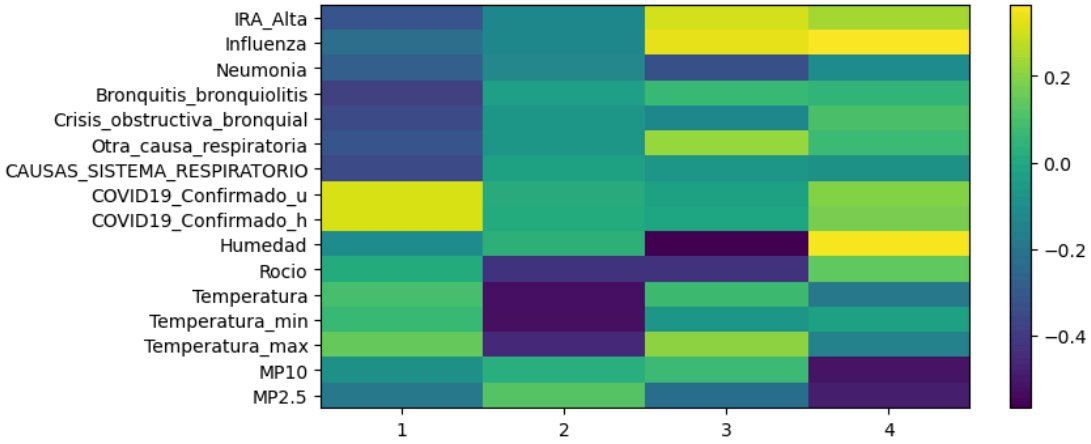
Varianza cada componente



Varianza explicada acumulada



Mapa calor PCA 4 primeras componentes



Periodo Septiembre-Octubre (2017-2021) para todas las edades

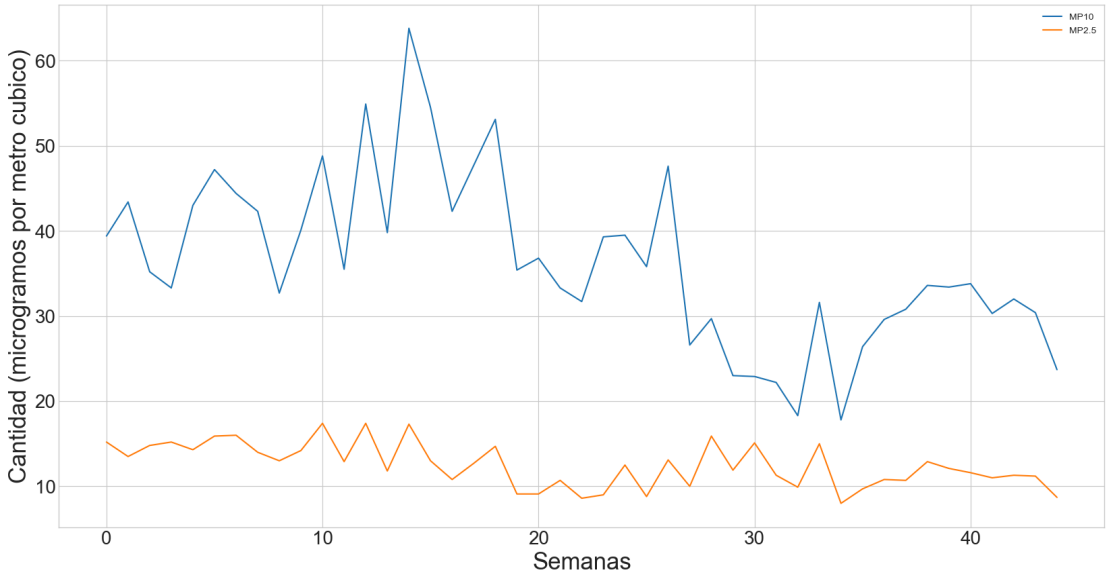
Media cada variable

```
----------------------
Media de cada variable
----------------------
IRA_Alta                        103.288889
Influenza                         7.977778
Neumonia                         10.133333
Bronquitis_bronquiolitis         45.933333
Crisis_obstructiva_bronquial     23.133333
Otra_causa_respiratoria          13.600000
CAUSAS_SISTEMA_RESPIRATORIO      11.066667
COVID19_Confirmado_u              2.355556
COVID19_Confirmado_h              1.200000
Humedad                          66.411111
Rocio                             8.622222
Temperatura                      15.677778
Temperatura_min                  10.555556
Temperatura_max                  24.251111
MP10                             36.375556
MP2.5                            12.491111
dtype: float64
```

