UNIVERSIDAD
DE ATACAMA

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

Análisis del comportamiento de las enfermedades respiratorias en la comuna de Copiapó, utilizando algoritmos de clustering.

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

Profesor Guía: Mg. Andrés Alfaro Avalos

Luis Gonzalo Espejo Tapia

Copiapó, 2022, Chile.

UNIVERSIDAD
DE ATACAMA

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

Análisis del comportamiento de las enfermedades respiratorias en la comuna de Copiapó, utilizando algoritmos de clustering.

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

Profesor Guía:
Mg. Andrés Alfaro Avalos

Miembros del Comité:
Mg. Servando Campillay Briones
Dr. Wilson Castillo Rojas

Luis Gonzalo Espejo Tapia

Copiapó, 2022, Chile.

Esta tesis se la dedico a mis padres,
que vean reflejado el fruto
de su trabajo día a día.

Agradecimientos

Agradezco a mi madre y mi padre por apoyarme de manera incondicional durante mi formación como profesional. Agradezco también a mis amigos dentro de la universidad quiénes confiaron en mí y me ayudaron a pasar desafíos tanto académicos como personales, en especial a mi mejor amiga Anita, quien me animó a no rendirme y me acompañó sin condición en cada proceso de mi formación.

También agradezco de gran manera la enseñanza que me transmitieron mis profesores durante todos los años de mi paso por esta hermosa carrera, que con dedicación y paciencia me educaron. De manera especial al profesor Andrés Alfaro por creer en mí y ofrecerme esta investigación, por asistir y guiarme en este trabajo.

Resumen

Las enfermedades que afectan al sistema respiratorio son unas de las principales causas de muertes humanas alrededor del mundo. Estas enfermedades derivan en efectos negativos a los pulmones y a otros órganos del sistema respiratorio. Comúnmente estas enfermedades respiratorias suelen ser causadas por infecciones respiratorias, consumo de tabaco, inhalación de humo, entre otras formas de contaminación del aire.

A través del tiempo, el aumento de casos de las enfermedades respiratorias y el crecimiento exponencial respecto a la cantidad de datos generados por las entidades relacionadas al área de la salud en Chile y la oportunidad que nos ofrecen las nuevas tecnologías como la minería de datos o la inteligencia artificial, permiten combinar la salud y la informática en este trabajo.

Bajo este contexto es que la presente investigación se fundamenta en la aplicación de minería de datos por medio del algoritmo de clustering K-Means que posibilita la generación de un modelo descriptivo acerca del comportamiento de las enfermedades respiratorias en la comuna de Copiapó en Chile.

El conjunto de datos abarca un periodo de 62 semanas correspondientes a los meses desde 29 de diciembre del año 2019 al 6 de marzo del año 2021, donde se consideran un total de 16 variables. Estas corresponden a 8 variables que representan a enfermedades respiratorias (incluyendo COVID-19), además de 5 que corresponden a variables climáticas y otras 3 que corresponden a contaminantes. Utilizando las bases de datos proporcionadas por las entidades correspondientes, y los grupos de edad predefinidos por estas bases de datos, se cuenta con un total de 9.528 pacientes divididos por rango de edad, con 391 Lactantes, 693 Primera infancia, 661 Infantes, 6.530 Jóvenes y adultos, y 1.253 Adultos mayores.

Además la investigación se lleva a cabo utilizando una metodología paso a paso, que permite generar clústeres de la información, además de permitir un análisis de los resultados y la relación de las enfermedades respiratorias con las variables climáticas y variables contaminantes.

Índice General

Índice de Figuras

Índice de tablas

Capítulo I
Introducción

En el año 2006, la Organización Mundial de la Salud (OMS) promueve la Alianza Global contra las Enfermedades Respiratorias (GARD, *global alliance against chronic respiratory deseases* del inglés), con el objetivo de sanar y mejorar la calidad de vida de las personas afectadas por enfermedades crónicas y respiratorias. Para el año 2017 y según lo expuesto por la Asociación Latinoamericana de Tórax (2017) más de 65 millones de personas habrían sufrido de enfermedades pulmonares y 334 millones de asma, siendo esta la enfermedad crónica más común en la niñez. Todo esto causado principalmente por la exposición a contaminantes que van enfermando a miles de millones de seres humanos, posicionando a estas enfermedades como una de las principales causas de muerte a nivel mundial.

En Chile la situación no es muy diferente, según lo expuesto por R. González, R. Pinto y J.P. Álvarez (2017), nuestro país ha tenido una larga lucha contra las enfermedades respiratorias. En los años 30 la tuberculosis fue la causante de 15.000 muertos; la pulmonía fue el verdugo en la década de los 60; y a finales de los años 80 la mortalidad por neumonía es 17 cada 1000 habitantes, cifras muy altas para esos tiempos. En la actualidad, Chile comienza a reaccionar y en 1993 se implementan las salas IRA (Infecciones Respiratorias Agudas) en todo el país, para luego desarrollar planes como el AUGE/GES, y campañas de prevención (M. Barros Monge, 2005). Todo esto, con la finalidad de disminuir cualquier causalidad negativa que pueden provocar estas enfermedades

Hace un tiempo atrás el gobierno lanza la Estrategia Nacional de Salud, que consiste en un plan que traza los objetivos sanitarios de Chile durante una década, desde el año 2011 al 2020. En esta reforma, se destaca el objetivo sanitario, "Fortalecimiento del Sector Salud", que promueve la mejora sustancial de los sistemas de información en salud, dando hincapié en mejorar la gestión de los datos y por lo tanto de la información que se genera, destacando que "*En el ámbito de la salud pública, esta información permite caracterizar la población, evaluar la magnitud, frecuencia y gravedad de diversos fenómenos que afectan a la salud de las personas, y contribuye a formular políticas sanitarias, ya sea con medidas preventivas o de respuesta de servicios de salud*" (Estrategia Nacional de Salud, 2010). Así mismo, se destaca la gran cantidad de datos que se generan en este sector y la importancia que tiene la información inferida, para ser una alternativa importante en la lucha contra estas enfermedades.

Según lo expuesto anteriormente, el gran volumen de datos que dispone este sector, sumado al mejoramiento y avances en el procesamiento de la información, ofrecen una oportunidad valiosa y única para aportar de manera innovadora en la prevención y mitigación de las consecuencias que pueden causar las enfermedades respiratorias. Para esto, emergen nuevas disciplinas computacionales capaces de manejar y trabajar con estos datos, destacando el *Big Data* (BD), tecnología capaz de gestionar y extraer conocimiento que se genera en grandes

cantidades de datos. Según E. Menasalvas y C. Rodríguez-González (2017), si Estados Unidos utilizara este tipo de tecnologías puede eventualmente disminuir sus gastos en salud en un 8%. Por otro lado, la Inteligencia Artificial (IA) busca emular los procesos de la inteligencia humana, a través de máquinas computacionales. Estos procesos, requieren gran cantidad de datos y reglas algorítmicas que permiten procesar esta información. La IA camina a convertirse en la tecnología con mayor impacto en la humanidad. Y por último, la Minería de Datos (*Data Mining*), consiste en el análisis y manejo de grandes volúmenes de datos, permitiendo tomar decisiones importantes, como por ejemplo en el sector salud, diagnósticos y tratamientos. En síntesis, una serie de algoritmos inteligentes, capaces de extraer valor de los datos y así mejorar la calidad de vida de las personas, y en este caso de los que padecen enfermedades respiratorias.

Bajo este contexto, el presente trabajo de apoyo a la investigación, busca generar las primeras aproximaciones en el descubrimiento de patrones o comportamiento particular que pueden tener enfermedades respiratorias en la comuna de Copiapó, a través del algoritmo de Clustering con el objetivo de generar nuevo conocimiento y en un futuro mejorar las acciones de prevención y así anticiparse a las consecuencias que causan estas patologías en la población. Tomando en cuenta lo expuesto por la Asociación Latinoamericana de Tórax (2017), sobre la estrecha relación entre las enfermedades respiratorias y variables medioambientales, es que, se utilizan además, datos ambientales del aire (temperatura, punto de rocío y humedad) y datos de contaminación del aire (materiales particulados y SO_2) para describir con mayor detalle el comportamiento de este tipo de enfermedades.

1.1. Objetivos

A continuación se detalla el objetivo general y los objetivos específicos que se pretenden lograr con esta investigación

1.1.1. Objetivo General

El objetivo general de esta investigación es analizar el comportamiento de las enfermedades respiratorias en la comuna de Copiapó utilizando algoritmos de clustering.

1.1.2. Objetivos Específicos

- Diseñar y construir una estructura de datos.
- Preparar y explorar los datos.
- Determinar modelo de agrupamiento considerando rango etario y tipo de variables.
- Desarrollar una implementación de algoritmo de clustering.
- Interpretar y evaluar los resultados preliminares obtenidos

1.2. Resultados Preliminares

Como resultados preliminares de esta investigación se espera en primera instancia una exploración de datos, para luego obtener una primera aproximación al comportamiento de las enfermedades respiratorias en la comuna de Copiapó, a través del desarrollo de un proceso de minería de datos, utilizando técnicas de clustering y de visualización de datos.

Capítulo II
Marco teórico

El siguiente capítulo contiene los conceptos claves para entender el contexto de este trabajo de investigación. En primer lugar se expone una base teórica de las enfermedades respiratorias, variables ambientales y contaminantes y el impacto que tienen estas en la salud del ser humano. Para luego, tener un acercamiento teórico a las tecnologías de análisis en especial a la minería de datos, profundizando en conceptos y herramientas relacionados a esta disciplina.

2.1. Enfermedades respiratorias

Según la Organización Mundial de la Salud (OMS), las enfermedades respiratorias son todas aquellas 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 crónica*" (W.H Organization, OMS, 2015). Además según la Asociación Latinoamericana de Tórax (2017) este tipo de patologías afecta a más de mil millones de personas a nivel mundial y es la tercera causa de muerte a nivel nacional, causada principalmente por agentes contaminantes. Esta enfermedad ha requerido una larga lucha e incontables esfuerzos para aplacar sus efectos. Así lo presentan R. Gonzales, R. Pinto y J.P. Alvarez (2017) que describen como Chile ha enfrentado durante su historia los ataques de estas patologías.

En la actualidad la lucha continúa. Así lo demuestran una serie de medidas y planes como el AUGE/GES, campañas de prevención y vacunación, programas de vigilancia epidemiológica y La Estrategia Nacional de Salud para los años entre 2011 al 2020. Este último documento, es una guía que propone una serie de metas a lograr en la salud pública. Contempla dentro de sus objetivos el "Fortalecer la institucionalidad del sector salud", que busca entre otras cosas fomentar la investigación y mejorar los sistemas de información en salud, lo que conlleva a disponer de datos que permite, por una parte, apoyar la toma de decisión y coordinar la organización y, por otra parte, describir y evaluar los diferentes fenómenos que afectan a la salud de las personas.

Para poder clasificar este tipo de enfermedades, existe el sistema denominado Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud, conocido mundialmente por sus siglas CIE-10 (CIE-10 ES). Este sistema clasifica a través de códigos diferentes tipos de diagnósticos, contiene 22 capítulos que permiten así, organizar las distintas enfermedades. En particular, el capítulo X, está dedicado a clasificar las enfermedades respiratorias o "Enfermedades del aparato respiratorio", clasificándolas desde la J00 - J99 (Ver Tabla 1). Para este trabajo se consideran las enfermedades respiratorias presentes en este capítulo, 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 a las definiciones de estas enfermedades también se puede ver el Anexo A, correspondiente al capítulo X Enfermedades del aparato respiratorio.

Tabla 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 variables 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).

Algunos trabajos determinan la existencia de una estrecha relación entre las variables ambientales y las enfermedades respiratorias tanto en las zonas frías como las tropicales, mientras en las zonas frías se presenta un aumento de casos de enfermedades respiratorias en los meses de invierno (Dowell SF, Whitney CG, Wright C, Rose CE Jr, Schuchat, 2003), en las zonas tropicales aumentan estos casos en los meses de lluvia. (Omer SB, Sutanto A, Sarwo H, Linehan M, Djelantik IG, Mercer D, et al., 2008).

Por otra parte, estas variables ambientales son medidas a través de estaciones climáticas las cuales cuentan con distintos instrumentos para el registro de estas variables. En el plano nacional la entidad encargada de estas estaciones de medición son pertenecientes a la Dirección Meteorológica de Chile la cual pertenece a su vez a la Dirección General de Aeronáutica Civil (DGAC).

Las variables ambientales consideradas para esta investigación son temperatura, rocío, humedad, temperatura mínima y temperatura máxima. No se utiliza la variable precipitación, ya que, en el periodo de análisis para esta investigación (29 de diciembre del 2019 al 6 de marzo del 2021) no se registran lluvias en la zona.

Para tener un mayor acercamiento a estas variables ambientales, sus definiciones y unidades de medidas ver el Anexo B, el cual corresponde al diccionario de datos de esta investigación.

2.3. Variables contaminantes

Las variables contaminantes son aquellas variables que 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 (O_3), el dióxido de nitrógeno (NO_2) y el dióxido de azufre (SO_2).

A nivel global, bajo el reporte especial de la Agencia Internacional de Energía (IEA, 2016), se informa que cada día 18.000 personas mueren por causas asociadas a la contaminación del aire, un total de 6,5 millones de personas por año. Por su parte la OMS (WHO, 2016), menciona que la contaminación es un aspecto preocupante, sobre todo en los países en vías de desarrollo y países desarrollados donde las emisiones han aumentado a causas de la combustión, las industrias, alto tráfico vehícular, entre otras. Esto ha provocado el aumento en la mala calidad del aire, especialmente en las zonas urbanas del país. Este aumento de la contaminación ambiental es preocupante para el área de la salúd pública debido a que existen estudios epidemiológicos que relacionan directamente la exposición a material particulado (MP2.5 y MP10) con enfermedades respiratorias, cardiovasculares y salud mental.

El material particulado es una de las principales variables contaminantes en el aire, se reconocen 2 tipos de este material dependiendo de su medida en micras, por un lado el material particulado de 2.5 micras (MP2.5) provienen en su mayoría de las emisiones de los vehículos diésel, mientras que el material particulado de 10 micras (MP10) es proveniente en su mayoría de polvo u origen natural (Legarreta et al., 2015).

Por otro lado, los efectos que producen las partículas en la salud de las personas han estado históricamente asociados a la exacerbación de las enfermedades respiratorias e incluso del tipo cardiovascular (C. Linares Gil y J. Díaz Jiménez, 2008).

En el plano nacional, la entidad encargada de las estaciones de medición de estas variables contaminantes es el Sistema de Información Nacional de Calidad del Aire (SINCA). Para

encontrar las variables de contaminación y sus definiciones, que se utilizaron en este trabajo se recomienda ver el Anexo B, correspondiente al diccionario de datos.

Las variables ambientales consideradas para esta investigación son material particulado de 2.5 micras y 10 micras, y dióxido de azufre. No se consideran las variables de ozono debido a que no existen registros para el periodo investigado, y tampoco la variable dióxido de nitrógeno debido a que sus registros son escasos para el periodo de tiempo que abarca la investigación.

Para tener un mayor acercamiento a estas variables contaminantes, sus definiciones y unidades de medidas ver el Anexo B, el cual corresponde al diccionario de datos de esta investigación.

2.4. Tecnologías de análisis de datos

La irrupción de nuevas tecnologías de datos crea oportunidades y abre el paso a la innovación, elementos y técnicas únicas en su estilo que dejan obsoletas a las tecnologías utilizadas previas a estas. En lo que respecta al área de la salud se puede hablar de distintas tecnologías disruptivas que llegaron a cambiar la escena, algunos ejemplos son; la Inteligencia Artificial (IA), *Big Data* y la Minería de Datos (*Data Mining)*.

En la actualidad, el sector del área de la salud genera un gran volumen de datos, esto es consecuencia principalmente de la transformación digital e implementación de herramientas electrónicas que permiten generar esta gran cantidad de datos (Bellinger, M. Jabbar, O. Zaïane, A. Osornio-Vargas, 2017).

En este contexto, el uso de tecnologías y herramientas para tratar, procesar y obtener información de esta nube de datos, está tomando cada vez mayor relevancia. No obstante, a nivel nacional el uso de esta tecnología no es frecuente (Estrategia nacional de salud, 2010).

A continuación se describen las tecnologías y herramientas que se utilizaron en este trabajo.

2.4.1. Minería de datos

La Minería de Datos es un conjunto de metodologías y herramientas que mediante el análisis de grandes volúmenes de datos ayudan a obtener patrones de comportamiento o tendencias invisibles que pueden ser de gran ayuda en la toma de decisiones (Marchán E, Salcedo J, Aza T, Figuera L, Martínez de Pisón F, G. P., 2011).

Para lograr buenos resultados, se debe comprender que la minería de datos al igual que la programación no se basa en una metodología estándar que resuelva cualquier problema que se presente, sino, más bien consiste en una metodología dinámica e iterativa que dependerá del problema, de las fuentes de datos, del conocimiento de herramientas, de la metodología desarrollada y los recursos que se tenga (Marchán E, Salcedo J, Aza T, Figuera L, Martínez de Pisón F, G. P., 2011).

2.4.2. Técnicas de Minería de datos

Las técnicas de Minería de Datos pueden clasificarse en: técnicas de modelado originado por la teoría (Técnicas Predictivas), técnicas de modelo originado por los datos (Técnicas Descriptivas) y en técnicas auxiliares (Técnicas auxiliares o Prescriptivas).

Por un lado las técnicas de modelado originado por la teoría especifican el modelo para los datos con base en un conocimiento previo que lleva a la predicción. En cambio el modelo originado por los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como válido (M. Pérez, 2014).

Como se menciona, las técnicas de Minería de Datos se pueden clasificar, de la siguiente manera.

2.4.2.1. Técnicas Predictivas

Especifican el modelo para los datos en base a un conocimiento técnico previo. Estas técnicas se utilizan para prever el comportamiento futuro de alguna entidad (J. Molina & García, 2008).

2.4.2.2. Técnicas Descriptivas

En esta técnica no se asigna ningún rol predeterminado a las variables. Tampoco se emplean situaciones en la existencia de variables dependientes ni independientes, y tampoco supone la existencia de un modelo previo a los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones (C. P. López, 2017).

Dentro de las Técnicas Descriptivas se encuentran los algoritmos de agrupamiento (*clustering*), estos algoritmos corresponden a una técnica de análisis de datos que ayudan a resolver problemas de clasificación, que intentan identificar agrupaciones de datos (*clústeres*) de acuerdo a una medida que puede ser de distancia o de similitud entre ellos (Rendón et al, 2015).

2.2.2.3. Técnicas Auxiliares o Prescriptivas

El análisis utilizando técnicas prescriptivas permite simular escenarios para optimizar resultados, estos análisis son desarrollados utilizando algoritmos de aprendizaje de máquina y técnicas estadísticas (Oviedo, Oviedo, & Vélez, 2015).

Además las técnicas prescriptivas corresponden a técnicas más superficiales y limitadas, es decir, que responden a nuevos métodos pero basados en técnicas estadísticas descriptivas e informes (M. Pérez, 2014).

2.5. Algoritmo de Clustering

El *Clustering* (o algoritmo de agrupamiento) consiste en agrupar una serie de datos según un criterio en grupos o clústeres. Generalmente el criterio suele ser la similitud. Está considerado

como una Técnica Descriptiva además de un aprendizaje no supervisado dentro de la minería de datos (E. J. Blanco-Herminda Sanz, 2016).

Según M. Garre, J. Cuadrado y M. Sicilia (2007); una gran variedad de algoritmos de clustering han surgido en los últimos años, los cuales se clasifican de la siguiente manera:

- Métodos Jerárquicos
    - Algoritmos Aglomerativos
    - Algoritmos Divisivos
- Métodos de Particionado y Recolocación
    - Clustering Probabilístico
    - Método de *K-medoids*
    - Método de *K-means*
    - Algoritmos Basados en Densidad
- Métodos Basados en Rejillas
- Métodos Basados en la Co-Ocurrencia de Datos Categóricos
- Clustering Basado en Restricciones
- Algoritmos para Datos de Grandes Dimensiones
- Clustering Subespacial
- Técnicas de Co-Clustering

Jiawei Hand And Micheline Kamber (2006) por otro lado simplifican la clasificación de algoritmos de agrupamiento de la siguiente manera:

- Métodos basados en particiones
- Métodos Jerárquicos
- Métodos basados en modelos

Además Jiawei Hand and Micheline Kamber (2006) indican que el resto de métodos y algoritmos que no entran en esta clasificación son nada más que combinaciones de los anteriores. A continuación se definen los distintos métodos mencionados, los cuales corresponden a los tres métodos comúnmente más utilizados cuando se trabaja con algoritmos de *clustering*.

2.5.1. Métodos basados en particiones

P. Larrañaga, I. Inza y A. Moujahid (2012) explican que en el *clustering* particional el objetivo es obtener una partición de los objetos en *clústeres* de tal manera que todos los objetos pertenezcan a algunos de los *K* posibles *clústeres* y que por otra parte sean disjuntos.

Los algoritmos basados en particiones asignan a un conjunto de datos *K* grupos sin estructura jerárquica, siendo K un número menor que el número total de objetos.

Dentro de los algoritmos basados en particiones, los más utilizados son *K-means* y *K-medoids.* A continuación se describen brevemente cada uno de ellos.

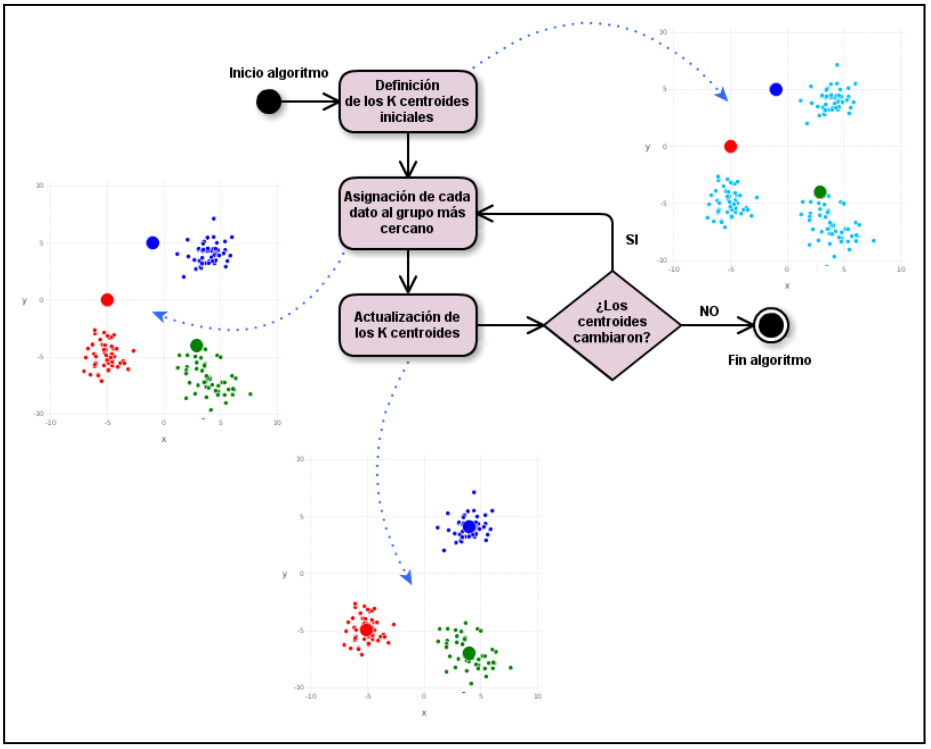
2.5.1.1. Algoritmo K-Means

Este algoritmo fue creado por MacQueen en el año 1967, es el algoritmo de agrupamiento más conocido y utilizado, esto debido a que su aplicación es muy simple y eficaz. Su procedimiento consiste en la clasificación de un conjunto en un determinado número K de *clústeres*, este algoritmo pide que K sea entregado antes (C. García Cambronero, I. Gómez Moreno, 2006).

Su nombre es dado porque representa cada uno de los grupos por la media o media ponderada de sus datos. Cada *clúster* es característico por su centro que se encuentra en el centro o medio de los elementos que componen el *clúster*.

El algoritmo k-means contiene 4 pasos principales que se consideran esenciales para definir su proceso (Pérez, Henriques, Pazos, Cruz, Reyes, Salinas y Mexicano, 2007). Estos pasos están representados en la Figura 1, y el diagrama de flujo muestra como es el funcionamiento del algoritmo k-means.

