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

ANÁLISIS LONGITUDINAL SOBRE TRAYECTORIAS SINTOMÁTICAS DE PACIENTES COVID-19

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

ALEJANDRO MATÍAS BULL CUEVAS

PROFESOR GUÍA:
ÁNGEL JIMÉNEZ-MOLINA

MIEMBROS DE LA COMISIÓN:
JUAN PABLO ROMERO GODOY
MAURICIO SOTO DURÁN

Este trabajo ha sido financiado por el proyecto COVID 0251 de la Agencia Nacional de Investigación y Desarrollo.

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL
INDUSTRIAL
POR: ALEJANDRO MATÍAS BULL CUEVAS
FECHA: 2021
PROF. GUÍA: ÁNGEL JIMÉNEZ-MOLINA

ANÁLISIS LONGITUDINAL SOBRE TRAYECTORIAS SINTOMÁTICAS DE PACIENTES COVID-19

El COVID-19 ha sido el causante de la pandemia del 2020, que a su vez ha traído consigo principalmente una crisis sanitaria y económica.

A través de un fondo de la Agencia Nacional de Investigación y Desarrollo, es que profesionales de la Facultad de Ciencias Físicas y Matemáticas están implementando un Sistema integrado de información para el seguimiento domiciliario de pacientes COVID-19, en el Servicio de Salud Metropolitano Sur Oriente. El SSMSO cuenta con una plataforma de registro y seguimiento de pacientes COVID-19, no obstante, se encuentran limitados para realizar el análisis de dichos datos, comprometiendo la efectividad de enfrentar esta pandemia.

Dentro de este contexto se enmarca el siguiente trabajo, el cual busca aumentar la efectividad del seguimiento a los pacientes COVID-19 a través de analítica sintomática que potencie la toma de decisiones.

Se utiliza la metodología de Knowledge Discovery in Databases con los datos de la plataforma COVID-19 del SSMSO. En la fase de Data Mining se emplean algoritmos de aprendizaje no supervisado para encontrar las trayectorias sintomáticas que presentaron los pacientes. Luego a través de algoritmos de lenguaje supervisado se busca predecir la trayectoria sintomática que tendrán nuevos pacientes en el futuro.

Como principales resultados de la investigación se encontraron 7 trayectorias sintomáticas, donde dos están caracterizadas por anosmia y disgeusia, dos por la combinación de cefalea, mialgias y tos seca, uno por cefalea y mialgias, uno por tos seca y el último por solo cefalea de no más de una semana. Tras el análisis, queda en evidencia que las trayectorias caracterizadas por anosmia y disgeusia presentan un menor riesgo para los distintos grupos etarios en comparación a las otras trayectorias sintomáticas. Adicionalmente, las trayectorias más riesgosas se caracterizan por presentar disnea, compromiso del estado general y decaimiento aproximadamente una semana después del inicio de los síntomas. La herramienta de predicción obtiene buenos resultados para predecir los síntomas que va a tener un determinado paciente en el futuro, a pesar de no ser tan efectiva con la duración exacta.

La principal conclusión del trabajo es que al conocer como fueron las trayectorias sintomáticas de distintos pacientes en el pasado, se puede predecir como será la evolución sintomática que tendrán los nuevos pacientes en el futuro, lo cual entrega un insumo al equipo clínico que permite anticiparse a las evoluciones que tendrán las nuevas personas afectadas por la pandemia.

Agradecimientos

Incontables son las personas, experiencias y momentos que me han ayudado a llegar hasta este punto y a ser la persona que hoy escribe esta memoria. En primer lugar, agradecer profundamente a mi familia y en especial a mis padres, Soledad y Ricardo. Gracias a su apoyo incondicional, tanto en las buenas como en las malas, han permitido que pueda salir adelante una y otra vez, ayudándome a convertirme en una mejor persona y dándome las facilidades para obtener un título universitario.

Gracias de todo corazón a mis amistades que me han acompañado desde tiempos inmemoriales, mis queridos *Tuleks*, por tantas risas, consejos y buenas vacaciones que nos hemos mandado. No hay palabras para expresar todo lo que hemos vivido y lo mucho que he aprendido junto a ustedes, gracias por estar ahí en los distintos desafíos que me ha puesto la vida, ¡y vamos por muchos más!

Agradezco también profundamente a las distintas amistades que logré forjar a lo largo de mi paso por la universidad, desde los que hice siendo un mechón de la 5 hasta los que hice en el Departamento de Industrias. En primer lugar, destaco a mis cabros que se fueron a Civil, que buenos *mochileos* que nos mandamos que hacían que cada vez que nos viéramos en la U recordáramos y nos riéramos de las distintas anécdotas y experiencias de nuestros viajes. En segundo lugar, destaco a las distintas personas con las que nos quedamos *vacilando* algún viernes en la U y después organizábamos alguna junta o continuábamos la celebración en alguna otra facultad o lugar. En tercer lugar, pero no por ello menos importante, destaco a todas y todos los que alguna vez nos hicimos llamar los *Macacos*. Gracias por tanta buena onda, tanto aguante y tanto apañe, no cualquiera puede contar que se quedó hasta las 12 de la noche en la universidad estudiando o haciendo una tarea/trabajo un día de semana (entiéndase estudiando como *realmente estudiando)* y que, dentro de todo, haya sido una jornada más que agradable. He aprendido mucho de ustedes y espero seguir creciendo a su lado, se les quiere un montón.

Quiero agradecer a la Rama de Natación de la FCFM y a mi Grupo Scout por haber sido pilares fundamentales en lo que fue mi crecimiento a lo largo de todos estos años, me han ayudado a ser más disciplinado y a expandir mi conciencia conociendo distintos tipos de personas y formas de vivir la vida.

Finalmente, agradecer a mi comisión y al equipo detrás del proyecto COVID 0251, siempre pude contar con ustedes al momento de enfrentar las distintas dificultades de este trabajo, dificultades que podían ser tan simples como importar una base de datos en Python, o tan complejas como entender en profundidad el contexto en el que estábamos inmersos y todo lo que estaba en juego.

Alejandro Bull Cuevas

Tabla de contenido

1. Introducción

Para una mejor comprensión de este documento primero se describe en la Sección 1.1 el virus causante de la pandemia, además de las principales consecuencias y desafíos que esto ha traído a los distintos servicios de salud.

Luego, en la Sección 1.2 se describe un proyecto financiado por la ANID con la finalidad de enfrentar de mejor manera a la pandemia, proyecto en el cual se enmarca la investigación desarrollada a lo largo de este documento. En la sección 1.3 se describe el servicio de salud con el cual se trabaja.

Finalmente, en las siguientes secciones de la Introducción, se declaran los objetivos del trabajo, la hipótesis que se busca responder con dichos objetivos, la metodología para cumplirlos y finalmente los resultados finales, que vendrían a ser el fruto de la investigación

1.1. SARS-CoV-2

El año 2020 y el 2021 se han caracterizado por la enfermedad Coronavirus Disease 2019 (COVID-19), causada por el virus Severe Acute Respiratoy Syndrome Coronavirus 2 (SARS-CoV-2). Esta crisis sanitaria ha ocasionado el surgimiento de distintas necesidades, tanto para enfrentar las consecuencias económicas que esta ha producido, como también para enfrentar de mejor manera la enfermedad en sí. Dentro de este contexto, distintas instituciones han buscado cómo aportar para enfrentar de mejor manera esta pandemia y poder resguardar la integridad de los ciudadanos.

1.2. Proyecto ANID

1.2.1. Concurso y adjudicación de fondos

Con el fin de enfrentar la crisis sanitara desde distintas áreas del conocimiento, la Agencia Nacional de Investigación y Desarrollo, ANID, lanzó el 29 de abril del 2020 un concurso para la asignación rápida de recursos para proyectos de investigación sobre el Coronavirus [1]. Precisamente, de 1056 postulaciones 63 proyectos fueron seleccionados y su distribución dentro de las áreas del conocimiento es la siguiente: [2]

Área OCDE	Porcentaje
Medicina	46%
Ciencias Sociales	32%
Ingeniería y Tecnología	12%
Ciencias Naturales	6,3%
Humanidades y Ciencias Agrícolas	3%

Tabla 1.1: Áreas de conocimiento que adjudicaron fondos ANID

La presente investigación se enmarca en el proyecto COVID 0251, financiado por la Agencia Nacional de Investigación y Desarrollo. El proyecto se titula "Sistema integrado de información para el seguimiento domiciliario de pacientes COVID-19 en servicios de salud", cuyo director es Richard Weber, Profesor Titular de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. En este proyecto participaron investigadores del Departamento de Ingeniería Industrial (Susana Mondschein, Ángel Jiménez y Fernando Ordóñez), además de profesionales del Centro de Sistemas Públicos de esta unidad.

El objetivo general de este proyecto es: "aumentar la efectividad del seguimiento a los pacientes COVID-19 confirmados, los casos sospechosos, los casos probables y sus contactos mediante una plataforma analítica escalable que integre diferentes fuentes de información, incluyendo la autodeclarada por pacientes, y entregue analítica relevante para potenciar la toma de decisiones". Es posible encontrar más información sobre el proyecto en el siguiente enlace:

https://www.sistemaspublicos.cl/gproyecto/covid0251/

1.2.2. Áreas del Proyecto

El proyecto se divide en las siguientes tres áreas de trabajo, cada una con sus propios objetivos específicos y resultados esperados:

1. **Auto reporte**: Generar un sistema de seguimiento multicanal de pacientes COVID-19 en cuarentena domiciliaria, mediante auto reporte de síntomas y signos, asistido por canales digitales bidireccionales activos y pasivos.
2. **Integración**: Automatizar la integración de las diferentes fuentes de información disponibles, que sean relevantes para el monitoreo de pacientes COVID-19 en cuarentena domiciliaria y sus contactos.
3. **Analítica**: Crear un sistema de monitoreo prospectivo con la información integrada, que permita mejorar la toma de decisiones con respecto al seguimiento de pacientes COVID-19, aumentando su eficiencia y eficacia. Además, mediante algoritmos de aprendizaje de máquina, se espera que el sistema pueda generar alertas de pacientes críticos, posibles focos de contagio, y otras situaciones relevantes, que permitan orientar la toma de decisiones de los directivos del servicio.

El trabajo desarrollado a lo largo de este documento se centra en la sección de analítica del proyecto, específicamente siendo un análisis descriptivo a través de aprendizaje no supervisado de los síntomas reportados por los pacientes a lo largo de los seguimientos.

1.3. Servicio de Salud Metropolitano Sur Oriente

1.3.1. Información general

El proyecto se desarrolla en el Servicio de Salud Metropolitano Sur Oriente (desde ahora en adelante SSMSO).

El SSMSO es la Red de Atención de Salud Pública que esta inserta en el área sur oriente de Santiago, específicamente en las comunas de Puente Alto, La Florida, San Ramón, La Granja, La Pintana, San José de Maipo y Pirque. En total, cuenta con 35 Centros de Salud (CES, CESFAM, ANCORA), 5 CECOSF, 8 Postas Rurales, 10 COSAM, 1 centro de Especialidades y un Centro de Imágenes. La dotación destinada para estos centros de salud es de un total de 9331 funcionarios para el año 2020, un 11% superior a lo que se tuvo para el 2019. [3,4]

1.3.2. Unidad de Salud Digital

Entre los distintos organismos del SSMSO, se forma una alianza de trabajo con la Unidad de Salud Digital (desde ahora en adelante USD), quienes son los encargados de impulsar la digitalización, el desarrollo y la inclusión de nuevas tecnologías a los establecimientos que conforman la red, con el fin de que los distintos actores de los procesos asistenciales cuenten con la información oportuna y a tiempo para tomar las mejores decisiones.

La USD lanzó el 29 de mayo del 2020 una plataforma web denominada “Plataforma COVID19 SSMSO”, con el fin de apoyar a los establecimientos de la red con el seguimiento de los pacientes. A través de la plataforma es posible que tanto el personal clínico como los trazadores consulten, registren y actualicen los resultados de exámenes, casos confirmados junto a sus respectivos contactos, signos, antecedentes, síntomas y el desenlace de cada paciente. [5]

1.3.3. Situación actual y oportunidad

Si bien el SSMSO actualmente cuenta con una plataforma que permite realizar los seguimientos de manera segura y simple, existen factores que dificultan el aprovechamiento de la información generada como insumo para la toma de decisiones a nivel servicio, tales como el escaso tiempo y recursos que pueden destinar a la labor de analizar los datos o de integrar la información recopilada con los sistemas tecnológicos que actualmente poseen.

Este trabajo se centrará en que, a través de la plataforma, están documentados los distintos síntomas que han presentado los pacientes a través del tiempo, el cual ha quedado únicamente como un registro y no se ha trabajado en extraer el valor que este pudiese tener. Para efectos de esta investigación, una *trayectoria sintomática* se define como los distintos síntomas presentados a lo largo del tiempo en el que un paciente se encuentra sintomático, tomando en consideración el orden de la aparición de los síntomas y la prevalencia de estos.

A raíz de esto, nace la oportunidad de aprovechar los registros sintomáticos para generar conocimiento sobre como son las distintas trayectorias sintomáticas que las personas que sufren de COVID-19 pueden presentar. Conocer a que trayectoria pertenece un paciente le entrega un insumo más tanto a los equipos clínicos para tomar decisiones sobre el tratamiento, como a los equipos de trazabilidad para decidir a quién hacerle un seguimiento más exhaustivo. Además, tiene un valor per se al extender el conocimiento sobre el comportamiento de la enfermedad.

1.4. Objetivos

1.4.1. Objetivo general

Aumentar la efectividad del seguimiento y tratamiento a pacientes COVID-19 del SSMSO a través de analítica sintomática que permita entender en mayor profundidad la evolución de la enfermedad para distintos pacientes, potenciando así la toma de decisiones.

1.4.2. Objetivos específicos

1. Encontrar los distintos tipos de trayectorias sintomáticas y los patrones que estas siguen.
2. Identificar qué tipo de personas, con relación a la edad, género y antecedentes médicos, se asocian a cada trayectoria sintomática.
3. Facilitar los resultados desarrollados a los distintos equipos del SSMSO para que puedan asociar los nuevos pacientes a las trayectorias sintomáticas ya identificadas.

1.5. Hipótesis de investigación

Basado en la literatura existente relacionada a como el virus afecta a los humanos, es que nacen las siguientes tres hipótesis de investigación:

"*Un registro sintomático es insuficiente para describir la enfermedad y se requiere de un análisis longitudinal sobre los distintos síntomas"*

"*Si bien todas las personas reaccionan de manera diferente al SARS-CoV-2, existen grupos de personas similares que reaccionan en relación con su secuencia de síntomas de manera similar*"

"*Se puede predecir como será la evolución sintomática de una persona en base a los síntomas que ya ha tenido"*

1.6. Metodología

En primera instancia se realiza una revisión sistemática del estado del arte, específicamente en lo que son las metodologías utilizadas para encontrar grupos similares y patrones en los datos, además de actuales investigaciones sobre el COVID-19.

Luego, con dichas directrices estudiadas, se procede a aplicar la metodología Knowledge Discovery in Databases (desde ahora KDD), la cual busca generar conocimiento desde los datos, contemplando los siguientes pasos: [5]

1. **Comprensión del dominio**: Caracterizar los objetivos del usuario final, entendiendo el entorno en el cual se inserta la investigación y que ocurrirá con el descubrimiento de conocimiento
2. **Elección y creación del set de datos**: Determinar con que base o con cuales bases de datos se trabajará, verificando que se cuente con todos los atributos necesarios para poder llegar a los resultados buscados.
3. **Preprocesamiento y limpieza**: Mejorar la calidad de los datos seleccionados a través del trato de los valores nulos, ya sea eliminando dichas filas o asignándole valores. Así mismo, eliminando variables que no aporten a la hora de trabajar la base de dato
4. **Transformación de los datos**: Creación de nuevas variables en función de las ya existentes, discretizar variables numéricas y darle una forma a la base de datos que permita aplicar los modelos.
5. **Data Mining**: Esta etapa incluye, en primer lugar, la definición del objetivo principal de la minería de los datos. Principalmente si esta será de carácter predictiva o descriptiva. En función de esto, la elección de uno o más algoritmos de Data Mining para encontrar patrones. Finalmente, su implementación y siguientes iteraciones, hasta poder obtener resultados satisfactorios.
6. **Interpretación**: Evaluación e interpretación de los patrones obtenidos, así como su confiabilidad en función del objetivo definido. A raíz de esto, puede surgir de que se requiera volver al paso 4, realizar modificaciones y comenzar desde ese punto nuevamente.
7. **Uso del conocimiento descubierto**: Integrar el conocimiento a otro sistema para futuras actividades. El riesgo recae en que todo este conocimiento fue descubierto a partir de datos estáticos y el desafío está en ver que dicho conocimiento converse con la realidad dinámica.

Cabe destacar que el proceso KDD es interactivo e iterativo, dado que involucra al usuario final en las distintas fases, especialmente en la de interpretación y puede resultar en tener que volver a realizar los primeros pasos.

1.7. Resultados esperados

Para cumplir con los objetivos declarados anteriormente y con la metodología declarada como directriz, se esperan los siguientes resultados asociados a cada objetivo:

1. Implementación de algoritmo no supervisado que permita agrupar a los pacientes con trayectorias sintomáticas similares, identificando así las trayectorias características del COVID-19.
2. Análisis detallado de las características individuales (antecedentes, edad, genero, variables sociodemográficas) asociadas a cada trayectoria.
3. Interfaz en la que se puedan visualizar e interactuar con las trayectorias definidas.

2. Marco Teórico

El desarrollo de la solución descrita en este documento se basa principalmente en los siguientes conceptos. Los puntos 2.1 y 2.2 se centrarán en el manejo y tratamiento de datos, los puntos 2.3 y 2.4 tratarán sobre técnicas de Machine Learning y la evaluación de estas, mientras que el último punto tratará sobre el estado del arte con respecto al COVID-19:

2.1. Feature Engineering

Los algoritmos de *machine learning* están basados principalmente en generar datos de salida a través de un set de datos de entrada. Para que el algoritmo funcione de manera apropiada y el resultado de salida brinde valor, es fundamental que los datos de entrada estén correctamente configurados.

El objetivo de *Feature Engineering* vendría a ser, en primer lugar, preparar un set de datos de entrada adecuado, de modo de que este sea compatible con los requerimientos que requiera el algoritmo de *machine learning* a implementar. En segundo lugar, mejorar el desempeño que este algoritmo pudiese llegar a tener.

Existe una serie de técnicas asociadas que, dependiendo del set de datos que se tenga, de los algortimos a implementar y de los resultados buscados, pueden tener una mayor o menor relevancia. A continuación, se mencionarán las principales técnicas de *Feature Engineering* que son indispensables para el trabajo desarrollado. [7]

2.1.1. Imputación de datos

Un problema común dentro de las bases de datos es que estén incompletas, es decir, que haya lo denominado *missing values*. Esto puede deberse a errores humanos, problemas en el flujo de obtención de datos o en el caso de las series de tiempo, días en los que no se pudieron obtener datos.

Los principales métodos para hacerse cargo de esta problemática son los siguientes:

1. Eliminación: Directamente eliminar el registro (fila) que no esté completo. Notar que esta opción puede reducir considerablemente el tamaño de la muestra en caso de que haya muchos registros incompletos. Dentro de este punto, también se pueden definir umbrales de eliminación, por ejemplo, la eliminación de filas que tengan más de un cierto porcentaje de valores nulos.
2. *Numerical Imputation*: Imputar con algún número que tenga sentido dado el contexto de la base de datos. A modo de ejemplo, si se trata de datos binarios donde "1" corresponde a si y "0" a no, es probable que si hay un valor faltante este corresponda a 0. En ejemplos con variables numéricas la mediana o el promedio son candidatos. De esta forma, no se reduce el tamaño de la muestra.
3. *Backward Filling:* Cuando hay valores nulos en una secuencia de datos o series de tiempo, estos pueden ser construidos a través de este método, el cual completa los valores nulos de la secuencia copiando el siguiente valor no nulo.

4. *Forward Filling:* Análogo al método anterior, pero en dirección contraria.
5. Interpolación: Completar el dato faltante a través de los registros adyacentes. La interpolación más conocida es la lineal, siendo prácticamente el promedio entre el registro anterior y el siguiente.

Otra solución propuesta por Zhang [8] denominada CMI (Clustering-based Missing Value Imputation) consiste en aplicar algoritmos de clusterización a toda la base, obteniendo así distintos grupos similares. Luego, los valores faltantes de un individuo perteneciente a un clúster se completan con los valores de dicho clúster.

2.1.2. Manejo de valores atípicos

Ampliamente conocidos como *outliers*, son datos cuyos valores están considerablemente desviados del resto de los datos de la muestra. Dependiendo del tipo de datos con los que se esté trabajando, los *outliers* pueden ser interpretados como errores que deben ser "arreglados", o, como fenómenos de interés a ser estudiados. Independiente del caso, son datos que deben ser tratados de forma diferente al resto de la muestra. [9]

Para identificar los *outliers* las principales técnicas son:

1. Visualizaciones:
    a. Diagrama de caja.
    b. Diagrama de dispersión.
    c. Gráfico Q-Q (cuantil) o Gráfico P-P (probabilidad).
2. Metodologías estadísticas:
    a. Desviación estándar.
    b. Percentiles.
    c. Distancias con el resto de los puntos (tales como la Euclidiana o la de Mahalanobis).

Para tratar con los *outliers*, las principales técnicas consisten en:

1. Corregir el valor en caso de que sea producto de un error.
2. Mantener el registro.
3. Remover el registro de modo de que no pueda influir en el análisis.
4. Reportar 2 tipos de hallazgos, uno para la base con *outliers* y otro para la base sin los *outliers*. Luego comparar los resultados y evaluar las diferencias.
5. Aplicar alguna transformación o tratamiento matemático a los *outliers* que permita suavizar la desviación y a la vez mantener la esencia del valor.

Lo crucial a la hora de definir cómo tratar con los outliers es identificar previamente si estos corresponden a casos poco frecuentes de interés o a "ruido" que no representa las propiedades de los datos.

2.1.3. Binning

Técnica utilizada para reducir los efectos de observaciones menores. Los datos originales se pueden agrupar en pequeños intervalos llamados *bins*, los cuales permiten crear un modelo más robusto que reduce la probabilidad de overfitting. Un ejemplo típico de binning es para representar la edad, la cual originalmente viene como un valor numérico entero mayor que cero, el cual a través de *binning* puede ser representado en una cantidad considerablemente menor de rangos etáreos.

2.1.4. One-Hot encoding

Los datos categóricos, también conocidos como variables cualitativas o variables de atributos, tienden a no ser compatibles directamente con los algoritmos de machine learning. Esta metodología busca convertir una columna de datos categóricos en múltiples columnas de datos binarios. De forma más concreta, si se tienen N atributos o categorías distintas en la columna objetivo, a través del One-Hot encoding esa columna se transformaría en N columnas, donde cada una está asociada a un atributo en particular. De esta manera, si un usuario tenía el atributo $n \in N$, luego de la transformación este tomaría el valor de 1 en la columna nueva columna asociada a n y el valor de 0 en las otras.

Finalmente, se logra representar toda la información de los datos categóricos a través de valores que pueden ser incorporados en los algoritmos de machine learning.

2.2. Dimensionalidad

Existe un término llamado *Curse of Dimensionality*, inicialmente introducido por Bellman [10], el cual indica que el número de muestras necesarias para estimar una función con un determinado nivel de precisión crece exponencialmente con respecto al número de variables independientes de la función. En otras palabras, el valor añadido por agregar una nueva dimensión es mucho menor en comparación con la sobrecarga que agrega al algoritmo. Más aún, dicha sobrecarga al algoritmo puede resultar en un mal funcionamiento de este conllevando a resultados erróneos. [11]

Ante esta problemática es que surgen distintas alternativas para poder lidiar con ella. En primer lugar, se tiene la alternativa más fácil y directa, que es extraer las variables que directamente no aportan al modelo. La forma de definir si una variable aporta o no puede ser de distintas formas, como por ejemplo a través de regresiones donde dicha variable no represente significancia alguna, concluyendo así que su incorporación no aportaría en los resultados del obtenidos.

Otra forma de lidiar con el problema de la dimensionalidad es a través de algoritmos específicamente diseñados para esta labor, los cuales, en vez de eliminar directamente las variables, proyectan el espacio dimensional inicial en uno de menor dimensión manteniendo la información más relevante y explicativa de las variables, lo que permite una implementación más fácil y directa de los futuros algoritmos.

Un algoritmo de reducción de dimensionalidad ampliamente utilizado en distintos estudios y modelos de distintos rubros, destacando su empleabilidad en estudios del campo de la medicina, es el denominado *Principal Component Analysis. [12,13,14,15]*

Principal Component Analysis (desde ahora PCA) es un método estadístico no supervisado que funciona según lo declarado en el párrafo anterior. Ante una muestra con n individuos y v variables (o dimensiones), a través de PCA se puede encontrar un número de factores f (donde $f < v$) que logran explicar en casi su totalidad lo mismo que las v variables iniciales. Cada uno de estos f factores llevan el nombre de componente principal y son construidos a partir de combinaciones de las variables originales. [16]

Este método se basa en una serie de pasos y procesamientos matemáticos. Como primer requisito se tiene que los datos deben estar estandarizados o normalizados, de modo de que cada uno contribuya de igual forma al análisis. El método es extremadamente sensible a las varianzas de las variables iniciales, por lo que si hay grandes diferencias numéricas entre las variables, las que tengan valores mayores van a dominar a las que tengan valores menores, traduciéndose en resultados sesgados.

El siguiente paso consiste en evaluar las relaciones existentes entre las variables, con el fin de identificar si hay variables altamente correlacionadas que como consecuencia entreguen información redundante al modelo. Esta información se obtiene a partir de la matriz de covarianzas, la cual es una matriz simétrica de v x v, con v como la cantidad de variables iniciales. Si hubiera 3 variables (x,y,z) $\in v$, la matriz se vería de esta forma:

Cov (x,x)	Cov (x,y)	Cov (x,z)
Cov (y,x)	Cov (y,y)	Cov (y,z)
Cov (z,x)	Cov (z,y)	Cov (z,z)

Donde la fórmula de la covarianza viene dada por:

$$Cov(X,Y) = \frac{\sum_i^n (xi - \bar{x})(yi - \bar{y})}{n}$$

Los valores positivos están asociados a que ambas variables incrementan (o disminuyen) a la par, es decir, que están correlacionados. Los valores negativos serían el caso contrario donde están inversamente correlacionados y los valores cercanos a 0 sugerirían que no hay una correlación asociada entre ambas variables.

A continuación, se procede a calcular los vectores y los valores propios de la matriz de covarianza. Los vectores propios son todos aquellos vectores que, multiplicados con la matriz de covarianza resultan en un múltiplo entero de sí mismos. A este múltiplo entero del vector propio se puede acceder a través de la multiplicación de un escalar con el vector propio. Este escalar es denominado valor propio. Esto se puede representar de la siguiente manera, con M como la matriz de covarianza, f_j como un vector propio y λ_j como su respectivo valor propio:

$$M \times f_j = \lambda_j \times f_j$$

Los vectores propios tienen la propiedad de ser ortogonales entre si y existen tantos vectores propios como dimensiones tenga la matriz. Para el ejemplo anterior de una matriz de 3x3, existirían 3 vectores propios con sus respectivos valores propios. Otra

forma de interpretar estos parámetros es que los vectores propios describen las direcciones en las que están distribuidos los datos, mientras que los valores propios determinan la magnitud dada dicha dirección.

PCA busca reducir la dimensionalidad de las variables sin perder demasiada información de estas. Dicho de una forma más geométrica, se busca mantener las direcciones que expliquen la mayor cantidad de varianza. En base a la definición del punto anterior, como los vectores y valores propios representan las direcciones en las que están distribuidos los datos y sus respectivas magnitudes, es que se emplean estos mismos para definir los nuevos componentes principales. De forma más precisa, se ordenan de mayor a menor los λ_j valores propios y se definen como componentes principales los f_j vectores propios asociados a los valores propios. De esta manera, el primer componente principal (PC1) vendría a ser el vector propio asociado al valor propio de mayor magnitud, el segundo componente principal (PC2) el asociado al de segunda mayor magnitud y así sucesivamente.

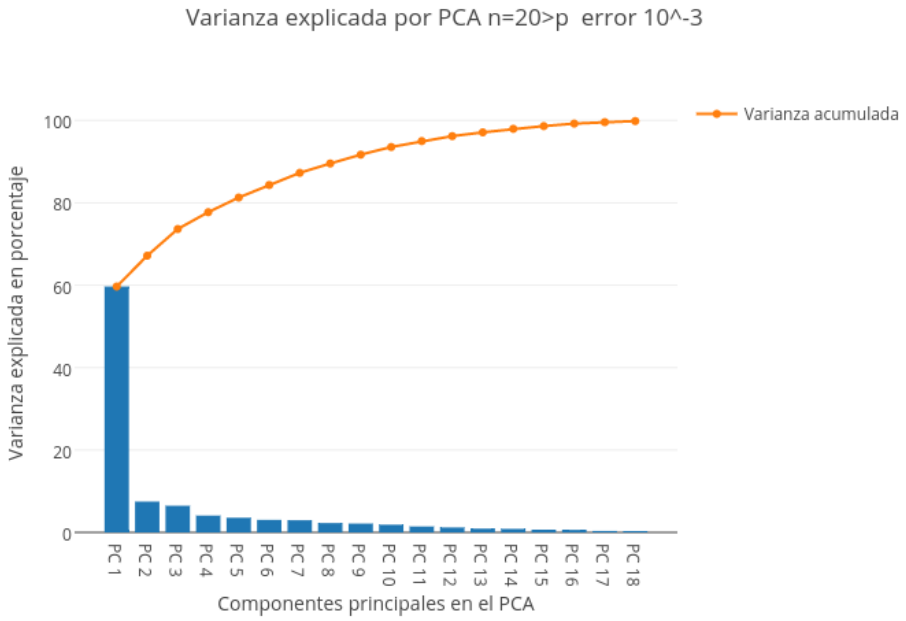


Figura 2.1: Varianza explicada por PCA. Fuente: Chart-Studio

Finalmente, no existe una única forma de definir el número óptimo de componentes principales a conservar. Como el objetivo principal es reducir la dimensionalidad, es de esperarse que se elija una cifra considerablemente menor al número de variables originales. En ocasiones donde se busca representar gráficamente un set de datos de múltiples variables, se pueden conservar únicamente los primeros 2 componentes principales de modo de poder representarlos en un gráfico con 2 ejes. No obstante, en ocasiones en las que el foco no es una representación visual si no que preparar la base de datos para que esta tenga un mejor funcionamiento al ser trabajada con algoritmos de machine learning, distintos autores sugieren elegir una cantidad de componentes principales que permita explicar alrededor de un 95% de la varianza de la base de datos original. De esta forma, se logra reducir la cantidad de variables, eliminar el problema de variables correlacionadas dentro del modelo y se pierde solamente una parte ínfima

de la información contenida en la base original.

Otros ejemplos de métodos para reducir la dimensionalidad que también son conocidos por sus resultados son el método supervisado Linear Discriminant Analysis (LCA) y el método no linear, probabilístico, t-distributed Stochastic Neighbour Embedding (t-SNE).

2.3. Análisis de clústeres

2.3.1. Objetivo general

Clustering es la asociación de objetos a grupos (clústeres) de objetos similares. Estos objetos por lo general son vectores de datos caracterizados por una serie de atributos, a través de los cuales se determina como realizar la partición. Estos atributos pueden ser tanto numéricos, como binarios y categóricos. Lo fundamental de los clústeres, es que los objetos se asocian a un clúster minimizando la distancia hacia este y maximizando la distancia hacia otros clústeres. [17]

La definición de dicha distancia (o equivalentemente, similitud) es una función que se puede definir de manera diferente dependiendo de los tipos de datos que se quieran comparar. Ejemplos típicos son la distancia Euclidiana, la distancia de Mahalanobis o la distancia de Hamming.

La asignación de los clústeres se puede definir como una asignación *fuerte*, donde cada objeto pertenece solamente a un clúster o puede ser *difusa*, cuando un objeto puede pertenecer a más de un clúster con ciertas probabilidades asociadas.

2.3.2. Principales familias de algoritmos

Existe una amplia gama de algoritmos diseñados para la clusterización. Estos se agrupan en familias dependiendo de su funcionamiento y sobre qué tipos de espacios de datos funcionan de mejor manera. Estas familias son las siguientes: [18,19]

1. Métodos Jerárquicos: Corresponde a uno de los primeros métodos de clusterización. La idea general detrás de los algoritmos de esta familia consiste en asignar a todos los objetos un clúster. Es decir, existe una igual cantidad de clústers como de objetos. Luego, el segundo paso consiste en encontrar el par de clústeres más cercanos entre sí y combinarlos en un solo clúster. Finalmente, se repite el segundo paso hasta que todos los objetos estén dentro del mismo clúster. De esta forma, se arma una jerarquía de particiones.
2. Métodos basados en prototipos: Estos tipos de algoritmos usan el concepto físico de centro de masa, al cual llaman centroide. Existe un centroide por cada clúster, el cual representa el prototipo de objeto "promedio" de dicho clúster. A su vez, las distancias de los objetos se miden con respecto al centroide. Dentro de los algoritmos de este tipo destacan los Modelos de Mezclas Gaussianas, Fuzzy C-Means, K-Means, K-Medoids y K-Modes.
3. Clustering basado en densidad: Estos algoritmos se basan en la teoría de que los puntos aparecen dentro de un clúster, el cual es considerablemente más

denso que el espacio en el que no está comprendido el clúster, lo que permite identificar el ruido (datos que no están dentro de ningún clúster) además de clústeres adyacentes. El algoritmo más conocido de este tipo es el DBSCAN.

4. Clustering basado en grafos: También llamado como detector de comunidades, tiene como objetivo encontrar subgrafos dentro de un grafo. Esto es únicamente posible cuando la cantidad de nodos es muy similar a la cantidad de enlaces.

2.3.3. Métricas de desempeño

Con el fin de determinar si el algoritmo de clustering implementado cumple con el objetivo señalado en el punto 2.4.1, además de poder definir la cantidad óptima de clústeres, es que existen distintos índices y métricas para validarlo. Para efectos de este trabajo, se explicarán 3 índices que permiten validar el desempeño de los clústeres cuando no hay un conocimiento de las verdaderas etiquetas, es decir, las únicas etiquetas existentes son las definidas a través del algoritmo de clusterización. [20]

En primer lugar, se tiene el índice de Calinski-Harabasz, también conocido como el *Variance Ratio Criterion*. El índice es una medida donde la cohesión (que tan similar es un punto con respecto a su clúster) se estima en base a la distancia que tienen los puntos del clúster hacia su centroide, mientras que la separabilidad se basa en la distancia que tienen los centroides de los clústeres con respecto al centroide global. La forma del índice es la siguiente:

$$\frac{Separabilidad}{Cohesión}$$

De este modo, mientras haya una mayor separabilidad y una menor cohesión, se tendrá un mayor índice de Calinski-Harabasz, traduciéndose en un mejor desempeño para una determinada cantidad de clústeres. De manera más detallada, el índice se obtiene de la siguiente fórmula:

$$CH = \frac{\frac{\sum_k^K n_k \|c_k - c\|^2}{K-1}}{\frac{\sum_k^K \sum_i^{n_k} \|d_i - c_k\|^2}{N-K}}$$

Donde K es el número de clústeres en el set de datos D = $[d_1, d_2, \dots d_N]$, n_k es el número de puntos del clúster k, c_k es el centroide del clúster k, c es el centroide global y N el total de puntos.