Varianza cada variable

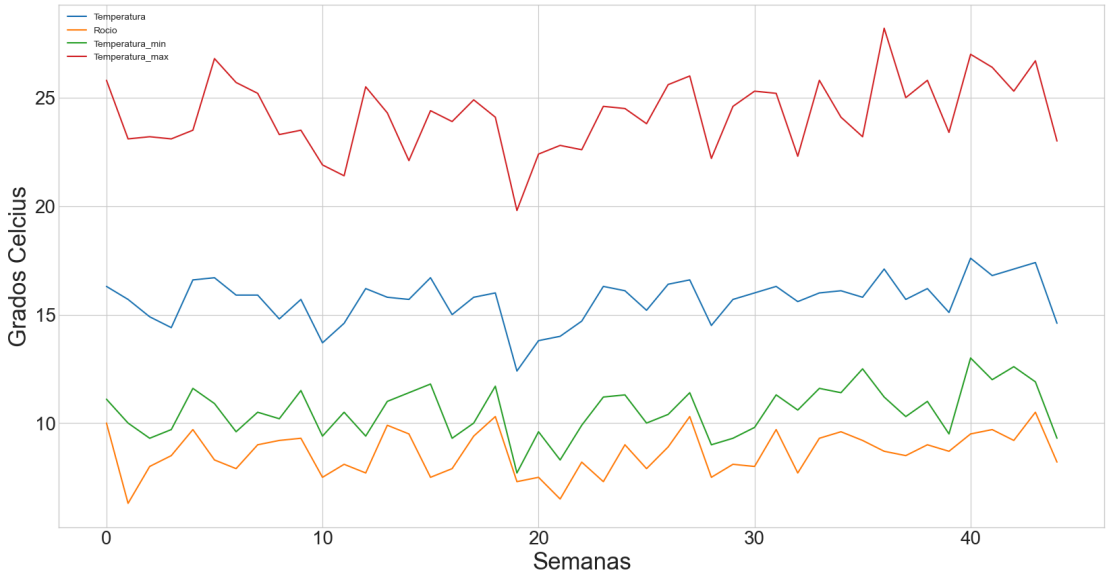
```
-------------------------
Varianza de cada variable
-------------------------
IRA_Alta                        4074.710101
Influenza                        161.840404
Neumonia                          35.981818
Bronquitis_bronquiolitis        1436.063636
Crisis_obstructiva_bronquial     174.209091
Otra_causa_respiratoria           29.427273
CAUSAS_SISTEMA_RESPIRATORIO       24.290909
COVID19_Confirmado_u              14.279798
COVID19_Confirmado_h               3.800000
Humedad                            9.983737
Rocio                              1.027677
Temperatura                        1.102222
Temperatura_min                    1.343434
Temperatura_max                    2.846192
MP10                             100.370071
MP2.5                              6.714919
dtype: float64
```