Figura 1: Diagrama de flujo algoritmo k-means



Fuente: Estrategia K-medias, Facultad de informática. UNLP (2016)

Cabe mencionar que cuando se trabaja con este tipo de algoritmo, también se suele trabajar con Análisis de Componentes Principales (PCA) (ver la sección 2.6), esto debido a que facilitan el análisis de los clústeres formados por el algoritmo, para saber cómo es que funciona y su definición.

Además del algoritmo *k-means*, dentro de los métodos basados en particiones más conocidos se encuentra el algoritmo *k-medoids*, definido a continuación.

2.5.1.2 Algoritmo K-Medoids

El algoritmo *K-Medoids* a diferencia de *K-means*, trabaja con la distancia mediana y no con la distancia media. *K-Medoids* tiene como propósito principal dividir un conjunto de datos en grupos, los representantes de estos grupos son llamados *medoids*. Cada dato observado es agrupado al *medoids* más cercano (Gonzáles , Heynz Roberth, Ticona Gonzáles, Ulises Amaru, 2019). Este algoritmo es más robusto al ruido, y también es necesario que se le asigne un K para trabajar.

2.5.2. Métodos Jerárquicos

En el método jerárquico se encuentran *clústeres* sucesivos utilizando los establecidos previamente, es decir, el algoritmo es divisivo, comienzan con todos los puntos en un solo *cluster* y luego los va dividiendo en grupos hasta que no tengan ninguna semejanza.

Es un proceso en el que la jerarquía de *clúster* se va creando en función de la distancia entre los puntos de los datos. Como resultado se obtiene una agrupación de manera jerárquica en un diagrama de árbol con diferentes *clústeres* (Jain, Brijnesh & Obermayer, Klaus, 2010).

Dentro de los algoritmos de Método Jerárquico, los más utilizados son enlace simple, enlace medio y enlace completo.

2.5.3. Métodos basados en modelos

Según Jiawei Hand y Micheline Kamber (2006) este método tiene como principal eje, crear un modelo nuevo pero basado en otro, es decir, es necesario suponer un modelo de *clúster* y ajustar los nuevos parámetros en base a este.

2.6. Análisis de componentes principales

Como se menciona en la sección 2.5.1.1. Algoritmo *K-Means*, el Análisis de Componentes Principales (PCA, *principal component analysis* del inglés) juega un rol importante a la hora de trabajar con dicho algoritmo.

Esta técnica es utilizada cuando se requiere disminuir la cantidad de variables a una cantidad de componentes de manera que facilite el trabajo, las variables representadas en estos nuevos componentes principales o factores son una combinación lineal de las variables originales, y además independientes entre sí (Terrádez Gurrea. M, 2002).

La interpretación de los nuevos factores es un aspecto clave en el PCA, ya que no viene dada a priori, sino que debe ser deducida tras observar la relación de los factores con las variables que se tienen al inicio (correlación) (Walpole et al, 2007).

Capítulo III
Estado del arte

En este capítulo se presentan trabajos correspondientes al estado del arte. Si bien la Estrategia Nacional de salud 2011-2020 traza un camino para mejorar los sistema de información y promover el uso de datos, para fortalecer la investigación en el sector de la salud, estos esfuerzos no se ven reflejados en trabajos de investigación. Por el contrario, a nivel internacional se pueden encontrar investigaciones que utilizan grandes volúmenes de datos, demostrando enormes avances en el manejo de contención de enfermedades respiratorias. A continuación se describen algunos trabajos:

1. En el trabajo realizado por Songjing Chen y Sizhu Wu (2020), denominado "Deep learning for identifying environmental risk factors of acute respiratory diseases in Beijing, China: implications for population with different age and gender", se plantea un análisis cuantitativo que identifica factores de riesgo relacionados a las enfermedades respiratorias, en la ciudad de Beijing, China. Utilizando variables médicas y medioambientales, combinado con redes neuronales profundas (Deep Learning), se pudo estratificar por edad y sexo a la población identificando, por ejemplo, que las mujeres menores de 50 años son más sensibles a contaminantes generales del aire como SO_2 y NO_2 que los hombres. Los resultados de este estudio pueden mejorar la calidad de vida de las personas en Beijing y prevenir estas enfermedades.

2. En la investigación realizada por Bellinger, Colin Mohomed Jabbar, Mohomed Shazan Zaïane, Osmar Osornio-Vargas, Alvaro (2017), llamada "A systematic review of data mining and machine learning for air pollution epidemiology", se realiza una revisión sistemática de la literatura sobre la aplicación de métodos de minería de datos y aprendizaje automático utilizados para la predicción e identificación de patrones en enfermedades respiratorias. Este trabajo es particularmente interesante ya que, por una parte, expone numerosos ejemplos que utilizan estos conceptos tecnológicos y a la vez describe cuáles son los algoritmos más utilizados para identificar y predecir este tipo de patologías.

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 (2017), si bien no estudia directamente enfermedades respiratorias, utiliza los mismos conceptos tecnológicos (minería de datos) para identificar factores de riesgos 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 (2017), 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 e identificar, por ejemplo, que en el año 2014 se observa un aumento en los diagnosticados con ASMA, siendo el género masculino entre 4 y 5 años el más afectado.

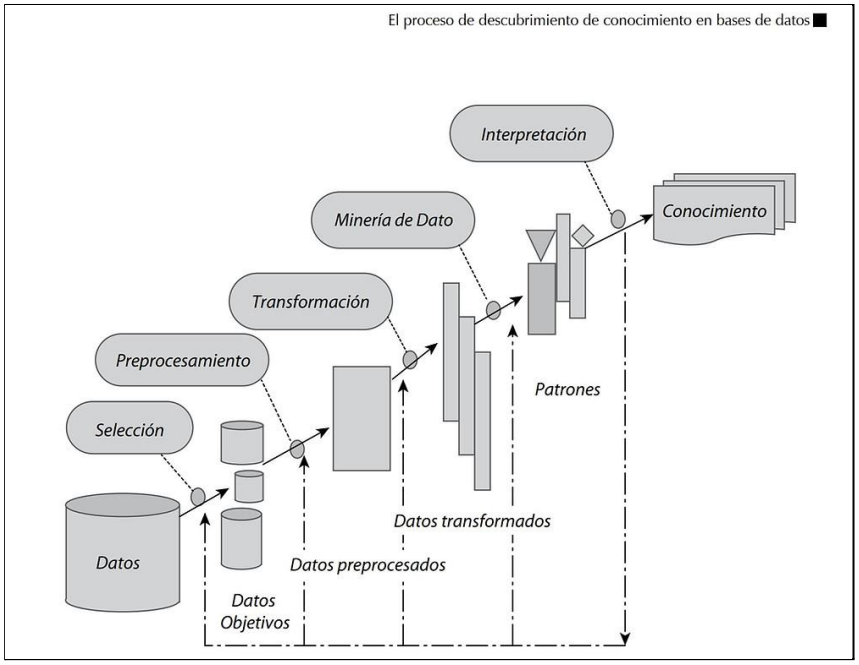
5. Y por último, a nivel nacional se encuentra el proyecto FONDEF de investigación y desarrollo denominado “Desarrollo y Evaluación de algoritmos de Data Mining para la predicción del Riesgo de Crisis en Pacientes Ambulatorios de un Hospital Pediátrico” (S. Alejandro, R. Perez, 2017). Esta investigación es una de las pocas iniciativas nacionales que utiliza estas tecnologías, tiene como objetivo desarrollar un paquete de medidas tecnologicas (algoritmos y sistemas de predicción de riesgos) que permiten detectar el peligro de crisis respiratorias y así monitorear a más de 33 cunas en la unidad de lactancia del Hospital.

Capítulo IV
Metodología

Las metodologías permiten trabajar un proceso de minería de datos en forma sistemática y ordenada, sirven para ayudar al entendimiento del proceso, y la planificación y ejecución del proyecto. Las más utilizadas en la minería de datos son Knowledge Discovery in Database (KDD) y *Cross Industry Standard Process for Data Mining* (CRISP-DM);

1. *Knowledge Discovery in Database* (KDD) consiste en un proceso automático en el cual se combinan análisis y descubrimiento, consiste en extraer patrones en forma de funciones a partir de los datos para que luego el usuario los analice, esta tarea implica preprocesar los datos, realizar minería de datos (*data mining*) y presentar resultados (Agrawal y Srikant, 1994).

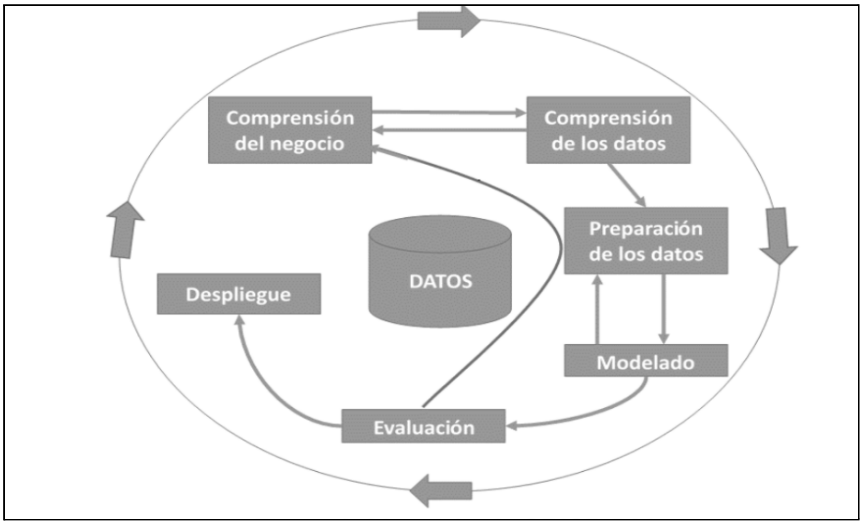
Figura 2: Etapas de KDD



Fuente: Timarán-pereira, S. et al. (2016).

2. *Cross Industry Standard Process for Data Mining* (CRISP-DM) consiste en una metodología que contempla el proceso de análisis de datos como un proyecto profesional. Este contexto considera la existencia de un cliente que no es parte del desarrollo, además considera que el proyecto no acaba cuando se encuentra el modelo ideal (ya que despues requiere despliegue y mantenimiento), sino que está relacionado con otros proyectos, y es necesario documentarlo de forma exhaustiva para que otros equipos de desarrollo trabajen a partir de él (Azevedo y Santos, 2008).

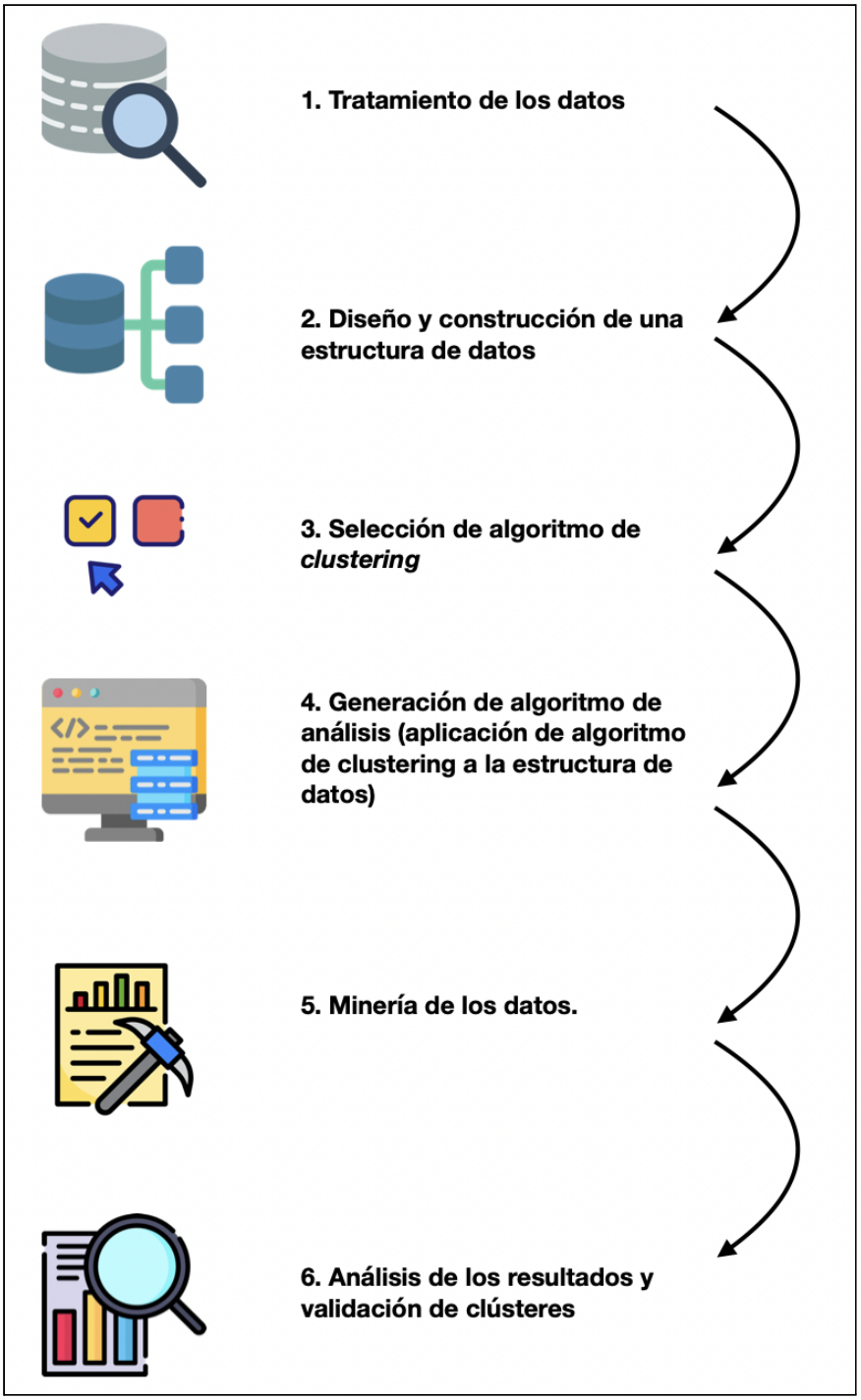
Figura 3: Etapas de CRISP-DM



Fuente: Hernández Cedon, J. A. (2015).

Con base en estos dos modelos se propone una metodología de 6 etapas: 1. Obtención de las bases de datos, 2. Diseño y Construcción de una Estructura de Datos, 3. Selección de Algoritmo de Clustering, 4. Programación de Algoritmo de Análisis (aplicación del algoritmo de clustering a la estructura de datos), 5. Minería de los datos y 6. Análisis de los resultados y validación de clústeres. En la Figura 4 se puede apreciar un diagrama que representa la metodología utilizada.

Figura 4: Diagrama de Metodología utilizada



Fuente: Elaboración propia.

En las siguientes secciones se describen detalladamente cada una de estas etapas, además de las acciones desarrolladas y los resultados obtenidos en cada una de ellas.

4.1. Tratamiento de los datos

Esta primera fase consiste en capturar u obtener los datos de las diferentes fuentes de datos. Estas variables extraídas son guardadas en su forma original para luego ser tratadas (Ralph Kimball, Joe Caserta, 2004).

Se cuentan con tres principales fuentes de datos para extraer la información necesaria para esta investigación, la tabla 2 expone las fuentes y qué variables se consiguen de dichas fuentes.

Tabla 2: Fuentes de Datos

Fuente	**Datos**	**Representación**
Hospital Regional de Copiapó	Enfermedades respiratorias	Adultos entre 15 - 64 a□ **Edad y Tipo de Atención** / **1** TOTAL DEMANDA 935 **SECCIÓN 1. TOTAL ATENCIONES DE URGENCIA 810** **TOTAL CAUSAS SISTEMA RESPIRATORIO 42** IRA Alta (J00-J06) 24 Influenza (J09-J11) 0 Neumonía (J12-J18) 5 Bronquitis/bronquiolitis aguda (J20-J21) 1 Crisis obstructiva bronquial (J40-J46) 7 Otra causa respiratoria (J22, J30-J39, J47, J60-J98) 5 COVID 19 Sospechoso (U07.2) 50 COVID 19 Confirmado (U07.1) 12
Dirección General de Aeronáutica Civil (DGAC)	Variables ambientales	**CodigoNacional** / **momento** / **Ts_Valor** **270009** 01-11-2020 00:00:00 14,4 **270009** 01-11-2020 01:00:00 13,2 **270009** 01-11-2020 02:00:00 12,7 **270009** 01-11-2020 03:00:00 12,2 **270009** 01-11-2020 04:00:00 12,0 **270009** 01-11-2020 05:00:00 11,5 **270009** 01-11-2020 06:00:00 11,3 **270009** 01-11-2020 07:00:00 11,5
Sistema de Información Nacional de Calidad del Aire (SINCA)	Variables contaminantes	**FECHA (YYMMDD)** / **HORA (HHMM)** / **Registros preliminares** **201101** 0 12 **201102** 0 10 **201103** 0 13 **201104** 0 9 **201105** 0 10 **201106** 0 8 **201107** 0 7 **201108** 0 6

Para los datos de las enfermedades respiratorias, se extraen el total de 8 variables disponibles, correspondientes a enfermedades respiratorias diagnosticadas en el Hospital Regional de Copiapó, incluidas la nueva afección respiratoria COVID-19, estas son: Infección respiratoria aguda (IRA), Influenza, Neumonía, Bronquitis, Crisis obstructiva bronquial (COB), Otra

causa respiratoria (OCR), Causas sistema respiratorio (CSR) y Covid19. Las variables respiratorias se encuentran agrupadas desde la base de datos por rango de edad y también de forma general, los grupos existentes son; general (todas las edades), menores de 1 año (llamados "Lactantes"), niños de 1 a 4 años (llamados "Primera infancia"), niños de 5 a 14 años (llamados "Infantes"), adultos de 15 a 64 años (llamados "Jóvenes y adultos") y adultos de 65 o más años (llamados "Adultos mayores"). Se trabaja con estos mismos grupos de edad.

Para el caso de los datos de variables ambientales, se captaron el total de variables disponibles, 5 variables, las cuales son: temperatura, rocío, humedad, temperatura mínima y temperatura máxima.

Y por último, para las variables contaminantes, se extraen 3 de este tipo: material particulado de 2.5 micras (MP2.5), material particulado de 10 micras (MP10) y dióxido de azufre (SO_2).

El detalle y definición de cada una de estas variables se puede ver en el Anexo B correspondiente al Diccionario de Datos.

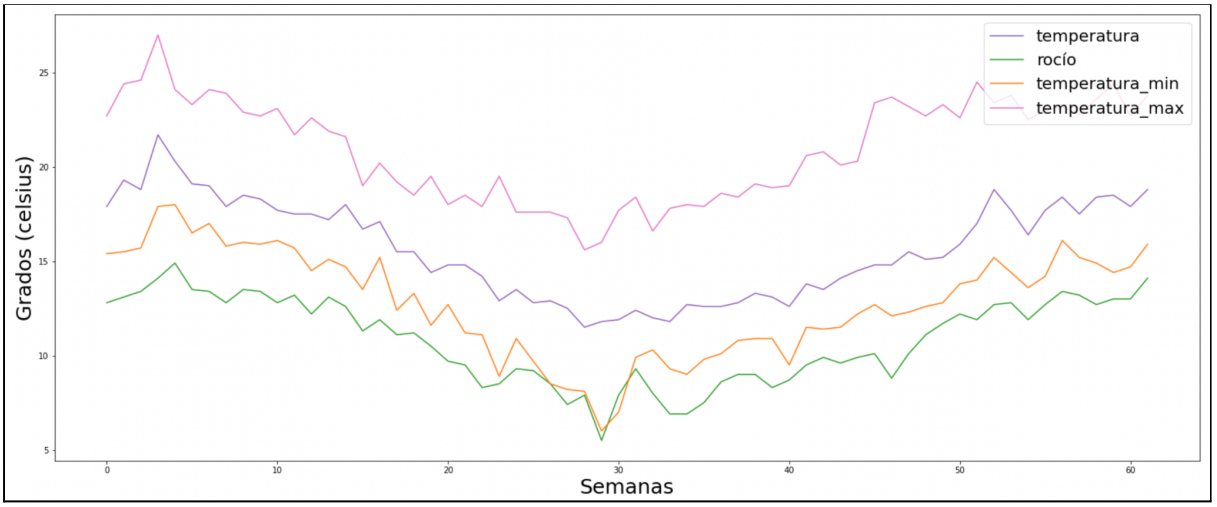
Una vez obtenidos los datos, es necesario explorar de forma individual cada una de estas variables, permitiendo tener una visión general. A continuación se expone una breve descripción de cada uno de estos grupos.

I. Variables Ambientales

En primer lugar el grupo de las variables ambientales se comporta de forma esperada, ya que según lo expuesto por Squeo, F. *et al* (2008), la región de Atacama y en especial la comuna de Copiapó cuenta con una temperatura media anual de 15,7 °C, variando en verano entre los 18°C a 20°C y el resto del año entre los 18°C a 20,5 °C. A su vez, Copiapó tiene un comportamiento típico de ciudades-oasis con existencia de islas de calor y de fresco urbano, lo que causaría aumento o disminución de las temperaturas (Gómez Sarria, N., 2014). Se obtiene además, que la media de la temperatura es de 15.67 °C, con una mínima de 11.5 °C y una máxima de 21.7 °C. Destacando la temperatura máxima de 27 °C y la mínima de 6°C, lo que a simple vista puede ser consecuencia de las islas de calor y fresco urbano que, como se menciona anteriormente, provocan aumento o disminución de las temperaturas.

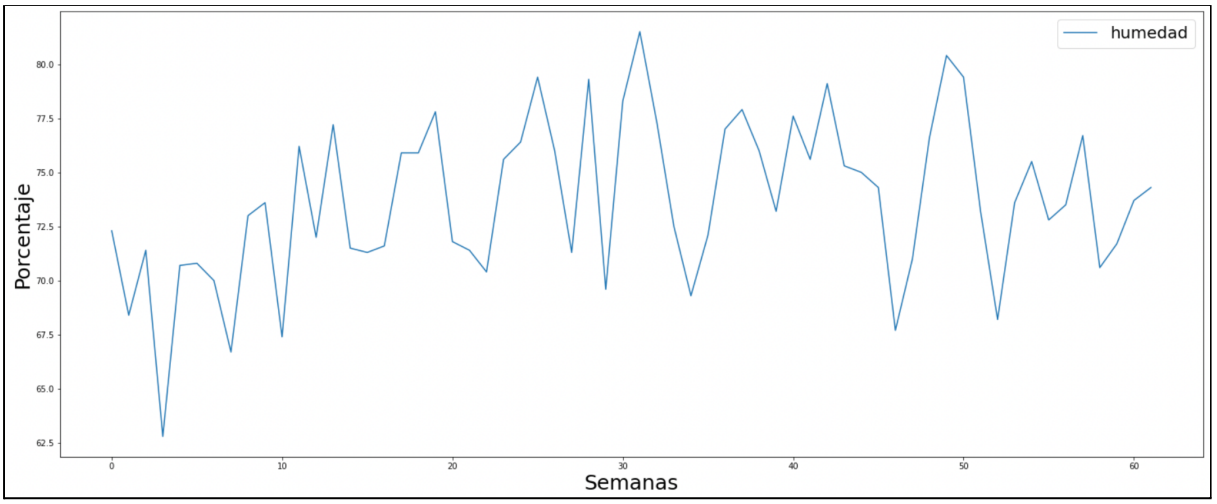
De forma un poco más general, se puede visualizar en la Figura 5, que tanto la *Temperatura, Rocío, Temperatura_max* y *Temperatura_min*, se comportan muy parecido en especial las variables de *Rocío* y *Temperatura_min*. Todas ellas aumentan en los meses de enero, febrero y declinan en los meses de junio y julio. Y por último la variable humedad (Figura 6), medida en porcentajes, no presenta grandes variaciones a lo largo del tiempo medido, evidenciando esto en su desviación estándar.

Figura 5: Análisis exploratorio de temperaturas y punto de rocío



Fuente: Elaboración propia.

Figura 6: Análisis exploratorio de humedad



Fuente: Elaboración propia.