En segundo lugar, se tiene el índice de Davies-Bouldin [21], cuyo valor no depende directamente del número de clústers. El índice indica la similitud promedio entre clústeres, donde dicha similitud es una medida que compara la distancia entre los clústeres con el tamaño de ellos mismos. La formulación matemática de dicha similitud es la siguiente:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Donde s_i es la distancia promedio que hay entre cada punto del clúster i y el centroide

de dicho clúster. El denominador d_{ij} representa la distancia que hay entre los centroides de los clústeres i y j respectivamente. Luego, partir de dicha definición se construye el índice:

$$DB = \frac{1}{K}\sum_{i}^{K} max\, R_{ij}$$

El valor que indica este índice corresponde al promedio de similitud que hay entre cada clúster y su clúster más cercano. En consecuencia, la mejor elección de clústeres será la que minimice el índice de Davies-Bouldin.

En tercer lugar, se tiene el coeficiente de siluetas. Este coeficiente se mide para cada muestra a través de la siguiente fórmula:

$$s = \frac{b - a}{\max(a, b)}$$

Donde "a" corresponde a la distancia promedio entre el punto y todos los otros puntos del clúster al cual pertenece y "b" corresponde a la distancia promedio entre el punto y todos los otros puntos pertenecientes al clúster más cercano. Para medir el desempeño total de la clusterización, se toma el promedio de todos los "s". Este valor va desde el -1 al 1, donde los valores negativos se asocian a un clustering incorrecto, mientras que los valores cercanos al 1 se asocian a clústeres con alta cohesión y separabilidad.

Finalmente, de forma adicional a los coeficientes de validación de clustering, es fundamental la evaluación externa por parte de algún experto en el rubro, con el fin de que se pueda corroborar de que la forma en la que se agrupan los datos en base a las distancias que tienen entre si tenga sentido fuera del punto de vista matemático.

2.4. Aprendizaje supervisado

2.4.1. Objetivo general

A través de la implementación de algoritmos sobre datos etiquetados, se busca asignarle una etiqueta de salida adecuada a nuevos datos. En otras palabras, se entrena un algoritmo el cual está compuesto por distintas variables y una clasificación y luego, ante un nuevo conjunto de datos con las mismas variables, se sugiere la clasificación más adecuada en función de los datos con los que fue entrenado y de los patrones de estos que aprendió.

2.4.2. Tipos de error

Para la evaluación de un modelo existen distintas medidas de error para medir y comparar el desempeño de la clasificación. Los siguientes términos son claves para poder definir métricas más avanzadas que cuantifiquen a un buen predictor

- **Verdadero Positivo (TP)**: Número de casos en los que el algoritmo predijo una clase positiva cuando efectivamente era una clase positiva
- **Verdadero Negativo (TN)**: Número de casos en los que el algoritmo predijo que era una clase negativa cuando efectivamente era una clase negativa

- **Falso Positivo (FP)**: Número de casos en los que se predijo una clase positiva y realmente era una clase negativa
- **Falso Negativo (FN)**: Número de casos en los que se predijo una clase negativa y realmente era una clase positiva

2.4.3. Métricas de desempeño

En base a los tipos de error mencionados en la sección anterior, existen las siguientes métricas para medir el desempeño de una clasificación

- **Accuracy**: Se define como el porcentaje de aciertos sobre el total de los casos

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Recall**: También conocido como especificidad, corresponde a la proporción de observaciones correctamente predichas

$$Recall = \frac{TP}{TP + FN}$$

- **Precisión**: También conocido como sensibilidad, corresponde a la proporción etiquetadas como una clase que realmente correspondían a dicha clase

$$Precision = \frac{TP}{TP + FP}$$

- **F1-Score**: Es una combinación de la precisión y el recall que busca incorporar información relevante de ambas métricas en solo una

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

2.5. Trabajos relacionados

Hay una amplia gama de estudios relacionados con el COVID-19. En los siguientes puntos, se desarrollará lo que estos estudios señalan sobre la respuesta inmunológica que tiene el cuerpo ante el virus, los síntomas que presenta una persona infectada, los principales factores de riesgo y finalmente, una caracterización de las variables sociodemográficas chilenas que podrían presentar diferencias en comparación a otros países del mundo en los que se han hecho los distintos estudios.

2.5.1. Respuesta inmunológica ante el SARS-CoV-2

Normalmente la respuesta inmune del ser humano ante la mayoría de los virus consiste en una fase de contención rápida dirigida por la inmunidad innata la cual, en caso de no ser suficiente, deriva a una fase de contención tardía adaptativa, sofisticada que debería eliminar el virus y en el mejor de los casos, generar memoria inmunológica que permita enfrentar a dicho patógeno en caso de que vuelva a ingresar al cuerpo. [22]

Se ha mostrado que el SARS-CoV-2 es capaz de emplear mecanismos de evasión del

sistema inmune innato, especialmente en pacientes cuyo sistema inmune esta debilitado ya sea por la edad o por comorbilidades, lo que produce que la patología pueda converger en una neumonía, fallo de múltiples órganos y finalmente, la muerte. [23]

El periodo de incubación varía entre 2 a 14 días con 5 días como promedio y en casos extremos puede tardar hasta 24 días la incubación. Una vez que el virus está incubado e infecta al organismo, en la mayoría de los casos (más del 80%) se manifiesta una enfermedad leve, la cual puede ser asintomática o presentar síntomas a lo largo de los siguientes 11.5 días (CI, 8.2 – 15.6). La duración promedio que tarda el cuerpo en eliminar completamente este virus es alrededor de 20 días. [24]

El síndrome de dificultad respiratoria aguda (SDRA) es la mayor complicación asociada al COVID-19 y es cuando se categoriza como que el paciente tiene una enfermedad severa. Esta complicación suele darse después de una mediana de 8 días en un 20% de los pacientes, aunque la ventilación mecánica es requerida solo en un 12.3% de los casos. [22]

Finalmente, si bien el COVID-19 es mundialmente conocido por ser una enfermedad respiratoria, también existe evidencia de que los síntomas que produce y los órganos que afecta no son únicamente los del sistema respiratorio. Últimos estudios han demostrado que el COVID-19 puede definirse como una enfermedad vascular, es decir que afecta directamente al aparato circulatorio. Las glicoproteínas espiga, también conocidas como peplómeros que tiene el SARS-CoV-2 se unen a la enzima convertidora de angiotensina 2 (ACE2) que está presente en el cuerpo. Dentro de las funciones de esta enzima, se encuentra la regulación de la presión sanguínea, la función cardíaca y la función pulmonar. A través de esta unión, el SARS-CoV-2 comienza a disminuir la cantidad de ACE2 en los pulmones, provocando de esta manera daño al sistema respiratorio de distintas magnitudes. Además, desde está unión puede infectar el endotelio vascular, lo que desencadena un daño a nivel mitocondrial que finalmente repercute en una variedad de secuelas en el organismo que se manifiestan a través de distintos y aparentemente inconexos síntomas. [25, 26]

2.5.2. Dinámicas sintomáticas

Según el Centro Estatal de Vigilancia Epidemiológica y Control de Enfermedades de México, heterogeneidad es una de las palabras que mejor define la infección causada por el SARS-CoV-2. Las secuelas del COVID-19 en el organismo se presentan principalmente en las distintas áreas:

1. **Respiratorios**: Tos, dificultad para respirar, neumonía, fibrosis pulmonar
2. **Neurológicos**: Anosmia (pérdida de olfato), disgeusia (perdida de gusto), cefaleas (dolores de cabeza)
3. **Digestivos**: Vómitos, náuseas, diarreas, dolor abdominal
4. **Corporales**: Mialgias, Fatiga, Decaimiento

No obstante, también se han declarado síntomas asociados a problemas cardíacos, oftalmológicos, osteomusculares, alteraciones endocrinas, coagulación, insuficiencia renal. De forma concisa, puede afectar gran parte del organismo. [27]

Existe una amplia gama de documentos en la literatura que señalan cuales son los síntomas más comunes, cuales se asocian a casos más riesgosos y la prevalencia que

estos tienen a lo largo del tiempo. A través de una revisión sistemática de los distintos papers desarrollados en base a los síntomas auto reportados por las personas, lo primero a destacar es que existe una gran varianza entre los síntomas señalados en cada estudio, lo cual podría sugerir que los tipos de síntomas que las personas perciben y reportan, dependen de la ubicación y de la realidad que estos viven.

En [28] se analizan 2.471 casos positivos obtenidos de la segunda HMO (Health Management Organization) más grande de Israel entre el 1 de marzo del 2020 y el 7 de junio del mismo año. Dentro de estos datos existen registros sintomáticos por personal de la salud, donde del ranking de síntomas identificados destacan: tos (11.6%), fiebre (10.3%), mialgias (7.7%) y fatiga (5.9%). Además de registros sintomáticos obtenidos por el personal de salud, este estudio contempla y contrasta con los síntomas autos reportados por los pacientes. Dentro de los síntomas auto reportados, la mayor prevalencia fue por parte de la tos (21%), fatiga (19%), congestión nasal (16%) y mialgias (11%). La duración de la enfermedad de los niños fue significativamente más corta. Adicionalmente, pérdidas de olfato y gusto fueron documentadas por los profesionales únicamente en un 1.1% de los casos, aunque estos síntomas fueron auto reportados por un 10% de los pacientes, sugiriendo que hay una diferencia entre la percepción de este síntoma y que este sea validado por un especialista. Con respecto al tiempo en el que estos síntomas fueron identificados, fatiga, mialgias congestión nasal y falta de aire fueron reportados como síntomas tardíos.

En [29] se analizan 105 pacientes COVID-19 positivos de la República Checa entre el 20 de abril del 2020 hasta el 2 de septiembre del mismo año. En total se realizaron 1223 entrevistas telefónicas (promedio 12, mediana 11, mínimo 1 y máximo 25), a pacientes con una edad promedio de 40 años, 52% de mujeres y donde las comorbilidades más frecuentes fueron alergias (43%) e hipertensión (24%). Con respecto a la incidencia de los síntomas, los reportes más comunes fueron síntomas generales del tracto respiratorio (71%), fatiga (65%), fiebre (60%), anosmia (59%), dolor de cabeza (58%), mialgias (55%), ageusia (47%) y tos seca (43%). Adicionalmente, las mujeres reportaron una mayor frecuencia de anosmia que los hombres (66% contra 52%). Con respecto a las duraciones de estos síntomas, pacientes que tuvieron una mediana de 5 días sintomáticos tuvieron fiebre, dolor de cabeza, mialgias, diarrea, dolor abdominal entre otros, mientras que los pacientes que tuvieron una mediana de 10 días sintomáticos reportaron principalmente dificultades para respirar, tos tanto seca o productiva, anosmia y ageusia. La investigación concluye en que no se puede predecir una trayectoria de forma fiable debido a que el curso de la enfermedad es notoriamente variable, enfatizando en la necesidad de monitoreo individualizado para los distintos pacientes.

En [30] se analizan 893 reportes sintomáticos de 270 pacientes COVID-19 positivos de Arabia, entre el 3 de marzo del 2020 y el 27 de mayo del mismo año. De este estudio se desprende una frecuencia de síntomas consistente entre hombres y mujeres, donde destaca que la fiebre fue auto reportada por el 60% de los usuarios, además de ser típicamente junto al dolor de cabeza los síntomas iniciales. El top 4 de síntomas que fueron reportados junto a la fiebre fueron el dolor de cabeza (23.7%), tos seca (14.4%), anosmia (13.7%) y ageusia (12.2%). La fatiga y el dolor de garganta fueron síntomas menos frecuentes y tardíos.

En [31] se busca el posible orden de síntomas discernibles de COVID-19. Se utilizan los datos de WHO-China Joint Report, donde hay 55.924 casos confirmados entre el 16

de febrero del 2020 y el 24 del mismo mes y año. De estos registros sintomáticos, se identificaron las secuencias más y menos probables que un paciente cualquiera podría presentar. La secuencia más probable (76,9%) viene dada por comenzar con fiebre, luego tos y en un 57,5% continuar con nauseas/vómitos y finalmente diarrea. La secuencia menos probable (1%) comienza con diarrea, náuseas, vómitos, sigue en un 23,8% de los casos con tos y finaliza con fiebre. Estas secuencias fueron confirmadas con un grupo de datos de validación de 1.099 pacientes, en donde además se analizaron las diferencias entre los pacientes que tuvieron una enfermedad leve y los que tuvieron una enfermedad más severa (requerir hospitalización y/o ventilación mecánica). De este análisis, se sugiere que las secuencias más y menos comunes de presentar estos síntomas son idénticas para pacientes con una enfermedad leve en comparación con los pacientes con una enfermedad severa. En otras palabras, estos resultados sugieren que la severidad del COVID-19 no altera el orden de aparición de los síntomas discernibles.

En [32] a través de algoritmos no supervisados se encuentran 6 clústeres sintomáticos, con los datos generados por 1.653 usuarios en la aplicación COVID Symptom Study App. Estos usuarios provienen de Reino Unido, Estados Unidos y Suecia, tuvieron síntomas persistentes y un registro regular en la aplicación hasta fines de abril del 2020. De este trabajo se obtuvieron 6 clústeres sintomáticos, los cuales están enumerados de forma que, mientras mayor sea el número del clúster, mayor es la tasa de hospitalización. En otras palabras, el clúster 1 tiene las manifestaciones más leves de la enfermedad mientras que el clúster 6 tiene las más graves. Los primeros dos grupos sintomáticos destacan por ser versiones leves de la enfermedad, donde los principales síntomas se asocian al tracto respiratorio (tos seca, dolor de garganta) y se diferencian en que el segundo clúster no presenta dolores musculares y si presenta fiebre. El tercer clúster presenta una mayor cantidad de síntomas gastrointestinales tales como diarrea o anorexia y no presenta tos. Los últimos 3 clústeres tienen tos persistente y se diferencian de los anteriores por la presencia de fatiga y dolor de pecho. Adicionalmente, los usuarios del clúster 5 presentan confusión, dolor de garganta y los usuarios del clúster 6 declaran disnea, tanto temprana como tardía. En todos los clústeres sintomáticos se presencia dolor de cabeza a lo largo de la enfermedad y pérdida del olfato, donde este último síntoma aumenta su frecuencia con el paso de los días y está ligeramente más presente en los primeros 4 clústers que en los últimos 2. Los principales frutos de estos clústeres sintomáticos son que el dolor de cabeza es un síntoma trascendental al clúster, que la fatiga está asociada a clústeres con desenlaces más severos y que la pérdida de gusto y olfato está más asociada a cuadros largos y menos graves de COVID-19.

De los distintos estudios, hay una tendencia de que los síntomas más comunes corresponden a la tos seca, fiebre, dolor de cabeza, fatiga y pérdida del olfato. No necesariamente se presentan todos estos síntomas en conjunto, además de que los síntomas previamente señalados no son exclusivos del COVID-19, lo que dificulta en gran medida su diagnóstico a través de únicamente el cuadro sintomático, por lo que es importante mantener estrategias de testeo, trazabilidad y aislamiento con el fin de la correcta identificación de casos y evitar la propagación del virus.

A pesar del punto anterior, diversos estudios han señalado que la anosmia y la disgeusia son uno de los síntomas claves del COVID-19, presentando la mayor razón de momios asociados a la positividad de los casos. Estos 2 síntomas, si bien todo tipo de pacientes los han reportado, están mayormente asociados a mujeres jóvenes que

tuvieron una enfermedad leve o moderada. [33, 34, 35, 36, 37]

Los síntomas gastrointestinales son síntomas atípicos de la enfermedad y en los casos en que han aparecido, suelen manifestarse como síntomas iniciales y de escasa duración [38, 39, 40]. Finalmente, la disnea se asocia a un síntoma de riesgo, el cual usualmente se manifiesta alrededor de una semana (mediana de 5 hasta 8 días) de iniciados los síntomas. Cabe destacar que la relación entre reportar disnea y tener un desenlace severo (manifestar el SDRA, requerir suplementación de oxígeno o fallecer) no es estricta, dado que no todos los pacientes que tuvieron dicho síntoma tuvieron un desenlace severo, así como tampoco todos los pacientes con un desenlace severo reportaron haber tenido disnea. [22, 41]

2.5.3. Factores de riesgo

Además de la disnea como síntoma asociado al riesgo de una enfermedad severa, existe una serie de características personales que se relacionan con un desenlace riesgoso. En primer lugar, se tiene el riesgo asociado a una avanzada edad, siendo particularmente más riesgoso el grupo de personas que tienen más de 60 años. [42]

Dentro de las comorbilidades asociadas a complicaciones con el virus destacan la diabetes mellitus, la hipertensión, bajo nivel de linfocitos (Lymphocytopenia), las cuales a través de estudios de regresión de múltiples variables resultaron ser predictores significativos de una enfermedad severa. Adicionalmente, el asma es una condición preexistente asociada con cuadros largos (más de 3 semanas) de COVID-19, aunque no está asociado con un desenlace riesgoso. [43]

Hay una importante cantidad de evidencia que señala que el COVID-19 produce cuadros sintomáticos más severos y asociados a una mayor tasa de mortalidad (aproximadamente el 60% de las muertes por COVID-19) en los hombres [44, 45]. Una posible explicación a esta diferencia en el sexo es desarrollada en [46], donde se encontró que las mujeres tienen una activación más robusta de linfocitos ante la infección de SARS-CoV-2, la cual está correlacionada con un mejor desenlace.

2.5.4. Variables sociales en Chile

De forma complementaria a los puntos anteriores, es crucial tener en consideración las variables socioeconómicas y demográficas del país y particularmente, de Santiago, dado que estas no necesariamente son comparables con las del resto del mundo, teniendo como consecuencia diferentes implicancias en el acceso a tratamientos, presencia de comorbilidades y una distinta tasa de mortalidad.

En septiembre del 2020, Chile tenía más de 400.000 casos y 590 muertes por COVID-19, teniendo una de las peores tasas a nivel mundial. Para dicha fecha, el 70% de los casos eran reportados por las comunas de Santiago, convirtiéndolo en uno de los brotes urbanos más grandes del mundo. [47]

Si bien se aplicaron distintas intervenciones no farmacéuticas (NPI) para disminuir el aumento de casos, tales como cuarentenas dinámicas, el efecto de estas fue considerablemente diferente en las distintas comunas de Santiago. Particularmente, las personas de escasos recursos y, en consecuencia, las comunas en las que ellos residen presentaron mayores tasas de mortalidad y contagio. Los motivos que potencialmente explican este fenómeno son la escasa capacidad de mantener el distanciamiento social en comunas de menores recursos y la menor capacidad de acceder a atención

hospitalaria. [47]

En otro estudio se encontraron relaciones robustas entre el desenlace de los pacientes y su estatus socioeconómico, siendo esta relación aún más fuerte en grupos de menores edades. Personas entre 0 y 40 años pertenecientes al estatus socioeconómico más bajo tienen una tasa de mortalidad hasta 10 veces mayor que la de las personas pertenecientes al estatus socioeconómico más alto. [48]

Finalmente, en un estudio publicado por la Comisión Económica para América Latina y el Caribe (CEPAL) se puede visualizar la evolución temporal de la enfermedad en Santiago de Chile y en particular el efecto de las variables sociales y económicas planteadas anteriormente. En este estudio, se clasifican las comunas según los percentiles de ingresos de la población, teniendo así a las comunas de barrio alto, las comunas de clase media y las comunas populares. [49]

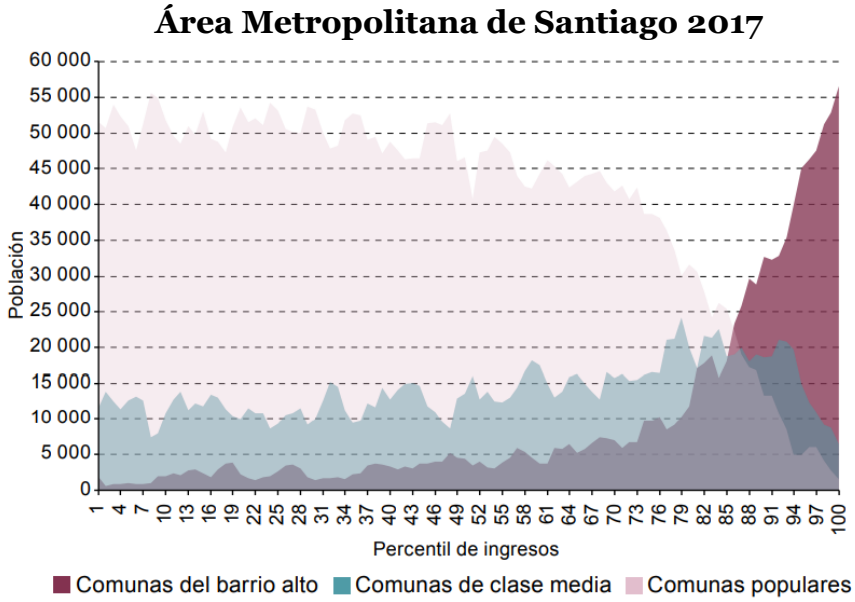


Figura 2.2: Distribución de la población según percentil de ingresos y estrato socioeconómico de las comunas. Fuente: Elaboración de Alejandro I. Canales sobre la base de datos de la Encuesta de Caracterización Socioeconómica Nacional (CASEN), 2017.

Sobre dicha clasificación, en dicho estudio se desarrolla el siguiente gráfico para ver la incidencia del COVID-19 según el estrato socioeconómico.

Área Metropolitana de Santiago 2020
Casos por cada 100.000 habitantes.

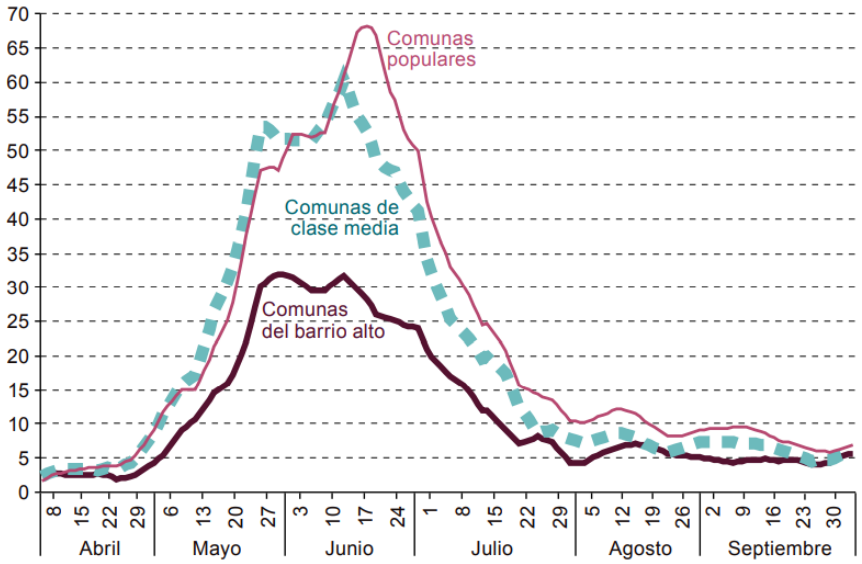


Figura 2.3: Incidencia del COVID-19 según estrato socioeconómico de las comunas.
Fuente: Elaboración de Alejandro I. Canales sobre la base de Ministerio de Ciencia, Tecnología, Conocimiento e Innovación, "Datos- COVID19" [en línea] https://github.com/MinCiencia/Datos-COVID19.

Adicionalmente, en el estudio se diseña la siguiente tabla que permite visualizar las diferencias en la tasa bruta de mortalidad por COVID-19, de donde se destaca la superior proporción de defunciones en las comunas populares, llegando hasta casi 189 fallecimientos por 100.000 habitantes, en contraste con la proporción de las comunas del barrio alto que es menos de la mitad.

Chile al 30 de septiembre 2020

	Defunciones	Tasa bruta de mortalidad (ajustada)
Nacional	17 522	90,1
Área Metropolitana de Santiago	10 978	164,6
Comunas del barrio alto	1 309	90,9
Comunas de estratos medios	1 965	155,8
Comunas populares	7 704	188,6
Otras regiones de Chile	6 544	45,8

Tabla 2.1: Defunciones y tasa bruta de mortalidad por COVID-19 (ajustada por edad)
Fuente: Elaboración de Alejandro I. Canales sobre la base de Departamento de Estadísticas e Información de Salud del Ministerio de Salud (DEIS), "Defunciones por causa de muerte 2016-2020", Ministerio de Salud, 2020 [en línea] https:// deis.minsal.cl/#datosabiertos.

Es importante tener estas consideraciones a la hora de sacar conclusiones, sobre todo porque los datos en los que se desarrolla este trabajo de título corresponden a personas que habitan en comunas populares, con la excepción de La Florida que según el estudio de la CEPAL correspondería a una comuna de clase media.

3 Desarrollo del trabajo

En esta sección se verán las primeras 4 etapas de la metodología KDD, siendo estas la comprensión del dominio, la elección y creación del set de datos, el procesamiento y limpieza y finalmente, la transformación de los datos.

3.1 Comprensión del dominio

La investigación se inserta en el rubro de la salud, específicamente la salud pública de Chile en el SSMSO. El usuario principal del conocimiento que vaya a ser obtenido a lo largo de esta investigación va a ser el equipo clínico y de trazabilidad del SSMSO, sin embargo, esto no se reduce exclusivamente a ellos, dado que el conocimiento que pueda surgir es escalable a otros servicios de salud del país.

3.2 Elección y creación del set de datos

A través de los datos registrados en la Plataforma COVID-19 SSMSO, los datos de las Defunciones registradas por el Departamento de Estadísticas e Información de Salud (desde ahora DEIS) y los datos de los Egresos Hospitalarios por parte de Más Inteligencia en Salud (desde ahora MIS), la Unidad de Salud Digital genera un set de datos anonimizados, que es de la cual se dispone para trabajar. Cabe destacar que la implementación de dichos datos para la investigación y difusión científica del proyecto cuenta con la autorización del Comité Ético Científico del SSMSO.

Esta base de datos tiene 10 tablas (hojas de Excel) diferentes con información sobre los establecimientos en los que se atienden los pacientes, datos personales, resultados de exámenes, contactos de cada paciente, antecedentes auto declarados, síntomas registrados durante los seguimientos, datos de los seguimientos, egresos de la plataforma (entendiendo como egreso de plataforma cuando se le deja de hacer seguimiento), egresos hospitalarios (entendiendo como egreso hospitalario los egresos de los pacientes que fueron hospitalizados) y las defunciones.

Considerando todas las distintas hojas, se tiene un total de:

Tipo de Variable	Cantidad
Categórica	15
ID	13
Fecha	8
Numérica	5
Binaria	5
Texto libre	2

Tabla 3.1: Tipo de variables de la base de datos sin filtrar

La llave principal que conecta las distintas bases de datos es el IDINGRESO, el cual se le asigna a un paciente cuando ingresa a la plataforma. Esta llave es diferente a la de IDPER, asociada al ID de una persona, dado que una persona puede reingresar a la plataforma, teniendo así distintos IDINGRESO para la misma persona. Esto es común en casos donde el paciente ingresa a la plataforma como un paciente "Descartado" gracias a un test de reacción en cadena de polimerasa (PCR) negativo, pero que después de unos meses efectivamente se contagia e ingresa a la plataforma como un paciente "Confirmado", el cual es designado tras recibir un resultado positivo de PCR.

Como lo que se busca con esta investigación es encontrar las distintas trayectorias sintomáticas y las características de los pacientes que se asocian a ellas, es que la selección de los datos debe ir en dicha línea. A raíz de esto, dentro todos los datos disponibles, se conservan únicamente los siguientes:

Categóricas:

1. **Síntomas:** Síntomas declarados por el paciente durante el seguimiento. Existen 26 opciones de síntomas, además de la opción "OTRO" que los pacientes pueden declarar. (Anexo A)
2. **Antecedentes:** Antecedentes o comorbilidades declaradas por el paciente al momento del ingreso. Existen 19 opciones diferentes de antecedentes a declarar. (Anexo B)
3. **Causal de Egreso (Plataforma):** Motivo por el cual se completa el seguimiento. Las distintas causales corresponden al alta, fallecimiento, fallecimiento por otras causas, traslado y abandono. El traslado está asociado a si el paciente fue trasladado a algún centro hospitalario o si fue trasladado fuera del SSMSO.
4. **Destino de Egreso (Plataforma):** Variable que responde el destino producto de la causal de egreso. Los destinos corresponden al domicilio, a controles de atención, hospitalización, urgencia u otro.
5. **Tipo de Egreso (Hospitalario):** Variable exclusiva para pacientes que fueron hospitalizados. Los distintos tipos son si el paciente es dado de alta, fallecimiento o si es trasladado a algún centro público o privado.
6. **Tipo de Ingreso:** Variable que clasifica, dado un determinado ingreso, si la persona es contagiada de COVID-19 o no. Los distintos tipos de ingresos corresponden a Descartado, Confirmado, Probable, Sospechoso y Contacto.
7. **Tipo de Movimiento:** Tipo de operación realizada en la plataforma, donde estas operaciones pueden ser el ingreso de un paciente, los seguimientos de un paciente o el egreso de un paciente. Los síntomas se declaran tanto en el ingreso como en los seguimientos, el egreso es administrativo y típicamente se realiza una vez que el paciente deja de declarar síntomas en los últimos seguimientos.
8. **Tipo de Vacuna:** Variable que representa con que vacuna fue inoculada una persona. Las opciones disponibles son las vacunas provenientes de Sinovac, Pfizer, AstraZeneca y CanSino.

9. **Dosis de Vacuna:** Variable que representa la dosis con la que una persona fue inoculada en un tiempo dado.
10. **Género:** Variable que indica si una persona es hombre o mujer.

ID:

1. **ID de Ingreso:** Cada vez que una persona es ingresada a la Plataforma COVID-19 se le genera un nuevo ID de ingreso, el cual se mantiene hasta que dicha persona es egresada de la plataforma. De esta manera se pueden diferenciar los reingresos de una misma persona a la plataforma.
2. **ID de Persona:** Cada persona tiene asociado un único ID de persona, donde si el paciente reingresa a la plataforma, conservará el mismo ID de persona de la primera vez, pero se le asignará un nuevo ID de ingreso.
3. **ID de Movimiento:** Está asociado a cada movimiento de la Plataforma COVID-19. Como se mencionó en la sección anterior, estos movimientos corresponden al ingreso del paciente, a los distintos seguimientos que se le realice y al egreso en caso de estar egresado.
4. **ID de Egreso:** Asociado al egreso de la plataforma. La existencia de este ID para un ID de ingreso señala que este paciente ya completó su seguimiento.

Fechas:

1. **Inicio Síntomas:** Fecha en la que una persona declara que comenzó a tener síntomas asociados al COVID-19.
2. **Ingreso:** Fecha en la que una persona ingresa la plataforma.
3. **Movimiento:** Fecha asociada a cada movimiento realizado sobre una persona.
4. **Egreso Plataforma:** Fecha en la que una persona completa el seguimiento y es egresada de la plataforma.
5. **Egreso Hospitalario:** En el caso de los pacientes que requirieron hospitalización, esta fecha corresponde a cuando estas egresan del recinto hospitalario.
6. **Vacunación:** Fecha asociada a la inoculación de alguna dosis.

Numéricas:

1. **Edad:** Variable que señala la cantidad de años que tiene una persona.

Sobre esta muestra de los datos originales, en primer lugar, se filtra por el *Tipo de Ingreso* de los pacientes, conservando únicamente los que ingresaron como *Confirmados* y *Probables* y fueron sintomáticos. Los casos *Descartados* de COVID-19 quedan fuera del análisis dado que en su mayoría no presentan registros sintomáticos y en caso de que presenten, los síntomas presentados no se pueden asociar a que estos hayan sido producidos a causa del SARS-CoV-2. Los casos *Probables* se consideran

dentro del análisis dado que, según el Protocolo de Seguimiento de Casos COVID-19 de ANCORA UC [50], que es por el cual se guía el SSMSO, un paciente clasificado como *Probable* es un contacto estrecho sintomático de un paciente COVID-19 positivo, el cual no requiere de una toma de PCR para confirmar la presencia del virus y para efectos de seguimiento es contactado y evaluado de la misma forma que un caso confirmado. El resto de los contactos quedan fuera del análisis dado que no son contactos estrechos o no son contactos sintomáticos. Los casos sospechosos también quedan fuera, debido a la escasa cantidad de seguimientos realizados sobre estos.

Adicionalmente, se mantienen únicamente los pacientes que fueron egresados, con la justificación de que para que el análisis sea correcto, se debe conocer la trayectoria sintomática completa de los pacientes e incorporar pacientes que aún están en seguimiento conduciría a resultados erróneos. Tras estos filtros iniciales, la base seleccionada de momento cuenta con 21.738 IDs de ingreso diferentes, correspondientes a pacientes confirmados o probables que ya fueron dados de alta del seguimiento.

De forma exploratoria, la distribución de la cantidad de síntomas reportados es la siguiente:

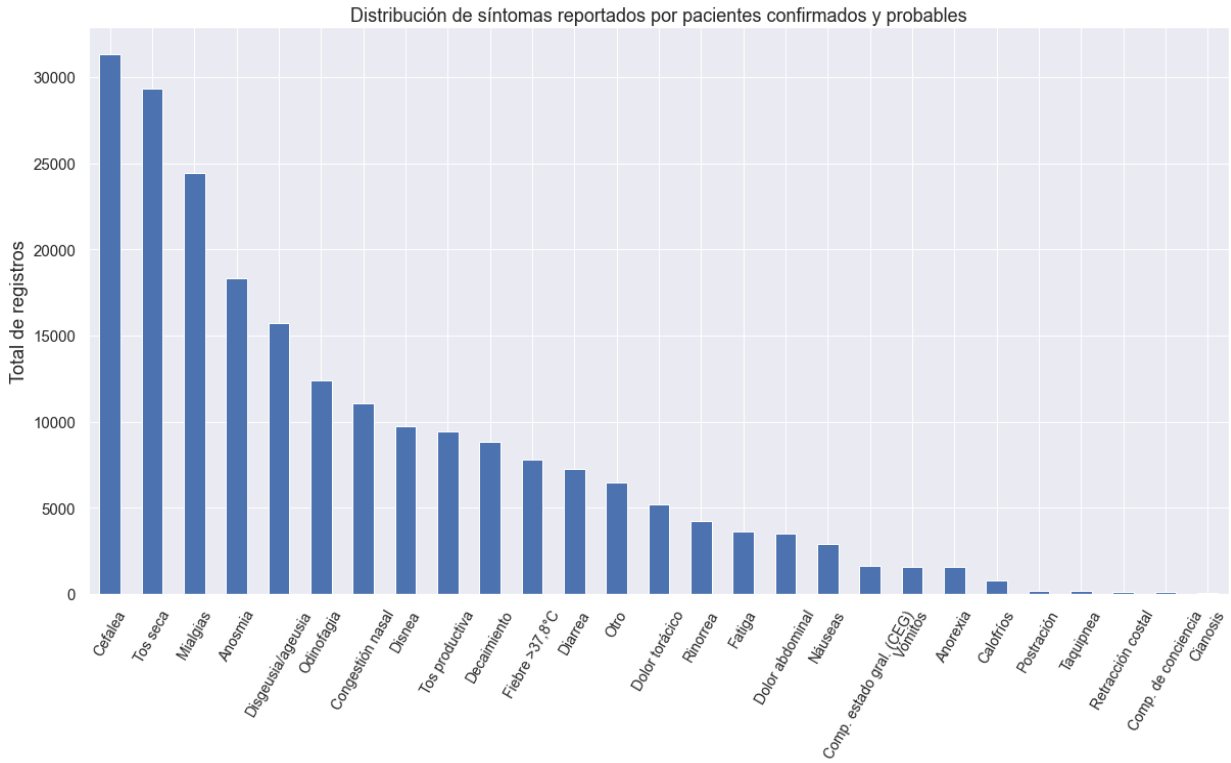


Figura 3.1: Distribución total de los síntomas auto declarados por los pacientes confirmados y probables del SSMSO

De dicho gráfico se puede observar que hay un fuerte desbalance en lo que respecta a los síntomas reportados. Mayoritariamente se reporta cefalea, mialgias y tos seca, mientras que otros síntomas como la taquipnea, postración o la retracción costal son casos aislados.