Variables contaminantes

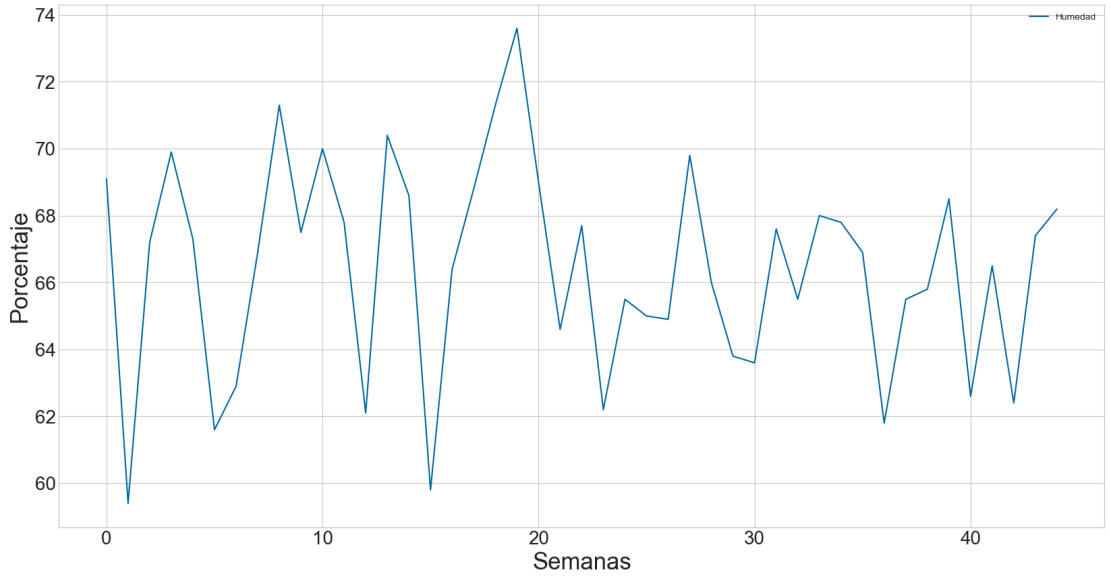


Periodo Septiembre-Octubre (2017-2021) para todas las edades

Variables ambientales

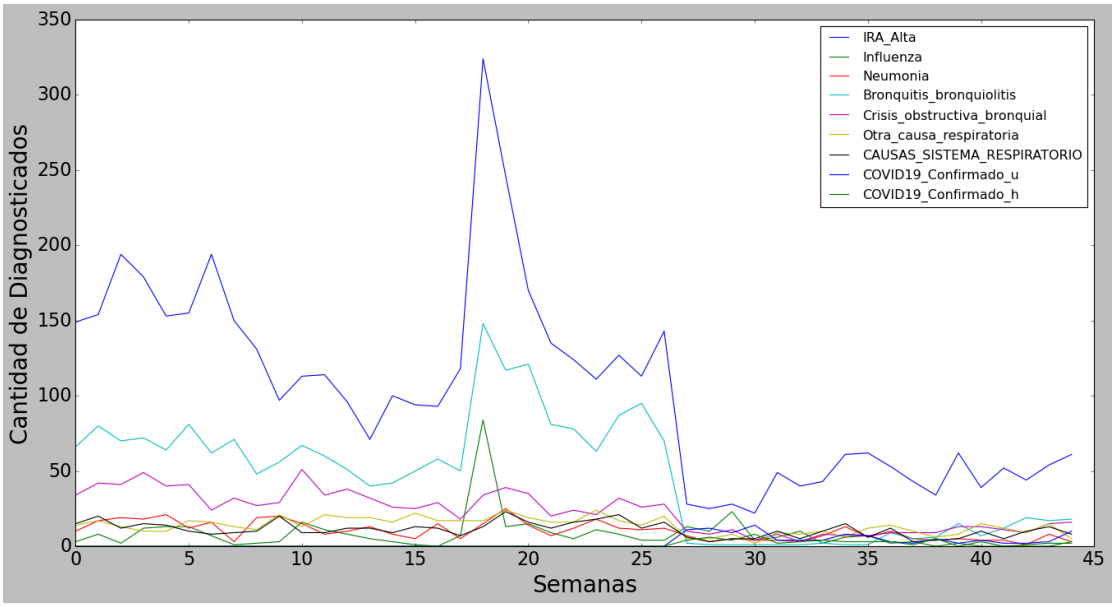


Humedad

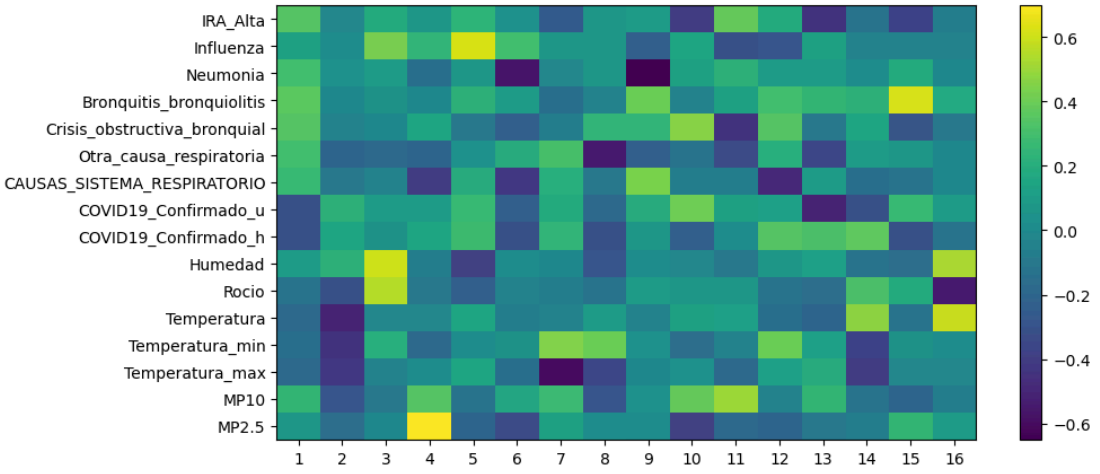


Periodo Septiembre-Octubre (2017-2021) para todas las edades

Variables enfermedades respiratorias

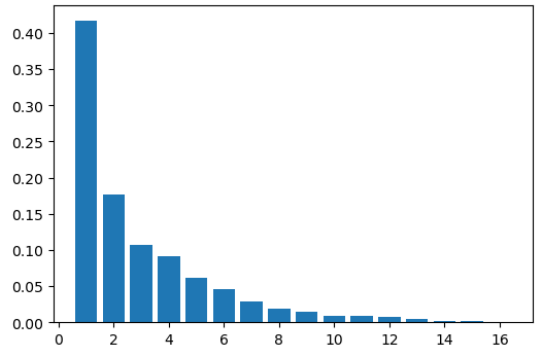


Mapa calor PCA 16 componentes

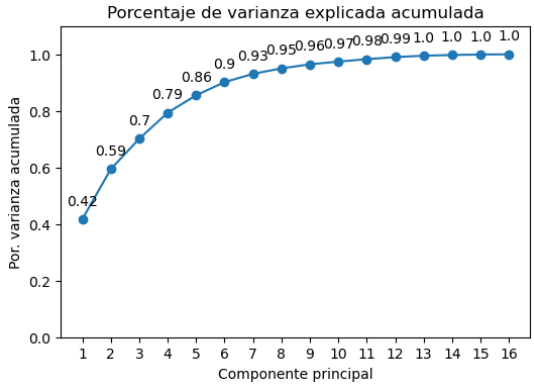


Periodo Septiembre-Octubre (2017-2021) para todas las edades

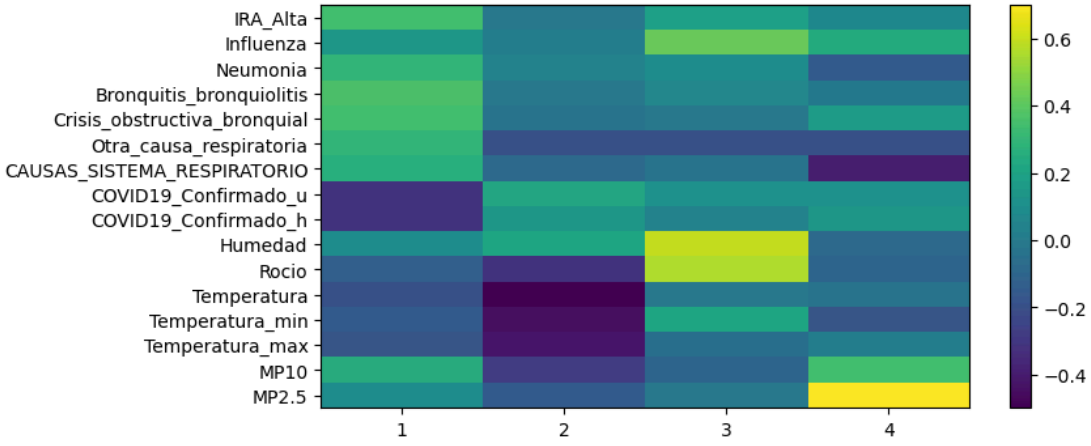
Varianza cada componente



Varianza explicada acumulada



Mapa calor PCA 4 primeras componentes



Periodo Noviembre-Diciembre (2017-2021) para todas las edades

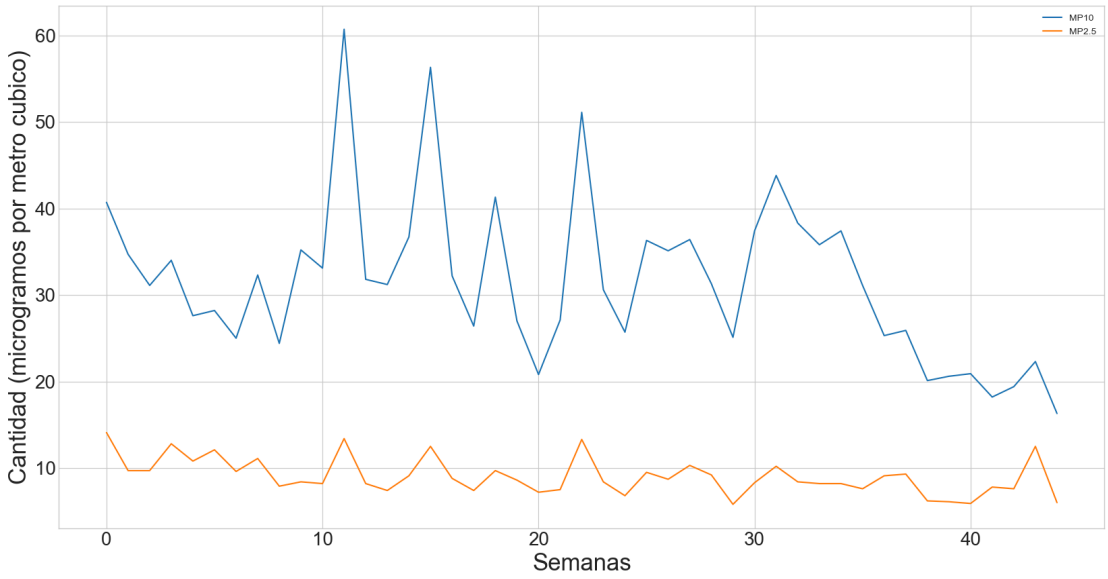
Media cada variable

```
----------------------