II. Variables Contaminantes

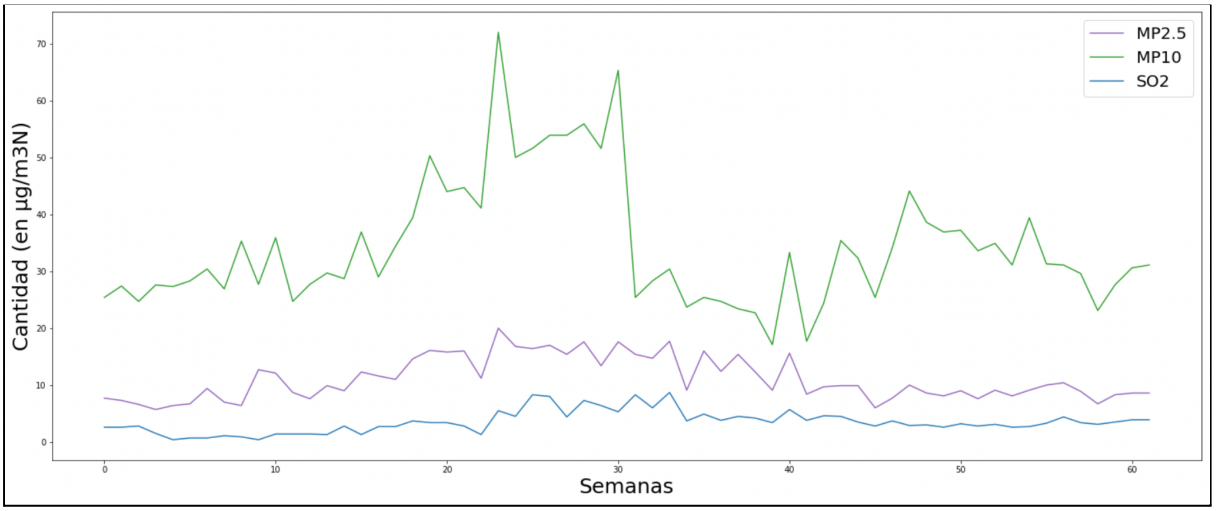
Para el caso de las variables contaminantes, se puede apreciar un comportamiento dentro de las normas establecidas (Ver Tabla 3). Tal es el caso del MP2.5 y MP10 que registran máximas de 20 µg/m3N y 70 µg/m3N respectivamente, mediciones dentro del rango esperado (Ver tabla 3). Del mismo modo la variable SO_2 se encuentra en los rangos o parámetros deseados con 8.7 ppbv, destacando su bajo nivel de emisión, muy por debajo de la norma (Ver Tabla 3).

Tabla 3: Norma primaria de calidad de aire para MP2.5, MP10 y SO_2 como concentración de 24 horas

Variables de calidad de aire	Norma	Regulación
MP2.5	50 µg/m3N	Decreto 12/2011, ESTABLECE NORMA PRIMARIA DE CALIDAD AMBIENTAL PARA MATERIAL PARTICULADO FINO RESPIRABLE MP 2,5
MP10	150 µg/m3N	(DS Nº59/1998 MINSEGPRES).
SO_2	96 ppbv (250 µg/m3N)	(DS Nº104/2018 MINSEGPRES).

Si bien estas variables se encuentran dentro de los parámetros esperados, se puede evidenciar que en los meses de mayo, julio y julio aumentan las emisiones de MP10 y disminuye a finales de septiembre como se muestra en la Figura 7. Del mismo modo, se puede apreciar que las variables MP2.5 y SO_2 se mantienen estables en el tiempo, con un leve aumento en los meses de mayo, junio y julio.

Figura 7: Análisis exploratorio de variables contaminantes



Fuente: Elaboración propia.

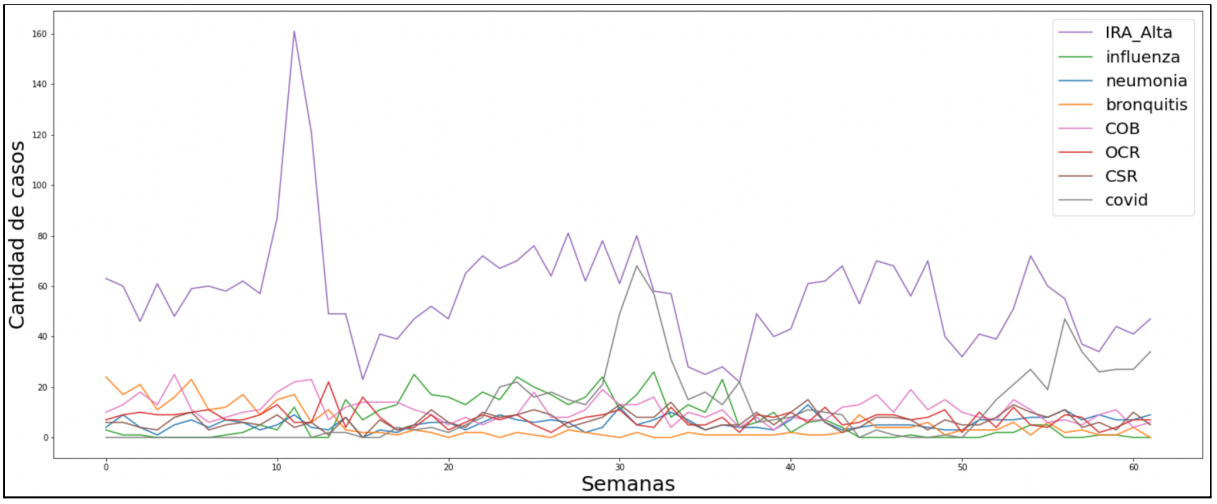
III. Enfermedades Respiratorias

Finalmente en lo que respecta a las enfermedades respiratorias, se puede observar que las variables con mayor cantidad de diagnósticos son *IRA_alta* y *Covid* con 161 y 68 registros respectivamente. Y por el contrario, las enfermedades menos diagnosticadas son *Neumonía* y *CSR*. Esto se puede ver reflejado en la Figura 8, donde se aprecia claramente que la variable

IRA_alta, que representa a los diagnósticos de infecciones agudas de las vías respiratorias superiores, predominan durante todo el periodo de análisis, incrementando el número de casos en los meses de febrero y marzo, decayendo en los meses de abril y septiembre. Si bien con algunas diferencias de números de casos, la variable Covid también se puede destacar, en primera instancia, porque aparece en marzo aproximadamente, no existen registros anteriores de esta enfermedad y en segunda instancia relacionada al número de casos, llegando a su número máximo de diagnósticos en los meses de julio, agosto y en menor cantidad en enero. Y por el contrario disminuye en el mes de noviembre.

Del mismo modo se puede apreciar que las otras variables de *Influenza, Neumonía, Bronquitis, CSR, OCR y CSR* no son muy predominantes, no superando como máximo los 26 casos en la semana.

Figura 8: Análisis exploratorio de enfermedades respiratorias



Fuente: Elaboración propia.

4.2. Diseño y construcción de una estructura de datos

Luego de obtener y analizar los datos, se comienza un proceso de transformación que según lo expuesto por Ralph Kimball, Joe Caserta (2004), permite procesar las variables de tal forma que sean coherentes y útiles para la necesidad o problemas que se tengan. Luego de la transformación, se comienza un proceso de carga de datos para crear una estructura con la cual trabajar. La fase de carga de datos consiste en almacenar los datos que ya fueron transformados en un sistema de destino (Ralph Kimball, Joe Caserta, 2004).

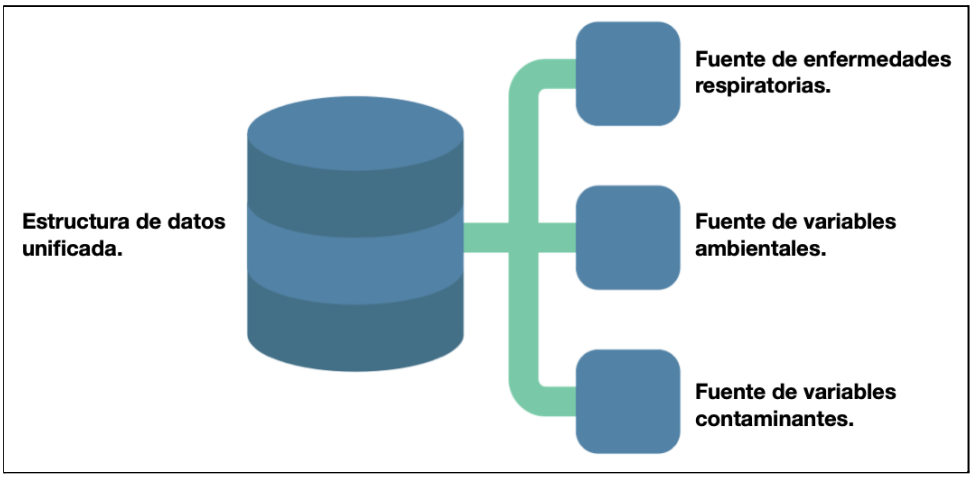
Para esto, en una primera instancia se comienza un proceso de limpieza de datos. Se conoce como limpieza de datos al proceso encargado de corregir los errores en los datos, convirtiéndose en un mecanismo necesario para que las estadísticas, los informes y las decisiones que se tomen sean confiables, ya que, garantizando la calidad de los datos puede existir una seguridad y fiabilidad en las acciones que se produzcan a partir del análisis de esos datos (López, B., 2011).

En el proceso de transformación de datos se analizaron el total de variables disponibles de todas las fuentes, con la idea de llegar a construir una estructura de datos unificando todas las fuentes de datos. Se ajustan los formatos de los archivos de descarga para poder trabajar más fácilmente en su posterior programación. Debido a la compatibilidad de la mayoría de los archivos, estos son transformados al tipo *Comma Separated Values* (CSV).

Como resultado se obtiene una estructura que reúne a todas las variables, es decir, 16 columnas (8 enfermedades respiratorias, 5 variables ambientales y 3 variables contaminantes) y donde las filas representan los datos recolectados separados por semanas, con un total de 62 filas (62 semanas). Se puede revisar una hoja (todas las edades) de la estructura de datos en el Anexo D de este trabajo.

En la Figura 9 se puede observar la forma en que la estructura de datos unifica las 3 fuentes de datos para esta investigación.

Figura 9: Dinámica para la estructura de datos



Fuente: Elaboración propia.

4.3. Selección de algoritmo de *clustering*

Esta fase tiene como objetivo determinar la técnica de agrupamiento (*clustering*) a utilizar en esta investigación. *Clustering* hace referencia a un gran abanico de técnicas no supervisadas que tienen como finalidad encontrar patrones o grupos dentro de un conjunto de observaciones (Amat Rodrigo, J., 2017).

A la hora de aplicar una técnica de agrupación es necesario seleccionar cual de todos los existentes es adecuado para trabajar dado los parámetros para este caso. Existen una gran cantidad de algoritmos, indicados en en la sección 2.5, pero de manera simplificada estas son las 3 principales ramas de tipos de algoritmos de agrupación:

- Métodos basados en particiones
- Métodos Jerárquicos
- Métodos basados en modelos

Como base para la selección del tipo de técnica y considerando las 3 principales ramas de los tipos de técnicas de agrupación (métodos basados en particiones, métodos jerárquicos, métodos basados en modelos) se puede determinar lo siguiente:

1) No se tiene ni se busca trabajar con variables jerárquicas, es decir, las variables son uniformes en jerarquía, ni tampoco se busca como resultado una forma árbol jerárquico para esta investigación.

2) No se tiene como base un modelo anterior, es decir, esta investigación no tiene ni busca trabajar sobre un modelo de método de agrupamiento anterior.

De acuerdo con estas restricciones se determina trabajar con basados en particiones, además en la sección 2.5.1 se menciona que los algoritmos más utilizados basados en particiones son *K-means* y *K-medoids,* para lograr definir cual de estos algoritmos se utiliza se aplican criterios de selección:

Como criterios de selección se tiene lo siguiente:

A) Permite analizar pequeños y grandes volúmenes de datos.

   Dentro del trabajo realizado por Aguilar-Aldana, J. S. (2017) concluyen como criterio de selección la capacidad de analizar tanto grandes como pequeñas cantidades de datos. Además para este trabajo se considera el realizar pruebas preliminares, por lo que se necesita que el algoritmo seleccionado sea capaz de analizar pequeños volúmenes de datos, y que también tenga la capacidad de analizar grandes volúmenes de datos, para realizar pruebas superiores a 1 año de *data.*

B) Acepta la asignación de la cantidad de clústeres.

   Según las conclusiones de Viera, Angel Freddy Godoy. (2017) se consideran importantes en este tipo de investigaciones al trabajar con minería de datos, algoritmos que tengan la característica de poder asignar previamente un número determinado de clústeres. Para este trabajo, en la asignación óptima de clústeres se utilizan métodos para esto, para lo cual el algoritmo seleccionado debe permitir que se le proporcione el número de clústeres a trabajar.

C) Alto rendimiento con Bases de datos reales.

   En el trabajo de Pascual, D., Pla, F y Sánchez, S. (2007) se describe que como característica relevante de un algoritmo, es su capacidad de trabajar con bases de

datos reales. En su trabajo se implementan y comparan diferentes algoritmos de clasificación, utilizando bases de datos reales y artificiales obteniendo los mejores de clasificación cuando se utilizan bases con datos reales.

D) Permite el descubrimiento de datos anómalos.

Tanto en el trabajo realizado por Aguilar-Aldana, J. S. (2017) y en el de Pascual, D., Pla, F y Sánchez, S. (2007) se describe la importancia a la hora de trabajar con algoritmo de agrupamiento la capacidad de detectar ruido, es decir datos anómalos, para así evitar una influencia negativa dentro de los grupos.

E) Algoritmo simple y eficiente.

Al momento de buscar patrones sobre los datos es importante que el algoritmo sea simple de aplicar y eficiente, esto con el fin de tener la ventaja de realizar de manera rápida cálculos simples y el procesamiento secuencial de los datos (Prado, Pedro & Monteiro, António, 2008)

Basados en estos criterios de selección se crea una tabla comparativa entre los algoritmos K-Means y K-Medoids con el fin de lograr determinar cuál de estos Métodos basados en particiones es el adecuado para este trabajo. La tabla 4 muestra esta comparación de ambos algoritmos.

Tabla 4: Cuadro comparativo de algoritmos *k-means* y *k-medoids*

Criterios	Algoritmos		Observaciones
	K-means	K-medoids	
A) Permite analizar pequeños y grandes volúmenes de datos	✔	✘	Algoritmo K-medoids no trabaja bien con volúmenes grandes. (Balabantaray, R. C., Sarma, C., & Jha, M., 2015).
B) Acepta la asignación de la cantidad de clusters.	✔	✔	(Preeti Arora, Deepali, Shipra Varshney, 2016).
C) Alto rendimiento con Bases de datos reales.	✔	✔	(Amat Rodrigo, J., 2017)
D) Permite el descubrimiento de datos anómalos.	✔	✘	K-medoids es más robusto al ruido, por lo que no se verán reflejados datos *outliers* en los *clusters.* (Preeti Arora, Deepali, Shipra Varshney, 2016).
E) Algoritmo simple y eficiente.	✔	✘	K-medoids necesita mucho más tiempo de ejecución con grandes volúmenes de datos. (Cristina García Cambronero, Irene Gómez Moreno 2006).

✔ Si cumple con el criterio.

✘ No cumple con el criterio.

Considerando la tabla comparativa, además de los criterios de selección y considerando también el análisis previo, el algoritmo a utilizar para este trabajo es el algoritmo de agrupamiento ***K-means.***

4.4. Generación de algoritmo de análisis

Una vez seleccionado el algoritmo, se comienza la etapa de implementación. En esta etapa se utilizaron diferentes herramientas, librerías y acciones, que permitieron implementar el algoritmo de clustering *k-means*. A continuación se describen las más relevantes.

4.4.1. Herramientas

Como herramienta principal de para la programación se utiliza la plataforma *Anaconda*, que permite trabajar con el lenguaje de programación *Python*, utilizado comúnmente en las áreas de ciencias de datos y aprendizaje automático. Además de esta plataforma se utiliza *Jupyter*

Notebook, para el desarrollo del algoritmo principal y *Spyder*, como entorno de desarrollo para *Python* utilizado para codificar de manera preliminar el algoritmo.

Además de estas herramientas, se utilizan librerías específicas de Python que contienen algoritmos y elementos de programación esenciales para esta tarea.

4.4.2. Librerías

Las librerías corresponden a un conjunto de funciones específicas que ayudan al momento de programar, actualmente se puede observar una gran tendencia sobre el uso de Python, esto es debido a la disponibilidad de librerías de visualización, procesamiento de señales, estadísticas y álgebra entre otras; de fácil utilización y que cuentan con buena documentación (Challenger-Pérez et al., 2014).

Las librerías utilizadas para este proyecto son las siguientes:

A) Numpy: Empleada para trabajar con cálculos numéricos.
B) Pandas: Contiene un conjunto de funciones para trabajar tablas de forma más rápida y sencilla.
C) Matplotlib.pyplot: En ella se encuentran funciones que permiten graficar con lenguaje de programación Python.
D) Sklearn.cluster: Contiene el algoritmo de agrupación *k-means*, seleccionado para esta investigación.
E) Sklearn.decomposition: Se encuentran funciones para el Análisis de componentes principales (PCA).
F) Funpymodeling.model_validation: Se encuentra el modelo de validación de clustering.

4.4.3. Lectura de los datos

Esta acción permite cargar los datos, utilizando códigos en programación para un análisis exploratorio y así tener un contexto general del comportamiento de las variables utilizadas en esta investigación. La Figura 10 muestra la estructura de datos al momento de ser leída.