El desarrollo de este trabajo se realizó con los pacientes que ingresaron a la Plataforma COVID-19 desde el 1 de junio del 2020 hasta el 15 de abril del 2021.

Adicionalmente, se replicaron todas las etapas de la metodología en una muestra independiente de mayo del 2021, con la finalidad de validar la continuidad en el tiempo de los resultados obtenidos. Cabe destacar que, en la muestra original se analizaron pacientes no vacunados o sin una campaña de vacunación completa, mientras que en la muestra de validación de mayo hay datos de pacientes con sus respectivas campañas de vacunación finalizadas.

Los gráficos, cifras y referencias que se harán en las siguientes etapas corresponden a los datos de los pacientes de la muestra original.

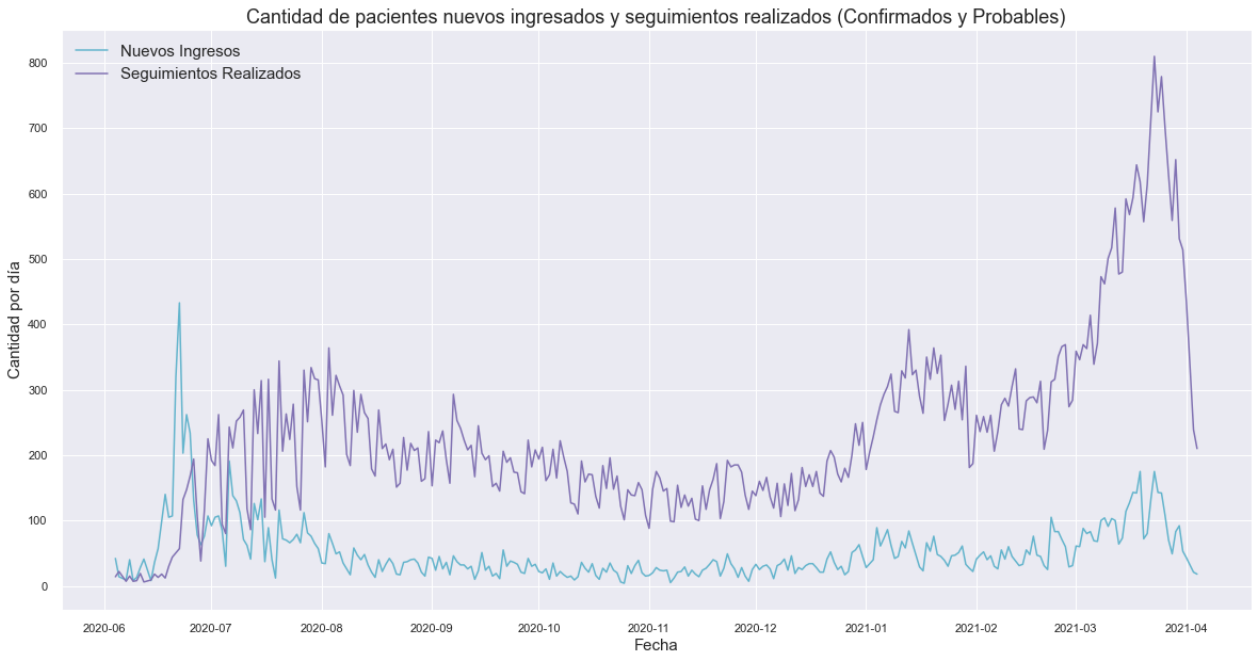
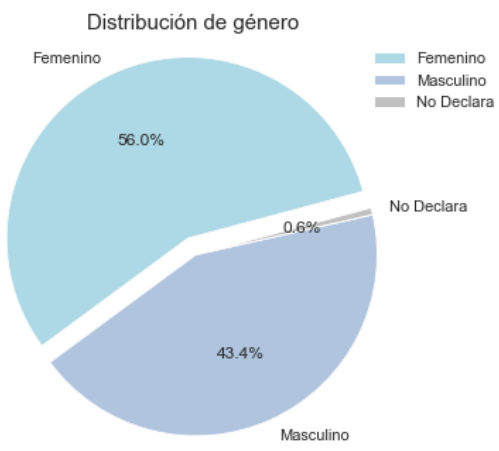


Figura 3.2: Cantidad diaria de nuevos ingresos a la plataforma y seguimientos realizados



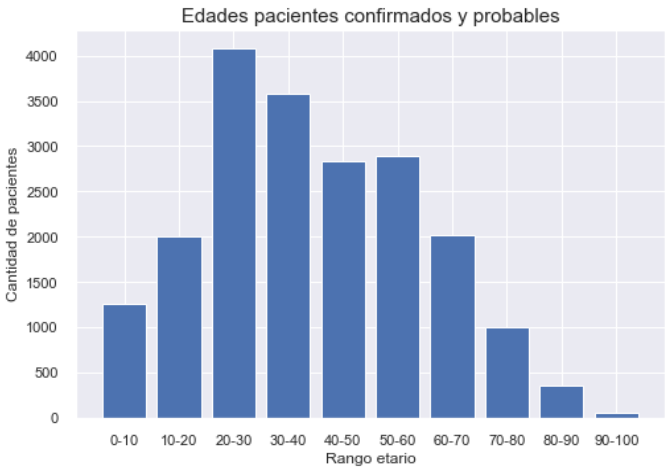


Figura 3.3 y 3.4: Distribución de género y de edades de los pacientes confirmados y probables

3.3 Preprocesamiento y limpieza

Con la base ya seleccionada, la primera fase de preprocesamiento y limpieza consiste, en primer lugar, eliminar los registros con edades negativas.

Luego de este paso, se procede a limpiar los datos asociados a la fecha de inicio de síntomas. En particular se tienen 1.647 registros en el que las fechas de inicio de síntomas tienen valores sin información, a modo de ejemplo: "100-02-13 00:00:00". Por simplicidad, estos valores se llevan a nulos y finalmente se procede a eliminar dichos registros de la base.

Adicionalmente con respecto a esta misma variable, hay 338 pacientes que tienen el inicio de la fecha de sus síntomas 1 mes o más antes de recién ser registrados en el sistema. Registros con estas características también quedan fuera del análisis.

Finalmente se realiza una limpieza general de la hoja de egresos, donde hay registros de pacientes que tienen los datos cambiados entre sus columnas. A modo de ejemplo, el campo de texto libre asociado al seguimiento colocado en la columna correspondiente a la causal de egreso. Estos registros también son filtrados de la base de datos.

3.4 Transformación de los datos

Con los datos seleccionados y la limpieza realizada, lo que viene a continuación es realizar todas las transformaciones que permitan aplicar los algoritmos para realizar la clusterización. En primer lugar, las nuevas variables definidas y transformaciones son las siguientes:

1. Discretización de edad: Al ser esta originalmente una variable numérica que varía entre 5 y 100 aproximadamente, es difícil ver como cada trayectoria sintomática distribuye en base a las edades. Bajo este pretexto, es que se discretiza llevándola a intervalos de 20 días.
2. Desenlace: Se tiene la variable de *Causal Egreso, Destino y Egreso Hospitalario*, donde cada una contiene parcialmente la información del desenlace de cada paciente. Combinando las tres variables, se logra definir cuál fue el estado final de cada paciente, siendo estos alta, hospitalización o fallecimiento por COVID-19. Alta se entiende como pacientes que no requirieron hospitalización, hospitalización como pacientes que fueron dados de alta del hospital y fallecimiento los pacientes que fallecieron, independiente de si hayan recibido o no hospitalización. Los pacientes que fallecieron por alguna causa externa al COVID-19 fueron extraídos de la muestra.

3. *One-Hot Encoding Síntomas:* Originalmente la variable *Síntomas* es una variable categórica con el nombre de un síntoma declarado que esté asociado a el IDMOVIMIENTO de un paciente. Bajo esta configuración, cada síntoma declarado en un seguimiento corresponde a una fila en la base de datos. A través de esta técnica esencialmente lo que se hace es cambiar la forma de la base de datos de modo que tenga menos filas, a cambio de incorporar más columnas. De este modo, cada síntoma pasa a ser una columna y si un paciente presenta determinado síntoma en un determinado seguimiento, la celda correspondiente tomaría el valor de 1.

IDINGRESO	IDMOVIMIENTO.1	Síntoma 1.1
		Síntoma 1.2
	IDMOVIMIENTO.2	Síntoma 2.1
		Síntoma 2.2
		Síntoma 2.3
	IDMOVIMIENTO.3	Síntoma 3.1

Tabla 3.2: Ejemplo de forma "Long"

		Anosmia	Cefalea	…	Tos
IDINGRESO	IDMOVIMIENTO.1	1	0	0	1
	IDMOVIMIENTO.2	1	1	0	1
	IDMOVIMIENTO.3	1	0	0	0

Tabla 3.3: Ejemplo de forma "Wide"

4. *One-Hot Encoding Antecedentes:* De forma análoga con la transformación de los síntomas, se realiza esto mismo con los registros de los antecedentes. La diferencia fundamental viene en que los antecedentes se registran una única vez al principio del seguimiento y estos no se van actualizando en cada seguimiento.
5. Cantidad de seguimientos: Variable definida para entender cuántos seguimientos se tienen por cada paciente, con el fin de poder estimar la cantidad de información que tendrán las trayectorias sintomáticas. En promedio son 4,53 seguimientos por paciente con una distribución estándar de 3,73.

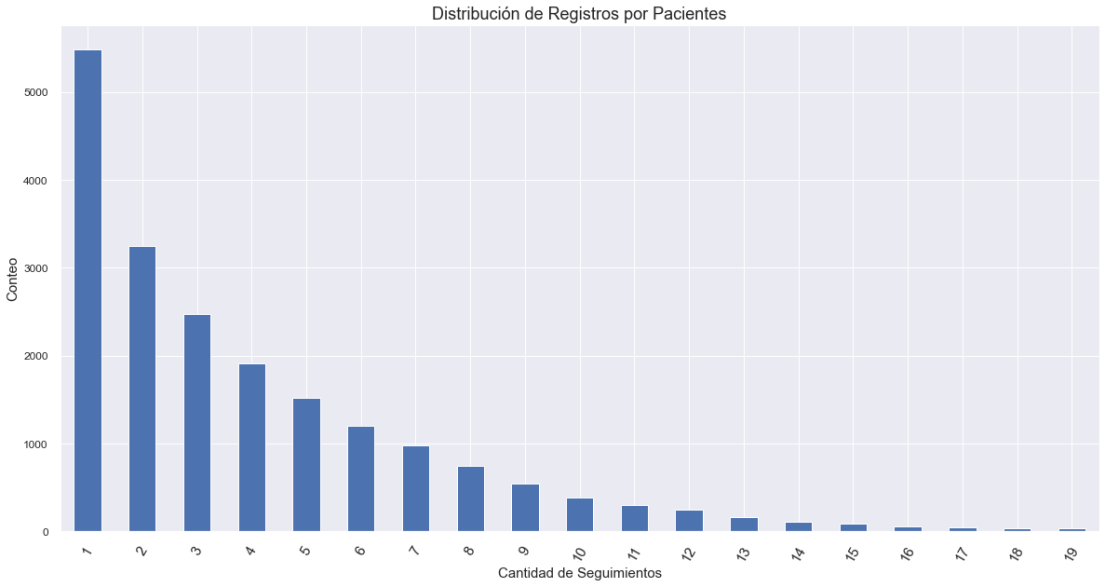


Figura 3.5: Distribución de la cantidad de registros sintomáticos para cada paciente

6. Eliminar pacientes con menos de 2 seguimientos: Los pacientes que tienen únicamente 1 registro sintomático aportan escasa información para generar trayectorias sintomáticas y son filtradas del modelo. Luego de este filtro, la cantidad de seguimientos sintomáticos por paciente aumenta a 5,46 y la desviación estándar queda en 3,66.
7. Días entre seguimientos: Variable definida para entender cada cuanto se realizan lo seguimientos, con el fin de clarificar cada cuanto se tiene una nueva actualización del cuadro sintomático de los pacientes. En promedio pasan 1,73 días entre los seguimientos, con una desviación estándar de 1,62 y una mediana de 1 día.

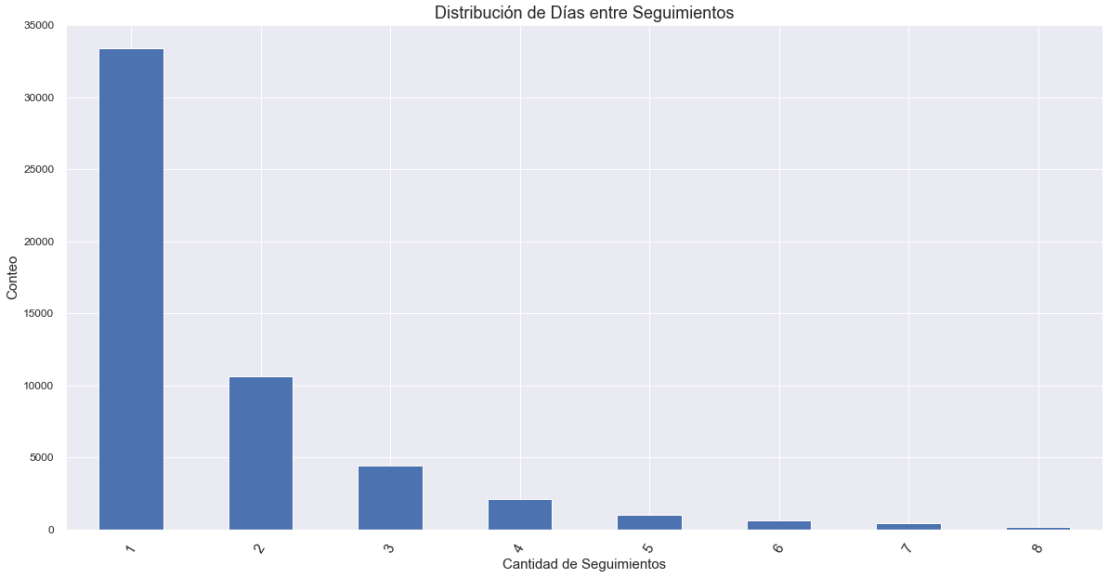


Figura 3.6: Distribución de días entre seguimientos

8. Diferencia entre Fecha de Seguimiento y Fecha de Inicio Síntomas: Variable definida para ubicar en la línea temporal a que día corresponde un determinado cuadro sintomático y poder ubicarla después en la trayectoria sintomática. De esta variable nacen dos subproductos, siendo el primero la diferencia entre la fecha del inicio de los síntomas y la fecha del último seguimiento con registro sintomático. De esta diferencia, nace lo que vendría a ser la duración de la trayectoria sintomática para los pacientes. Como segundo subproducto de esta variable, surge la variable correspondiente a la diferencia entre la fecha de ingreso (primer seguimiento) y la fecha de inicio síntomas. Esta variable es fundamental, dado que entrega información de cuanto se conoce de la trayectoria sintomática de un determinado paciente. Por ejemplo, si esta diferencia es 0, corresponde a que los síntomas del paciente se registraron justo en el día en el que estos comenzaron a manifestarse. No obstante, para valores mayores a 0, significaría que el primer registro sintomático comenzó después de la manifestación de estos, provocando que los primeros días no se tenga registro de los síntomas que presentó dicho paciente.

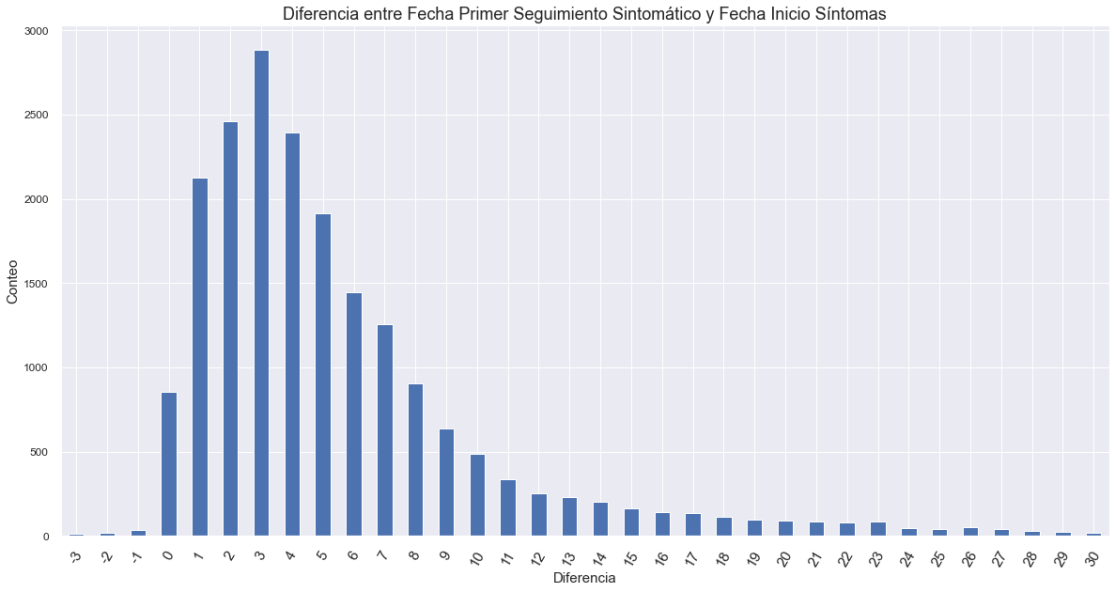


Figura 3.7: Distribución de la diferencia entre el primer registro sintomático y la fecha en la que inician los síntomas

9. Clasificación dependiente de la diferencia entre el primer seguimiento y la fecha de inicio de síntomas: Para los efectos de esta investigación la clasificación consistió en conservar a los pacientes que ingresan al seguimiento hasta 5 días después de que inicien sus síntomas. Es decir, si un paciente ingreso a los 6 o más días al seguimiento después de que hayan iniciado sus síntomas, este paciente queda fuera de la base de datos para la iteración del algoritmo. El motivo por el cual se consideran 5 días nace producto de interacciones con personal del SSMSO, donde se planteó que para los pacientes que ingresan días después de que sus síntomas ya iniciaron, se les preguntan cuáles fueron sus síntomas iniciales y estos quedan registrados en el primer registro. En base dicho punto, se tienen registrados los síntomas iniciales y se mantiene la esencia de la trayectoria que el paciente tuvo. Finalmente, tras esta medida se reduce considerablemente la cantidad de

datos a analizar, pasando a 10.057 IDs de ingreso distintas.

10. Variable de estado de vacunación: El estado de vacunación se mide al momento en el que la persona fue infectada por el virus. Dado que la fecha exacta de infección no está identificada, se utiliza la fecha de inicio de los síntomas como su reemplazo. Además de la fecha de inicio de los síntomas, esta variable depende de los siguientes 3 aspectos. En primer lugar, el tipo de vacuna suministrada, que en base a los datos disponibles puede corresponder a Pfizer o Sinovac. En segundo lugar, la última dosis suministrada y finalmente, cuantos días han pasado desde la última dosis. De esta forma, el estado de vacunación hace referencia a que tipo de protección tenía el paciente a la hora de haber contraído sus primeros síntomas, donde el caso ideal sería que hubiesen pasado más de 2 semanas desde la inoculación de la segunda dosis (o primera en caso de CanSino). Esta variable es relevante para la muestra de validación de mayo.

De los pacientes finalmente conservados se tienen las variables correspondientes a su género, desenlace, duración de síntomas y antecedentes, donde todas estas variables están separadas en los distintos rangos etarios previamente definidos. Con respecto a los desenlaces, los pacientes que no entran dentro de la categoría de hospitalizados ni fallecimientos son los pacientes que fueron dados de alta del seguimiento sin mayores inconvenientes. Con respecto a los antecedentes, no todos los pacientes declararon tener antecedentes, así como también existen pacientes que declararon tener más de uno.

La siguiente tabla contiene la información distribuida a lo largo de los rangos etarios definidos. Los p-valor señalados, para todas las variables a excepción de la duración de los síntomas, están calculados mediante tests de Chi-Cuadrado. Para la duración de los síntomas se implementa un Test-T. Los grupos que se comparan son los distintos rangos etarios, teniendo como un grupo a los pacientes pertenecientes a un rango etario y como el otro grupo a todos los pacientes que no pertenecen a dicho rango etario. De esta manera, se puede corroborar la presencia de diferencias estadísticamente significativas mientras mayor sea la edad. En este caso, un p-valor menor a 0.05 sugeriría que dicha variable difiere de los otros grupos de forma estadísticamente significativa con un 95% de confianza.

N = 10.057	Rango etario														
	0-19 N = 1503 (15%)			20-39 N = 4049 (40%)			40-59 N = 2881 (29%)			60-79 N = 1430 (14%)			80+ N = 192 (2%)		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Femenino	788	52%	0	2261	56%	0.077	1701	59%	0.007	849	59%	0.046	125	64%	0.039
Desenlace															
Hospitalizados	10	1%	0	71	2%	0	177	6%	0	197	14%	0	36	19%	0
Fallecimientos	0	0%	0	2	0%	0	10	0%	0.001	42	3%	0	31	16%	0
Duración Síntomas	**Std**			**Std**			**Std**			**Std**			**Std**		
Promedio + Std	7.7	4.1	0	8.9	4.7	0	10.2	6.3	0	11.2	8.2	0	11.8	9.5	0
Antecedentes															
Asma	77	5%	0	112	3%	0.001	84	3%	0.046	74	5%	0	6	3%	0.903
Cardiopatías	2	0%	0.003	1	0%	0	11	0%	0.005	50	3%	0	15	8%	0
Diabetes	15	1%	0	57	1%	0	264	9%	0	311	22%	0	50	26%	0
Embarazada	2	0%	0.03	52	1%	0	1	0%	0	0	0%	0.005	0	0%	0.581
Enfermedad Renal	0	0%	0.046	1	0%	0	8	0%	0.937	12	1%	0	8	4%	0
EPOC	0	0%	0.002	1	0%	0	10	0%	0.035	35	2%	0	17	9%	0
HTA	17	1%	0	98	2%	0	465	16%	0	604	42%	0	110	57%	0
Obesidad	13	1%	0.012	72	2%	0.497	50	2%	0.774	29	2%	0.288	3	2%	0.874
Tabaco	17	1%	0	220	5%	0	137	5%	0.099	50	3%	0.164	0	0%	0.006
Vacuna Influenza	129	9%	0.002	291	7%	0	293	10%	0.159	326	23%	0	55	28%	0

Tabla 3.4: Tabla de variables personales de pacientes junio 2020 a abril 2021

De la tabla, se obtiene que hay una asociación entre la mayoría de las variables y el rango etario. Las más notorias son el porcentaje de hospitalizados y fallecidos asociados a cada edad, así como también la duración de los síntomas y la presencia de antecedentes tales como diabetes, cardiopatías e hipertensión (HTA).

A continuación, para crear la base final sobre la cual actuará el algoritmo de minería de datos, como el objetivo es que sea no supervisado (es decir ver cómo se comportan las trayectorias sintomáticas sin condicionarlas con el desenlace, la edad ni los antecedentes de los pacientes), se crea un *subset* de la base de datos mencionada anteriormente que contenga únicamente los registros sintomáticos. La base completa transformada se volverá a utilizar en la fase de evaluación, interpretación y predicción una vez que se tengan los resultados de la minería. La creación de este nuevo *subset* consiste en los siguientes pasos:

- Filtrar variables: Conservar únicamente las variables asociadas a síntomas, IDINGRESO, IDMOVIMIENTO y la diferencia entre la fecha de seguimiento e inicio de síntomas, con el fin de saber en que momento se manifestó un determinado síntoma para un paciente dado.
- Discretización de diferencia entre Fecha de Seguimiento y Fecha de Inicio Síntomas: Los pacientes que tienen seguimientos sintomáticos más allá de los 19 días se agrupan en un solo intervalo final.
- Nueva disposición de base de datos: Para poder aplicar un algoritmo que detecte la trayectoria completa, se hace un nuevo *reshape* que permita tener en una fila toda la evolución sintomática de un paciente. Esto se puede

entender como que para cada fila se define un S(i,n,t) que puede tomar el valor de 1 si el individuo i presenta el síntoma *n* en el intervalo *t*, o el valor de 0 en su defecto. En este caso, n puede tomar 27 valores diferentes correspondiente a los 27 síntomas disponibles en la Plataforma COVID-19 SSMSO y en base a la discretización definida en el punto anterior, hay 20 unidades distintas de tiempos, que están conformadas por 1 día a excepción del intervalo final que es de 19 días o más.

IDINGRESO	S(1,1)	S(2,1)	S(3,1)	…	S(25,T)	S(26,T)	S(27,T)
A	0	1	1	…	1	0	1
B	1	0	1	…	0	0	0
C	0	0	1	…	0	0	0

Tabla 3.5: Ejemplo de matriz de datos definitiva

- Interpolación Lineal: En vista de que no existió un seguimiento para cada día, existen días en los que no se tiene información sobre el cuadro sintomático que padeció un determinado paciente. Para lidiar con este problema se utilizó la técnica de interpolación lineal para imputar los datos faltantes. De esta forma, se puede mantener la esencia de cómo fue la trayectoria sintomática. Cabe destacar que esta técnica también fue implementada en [32] para imputar síntomas de COVID-19 y en otros estudios sintomáticos, por ejemplo, para apoyar la predicción de síntomas tardíos de cáncer de cabeza y cuello [51].

ID	Anosmia 4	Anosmia 5	Anosmia 6	Anosmia 7	Anosmia 8
A	1	NaN	NaN	NaN	0

Tabla 3.6: Ejemplo de vector con valores nulos

ID	Anosmia 4	Anosmia 5	Anosmia 6	Anosmia 7	Anosmia 8
A	1	0.75	0.5	0.25	0

Tabla 3.7: Ejemplo de vector con datos interpolados

- Backward Filling: Dado que no todos los pacientes ingresaron a la plataforma el día exacto en el que comenzaron sus síntomas, hay ciertos pacientes que no tienen registros sintomáticos durante los primeros días. Para lidiar con este problema, se utilizó la técnica *backward filliing*, básicamente completando los registros sintomáticos faltantes con el primer registro sintomático siguiente. Todo esto bajo el supuesto administrativo de que, en teoría, los pacientes que ingresan hasta 5 días después de que inician sus síntomas a la plataforma se les pregunta tanto por sus síntomas actuales como por sus síntomas iniciales.

ID	Anosmia 1	Anosmia 2	Anosmia 3
A	NaN	NaN	1

Tabla 3.8: Ejemplo de vector con datos iniciales vacíos

ID	Anosmia 1	Anosmia 2	Anosmia 3
A	1	1	1

Tabla 3.9: Ejemplo de vector aplicando backward filling

- Missing Values futuros: Finalmente, cuando un paciente deja de presentar síntomas y es egresado de la plataforma, naturalmente deja de tener registros sintomáticos. A modo de ejemplo, si un paciente dejo de tener síntomas al décimo día, actualmente en la base de datos aparecerían solo valores vacíos después del décimo día. Estos valores vacíos se reemplazan por 0, indicando que el paciente no tuvo determinados síntomas en un determinado tiempo.

ID	Anosmia 15	Anosmia 16	Anosmia 17	Anosmia 18	Anosmia 19+
A	1	0	NaN	NaN	NaN

Tabla 3.10: Ejemplo de vector con datos finales vacíos

ID	Anosmia 15	Anosmia 16	Anosmia 17	Anosmia 18	Anosmia 19+
A	1	0	0	0	0

Tabla 3.11: Ejemplo de vector reemplazando por 0 valores finales

A través de estas transformaciones, la base resultante es una matriz de 10.057 filas correspondientes a pacientes y 540 columnas correspondientes a los 27 síntomas a lo largo de 20 intervalos de tiempo. Los valores de cada punto varían entre el 0 y el 1, por lo que no hace falta estandarizar o normalizar la base de datos. La notación formal de esto último corresponde a:

Conjunto I: Conjunto de i pacientes. Un total de 10.057.

Conjunto T: Conjunto de t días desde el inicio de los síntomas. Un total de 20.

Conjunto N: Conjunto de n síntomas. Un total de 27.

Luego, se define $S_{i,t,n}$ como 1 si el paciente i presenta en el tiempo t el síntoma n, o como 0 en el caso contrario.

$$S_{i,t,n} \in \{0,1\}, \qquad \forall i \in I, \qquad \forall t \in T, \qquad \forall n \in N$$

En base a esta definición, la notación formal de una trayectoria sintomática vendría a ser para un paciente $\hat{\imath}$, el conjunto de todos los $S_{\hat{\imath},t,n}$ asociados a dicho paciente.

$$TS_i = \{S_{i,0,1}, S_{i,0,2}, S_{i,0,3}, \dots, S_{i,19,26}, S_{i,19,27}\}, \qquad \forall i \in I$$

De forma exploratoria, se ve que el comportamiento sintomático de los 16 síntomas más reportados para los siguientes tipos de pacientes es el siguiente. Para efectos de estos gráficos, se entiende como desenlace leve los pacientes que fueron dados de alta del seguimiento sin haber requerido hospitalización, como pacientes hospitalizados los pacientes que requirieron hospitalización y fueron dados de alta de ella y finalmente los pacientes fallecidos son los que fallecieron, independiente de haber sido hospitalizados o no.

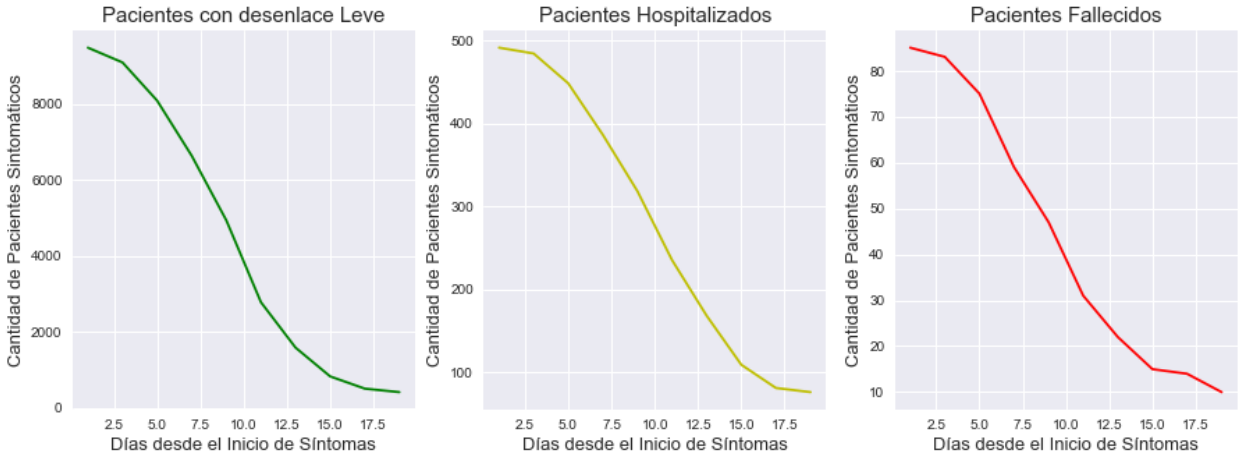


Figura 3.8: Cantidad de pacientes sintomáticos según la cantidad de días desde el inicio de los síntomas

Síntomas predominantes en pacientes con desenlace leve:

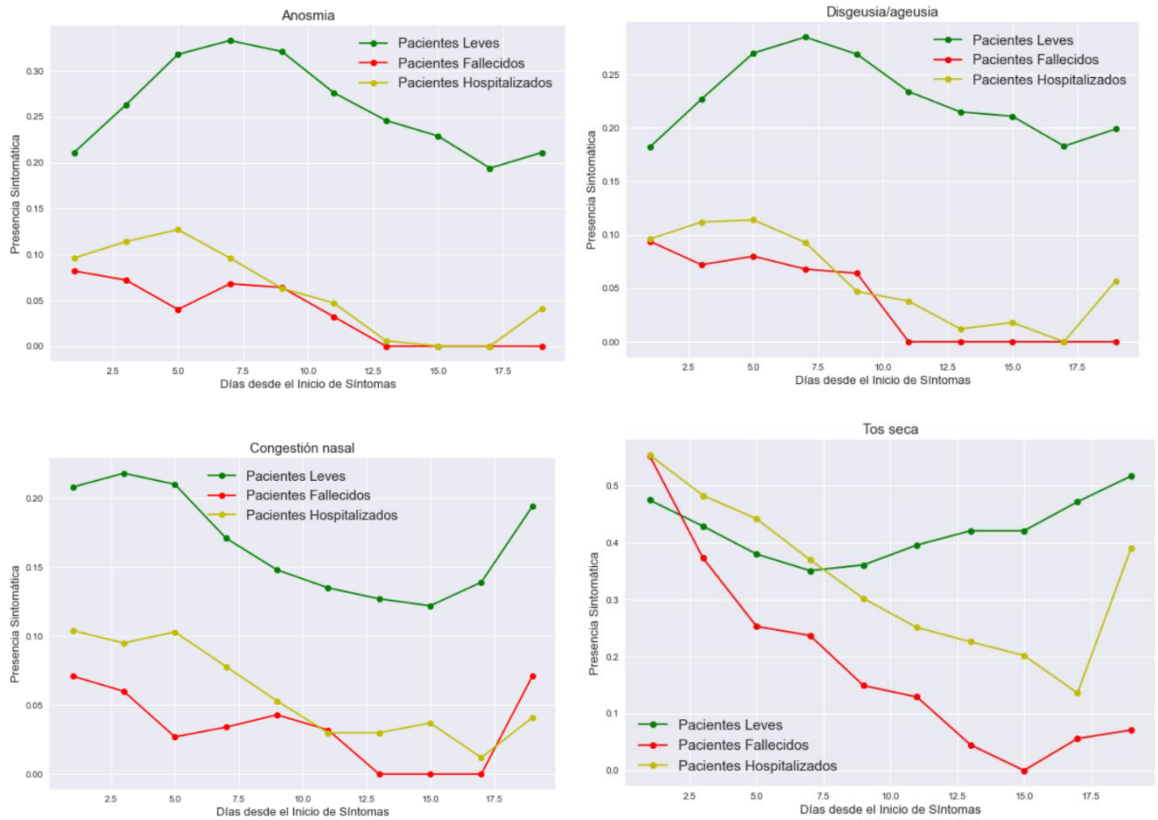


Figura 3.9: Porcentaje de personas sintomáticas que presentan Anosmia, Disgeusia/Ageusia, Congestión Nasal y Tos Seca

Síntomas predominantes en pacientes hospitalizados y fallecidos.