Media de cada variable
----------------------
IRA_Alta                          76.111111
Influenza                          2.844444
Neumonia                           6.533333
Bronquitis_bronquiolitis          24.177778
Crisis_obstructiva_bronquial      16.955556
Otra_causa_respiratoria           11.177778
CAUSAS_SISTEMA_RESPIRATORIO        7.711111
COVID19_Confirmado_u               1.555556
COVID19_Confirmado_h               0.933333
Humedad                           65.597778
Rocio                             10.406667
Temperatura                       17.664444
Temperatura_min                   12.657778
Temperatura_max                   25.744444
MP10                              31.604444
MP2.5                              9.057778
dtype: float64
```

Varianza cada variable

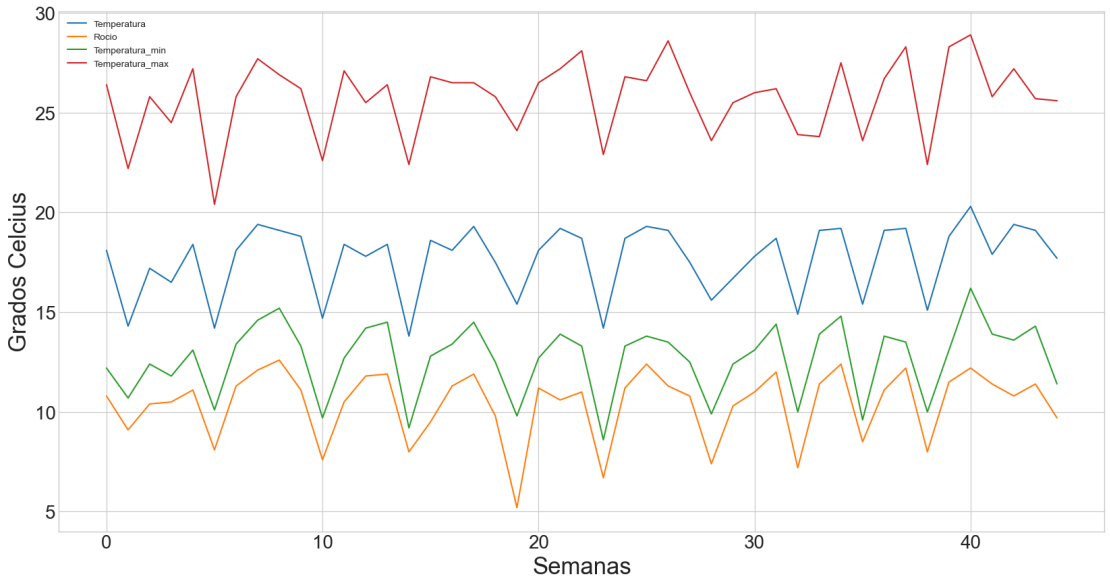
```
-------------------------
Varianza de cada variable
-------------------------
IRA_Alta                         613.737374
Influenza                          8.907071
Neumonia                          20.981818
Bronquitis_bronquiolitis         234.285859
Crisis_obstructiva_bronquial      42.225253
Otra_causa_respiratoria           15.376768
CAUSAS_SISTEMA_RESPIRATORIO       10.391919
COVID19_Confirmado_u               7.616162
COVID19_Confirmado_h               3.018182
Humedad                            7.812949
Rocio                              3.033364
Temperatura                        3.072343
Temperatura_min                    3.278859
Temperatura_max                    3.691162
MP10                              88.257253
MP2.5                              4.472949
dtype: float64
```