Figura 10: Lectura de la estructura de datos

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Temperatura      62 non-null     float64
 1   Rocio            62 non-null     float64
 2   Humedad          62 non-null     float64
 3   Temperatura_min  62 non-null     float64
 4   Temperatura_max  62 non-null     float64
 5   MP2.5            62 non-null     float64
 6   MP10             62 non-null     float64
 7   SO2              62 non-null     float64
 8   IRA_Alta         62 non-null     int64
 9   Influenza        62 non-null     int64
 10  Neumonia         62 non-null     int64
 11  Bronquitis       62 non-null     int64
 12  COB              62 non-null     int64
 13  OCR              62 non-null     int64
 14  CSR              62 non-null     int64
 15  Covid            62 non-null     int64
dtypes: float64(8), int64(8)
memory usage: 7.9 KB
```

Fuente: Elaboración propia.

En esta Figura se puede observar una descripción de los datos, en la parte izquierda el número que utiliza Python para identificar la cantidad de variables asignando desde el 0 hasta el 15, a su derecha se encuentra el nombre de las columnas (*Column*) que corresponden al nombre de las variables, continuando a la derecha de éstas, se encuentran la cantidad de datos *Non-null* indicando que no se cuenta con datos nulos en el conjunto de datos, y finalmente *Dtype* indica a qué tipo de dato corresponde a cada variable.

Posterior a esta lectura, comienza un proceso de normalización de los datos, descrito en la siguiente sección.

4.4.4. Normalización de los datos

La normalización tiene como propósito producir un conjunto de datos estable (Ricardo, Catherine. M., 2009). Los principios de la normalización permiten que se logre un diseño más flexible, así el modelo puede extenderse cuando necesite representar nuevas variables, conjunto de entidades y relaciones. También sirve para reducir la redundancia en la base de datos, tanto para ahorrar espacio como para evitar inconsistencias en los datos (Ricardo, Catherine. M., 2009)

Además la normalización asegura que el diseño esté libre de anomalías. Una anomalía es un estado inconsistente, incompleto o contradictorio de la base de datos.

Los valores contenidos en las variables de este trabajo son muy distintos en lo que respecta a la distancia numérica que existe entre ellos, es decir, existen valores iguales o cercanos a 0 mientras que al mismo tiempo otras variables presentan valores cercanos a los 200. Es por esta razón que se lleva a cabo una normalización para que así todos los valores estén entre 0 y 1.

La Figura 11 muestra la descripción de todas las variables, antes de realizada la normalización, considerar los siguientes significados:

- count: Cantidad de datos dentro de la variable correspondiente.
- mean: Valor medio dentro de cada variable.
- std: Desviación estándar dentro de cada variable.
- min: Valor mínimo dentro de cada variable.
- max: Valor máximo dentro de cada variable.

Figura 11: Datos antes de la normalización

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid
count	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000
mean	15.672581	10.854839	73.687097	12.808065	20.990323	11.070968	34.283871	3.508065	56.725806	7.354839	5.741935	5.032258	10.903226	7.564516	6.774194	12.387097
std	2.601082	2.272550	3.735847	2.825023	2.707713	3.730485	11.093170	1.947365	21.839331	7.917809	2.592141	6.174957	4.874523	3.509336	3.164451	15.303867
min	11.500000	5.500000	62.800000	6.000000	15.600000	5.700000	17.100000	0.400000	22.000000	0.000000	0.000000	0.000000	3.000000	2.000000	0.000000	0.000000
25%	13.150000	9.000000	71.325000	10.900000	18.500000	8.325000	27.325000	2.600000	43.250000	0.000000	4.000000	1.000000	8.000000	5.000000	5.000000	0.000000
50%	15.500000	11.150000	73.550000	12.750000	21.650000	9.900000	31.100000	3.350000	57.000000	4.500000	6.000000	2.000000	10.000000	7.500000	6.000000	7.000000
75%	17.900000	12.800000	76.350000	15.200000	23.275000	14.675000	38.250000	4.400000	64.750000	13.000000	7.000000	6.000000	13.000000	9.000000	8.750000	19.750000
max	21.700000	14.900000	81.500000	18.000000	27.000000	20.000000	72.000000	8.700000	161.000000	26.000000	13.000000	24.000000	25.000000	22.000000	15.000000	68.000000

Fuente: Elaboración propia.

Cómo se logra observar en la Figura los valores más bajos (min) y los valores más altos (max) de cada variable son muy distantes entre sí, por ejemplo: mientras se tiene valores mínimos de 0.4 (min SO_2) también existen valores máximos de 161 (max IRA_Alta).

Por otro lado la Figura 12, muestra un extracto de las mismas variables después de aplicada la normalización.

Figura 12: Datos después de la normalización

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid
0	0.627451	0.776596	0.508021	0.783333	0.622807	0.139860	0.151184	0.265060	0.294964	0.115385	0.307692	1.000000	0.318182	0.25	0.400000	0.000000
1	0.764706	0.808511	0.299465	0.791667	0.771930	0.111888	0.187614	0.265060	0.273381	0.038462	0.692308	0.708333	0.454545	0.35	0.400000	0.000000
2	0.715686	0.840426	0.459893	0.808333	0.789474	0.062937	0.138434	0.289157	0.172662	0.038462	0.307692	0.875000	0.681818	0.40	0.266667	0.000000
3	1.000000	0.914894	0.000000	0.991667	1.000000	0.000000	0.191257	0.132530	0.280576	0.000000	0.076923	0.458333	0.454545	0.35	0.200000	0.000000
4	0.862745	1.000000	0.422460	1.000000	0.745614	0.048951	0.185792	0.000000	0.187050	0.000000	0.384615	0.666667	1.000000	0.35	0.533333	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
57	0.588235	0.819149	0.743316	0.766667	0.631579	0.223776	0.227687	0.361446	0.107914	0.000000	0.538462	0.125000	0.090909	0.30	0.266667	0.500000
58	0.676471	0.765957	0.417112	0.741667	0.701754	0.069930	0.109290	0.325301	0.086331	0.038462	0.692308	0.041667	0.272727	0.00	0.400000	0.382353
59	0.686275	0.797872	0.475936	0.700000	0.771930	0.181818	0.191257	0.373494	0.158273	0.038462	0.538462	0.041667	0.363636	0.10	0.200000	0.397059
60	0.627451	0.797872	0.582888	0.725000	0.622807	0.202797	0.245902	0.421687	0.136691	0.000000	0.538462	0.166667	0.045455	0.25	0.666667	0.397059
61	0.715686	0.914894	0.614973	0.825000	0.719298	0.202797	0.255009	0.421687	0.179856	0.000000	0.692308	0.000000	0.136364	0.25	0.333333	0.500000

Fuente: Elaboración propia.

Al aplicar la normalización de los datos se puede observar en la Figura que los valores ahora se encuentran en el rango entre 0 y 1.

4.4.5. Implementación de algoritmo k-means

La fase de programación conlleva una secuencia de acciones para desarrollar e implementar el algoritmo *K-Means*, esta tarea se facilita con la librería *sklearn,* ya que*,* que contiene dicho algoritmo. Sin embargo, como se menciona en la sección 2.5.1.1. Algoritmo K-Means, es necesario proporcionar al algoritmo la cantidad de *clústeres* (K) a utilizar, es por esto que a continuación se describen los pasos utilizados para determinar K mediante programación y cuáles son los métodos que se utilizan.

4.4.5.1. Definición de los centroides

Como se menciona anteriormente, el algoritmo *K-Means* solicita como inicio una cantidad determinada de *clústeres* o K. A continuación se exponen los diferentes métodos que se utilizan en este trabajo para determinar K.

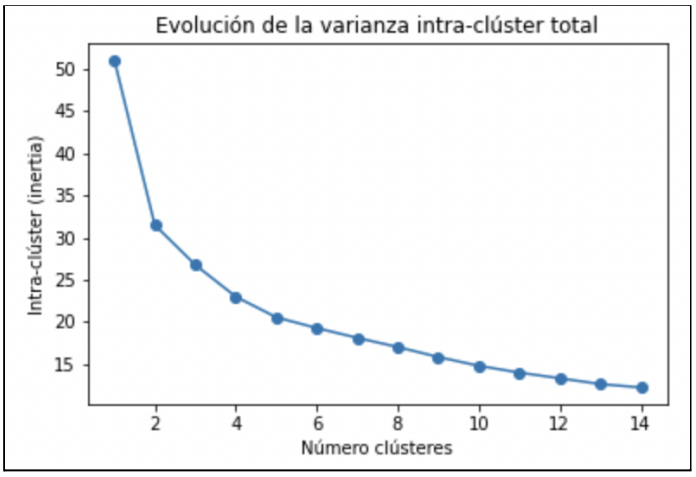
En general no existe un modo exacto para determinar K, pero se puede estimar con aceptable precisión a través de algunos métodos.

4.4.5.1.1 Método del Codo

Uno de los Métodos más usados para esto, es la distancia media entre los puntos de los datos y su centroide o más conocido como el método del codo. Así el valor de la media disminuye a medida que se aumenta el valor de K. Se debe utilizar la distancia media al centroide en función de K y finalmente encontrar el “punto de codo”, donde la tasa de descenso disminuye considerablemente (Langfelder, P. Horvath, S., 2008).

Para este trabajo se aplica el método del codo, indicando un número máximo de clústeres a iterar de 15 y el conjunto de datos normalizados. La Figura 13 muestra gráficamente el resultado de la aplicación de este método aplicado a la totalidad de los 9.528 casos médicos, sin distinción de edad.

Figura 13: Método del codo aplicado a la totalidad de los casos



Fuente: Elaboración propia.

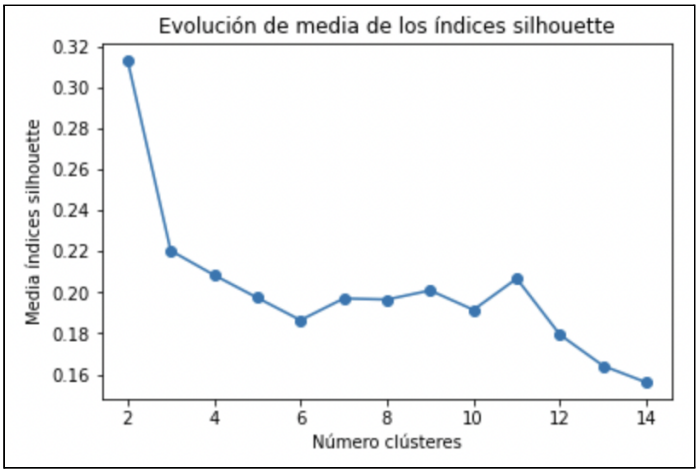
Se observa que la tasa de descenso baja significativamente en el número de clúster = 2, indicando el número de clústeres óptimos para trabajar.

4.4.5.1.2 Método de Silhouette

Otro método con el que se puede determinar el número de K es el método de Silhouette, este es muy similar al del codo, pero con la diferencia de que en lugar de minimizar la distancia entre los puntos y su centroides, este maximiza la media de los coeficientes de silhouette. Este coeficiente cuantifica la calidad de la asignación que se ha hecho comparando su similitud con el resto de las observaciones de su cluster frente a la de los otros clústeres. Su valor puede estar entre -1 y 1, siendo los valores altos un indicativo de que la asignación al cluster es correcta (Amat Rodrigo, J., 2017).

La Figura 14 muestra de manera gráfica este método aplicado a la totalidad de los 9.528 casos médicos sin distinción de edad.

Figura 14: Método de Silhouette aplicado a la totalidad de los casos



Fuente: Elaboración propia.

Se observa que el valor máximo en la media de los índices de Silhouette ocurre en el Número de clústeres = 2, indicando el número óptimo de clústeres a trabajar, además de coincidir con el método del codo.

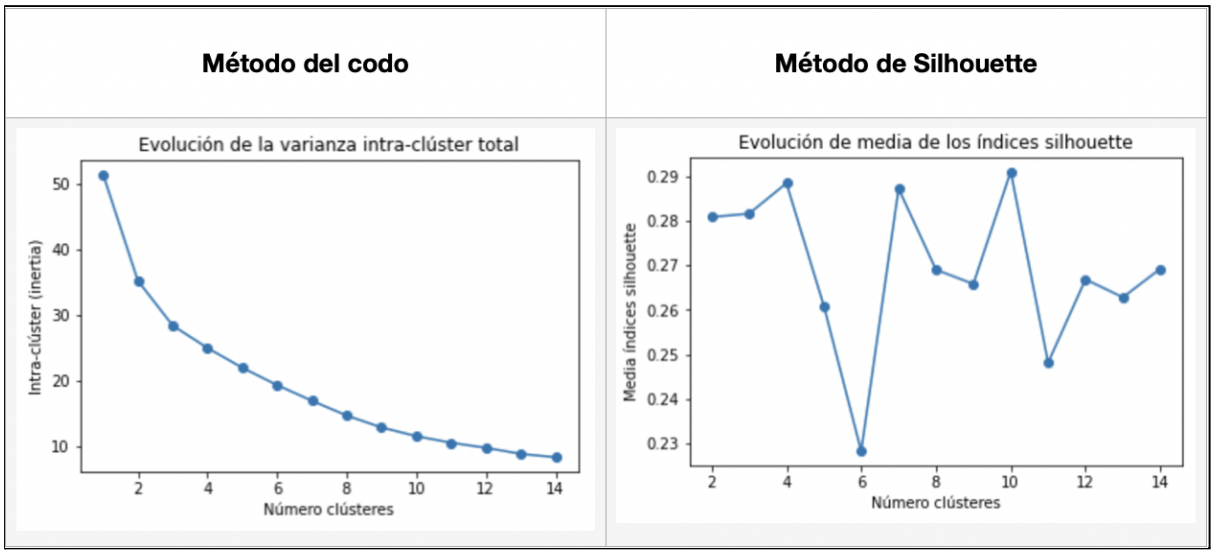
Finalmente el algoritmo de clustering k-means tiene como variables de entrada, la cantidad de *clústeres* óptimos, 300 iteraciones y el conjunto de datos normalizados.

La estructura de datos, además de contener tanto las variables enfermedades respiratorias, variables ambientales y variables contaminantes, se encuentra segmentada en rangos de edades, como se menciona en la sección 4.1, estos rangos fueron mantenidos desde las bases de datos, lo que permite tener una descripción más detallada del comportamiento de las enfermedades respiratorias. Es por esto, que se hace necesario determinar a la vez el K o número de clústeres por rango de edad.

A continuación se expone la aplicación de los métodos del codo y silhouette aplicados por rangos de edad.

I. Para el caso de los Lactantes y para un total de 391 registros médicos, se obtienen los siguientes *clústeres*. La Figura 15 muestra el número óptimo de *clústeres* para este rango de edad.

Figura 15: Métodos aplicados a Lactantes

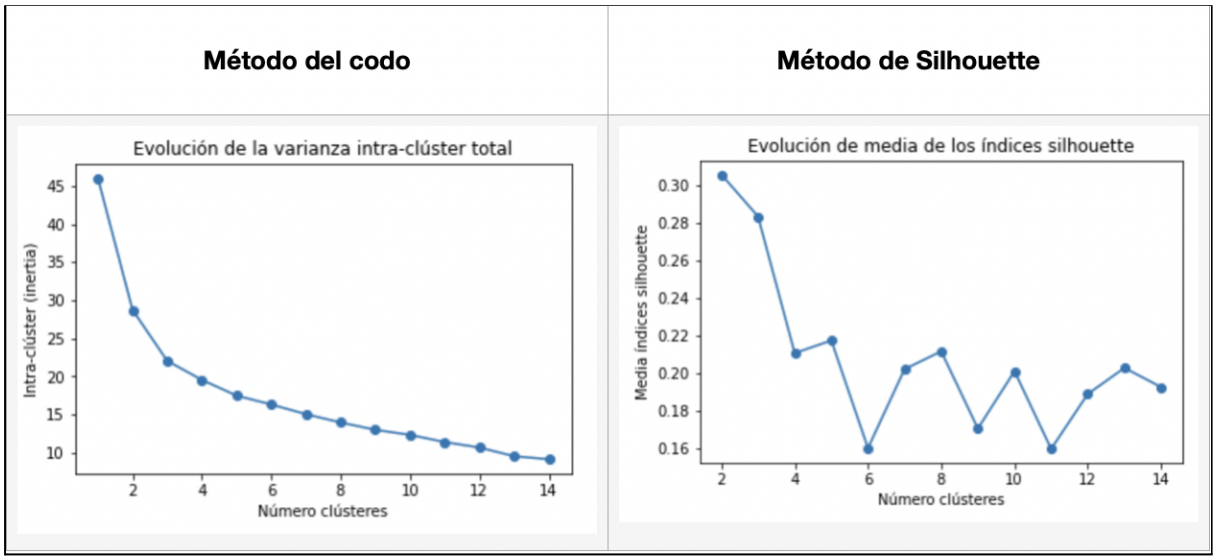


Fuente: Elaboración propia.

En este grupo de edad se observa mediante el método del codo que el número de clústeres óptimos para trabajar podría ser de 3 o 4, mientras tanto el Método de Silhouette muestra que el número óptimo de *clúster* a trabajar se encuentra en 4, 7 o 10, se ratifica entonces el número 4 para la cantidad de *clústeres* a trabajar para el grupo de Lactantes.

II. Para el caso de la Primera infancia, para un total de 693 registros médicos, se obtienen los siguientes *clústeres*. La Figura 16 muestra el número óptimo de *clústeres* para este rango de edad.

Figura 16: Métodos aplicados a Primera infancia

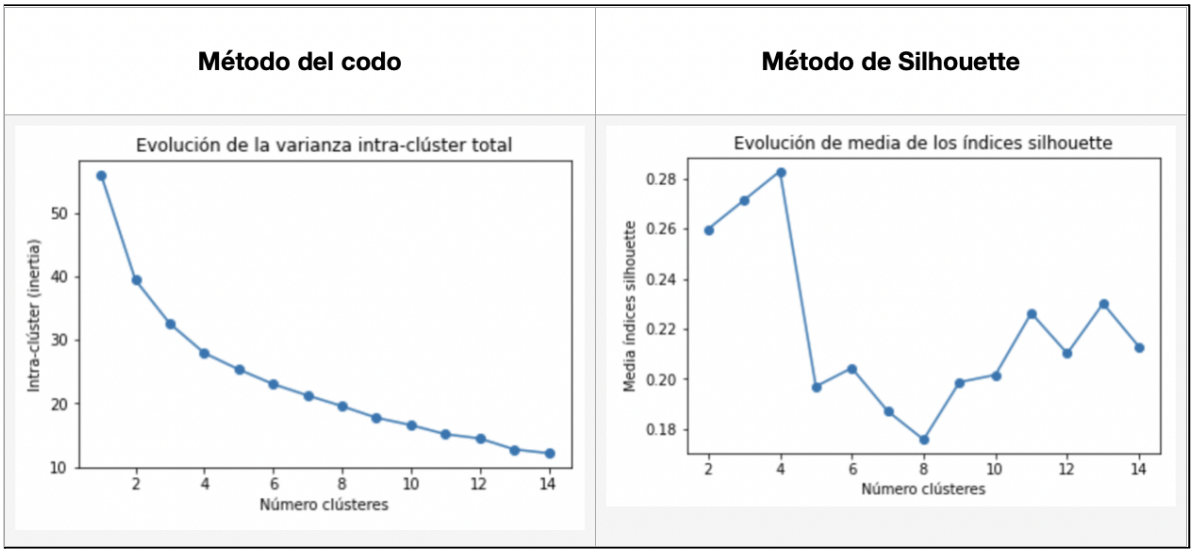


Fuente: Elaboración propia.

En este grupo de edad se observa que el número óptimo de *clústeres* para trabajar es de 2. Además se logra observar que con 3 *clústeres* se acerca a la cantidad óptima para trabajar, pero 2 *clústeres* continúa siendo la cantidad óptima.

III. En el caso de los Infantes, para un total de 661 registros médicos, se obtienen los siguientes *clústeres*. La Figura 17 muestra el número óptimo de *clústeres* para este rango de edad.

Figura 17: Métodos aplicados a Infantes

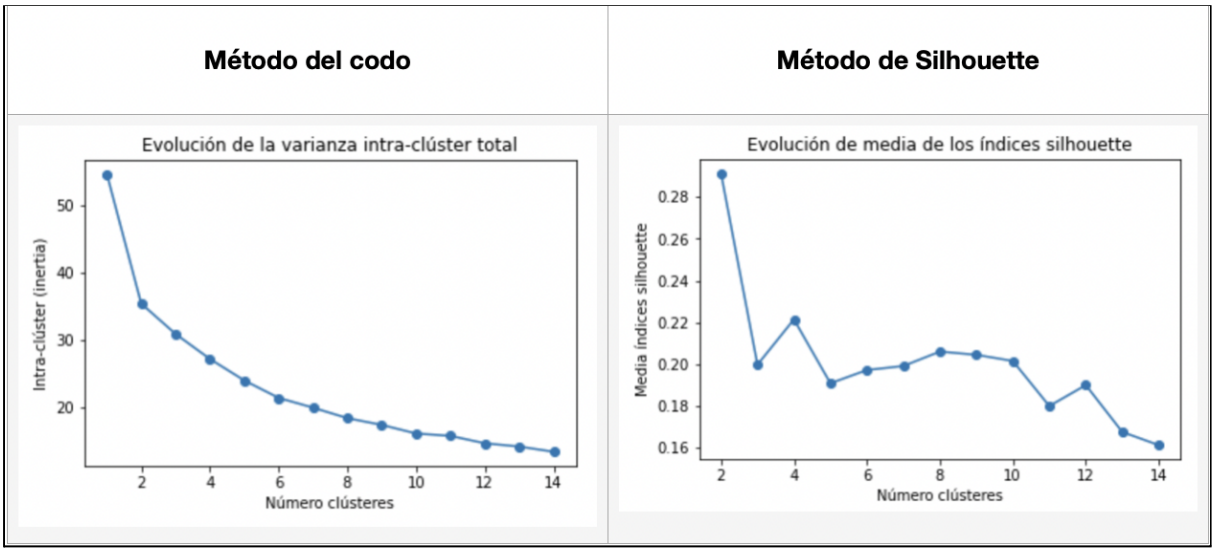


Fuente: Elaboración propia.

En este grupo de edad como muestra la Figura 16 se observa mediante el Método del codo que 2, 3 o 4 puede ser una cantidad óptima de clústeres, pero el Método de Silhouette ratifica que con 4 clústeres es la mejor forma de trabajar.

IV. Para el caso de los Jóvenes y adultos, para un total de 6.530 registros médicos, se obtienen los siguientes *clústeres*. La Figura 18 muestra el número óptimo de *clústeres* para este rango de edad.

Figura 18: Métodos aplicados a Jóvenes y adultos

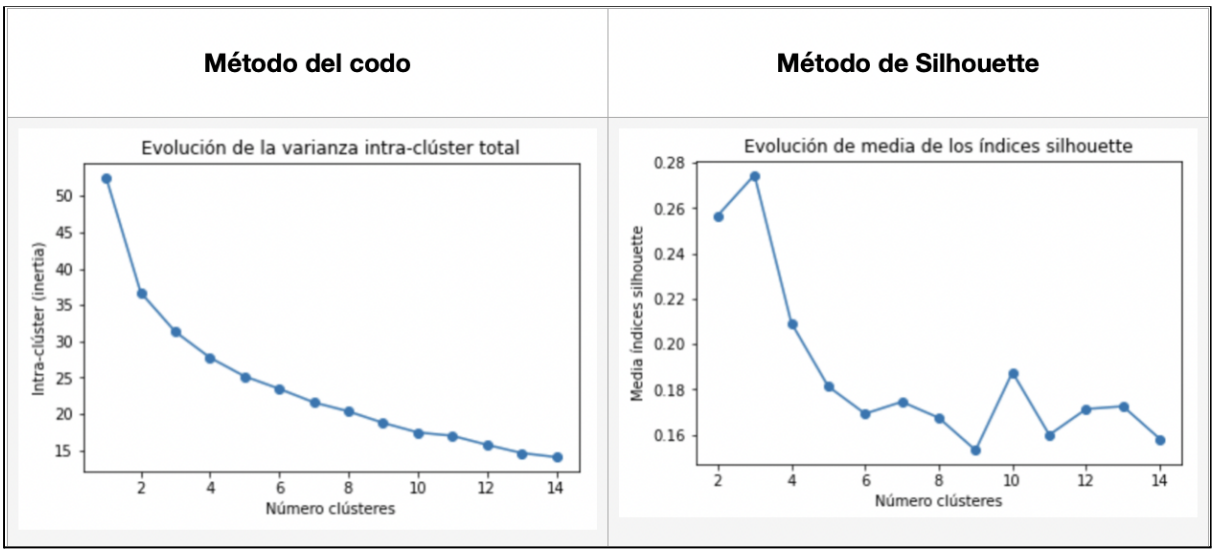


Fuente: Elaboración propia.

Con resultados similares (visualmente) al grupo anterior, este grupo de edad se repite el número óptimo de *clústeres* para trabajar de 2, siendo además 4 la segunda opción a trabajar, pero alejado del óptimo.

V. En el último rango de edad, el caso de los Adultos mayores, para un total de 1.253 registros médicos, se obtienen los siguientes *clústeres*. La Figura 19 muestra el número óptimo de *clústeres* para este rango de edad.

Figura 19: Métodos aplicados a Adultos mayores



Fuente: Elaboración propia.

Para este grupo de edad el número óptimo de *clústeres* para trabajar es de 3. Se logra observar en el método del codo que el descenso continúa siendo grande en 2, pero en el método de silhouette se logra apreciar de mejor manera que el punto más alto se encuentra cuando el número de clústeres es 3, determinando así la cantidad de *clústeres* en 3.

A continuación se resumen en una tabla la cantidad de *clústeres* determinados mediante ambos métodos utilizados por grupo de edad.

Tabla 5: Resumen de clústeres por grupo

Grupo de edad	**Cantidad de clústeres**
General (todas las edades).	2
Lactantes.	4
Primera infancia.	2
Infantes.	4
Jóvenes y adultos.	2
Adultos mayores.	3

4.5. Minería de los datos

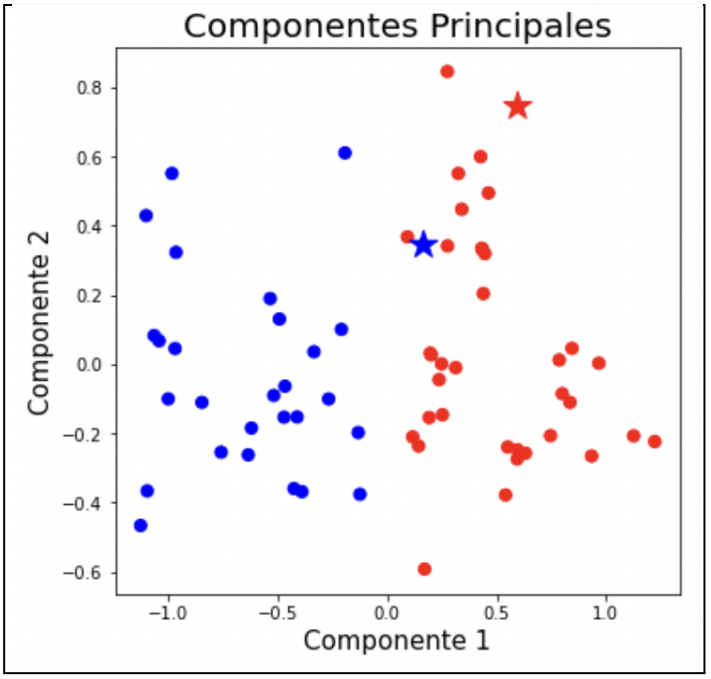
En esta fase se generan los agrupamientos dada la cantidad de *K* previamente determinado en la fase anterior, proporcionado lo necesario al algoritmo para que éste genere los grupos según la estructura de datos.

Como se menciona en apartados anteriores, se cuenta con un total de 16 variables, por lo que se hace necesario disminuir la cantidad de estas para graficar. Para esto, es necesario aplicar el proceso de PCA, indicando al algoritmo el número de componentes (Componente 1 y Componente 2), los cuales ayudan a interpretar y graficar los datos.

A continuación se muestran los clústeres obtenidos por grupo de edad;

I. Para todas las edades: En la Figura 20 se puede observar gráficamente el resultado de los clústeres formados para la totalidad de los 9.528 casos registrados, se observa además que los centroides se ven representados por estrellas con el color referente al *cluster* del cual pertenecen. Recordando también que el número de clústeres entregados al algoritmo *K-Means* para este grupo es de 2 (Ver tabla 5).

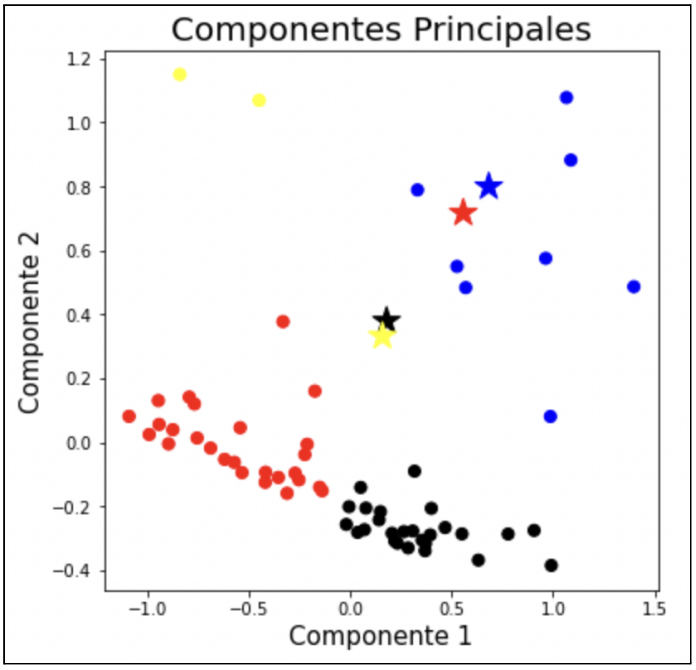
Figura 20: Clústeres formados con la totalidad de los casos



Fuente: Elaboración propia.