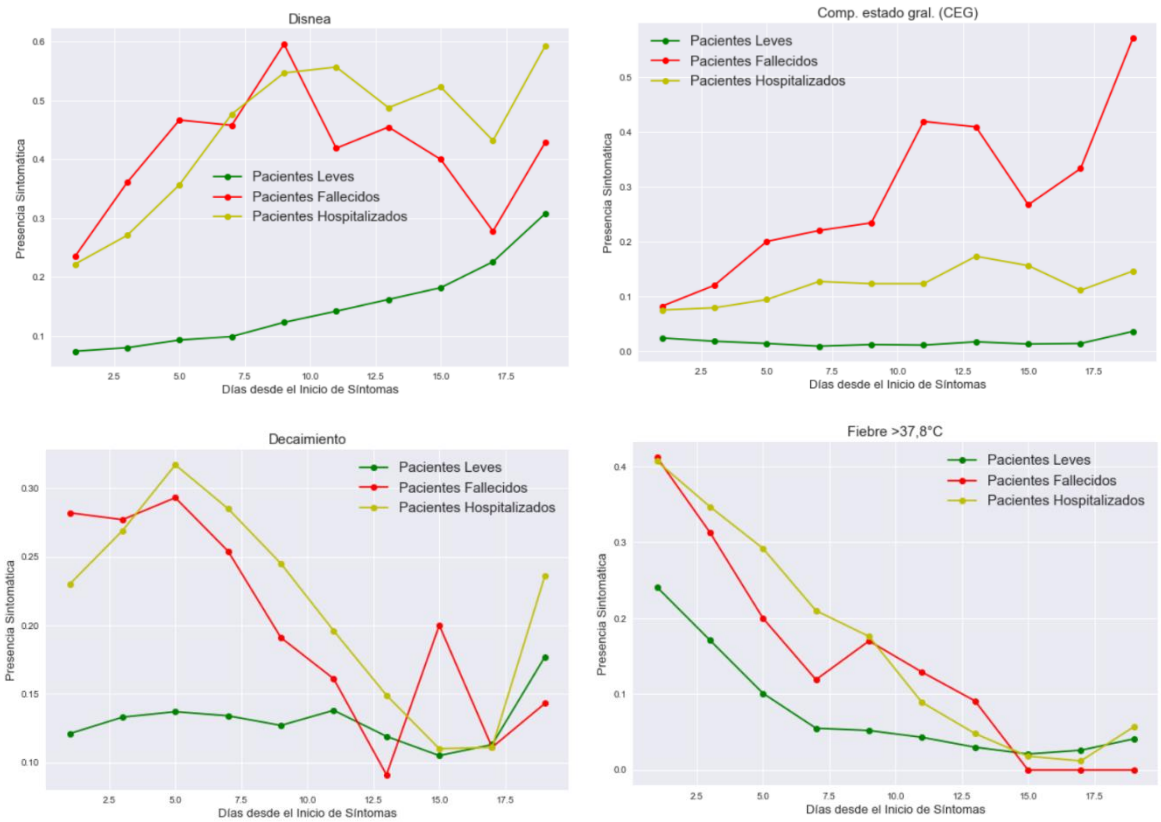


Figura 3.10: Porcentaje de personas sintomáticas que presentan Disnea, CEG, Decaimiento y Fiebre

Síntomas predominantes sin una asociación directa al riesgo:

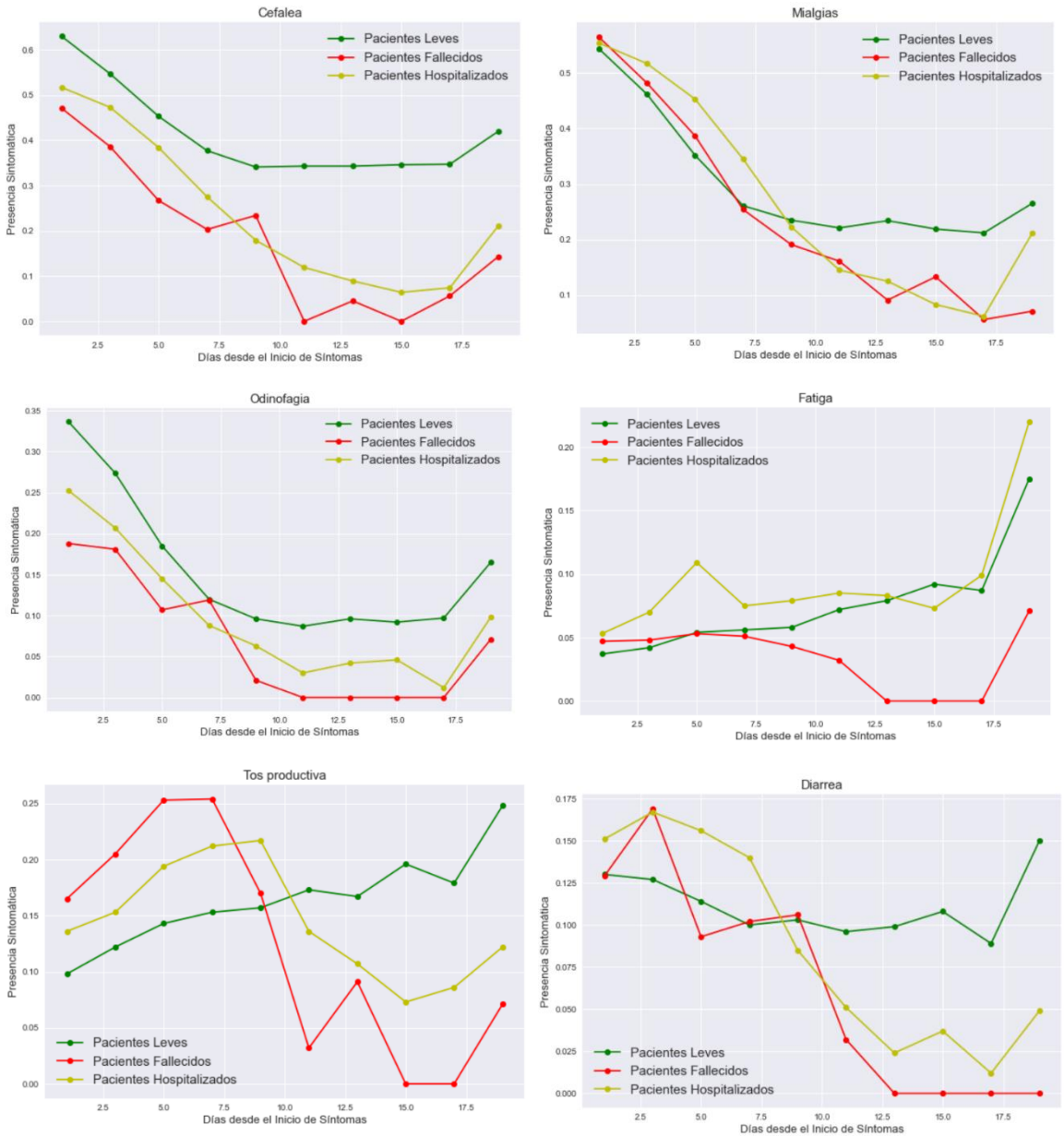


Figura 3.11: Porcentaje de personas sintomáticas que presentan Cefalea, Mialgias, Odinofagia, Fatiga, Tos Productiva y Diarrea

De acuerdo con los gráficos presentados, se observa que existen diferencias en cuanto a los síntomas que presentaron los pacientes en función del desenlace que estos tuvieron.

Es importante recalcar que existe una fuerte correlación entre presentar un síntoma y volver a tener dicho síntoma al día siguiente. A modo de ejemplo, se puede visualizar la matriz de correlación de presentar anosmia en los distintos intervalos de tiempo, donde Anosmia_t corresponde a presentar Anosmia en el día t desde el inicio de los síntomas:

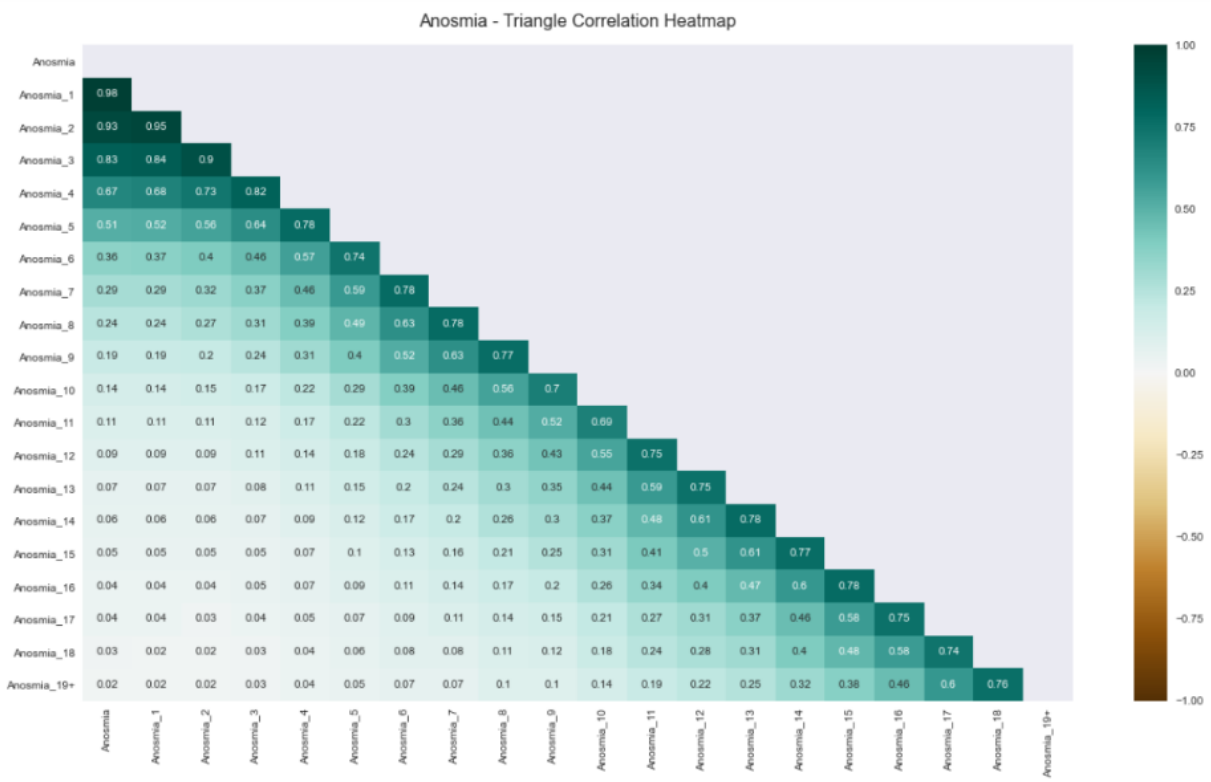


Figura 3.12: Matriz de correlación de la anosmia a través del tiempo

Un patrón similar se comparte para los otros 26 síntomas, por lo que el último paso de la transformación de datos consiste en la implementación de PCA, con el fin de tener únicamente variables no correlacionadas entre sí.

Aplicando PCA y manteniendo el 95% de la varianza de los datos se logra reducir de 540 a 188 variables no correlacionadas, lo que potencialmente mejora tanto la velocidad como el rendimiento de algún algoritmo de Machine Learning aplicado sobre los datos. Estas variables no correlacionadas, también conocidas como componentes principales, permiten representar la trayectoria sintomática de cada paciente. Formalmente esto se define como:

Conjunto P: Conjunto de p componentes principales. Un total de 188.

Luego, se define $PC_{i,p}$ como el valor que toma el componente principal p para un paciente i. Estos valores corresponden al conjunto de los números reales y fluctúan entre -3,26 y 6,07.

$$PC_{i,p} \in [-3.26, 6.07] \ \forall i \in I, \qquad \forall p \in P$$

En base a esta definición, la notación formal de una trayectoria sintomática tratada con PCA vendría a ser para un paciente $\hat{\imath}$, el conjunto de todos los $PC_{\hat{\imath},p}$ asociados a dicho paciente, siendo este el objeto a clusterizar.

$$TS\_PCA_i = \{PC_{i,1}, PC_{i,2}, PC_{i,3}, \dots, PC_{i,187}, PC_{i,188}\}, \qquad \forall i \in I$$

Finalmente, se aplica el Test de Hopkins sobre la base de datos interpolada. Esta prueba da un puntaje que determina la distribución que tiene una base de datos y señala si realizar clustering aporta valor [52]. Concretamente, valores cercanos a 0 implican que los datos no están distribuidos de forma uniforme, por lo que, en consecuencia, tendría sentido aplicar modelos de clustering para clasificar los datos. Si el puntaje está entre 0.3 y 0.5, se pueden aplicar modelos de clustering, pero la relevancia de estos clústers es cuestionable. Si el puntaje es superior a 0.5, no vale la pena aplicar algoritmos de clustering. El puntaje obtenido fue de **0.2**, por lo que se descarta la hipótesis de que los datos estén distribuidos de forma uniforme.

4 Resultados

En esta sección se verán los resultados de la etapa de minería de datos planteada en la metodología. En primer lugar, se mostrarán los resultados del clustering sintomático, luego los resultados asociados a un modelo predictivo para predecir el clúster sintomático definido en la primera parte y finalmente, si estos resultados son consistentes con los de otra iteración de clustering sobre una muestra independiente de pacientes de mayo del 2021.

4.1 Clustering

4.1.1 Cantidad de clústers y desempeño

Con la finalidad de poder encontrar trayectorias sintomáticas características, se procede a trabajar la base de datos transformada con los siguientes algoritmos no supervisados de *clusterización*:

- K-Modes
- K-Means
- Mezclas Gaussianas
- Hierarchical Clustering
- Binary Matrix Decomposition

Para validar la cantidad de clústeres a definir y evaluar cual algoritmo entrega los mejores resultados, se utilizaron las métricas de Calinski-Harabasz, Davies Bouldin y el Coeficiente de Siluetas

Adicionalmente, se corrieron los algoritmos con y sin la reducción de dimensionalidad, con el objetivo de poder cuantificar la mejora en las segmentaciones sintomáticas.

De forma análoga a PCA, se generó una base con el método de reducción de dimensionalidad denominado Multiple Correspondence Analysis (MCA), con el fin de poder comparar estas dos técnicas y conservar la que segmentara de mejor manera los grupos.

Recordando que se busca minimizar el índice de Davies-Bouldin y maximizar tanto Calinski-Harabasz como el coeficiente de siluetas, los mejores resultados para cada iteración son los siguientes:

			Métricas de desempeño		
Dimensionalidad	**Algoritmo**	**# Clústers**	**Calinski**	**Davies**	**Silhouette**
Sin trato	K-Modes	6	278,2	4,2	0,026
Sin trato	K-Means	6	287,2	4,6	0,014
Sin trato	GM	8	66,7	7,3	0,080
Sin trato	Hierarchical	7	24,4	4,2	-0,27
Sin trato	BMD	7	14,6	4,1	-0,006

PCA	K-Means	7	355,4	3,2	0,027
PCA	K-Means	6	392,8	3,3	0,027
PCA	GM	7	117,1	7,5	-0,01
PCA	GM	6	143,5	7,3	-0,02
PCA	Hierarchical	7	214,2	4,5	-0,087
PCA	Hierarchical	6	238,6	4,8	-0,088
MCA	K-Means	7	168,7	4,1	-0,079
MCA	K-Means	6	180,3	4,0	-0,077

Tabla 4.1: Comparación de desempeño de clustering

En primer lugar, cabe destacar que tanto K-Modes como Binary Matrix Decomposition solo fueron implementados sobre la base de datos sin reducción de dimensionalidad y aproximando todos los valores a cero o a uno, debido a que estos algoritmos están diseñados para trabajar con bases de datos de valores binarios. Tanto PCA como MCA generan nuevas variables, las cuales tienen valores continuos en vez de binarios.

En segundo lugar, en base a las métricas obtenidas, la mejor combinación fue la de implementar PCA para la reducción de dimensionalidad y luego el algoritmo K-Means sobre dicha base de datos. Considerando dicho caso, en los siguientes gráficos, se puede visualizar como varía cada métrica de desempeño en función de la cantidad de grupos en los que la base de datos se segmenta.

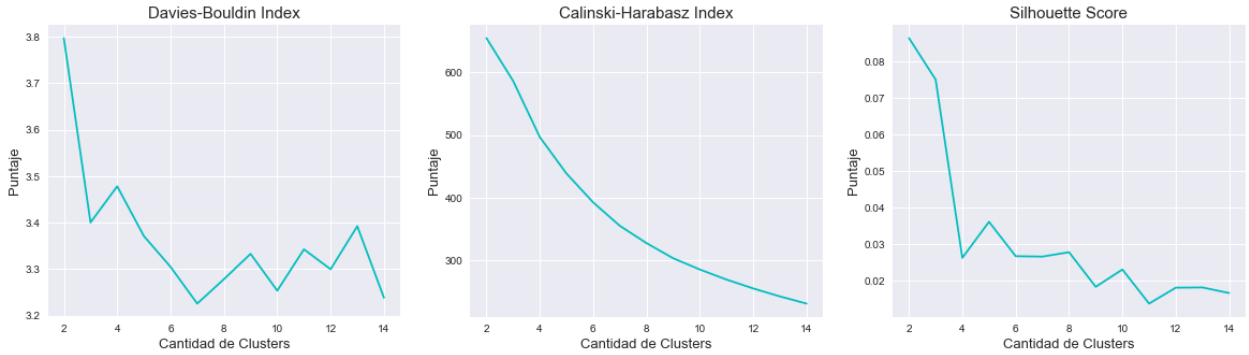


Figura 4.1: Métricas de desempeño de clustering

De estos 3 gráficos se desprende que la cantidad óptima de clústers está entre 6 y 7. Comparando estas dos configuraciones, la segmentación de la base de datos en 6 grupos tiene un mejor índice de Calinski-Harabasz mientras que si esta se segmenta en 7 se tiene un mejor índice de Davies-Bouldin. Ambos tienen un coeficiente de siluetas casi idéntico. Finalmente, para decidir cuál de las 2 configuraciones mantener, se tuvieron conversaciones con personal de la salud del SSMSO donde se evaluaron las trayectorias sintomáticas asociadas a los clústeres. De esta evaluación, se llegó al consenso de que era más valioso considerar 7 clústeres en vez de 6, básicamente porque la séptima trayectoria contiene la información de un grupo de pacientes que se caracterizan por presentar tos seca como síntoma predominante. Al considerar 6 clústeres, los pacientes que podrían haber sido asignados a la trayectoria sintomática de tos seca como síntoma predominante se asignan a otros grupos, sobre generalizando la trayectoria sintomática que estos pacientes tuvieron.

4.1.2 Trayectorias sintomáticas y caracterización

En esta sección se verán los síntomas presentados en cada clúster, encontrando así las trayectorias sintomáticas que tuvieron los distintos pacientes. Adicionalmente, se mostrarán las características personales de los pacientes que pertenecieron a cada trayectoria. Finalmente, se muestra como varían los síntomas dentro de un mismo clúster para los pacientes que tuvieron un desenlace crítico (hospitalización o fallecimiento) en contraste a los que no tuvieron complicaciones con su desenlace.

Las 7 trayectorias sintomáticas se pueden resumir de la siguiente forma:

- Una trayectoria caracterizada por cefalea y mialgias (18% de los pacientes)
- Una trayectoria caracterizada por cefalea y otros síntomas por pocos días (23.7% de los pacientes)
- Una trayectoria caracterizada por tos seca (13.7% de los pacientes)
- Dos trayectorias caracterizadas por cefalea, mialgias y tos seca, siendo una de larga duración y la otra de corta duración (7.6% y 17% de los pacientes respectivamente)
- Dos trayectorias caracterizadas por anosmia y disgeusia/ageusia, siendo una de larga duración y la otra de corta duración (8.4% y 11.6% de los pacientes respectivamente)

A continuación, se exponen en detalle los resultados asociados a cada trayectoria. Específicamente, se muestra un gráfico de calor para cada clúster los cuales facilitan la visualización de los distintos síntomas presentados en cada tiempo. Cada celda representa la proporción de pacientes que presentaron un determinado síntoma en un determinado día desde el inicio de sus síntomas. Mientras más intenso sea el color morado mayor es la proporción

Porcentaje de pacientes que reportaron tener un síntoma S en el tiempo T

Debajo de ambos gráficos hay una tabla que permite ver las principales variables de los pacientes asignados al clúster por rango etario, donde se destaca el desenlace que estos tuvieron, la duración promedio de los síntomas y los distintos antecedentes que estos presentan. Los p-valor se obtuvieron mediante tests de Chi-Cuadrado para todas las variables, a excepción de la duración de los síntomas que fue mediante un Test-T. Los grupos que se comparan son, por un lado, los pacientes pertenecientes al clúster y, por el otro lado, todos los pacientes que no pertenecen a dicho clúster. La finalidad de estos tests es para corroborar si los pacientes pertenecientes a un clúster tienen diferencias significativas sobre alguna variable en contraste con los pacientes pertenecientes los otros clúster.

Clúster 0: Cefalea y Mialgias

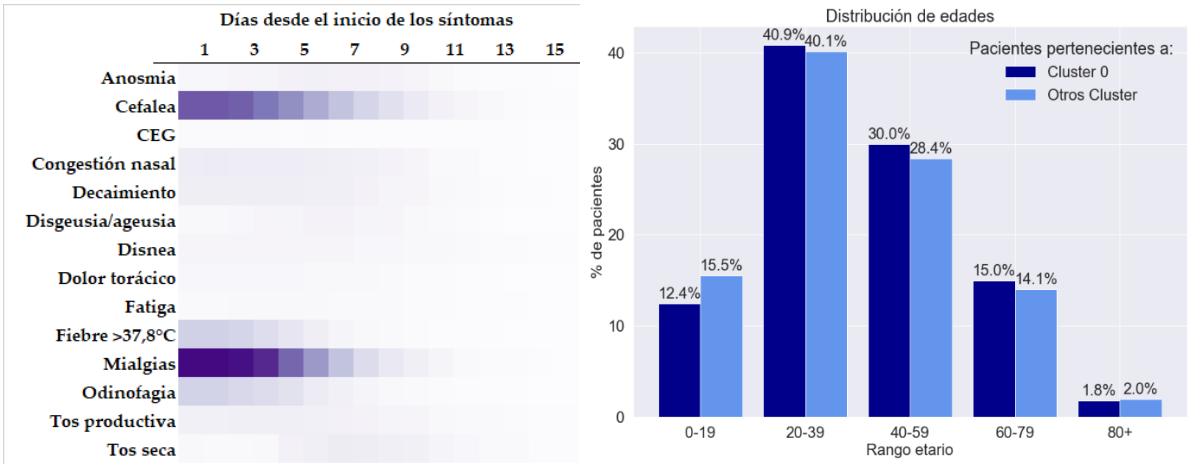


Figura 4.2: Gráfico de calor de los síntomas y distribución de los rangos etarios del Clúster 0

	Rango etario														
N = 1.809 (18,0%)	**0-19**			**20-39**			**40-59**			**60-79**			**80+**		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**225**	**12%**	**0.001***	**739**	**41%**	**0.59**	**542**	**30%**	**0.181**	**271**	**15%**	**0.324**	**32**	**2%**	**0.651**
F	115	51%	0.721	382	52%	0.013*	309	57%	0.308	163	60%	0.825	23	72%	0.447
Desenlace															
Hospitalizados	2	1%	0.998	13	2%	0.887	28	5%	0.341	33	12%	0.453	6	19%	0.827
Fallecimientos	0	0%	NaN	0	0%	0.805	1	0%	0.757	6	2%	0.56	6	19%	0.838
Duración Síntomas		**Std**			**Std**			**Std**			**Std**			**Std**	
Promedio + Std	7.6	3	0.714	8	3.4	0*	8.7	4	0*	9.6	5.7	0*	12.8	11.2	0.528
Antecedentes															
Asma	9	4%	0.506	23	3%	0.61	17	3%	0.843	8	3%	0.092	0	0%	0.584
Cardiopatías	1	0%	0.691	0	0%	0.411	1	0%	0.66	9	3%	0.993	1	3%	0.48
Diabetes	4	2%	0.362	10	1%	0.973	39	7%	0.093	58	21%	0.943	10	31%	0.58
Enfermedad Renal	0	0%	NaN	0	0%	0.411	2	0%	0.996	4	1%	0.365	1	3%	0.861
EPOC	0	0%	NaN	0	0%	0.411	5	1%	0.034*	6	2%	0.954	3	9%	0.835
HTA	4	2%	0.514	17	2%	0.919	79	15%	0.301	128	47%	0.075	15	47%	0.302
Obesidad	3	1%	0.665	8	1%	0.153	9	2%	0.973	3	1%	0.339	0	0%	0.994
Tabaco	6	3%	0.043*	48	6%	0.187	26	5%	0.951	7	3%	0.468	0	0%	NaN

Tabla 4.2: Tabla de variables personales de pacientes del Clúster 0

Trayectoria caracterizada por cefalea y mialgias. Secundariamente algunos de los pacientes asignados a este clúster presentaron fiebre y odinofagia los primeros días desde el inicio de sus síntomas. En comparación a los siguientes clústeres, no presenta diferencias estadísticamente significativas relacionadas al desenlace de los pacientes para los distintos rangos etarios. Tampoco presenta diferencias significativas relacionadas a los antecedentes que los pacientes declararon.

Clúster 1: Cefalea y otros síntomas

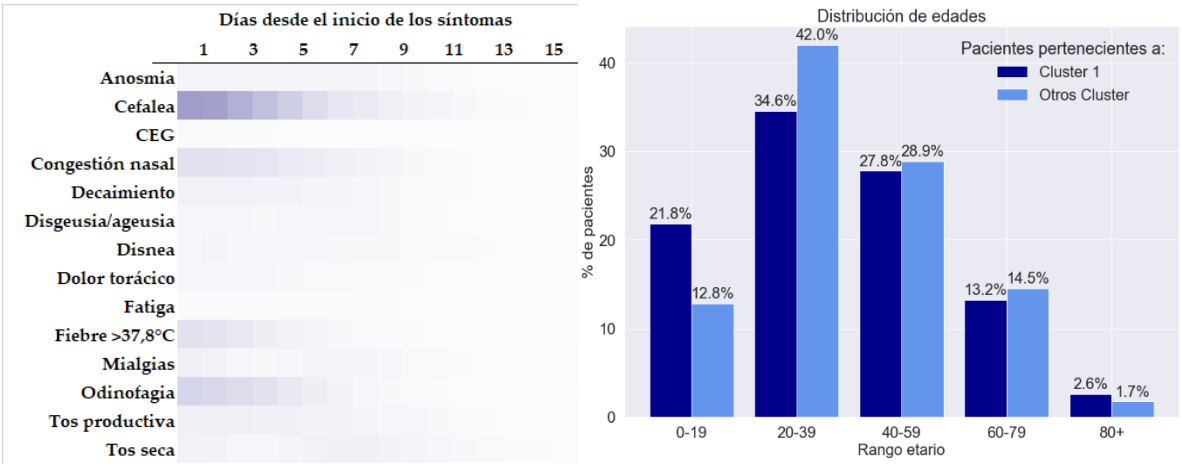


Figura 4.3: Gráfico de calor de los síntomas y distribución de los rangos etarios del Clúster 1

	Rango etario														
N = 2.393 (23,8%)	**0-19**			**20-39**			**40-59**			**60-79**			**80+**		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**522**	**22%**	**0***	**827**	**35%**	**0***	**665**	**28%**	**0.3**	**317**	**13%**	**0.127**	**62**	**3%**	**0.009***
F	286	55%	0.2	444	54%	0.174	390	59%	0.848	177	56%	0.165	40	65%	0.885
Desenlace															
Hospitalizados	6	1%	0.177	11	1%	0.373	43	6%	0.762	48	15%	0.479	16	26%	0.114
Fallecimientos	0	0%	NaN	0	0%	0.872	4	1%	0.37	10	3%	0.943	9	15%	0.864
Duración Síntomas		**Std**			**Std**			**Std**			**Std**			**Std**	
Promedio + Std	6.2	3.9	0*	6.3	3.7	0*	7.6	5.9	0*	9.1	7.1	0*	10.3	11	0.144
Antecedentes															
Asma	29	6%	0.666	17	2%	0.201	15	2%	0.307	18	6%	0.753	1	2%	0.71
Cardiopatías	0	0%	0.772	0	0%	0.463	4	1%	0.491	13	4%	0.624	4	6%	0.866
Diabetes	1	0%	0.043*	8	1%	0.298	53	8%	0.254	65	21%	0.595	15	24%	0.866
Enfermedad Renal	0	0%	NaN	1	0%	0.463	4	1%	0.165	3	1%	0.911	5	8%	0.132
EPOC	0	0%	NaN	0	0%	0.463	0	0%	0.174	4	1%	0.179	5	8%	0.971
HTA	2	0%	0.081	19	2%	0.896	114	17%	0.459	129	41%	0.571	30	48%	0.148
Obesidad	4	1%	0.993	12	1%	0.515	8	1%	0.303	10	3%	0.165	0	0%	0.567
Tabaco	3	1%	0.218	43	5%	0.805	32	5%	0.98	17	5%	0.061	0	0%	NaN

Table 4.3: Tabla de variables personales de pacientes del Clúster 1

Trayectoria caracterizada principalmente por la presencia de cefalea por un intervalo corto de tiempo. Este síntoma también suele ser acompañado secundariamente por congestión nasal, fiebre u odinofagia. Esta trayectoria para todos los rangos etarios a excepción del último tiene una duración más corta que la de los otros clúster, siendo esta diferencia en la duración estadísticamente significativa. Con respecto a las tasas de hospitalización y mortalidad se encuentra dentro del promedio y sin diferencias considerables, a excepción de la tasa de hospitalización de los pacientes con más de 80 años, donde presenta el mayor porcentaje de los grupos, aunque no presenta diferencias estadísticamente significativas.

Clúster 2: Tos Seca

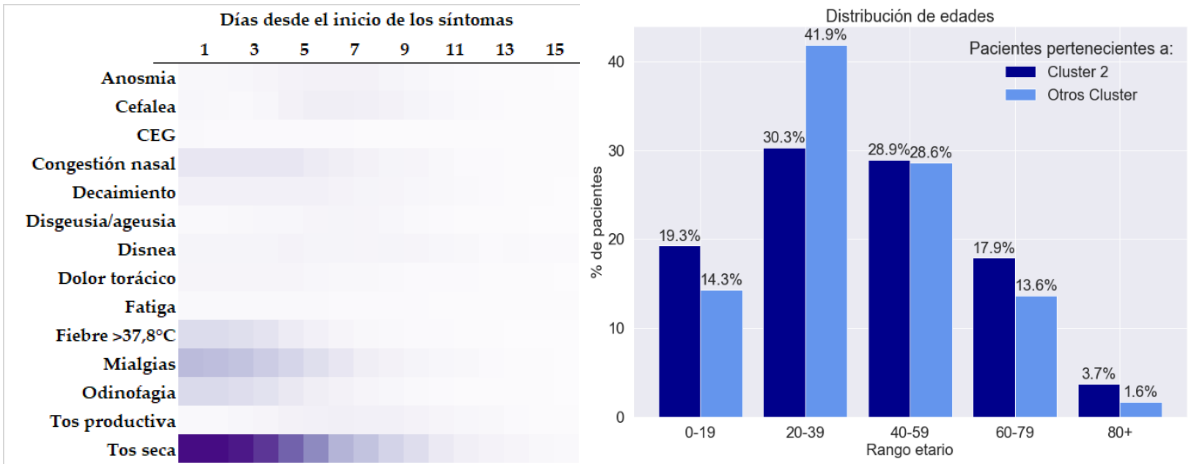


Figura 4.4: Gráfico de calor de los síntomas y distribución de los rangos etarios del Clúster 2

	Rango etario														
N = 1.381 (13,7%)	**0-19**			**20-39**			**40-59**			**60-79**			**80+**		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**266**	**19%**	**0***	**418**	**30%**	**0***	**399**	**29%**	**0.853**	**247**	**18%**	**0***	**51**	**4%**	**0***
F	119	45%	0.007*	196	47%	0*	206	52%	0.001*	136	55%	0.148	31	61%	0.643
Desenlace															
Hospitalizados	0	0%	0.291	12	3%	0.101	36	9%	0.014*	36	15%	0.765	9	18%	0.988
Fallecimientos	0	0%	NaN	1	0%	0.495	2	1%	0.916	8	3%	0.919	10	20%	0.548
Duración Síntomas		**Std**			**Std**			**Std**			**Std**			**Std**	
Promedio + Std	7.6	3.8	0.571	8.2	3.4	0.001*	9.7	5.9	0.084	10.5	9.7	0.139	10.6	6.6	0.291
Antecedentes															
Asma	20	8%	0.072	16	4%	0.215	14	4%	0.55	10	4%	0.471	2	4%	0.942
Cardiopatías	0	0%	0.787	0	0%	0.192	1	0%	0.984	10	4%	0.742	3	6%	0.787
Diabetes	5	2%	0.21	2	0%	0.138	41	10%	0.462	64	26%	0.097	12	24%	0.81
Enfermedad Renal	0	0%	NaN	0	0%	0.192	1	0%	0.688	1	0%	0.661	1	2%	0.621
EPOC	0	0%	NaN	1	0%	0.192	1	0%	0.916	8	3%	0.51	4	8%	0.986
HTA	3	1%	0.753	11	3%	0.898	66	17%	0.872	104	42%	0.98	35	69%	0.066
Obesidad	3	1%	0.884	14	3%	0.018*	12	3%	0.059	4	2%	0.801	1	2%	0.703
Tabaco	1	0%	0.335	23	6%	0.962	17	4%	0.709	9	4%	0.959	0	0%	NaN

Tabla 4.4: Tabla de variables personales de pacientes del Clúster 2

Trayectoria caracterizada por la tos seca como síntoma predominante. Algunos pacientes también manifiestan mialgias, odinofagia, fiebre o congestión nasal. Hay una mayor tasa de obesidad para el rango de 20 a 39 y el de 40 a 59 años, siendo estadísticamente significativa para el primero. Con respecto al rango de 40 a 59 años, se observa que el p-valor asociado a los pacientes hospitalizados es menor a 0.05. Adicionalmente, los pacientes que presentan hipertensión (HTA) del último rango etario también son superiores a la media, con un p-valor de 0.066 cercano a 0.05.

Clúster 3: Cefalea, Mialgias y Tos Seca – Larga Duración

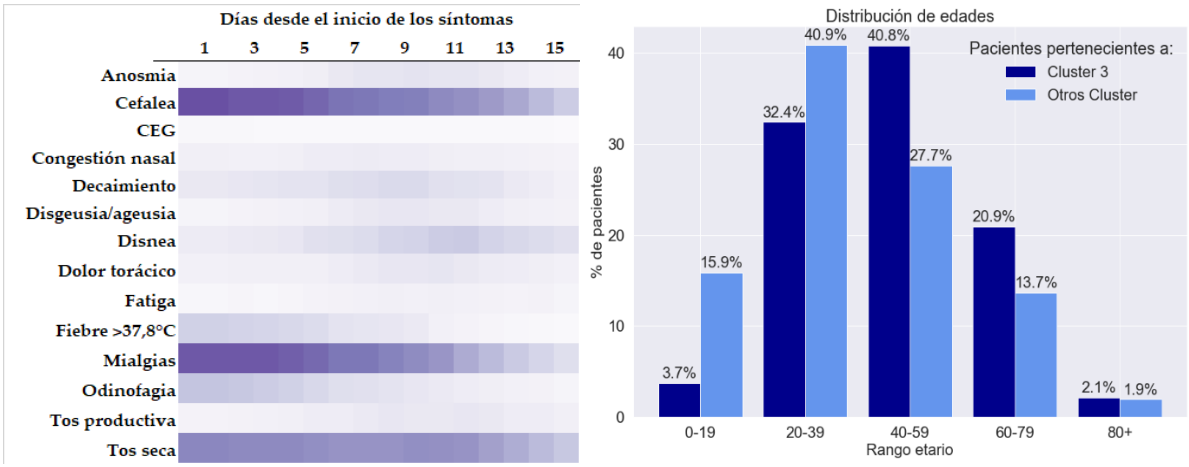


Figura 4.5: Gráfico de calor de los síntomas y distribución de los rangos etarios del Clúster 3

	Rango etario														
N = 759 (7,6%)	**0-19**			**20-39**			**40-59**			**60-79**			**80+**		
Variables	**N**	**%**	**p valor**	**N**	**%**	**P valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**28**	**4%**	**0***	**246**	**32%**	**0***	**310**	**41%**	**0***	**159**	**21%**	**0***	**16**	**2%**	**0.814**
F	12	43%	0.405	169	69%	0*	199	64%	0.059	102	64%	0.224	9	56%	0.659
Desenlace															
Hospitalizados	1	4%	0.462	11	4%	0.002*	26	8%	0.106	32	20%	0.019*	3	19%	0.753
Fallecimientos	0	0%	NaN	0	0%	0.262	1	0%	0.665	4	3%	0.933	2	12%	0.968
Duración Síntomas		**Std**			**Std**			**Std**			**Std**			**Std**	
Promedio + Std	18	4.9	0*	17.8	7.7	0*	19.3	8	0*	20.3	9.1	0*	22.2	9.3	0*
Antecedentes															
Asma	1	4%	0.955	9	4%	0.496	8	3%	0.847	13	8%	0.105	1	6%	0.994
Cardiopatías	0	0%	0.015*	0	0%	0.066	2	1%	0.758	3	2%	0.346	4	25%	0.027*
Diabetes	1	4%	0.672	10	4%	0.001*	39	13%	0.035*	37	23%	0.695	7	44%	0.156
Enfermedad Renal	0	0%	NaN	0	0%	0.066	1	0%	0.68	1	1%	0.879	0	0%	0.834
EPOC	0	0%	NaN	0	0%	0.066	1	0%	0.665	5	3%	0.74	2	12%	0.928
HTA	3	11%	0*	9	4%	0.276	64	21%	0.028*	70	44%	0.69	13	81%	0.071
Obesidad	0	0%	0.595	4	2%	0.95	4	1%	0.685	3	2%	0.869	1	6%	0.593
Tabaco	1	4%	0.741	12	5%	0.801	18	6%	0.436	4	3%	0.628	0	0%	NaN

Table 4.5: Tabla de variables personales de pacientes del Clúster 3

Trayectoria caracterizada por presentar principalmente cefalea, mialgias y tos seca por una larga duración. Fiebre y odinofagia suelen verse al comienzo, así como el decaimiento y la disnea se manifiestan después de la primera semana en un 20% de los casos aproximadamente. Con respecto al desenlace, se destaca el p-valor asociado a los pacientes hospitalizados de los rangos etarios de 20 a 39 y de 60 a 79 años los cuales son de 0.002 y de 0.019 respectivamente. Con respecto a los antecedentes, se destacan la diabetes y la hipertensión, presentando p-valores menores a 0.05 en distintos intervalos. Adicionalmente, a este grupo sintomático se asocian considerablemente más pacientes de 40 a 59 años.