Variables contaminantes

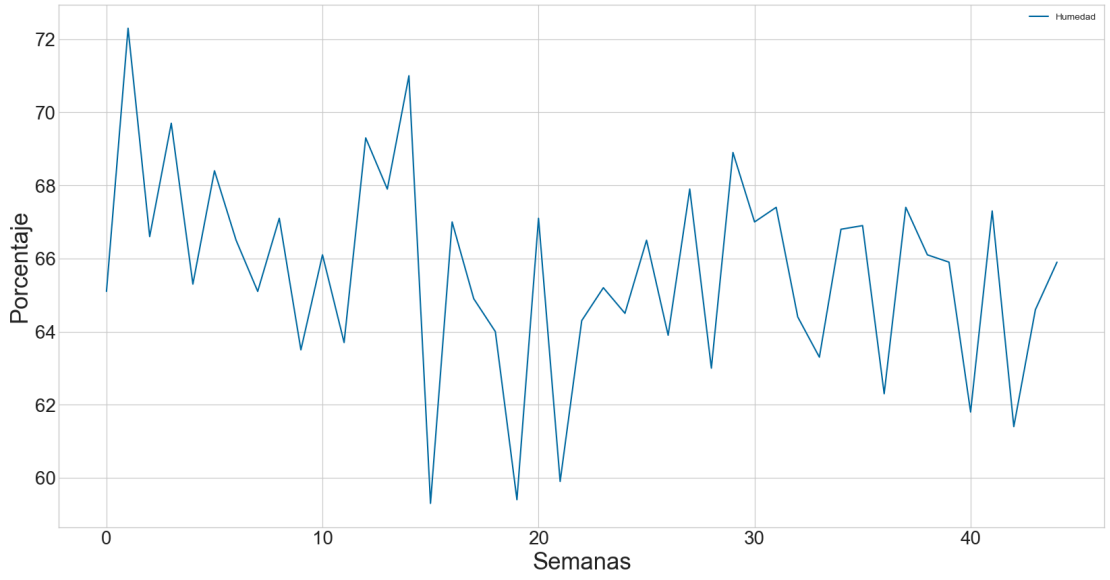


Periodo Noviembre-Diciembre (2017-2021) para todas las edades

Variables ambientales

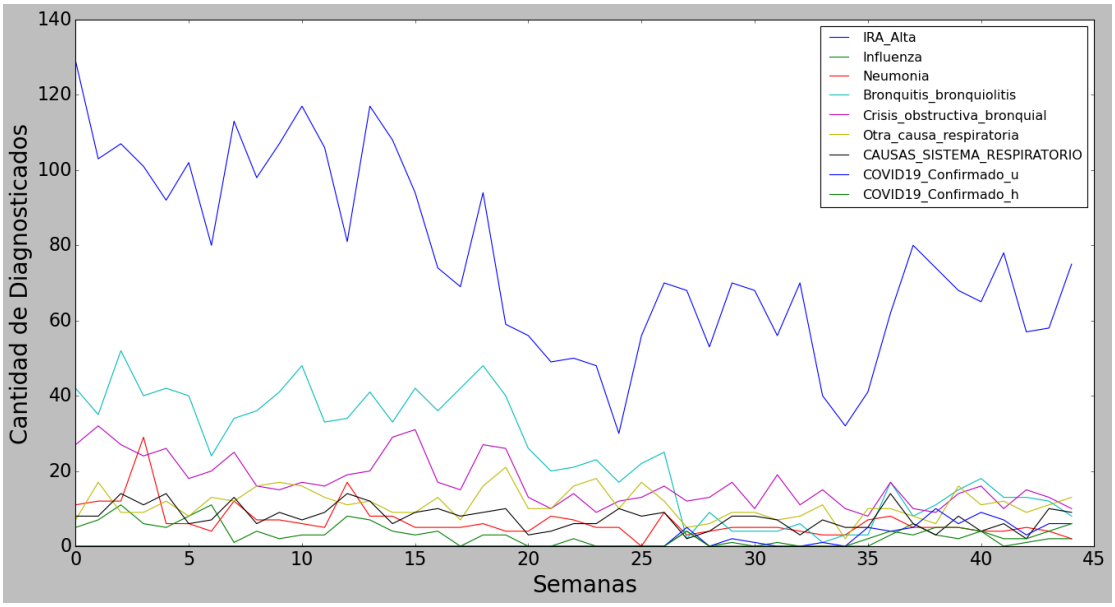


Humedad

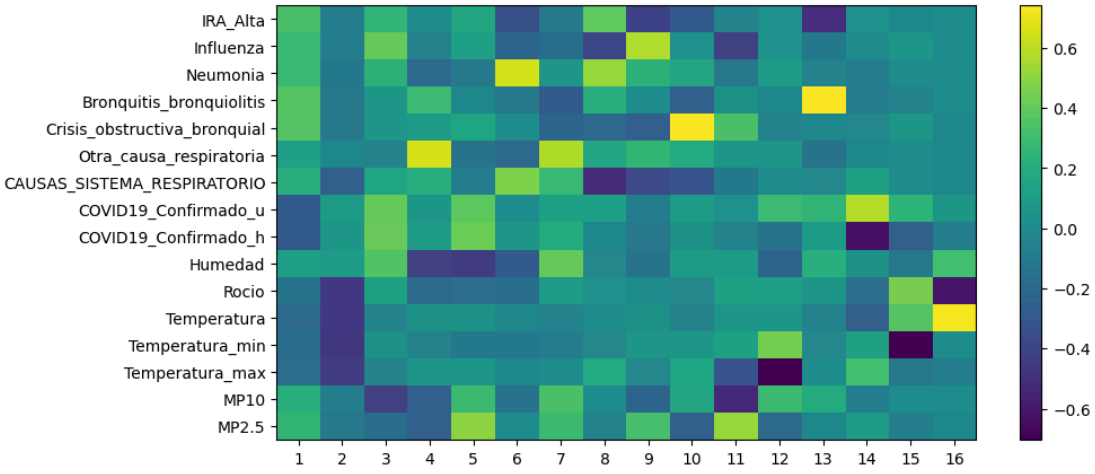


Periodo Noviembre-Diciembre (2017-2021) para todas las edades

Variables enfermedades respiratorias

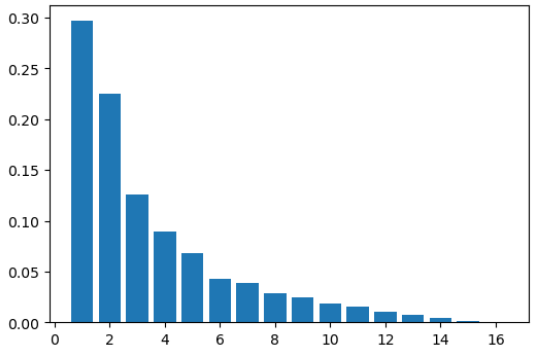