II. Lactantes: En la Figura 21 se puede observar el resultado de los clústeres formados para el grupo de Lactantes correspondiente a 391 casos registrados, los centroides se ven representados por estrellas del color que corresponden a cada clúster. La cantidad de cluster entregados al algoritmo K-Means es de 4 (Ver Tabla 5). Se observa además que los grupos de color amarillo y azul tienen menos cantidad de datos en comparación al rojo y al negro.

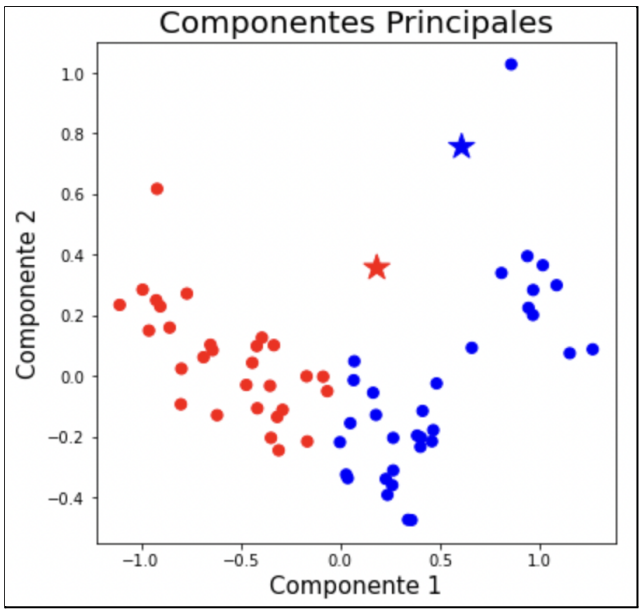
Figura 21: Clústeres formados para Lactantes



Fuente: Elaboración propia.

III. Primera infancia: Representados en la Figura 22 se puede observar el resultado de los clústeres formados para el grupo de Primera infancia correspondiente a 693 casos registrados, los centroides se ven representados por estrellas del color que corresponden a cada cluster. La cantidad de clústeres entregados al algoritmo K-Means es de 2 (Ver Tabla 5). Se observa además un agrupamiento claro y diferenciado uno del otro.

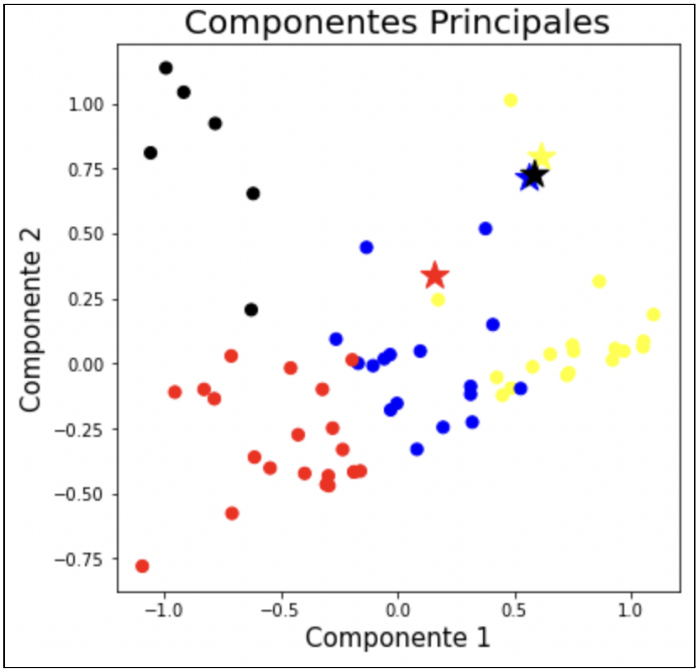
Figura 22: Clústeres formados para Primera infancia



Fuente: Elaboración propia.

IV. Infantes: En la Figura 23 se puede observar el resultado de los clústeres formados para el grupo de Infantes correspondiente a 661 casos registrados, los centroides se ven representados por estrellas del color que corresponden a cada *clúster*. La cantidad de clústeres entregados al algoritmo *K-Means* es de 4 (Ver Tabla 5). Se observa también que la cantidad de datos asignados por *K-Means* a cada grupo es similar entre uno y otro. Además los centroides para los *clústeres* azul, negro y amarillo están muy cerca entre ellos.

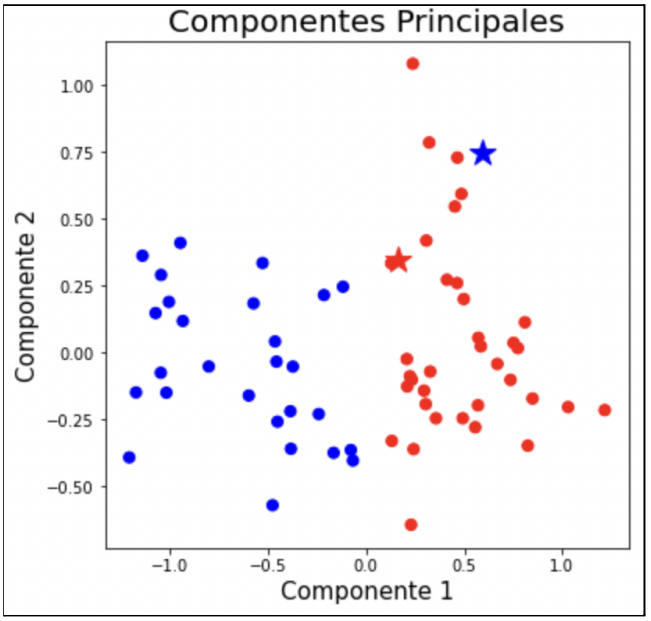
Figura 23: Clústeres formados para Infantes



Fuente: Elaboración propia.

V. Jóvenes y adultos: Para el caso del grupo de Jóvenes y adultos correspondiente a 6.530 casos registrados los *clústeres* formados se observan en la Figura 24, los centroides se ven representados por estrellas del color que corresponden a cada *clúster*. La cantidad de *clústeres* entregados al algoritmo *K-Means* es de 2 (Ver Tabla 5). Este grupo contiene la mayor cantidad de casos, y corresponde al grupo que tiene mayor representatividad en el análisis general de los casos, se puede apreciar por esto una distribución similar al grupo I. Para todos los casos, con *clústeres* bien definidos y claros.

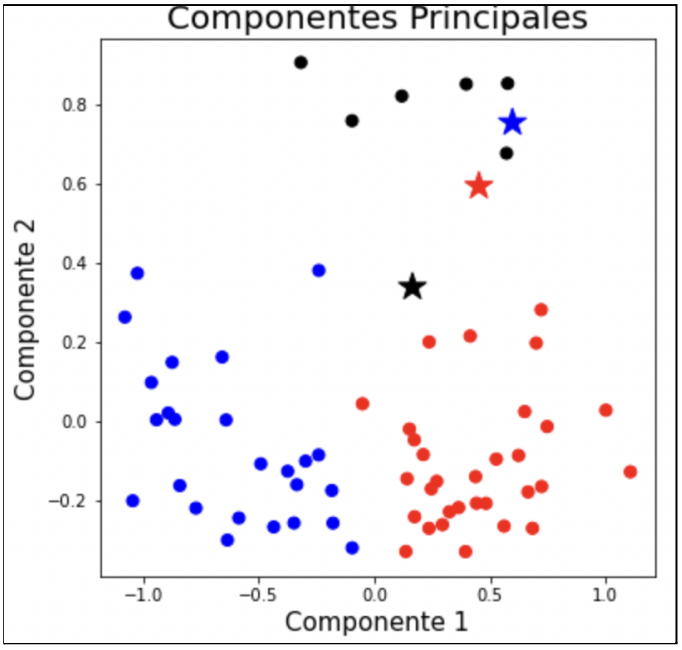
Figura 24: Clústeres formados para Jóvenes y adultos



Fuente: Elaboración propia.

VI. Adultos mayores: Para el último caso, del grupo de adultos mayores correspondiente a 1.253 casos registrados, los *clústeres* formados se muestran en la Figura 25, los centroides se ven representados por estrellas del color que corresponden a cada clúster. La cantidad de clústeres entregados al algoritmo K-Means es de 3 (Ver Tabla 5). Se observa además que el grupo de color negro cuenta con menos datos en comparación a los de otros colores.

Figura 25: Clústeres formados para Adultos mayores



Fuente: Elaboración propia.

4.6. Análisis de los resultados y validación de clústeres

En esta última etapa se analizan los componentes creados a través del PCA y validación de los clústeres, resultantes de la etapa anterior.

Estos componentes contienen pesos asignados a cada variable, para determinar qué variable aporta una mayor información al componente, es necesario observar su valor absoluto, en otras palabras, los mayores valores absolutos son los que representan o caracterizan de mejor manera al componente (Amat Rodrigo, J., 2020).

Para realizar el análisis de los *clústeres*, es necesario relacionar las variables con los componentes que las representan (componente 1 y componente 2), para esto se crea una matriz de correlación entre ambas métricas (componentes y variables).

Con la idea de apoyar los resultados que se obtienen de la matriz de correlación, se genera un mapa de calor para cada componente y la significancia que tiene cada variable en ellos.

Los mapas de calor usualmente son utilizados para explorar datos, ayudar a identificar patrones y cambios (Leland Wilkinson & Michael Friendly, 2009).

A continuación se exponen los análisis de los resultados, estas observaciones según segmentación de edad.

I. **Para todas las edades**: Para analizar la totalidad de los casos se crea una matriz de correlación entre los componentes y las variables. Los pesos asignados a cada variable son leídos en valor absoluto. La Tabla 6 muestra esta correlación y los pesos asignados por el algoritmo.

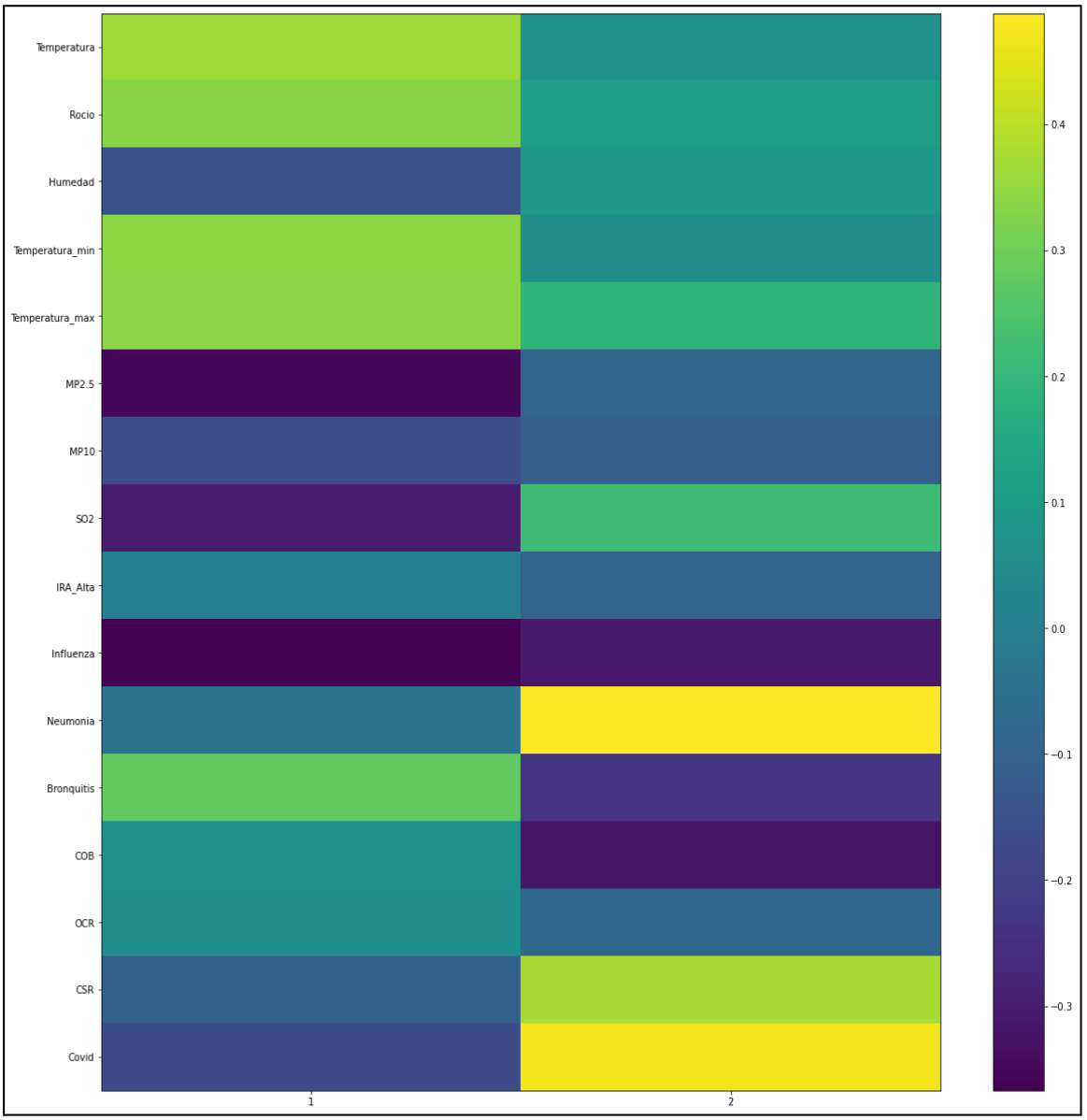
Tabla 6: Matriz de correlación para todos los casos.

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid
PC1	0.367437	0.334518	-0.154872	0.341351	0.338469	-0.347653	-0.163756	-0.297896	0.003560	-0.367020	-0.036641	0.280116	0.068321	0.059640	-0.100322	-0.169667
PC2	0.070403	0.108109	0.088469	0.054354	0.191230	-0.087657	-0.108968	0.217728	-0.096385	-0.307731	0.488164	-0.234452	-0.321645	-0.078689	0.375963	0.469880

Los pesos asignados en la primera componente (PC1) a las variables temperatura, rocío, temperatura_min, temperatura_max, MP2.5, SO_2, influenza y bronquitis son similares entre ellas (en valor absoluto). Esto quiere decir que la primera componente recoge mayoritariamente la información correspondiente a esas variables. Por otro lado, los pesos asignados (valores numéricos dentro del componente) en la segunda componente (PC2) a las variables neumonía, y covid son similares entre ellas, esto quiere decir que la segunda componente recoge información mayoritariamente de esas variables.

Para este caso, se confirma que existe una relación directa entre neumonía y COVID-19 sin influencias de factores ambientales como temperatura, rocío o humedad, ni contaminantes como material particulado o dióxido de azufre. Además las enfermedades respiratorias Bronquitis e Influenza tienen una relación con las variables de Temperatura, Rocío y con las variables de Material particulado de 2.5 micras (MP2.5) y el Dióxido de azufre (SO_2). Para observar con mayor claridad las relaciones entre la variables y componentes ver Figura 26.

Figura 26: Mapa de calor para todos los casos



Fuente: Elaboración propia.

Para el total de los casos se observa una similitud dentro del componente 1 (PC1) entre las variables temperatura, rocío, temperatura_min, temperatura_max y bronquitis representadas por un color más cálido que el resto de variables, pero a la vez también se encuentra una similitud entre las variables MP2.5, SO_2 e influenza representadas por un color más oscuro.

En el caso del componente 2 (PC2) se observa una clara similitud entre las variables neumonía y covid, seguido por CSR representadas por los colores más cálidos, y de la misma manera pero con colores fríos u oscuros se observan influenza, bronquitis y crisis obstructiva bronquial (COB).

II. **Lactantes**: Para analizar este grupo de edad se crea una matriz de correlación entre los componentes y las variables. Los pesos asignados a cada variable son leídos en valor absoluto. La Tabla 7 muestra esta correlación y los pesos asignados por el algoritmo.

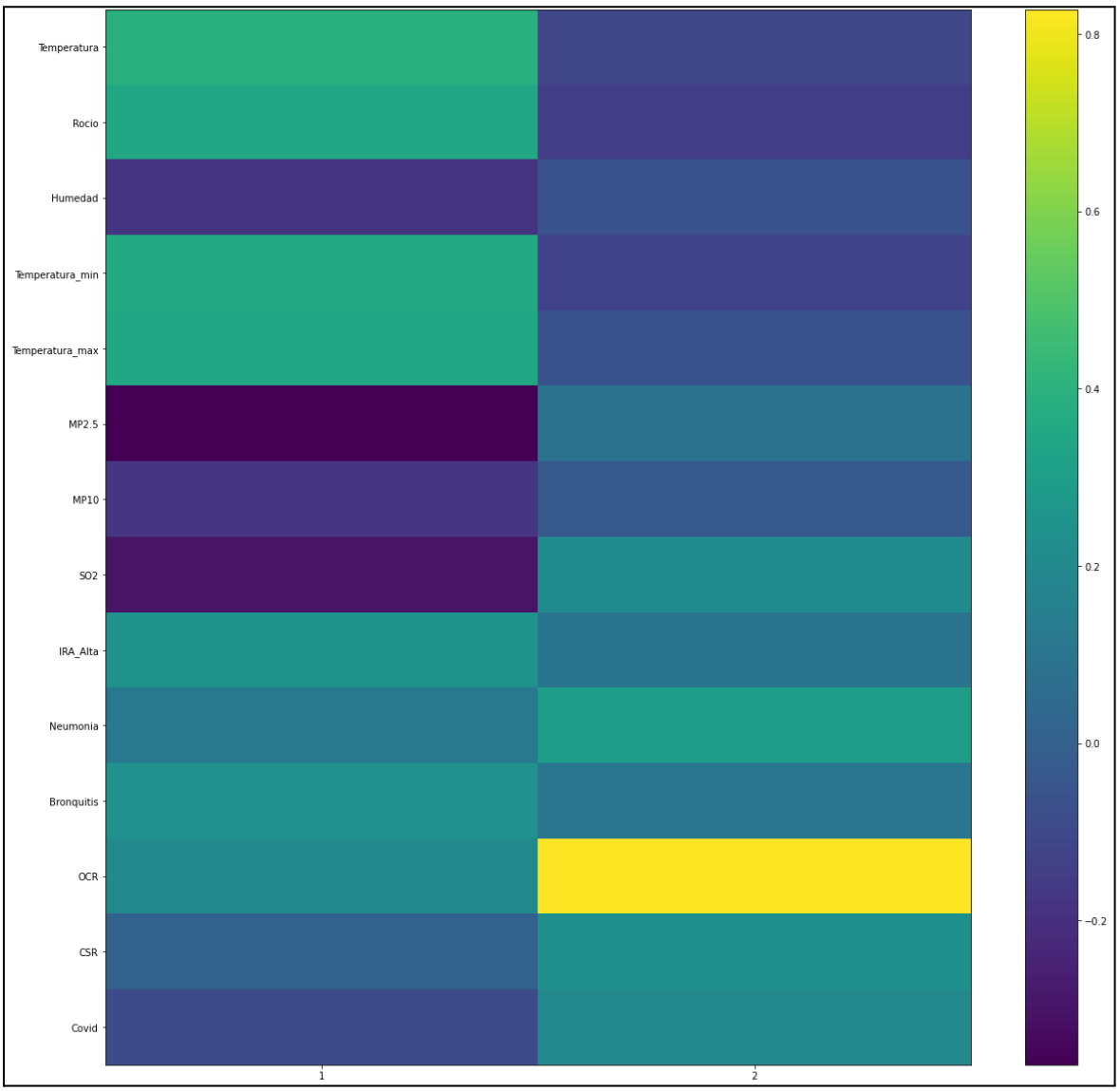
Tabla 7: Matriz de correlación para Lactantes

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Neumonia	Bronquitis	OCR	CSR	Covid
PC1	0.393484	0.351510	-0.180909	0.361482	0.352943	-0.362628	-0.175227	-0.300748	0.243906	0.117739	0.239178	0.207960	0.008179	-0.084371
PC2	-0.107282	-0.146355	-0.061230	-0.122929	-0.062846	0.091825	-0.028050	0.210648	0.094530	0.302101	0.104866	0.827857	0.233004	0.199584

Para el grupo de Lactantes no se presentaron casos de influenza ni de COB en el periodo de tiempo analizado. En la primera componente (PC1) los pesos asignados a las variables temperatura, rocío, temperatura_min, temperatura_max, MP2.5, SO_2, IRA_Alta y bronquitis son similares, es decir que de esas variables se recoge en mayor cantidad la información. Por otro lado, en la segunda componente (PC2) existe un mayor peso en Otra Causa Respiratoria (OCR) y seguido por neumonía, recogiendo de esas variables la mayor cantidad de información.

Para los datos correspondientes a Lactantes (391 casos), se puede concluir que las enfermedades respiratorias Infección respiratoria alta (IRA) tienen una relación con las variables de temperatura, rocío y con las variables de Material particulado de 2.5 micras (MP2.5) y el Dióxido de azufre (SO_2). Además existe una relación en los casos de neumonía y los casos de Otras Causas Respiratorias (OCR), es decir, de otras enfermedades respiratorias atendidas de urgencia, de las cuales no se encuentran entre las más comunes o no se encuentran detalladas en el sistema del Hospital. Para observar con mayor claridad las relaciones entre la variables y componentes ver Figura 27.

Figura 27: Mapa de calor para Lactantes



Fuente: Elaboración propia.

En el mapa de calor generado para el grupo de Lactantes se observa que en el componente 1 (PC1) hay una relevancia mediante colores cálidos por las variables ambientales de temperatura, rocío, temperatura_min, temperatura_max, y las enfermedades de IRA_Alta y bronquitis, mientras que los colores más oscuros muestran una relevancia por las variables contaminantes de MP2.5 y SO_2 .

Por otro lado, en la segunda componente (PC2) se observa una mayor relevancia dentro de los colores cálidos por la variable de OCR y neumonía, mientras que en colores oscuros se observa una relevancia por las variables ambientales de rocío y temperatura_min.

III. **Primera infancia**: Para analizar este grupo de edad se crea una matriz de correlación entre los componentes y las variables. Los pesos asignados a cada variable son leídos en valor absoluto. La Tabla 8 muestra esta correlación y los pesos asignados por el algoritmo.

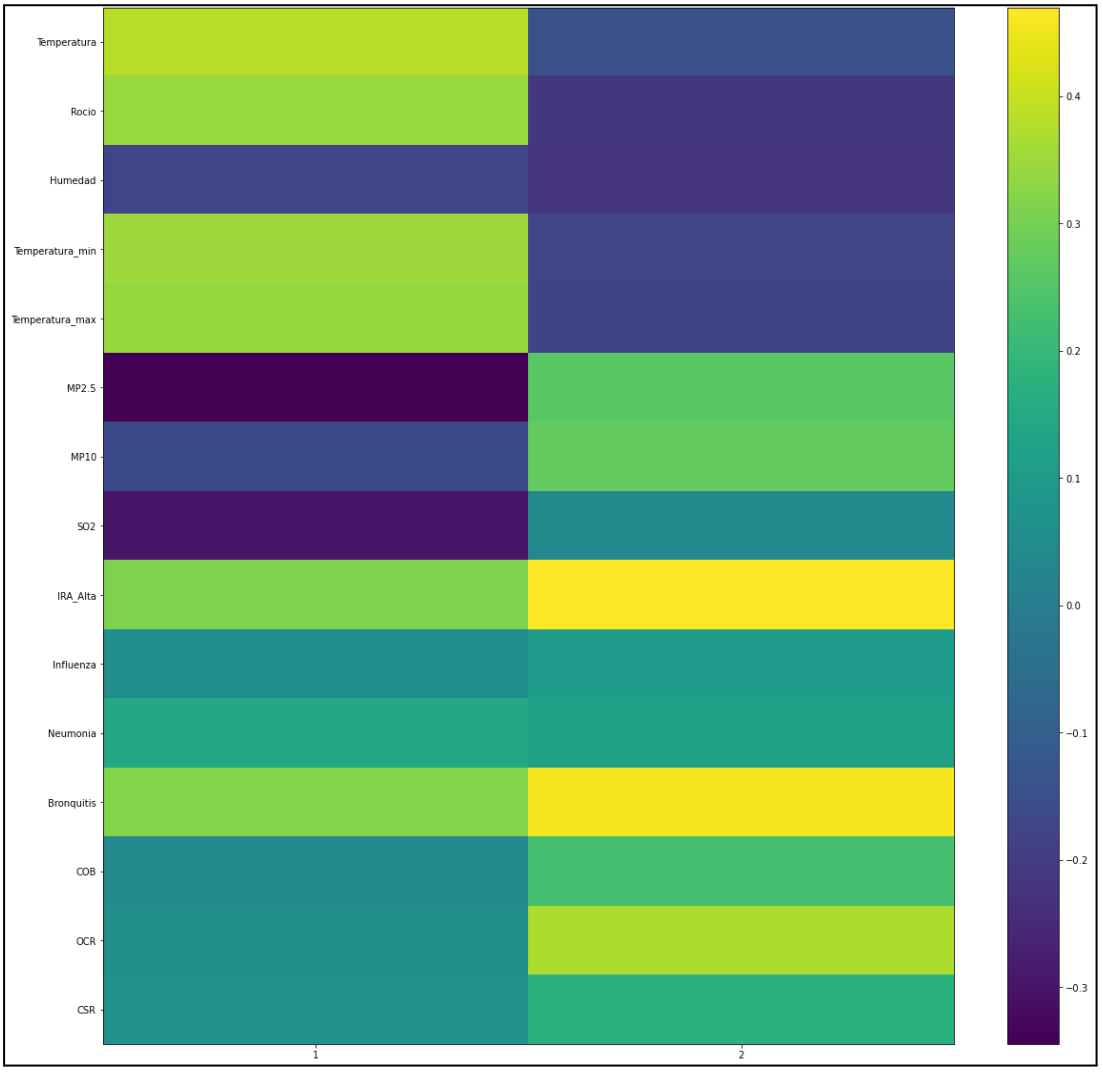
Tabla 8: Matriz de correlación para Primera infancia

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR
PC1	0.381200	0.343333	-0.173114	0.351599	0.340092	-0.344628	-0.160632	-0.298147	0.312347	0.055114	0.148256	0.319023	0.046058	0.056589	0.066451
PC2	-0.142895	-0.213390	-0.219344	-0.173026	-0.181431	0.260969	0.274129	0.046024	0.469894	0.107027	0.117417	0.460144	0.230942	0.368873	0.174738

Para el grupo de Primera infancia no se presentaron casos de covid, en el periodo de tiempo analizado por la investigación. En la primera componente (PC1) los pesos asignados a las variables temperatura, rocío, temperatura_min, temperatura_max, MP2.5, SO_2, IRA_Alta y bronquitis son similares, esto quiere decir que de esas variables se recoge mayoritariamente información. Por su parte, el segundo componente (PC2) obtiene mayormente la información de las variables IRA_Alta, bronquitis y OCR.

En el grupo de Primera infancia (693 casos), se concluye algo similar al grupo de Lactantes, es decir, que las enfermedades de infección respiratoria alta y bronquitis tienen una relación directa con la temperatura, rocío y con las variables de material particulado de 2.5 micras (MP2.5) y el dióxido de azufre (SO_2). Mencionando también una relación similar al caso de Lactantes, que OCR tienen una relación con Infección respiratoria alta y bronquitis, mostrando así que estas dos enfermedades respiratorias son las que más se destacan en este grupo de edad. Para observar con mayor claridad las relaciones entre la variables y componentes ver Figura 28.

Figura 28: Mapa de calor para Primera infancia



Fuente: Elaboración propia.

En el mapa de calor para el grupo de Primera infancia se observa que en la primera componente (PC1) resalta en colores más cálidos las variables de temperatura, rocío, temperatura_min, temperatura_max, IRA_Alta y bronquitis, además por el lado de los colores más fríos se observa la relevancia de MP2.5 y SO_2.

En el caso de la segunda componente (PC2) se observa la relevancia de los colores más cálidos representando a las variables de IRA_Alta, bronquitis y OCR, en cambio por los colores más fríos resaltan las variables de rocío y humedad.

IV. **Infantes**: Para analizar este grupo de edad se crea una matriz de correlación entre los componentes y las variables. Los pesos asignados a cada variable son leídos en valor absoluto. La Tabla 9 muestra esta correlación y los pesos asignados por el algoritmo.

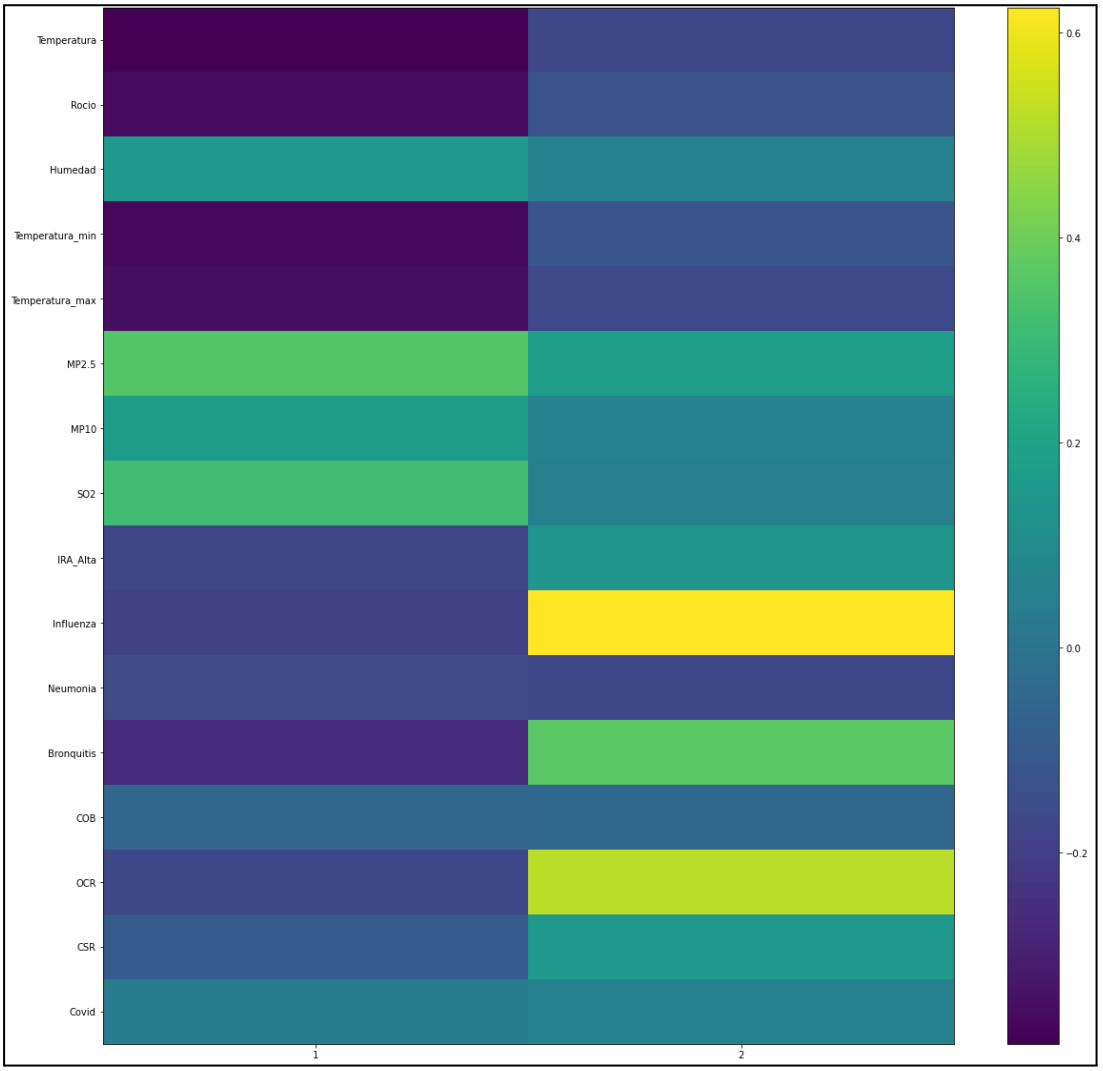
Tabla 9: Matriz de correlación para Infantes

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid
PC1	-0.386368	-0.356126	0.156558	-0.363699	-0.343674	0.352303	0.172928	0.313651	-0.17436	-0.188018	-0.150394	-0.258687	-0.054250	-0.166711	-0.095598	0.037845
PC2	-0.165625	-0.128642	0.063723	-0.120141	-0.162381	0.176535	0.062437	0.052098	0.14605	0.624749	-0.168417	0.364073	-0.047596	0.520617	0.155210	0.058548

Para el caso del grupo de Infantes, en la primera componente (PC1) los pesos asignados a las variables temperatura, rocío, temperatura_min, temperatura_max, MP2.5, SO2, y bronquitis son similares, esto quiere decir que de esas variables se recoge la mayor cantidad de información. Por su parte, el segundo componente (PC2) obtiene mayormente la información de las variables influenza, bronquitis y OCR.

En este grupo de Infantes (661 casos), se concluye que la enfermedad de bronquitis tiene una relación con las variables de temperatura, rocío y con las variables de material particulado de 2.5 micras (MP2.5) y el dióxido de azufre (SO_2). Además de una relación existente entre influenza y OCR de atención urgencias, seguidas en esta relación por bronquitis. Aparece así un patrón con respecto del grupo de edad de Primera infancia para OCR y bronquitis. Para observar con mayor claridad las relaciones entre la variables y componentes ver Figura 29.

Figura 29: Mapa de calor para Infantes



Fuente: Elaboración propia.

Observando el mapa de calor para el grupo de Infantes, en la primera componente (PC1) se resaltan a través de los colores más cálidos las variables de MP2.5 y SO_2, mientras tanto que, por los colores más fríos se resalta la relevancia de las variables temperatura, Rocío, temperatura_min, temperatura_max, y bronquitis

Por su parte, el segundo componente (PC2) los colores más cálidos resaltan las variables influenza, bronquitis y OCR, mientras que los colores más oscuros no resalta variable alguna sobre otra de manera significativa.

V. **Jóvenes y adultos**: Para analizar este grupo de edad se crea una matriz de correlación entre los componentes y las variables. Los pesos asignados a cada variable son leídos en valor absoluto. La Tabla 10 muestra esta correlación y los pesos asignados por el algoritmo.

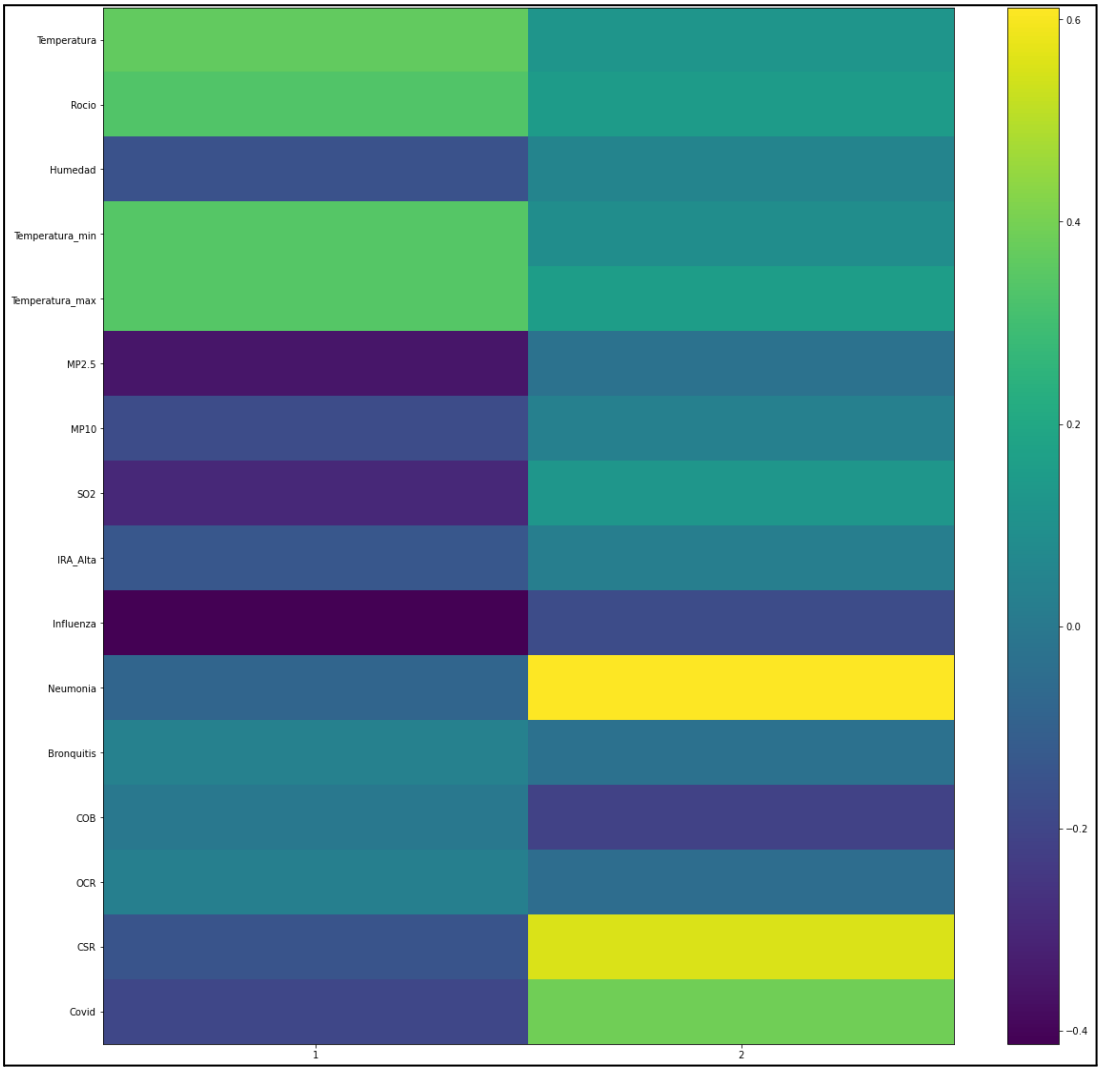
Tabla 10: Matriz de correlación para Jóvenes y adultos

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid
PC1	0.365421	0.332134	-0.154866	0.340626	0.342528	-0.352876	-0.175678	-0.296056	-0.134122	-0.413145	-0.079396	0.036721	-0.004405	0.031168	-0.145286	-0.193283
PC2	0.119731	0.151099	0.049310	0.089015	0.158260	-0.028042	0.034329	0.125675	0.023373	-0.172897	0.611972	-0.030820	-0.209828	-0.045990	0.553765	0.388556