Clúster 4: Anosmia, Disgeusia/Ageusia – Tardía

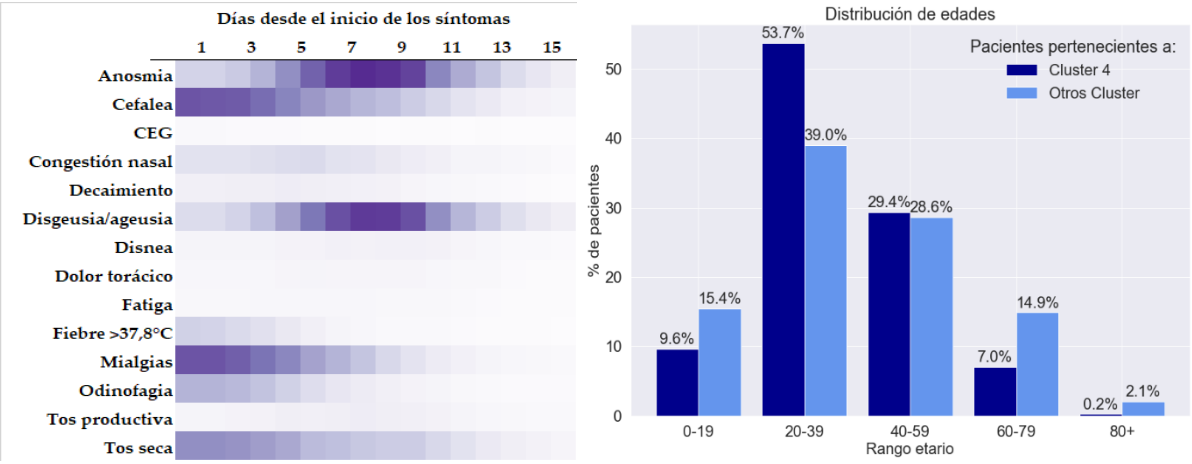


Figura 4.6: Gráfico de calor de los síntomas y distribución de los rangos etarios del Clúster 4

	Rango etario														
N = 841 (8,4%)	**0-19**			**20-39**			**40-59**			**60-79**			**80+**		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**81**	**10%**	**0***	**452**	**54%**	**0***	**247**	**29%**	**0.657**	**59**	**7%**	**0***	**2**	**0%**	**0***
F	44	54%	0.813	283	63%	0.002*	153	62%	0.367	46	78%	0.005*	2	100%	0.754
Desenlace															
Hospitalizados	0	0%	0.956	2	0%	0.039*	3	1%	0.001*	5	8%	0.311	0	0%	0.814
Fallecimientos	0	0%	NaN	0	0%	0.534	0	0%	0.686	1	2%	0.855	0	0%	0.726
Duración Síntomas		**Std**			**Std**			**Std**			**Std**			**Std**	
Promedio + Std	12.3	4.1	0*	12.2	3.5	0*	13	4.4	0*	13.5	5.2	0.031*	15	1.4	0.631
Antecedentes															
Asma	1	1%	0.17	14	3%	0.762	8	3%	0.906	2	3%	0.74	0	0%	0.072
Cardiopatías	0	0%	0.219	0	0%	0.217	2	1%	0.548	2	3%	0.752	0	0%	0.358
Diabetes	0	0%	0.723	7	2%	0.954	16	6%	0.157	13	22%	0.915	0	0%	0.98
Enfermedad Renal	0	0%	NaN	0	0%	0.217	0	0%	0.814	1	2%	0.994	0	0%	0.136
EPOC	0	0%	NaN	0	0%	0.217	0	0%	0.686	0	0%	0.417	1	50%	0.414
HTA	1	1%	0.653	11	2%	0.886	31	13%	0.13	27	46%	0.671	1	50%	0.6
Obesidad	1	1%	0.805	10	2%	0.581	3	1%	0.689	2	3%	0.775	0	0%	0.007*
Tabaco	1	1%	0.653	22	5%	0.65	12	5%	0.939	1	2%	0.684	0	0%	NaN

Tabla 4.6: Tabla de variables personales de pacientes del Clúster 4

Trayectoria caracterizada por comenzar con cefalea, mialgias y tos seca para luego manifestar anosmia y disgeusia/ageusia. Hay diferencias significativas en la cantidad de pacientes asociados a los rangos etarios, siendo mayoritariamente mujeres jóvenes entre 20 a 39 años. En relación con los desenlaces, el p-valor asociado a la hospitalización es menor a 0.05 para los pacientes entre 20 y 59 años. Adicionalmente, esta trayectoria es estadísticamente significativa más larga que el promedio de las otras trayectorias para todos los rangos etarios a excepción del último. No se aprecian diferencias importantes relacionadas a los antecedentes

Clúster 5: Anosmia, Disgeusia/Ageusia – Temprana

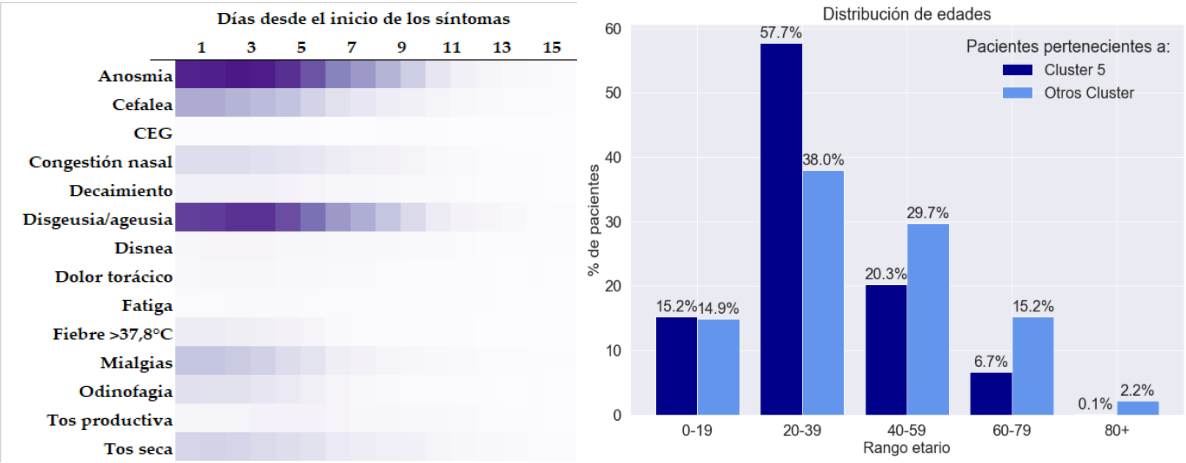


Figura 4.7: Gráfico de calor de los síntomas y distribución de los rangos etarios del Clúster 5

	Rango etario														
N = 1.164 (11,6%)	**0-19**			**20-39**			**40-59**			**60-79**			**80+**		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**177**	**15%**	**0.824**	**672**	**58%**	**0***	**236**	**20%**	**0***	**78**	**7%**	**0***	**1**	**0%**	**0***
F	101	57%	0.217	405	60%	0.013*	162	69%	0.002*	47	60%	0.964	1	100%	0.762
Desenlace															
Hospitalizados	1	1%	0.751	7	1%	0.168	6	3%	0.024*	4	5%	0.035*	0	0%	0.417
Fallecimientos	0	0%	NaN	0	0%	0.749	0	0%	0.712	3	4%	0.885	0	0%	0.352
Duración Síntomas		**Std**			**Std**			**Std**			**Std**			**Std**	
Promedio + Std	8.2	3.5	0.073	8.6	3.2	0.074	9.4	4	0.053	10.3	4.5	0.285	15	NaN	0.632
Antecedentes															
Asma	4	2%	0.097	13	2%	0.19	6	3%	0.878	3	4%	0.778	1	100%	0.007*
Cardiopatías	0	0%	0.562	0	0%	0.369	0	0%	0.659	3	4%	0.885	0	0%	0.113
Diabetes	1	1%	0.83	11	2%	0.709	21	9%	0.976	11	14%	0.123	1	100%	0.579
Enfermedad Renal	0	0%	NaN	0	0%	0.369	0	0%	0.841	0	0%	0.844	0	0%	0.021*
EPOC	0	0%	NaN	0	0%	0.369	1	0%	0.712	4	5%	0.231	0	0%	0.144
HTA	1	1%	0.704	13	2%	0.447	34	14%	0.507	27	35%	0.199	1	100%	0.892
Obesidad	1	1%	0.979	14	2%	0.62	2	1%	0.406	3	4%	0.448	0	0%	0*
Tabaco	1	1%	0.704	36	5%	0.998	15	6%	0.295	4	5%	0.624	0	0%	NaN

Tabla 4.7: Tabla de variables personales de pacientes del Clúster 5

Trayectoria caracterizada por presentar anosmia junto a disgeusia/ageusia desde el comienzo. Los p-valores asociados a la cantidad de pacientes son menores a 0.05 para todos los rangos etarios a excepción del primero. También se presenta una mayor tasa de mujeres entre los 20 a 59 años. Con respecto al desenlace, se aprecian p-valores menores a 0.05 para pacientes hospitalizados entre 40 a 79 años. En relación con los antecedentes, no hay diferencias considerables.

Clúster 6: Cefalea, Tos Seca y Mialgias – Corta Duración

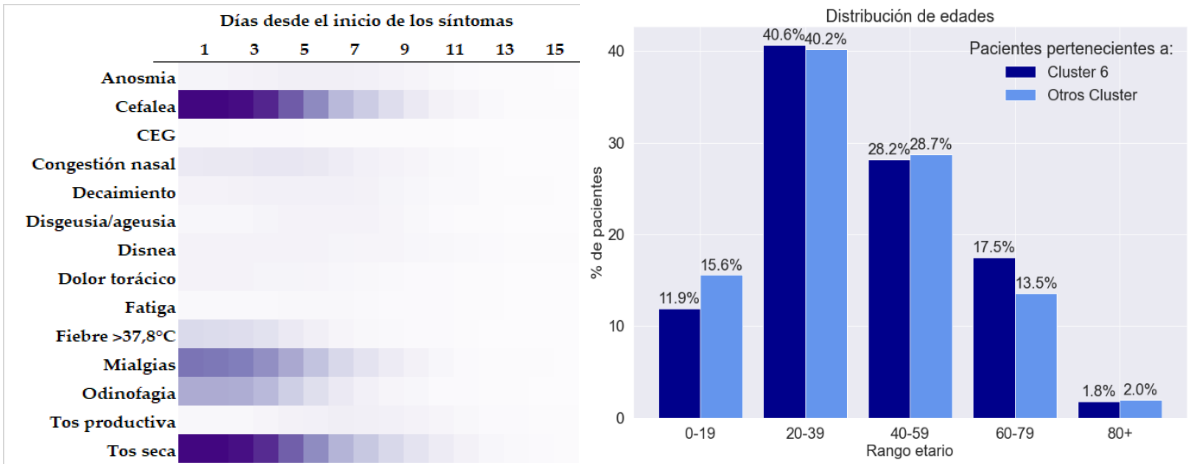


Figura 4.8: Gráfico de calor de los síntomas y distribución de los rangos etarios del Clúster 6

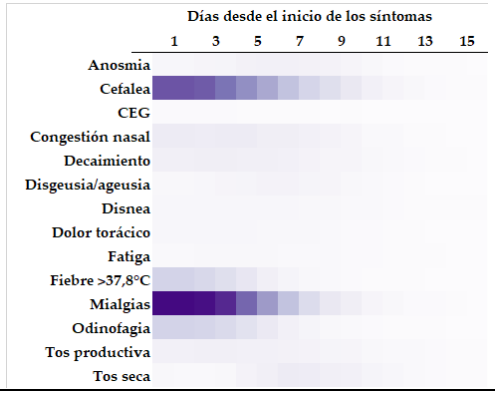
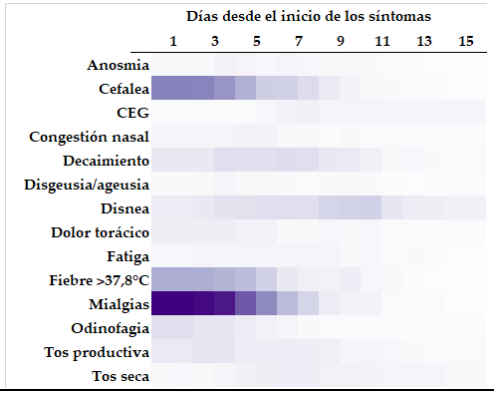
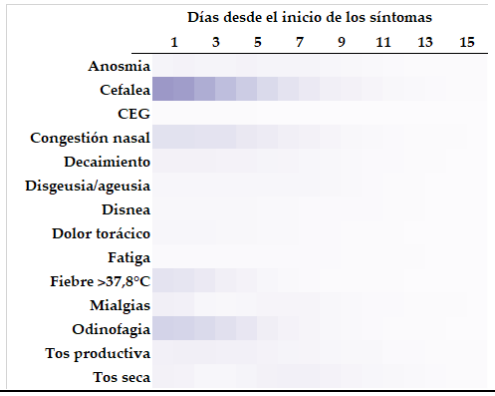
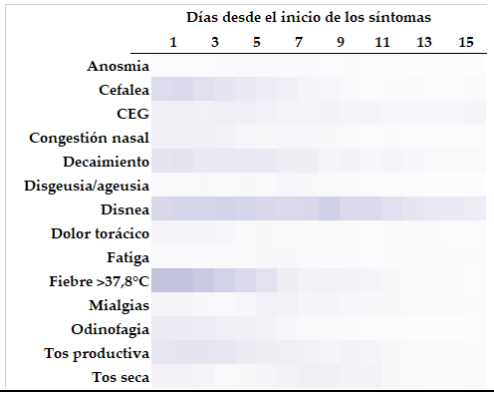
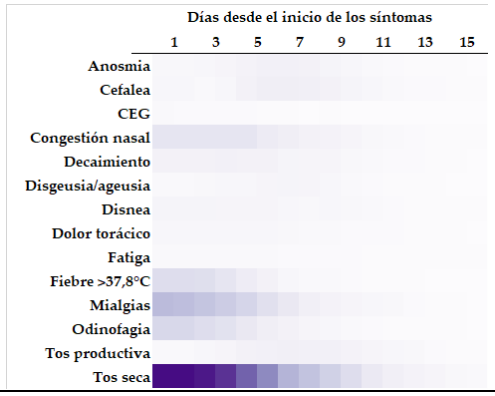
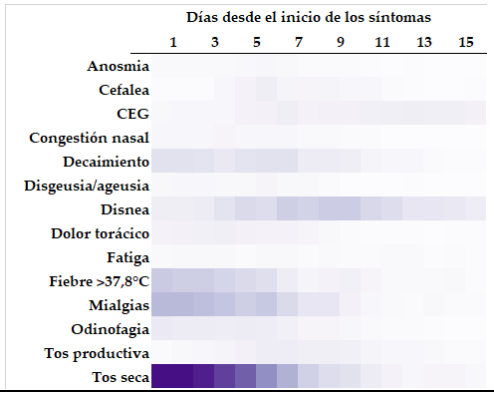
	Rango etario														
N = 1.710 (17,0%)	**0-19**			**20-39**			**40-59**			**60-79**			**80+**		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**204**	**12%**	**0***	**695**	**41%**	**0.744**	**482**	**28%**	**0.666**	**299**	**17%**	**0***	**30**	**2%**	**0.631**
F	111	54%	0.593	382	55%	0.639	282	59%	0.833	178	60%	0.998	19	63%	0.944
Desenlace															
Hospitalizados	0	0%	0.427	15	2%	0.463	35	7%	0.31	39	13%	0.75	2	7%	0.117
Fallecimientos	0	0%	NaN	1	0%	0.769	2	0%	0.883	10	3%	0.782	4	13%	0.873
Duración Síntomas		**Std**			**Std**			**Std**			**Std**			**Std**	
Promedio + Std	8	3	0.268	8.3	3.2	0.001*	9.1	3.5	0*	10.5	7.3	0.074	9.9	4.5	0.24
Antecedentes															
Asma	13	6%	0.484	20	3%	0.944	16	3%	0.668	20	7%	0.237	1	3%	0.624
Cardiopatías	1	0%	0.637	1	0%	0.384	1	0%	0.783	10	3%	0.987	3	10%	0.893
Diabetes	3	1%	0.725	9	1%	0.92	55	11%	0.074	63	21%	0.81	5	17%	0.311
Enfermedad Renal	0	0%	NaN	0	0%	0.384	0	0%	0.426	2	1%	0.995	1	3%	0.793
EPOC	0	0%	NaN	0	0%	0.384	2	0%	0.883	8	3%	0.939	2	7%	0.928
HTA	3	1%	0.891	18	3%	0.854	77	16%	0.968	119	40%	0.371	15	50%	0.545
Obesidad	1	0%	0.83	10	1%	0.558	12	2%	0.231	4	1%	0.471	1	3%	0.954
Tabaco	4	2%	0.396	36	5%	0.816	17	4%	0.204	8	3%	0.489	0	0%	NaN

Tabla 4.8: Tabla de variables personales de pacientes del Clúster 6

Trayectoria caracterizada por cefalea, tos seca y mialgias. Nuevamente la fiebre y la odinofagia como síntomas secundarios durante los primeros días. No se observan diferencias estadísticamente significativas con relación al desenlace de los pacientes ni de los antecedentes. Hay menos pacientes en el rango de 0 a 19 años y más pacientes en el rango de 60 a 79 años que el promedio.

4.1.3 Trayectorias con desenlaces críticos

En esta sección se verán las diferencias sintomáticas dentro de los clústeres que hay entre los pacientes que tuvieron un desenlace leve y los que tuvieron un desenlace crítico, es decir, que requirieron hospitalización o fallecieron

Cluster	Pacientes con desenlace leve	Pacientes con desenlace crítico
Clúster 0		
Clúster 1		
Clúster 2		

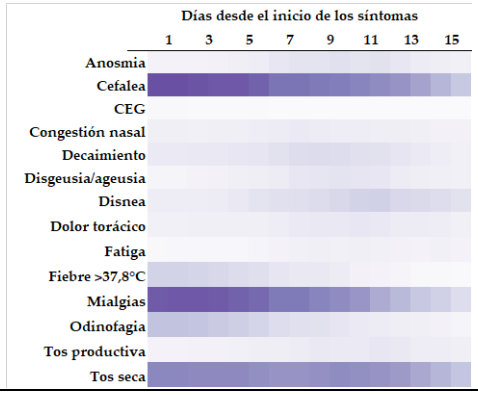
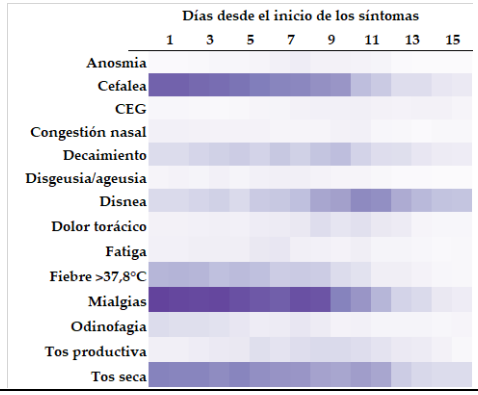
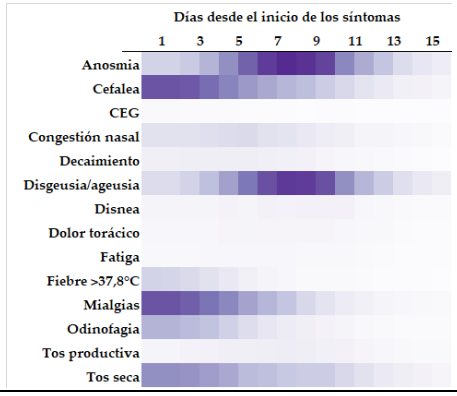
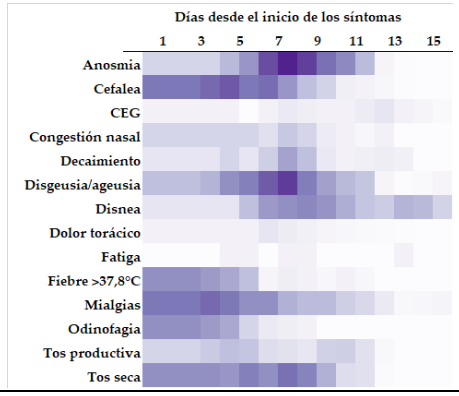
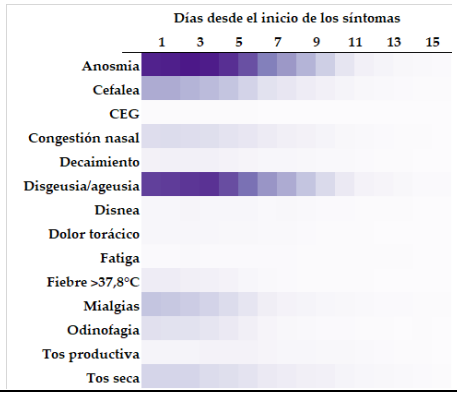
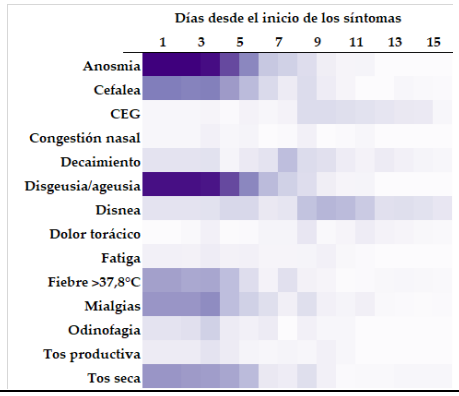
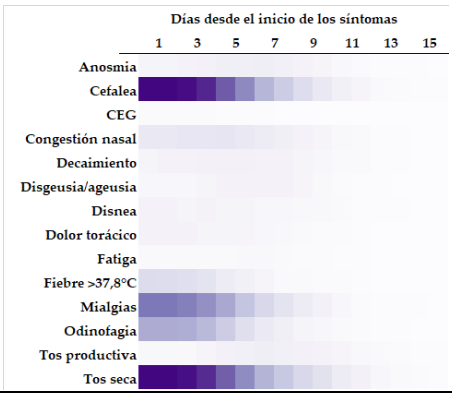
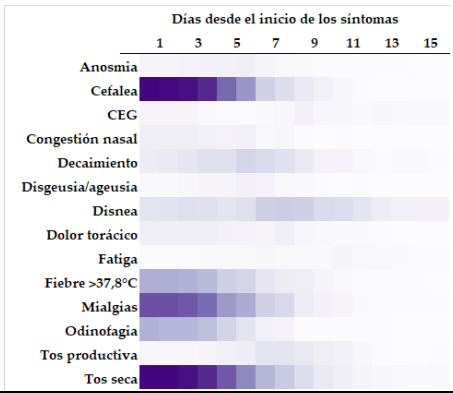
Clúster 3		
Clúster 4		
Clúster 5		
Clúster 6		

Tabla 4.9: Comparación de trayectorias sintomáticas dentro de un mismo clúster entre pacientes con desenlace leve y pacientes que requirieron hospitalización o fallecieron

		Cluster (Normalizado en columnas)						
	Desenlace	0	1	2	3	4	5	6
0.8%	Fallecimiento	1%	1%	2%	1%	0%	0%	1%
4.9%	Hospitalizado	5%	5%	7%	10%	1%	2%	5%
94.3%	Alta	94%	94%	92%	89%	99%	98%	94%

Tabla 4.10: Distribución de desenlaces por cluster

		Cluster (Normalizado en filas)						
	Desenlace	0	1	2	3	4	5	6
0.8%	Fallecimiento	15%	27%	25%	8%	1%	4%	20%
4.9%	Hospitalizado	17%	25%	19%	15%	2%	4%	19%
94.3%	Alta	18%	24%	13%	7%	9%	12%	17%

Tabla 4.11: Distribución de clústeres por desenlace

De la comparación sintomática dentro de los mismos clústeres se puede ver que hay diferencias particularmente en lo que respecta al compromiso del estado general (CEG), al decaimiento, disnea y fiebre mayoritariamente. Este mismo patrón se comparte en todos los clústeres. Con el fin de cuantificar de manera más precisa esta diferencia, se genera una tabla con el promedio de los puntos porcentuales entre la versión riesgosa de un clúster y su versión leve. El punto porcentual es básicamente la diferencia aritmética entre dos porcentajes.

	Días desde el comienzo de los síntomas														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Anosmia	-0.01	-0.01	-0.02	-0.04	-0.07	-0.09	-0.09	-0.07	-0.08	-0.09	-0.05	-0.04	-0.09	-0.07	-0.06
Cefalea	-0.05	-0.04	-0.04	0.00	0.00	0.00	-0.01	-0.03	-0.03	-0.04	-0.08	-0.08	-0.08	-0.06	-0.05
CEG	0.04	0.05	0.05	0.04	0.04	0.04	0.06	0.08	0.12	0.10	0.10	0.10	0.11	0.09	0.08
Congestión nasal	-0.07	-0.07	-0.07	-0.07	-0.08	-0.06	-0.07	-0.03	-0.02	-0.04	-0.02	-0.02	-0.02	-0.03	-0.02
Decaimiento	0.09	0.09	0.09	0.09	0.10	0.10	0.14	0.17	0.13	0.10	0.06	0.04	0.06	0.04	0.02
Disgeusia/ageusia	0.04	0.04	0.03	0.02	0.00	-0.02	-0.04	-0.05	-0.08	-0.10	-0.07	-0.04	-0.07	-0.06	-0.04
Disnea	0.13	0.13	0.14	0.15	0.16	0.20	0.23	0.25	0.30	0.29	0.28	0.24	0.18	0.19	0.18
Dolor torácico	0.01	0.01	0.02	0.03	0.01	0.01	0.04	0.04	0.05	0.01	0.02	0.02	0.02	0.00	-0.01
Fatiga	0.02	0.02	0.02	0.03	0.03	0.04	0.02	0.02	0.02	0.01	0.01	0.00	0.01	0.02	-0.01
Fiebre >37,8°C	0.20	0.20	0.20	0.20	0.20	0.17	0.08	0.11	0.09	0.06	0.05	0.02	0.03	0.02	0.02
Mialgias	0.04	0.04	0.05	0.09	0.05	0.08	0.08	0.06	0.09	0.06	0.04	0.04	0.00	-0.01	-0.02
Odinofagia	-0.04	-0.05	-0.03	-0.01	-0.02	-0.03	-0.02	-0.03	-0.01	-0.03	-0.02	-0.02	-0.02	-0.01	-0.01
Tos productiva	0.06	0.07	0.06	0.08	0.07	0.07	0.05	0.05	0.04	0.08	0.06	0.03	-0.01	-0.01	-0.01
Tos seca	0.04	0.04	0.05	0.04	0.06	0.06	0.03	0.04	0.04	-0.01	-0.03	-0.02	-0.06	-0.05	-0.03

0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30

Tabla 4.12: Punto porcentual entre pacientes con desenlace crítico y pacientes con desenlace leve

A través de este gráfico, se ve que en promedio los pacientes riesgosos asociados a una trayectoria sintomática presentan 20 puntos porcentuales más fiebre al comienzo de los síntomas que los pacientes leves. Con respecto a la disnea, el aumento de puntos porcentuales va desde 13 hasta 30 al superar los 7 días desde el comienzo de los síntomas. Por otro lado, el decaimiento y el CEG también marcan diferencias, donde el primero fluctúa de los 9 puntos porcentuales hasta los 17 a lo largo de la primera semana, mientras que el CEG va desde los 4 puntos hasta los 12 después y manteniéndose alrededor de dicha cifra después de la primera semana.

Síntomas como la anosmia, disgeusia/ageusia y la congestión nasal fueron

reportados con menor frecuencia en los casos de pacientes con desenlaces críticos, sin embargo, estas diferencias en puntos porcentuales llegan hasta 10 como máximo, siendo considerablemente menores a las asociadas a los síntomas anteriormente descritos.

4.2 Modelo de predicción

Con el grupo de pacientes ya segmentados en base a los síntomas que estos reportaron, lo que se busca en esta sección es poder predecir a que clúster sintomático va a pertenecer un paciente. El primer desafío a la hora de entrenar un modelo que prediga los síntomas que tendrá un paciente en el futuro, es la cantidad de datos a considerar. Si se incorporan los síntomas que un paciente ha presentado hasta el momento, naturalmente mientras más días sintomáticos se conozcan más certera será la predicción. Además de considerar los síntomas, los distintos modelos contemplan el rango etario, el género y los antecedentes declarados por cada paciente.

A continuación, considerando desde 1 a 14 días sintomáticos y separando la base de datos en un 75% de entrenamiento y un 25% de testeo, se muestra el desempeño de los siguientes 4 modelos de clasificación:

- Support Vector Machine (SVM)
- Artificial Neural Networks (ANN)
- Random Forest Classifier (RFC)
- ADA Boost Classifier (ADA)

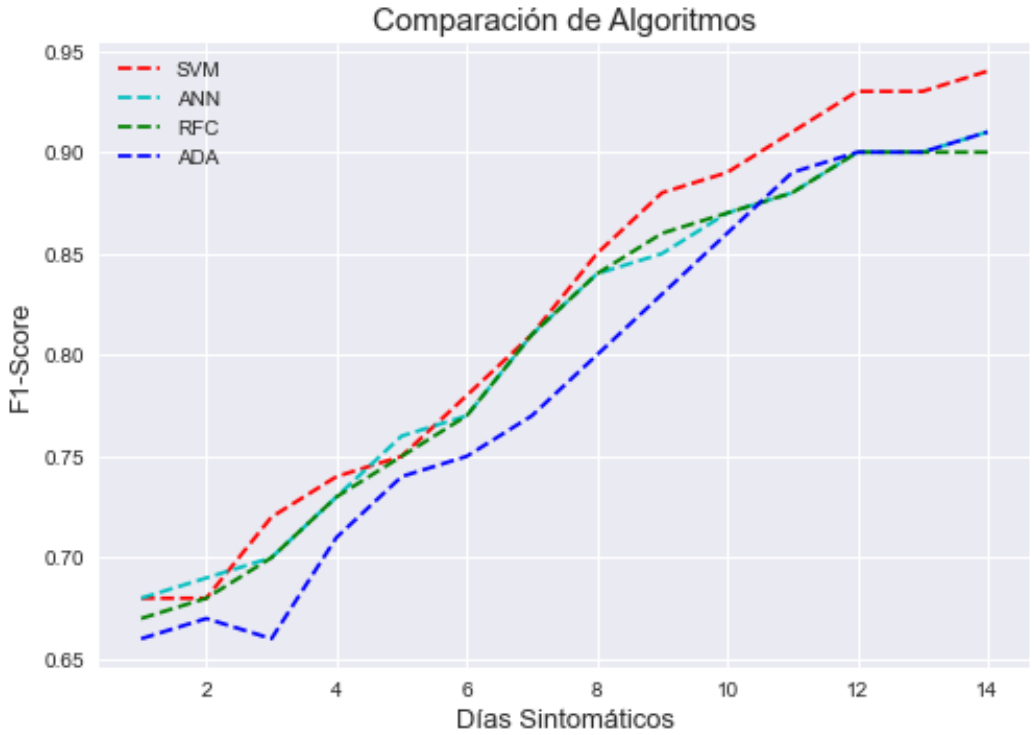


Figura 4.9: Comparación del F1-Score de distintos algoritmos de aprendizaje supervisado para disinta cantidad de días sintomáticos

Del gráfico se puede apreciar que el F1-Score de los 4 modelos de clasificación para la muestra de testeo es similar para una misma cantidad de días sintomáticos. SVM muestra el mejor F1-Score, seguido por ANN, RFC y finalmente ADA. Al ser un problema de múltiples clases (clústeres), es crucial ver la matriz de confusión y las distintas métricas asociadas al desempeño para validar de que el modelo predictivo asigne correctamente a los pacientes.

A continuación, se muestran las matrices de confusión obtenidas a través SVM. Estas matrices de confusión están normalizadas en las filas, por lo que los valores de la diagonal representan la precisión de la predicción para cada clúster. Adicionalmente, estas matrices cuentan con formato condicional de color, donde los valores amarillos claro son cercanos a 0 y los valores rojos intenso son cercanos al 1.