Mapa calor PCA 16 componentes

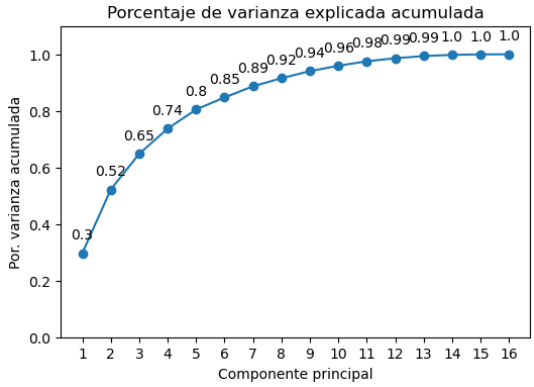


Periodo Noviembre-Diciembre (2017-2021) para todas las edades

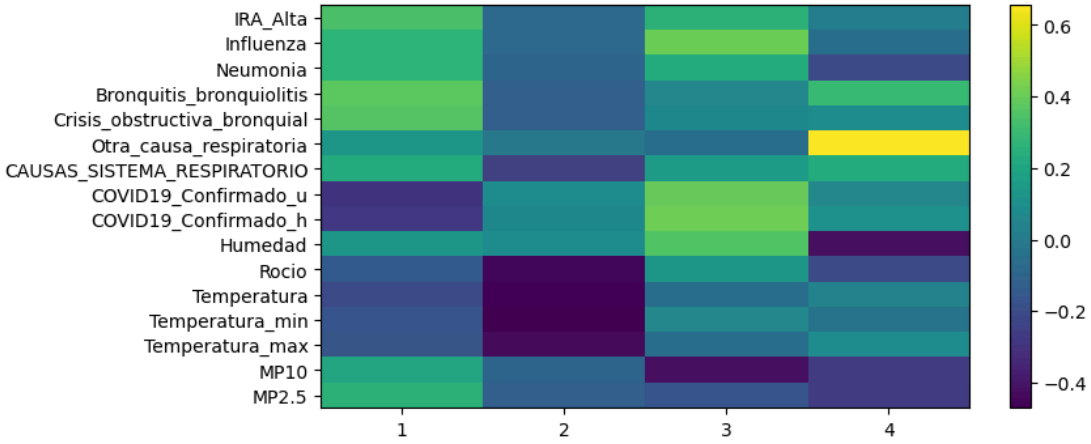
Varianza cada componente



Varianza explicada acumulada



Mapa calor PCA 4 primeras componentes



Anexo C CAPÍTULO X CIE-10

Infecciones agudas de las vías respiratorias superiores:

- J00 Rinofaringitis aguda [resfriado común]
- J01 Sinusitis aguda
- J02 Faringitis aguda
- J03 Amigdalitis aguda
- J04 Laringitis y traqueitis agudas
- J05 Laringitis obstructiva aguda [crup] y epiglotitis
- J06 Infecciones agudas de las vías respiratorias superiores, de sitios múltiples o no especificados

Influenza (gripe) y neumonía:

- J10 Influenza debida a virus de la influenza identificado
- J11 Influenza debida a virus no identificado
- J12 Neumonía viral, no clasificada en otra parte
- J13 Neumonía debida a Streptococcus pneumoniae
- J14 Neumonía debida a Haemophilus influenzae
- J15 Neumonía bacteriana, no clasificada en otra parte
- J16 Neumonía debida a oíros microorganismos infecciosos, no clasificados en otra parte
- J17 Neumonía en enfermedades clasificadas en otra parte
- J18 Neumonía, organismo no especificado

Otras infecciones agudas de las vías respiratorias inferiores:

- J20 Bronquitis aguda
- J21 Bronquiolitis aguda
- J22 Infección aguda no especificada de las vías respiratorias inferiores

Otras enfermedades de las vías respiratorias superiores:

- J30 Rinitis alérgica y vasomotora
- J31 Rinitis, rinofaringitis y faringitis crónicas
- J32 Sinusitis crónica
- J33 Pólipo nasal
- J34 Otros trastornos de la nariz y de los senos paranasales
- J35 Enfermedades crónicas de las amígdalas y de las adenoides
- J36 Absceso periamigdalino
- J37 Laringitis y laringotraqueitis crónicas
- J38 Enfermedades de las cuerdas vocales y de la laringe, no clasificadas en otra parte
- J39 Otras enfermedades de las vías respiratorias superiores
- Enfermedades crónicas de las vías respiratorias inferiores
- J40 Bronquitis, no especificada como aguda o crónica
- J41 Bronquitis crónica simple y mucopurulenta

- J42 Bronquitis crónica no especificada
- J43 Enfisema
- J44 Otras enfermedades pulmonares obstructivas crónicas
- J45 Asma
- J46 Estado asmático
- J47 Bronquiectasia

Enfermedades del pulmón debida a agentes externos:

- J60 Neumoconiosis de los mineros del carbón
- J61 Neumoconiosis debida al asbesto y a otras fibras minerales
- J62 Neumoconiosis debida a polvo de sílice
- J63 Neumoconiosis debida a otros polvos inorgánicos
- J64 Neumoconiosis, no especificada
- J65 Neumoconiosis asociada con tuberculosis
- J66 Enfermedades de las vías aéreas debidas a polvos orgánicos específicos
- J67 Neumonitis debida a hipersensibilidad al polvo orgánico
- J68 Afecciones respiratorias debidas a inhalación de gases, humos, vapores y sustancias químicas
- J69 Neumonitis debida a sólidos y líquidos
- J70 Afecciones respiratorias debidas a otros agentes externos

Otras enfermedades respiratorias que afectan principalmente al intersticio:

- J80 Síndrome de dificultad respiratoria del adulto
- J81 Edema pulmonar
- J82 Eosinofilia pulmonar, no clasificada en otra parte
- J84 Otras enfermedades pulmonares intersticiales
- Afecciones supurativas y necróticas de las vías respiratorias inferiores
- J85 Absceso del pulmón y del mediastino
- J86 Piotorax
- Otras enfermedades de la pleura
- J90 Derrame pleural no clasificado en otra parte
- J91 Derrame pleural en afecciones clasificadas en otra parte
- J92 Paquipleuritis
- J93 Neumotorax
- J94 Otras afecciones de la pleura

Otras enfermedades del sistema respiratorio:

- J95 Trastornos del sistema respiratorio consecutivos a procedimientos, no
- J96 Insuficiencia respiratoria, no clasificada en otra parte
- J98 Otros trastornos respiratorios
- J99 Trastornos respiratorios en enfermedades clasificadas en otra parte