Para el caso del grupo de Jóvenes y adultos (6.530 casos), en la primera componente (PC1) los pesos asignados a las variables temperatura, rocío, temperatura_min, temperatura_max, MP2.5, SO_2 e influenza son similares, esto quiere decir que de esas variables se recoge la mayor cantidad de información. Por su parte, el segundo componente (PC2) obtiene mayormente la información de las variables neumonía, Causa Sistema Respiratorio (CSR) y covid por amplia mayoría sobre las demás variables.

Este grupo de edad, Jóvenes y adultos concentra la mayor cantidad de casos por grupo de edad (6.530 casos). Se concluye que la enfermedad de influenza tiene relación con las variables de temperatura, rocío y con las variables de material particulado de 2.5 micras (MP2.5) y el dióxido de azufre (SO_2). Se puede apreciar además una relación interesante entre los casos de neumonía, CSR, es decir, hospitalizaciones alusivas al sistema respiratorio y el nuevo síndrome respiratorio SARS-CoV-2 o COVID19. Para observar con mayor claridad las relaciones entre la variables y componentes ver Figura 30.

Figura 30: Mapa de calor para Jóvenes y adultos



Fuente: Elaboración propia.

El mapa de calor correspondiente para grupo de Jóvenes y adultos, muestra que en la primera componente (PC1) los colores cálidos dan relevancia a las variables temperatura, rocío, temperatura_min y temperatura_max, mientras que los colores fríos resaltan las variables de MP2.5, SO_2 e influenza por sobre el resto.

En cambio el segundo componente (PC2) a través de colores cálidos resalta las variables neumonía, CSR y covid de manera más significativa que el resto. Por otro lado, los colores oscuros resaltan las variables de Influenza y COB pero sin colores tan resaltados del resto.

VI. **Adultos mayores**: Para analizar este grupo de edad se crea una matriz de correlación entre los componentes y las variables. Los pesos asignados a cada variable son leídos en valor absoluto. La Tabla 11 muestra esta correlación y los pesos asignados por el algoritmo.

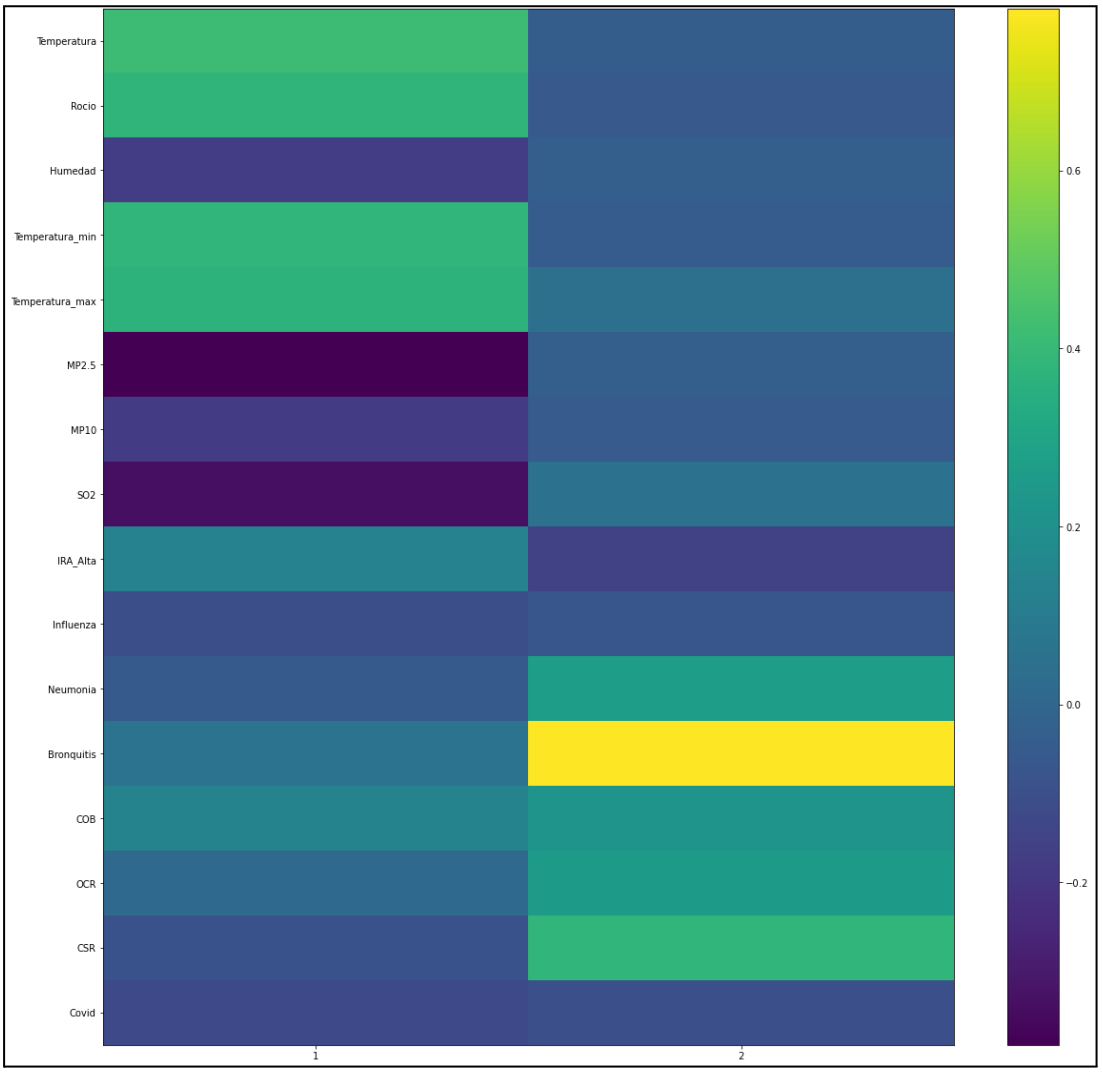
Tabla 11: Matriz de correlación para Adultos mayores

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid
PC1	0.415688	0.378812	-0.171201	0.38461	0.371518	-0.382820	-0.174841	-0.335088	0.135704	-0.104781	-0.050673	0.060401	0.137158	0.014942	-0.091327	-0.126700
PC2	-0.038547	-0.058572	-0.027845	-0.04143	0.047616	-0.030432	-0.047093	0.058016	-0.155784	-0.068775	0.271550	0.782456	0.219764	0.252651	0.382903	-0.100477

Para el último grupo, el de Adultos mayores, en la primera componente los pesos asignados a las variables temperatura, rocío, temperatura_min, temperatura_max, MP2.5 y SO_2, son similares, esto quiere decir que de esas variables se recoge la mayoritariamente la información. Por otro lado, el segundo componente obtiene en su mayoría la información de las variables bronquitis y CSR por una amplia diferencia sobre las demás variables.

Para este grupo de edad, Adultos mayores (1.253 casos). Se concluye que no se genera relación entre alguna enfermedad respiratoria y variables ambientales ni tampoco con variables contaminantes. Sin embargo si se aprecia una relación entre las enfermedades de bronquitis y CSR, es decir, casos de hospitalización derivados por el sistema respiratorio. Para observar con mayor claridad las relaciones entre la variables y componentes ver Figura 31.

Figura 31: Mapa de calor para Adultos mayores



Fuente: Elaboración propia.

En el caso del mapa de calor para el último grupo, el de Adultos mayores, en la primera componente (PC1) los colores más cálidos resaltan a las variables temperatura, rocío, temperatura_min y temperatura_max, mientras que los colores más fríos resaltan las variables MP2.5 y SO_2.

Por otro lado, en el segundo componente (PC2) se puede observar la relevancia mediante colores más cálidos de las variables bronquitis y CSR por una amplia diferencia sobre las demás variables, mientras que no hay una preferencia clara por los colores más fríos.

En resumen se puede apreciar lo siguiente:

Tabla 12: Resumen de análisis de componentes

Grupo	PC1	PC2
Para todas las edades	Temperatura, Rocío, Temperatura_min, Temperatura_max, MP2.5, SO2, Influenza y Bronquitis	Neumonía, y Covid
Lactantes	Temperatura, Rocio, Temperatura_min, Temperatura_max MP2.5, SO2, IRA_Alta y Bronquitis	OCR y Neumonía
Primera infancia.	Temperatura, Rocio, Temperatura_min, Temperatura_max MP2.5, SO2, IRA_Alta y Bronquitis	IRA_Alta, Bronquitis y OCR
Infantes.	Temperatura, Rocio, Temperatura_min, Temperatura_max MP2.5, SO2 y Bronquitis	Influenza, Bronquitis y OCR
Jóvenes y adultos.	Temperatura, Rocio, Temperatura_min, Temperatura_max MP2.5, SO2 e Influenza	Neumonía, CSR y Covid.
Adultos mayores.	Temperatura, Rocio, Temperatura_min, Temperatura_max MP2.5 y SO2	Bronquitis y CSR

4.6.1. Validación de clustering

La evaluación de los resultados de un algoritmo de clustering es importante, sin embargo definir si el resultado del agrupamiento es aceptable o no se vuelve difícil, es por esto que existen diversas formas de validar los clústeres (E. L. Guzmán, 2016). En esta sección se muestra una forma de validar los clústeres, analizar cómo están agrupadas las variables y una comparación en cuanto a los resultados del PCA.

Con el módulo “model_validation” de la librería “funpymodeling” se puede validar el modelo de clúster, para esto se indica el *data frame* y el nombre de la variable por la cual se hace el clúster. Esta función devuelve una tupla de tablas, la primera tabla muestra los promedios de

los valores reales de las variables y la segunda tabla muestra estos valores normalizados, asignando un valor 0 a las variables menos representativas en el clúster, un valor 1 a las variables más representativas y un valor intermedio a las variables dependiendo de su representatividad en el clúster.

I. **Para todas las edades**: En este grupo de edad se realiza una validación de clústeres obteniendo la siguiente tupla de tablas:

Tabla 13: Validación de clustering para todas las edades parte 1

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid	Cluster
0	16.320000	11.286667	72.590000	13.37000	20.753333	11.613333	37.8800	2.923333	64.166667	9.733333	5.033333	7.866667	12.266667	8.1000	6.166667	5.166667	0
1	15.065625	10.450000	74.715625	12.28125	21.212500	10.562500	30.9125	4.056250	49.750000	5.125000	6.406250	2.375000	9.625000	7.0625	7.343750	19.156250	1

Se observa en la Tabla 13 generada por el modelo de validación, el promedio de los valores reales de las variables en cada clúster.

Tabla 14: Validación de clustering para todas las edades parte 2

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid	Cluster
0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0
1	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	1

La Tabla 14 es la segunda tabla formada por el modelo de validación. Esta tabla confirma lo que se obtiene del PCA y mapa de calor para la totalidad de los casos; en el clúster 0 se destaca el agrupamiento de bronquitis junto a las variables ambientales. También el modelo muestra que las enfermedades neumonía y covid se encuentran agrupadas en el mismo clúster 1.

El método de validación de clustering ofrece un mayor detalle en cuanto a la distribución de las variables en los clúster, en comparación al PCA. Comparando ambos resultados se obtiene lo siguiente:

Tabla 15: Comparación de resultados para todas las edades

Variables	Comparación			
	PCA		Validación de clustering	
	Componente 1	Componente 2	Clúster 0	Clúster 1
Temperatura	●		●	
Rocío	●		●	
Humedad				●
Temperatura_min	●		●	
Temperatura_max	●			●
MP2.5	●		●	
MP10			●	
SO2	●			●
IRA_Alta			●	
Influenza	●		●	
Neumonía		●		●
Bronquitis	●		●	
COB			●	
OCR			●	
CSR		●		●
Covid		●		●

● Mayor representatividad de la variable.

Se observa que el método de validación de clustering muestra la representatividad de todas las variables, en cambio el PCA deja fuera aquellas variables que no tienen mayor representación para los componentes. Sin embargo se observa un patrón en ambos métodos, describiendo el comportamiento de bronquitis y las variables ambientales, y de la misma manera el comportamiento de neumonía y covid.

II. **Lactantes**: En este grupo de edad se realiza una validación de clústeres obteniendo la siguiente tupla de tablas:

Tabla 16: Validación de clustering para Lactantes parte 1

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Neumonia	Bronquitis	OCR	CSR	Covid	Cluster
0	15.246154	10.584615	74.607692	12.538462	20.446154	11.192308	38.800000	2.892308	3.846154	0.000000	0.230769	0.076923	0.230769	0.153846	0
1	12.766667	8.466667	75.758333	10.283333	18.675000	12.983333	24.708333	5.133333	2.000000	0.166667	0.166667	0.166667	0.250000	0.333333	1
2	18.075000	12.825000	71.639286	15.314286	23.325000	8.385714	29.292857	2.328571	6.142857	0.107143	2.071429	0.214286	0.071429	0.071429	2
3	12.688889	8.300000	75.966667	8.766667	17.600000	16.700000	56.055556	5.900000	2.666667	0.000000	0.222222	0.000000	0.000000	0.222222	3

Se observa en la Tabla 16 generada por el modelo de validación, el promedio de los valores reales de las variables en cada clúster.

Tabla 17: Validación de clustering para Lactantes parte 2

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Neumonia	Bronquitis	OCR	CSR	Covid	Cluster
0	0.474789	0.504887	0.685959	0.576056	0.497145	0.337563	0.449535	0.157846	0.445623	0.000000	0.033654	0.358974	0.923077	0.314685	0
1	0.014440	0.036832	0.951857	0.231636	0.187773	0.552978	0.000000	0.785333	0.000000	1.000000	0.000000	0.777778	1.000000	1.000000	1
2	1.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.146250	0.000000	1.000000	0.642857	1.000000	1.000000	0.285714	0.000000	2
3	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	0.160920	0.000000	0.029167	0.000000	0.000000	0.575758	3

En la Tabla 17 se observa que en el clúster 0 se encuentra el índice más bajo de neumonía y el segundo más alto de CSR. En el clúster 1 se encuentra el índice más alto de neumonía, CSR y covid. En el clúster 2 se encuentran los índices más altos de temperatura, rocío, temperatura_min, temperatura_max, IRA_Alta, bronquitis y OCR. Por último en el clúster 3 se observan los índices más altos de humedad, MP2.5, MP10 y SO_2. Todas estos índices confirman los resultados del PCA y mapas de calor.

En resumen se tiene lo siguiente:

Tabla 18: Comparación de resultados para Lactantes

Variables	Comparación					
	PCA		Validación de clustering			
	Componente 1	Componente 2	Clúster 0	Clúster 1	Clúster 2	Clúster 3
Temperatura	●				●	
Rocío	●				●	
Humedad						●
Temperatura_min	●				●	
Temperatura_max	●				●	
MP2.5	●					●
MP10						●
SO2	●					●
IRA_Alta	●				●	
Neumonía		●		●		
Bronquitis	●				●	
OCR		●			●	
CSR		●		●		
Covid		●		●		

● Mayor representatividad de la variable.

Para este grupo de edad se repite el comportamiento de bronquitis con respecto a las variables ambientales, además del comportamiento de neumonía y covid. El algoritmo los agrupa siempre juntos. Además se puede observar que en el método de validación de clustering, el clúster 0 solo contiene valores intermedios de las variables. La comparación muestra nuevamente similitud en el agrupamiento de las variables, el clúster 1 confirma lo obtenido mediante el componente 2, y a su vez el clúster 2 confirma lo obtenido a través del componente 1.

III. **Primera infancia**: En este grupo de edad se realiza una validación de clústeres obteniendo la siguiente tupla de tablas:

Tabla 19: Validación de clustering para Primera infancia parte 1

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Cluster
0	16.30625	11.447917	73.425000	13.637500	21.810417	9.760417	29.329167	3.13125	8.395833	0.125000	0.187500	3.062500	0.145833	0.166667	0.145833	0
1	13.50000	8.821429	74.585714	9.964286	18.178571	15.564286	51.271429	4.80000	3.642857	0.071429	0.071429	0.714286	0.000000	0.071429	0.142857	1

Se observa en la Tabla 19 generada por el modelo de validación, el promedio de los valores reales de las variables en cada clúster.

Tabla 20: Validación de clustering para Primera infancia parte 2

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Cluster
0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0
1	0.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1

En la Tabla 20 se observa que el clúster 0 contiene los índices más altos de temperatura, rocío, tempratura_min, temperatura_max, IRA_Alta, influenza, neumonía, bronquitis, COB, OCR y CSR. Por otro lado el clúster 1 contiene los índices más altos de humedad y variables contaminantes (MP2.5, MP10 y SO_2). Estos índices confirman y ayudan a comprender más el PCA y el mapa de calor.

En resumen se tiene lo siguiente:

Tabla 21: Comparación de resultados para Primera infancia

Variables	Comparación			
	PCA		Validación de clustering	
	Componente 1	Componente 2	Clúster 0	Clúster 1
Temperatura	●		●	
Rocío	●		●	
Humedad				●
Temperatura_min	●		●	
Temperatura_max	●		●	
MP2.5				●
MP10				●
SO2				●
IRA_Alta	●	●	●	
Influenza			●	
Neumonía	●		●	
Bronquitis	●	●	●	
COB		●	●	
OCR		●	●	
CSR			●	

● Mayor representatividad de la variable.

Este grupo de edad muestra algo interesante, en primer lugar aparece nuevamente el comportamiento de bronquitis con las variables ambientales, y en segundo lugar el PCA muestra a su vez el comportamiento de bronquitis con enfermedades como IRA, COB y OCR. Además el modelo de validación muestra que todas estas variables mencionadas están agrupadas en el mismo clúster 0, y que las variables menos representativas en el PCA están agrupadas a su vez en el clúster 1 del modelo de validación.

IV. **Infantes**: en este grupo de edad se realiza una validación de clústeres obteniendo la siguiente tupla de tablas:

Tabla 22: Validación de clustering para Infantes parte 1

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid	Cluster
0	13.072727	8.536364	75.172727	9.345455	17.718182	16.554545	53.927273	5.390909	4.272727	0.181818	0.000000	0.454545	0.090909	0.090909	0.000000	0.000000	0
1	12.923077	8.592308	75.646154	10.469231	19.038462	12.446154	24.761538	4.953846	4.538462	0.000000	0.000000	0.538462	0.538462	0.461538	0.076923	0.307692	1
2	16.534783	11.686957	73.860870	13.734783	21.934783	9.713043	34.217391	3.052174	4.217391	0.173913	0.130435	1.043478	0.391304	0.304348	0.217391	0.000000	2
3	18.640000	13.240000	70.633333	15.953333	23.633333	7.940000	28.233333	1.573333	12.266667	0.600000	0.200000	2.866667	0.733333	0.666667	0.266667	0.000000	3

Se observa en la Tabla 22 generada por el modelo de validación, el promedio de los valores reales de las variables en cada clúster.

Tabla 23: Validación de clustering para Infantes parte 2

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid	Cluster
0	0.026177	0.000000	0.905557	0.000000	0.000000	1.000000	1.000000	1.000000	0.006875	0.303030	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0
1	0.000000	0.011894	1.000000	0.170066	0.223203	0.523087	0.000000	0.885513	0.039888	0.000000	0.000000	0.034789	0.696662	0.643725	0.288462	1.0	1