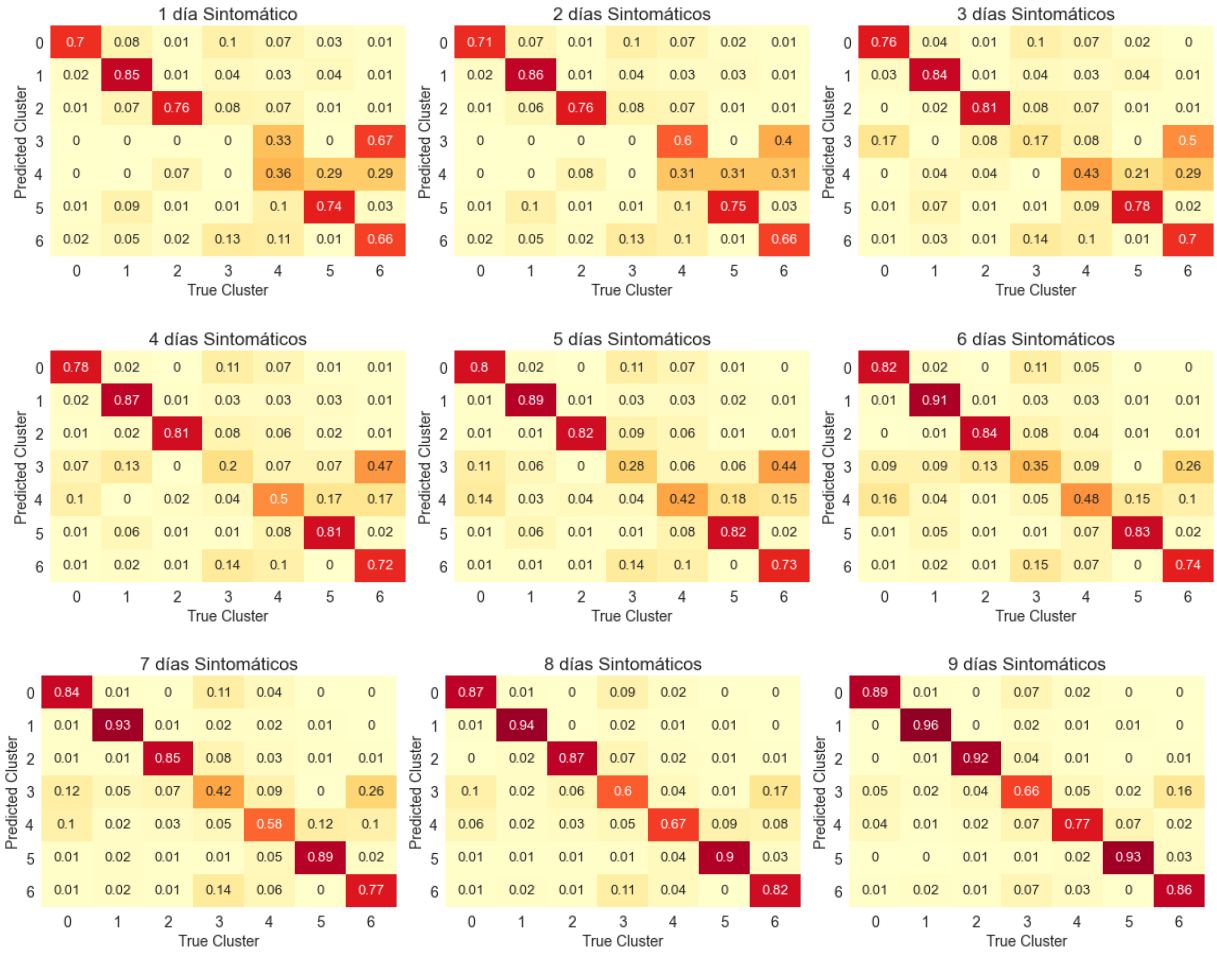


Figura 4.10: Gráficos de calor asociados a la precisión de los modelos predictivos construidos con SVM para distinta cantidad de días sintomáticos

Con relación a las distintas métricas asociadas a la predicción, se mostrarán las obtenidas al entrenar el modelo con 1, 2, 3, 4, 5, 10 y 15 días sintomáticos.

cluster	precision	recall	f1-score	support
0	0.70	0.93	0.80	456.00
1	0.85	0.76	0.81	599.00
2	0.76	0.91	0.83	329.00
3	0.00	0.00	0.00	198.00
4	0.36	0.03	0.05	191.00
5	0.74	0.83	0.78	315.00
6	0.66	0.93	0.77	427.00
accuracy			0.73	
macro avg	0.58	0.63	0.58	2515.00
weighted avg	0.66	0.73	0.68	2515.00

Tabla 4.12: Reporte de clasificación de SVM con 1 día sintomático

cluster	precision	recall	f1-score	support
0	0.71	0.93	0.81	456.00
1	0.86	0.78	0.82	599.00
2	0.76	0.92	0.84	329.00
3	0.00	0.00	0.00	198.00
4	0.31	0.02	0.04	191.00
5	0.75	0.84	0.79	315.00
6	0.66	0.93	0.77	427.00
accuracy			0.74	
macro avg	0.58	0.63	0.58	2515.00
weighted avg	0.66	0.74	0.68	2515.00

Tabla 4.13: Reporte de clasificación de SVM con 2 días sintomáticos

cluster	precision	recall	f1-score	support
0	0.76	0.93	0.84	456.00
1	0.84	0.89	0.86	599.00
2	0.81	0.93	0.86	329.00
3	0.17	0.01	0.02	198.00
4	0.43	0.06	0.11	191.00
5	0.78	0.84	0.81	315.00
6	0.70	0.92	0.79	427.00
accuracy			0.77	
macro avg	0.64	0.65	0.61	2515.00
weighted avg	0.70	0.77	0.72	2515.00

Tabla 4.14: Reporte de clasificación de SVM con 3 días sintomáticos

cluster	precision	recall	f1-score	support
0	0.78	0.94	0.85	456.00
1	0.87	0.91	0.89	599.00
2	0.81	0.95	0.87	329.00
3	0.20	0.02	0.03	198.00
4	0.50	0.13	0.20	191.00
5	0.81	0.86	0.83	315.00
6	0.72	0.93	0.81	427.00
accuracy			0.79	
macro avg	0.67	0.68	0.64	2515.00
weighted avg	0.73	0.79	0.74	2515.00

Tabla 4.15: Reporte de clasificación de SVM con 4 días sintomáticos

cluster	precision	recall	f1-score	support
0	0.80	0.94	0.86	456.0
1	0.89	0.92	0.91	599.0
2	0.82	0.95	0.88	329.0
3	0.28	0.03	0.05	198.0
4	0.42	0.16	0.23	191.0
5	0.82	0.89	0.85	315.0
6	0.73	0.93	0.81	427.0
accuracy			0.80	
macro avg	0.68	0.69	0.66	2515.0
weighted avg	0.74	0.80	0.75	2515.0

Tabla 4.16: Reporte de clasificación de SVM con 5 días sintomáticos

cluster	precision	recall	f1-score	support
0	0.89	0.95	0.92	456.0
1	0.95	0.95	0.95	599.0
2	0.91	0.95	0.93	329.0
3	0.68	0.54	0.60	198.0
4	0.83	0.77	0.80	191.0
5	0.94	0.93	0.94	315.0
6	0.89	0.93	0.91	427.0
accuracy			0.90	
macro avg	0.87	0.86	0.86	2515.0
weighted avg	0.89	0.90	0.89	2515.0

Tabla 4.17: Reporte de clasificación de SVM con 10 días sintomáticos

cluster	precision	recall	f1-score	support
0	0.94	0.96	0.95	456.00
1	0.96	0.96	0.96	599.00
2	0.96	0.95	0.96	329.00
3	0.87	0.84	0.85	198.00
4	0.90	0.85	0.88	191.00
5	0.94	0.95	0.95	315.00
6	0.94	0.96	0.95	427.00
accuracy			0.94	
macro avg	0.93	0.92	0.93	2515.00
weighted avg	0.94	0.94	0.94	2515.00

Tabla 4.18: Reporte de clasificación de SVM con 15 días sintomáticos

Con estos modelos se desarrolló una interfaz interactiva en Streamlit, la cual quedó a disposición de la Unidad de Salud Digital del SSMSO. A través de la interfaz, se pueden introducir los datos de nuevos pacientes para así predecir la trayectoria sintomática que este tendrá en el futuro. (Anexo E)

Finalmente, también se entrenó el modelo sin ningún día sintomático, es decir, tratar de predecir la trayectoria sintomática de un paciente únicamente conociendo sus antecedentes (en caso de tener), su edad y su género. La matriz de confusión y las distintas métricas asociadas son las siguientes:

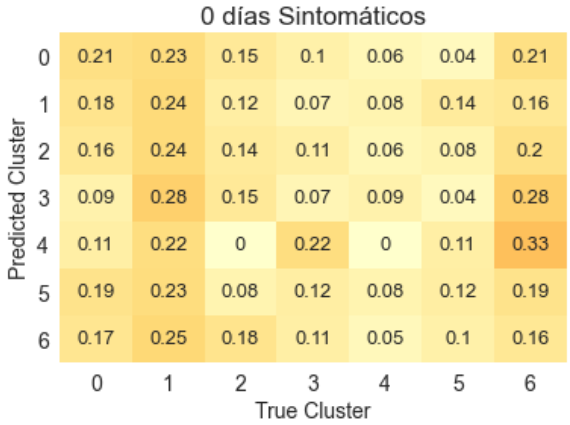


Figura 4.11: Gráfico de calor asociado a la predicción sintomática únicamente con variables personales

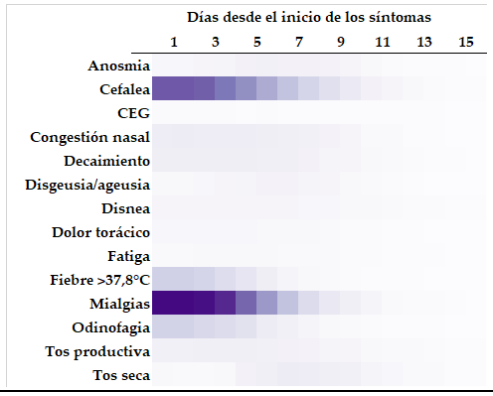
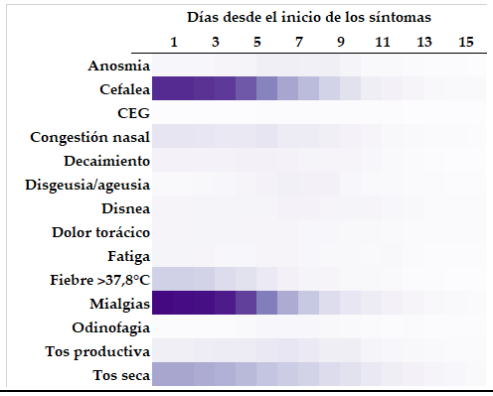
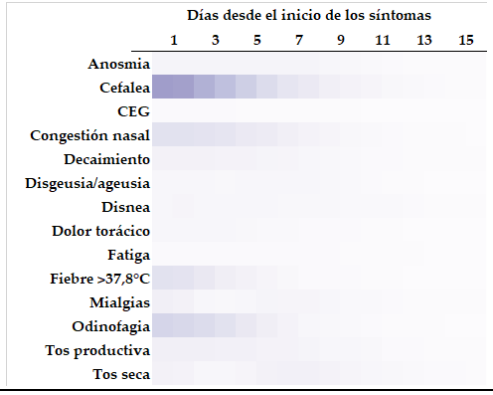
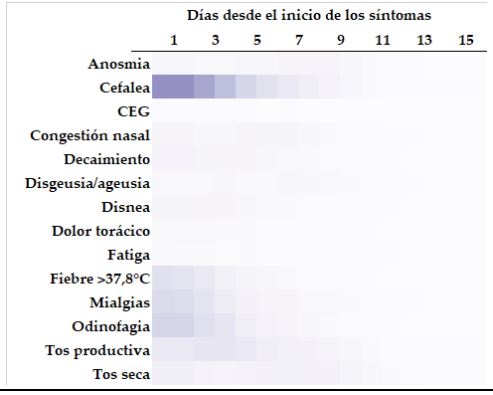
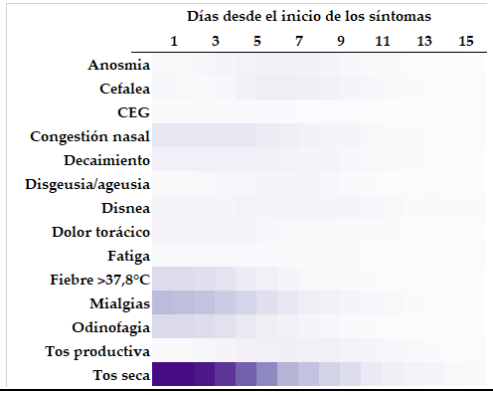
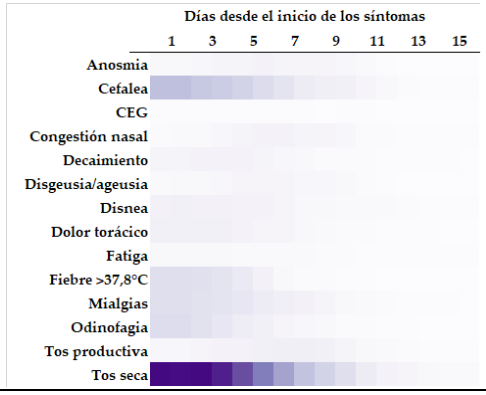
cluster	precision	recall	f1-score	support
0	0.21	0.08	0.11	456.00
1	0.24	0.78	0.36	599.00
2	0.14	0.05	0.07	329.00
3	0.07	0.02	0.02	198.00
4	0.00	0.00	0.00	191.00
5	0.12	0.01	0.02	315.00
6	0.16	0.07	0.09	427.00
accuracy			0.22	
macro avg	0.13	0.14	0.10	2515.00
weighted avg	0.16	0.22	0.14	2515.00

Tabla 4.19: Reporte de clasificación de SVM con 0 días sintomáticos

Como se puede apreciar, el desempeño del modelo sin ningún indicio de los síntomas que tuvo el paciente no entrega resultados que permitan tomar decisiones.

4.3 Muestra de validación - Clustering

Como última parte de la sección de los resultados, a la muestra independiente de mayo se le replicó la transformación de los datos el modelo de clusterización desde cero. Los resultados obtenidos son los siguientes:

Cluster	Muestra de junio 2020 a abril 2021	Muestra de mayo 2021
Clúster 0		
Clúster 1		
Clúster 2		

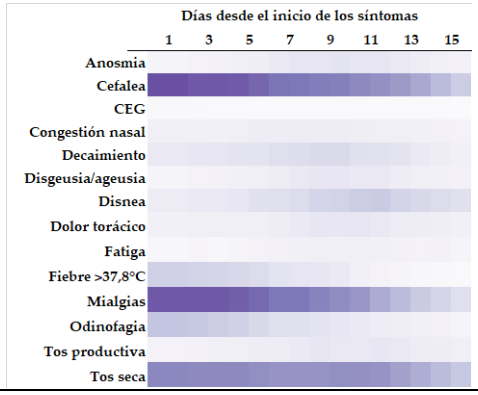
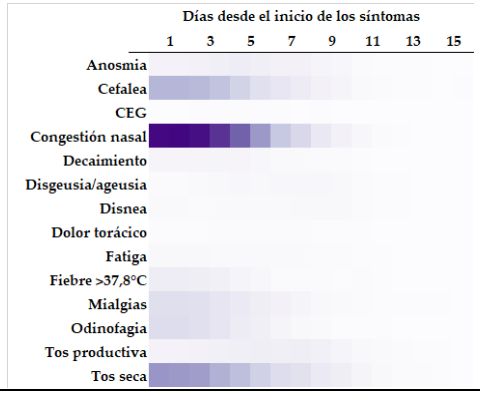
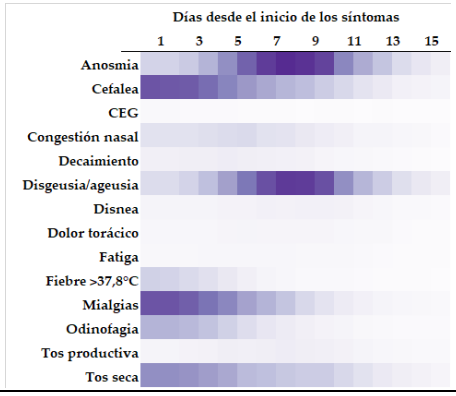
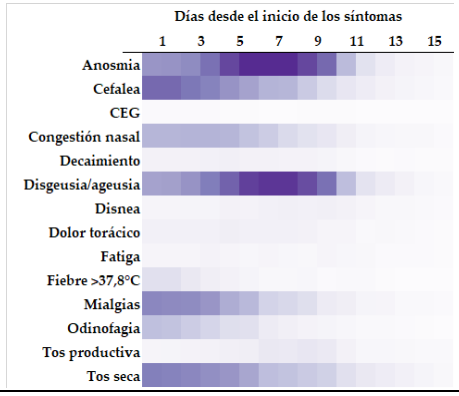
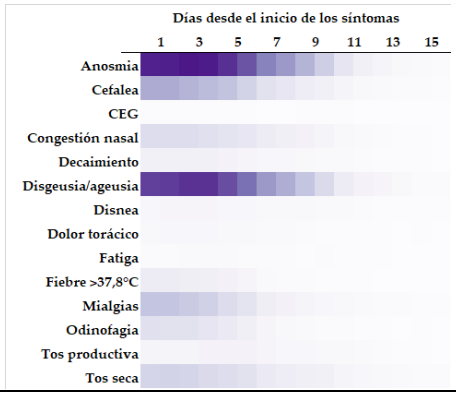
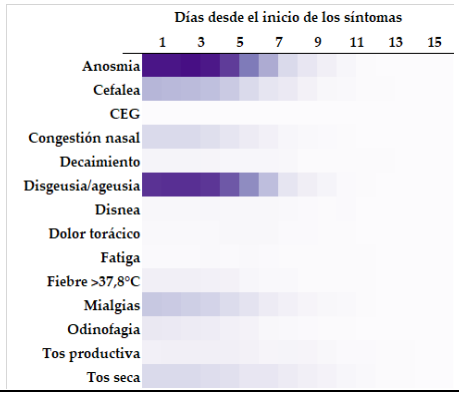
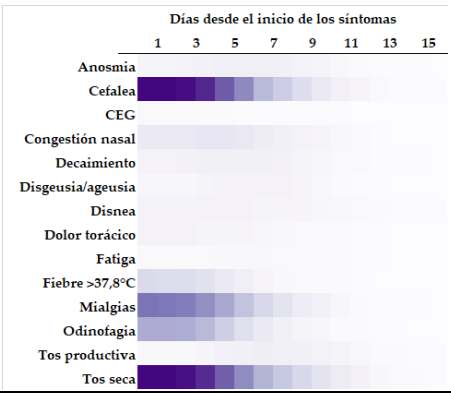
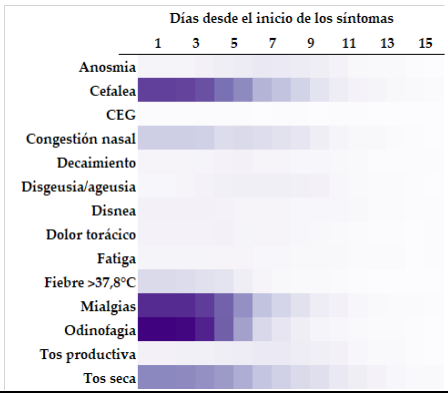
Clúster 3		
Clúster 4		
Clúster 5		
Clúster 6		

Tabla 4.20: Comparación de trayectorias sintomáticas de clústeres similares obtenidos en la muestra original y la muestra de mayo 2021

	Cluster (Normalizado en filas)						
Vacunación	0	1	2	3	4	5	6
No Vacunados	0.16	0.21	0.15	0.15	0.09	0.11	0.13
Sinovac	0.14	0.2	0.18	0.17	0.11	0.08	0.12

Tabla 4.21: Distribución de clústeres para pacientes de la muestra de mayo 2021

De dichos gráficos de calor se puede visualizar que en esencia las trayectorias sintomáticas se mantuvieron en la muestra independiente, a excepción del Clúster 3 asociado a cefalea, mialgias y tos seca, donde en su lugar surgió un clúster caracterizado por congestión nasal. No obstante, si bien se mantuvo la esencia de los síntomas característicos de los clústeres originales, es importante declarar las diferencias. Estas se pueden cuantificar a través de los puntos porcentuales entre la muestra de mayo y la muestra original para cada clúster (a excepción del Clúster 3). Las principales diferencias sintomáticas dentro de cada clúster son las siguientes:

	Días desde el comienzo de los síntomas														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Cluster 0															
Cefalea	0.13	0.13	0.13	0.18	0.18	0.16	0.11	0.11	0.07	0.05	0.03	0.03	0.02	0.02	0.02
Odinofagia	-0.27	-0.27	-0.25	-0.23	-0.17	-0.09	-0.05	-0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.01
Tos seca	0.42	0.43	0.41	0.38	0.30	0.23	0.17	0.15	0.13	0.10	0.09	0.07	0.06	0.04	0.04
Cluster 1															
Congestión nasal	-0.14	-0.14	-0.14	-0.13	-0.09	-0.07	-0.04	-0.03	-0.03	-0.02	-0.01	-0.01	0.00	-0.01	-0.01
Mialgias	0.14	0.15	0.14	0.09	0.04	-0.01	0.00	-0.02	-0.01	0.00	-0.01	-0.01	0.00	-0.01	-0.01
Cluster 2															
Cefalea	0.31	0.32	0.30	0.26	0.19	0.12	0.07	0.03	0.02	0.03	0.01	0.01	0.00	0.00	0.00
Congestión nasal	-0.15	-0.14	-0.14	-0.12	-0.10	-0.06	-0.04	-0.02	-0.01	-0.01	-0.02	0.00	-0.01	0.00	0.00
Mialgias	-0.16	-0.15	-0.15	-0.12	-0.10	-0.08	-0.06	-0.03	-0.02	-0.03	-0.02	-0.02	-0.01	0.00	0.00
Cluster 4															
Anosmia	0.24	0.25	0.24	0.25	0.23	0.16	0.05	0.00	-0.06	-0.11	-0.20	-0.24	-0.21	-0.16	-0.11
Congestión nasal	0.20	0.20	0.21	0.19	0.16	0.10	0.11	0.06	0.04	0.04	0.01	0.00	-0.02	-0.01	0.00
Disgeusia/ageusia	0.23	0.24	0.25	0.26	0.22	0.16	0.08	0.01	-0.05	-0.09	-0.18	-0.21	-0.17	-0.13	-0.11
Mialgias	-0.16	-0.17	-0.15	-0.13	-0.15	-0.08	-0.13	-0.08	-0.04	-0.05	-0.02	-0.03	-0.01	-0.02	-0.01
Cluster 5															
Anosmia	0.04	0.03	0.03	0.01	-0.04	-0.11	-0.16	-0.26	-0.24	-0.19	-0.13	-0.09	-0.07	-0.04	-0.03
Disgeusia/ageusia	0.04	0.04	0.01	-0.01	-0.03	-0.10	-0.14	-0.25	-0.22	-0.18	-0.12	-0.07	-0.06	-0.04	-0.02
Cluster 6															
Mialgias	0.21	0.22	0.24	0.26	0.25	0.19	0.09	0.08	0.06	0.04	0.04	0.03	0.02	0.01	0.01
Odinofagia	0.55	0.55	0.55	0.50	0.41	0.26	0.11	0.08	0.06	0.02	0.03	0.01	0.01	0.01	0.01
Tos seca	-0.40	-0.40	-0.38	-0.32	-0.21	-0.13	-0.07	-0.04	-0.01	0.03	0.02	0.03	0.03	0.02	0.01

0.50 0.45 0.40 0.35 0.30 0.25 0.20 0.15 0.10 0.05 0.00 -0.05 -0.10 -0.15 -0.20 -0.25 -0.30 -0.35 -0.45 -0.50

Tabla 4.22: Punto porcentual entre la muestra de mayo y la muestra original para las principales variaciones sintomáticas

Además de validar el resultado de los clústeres con esta muestra, hay resultados relevantes asociados al estado de vacunación de los pacientes. Específicamente hay 3.225 pacientes sin vacuna y 863 pacientes con más de 15 días desde su segunda dosis de Sinovac. La clusterización fue iterada 3 veces, una sobre la base considerando a ambos tipos de pacientes, otra sobre únicamente los pacientes sin ninguna vacuna y finalmente sobre únicamente los pacientes con la campaña de vacunación completa de Sinovac. Las trayectorias sintomáticas encontradas en los 3 casos son las mismas, únicamente con diferencias decimales despreciables. A raíz de esto, es que a continuación se mostrarán por separado los resultados asociados a los pacientes vacunados y a los no vacunados. Las siguientes dos tablas de variables personales siguen el mismo formato de las tablas anteriores, solo con la diferencia de que esta vez los p-valores están calculados en función del grupo de pacientes vacunados por un lado

y por el otro el grupo de los pacientes no vacunados. Todo esto sub segmentado por los rangos etarios en los que se encuentran.

	Rango etario														
	0-19			20-39			40-59			60-79			80+		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**640**	**24%**	**0***	**1460**	**56%**	**0***	**448**	**17%**	**0***	**67**	**3%**	**0***	**4**	**0%**	**0***
F	322	50%	0.988	720	49%	0.002*	228	51%	0*	35	52%	0.365	3	75%	0.727
Desenlace															
Hospitalizados	12	2%	0.502	42	3%	0.157	45	10%	0*	17	25%	0*	2	50%	0.243
Fallecimientos	0	0%	NaN	1	0%	0.47	3	1%	0.611	1	1%	0.702	0	0%	0.59
Duración Síntomas		**Std**			**Std**			**Std**			**Std**			**Std**	
Promedio + Std	7.1	2.9	0.247	8	3.4	0.028*	8.5	4.1	0.197	7.6	3.6	0.108	6.5	1.7	0.744
Antecedentes															
Asma	34	5%	0.919	31	2%	0.004*	7	2%	0.042*	4	6%	0.858	0	0%	0.384
Cardiopatías	7	1%	0.26	1	0%	0.177	3	1%	0.996	2	3%	0.748	0	0%	0.753
Diabetes	2	0%	0.01*	9	1%	0*	23	5%	0*	11	16%	0.637	1	25%	0.815
Enfermedad Renal	1	0%	0*	0	0%	NaN	1	0%	0.664	0	0%	0.741	0	0%	0.384
EPOC	0	0%	NaN	0	0%	NaN	0	0%	NaN	1	1%	0.864	0	0%	0.753
HTA	3	0%	0.092	12	1%	0*	33	7%	0*	18	27%	0.208	2	50%	0.727
Obesidad	7	1%	0.26	19	1%	0.381	3	1%	0.015*	1	1%	0.702	0	0%	NaN
Tabaco	3	0%	0.044*	51	3%	0.718	9	2%	0.575	0	0%	0.504	0	0%	NaN

Tabla 4.23: Tabla de variables personales de pacientes No Vacunados de la muestra de mayo 2021

	Rango etario														
	0-19			20-39			40-59			60-79			80+		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**11**	**2%**	**0***	**159**	**23%**	**0***	**193**	**28%**	**0***	**287**	**41%**	**0***	**44**	**6%**	**0***
F	5	45%	0.988	100	63%	0.002*	130	67%	0*	170	59%	0.365	23	52%	0.727
Desenlace															
Hospitalizados	0	0%	0.502	1	1%	0.157	2	1%	0*	21	7%	0*	6	14%	0.243
Fallecimientos	0	0%	NaN	1	1%	0.47	0	0%	0.611	5	2%	0.702	3	7%	0.59
Duración Síntomas		**Std**			**Std**			**Std**			**Std**			**Std**	
Promedio + Std	8.2	3.9	0.247	7.3	3.7	0.028*	8.1	3.5	0.197	8.4	3.9	0.108	7	3	0.744
Antecedentes															
Asma	0	0%	0.919	10	6%	0.004*	9	5%	0.042*	13	5%	0.858	2	5%	0.384
Cardiopatías	0	0%	0.26	0	0%	0.177	2	1%	0.996	8	3%	0.748	4	9%	0.753
Diabetes	0	0%	0.01*	6	4%	0*	27	14%	0*	57	20%	0.637	7	16%	0.815
Enfermedad Renal	0	0%	0*	0	0%	NaN	0	0%	0.664	4	1%	0.741	2	5%	0.384
EPOC	0	0%	NaN	0	0%	NaN	0	0%	NaN	6	2%	0.864	4	9%	0.753
HTA	1	9%	0.092	14	9%	0*	44	23%	0*	103	36%	0.208	24	55%	0.727
Obesidad	0	0%	0.26	4	3%	0.381	7	4%	0.015*	5	2%	0.702	0	0%	NaN
Tabaco	0	0%	0.044*	7	4%	0.718	6	3%	0.575	6	2%	0.504	0	0%	NaN

Tabla 4.24: Tabla de variables personales de pacientes vacunados con Sinovac de la muestra de mayo 2021

A través de la comparación entre estos dos grupos de pacientes, se ven diferencias estadísticamente significativas asociadas la hospitalización.

Complementariamente a los resultados anteriores, se hizo un test de Chi-Cuadrado para el desenlace de cada clúster, separando por rangos etarios y por estados de vacunación. Estos test Chi-Cuadrado contemplan 3 grados de libertad asociados a los 3 tipos distintos de desenlace (Fallecimiento, Hospitalizado, Alta). Los resultados muestran menores porcentajes de fallecimiento y de hospitalización para la mayoría de las edades dentro de cada clúster y algunas diferencias significativas en ciertos casos, no obstante, hacen falta más datos para seguir validando estas tendencias:

Cluster 0

	Rango etario (No Vacunado)					Rango etario (Vacunado)					
Desenlace	**0-19**	**20-39**	**40-59**	**60-79**	**80+**	**Desenlace**	**0-19**	**20-39**	**40-59**	**60-79**	**80+**
Fallecimiento	0%	0%	1%	0%	0%	**Fallecimiento**	0%	0%	0%	0%	25%
Hospitalizado	7%	4%	10%	31%	0%	**Hospitalizado**	0%	0%	4%	2%	0%
Alta	93%	96%	89%	69%	100%	**Alta**	100%	100%	96%	98%	75%
p-valor	0.10	0.81	0.49	0.01	NaN	**p-valor**	0.10	0.81	0.49	0.01	NaN

Tabla 4.25: Tabla comparativa entre pacientes no vacunados y no vacunados para el Clúster Sintomático 0

Cluster 1

	Rango etario (No Vacunado)					Rango etario (Vacunado)					
Desenlace	**0-19**	**20-39**	**40-59**	**60-79**	**80+**	**Desenlace**	**0-19**	**20-39**	**40-59**	**60-79**	**80+**
Fallecimiento	0%	0%	0%	0%	0%	**Fallecimiento**	0%	0%	0%	3%	9%
Hospitalizado	0%	4%	11%	24%	0%	**Hospitalizado**	0%	3%	0%	7%	9%
Alta	100%	96%	89%	76%	0%	**Alta**	100%	97%	100%	90%	82%
p-valor	NaN	0.89	0.10	0.09	NaN	**p-valor**	NaN	0.89	0.10	0.09	NaN

Tabla 4.26: Tabla comparativa entre pacientes no vacunados y no vacunados para el Clúster Sintomático 1

Cluster 2

	Rango etario (No Vacunado)					Rango etario (Vacunado)					
Desenlace	**0-19**	**20-39**	**40-59**	**60-79**	**80+**	**Desenlace**	**0-19**	**20-39**	**40-59**	**60-79**	**80+**
Fallecimiento	0%	0%	1%	6%	0%	**Fallecimiento**	0%	5%	0%	3%	6%
Hospitalizado	3%	3%	14%	24%	0%	**Hospitalizado**	0%	0%	0%	6%	17%
Alta	97%	97%	84%	71%	0%	**Alta**	100%	95%	100%	90%	78%
p-valor	0.05	0.01	0.16	0.10	NaN	**p-valor**	0.05	0.01	0.16	0.10	NaN

Tabla 4.27: Tabla comparativa entre pacientes no vacunados y no vacunados para el Clúster Sintomático 2

Cluster 3

Rango etario (No Vacunado)					
Desenlace	0-19	20-39	40-59	60-79	80+
Hospitalizado	0%	2%	4%	20%	0%
Alta	100%	98%	96%	80%	0%
p-valor	NaN	0.86	0.66	0.99	NaN

Rango etario (Vacunado)					
Desenlace	0-19	20-39	40-59	60-79	80+
Hospitalizado	0%	0%	3%	9%	12%
Alta	100%	100%	97%	91%	88%
p-valor	NaN	0.86	0.66	0.99	NaN

Tabla 4.28: Tabla comparativa entre pacientes no vacunados y no vacunados para el Clúster Sintomático 3

Cluster 4

Rango etario (No Vacunado)					
Desenlace	0-19	20-39	40-59	60-79	80+
Hospitalizado	4%	1%	3%	100%	0%
Alta	96%	99%	97%	0%	0%
p-valor	0.02	0.27	0.99	0.32	NaN

Rango etario (Vacunado)					
Desenlace	0-19	20-39	40-59	60-79	80+
Hospitalizado	0%	0%	0%	11%	0%
Alta	100%	100%	100%	89%	0%
p-valor	0.02	0.27	0.99	0.32	NaN

Tabla 4.29: Tabla comparativa entre pacientes no vacunados y no vacunados para el Clúster Sintomático 4

Cluster 5

Rango etario (No Vacunado)					
Desenlace	0-19	20-39	40-59	60-79	80+
Hospitalizado	0%	2%	7%	20%	0%
Alta	100%	98%	93%	80%	100%
p-valor	NaN	0.74	0.59	0.58	1.00

Rango etario (Vacunado)					
Desenlace	0-19	20-39	40-59	60-79	80+
Hospitalizado	0%	0%	0%	0%	100%
Alta	100%	100%	100%	100%	0%
p-valor	NaN	0.74	0.59	0.58	1.00

Tabla 4.30: Tabla comparativa entre pacientes no vacunados y no vacunados para el Clúster Sintomático 5

Cluster 6

Rango etario (No Vacunado)					
Desenlace	0-19	20-39	40-59	60-79	80+
Fallecimiento	0%	0%	2%	0%	0%
Hospitalizado	5%	3%	15%	20%	100%
Alta	95%	97%	84%	80%	0%
p-valor	0.18	0.94	0.08	0.89	0.32

Rango etario (Vacunado)					
Desenlace	0-19	20-39	40-59	60-79	80+
Fallecimiento	0%	0%	0%	3%	0%
Hospitalizado	0%	0%	0%	15%	0%
Alta	100%	100%	100%	82%	100%
p-valor	0.18	0.94	0.08	0.89	0.32

Tabla 4.31: Tabla comparativa entre pacientes no vacunados y no vacunados para el Clúster Sintomático 6

5 Discusión

En esta sección se lleva a cabo la discusión y el análisis de los resultados plasmados en la sección anterior. En primer lugar, se analizarán los resultados del clustering, tanto los de la muestra original como los de la muestra de validación. Finalmente, se analizan los resultados de los modelos predictivos.

5.1 Análisis del clustering y validación

El primer punto para analizar sobre los resultados corresponde a las métricas de desempeño asociadas a los 7 clústeres definidos. En particular, el coeficiente de siluetas asociado a 7 grupos sintomáticos tiene un valor de 0.027, el cual está muy cercano a 0 y sugiere que no hay una gran distancia entre los vectores sintomáticos asociados a los pacientes. Los motivos que explican esta métrica son, en primer lugar, que los distintos clústeres están basados principalmente en los siguientes 5 síntomas: Cefalea, Mialgias, Tos Seca, Anosmia, Disgeusia/Ageusia y secundariamente en Odinofagia y Congestión Nasal. Lo que diferencia a un clúster de otro, en términos sintomáticos, es la presencia o ausencia de alguno de estos síntomas principales o la duración de estos en el caso de los clústeres 4 y 5 donde los síntomas prevalecen por más tiempo. En segundo lugar y siguiendo la línea de los síntomas que prevalecen por más tiempo, la base de datos está construida con pacientes que presentan síntomas a lo largo de distintas cantidades de días. Con el fin de poder usar un algoritmo que agrupe estos pacientes a pesar de dicha diferencia, se consideraron hasta 20 días sintomáticos, en donde si un paciente deja de presentar síntomas en un determinado día desde el inicio de sus síntomas, para efectos de la base de datos, el resto de los síntomas asociados a los días restantes para llegar a 20 días se rellenan con 0, indicando que no presentaron dichos síntomas en dichos días. De este modo, por ejemplo, pacientes que presentaron síntomas diferentes a lo largo de 10 días, comparten la misma sub-trayectoria sintomática para los 10 días finales, la cual corresponde a no presentar síntomas en dicho tramo final. A raíz de estos dos puntos, es que dada la configuración de la base de datos y a pesar de haber sido tratada con PCA, el coeficiente de siluetas entrega un valor bajo, a pesar de haber segmentando la base en el número óptimo de clústeres.

El segundo punto para analizar corresponde al Clúster 1, el cual se caracteriza por presentar cefalea por una escasa cantidad de días para la mayoría de los pacientes asignados. Este grupo, a pesar de no tener diferencias estadísticamente significativas con respecto a hospitalización y fallecimiento en comparación a los otros grupos, si presenta altas tasas para los distintos rangos etarios. En un principio, parece contra intuitivo que una trayectoria sintomática de corta duración tenga estas características, pero la explicación viene por los siguientes puntos:

El Clúster 1 está conformado por distintos tipos de personas, teniendo por un lado pacientes que efectivamente tuvieron una trayectoria sintomática leve de escasos días y escasos síntomas que pudieron dejar el seguimiento sin inconvenientes. Por otro lado, hay pacientes que a los pocos días de haber presentado síntomas se les agravó la enfermedad y requirieron de hospitalización (Ej: el 63% de los fallecidos del grupo muestran síntomas por menos de 6 días), en donde no se pudo continuar con el

seguimiento y administrativamente se ve como que hubiesen tenido síntomas por pocos días. Adicionalmente, hay pacientes que requirieron hospitalización y tuvieron una trayectoria sintomática caracterizada principalmente por disnea. Al haber tenido disnea como único síntoma predominante y no ser los suficientes pacientes como para poder conformar su propio clúster, es que el algoritmo los asigno al Clúster 1 al ser más similar que los otros clústeres que tienen varios otros síntomas predominantes.

El tercer punto es con respecto al Clúster 3 y el Clúster 6, los cuales están caracterizados por cefalea, mialgias y tos seca, siendo el primero de estos clústeres de larga duración (más de 2 semanas) y el segundo de duración más corta (menos de 2 semanas). Además de la diferencia significativa que hay asociada a la duración de estas trayectorias para todos los rangos etarios, es importante destacar que adicionalmente el Clúster 3 tiene una mayor proporción de pacientes pertenecientes a rangos etarios de más avanzada edad en comparación al Clúster 6. Además, si se comparan las distintas tasas de hospitalización y fallecimiento, se puede apreciar que las del Clúster 3 son considerablemente superiores y presentan diferencias estadísticamente significativas para el rango de 60 a 79 años. Básicamente, esto último se traduce en que la prevalencia de los síntomas esta asociada a una mayor probabilidad de requerir asistencia hospitalaria.

Con respecto al Clúster 0 de cefalea y mialgias, el Clúster 2 de tos seca y el Clúster 6, de cefalea, tos seca y mialgias se ve que ambos son trayectorias similares en varios aspectos. Por un lado, la duración que estos tienen es casi idéntica para los distintos rangos etarios. Por otro lado, no muestran diferencias concluyentes entre si con relación a las tasas de hospitalización y fallecimiento para los distintos rangos etarios. De esto último se desprende que la presencia o la ausencia de uno de estos síntomas principales en una trayectoria sintomática, no necesariamente se asocia a un desenlace más o menos riesgoso. A modo de ejemplo, la probabilidad de un desenlace riesgoso observada para un paciente de 65 años que ha tenido cefalea y mialgias por 5 días, no se diferencia de la probabilidad observada para un paciente de 65 años con los mismos antecedentes, pero, además de presentar cefalea y mialgias por 5 días, presenta también tos seca.

El cuarto punto para analizar corresponde a los Clúster 4 y 5, los cuales tienen como síntomas predominantes anosmia y disgeusia/ageusia. Al comparar los pacientes asignados a estos clúster con los pacientes que corresponden a otros clúster, se visualizan diferencias estadísticamente signficativas asociadas a la hospitalización para pacientes entre 20 y 59 años para el Clúster 4 y para pacientes entre 40 y 79 años para el Clúster 5. Adicionalmente, solo 3 personas de 80 o más años fueron asignadas a estos 2 grupos, donde cabe destacar que ninguna requirió hospitalización ni falleció. Al comparar estos resultados con lo que señala la literatura sobre la presencia de la anosmia como síntoma de COVID-19, se puede comprobar que para los pacientes que se atienden en el SSMSO tener trayectorias sintomáticas asociadas a este síntoma presentan un menor riesgo en comparación a padecer de otras trayectorias sintomáticas. Más aún si se analiza el Clúster 4, cuyos síntomas iniciales son cefalea, mialgias y tos seca, los cuales después del quinto día aproximadamente dejan de manifestarse y aparece anosmia y disgeusia/ageusia, se ve que, a pesar de haber comenzado con otros síntomas, la transición hacía anosmia y disgeusia/ageusia sugiere

una menor probabilidad de tener un desenlace riesgoso para todos los rangos etarios. Este último punto refuerza la idea de hacer un seguimiento a los pacientes que padecen de COVID-19, dado que cambios de este tipo en lo que son los síntomas reportados puede ser un indicador de que un paciente este pasando de presentar síntomas potencialmente riesgosos a otros síntomas donde la probabilidad de requerir hospitalización o fallecer disminuye considerablemente

El quinto punto para analizar corresponde a las trayectorias sintomáticas de los pacientes que requirieron hospitalización o que fallecieron. Además de lo descrito en el segundo y tercer punto, en base a los resultados expuestos, se puede ver que típicamente comparten las mismas trayectorias que los pacientes que tuvieron desenlaces leves. La principal diferencia es que a estas trayectorias se les suma el síntoma de disnea, el cual aproximadamente desde el séptimo día en adelante comienza a ser más notorio y predominante para todos los clústeres. Además de presentar disnea, también se logra diferenciar un aumento en lo que son los reportes del compromiso del estado general (CEG), el decaimiento y la fiebre en los primeros días desde el inicio de los síntomas.

Dentro de estos síntomas la disnea es reconocido como un síntoma asociado a un caso potencialmente grave. Por otro lado, la fiebre en la literatura está considerado como un síntoma común de COVID-19, por lo que considerarlo como un nuevo determinante de caso severo podría llevar a conclusiones erradas. El decaimiento y el CEG son síntomas que potencialmente podrían tenerse mayormente vigilados, dado que no son síntomas típicamente relacionados con evoluciones críticas de los pacientes. Nuevamente, estos resultados refuerzan la idea de mantener los seguimientos, dado que se puede identificar cuando un paciente comienza a desviarse hacia una versión más riesgosa de la trayectoria sintomática a la cual está asociado.

El sexto punto para analizar corresponde a las diferencias sintomáticas entre distintos grupos etarios. En base a lo expuesto, se ve que existen diferencias en lo que respecta a la asignación de pacientes a un grupo sintomático dependiendo de la edad. A modo de ejemplo, es poco frecuente que pacientes mayores de 60 años presenten exclusivamente anosmia y disgeusia/ageusia. No obstante, a pesar de que existan estas diferencias, ningún grupo sintomático es exclusivo para algún rango etario y en términos generales, cualquier persona puede presentar cualquier trayectoria. Lo que es importante destacar sobre este punto, es que, comparando los pacientes de un mismo grupo sintomático, hay diferencias estadísticamente significativas en lo que respecta a la duración de los síntomas. Más precisamente, a mayor edad que tenga un paciente, es más propenso a presentar síntomas por una mayor cantidad de días.

De los 27 síntomas que un paciente puede reportar al recibir el seguimiento del SSMSO, en este trabajo se ve que varios de ellos no destacan en ninguna de las trayectorias sintomáticas. Ejemplos tales como la taquipnea, cianosis, anorexia, compromiso de conciencia entre otros, fueron declarados con una frecuencia demasiado baja, por lo que fueron opacados por las otras trayectorias sintomáticas siendo finalmente asignadas a estas. Síntomas asociados a manifestaciones gastrointestinales tales como diarrea, náuseas, vómitos, dolor torácico y dolor abdominal distribuyeron de forma equitativa entre los distintos clústeres sin tener una inclinación preferente. La diarrea fue el síntoma más frecuente de los 5, manifestándose aproximadamente en un 10% de los casos durante los días iniciales y decayendo a lo

largo de la primera semana. Los vómitos y las náuseas tuvieron el mismo patrón, pero únicamente con un 3% de los pacientes de cada clúster y los dolores torácicos y abdominales con un 5% de los pacientes. Finalmente, como los síntomas son autodeclarados, la rinorrea, síntoma definido como un flujo o emisión abundante por la nariz, potencialmente puede haberse visto opacado por la declaración de congestión nasal, la cual tiene un nombre mucho más intuitivo desde un punto de vista coloquial.

Con respecto a los antecedentes que declararon los pacientes, no se observa ninguna inclinación o predominancia notoria para alguna trayectoria sintomática, si no que estas están presentes a lo largo de todas las trayectorias y rangos etarios. Existen algunos casos puntuales, tales como que el Clúster 3 tiene un mayor porcentaje de pacientes diabéticos en los rangos de 20 a 59 años y podría existir una asociación de que las personas diabéticas son más propensas a tener trayectorias sintomáticas más largas. No obstante, este resultado no es del todo concluyente, dado que, si bien puede existir esta propensión, también existen pacientes diabéticos de dichos rangos etarios que tuvieron otras trayectorias sintomáticas y de menor duración. Lo que, si se observa y cuenta con diferencias estadísticamente significativas, son la presencia de antecedentes para distintos rangos etarios, específicamente el aumento de ciertos antecedentes como hipertensión, cardiopatías y diabetes para edades más avanzadas. Finalmente, la presencia de los antecedentes para ciertos pacientes está más relacionado con el rango etario que con los síntomas que un paciente tuvo.

Comparando las trayectorias sintomáticas que fueron obtenidas a lo largo de este trabajo con las dinámicas sintomáticas que están planteadas en la literatura, se puede ver que en la mayoría de los casos calzan y comparten factores asociados. A modo de ejemplo, en este estudio también se ve como la tos seca, el dolor de cabeza (cefalea) y el dolor corporal (mialgias) son los síntomas más frecuentes asociados a la enfermedad. En la literatura se plantea que la fiebre también está dentro de los síntomas más frecuentes, pero en base a los resultados obtenidos se ve que los datos del SSMSO sugieren que este síntoma es menos frecuente y en caso de presentarse, está asociado al comienzo de los cuadros sintomáticos. Adicionalmente, en ciertos estudios se menciona la fatiga como un síntoma común o como síntoma asociado al riesgo. En esta investigación la fatiga es un síntoma que fue declarado en escasas oportunidades y no muestra una diferencia considerable entre los pacientes leves y los pacientes riesgosos. Este fenómeno podría explicarse dado que en el seguimiento hay otros síntomas similares que podrían explicar de manera más precisa el sentir de los pacientes, tales como el decaimiento y el compromiso del estado general, los cuales si están asociados a pacientes riesgosos. También, se comparte que la anosmia, síntoma el cual en este caso viene acompañado frecuentemente por disgeusia/ageusia es un síntoma asociado a un bajo riesgo y que la disnea es uno de los principales síntomas que pueden encarrilar una trayectoria sintomática leve hacia un desenlace crítico. Con respecto a los síntomas gastrointestinales, estos también se vieron como síntomas atípicos.

En la siguiente tabla comparativa se pueden apreciar los resultados comunes y las principales diferencias que hay entre las distintas investigaciones desarrolladas en la parte 2.5.2 de dinámicas sintomáticas y lo obtenido a lo largo de este trabajo:

Investigación	Resultados comunes	Principales diferencias
Longitudinal symptom dynamics of COVID-19 infection. [28]	• Tos y mialgias figuran dentro de los síntomas más reportados. • Disnea o falta de aire como síntoma tardío.	• Cefalea no figura dentro de los síntomas prevalentes. • Fatiga y congestión nasal como síntomas predominantes.
COVID-19's natural course among ambulatory monitored outpatients. [29]	• Mayor porcentaje de anosmia presente en mujeres que en hombres. • Disnea o dificultad para respirar fueron más característicos en cuadros largos.	• Fatiga y fiebre como síntomas más comunes que cefalea, mialgias, ageusia y tos seca.
Identifying and ranking common COVID-19 Symptoms From Tweets in Arabic: Content Analysis. [30]	• Cefalea como síntoma inicial frecuente. • Fatiga como síntoma menos frecuente y tardío.	• Fiebre como síntoma inicial frecuente.
Modeling the Onset of Symptoms of COVID-19. [31]	• La severidad del COVID-19 no altera el orden de aparición de los síntomas típicos discernibles. • Tos como síntoma frecuente.	• Contempla trayectorias sintomáticas improbables (comenzar con síntomas gastrointestinales, seguidos por tos y finalizando con fiebre). • Fiebre como síntoma inicial más probable.
Symptom clústeres in COVID-19: A potential clinical prediction tool from the COVID Symptom Study app. [32]	• Clústeres de trayectorias sintomáticas. • Cefalea como síntoma trascendental al clúster. • Anosmia como síntoma asociado a trayectorias menos riesgosas. • Disnea como síntoma tardío asociado a	• Existen diferencias notorias en el riesgo de hospitalización y de requerir asistencia respiratoria entre cada clúster. • Fatiga como síntoma asociado a trayectorias más riesgosas.

	trayectorias riesgosas.	

Tabla 5.1: Tabla comparativa entre el trabajo de título e investigaciones de dinámicas sintomáticas de COVID-19

Con respecto a los resultados del clustering obtenidos sobre la muestra independiente de mayo, se puede visualizar que se comparten gran parte de los patrones de las trayectorias sintomáticas obtenidas a través del clustering original. La principal diferencia es que el Clúster 3 asociado a cefalea, mialgias y tos seca por una larga duración no se forma con la nueva base y en su lugar aparece un nuevo clúster caracterizado por congestión nasal. Sumado a esta diferencia, los pacientes asignados a cefalea, tos seca y mialgias por una corta duración en el mes de mayo presentan considerablemente más odinofagia que los del clustering original.

Cabe destacar que en la muestra de mayo los pacientes efectivamente reportan tener más congestión nasal, siendo un 8.7% de los síntomas reportados, mientras que en la muestra original solo era un 5,6%. Adicionalmente, en la muestra de mayo se tienen registros de menos días sintomáticos para los pacientes. El motivo del primer punto puede deberse a que al considerar únicamente los pacientes de un mes caracterizado por el frío puede haber impulsado la presencia de la congestión nasal. El motivo del segundo punto puede deberse a distintas causas, tales como que al solo comparar un mes de datos y mantener únicamente los pacientes egresados para el análisis, pacientes que tuvieron una trayectoria sintomática larga fueron extraídos de la muestra al no haber finalizado su trayectoria antes de que terminase el mes de mayo.

En consecuencia, se ve que mayoritariamente las trayectorias sintomáticas y los patrones que estas siguen se mantienen en el tiempo, no obstante, el hecho de que haya habido cambios en ciertas configuraciones sugiere que es importante mantener una actualización constante de la clusterización con el fin de asegurar que los pacientes se asignen a grupos sintomáticos que representen de mejor manera los síntomas que estos presenten.

Siguiendo con la muestra de mayo, es importante destacar que las trayectorias sintomáticas obtenidas a través del clustering de los pacientes no vacunados haya dado los mismos patrones que la de los pacientes vacunados con más de 15 días desde su segunda dosis de Sinovac. Este resultado valida no solo que una persona con su campaña de vacunación completa de Sinovac pueda tener COVID-19, sino que también dicha persona probablemente presente los mismos síntomas considerando un curso de enfermedad no riesgoso. La clave para entender el efecto de la vacunación para enfrentar el COVID-19 está en este último punto: un curso de enfermedad no riesgoso. Tal como fue analizado en el quinto punto, los pacientes que requirieron hospitalización o fallecieron tienen trayectorias sintomáticas muy similares a la de los pacientes de un mismo clúster que tuvieron un desenlace leve, con la diferencia de que para el caso grave síntomas como el decaimiento, el CEG y la disnea marcaron la diferencia. En los pacientes con la campaña completa de vacunación de Sinovac, la cantidad de pacientes que terminan teniendo estos síntomas y que, por consecuencia, terminan requiriendo hospitalización o falleciendo son una cantidad considerablemente menor. Es más, comparando el grupo de pacientes no vacunados con el de pacientes vacunados con Sinovac, se puede ver que para el rango entre 40 a 59 años y el de 60 a 79 años, los pacientes no presentan diferencias estadísticamente significativas en cuanto a la duración de los síntomas, pero si las presentan en lo que es la tasa de pacientes que

requirieron hospitalización. En relación con los antecedentes asociados a ambos rangos etarios, el grupo de pacientes vacunados presenta mayores porcentajes de diabetes, de hipertensión, asma y de obesidad (estadísticamente significativos en el rango de 40 a 59 años para estos 4 antecedentes), los cuales son factores que se asocian con desenlaces riesgosos.

Al hacer la comparación dentro de un mismo clúster sintomático entre los pacientes vacunados contra los no vacunados la tendencia se repite, los pacientes vacunados, a pesar de haber presentado las mismas trayectorias sintomáticas, presentan menores tasas de hospitalización y de fallecimiento. Sin embargo, en este caso como cada clúster es aproximadamente un 15% de la muestra de mayo, hay pocos datos para validar de que haya diferencias estadísticamente significativas para todos los rangos etarios en todas las trayectorias sintomáticas.

En resumen, tanto vacunados como no vacunados comparten las mismas trayectorias sintomáticas por una cantidad casi idéntica de días, el grupo de los pacientes vacunados tiene antecedentes más riesgosos y, aun así, presenta porcentajes de hospitalización inferiores.

5.2 Análisis de los modelos predictivos

Previo al análisis en si de los modelos predictivos, es crucial definir los tiempos en los que la predicción de la trayectoria sintomática logra aportar valor. Si se logra predecir la trayectoria sintomática que va a tener un paciente previo a que este la tenga, el personal de salud se puede anticipar a la situación y estar preparado para los tratamientos que vayan a ser requeridos dependiendo del caso, dado que podrán conocer cómo será la evolución sintomática del paciente, además de asociarla con alguna probabilidad de riesgo dependiendo del rango etario en el cual se encuentre.

Si el modelo predictivo contempla demasiados días sintomáticos, básicamente en vez de predecir la trayectoria sintomática que tendrá un paciente, lo que estaría haciendo sería clasificar la trayectoria sintomática que el paciente ya tuvo.

En base a los resultados de las trayectorias sintomáticas, se tienen los siguientes aspectos relevantes a considerar

- El promedio de duración de síntomas de los pacientes entre 0 y 19 años es de 7.7 días, mientras que el del Clúster 1, caracterizado por ser de corta duración, para dicho rango etario es de 6.2 días
- El promedio de duración de síntomas de los pacientes de 80 o más años es de 11.8 días, mientras que el del Clúster 3, caracterizado por ser de larga duración, para dicho rango etario es de 22.2 días
- Los síntomas asociados a desenlaces críticos comienzan a ser más predominantes a partir del 7 día aproximadamente

A raíz de estos aspectos y de los puntos discutidos al comienzo de la sección, es que, si se predice considerando a lo más 5 días sintomáticos, se logra el objetivo de poder predecir los síntomas que tendrá una persona en el futuro, además de ser previo al día en el que típicamente comienzan a manifestarse de forma más predominante los síntomas asociados a desenlaces críticos. En consecuencia, utilizar los modelos predictivos considerando 10 o 15 días debería ser descartado o utilizado únicamente para registrar la trayectoria sintomática que un paciente ya tuvo.

Retomando los resultados obtenidos con el modelo entrenado con 5 días sintomáticos, las principales métricas fueron un accuracy del 80%, una precisión balanceada del 74% y un recall balanceado del 80%. No obstante, la precisión macro es del 68% y el recall macro es del 69%. La clave para entender las diferencias entre estos valores viene en lo que es la capacidad de predicción para los clústeres 3 y 4, los cuales son los clústeres de larga duración de, por un lado, cefalea mialgias y tos seca y, por el otro lado, de anosmia y disgeusia/ageusia. En promedio el modelo lográ clasificar de forma correcta a todas las clases a excepción de los clústeres 3 y 4, donde se destaca una baja precisión del 28% y 42% respectivamente y un deplorable recall del 3% y 16% respectivamente.

Si se analiza el modelo predictivo considerando 15 días, se puede visualizar como

estas métricas mejoran radicalmente, llegando para el clúster 3 y 4 a una precisión del 87% y 90% respectivamente y a un recall del 84% y 85% respectivamente.

Cabe destacar que hay un desbalance en los datos, donde tanto el clúster 3 como el 4 tienen menos registros asociados en comparación a los otros clústeres.

De estos 3 puntos se desprende que, de contar con pocos días sintomáticos para realizar la predicción, el modelo entrega buenas métricas a la hora de identificar cuáles serán los síntomas que presentará un determinado paciente en el futuro, no obstante, no logra discernir de forma correcta por cuanto tiempo va a manifestar dicho paciente dichos síntomas.

Al estar desbalanceados los datos y al aparentemente no haber suficientes diferencias relacionadas a los antecedentes, edad y genero entre un clúster de corta duración y su par de larga duración, el modelo predice que el paciente tendrá el clúster de corta duración. Esto explica los bajos recalls asociados, dado que rara vez el modelo predice que será de larga duración y las veces que lo hace, falla en más de la mitad de los casos.

Siguiendo la línea de que el modelo tiene dificultades para discernir basándose en las diferencias asociadas a antecedentes, género y edad, el último punto a analizar corresponde a los resultados obtenidos en la predicción sin ningún registro sintomático. Con un accuracy promedio del 22%, y con un recall inferior al 10% para todas las clases a excepción del Clúster 1, se ve que el modelo rara vez logra asociar un set de variables personales a un determinado clúster y la mayoría de las veces, ante la incertidumbre, únicamente predice que el paciente pertenece al Clúster 1 dado que es el grupo con más pacientes asignados.

De estos resultados, se ve que es necesario al menos un día sintomático para poder predecir de forma fidedigna los síntomas que vaya a tener un paciente en el futuro. A raíz de este punto, para implementar este trabajo en la práctica clínica diaria, la predicción sintomática debiese comenzar únicamente cuando se tenga al menos un registro sintomático y esta predicción debiese ir actualizándose a medida que se vayan teniendo más registros con el fin de mejorar la calidad de la predicción. Adicionalmente, si se cuenta únicamente con los síntomas iniciales, los resultados debiesen ser referenciales para conocer que tipo de síntomas se presentarán en el futuro, mas no para estimar cuanto tiempo perdurarán dichos síntomas. Finalmente, independiente del clúster sintomático, es crucial estar alerta a los distintos síntomas asociados al riesgo identificados en este trabajo.

6 Conclusiones

Como conclusiones generales de este trabajo, se tiene que existen diferencias en lo que son las trayectorias sintomáticas que presentan los pacientes infectados por Sars-CoV-2. Estas diferencias están dadas por la duración de los síntomas, la manifestación de determinados síntomas y los patrones que siguen.

No existe una única trayectoria sintomática asociada a un desenlace crítico, las trayectorias sintomáticas asociadas a casos críticos comparten la misma esencia que las asociadas a desenlaces leves, pero con la aparición de síntomas tales como el CEG, el decaimiento o la disnea que son los principales factores que llevan a que un paciente requiera asistencia hospitalaria.

Si bien hay síntomas asociados a mayor y menor riesgo, el factor que más influye a la hora de determinar si un paciente va a tener una trayectoria riesgosa, es el del rango etario al que dicho paciente pertenece. Adicionalmente, se asocia la efectividad de la vacunación a la hora de prevenir estos desenlaces riesgosos.

A continuación, se encuentran las conclusiones asociadas a puntos globales del trabajo, siendo estos en primer lugar los algoritmos implementados, luego los objetivos y resultados declarados, la hipótesis de investigación, las limitaciones y finalmente el trabajo a futuro.

6.1 Algoritmos, técnicas y métricas implementadas

El primer punto para analizar sobre los resultados corresponde a las métricas de desempeño asociadas a los 7 clústeres definidos. En particular, el coeficiente de siluetas asociado a 7 grupos

Con respecto a PCA, se concluye que mejora considerablemente los resultados de una base de datos con varias dimensiones correlacionadas.

Con respecto a las técnicas de clustering y predicción, se concluye que es importante realizar un estudio previo con el fin de identificar cuáles son los algoritmos y técnicas que mejores desempeños pueden tener en función de la base de datos con la que se cuenta y dado los resultados que se buscan obtener. Adicionalmente, si está el tiempo y los recursos para replicar el experimento, iterar con las distintas técnicas seleccionadas, comparar los resultados obtenidos y finalmente, después de eso quedarse con la mejor. Para efectos de este trabajo los mejores resultados se obtuvieron con K-Means y con Support Vector Machine, pero esto no necesariamente va a ser igual en otro problema, con otros datos y con otros resultados esperados.

Con respecto a las métricas de desempeño, tanto las implementadas para el clustering como para la predicción, se concluye la importancia de basarse en más que una sola métrica al momento de hacer interpretaciones. A modo de ejemplo, el coeficiente de siluetas asociado al clustering entregaba un puntaje muy bajo que no permitía discernir bien entre la cantidad óptima de clúster, a diferencia del índice de Davies-Bouldin que, para este caso en particular logró expresar diferencias en sus

puntajes que permitieron seleccionar la cantidad óptima de clústeres. Con respecto a las métricas asociadas a la predicción, considerando 5 días sintomáticos se llegaba a un accuracy del 80%, a pesar de que el modelo no estaba prediciendo bien 2 de las 7 clases que buscaba clasificar. Para hacer frente a eso, hizo falta ver el recall individual asociado a cada clase.

6.2 Objetivos y resultados

A lo largo de este trabajo se encontraron las trayectorias sintomáticas más características de los pacientes infectados por el virus Sars-CoV-2, además de asociar dichas trayectorias a las variables personales de los pacientes que las padecieron.

Se generó un modelo de clasificación que permite predecir los síntomas que tendrá un paciente en el futuro, el cual arroja resultados precisos para pacientes que tendrán síntomas por menos de 2 semanas.

Se dispuso de una herramienta que permite asignar pacientes nuevos a la trayectoria sintomática que más probablemente vayan a tener en el futuro.

A través de estos 3 frutos de la investigación, se cuenta con un insumo más a la hora de entender los pacientes que están en seguimiento, conocer cómo será su evolución y así poder anticiparse a situaciones. Esto permite potenciar decisiones asociadas a tratamiento y seguimiento. Con relación al tratamiento, ver que fármaco implementar para la reducción o disipación de síntomas, además de poder anticiparse a pacientes que potencialmente pudiesen requerir hospitalización o tratamientos más intensivos. Con relación al seguimiento, en base a los síntomas que vaya reportando un paciente y la trayectoria que este vaya desarrollando se puede definir si realizar un seguimiento más exhaustivo, em caso de que alguna paciente este teniendo trayectorias riesgosas o hacer seguimientos menos exhaustivos a pacientes en trayectorias de bajo riesgo, sobre todo en tiempos en los que los recursos y el personal sean escasos.

Adicionalmente, el poder brindarle a un paciente la información sobre cómo será su trayectoria sintomática, la cual fue predicha en base a las trayectorias sintomáticas que tuvieron pacientes con una evolución sintomática similar, siendo así más personalizada y precisa que información generalizada sobre la enfermedad, tiene el potencial de ayudar a tranquilizar a los distintos pacientes del SSMSO sobre cómo será el curso de su enfermedad, combatiendo la incertidumbre asociada a esta.

En conclusión, se cumplen con los objetivos planteados y con los resultados esperados.

6.3 Hipótesis

Recapitulando, las hipótesis de investigación fueron las siguientes:

"*Un registro sintomático es insuficiente para describir la enfermedad y se requiere de un análisis longitudinal sobre los distintos síntomas"*

"*Si bien todas las personas reaccionan de manera diferente al SARS-CoV-2, existen grupos de personas similares que reaccionan en relación con su secuencia de síntomas de manera similar*"

"*Se puede predecir como será la evolución sintomática de una persona en base a los síntomas que ya ha tenido"*

En relación con la primera, en el análisis de los resultados se vio como trayectorias sintomáticas con registros sintomáticos idénticos en ciertos días, convergen a resultados completamente diferentes. A modo de ejemplo, un paciente puede comenzar con cefalea, mialgias y tos seca, pero después de unos días dejar de manifestar estos síntomas y tener anosmia y disgeusia/ageusia. O en vez de dejar de manifestar dichos síntomas, puede que estos prevalezcan por más de dos semanas. A raíz de esto, se concluye que conocer uno o varios síntomas en un tiempo dado no solo contiene parcialmente la información del curso de la enfermedad de un paciente, sino que también puede conducir a conclusiones erradas, a diferencia de conocer una trayectoria sintomática en su totalidad.

En relación con la segunda, a lo largo del trabajo de mostró como los pacientes pertenecientes a un determinado rango etario son más propensos a presentar ciertas trayectorias sintomáticas sobre otras, además de presentar síntomas por menos días en comparación con los pacientes de otros rangos etarios que comparten las mismas trayectorias sintomáticas. Esto sugiere que efectivamente las personas similares tienden a reaccionar de manera similar a la enfermedad. No obstante, por otro lado, a través del modelo predictivo sin considerar ningún día sintomático, es decir, considerando únicamente las variables propias de cada persona, se vio como el nivel de predicción y de asignación a un determinado clúster sintomático era muy bajo. A raíz de esto, se concluye que, si bien la duración de los síntomas y la predominancia de estos es notoria en grupos de personas similares, con los datos considerados de los individuos no se puede conocer en su totalidad cuales van a ser los síntomas exactos que manifestará una persona dado que padece de COVID-19. Las variables consideradas no logran describir con exactitud el sistema inmunológico de los pacientes a la hora de contraer el virus.

En relación con la tercera hipótesis, como bien se desarrollo en la discusión, los modelos predictivos permiten ver como será la evolución sintomática de un paciente en función de los clústeres definidos. Cabe destacar que el desempeño de esta predicción debe ser mejorado para una predicción fidedigna de las trayectorias sintomáticas más largas.

En síntesis, se valida la primera y tercera hipótesis y con los datos actuales no se descarta la segunda, pero de momento no se da por validada.

6.4 Base de datos y limitaciones

Los virus cambian constantemente a través de la mutación. Actualmente ya se conocen diversas variantes del SARS-CoV-2 que causan la enfermedad COVID-19, las cuales potencialmente podrían influir en los síntomas que presenta una determinada persona, la prevalencia de estos síntomas e incluso, en el desenlace que un paciente pudiese tener. La información correspondiente a la variante con la que una persona fue infectada es un dato que no se tiene, por lo que no se pudo incorporar de ninguna manera en la investigación.

Tanto los síntomas como los antecedentes en los que está basada la investigación son auto reportados. El problema con el primer punto es que el auto reporte de síntomas es subjetivo, dado que no hay una medida estandarizada que defina si un paciente presenta o no presenta un determinado síntoma. En base al segundo punto, no se realizan exámenes para corroborar los antecedentes declarados. Por un lado, puede que un paciente tenga cierto antecedente y lo desconozca, por lo que declara no tenerlo. Por otro lado, hay antecedentes tales como el alcoholismo o el uso de drogas que un paciente podría simplemente preferir no declararlo

Además de ser auto reportados son registrados como variables binarias, por lo que se desconoce la severidad de los síntomas o de los antecedentes. A modo de ejemplo, un dolor de cabeza leve está registrado de la misma manera que un dolor de cabeza incapacitante, lo cual no permite ver la magnitud de los cuadros sintomáticos que presentan los pacientes.

Los seguimientos del SSMSO y, en consecuencia, los datos disponibles, están diseñados para chequear los síntomas de los pacientes en las primeras 2 a 3 semanas. Existen casos en los que a ciertos pacientes se les realizó seguimiento hasta que dejaron de presentar síntomas, extendiendo el seguimiento por varias semanas, pero estos presentan muy pocos casos para sacar conclusiones (menos de un 1%). La limitante asociada a este punto es que la literatura señala la existencia de casos crónicos de COVID-19, también conocido como COVID-19 prolongado, donde los síntomas se presentan por varias semanas e incluso meses. Las trayectorias sintomáticas definidas en este trabajo no capturan estos casos.

Finalmente, los clústeres obtenidos a lo largo de este trabajo representan las trayectorias sintomáticas más probables que presentaron los distintos pacientes, en base a los seguimientos realizados por el SSMSO. El hecho de que esta representación de las trayectorias sintomáticas se reduzca a 7 clústeres, deja fuera del estudio las eventuales trayectorias sintomáticas que son menos probables y que no están representadas en los clústeres.

6.5 Trabajo futuro

Cada vez hay más personas con sus campañas de vacunación completas y para vacunas no analizadas en esta investigación, tales como las producidas por AstraZeneca, Pfizer y CanSino. Las trayectorias sintomáticas fueron definidas en base a los síntomas reportados por personas no vacunadas y en base a la validación con los datos de mayo, se pudo ver que compartían las trayectorias sintomáticas con personas que tenían su campaña completa de Sinovac. Sin embargo, este resultado puede ser distinto para distintos tipos de inmunización.

Como se declaró en las conclusiones de las hipótesis, aún falta para poder validar de que personas similares tengan trayectorias sintomáticas similares. El trabajo futuro en esta línea correspondería a incorporar más variables personales, tales como variables sociodemográficas, otras morbilidades registradas en informes clínicos y no solo antecedentes auto reportados, entre otras variables.

Finalmente, otro trabajo podría ser la implementación de estas técnicas para el seguimiento después de ciertas cirugías que tengan el riesgo de complicaciones. Entender cómo van evolucionando los pacientes después de una cirugía riesgosa podría revelar ciertos patrones que ayuden al personal clínico a anticiparse a situaciones graves.

Bibliografía

[1] Página web de la Agencia Nacional de Investigación y Desarrollo https://www.anid.cl/concursos/concurso/?id=379

[2] Página web de la Agencia Nacional de Investigación y Desarrollo https://www.anid.cl/blog/2020/06/01/ministerio-de-ciencia-y-anid-dan-a-conocer-seleccionados-del-fondo-de-investigacion-cientifica-covid-19/

[3] Página web del Servicio de Salud Metropolitano Sur Oriente https://redsalud.ssmso.cl/

[4] Resumen Ejecutivo Servicio de Salud Metropolitano Sur Oriente 2019 https://www.ssmso.cl/FilesComunicaciones/Resumen%20Ejecutivo%20Cuenta%20P%C3%BAblica%20Participativa%20SSMSO%202019.pdf

[5] Página web de la Unidad de Salud Digital del SSMSO https://saluddigital.ssmso.cl/

[6] Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-Pérez, 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 /10.16925/978958760049

[7] Alice Zheng & Amanda Casari. (2018). Feature Engineering for Machine Learning. United States of America: O'Reilly Media http://yvesdaniel.fr/src/share/mlBooksPDF/featureengineeringformachinelearning_1ed.pdf

[8] Zhang S., Zhang J., Zhu X., Qin Y., Zhang C. (2008) Missing Value Imputation Based on Data Clustering. In: Gavrilova M.L., Tan C.J.K. (eds) Transactions on Computational Science I. Lecture Notes in Computer Science, vol 4750. Springer, Berlin, Heidelberg. https://doi-org.uchile.idm.oclc.org/10.1007/978-3-540-79299-4_7

[9] Herman Aguinis, Ryan K. Gottfredson, and Harry Joo. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. SAGE Journal http://www.hermanaguinis.com/ORMoutliers.pdf

[10] Bellman R.E. Adaptive Control Processes. Princeton University Press, Princeton, NJ, 1961.