2	0.631757	0.669821	0.643856	0.664257	0.712847	0.205820	0.324211	0.387377	0.000000	0.289855	0.652174	0.244156	0.467596	0.370709	0.815217	0.0	2
3	1.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.119037	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.0	3

En la Tabla 23 se observa que en el clúster 0 se encuentran los índices más altos de las variables contaminantes (MP2.5, MP10 y SO_2) y el segundo índice más alto de humedad. En el clúster 1 se encuentra el índice más alto de covid, cabe mencionar que este grupo de edad presentó una cantidad de casos de covid más bajo de que el resto de grupos de edad. En el clúster 2 se encuentra el índice más bajo de IRA_Alta. Finalmente en el clúster 3 se encuentran los índices más altos de temperatura, rocío, temperatura_min, temperatura_max, y todas las enfermedades respiratorias (IRA_Alta, influenza, neumonía, bronquitis, COB, OCR y CSR) a excepción del covid. Estos índices confirman y ayudan a comprender aún más el PCA y el mapa de calor.

En resumen se tiene lo siguiente:

Tabla 24: Comparación de resultados para Infantes

Variables	Comparación					
	PCA		Validación de clustering			
	Componente 1	Componente 2	Clúster 0	Clúster 1	Clúster 2	Clúster 3
Temperatura	●					●
Rocío	●					●
Humedad				●		
Temperatura_min	●					●
Temperatura_max	●					●
MP2.5	●		●			
MP10			●			
SO2	●		●			
IRA_Alta						●
Influenza		●				●
Neumonía						●
Bronquitis	●	●				●
COB						●
OCR		●				●
CSR						●
Covid				●		

● Mayor representatividad de la variable.

En este grupo de edad se observa algo similar al anterior, nuevamente se muestra el comportamiento de bronquitis con respecto a las variables ambientales, pero a su vez, el PCA, muestra el comportamiento entre bronquitis e influenza y OCR. Mientras tanto en el modelo de validación de clustering se observa que todas esas variables se encuentran agrupadas en el mismo clúster 3. En este grupo de edad no hay agrupación entre neumonía y covid, lo que se puede atribuir a la escasa cantidad de casos de covid.

V. **Jóvenes y adultos**: en este grupo de edad se realiza una validación de clústeres obteniendo la siguiente tupla de tablas:

Tabla 25: Validación de clustering para Jóvenes y adultos parte 1

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid	Cluster
0	13.442857	8.907143	75.335714	10.000000	18.185714	14.842857	46.150000	5.064286	66.214286	15.571429	3.142857	0.071429	9.285714	3.214286	3.357143	17.500000	0
1	16.322917	11.422917	73.206250	13.627083	21.808333	9.970833	30.822917	3.054167	28.791667	4.229167	2.604167	0.083333	6.187500	4.020833	2.625000	5.916667	1

Se observa en la Tabla 25 generada por el modelo de validación, el promedio de los valores reales de las variables en cada clúster.

Tabla 26: Validación de clustering para Jóvenes y adultos parte 2

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid	Cluster
0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0
1	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1

Sobre la Tabla 26 se observa que en clúster 0 los índices más altos son los de humedad, variables contaminantes (MP2.5, MP10 y SO_2), IRA_Alta, influenza, neumonía, COB, CSR y covid. Por otra parte en el clúster 1, se encuentran los índices más altos de temperatura, rocío, temperatura_min, temperatura_max, bronquitis y OCR. Estos índices confirman la misma información obtenida del PCA y del mapa de calor.

En resumen se tiene lo siguiente:

Tabla 27: Comparación de resultados para Jóvenes y adultos

Variables	Comparación			
	PCA		Validación de clustering	
	Componente 1	Componente 2	Clúster 0	Clúster 1
Temperatura	●			●
Rocío	●			●
Humedad			●	
Temperatura_min	●			●
Temperatura_max	●			●
MP2.5			●	
MP10			●	
SO2			●	
IRA_Alta			●	
Influenza			●	
Neumonía		●	●	
Bronquitis	●			●
COB			●	
OCR	●			●
CSR		●	●	
Covid		●	●	

● Mayor representatividad de la variable.

Para este grupo se identifica nuevamente el mismo patrón que en los otros grupos de edad, el comportamiento de la bronquitis con respecto a las variables ambientales agrupadas en el clúster 1, también en el caso de neumonía con respecto al covid agrupadas en el clúster 0, ambas situaciones confirman lo obtenido a través del PCA. Se puede observar nuevamente que la validación de clustering entrega información más completa, de cada variable, en contraste del PCA que solo muestra información de las variables más relevantes.

VI. **Adultos mayores**: en este grupo de edad se realiza una validación de clústeres obteniendo la siguiente tupla de tablas:

Tabla 28: Validación de clustering para Adultos mayores parte 1

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid	Cluster
0	17.127273	12.063636	72.487879	14.445455	22.481818	9.000000	32.206061	2.254545	1.939394	0.060606	2.393939	0.121212	4.272727	3.030303	3.151515	0.969697	0
1	13.072727	8.536364	75.172727	9.345455	17.718182	16.554545	53.927273	5.390909	0.818182	0.090909	2.818182	0.000000	2.818182	3.181818	4.090909	2.818182	1
2	14.594444	10.055556	74.977778	11.922222	20.255556	11.516667	26.088889	4.655556	1.333333	0.555556	3.111111	0.111111	2.333333	3.000000	3.944444	9.000000	2

Se observa en la Tabla 28 generada por el modelo de validación, el promedio de los valores reales de las variables en cada clúster.

Tabla 29: Validación de clustering para Adultos mayores parte 2

	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis	COB	OCR	CSR	Covid	Cluster
0	1.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.219739	0.000000	1.000000	0.000000	0.000000	1.000000	1.00	0.166667	0.000000	0.000000	0
1	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	0.000000	0.061224	0.591549	0.000000	0.25	1.000000	1.000000	0.230189	1
2	0.375311	0.430699	0.927389	0.505249	0.532655	0.333133	0.000000	0.765539	0.459459	1.000000	1.000000	0.916667	0.00	0.000000	0.844086	1.000000	2

Finalmente en la Tabla 29 se puede observar que en el clúster 0 se encuentran los índices más altos de temperatura, rocío, temperatura_min, temperatura_max, IRA_Alta bronquitis y COB. En el clúster 1 se encuentran los índices más altos de las variables contaminantes (MP2.5, MP10 y SO_2), OCR y CSR. Mientras que en clúster 2 se observan los índices más altos de influenza, neumonía y covid. Todos estos índices confirman los resultados dados por el PCA y el mapa de calor.
En resumen se tiene lo siguiente:

Tabla 30: Comparación de resultados para Adultos mayores

Variables	Comparación				
	PCA		Validación de clustering		
	Componente 1	Componente 2	Clúster 0	Clúster 1	Clúster 2
Temperatura	●		●		
Rocío	●		●		
Humedad				●	
Temperatura_min	●		●		
Temperatura_max	●		●		
MP2.5				●	
MP10				●	
SO2				●	
IRA_Alta	●		●		
Influenza					●
Neumonía					●
Bronquitis	●	●	●		
COB	●		●		
OCR				●	
CSR		●		●	
Covid					●

● Mayor representatividad de la variable.

En este grupo se observa el mismo patrón para bronquitis con las variables ambientales, agrupadas todas en el clúster 0, similar a lo representado por la componente 1 del PCA. Para este grupo de edad nuevamente el algoritmo agrupa a neumonía y covid en el mismo clúster, pero el PCA no muestra una relevancia significativa de estas variables.

Para observar de otra forma todos los resultados del modelo de validación de clustering ver Anexo C.

Capítulo V
Conclusiones

En este trabajo de apoyo a la investigación presentado como trabajo de titulación se presentó la utilización de un algoritmo de clustering para el análisis del comportamiento de las enfermedades respiratorias en la comuna de Copiapó. Se trabajó con algoritmo K-Means resultando eficiente para pruebas con bajo volumen de datos y con alto volumen de datos. Es posible utilizar algoritmos para la exploración del comportamiento de enfermedades respiratorias utilizando algoritmos de clustering.

En cuanto a la metodología utilizada, se vuelve imprescindible realizar un buen trabajo en la etapa 4, es la que conlleva más tiempo pero con un mal desempeño en la implementación del algoritmo se vuelve necesario volver a revisar la etapa para seguir avanzando a las siguientes.

Respecto a los resultados obtenidos, se consideran interesantes, en el contexto de pandemia, debido a la relación encontrada entre COVID-19 y neumonía. Además de que el rango de edad más afectado por enfermedades respiratorias en la comuna de Copiapó se encontró en el grupo de jóvenes y adultos. También se observa que en todos los grupos existe una relevancia de la variable "otra causa respiratoria" (otro diagnóstico no especificado por las bases de datos). Además de la relevancia de causas del sistema respiratorio, es decir, que un caso de hospitalización está relacionado con enfermedades como la neumonía o bronquitis. Además se identificó mediante el PCA una alta relación entre material particulado de 2.5 micras y las enfermedades respiratorias, ya que, estuvo presente en todos los grupos de edad, no así, el material particulado de 10 micras que no evidenció una gran relevancia en los análisis de resultados para ninguno de los grupos etarios. Además las enfermedades respiratorias que mayor cantidad de relaciones mostraron fueron Neumonía y Bronquitis, cada una relacionada a 3 grupos de edad distintos pero en ningún caso relacionadas entre sí.

Por otra parte, la validación del clustering ayudó a comprender de mejor manera los resultados obtenidos del PCA y del mapa de calor. Para todos los grupos etarios se encontró una validación aceptable por parte del modelo utilizado. En los resultados de este modelo se confirma nuevamente la relación que el algoritmo de clustering encontró entre neumonía y covid, esta relación estuvo presente en todos los grupos de edad que presentaban casos de covid a excepción de Infantes que podría ser atribuido a la escasa cantidad de casos de covid que presentó este grupo. También se considera interesante un patrón presentado en todos los grupos etarios, la relación entre bronquitis y las variables del medioambiente; temperatura, temperatura mínima, temperatura máxima y rocío, para todos los grupos de edad el algoritmo de clustering agrupó siempre de la misma forma a estas variables.

Como trabajos a futuro se piensa ampliar el periodo de los datos, además de aplicar otros algoritmos de *clustering,* como el algoritmo *K-Medoids.*

Referencias

Foro de las Sociedades Respiratorias Internacionales. El impacto mundial de la Enfermedad Respiratoria. 2da ed. México: Asociación Latinoamericana de Tórax; 2017. Disponible en: https://www.who.int/gard/publications/The_Global_Impact_of_Respiratory_Disease_ES.pdf

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.

E. Menasalvas and C. Gonzalo Alejandro Rodríguez-González, "BIG DATA EN SALUD: RETOS Y OPORTUNIDADES," Econ. Ind., vol. 405, pp. 87–97, 2017.

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].

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.

Dowell SF, Whitney CG, Wright C, Rose CE Jr, Schuchat A. Seasonal patterns of invasive pneumococcal disease. Emerg Infect Dis. 2003;9:574-9. http://dx.doi.org/10.3201/eid0905.020556

Omer SB, Sutanto A, Sarwo H, Linehan M, Djelantik IG, Mercer D, et al. Climatic, temporal, and geographic characteristics of respiratory syncytial virus disease in a tropical island population.Epidemiol Infect.2008;136:1319-27
http://dx.doi.org/10.1017/S0950268807000015

Organización Mundial de la Salud (WHO). (2021, 22 septiembre). Calidad del aire ambiente (exterior). https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

IEA 2016 World Energy Outlook Special Report https://www.iea.org/reports/world-energy-outlook-2016

WHO (2016). World Health Organization. Health risk assessment of air pollution – general principles. Regional Office for Europe. Copenhagen. 29p.

Legarreta, A., Corral, A., Delgado, M., Torres, J., Flores, J., y López, F., Material Particulado y Metales Pesados en Aire en Ciudades Mexicanas, Cultura Científica y Tecnológica, No 56(12) (2015)

Cristina Linares Gil, del Centro Nacional de Epidemiología, y Julio Díaz Jiménez, de la Escuela Nacional de Sanidad, ambos del Instituto de Salud Carlos III. Una versión de este artículo ha sido publicado previamente en: C. Linares, J. Díaz: "Las PM2,5 y su impacto sobre la salud. El caso de la ciudad de Madrid". Ecosostenible. 2008;35:32-37.

Vidal LMJ, Carnota LO, Rodríguez DA. Tecnologías e innovaciones disruptivas. Revista Cubana de Educación Médica Superior. 2019;33(1):1-13.

Marchán E, Salcedo J, Aza T, Figuera L, Martínez de Pisón F, G. P. (2011). Reglas de asociación para determinar factores de riesgo epidemiológico de transmisión de la enfermedad de Chagas. Revista de Ciencia E Ingeniería, (January), 55–60.

Pérez, M. (2014). MINERÍA DE DATOS A TRAVÉS DE EJEMPLOS

Molina, J., & García, J. (2008). Técnicas de Minería de Datos basadas en Aprendizaje Automático. Técnicas de Análisis de Datos, 96–266.

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

RENDÓN, Eréndira, ZEPEDA, Ricardo, BARRUETA, Elizabeth y ITZEL-MARÍA, Abundez. El algoritmo de agrupamiento K-Modas: Un caso de estudio. Revista de Tecnología e Innovación 2015, 2-5: 929-941

Oviedo, E., Oviedo, A., & Vélez, G. (2015). Minería de datos: aportes y tendencias en el servicio de salud de ciudades inteligentes. Revista Politécnica, 11(20), 111–120.

Blanco-Herminda Sanz, E.J. (2016). Prototipo de clustering orientado motor de búsqueda. (Bachelor's thesis, Universitat Politécnica de Catalunya).

Garre, Miguel; Cuadrado, Juan José; Sicilia, Miguel A.; Rodríguez, Daniel; Rejas, Ricardo Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software REICIS. Revista Española de Innovación, Calidad e Ingeniería del Software, vol. 3, núm. 1, abril, 2007, pp. 6-22

Jiawei Hand and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd edition, 2006.

Larrañaga, P., Inza, I., y Moujahid, A. (2012). Tema 14. Clustering [archivo PDF]. Recuperado de http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t14clustering.pdf

Cristina García Cambronero, Irene Gómez Moreno (2006), Algoritmo de Aprendizaje: KNN y Kmeans.

J. Pérez, M. F. Henriques, R. Pazos, L. Cruz, G. Reyes, J. Salinas, A. Mexicano. (January de 2007). Mejora al algoritmo de agrupamiento K-means mediante un nuevo criterio de convergencia y su aplicación a bases de datos poblacionales de cáncer. pág. 7.

Gonzáles Argote, Heynz Roberth, & Ticona Gonzáles, Ulises Amaru. (2019). Clustering, mediterraneidad y comercio internacional: aplicación empírica de los algoritmos Partitioning Around Medoids y K-means. Revista Latinoamericana de Desarrollo Económico, (32), 95-129.

Jain, Brijnesh & Obermayer, Klaus. (2010). Elkan's k-Means Algorithm for Graphs. 22-32. 10.1007/978-3-642-16773-7_2.

TERRADEZ GURREA, Manuel. Análisis de Componentes Principales. Cataluña: Universidad de Oberta. 2002. p. 11

Walpole Ronald E., Myers Raymond H., Myers Sharon L. y Ye Keying "Probabilidad y estadística para Ingeniería y ciencias". Octava Edición. Pearson Education. 2007

Chen S, Wu S. Deep learning for identifying environmental risk factors of acute respiratory diseases in Beijing, China: implications for population with different age and gender. Int J Environ Health Res. 2020 Aug;30(4):435-446. doi: 10.1080/09603123.2019.1597836. Epub 2019 Mar 31. PMID: 30929473.

Aguilar, J.S., & Gutiérrez, E. (2017). Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá, Colombia.

S. Alejandro and R. Perez, "DESARROLLO Y EVALUACIóN DE ALGORITMOS DE DATA MINING PARA LA PREDICCIÓN DEL RIESGO DE CRISIS EN PACIENTES AMBULATORIOS DE UN HOSPITAL PEDIÁTRICO," 2017.

Ralph Kimball, Joe Caserta, (2004). The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data

Squeo, Francisco & Arancio, Gina & Gutierrez, Julio. (2008). Libro Rojo de la Flora Nativa y de los Sitios Prioritarios para su Conservación: Región de Atacama.

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

Agrawal, R. y Srikant, R. (1994). Fast Algorithms for Mining Association Rules. vldb Conference, Santiago de Chile.

Azevedo, Ana; Zantos, Manuel Filipe. KDD, SEMMA and CRISP-DM: a parallel overview. 2008.

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y AlvaradoPérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia. doi: http://dx.doi.org/10.16925/9789587600490

LÓPEZ, B. 2011. Limpieza de Datos: Reemplazo de valores ausentes y Estandarización.
A. Silberschatz, et al., Fundamentos de Base de Datos, 4 ed., 2001.

Bellinger, C., Mohomed Jabbar, M., Zaïane, O. et al. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health 17, 907 (2017). https://doi.org/10.1186/s12889-017-4914-3

Ricardo, Catherine M. (2009), Bases de datos. McGraw-Hill. URI: http://up-rid2.up.ac.pa:8080/xmlui/handle/123456789/1354

Amat Rodrigo, J, (Septiembre 2017). Clustering y heatmaps: aprendizaje no supervisado. Recuperado de https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps

Hernández Cedon, J. A. (2015) MODELO DE MINERÍA DE DATOS PARA IDENTIFICACIÓN DE PATRONES QUE INFLUYEN EN EL APROVECHAMIENTO ACADÉMICO [Tesis de Maestría, Instituto Tecnológico de La Paz]. http://posgrado.lapaz.tecnm.mx/uploads/archivos/TesisHdzCedano.pdf

Aguilar-Aldana, J. S. (2017). Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá, Colombia. Facultad de Ingeniería.

Viera, Angel Freddy Godoy. (2017). Técnicas de aprendizaje de máquina utilizadas para la minería de texto. Investigación bibliotecológica, 31(71), 103-126. https://doi.org/10.22201/iibi.0187358xp.2017.71.57812

Pascual, D., Pla, F., & Sánchez, S. (2007). Algoritmos de agrupamiento. Métodos informáticos avanzados.

Prado, Pedro & Monteiro, António. (2008). Pattern recognition algorithms. 5.

Preeti Arora, Deepali, Shipra Varshney, (2016) Analysis of K-Means and K-Medoids Algorithm For Big Data, Procedia Computer Science, Volume 78, 2016, Pages 507-512, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.02.095. (https://www.sciencedirect.com/science/article/pii/S1877050916000971)

Balabantaray, R. C., Sarma, C., & Jha, M. (2015). Document clustering using k-means and k-medoids. arXiv preprint arXiv:1502.07938.

Challenger-Pérez, I., Díaz-Ricardo, Y., & Becerra-García, R. A. (2014). El lenguaje de programación Python. Ciencias Holguín, 20(2), 1-13.

Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008). https://doi.org/10.1186/1471-2105-9-559

Amat Rodrigo, J, (Diciembre 2020). PCA con Python. Recuperado de https://www.cienciadedatos.net/documentos/py19-pca-python.html

Leland Wilkinson & Michael Friendly (2009) The History of the Cluster Heat Map, The American Statistician, 63:2, 179-184, DOI: 10.1198/tas.2009.0033

E. L. Guzmán, "Métricas para la validación de Clustering," Elizabeth León Guzmán, 2016. https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf

Anexo A:

CIE-10 Capítulo X Enfermedades del aparato respiratorio (J00-J99)

Cap.10 ENFERMEDADES DEL APARATO RESPIRATORIO (J00-J99)

Nota:
- Cuando un problema respiratorio se describe como algo que ocurre en más de un sitio y no está específicamente representado, deberá clasificarse en la localización anatómica más baja (por ejemplo, traqueobronquitis a bronquitis se clasifica bajo J40).
Utilice código adicional, si procede, para identificar:
- dependencia del tabaco (F17.-)
- exposición a humo ambiental de tabaco (Z77.22)
- exposición al humo de tabaco en período perinatal (P96.81)
- exposición ocupacional al humo de tabaco ambiental (Z57.31)
- historia personal de dependencia del tabaco (Z87.891)
- tabaquismo activo (Z72.0)
Excluye 2:
- ciertas afecciones originadas en período perinatal (P04-P96)
- ciertas enfermedades infecciosas y parasitarias (A00-B99)
- complicaciones del embarazo, parto y puerperio (O00-O9A)
- enfermedades endocrinas, nutricionales y metabólicas (E00-E88)
- inhalación de humo (T59.81-)
- lesiones, envenenamientos y otras consecuencias de causas externas (S00-T88)
- malformaciones congénitas, deformaciones y anomalías cromosómicas (Q00-Q99)
- neoplasias (C00-D48)
- síntomas, signos y resultados anormales de pruebas complementarias, no clasificables bajo otro concepto (R00-R94)

Este capítulo contiene las siguientes secciones:
- J00-J06 Infecciones agudas del tracto respiratorio superior
- J09-J18 Gripe y neumonía
- J20-J22 Otras infecciones agudas del tracto respiratorio inferior
- J30-J39 Otras enfermedades del tracto respiratorio superior
- J40-J47 Enfermedades crónicas del tracto respiratorio inferior
- J60-J70 Enfermedades pulmonares debidas a agentes externos
- J80-J84 Otras enfermedades respiratorias que afectan principalmente al intersticio
- J85-J86 Trastornos supurativos y necróticos de vías respiratorias inferiores
- J90-J94 Otras enfermedades de la pleura
- J95 Complicaciones y trastornos intraoperatorios y posprocedimiento del aparato respiratorio, no clasificados bajo otro concepto
- J96-J99 Otras enfermedades del aparato respiratorio

J00-J06 INFECCIONES AGUDAS DEL TRACTO RESPIRATORIO SUPERIOR (J00-J06)
Excluye 1:
- enfermedad pulmonar obstructiva crónica con infección aguda del tracto respiratorio inferior (J44.0)

J09-J18 GRIPE Y NEUMONÍA (J09-J18)
Excluye 2:
- neumonía alérgica o eosinofílica (J82)
- neumonía asociada a ventilación mecánica (J95.851)
- neumonía congénita (P23.9)
- neumonía lipoidea (J69.1)
- neumonía neonatal por aspiración (P24.-)
- neumonía por aspiración de meconio (P24.01)
- neumonía por aspiración de sólidos y líquidos (J69.-)
- neumonía por aspiración NEOM (J69.0)
- neumonía reumática (I00)
J20-J22 OTRAS INFECCIONES AGUDAS DEL TRACTO RESPIRATORIO INFERIOR (J20-J22)
Excluye 2:
- enfermedad pulmonar obstructiva crónica con infección aguda del tracto respiratorio inferior (J44.0)
J30-J39 OTRAS ENFERMEDADES DEL TRACTO RESPIRATORIO SUPERIOR (J30-J39)
J40-J47 ENFERMEDADES CRÓNICAS DEL TRACTO RESPIRATORIO INFERIOR (J40-J47)
Excluye 1:
- bronquitis debida a productos químicos, gases, humos y vapores (J68.0)
Excluye 2:
- fibrosis quística (E84.-)
J60-J70 ENFERMEDADES PULMONARES POR AGENTES EXTERNOS (J60-J70)
Excluye 2:
- asma (J45.-)
- neoplasia maligna de bronquio y pulmón (C34.-)
J80-J84 OTRAS ENFERMEDADES RESPIRATORIAS QUE AFECTAN PRINCIPALMENTE AL INTERSTICIO (J80-J84)
J85-J86 ENFERMEDADES SUPURATIVAS Y NECRÓTICAS DEL TRACTO RESPIRATORIO INFERIOR (J85-J86)
J90-J94 OTRAS ENFERMEDADES DE LA PLEURA (J90-J94)
J95-J95 COMPLICACIONES Y TRASTORNOS INTRAOPERATORIOS Y POSPROCEDIMIENTO DE APARATO RESPIRATORIO, NO CLASIFICADOS BAJO OTRO CONCEPTO (J95)
J96-J99 OTRAS ENFERMEDADES DEL APARATO RESPIRATORIO (J96-J99)

Anexo B:

Diccionario de datos

Diccionario de datos

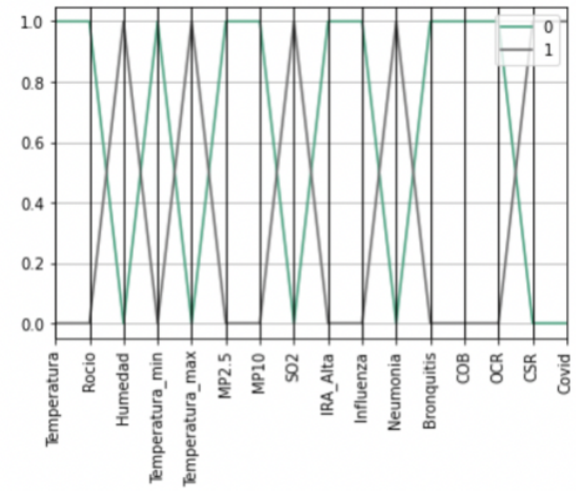
VARIABLES AMBIENTALES				
#	**Nombre de variable**	**Tipo**	**Definición breve**	**Unidad de medida**
0	Temperatura	Flotante	Magnitud física que expresa el grado o nivel de calor del ambiente.	[C°]
1	Rocio	Flotante	Temperatura a la cual se debe enfriar el aire para que el vapor de agua se condense.	[C°]
2	Humedad	Flotante	Agua de que está impregnado un cuerpo o que, vaporizada, se mezcla con el aire.	%
3	Temperatura_min	Flotante	Temperatura mínima que podría existir en un periodo de tiempo.	[C°]
4	Temperatura_max	Flotante	Temperatura máxima que podría existir en un periodo de tiempo.	[C°]
VARIABLES CONTAMINANTES				
#	**Nombre de variable**	**Tipo**	**Definición breve**	**Unidad de medida**
5	MP2.5	Flotante	Material particulado con diámetro aerodinámico menor o igual a 2,5 micrones.	[ug/m³]
6	MP10	Flotante	Material particulado con diámetro aerodinámico menor o igual a 10 micrones.	[µg/m³N]
7	SO2	Flotante	Gas producido por quema de combustibles.	[µg/m³N]
ENFERMEDADES RESPIRATORIAS				
#	**Nombre de variable**	**Tipo**	**Definición breve**	**Código CIE-10**
8	IRA_Alta	Entero	Infecciones agudas de las vías respiratorias superiores (atención urgencias).	(J00-J06)-U
9	Influenza	Entero	Enfermedad respiratoria contagiosa provocada por virus de la influenza (atención urgencias).	(J09-J11)-U
10	Neumonia	Entero	Influenza [gripe] y neumonía (atención urgencias).	(J12-J18)-U
11	Bronquitis_bronquiolitis	Entero	Otras infecciones agudas de las vías respiratorias inferiores (atención urgencias).	(J20-J21)-U
12	Crisis_obstructiva_bronquial	Entero	Enfermedades crónicas de las vías respiratorias inferiores (atención urgencias).	(J40-J46)-U
13	Otra_causa_respiratoria	Entero	Otras enfermedades respiratorias (atención urgencias).	(J22,J30,J39,J47,J60-J98)-U
14	COVID19_Sospechoso_u	Entero	Sospechoso de COVID19, virus no identificado (atención urgencias).	(U07.2)-U
15	COVID19_Confirmado_u	Entero	Confirmado de COVID19, virus identificado (atención urgencias).	(U07.1)-U
16	CAUSAS_SISTEMA_RESPIRATORIO	Entero	Enfermedades respiratorias general (hospitalización).	-H
17	COVID19_SOSPECHOSO_H	Entero	Sospechoso de COVID19, virus no identificado (hospitalización).	(U07.2)-H
18	COVID19_CONFIRMADO_H	Entero	Confirmado de COVID19, virus identificado (hospitalización).	(U07.1)-H

Anexo C:

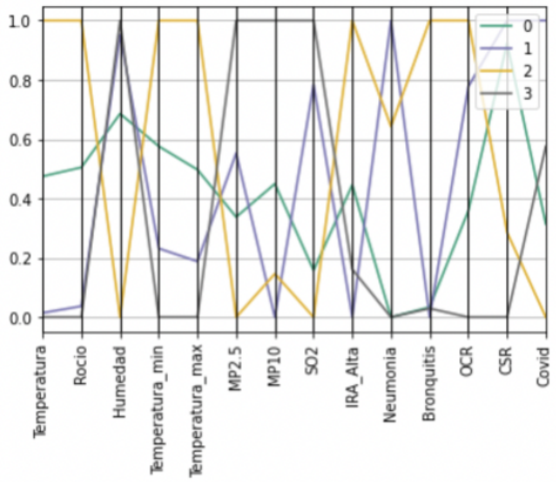
Validación de clustering en gráfico de coordenadas

Validación de clustering
Gráfico de coordenadas

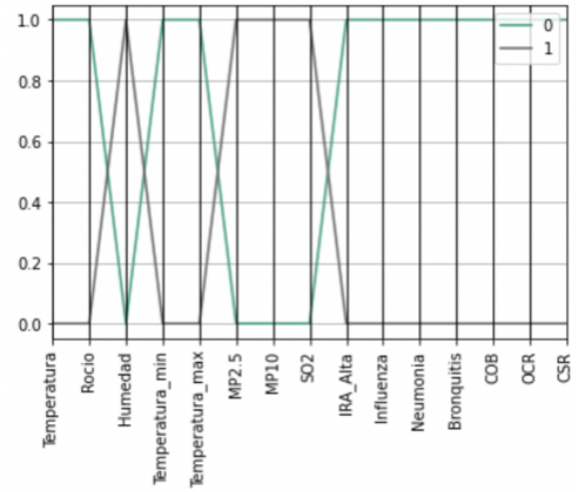
General



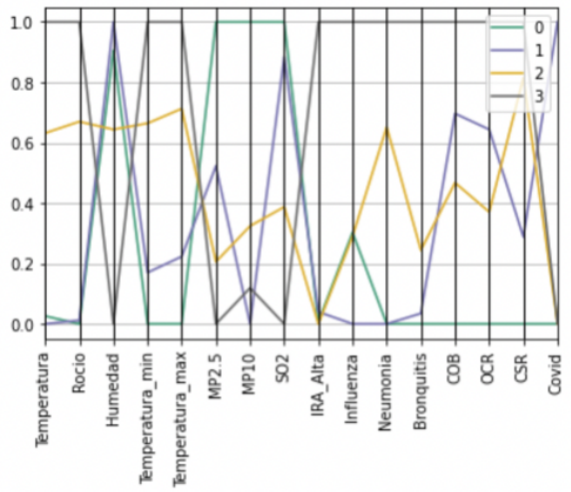
Menores de 1 año



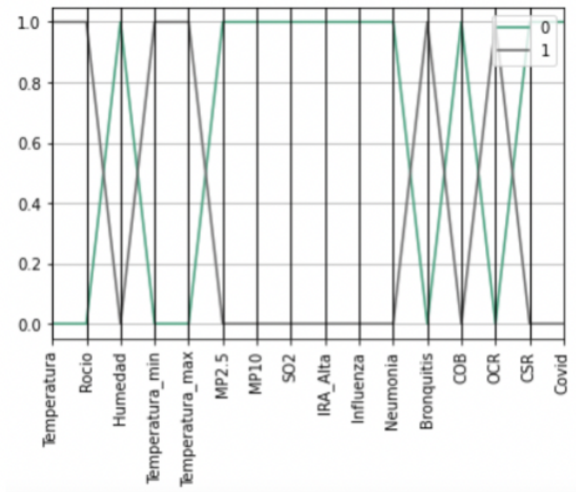
Niños de 1 a 4 años



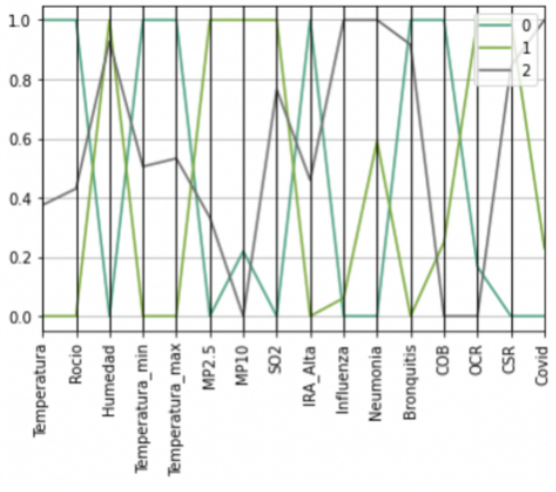
Niños de 5 a 14 años



Adultos de 15 a 64 años



Adultos de 65 o más años



Anexo D:

Estructura de datos

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Fecha	Temperatura	Rocio	Humedad	Temperatura_min	Temperatura_max	MP2.5	MP10	SO2	IRA_Alta	Influenza	Neumonia	Bronquitis_bronquiolitis	Crisis_obstructiva_bronquial	Otra_causa_respiratoria	COVID19_Sospechoso_u	COVID19_Confirmado_u	CAUSAS_SISTEMA_RESPIRATORIO	COVID19_SOSPECHOSO_H	COVID19_CONFIRMADO_H
2	1	17.9	12.8	72.3	15.4	22.7	7.7	25.4	2.6	63	3	4	24	10	7	0	0	6	0	0
3	2	19.3	13.1	68.4	15.5	24.4	7.3	27.4	2.6	60	1	9	17	13	9	0	0	6	0	0
4	3	18.8	13.4	71.4	15.7	24.6	6.6	24.7	2.8	46	1	4	21	18	10	0	0	4	0	0
5	4	21.7	14.1	62.8	17.9	27	5.7	27.6	1.5	61	0	1	11	13	9	0	0	3	0	0
6	5	20.3	14.9	70.7	18	24.1	6.4	27.3	0.4	48	0	5	16	25	9	0	0	8	0	0
7	6	19.1	13.5	70.8	16.5	23.3	6.7	28.3	0.7	59	0	7	23	11	10	0	0	10	0	0
8	7	19	13.4	70	17	24.1	9.4	30.4	0.7	60	0	4	11	6	11	0	0	3	0	0
9	8	17.9	12.8	66.7	15.8	23.9	7	26.9	1.1	58	1	7	12	8	7	0	0	5	0	0
10	9	18.5	13.5	73	16	22.9	6.4	35.3	0.9	62	2	6	17	10	7	0	0	6	0	0
11	10	18.3	13.4	73.6	15.9	22.7	12.7	27.7	0.4	57	5	3	9	11	9	0	0	5	0	0
12	11	17.7	12.8	67.4	16.1	23.1	12.1	35.9	1.4	87	3	5	15	18	13	1	0	9	1	0
13	12	17.5	13.2	76.2	15.7	21.7	8.7	24.7	1.4	161	12	9	17	22	6	26	0	4	3	0
14	13	17.5	12.2	72	14.5	22.6	7.6	27.7	1.4	121	0	4	6	23	6	36	0	6	6	0
15	14	17.2	13.1	77.2	15.1	21.9	9.9	29.7	1.3	49	0	3	11	7	22	37	1	1	5	1
16	15	18	12.6	71.5	14.7	21.6	9	28.7	2.8	49	15	8	3	12	4	41	1	8	5	1
17	16	16.7	11.3	71.3	13.5	19	12.3	36.9	1.3	23	7	0	2	14	16	19	0	0	0	0
18	17	17.1	11.9	71.6	15.2	20.2	11.6	29	2.7	41	11	3	2	14	8	20	0	7	2	0
19	18	15.5	11.1	75.9	12.4	19.2	11	34.4	2.7	39	13	2	1	14	3	69	3	3	0	1
20	19	15.5	11.2	75.9	13.3	18.5	14.6	39.4	3.7	47	25	5	3	11	4	46	2	5	1	1
21	20	14.4	10.5	77.8	11.6	19.5	16.1	50.3	3.4	52	17	6	2	9	9	46	2	11	0	2
22	21	14.8	9.7	71.8	12.7	18	15.8	44	3.4	47	16	6	0	5	3	34	1	6	3	1
23	22	14.8	9.5	71.4	11.2	18.5	16	44.7	2.8	65	13	3	2	8	6	35	4	4	1	1
24	23	14.2	8.3	70.4	11.1	17.9	11.2	41.1	1.3	72	18	6	2	5	9	36	5	10	3	3
25	24	12.9	8.5	75.6	8.9	19.5	20	72	5.5	67	15	9	0	8	7	30	13	8	1	7
26	25	13.5	9.3	76.4	10.9	17.6	16.8	50	4.5	70	24	7	2	9	9	75	15	9	0	7
27	26	12.8	9.2	79.4	9.7	17.6	16.4	51.6	8.3	76	20	6	1	18	5	121	11	11	3	5
28	27	12.9	8.5	76	8.5	17.6	17	53.9	8	64	17	7	0	8	2	56	11	9	3	7
29	28	12.5	7.4	71.3	8.2	17.3	15.4	53.9	4.4	81	13	6	3	8	6	62	11	4	3	4
30	29	11.5	7.9	79.3	8.1	15.6	17.6	55.9	7.3	62	16	2	2	11	8	62	10	6	2	3
31	30	11.8	5.5	69.6	6	16	13.4	51.6	6.4	78	24	4	1	19	9	190	15	8	7	6
32	31	11.9	7.9	78.3	7	17.7	17.6	65.3	5.3	61	11	12	0	13	11	145	33	13	2	16
33	32	12.4	9.3	81.5	9.9	18.4	15.4	25.4	8.3	80	17	5	2	13	5	124	46	8	7	22
34	33	12	8	77.3	10.3	16.6	14.7	28.3	6	58	26	7	0	16	4	90	39	8	4	18
35	34	11.8	6.9	72.5	9.3	17.8	17.7	30.4	8.7	57	8	10	0	4	12	66	20	14	5	11
36	35	12.7	6.9	69.3	9	18	9.1	23.7	3.7	28	13	7	2	10	5	52	11	6	2	4
37	36	12.6	7.5	72.1	9.8	17.9	16	25.4	4.9	25	10	3	1	8	5	40	12	3	4	6
38	37	12.6	8.6	77	10.1	18.6	12.4	24.7	3.8	28	23	5	1	11	8	36	9	5	2	4
39	38	12.8	9	77.9	10.8	18.4	15.4	23.4	4.5	22	4	4	1	4	2	38	14	5	1	8
40	39	13.3	9	76	10.9	19.1	12.3	22.7	4.2	49	6	4	1	8	9	55	4	10	0	2
41	40	13.1	8.3	73.2	10.9	18.9	9.1	17.1	3.4	40	10	3	1	3	8	84	4	5	2	3
42	41	12.6	8.7	77.6	9.5	19	15.6	33.3	5.7	43	2	7	2	8	10	46	4	10	4	4
43	42	13.8	9.5	75.6	11.5	20.6	8.4	17.7	3.8	61	6	13	1	7	6	32	8	15	0	3
44	43	13.5	9.9	79.1	11.4	20.8	9.7	24.4	4.6	62	7	6	1	7	12	40	7	6	1	3
45	44	14.1	9.6	75.3	11.5	20.1	9.9	35.4	4.5	68	4	3	2	12	5	24	5	2	0	4
46	45	14.5	9.9	75	12.2	20.3	9.9	32.3	3.5	53	0	4	9	13	6	20	0	4	1	0
47	46	14.8	10.1	74.3	12.7	23.4	6	25.4	2.8	70	0	5	4	17	9	28	2	8	0	1
48	47	14.8	8.8	67.7	12.1	23.7	7.7	34.1	3.7	68	0	5	4	10	9	5	1	8	0	0
49	48	15.5	10.1	71	12.3	23.2	10	44.1	2.9	56	1	5	4	19	7	7	0	7	0	0
50	49	15.1	11.1	76.6	12.6	22.7	8.6	38.6	3	70	0	4	6	11	8	27	0	3	1	0
51	50	15.2	11.7	80.4	12.8	23.3	8.1	36.9	2.6	40	0	3	1	15	11	23	1	7	0	0
52	51	15.9	12.2	79.4	13.8	22.6	9	37.2	3.2	32	0	3	3	10	2	32	0	5	1	0
53	52	17	11.9	73.2	14	24.5	7.6	33.6	2.8	41	0	7	3	8	10	23	5	5	0	2
54	53	18.8	12.7	68.2	15.2	23.4	9.1	34.9	3.1	39	2	7	3	7	4	75	9	8	0	6
55	54	17.7	12.8	73.6	14.4	23.8	8.1	31.1	2.6	51	2	7	6	15	12	56	15	13	2	6
56	55	16.4	11.9	75.5	13.6	22.5	9.1	39.4	2.7	72	5	8	1	11	5	60	16	10	2	11
57	56	17.7	12.7	72.8	14.2	23	10	31.3	3.3	60	5	8	6	6	4	56	13	8	0	6
58	57	18.4	13.4	73.5	16.1	23.2	10.4	31.1	4.4	55	0	11	2	7	9	52	29	11	0	18
59	58	17.5	13.2	76.7	15.2	22.8	8.9	29.6	3.4	37	0	7	3	5	8	40	23	4	1	11
60	59	18.4	12.7	70.6	14.9	23.6	6.7	23.1	3.1	34	1	9	1	9	2	34	15	6	0	11
61	60	18.5	13	71.7	14.4	24.4	8.3	27.6	3.5	44	1	7	1	11	4	45	17	3	4	10
62	61	17.9	13	73.7	14.7	22.7	8.6	30.6	3.9	41	0	7	4	4	7	34	16	10	1	11
63	62	18.8	14.1	74.3	15.9	23.8	8.6	31.1	3.9	47	0	9	0	6	7	44	23	5	0	11