[11] Chen L. (2009) Curse of Dimensionality. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi-org.uchile.idm.oclc.org/10.1007/978-0-387-39940-9_133

[12] Verma C.V., Ghosh S.M. (2020) Dimensionality Reduction Using PCA Algorithm for Improving Accuracy in Prediction of Cardiac Ailments in Diabetic Patients. In:

Vasudevan H., Gajic Z., Deshmukh A. (eds) Proceedings of International Conference on Wireless Communication. Lecture Notes on Data Engineering and Communications Technologies, vol 36. Springer, Singapore. https://doi-org.uchile.idm.oclc.org/10.1007/978-981-15-1002-1_45

[13] Chow, S., Wan, B. A., Pidduck, W., Zhang, L., DeAngelis, C., Chan, S., Yee, C., Drost, L., Leung, E., Sousa, P., Lewis, D., Lam, H., Chow, R., Lock, M., & Chow, E. (2019). Symptom clústeres in patients with breast cancer receiving radiation therapy. *European journal of oncology nursing : the official journal of European Oncology Nursing Society*, *42*, 14–20. https://doi-org.uchile.idm.oclc.org/10.1016/j.ejon.2019.07.004

[14] Martins, T. D., Annichino-Bizzacchi, J. M., Romano, A., & Filho, R. M. (2019). Principal Component Analysis on Recurrent Venous Thromboembolism. *Clinical and applied thrombosis/hemostasis : official journal of the International Academy of Clinical and Applied Thrombosis/Hemostasis*, *25*, 1076029619895323. https://doi-org.uchile.idm.oclc.org/10.1177/1076029619895323

[15] Salyer, J., Flattery, M., & Lyon, D. E. (2019). Heart failure symptom clústeres and quality of life. *Heart & lung : the journal of critical care*, *48*(5), 366–372. https://doi-org.uchile.idm.oclc.org/10.1016/j.hrtlng.2019.05.016

[16] Holland, S. (2008). PRINCIPAL COMPONENTS ANALYSIS (PCA). PCA (uga.edu)

[17] Gunopulos D. (2009) Clúster and Distance Measure. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi-org.uchile.idm.oclc.org/10.1007/978-0-387-39940-9_618

[18] Eynard D., Javarone M.A., Matteucci M. (2018) Clustering Algorithms. In: Alhajj R., Rokne J. (eds) Encyclopedia of Social Network Analysis and Mining. Springer, New York, NY. https://doi-org.uchile.idm.oclc.org/10.1007/978-1-4939-7131-2_138

[19] Peters G., Lampart M., Weber R. (2008) Evolutionary Rough k-Medoid Clustering. In: Peters J.F., Skowron A. (eds) Transactions on Rough Sets VIII. Lecture Notes in Computer Science, vol 5084. Springer, Berlin, Heidelberg. https://doi-org.uchile.idm.oclc.org/10.1007/978-3-540-85064-9_13

[20] García-Ordás M.T. et al. (2020) Clustering Techniques Performance Analysis for a Solar Thermal Collector Hybrid Model Implementation. In: de la Cal E.A., Villar Flecha J.R., Quintián H., Corchado E. (eds) Hybrid Artificial Intelligent Systems. HAIS 2020. Lecture Notes in Computer Science, vol 12344. Springer, Cham. https://doi-org.uchile.idm.oclc.org/10.1007/978-3-030-61705-9_27

[21] D. L. Davies and D. W. Bouldin, "A Clúster Separation Measure," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224-227, April 1979, doi: 10.1109/TPAMI.1979.4766909.

[22] Ortiz-Prado, E., Simbaña-Rivera, K., Gómez-Barreno, L., Rubio-Neira, M., Guaman, L. P., Kyriakidis, N. C., Muslin, C., Jaramillo, A., Barba-Ostria, C.,

Cevallos-Robalino, D., Sanches-SanMiguel, H., Unigarro, L., Zalakeviciute, R., Gadian, N., & López-Cortés, A. (2020). Clinical, molecular, and epidemiological characterization of the SARS-CoV-2 virus and the Coronavirus Disease 2019 (COVID-19), a comprehensive literature review. Diagnostic microbiology and infectious disease, 98(1), 115094. https://doi.org/10.1016/j.diagmicrobio.2020.115094

[23] Liang, Y., Wang, M. L., Chien, C. S., Yarmishyn, A. A., Yang, Y. P., Lai, W. Y., Luo, Y. H., Lin, Y. T., Chen, Y. J., Chang, P. C., & Chiou, S. H. (2020). Highlight of Immune Pathogenic Response and Hematopathologic Effect in SARS-CoV, MERS-CoV, and SARS-Cov-2 Infection. *Frontiers in immunology*, *11*, 1022. https://doi.org/10.3389/fimmu.2020.01022

[24] Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of internal medicine*, *172*(9), 577–582. https://doi.org/10.7326/M20-0504

[25] Duan, L., Zheng, Q., Zhang, H., Niu, Y., Lou, Y., & Wang, H. (2020). The SARS-CoV-2 Spike Glycoprotein Biosynthesis, Structure, Function, and Antigenicity: Implications for the Design of Spike-Based Vaccine Immunogens. *Frontiers in immunology*, *11*, 576622. https://doi.org/10.3389/fimmu.2020.576622

[26] Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., & Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science (New York, N.Y.)*, *367*(6485), 1444–1448. https://doi.org/10.1126/science.abb2762

[27] Centro Estatal de Vigilancia Epidemiológica y Control de Enfermedades. (2020). Secuelas por COVID-19. Visión CEVECE, Semana 42. https://salud.edomex.gob.mx/cevece/docs/tripticos/2020/Semana42.pdf

[28] Mizrahi, B., Shilo, S., Rossman, H. *et al.* Longitudinal symptom dynamics of COVID-19 infection. *Nat Commun* **11**, 6208 (2020). https://doi.org/10.1038/s41467-020-20053-y

[29] Weinbergerova, B., Mayer, J., Hrabovsky, S. et al. COVID-19's natural course among ambulatory monitored outpatients. Sci Rep 11, 10124 (2021). https://doi-org.uchile.idm.oclc.org/10.1038/s41598-021-89545-1

[30] Alanazi, E., Alashaikh, A., Alqurashi, S., & Alanazi, A. (2020). Identifying and Ranking Common COVID-19 Symptoms From Tweets in Arabic: Content Analysis. *Journal of medical Internet research*, *22*(11), e21329. https://doi.org/10.2196/21329

[31] Larsen JR, Martin MR, Martin JD, Kuhn P and Hicks JB (2020) Modeling the Onset of Symptoms of COVID-19. Front. Public Health 8:473. https://doi.org/10.3389/fpubh.2020.00473

[32] Sudre, C. H., Lee, K. A., Lochlainn, M. N., Varsavsky, T., Murray, B., Graham, M. S., Menni, C., Modat, M., Bowyer, R., Nguyen, L. H., Drew, D. A., Joshi, A. D., Ma,

W., Guo, C. G., Lo, C. H., Ganesh, S., Buwe, A., Pujol, J. C., du Cadet, J. L., Visconti, A., ... Ourselin, S. (2021). Symptom clústeres in COVID-19: A potential clinical prediction tool from the COVID Symptom Study app. *Science advances*, *7*(12), eabd4177. https://doi.org/10.1126/sciadv.abd4177

[33] Zahra, S. A., Iddawela, S., Pillai, K., Choudhury, R. Y., & Harky, A. (2020). Can symptoms of anosmia and dysgeusia be diagnostic for COVID-19?. *Brain and behavior*, *10*(11), e01839. https://doi.org/10.1002/brb3.1839

[34] Lee, D. J., Lockwood, J., Das, P., Wang, R., Grinspun, E., & Lee, J. M. (2020). Self-reported anosmia and dysgeusia as key symptoms of coronavirus disease 2019. CJEM, 22(5), 595–602. https://doi.org/10.1017/cem.2020.420

[35] Carignan, A., Valiquette, L., Grenier, C., Musonera, J. B., Nkengurutse, D., Marcil-Héguy, A., Vettese, K., Marcoux, D., Valiquette, C., Xiong, W. T., Fortier, P. H., Généreux, M., & Pépin, J. (2020). Anosmia and dysgeusia associated with SARS-CoV-2 infection: an age-matched case-control study. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, *192*(26), E702–E707. https://doi.org/10.1503/cmaj.200869

[36] Marshall M. (2021). COVID's toll on smell and taste: what scientists do and don't know. *Nature*, *589*(7842), 342–343. https://doi.org/10.1038/d41586-021-00055-6

[37] Kang, Y. J., Cho, J. H., Lee, M. H., Kim, Y. J., & Park, C. S. (2020). The diagnostic value of detecting sudden smell loss among asymptomatic COVID-19 patients in early stage: The possible early sign of COVID-19. *Auris, nasus, larynx*, *47*(4), 565–573. https://doi.org/10.1016/j.anl.2020.05.020

[38] Schmulson, M., Dávalos, M. F., & Berumen, J. (2020). Beware: Gastrointestinal symptoms can be a manifestation of COVID-19. Alerta: los síntomas gastrointestinales podrían ser una manifestación de la COVID-19. Revista de gastroenterologia de Mexico (English), 85(3), 282–287. https://doi.org/10.1016/j.rgmx.2020.04.001

[39] Xiaodong Yang, Jie Zhao, Qiang Yan, Shangxin Zhang, Yigao Wang, Yongxiang Li, A case of COVID-19 patient with the diarrhea as initial symptom and literature review, Clinics and Research in Hepatology and Gastroenterology, Volume 44, Issue 5, 2020, Pages e109-e112, ISSN 2210-7401, https://doi.org/10.1016/j.clinre.2020.03.013

[40] Kant, R., Chandra, L., Antony, M. A., & Verma, V. (2020). Case of COVID-19 presenting with gastrointestinal symptoms. *World journal of virology*, *9*(1), 1–4. https://doi.org/10.5501/wjv.v9.i1.1

[41] Hentsch L, Cocetta S, Allali G, Santana I, Eason R, Adam E, Janssens J, -P: Breathlessness and COVID-19: A Call for Research. Respiration 2021. https://doi.org/10.1159/000517400

[42] Tang, Q., Liu, Y., Fu, Y. et al. A comprehensive evaluation of early potential risk factors for disease aggravation in patients with COVID-19. Sci Rep 11, 8062 (2021).

https://doi-org.uchile.idm.oclc.org/10.1038/s41598-021-87413-6

[43] Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digit. Med. 4, 3 (2021). https://doi-org.uchile.idm.oclc.org/10.1038/s41746-020-00372-6

[44] Meng, Y., Wu, P., Lu, W., Liu, K., Ma, K., Huang, L., Cai, J., Zhang, H., Qin, Y., Sun, H., Ding, W., Gui, L., & Wu, P. (2020). Sex-specific clinical characteristics and prognosis of coronavirus disease-19 infection in Wuhan, China: A retrospective study of 168 severe patients. *PLoS pathogens*, *16*(4), e1008520. https://doi.org/10.1371/journal.ppat.1008520

[45] Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H. K., Morgan, R., & Klein, S. L. (2020). Impact of sex and gender on COVID-19 outcomes in Europe. *Biology of sex differences*, *11*(1), 29. https://doi.org/10.1186/s13293-020-00304-9

[46] Takahashi, T., Ellingson, M. K., Wong, P., Israelow, B., Lucas, C., Klein, J., Silva, J., Mao, T., Oh, J. E., Tokuyama, M., Lu, P., Venkataraman, A., Park, A., Liu, F., Meir, A., Sun, J., Wang, E. Y., Casanovas-Massana, A., Wyllie, A. L., Vogels, C., ... Iwasaki, A. (2020). Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature*, *588*(7837), 315–320. https://doi.org/10.1038/s41586-020-2700-3

[47] Gozzi, N., Tizzoni, M., Chinazzi, M. et al. Estimating the effect of social inequalities on the mitigation of COVID-19 across communities in Santiago de Chile. Nat Commun 12, 2429 (2021). https://doi-org.uchile.idm.oclc.org/10.1038/s41467-021-22601-6

[48] Mena, G. E., Martinez, P. P., Mahmud, A. S., Marquet, P. A., Buckee, C. O., & Santillana, M. (2021). Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile. Science (New York, N.Y.), 372(6545), eabg5298. https://doi.org/10.1126/science.abg5298

[49] Alejandro I. Canales. La desigualdad social frente al COVID-19 en el Área Metropolitana de Santiago (Chile). Notas de Población N° 111, págs. 13-42 (2020). http://hdl.handle.net/11362/46553

[50] Protocolo de Seguimiento de Casos COVID-19, ANCORA UC http://www.ancorauc.cl/wp-content/uploads/2020/06/Protocolo-de-Seguimiento-de-Casos-Covid-19-Adulto-e-Infantil-v1.4.pdf

[51] Yaohua Wang, Guadalupe M Canahuate, Lisanne V Van Dijk, Abdallah S. R. Mohamed, Clifton David Fuller, Xinhua Zhang, and Georgeta-Elisabeta Marai. 2021. Predicting late symptoms of head and neck cancer treatment using LSTM and patient reported outcomes. In 25th International Database Engineering & Applications Symposium (IDEAS 2021), July 14–16, 2021, Montreal, QC, Canada. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3472163.3472177

[52] Banerjee, A. (2004). "Validating clústeres using the Hopkins statistic". IEEE International Conference on Fuzzy Systems: 149–153. https://doi.org/10.1109/FUZZY.2004.1375706

Anexos

Anexo A. Glosario de síntomas

Síntoma	Descripción
Anorexia	Falta de apetito que origina una negativa a consumir alimentos.
Anosmia	Pérdida total del olfato. Puede ser temporal o crónica.
Calofríos	Indisposición del cuerpo en que se siente algún frío repentino (comienzo de hipotermia) y contracciones musculares, que puede preceder a la fiebre, a alguna emoción o terror, o a otras enfermedades.
Cefalea	Dolor o molestia en la cabeza, el cuero cabelludo o el cuello que a menudo está asociado con tensión de los músculos en estas zonas.
Cianosis	Coloración azulada de la piel debida a una oxigenación insuficiente de la sangre.
Comp. de conciencia	Disminución de la capacidad mental para responder adecuadamente a estímulos ambientales.
Comp. estado gral. (CEG)	Conjunto de manifestaciones que incluyen molestias anímicas y físicas de tipo sistémicas, que reducen las capacidades y el rendimiento del paciente, con repercusión en sus actividades cotidianas.
Congestión nasal	Tejidos nasales y adyacentes y los vasos sanguíneos se inflaman con el exceso de líquido y causan una sensación de "congestión".
Decaimiento	Pérdida progresiva de cualidades, de fuerza o de importancia que afecta a una situación o un hecho.
Diarrea	Heces acuosas y blandas.
Disgeusia/ageusia	La disgeusia de la lengua es un trastorno del gusto que se manifiesta en forma de sensación desagradable y persistente en la boca. La ageusia, un trastorno en el que la persona pierde por completo el gusto.
Disnea	Sensación de falta de aire.
Dolor abdominal	Dolor que se siente en el área entre el pecho y la ingle.
Dolor torácico	Sensación álgida localizada en la zona situada entre el dia- fragma y la fosa supraclavicular.
Fatiga	Trastorno caracterizado por cansancio extremo e incapacidad para funcionar debido a la falta de energía.
Mialgias	Conocida como dolor muscular.
Náuseas	Sensación de malestar en el estómago.
Odinofagia	Dolor en la faringe posterior que se produce con la deglución o sin ella. El dolor puede ser intenso; muchos pacientes rechazan la comida.

Postración	Afección por la que una persona está tan cansada o débil que es incapaz de hacer algo.
Retracción costal	Retracciones indican que una persona está realizando un esfuerzo para respirar. La zona debajo de las costillas, entre las costillas y en el cuello se hunde cada vez que se intenta inhalar.
Rinorrea	Descarga de una secreción mucosa, serosa o purulenta por las narinas (rinorrea anterior) o por las coanas (rinorrea posterior).
Fiebre >37,8°C	Aumento temporal en la temperatura del cuerpo en respuesta a alguna enfermedad o padecimiento.
Taquipnea	Respiración rápida por encima de los 20 ciclos por minuto.
Tos productiva	Una tos productiva es aquella en la que se arroja moco. Este también llamado flema o esputo.
Tos seca	Suele venir provocada por una sensación de cosquilleo en la garganta. La tos seca no produce mucosidad y por lo tanto se denomina tos no productiva.
Vómitos	Salida violenta del contenido del estómago a través de la boca.
Otro	Otros síntomas.

Anexo A: Glosario de síntomas. Fuentes: Medlineplus, MayoClinic y MSDManuals

Anexo B. Glosario de antecedentes

Antecedentes	Descripción
Alcohol	Consumo frecuente de alcohol. Diferente del alcoholismo, el cual es una enfermedad que genera una fuerte necesidad y ansiedad de ingerir alcohol, de forma que existe una dependencia física y psicológica del mismo individuo
Alergias	Una alergia es una reacción de su sistema inmunitario hacia algo que no molesta a la mayoría de las demás personas
Asma	Es una enfermedad crónica que provoca que las vías respiratorias de los pulmones se hinchen y se estrechen. Esto hace que se presente dificultad para respirar como sibilancias, falta de aliento, opresión en el pecho y tos.
Cáncer	El cáncer puede desarrollarse en cualquier parte del cuerpo. Se origina cuando las células crecen sin control y sobrepasan en número a las células normales. Esto hace que al cuerpo le resulte difícil funcionar de la manera que debería hacerlo.
Cardiopatías	La cardiopatía, también denominada arteriopatía coronaria, es una enfermedad progresiva del miocardio o músculo cardíaco. Se da a través del estrechamiento de los pequeños vasos sanguíneos que suministran la sangre y oxígeno al corazón.
Diabetes	Es una enfermedad prolongada (crónica) en la cual el cuerpo no puede regular la cantidad de azúcar en la sangre. Puede ser causada por muy poca producción de insulina, resistencia a la insulina o ambas.

Drogas	Consumo frecuente de drogas. Diferente de la drogadicción que es, la cual es la búsqueda y el consumo compulsivo o incontrolable de la droga a pesar de las consecuencias perjudiciales que acarrea y los cambios que causa en el cerebro, los cuales pueden ser duraderos.
Embarazada	El embarazo es el estado fisiológico de una mujer que comienza con la concepción del feto y continúa con el desarrollo fetal hasta el momento del parto.
Enfermedad Renal	La enfermedad renal crónica, también llamada insuficiencia renal crónica, describe la pérdida gradual de la función renal. Los riñones filtran los desechos y el exceso de líquido de la sangre, que luego se excretan con la orina.
Epoc	La enfermedad pulmonar obstructiva crónica (EPOC) es una enfermedad pulmonar inflamatoria crónica que causa la obstrucción del flujo de aire de los pulmones. Los síntomas incluyen dificultad para respirar, tos, producción de moco (esputo) y sibilancias.
Fibrosis Pulmonar	La fibrosis pulmonar es una enfermedad pulmonar que se produce cuando el tejido pulmonar se daña y se producen cicatrices. Este tejido engrosado y rígido hace que sea más difícil que tus pulmones funcionen correctamente. A medida que la fibrosis pulmonar empeora, se tienen cada vez más dificultades para respirar
HTA	La presión arterial es una medición de la fuerza ejercida contra las paredes de las arterias a medida que el corazón bombea sangre a su cuerpo. Hipertensión es el término que se utiliza para describir la presión arterial alta.
Obesidad	La obesidad es una enfermedad compleja que consiste en tener una cantidad excesiva de grasa corporal.
Paciente en Diálisis	La diálisis es el proceso artificial mediante el cual se extraen los productos de desecho y el exceso de agua del organismo. Este proceso es necesario cuando los riñones no funcionan correctamente.
Púerpera	Postparto o puerperio, mujer que hace muy poco que ha parido.
Tabaco	Consumo frecuente de tabaco. Diferente del tabaquismo, el cual es la adicción al consumo de tabaco
TACO	Tratamiento anticoagulante que previene la formación de coágulos o impide que los ya existentes se hagan más grandes. Los coágulos en las arterias, las venas y el corazón pueden causar ataques al corazón, derrames cerebrales y bloqueos
Vacuna Influenza	Vacuna contra la Influezna
Vacuna Pneumococo	Vacuna contra el Pneumococo

Anexo B: Glosario de antecedentes. Fuentes: Medlineplus, MayoClinic y MSDManuals

Anexo C. Tablas asociadas a los clústeres completas. Un total de 8 tablas:

Pacientes Totales. (1/8)

	Rango etáreo														
	0-19 N = 1503 (15%)			20-39 N = 4049 (40%)			40-59 N = 2881 (29%)			60-79 N = 1430 (14%)			80+ N = 192 (2%)		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Confirmados	789	52%	0	2844	70%	0.1	2090	73%	0	1093	76%	0	154	79%	0.003
Probables	714	48%	0	1205	30%	0.1	791	27%	0	337	24%	0	40	21%	0.003
F	788	52%	0	2261	56%	0.077	1701	59%	0.007	849	59%	0.046	125	64%	0.039
Edad	10.7	6.7	0	29.3	5.6	0	49.4	5.8	0	67.2	5.5	0	85.8	5.7	0
Desenlace															
Hospitalizados	10	1%	0	71	2%	0	177	6%	0	197	14%	0	36	19%	0
Fallecimientos	0	0%	0	2	0%	0	10	0%	0.001	42	3%	0	31	16%	0
Duración Síntomas															
Promedio + Std	7.7	4.1	0	8.9	4.7	0	10.2	6.3	0	11.2	8.2	0	11.8	9.5	0
Antecedentes															
Alcohol	1	0%	0	64	2%	0.001	46	2%	0.011	5	0%	0.003	0	0%	0.238
Alergias	24	2%	0.721	62	2%	0.695	38	1%	0.507	21	1%	0.924	2	1%	0.839
Asma	77	5%	0	112	3%	0.001	84	3%	0.046	74	5%	0	6	3%	0.903
Cardiopatías	2	0%	0.003	1	0%	0	11	0%	0.005	50	3%	0	15	8%	0
Cáncer	1	0%	0.026	2	0%	0	15	1%	0.666	25	2%	0	3	2%	0.083
Diabetes	15	1%	0	57	1%	0	264	9%	0	311	22%	0	50	26%	0
Drogas	3	0%	0.348	25	1%	0.001	9	0%	0.689	0	0%	0.025	0	0%	0.798
Embarazada	2	0%	0.03	52	1%	0	1	0%	0	0	0%	0.005	0	0%	0.581
Enfermedad Renal	0	0%	0.046	1	0%	0	8	0%	0.937	12	1%	0	8	4%	0
EPOC	0	0%	0.002	1	0%	0	10	0%	0.035	35	2%	0	17	9%	0
Fibrosis Pulmonar	0	0%	0.562	0	0%	0.074	0	0%	0.208	5	0%	0	2	1%	0
HTA	17	1%	0	98	2%	0	465	16%	0	604	42%	0	110	57%	0
Obesidad	13	1%	0.012	72	2%	0.497	50	2%	0.774	29	2%	0.288	3	2%	0.874
Paciente en Diálisis	0	0%	0.333	0	0%	0.016	3	0%	0.816	5	0%	0.011	3	2%	0
Púerpera	0	0%	0.933	3	0%	0.128	0	0%	0.646	0	0%	0.903	0	0%	0.063
TACO	1	0%	0.139	3	0%	0.002	3	0%	0.048	12	1%	0	10	5%	0
Tabaco	17	1%	0	220	5%	0	137	5%	0.099	50	3%	0.164	0	0%	0.006
Vacuna Influenza	129	9%	0.002	291	7%	0	293	10%	0.159	326	23%	0	55	28%	0
Vacuna Pneumoco	3	0%	0.051	2	0%	0	3	0%	0	41	3%	0	10	5%	0

Clúster 0. (2/8)

	Rango etáreo														
	0-19			20-39			40-59			60-79			80+		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**225**	**12%**	**0.001**	**739**	**41%**	**0.59**	**542**	**30%**	**0.181**	**271**	**15%**	**0.324**	**32**	**2%**	**0.651**
Confirmados	132	59%	0.053	520	70%	0.97	389	72%	0.693	191	70%	0.013	25	78%	0.963
Probables	93	41%	0.053	219	30%	0.97	153	28%	0.693	80	30%	0.013	7	22%	0.963
F	115	51%	0.721	382	52%	0.013	309	57%	0.308	163	60%	0.825	23	72%	0.447
Edad	11.9	6.4	0.005	29.4	5.4	0.656	49.5	5.7	0.657	67.1	5.2	0.611	86.2	7.5	0.676
Desenlace															
Hospitalizados	2	1%	0.998	13	2%	0.887	28	5%	0.341	33	12%	0.453	6	19%	0.827
Fallecimientos	0	0%	NaN	0	0%	0.805	1	0%	0.757	6	2%	0.56	6	19%	0.838
Duración Síntomas															
Promedio + Std	7.6	3	0.714	8	3.4	0	8.7	4	0	9.6	5.7	0	12.8	11.2	0.528
Antecedentes															
Alcohol	0	0%	0.326	13	2%	0.789	13	2%	0.144	0	0%	0.609	0	0%	NaN
Alergias	4	2%	0.957	12	2%	0.951	12	2%	0.069	8	3%	0.048	0	0%	0.745
Asma	9	4%	0.506	23	3%	0.61	17	3%	0.843	8	3%	0.092	0	0%	0.584
Cardiopatías	1	0%	0.691	0	0%	0.411	1	0%	0.66	9	3%	0.993	1	3%	0.48
Cáncer	0	0%	0.326	1	0%	0.805	2	0%	0.831	7	3%	0.364	1	3%	0.994
Diabetes	4	2%	0.362	10	1%	0.973	39	7%	0.093	58	21%	0.943	10	31%	0.58
Drogas	0	0%	0.934	6	1%	0.626	2	0%	0.869	0	0%	NaN	0	0%	NaN
Embarazada	0	0%	0.691	6	1%	0.28	1	0%	0.425	0	0%	NaN	0	0%	NaN
Enfermedad Renal	0	0%	NaN	0	0%	0.411	2	0%	0.996	4	1%	0.365	1	3%	0.861
EPOC	0	0%	NaN	0	0%	0.411	5	1%	0.034	6	2%	0.954	3	9%	0.835
Fibrosis Pulmonar	0	0%	NaN	0	0%	NaN	0	0%	NaN	1	0%	0.609	0	0%	0.745
HTA	4	2%	0.514	17	2%	0.919	79	15%	0.301	128	47%	0.075	15	47%	0.302
Obesidad	3	1%	0.665	8	1%	0.153	9	2%	0.973	3	1%	0.339	0	0%	0.994
Paciente en Diálisis	0	0%	NaN	0	0%	NaN	1	0%	0.924	1	0%	0.609	0	0%	0.994
Púerpera	0	0%	NaN	1	0%	0.943	0	0%	NaN	0	0%	NaN	0	0%	NaN
TACO	1	0%	0.326	0	0%	0.943	1	0%	0.924	3	1%	0.867	1	3%	0.896
Tabaco	6	3%	0.043	48	6%	0.187	26	5%	0.951	7	3%	0.468	0	0%	NaN
Vacuna Influenza	13	6%	0.134	55	7%	0.827	53	10%	0.798	58	21%	0.598	7	22%	0.5
Vacuna Pneumoco	1	0%	0.934	0	0%	0.805	1	0%	0.924	8	3%	0.913	1	3%	0.896

Clúster 1 (3/8)

	Rango etáreo														
	0-19			20-39			40-59			60-79			80+		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**522**	**22%**	**0**	**827**	**35%**	**0**	**665**	**28%**	**0.3**	**317**	**13%**	**0.127**	**62**	**3%**	**0.009**
Confirmados	238	46%	0	518	63%	0	443	67%	0	237	75%	0.472	49	79%	0.914
Probables	284	54%	0	309	37%	0	222	33%	0	80	25%	0.472	13	21%	0.914
F	286	55%	0.2	444	54%	0.174	390	59%	0.848	177	56%	0.165	40	65%	0.885
Edad	9.6	6.2	0	29.1	5.6	0.26	49.9	5.7	0.005	67.8	5.6	0.048	85.9	6.2	0.866
Desenlace															
Hospitalizados	6	1%	0.177	11	1%	0.373	43	6%	0.762	48	15%	0.479	16	26%	0.114
Fallecimientos	0	0%	NaN	0	0%	0.872	4	1%	0.37	10	3%	0.943	9	15%	0.864
Duración Síntomas															
Promedio + Std	6.2	3.9	0	6.3	3.7	0	7.6	5.9	0	9.1	7.1	0	10.3	11	0.144
Antecedentes															
Alcohol	0	0%	0.748	16	2%	0.448	10	2%	0.967	2	1%	0.673	0	0%	NaN
Alergias	8	2%	0.943	6	1%	0.05	9	1%	0.916	2	1%	0.254	0	0%	0.832
Asma	29	6%	0.666	17	2%	0.201	15	2%	0.307	18	6%	0.753	1	2%	0.71
Cardiopatías	0	0%	0.772	0	0%	0.463	4	1%	0.491	13	4%	0.624	4	6%	0.866
Cáncer	1	0%	0.748	0	0%	0.872	5	1%	0.524	5	2%	0.984	1	2%	0.567
Diabetes	1	0%	0.043	8	1%	0.298	53	8%	0.254	65	21%	0.595	15	24%	0.866
Drogas	1	0%	0.578	7	1%	0.488	3	0%	0.738	0	0%	NaN	0	0%	NaN
Embarazada	1	0%	0.772	13	2%	0.515	0	0%	0.523	0	0%	NaN	0	0%	NaN
Enfermedad Renal	0	0%	NaN	1	0%	0.463	4	1%	0.165	3	1%	0.911	5	8%	0.132
EPOC	0	0%	NaN	0	0%	0.463	0	0%	0.174	4	1%	0.179	5	8%	0.971
Fibrosis Pulmonar	0	0%	NaN	0	0%	NaN	0	0%	NaN	2	1%	0.673	1	2%	0.832
HTA	2	0%	0.081	19	2%	0.896	114	17%	0.459	129	41%	0.571	30	48%	0.148
Obesidad	4	1%	0.993	12	1%	0.515	8	1%	0.303	10	3%	0.165	0	0%	0.567
Paciente en Diálisis	0	0%	NaN	0	0%	NaN	1	0%	0.792	3	1%	0.133	1	2%	0.567
Púerpera	0	0%	NaN	1	0%	0.872	0	0%	NaN	0	0%	NaN	0	0%	NaN
TACO	0	0%	0.748	0	0%	0.872	0	0%	0.792	3	1%	0.911	3	5%	0.832
Tabaco	3	1%	0.218	43	5%	0.805	32	5%	0.98	17	5%	0.061	0	0%	NaN
Vacuna Influenza	49	9%	0.475	48	6%	0.099	50	8%	0.012	59	19%	0.053	16	26%	0.713
Vacuna Pneumoco	0	0%	0.511	0	0%	0.872	0	0%	0.792	13	4%	0.193	3	5%	0.832

Clúster 2 (4/8)

	Rango etáreo														
	0-19			20-39			40-59			60-79			80+		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**266**	**19%**	**0**	**418**	**30%**	**0**	**399**	**29%**	**0.853**	**247**	**18%**	**0**	**51**	**4%**	**0**
Confirmados	127	48%	0.1	293	70%	0.991	291	73%	0.899	194	79%	0.438	39	76%	0.691
Probables	139	52%	0.1	125	30%	0.991	108	27%	0.899	53	21%	0.438	12	24%	0.691
F	119	45%	0.007	196	47%	0	206	52%	0.001	136	55%	0.148	31	61%	0.643
Edad	7.9	6.5	0	29.4	5.6	0.815	49.9	5.9	0.044	68.3	5.8	0	85.6	4.8	0.718
Desenlace															
Hospitalizados	0	0%	0.291	12	3%	0.101	36	9%	0.014	36	15%	0.765	9	18%	0.988
Fallecimientos	0	0%	NaN	1	0%	0.495	2	1%	0.916	8	3%	0.919	10	20%	0.548
Duración Síntomas															
Promedio + Std	7.6	3.8	0.571	8.2	3.4	0.001	9.7	5.9	0.084	10.5	9.7	0.139	10.6	6.6	0.291
Antecedentes															
Alcohol	0	0%	0.397	6	1%	0.965	5	1%	0.708	1	0%	0.666	0	0%	NaN
Alergias	2	1%	0.346	4	1%	0.424	5	1%	0.911	1	0%	0.216	0	0%	0.967
Asma	20	8%	0.072	16	4%	0.215	14	4%	0.55	10	4%	0.471	2	4%	0.942
Cardiopatías	0	0%	0.787	0	0%	0.192	1	0%	0.984	10	4%	0.742	3	6%	0.787
Cáncer	0	0%	0.397	0	0%	0.495	1	0%	0.665	7	3%	0.244	0	0%	0.703
Diabetes	5	2%	0.21	2	0%	0.138	41	10%	0.462	64	26%	0.097	12	24%	0.81
Drogas	0	0%	0.963	2	0%	0.957	2	1%	0.806	0	0%	NaN	0	0%	NaN
Embarazada	1	0%	0.787	2	0%	0.188	0	0%	0.295	0	0%	NaN	0	0%	NaN
Enfermedad Renal	0	0%	NaN	0	0%	0.192	1	0%	0.688	1	0%	0.661	1	2%	0.621
EPOC	0	0%	NaN	1	0%	0.192	1	0%	0.916	8	3%	0.51	4	8%	0.986
Fibrosis Pulmonar	0	0%	NaN	0	0%	NaN	0	0%	NaN	1	0%	0.666	1	2%	0.967
HTA	3	1%	0.753	11	3%	0.898	66	17%	0.872	104	42%	0.98	35	69%	0.066
Obesidad	3	1%	0.884	14	3%	0.018	12	3%	0.059	4	2%	0.801	1	2%	0.703
Paciente en Diálisis	0	0%	NaN	0	0%	NaN	1	0%	0.888	1	0%	0.666	2	4%	0.347
Púerpera	0	0%	NaN	1	0%	0.718	0	0%	NaN	0	0%	NaN	0	0%	NaN
TACO	0	0%	0.397	1	0%	0.718	0	0%	0.888	4	2%	0.274	3	6%	0.924
Tabaco	1	0%	0.335	23	6%	0.962	17	4%	0.709	9	4%	0.959	0	0%	NaN
Vacuna Influenza	32	12%	0.036	21	5%	0.088	48	12%	0.217	49	20%	0.256	15	29%	0.988
Vacuna Pneumoco	1	0%	0.963	1	0%	0.495	0	0%	0.888	4	2%	0.279	2	4%	0.924

Clúster 3 (5/8)

	Rango etáreo														
	0-19			20-39			40-59			60-79			80+		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**28**	**4%**	**0**	**246**	**32%**	**0**	**310**	**41%**	**0**	**159**	**21%**	**0**	**16**	**2%**	**0.814**
Confirmados	18	64%	0.285	194	79%	0.003	250	81%	0.001	126	79%	0.431	14	88%	0.606
Probables	10	36%	0.285	52	21%	0.003	60	19%	0.001	33	21%	0.431	2	12%	0.606
F	12	43%	0.405	169	69%	0	199	64%	0.059	102	64%	0.224	9	56%	0.659
Edad	8.1	13	0.039	31.5	5.5	0	49.3	5.5	0.898	66.7	5.3	0.176	86.8	4.1	0.473
Desenlace															
Hospitalizados	1	4%	0.462	11	4%	0.002	26	8%	0.106	32	20%	0.019	3	19%	0.753
Fallecimientos	0	0%	NaN	0	0%	0.262	1	0%	0.665	4	3%	0.933	2	12%	0.968
Duración Síntomas															
Promedio + Std	18	4.9	0	17.8	7.7	0	19.3	8	0	20.3	9.1	0	22.2	9.3	0
Antecedentes															
Alcohol	0	0%	0	5	2%	0.747	3	1%	0.487	1	1%	0.936	0	0%	NaN
Alergias	0	0%	0.936	2	1%	0.497	3	1%	0.756	4	3%	0.415	1	6%	0.387
Asma	1	4%	0.955	9	4%	0.496	8	3%	0.847	13	8%	0.105	1	6%	0.994
Cardiopatías	0	0%	0.015	0	0%	0.066	2	1%	0.758	3	2%	0.346	4	25%	0.027
Cáncer	0	0%	0	0	0%	0.262	1	0%	0.924	1	1%	0.411	1	6%	0.593
Diabetes	1	4%	0.672	10	4%	0.001	39	13%	0.035	37	23%	0.695	7	44%	0.156
Drogas	0	0%	0.058	2	1%	0.987	0	0%	0.614	0	0%	NaN	0	0%	NaN
Embarazada	0	0%	0.015	3	1%	0.842	0	0%	0.205	0	0%	NaN	0	0%	NaN
Enfermedad Renal	0	0%	NaN	0	0%	0.066	1	0%	0.68	1	1%	0.879	0	0%	0.834
EPOC	0	0%	NaN	0	0%	0.066	1	0%	0.665	5	3%	0.74	2	12%	0.928
Fibrosis Pulmonar	0	0%	NaN	0	0%	NaN	0	0%	NaN	1	1%	0.936	0	0%	0.387
HTA	3	11%	0	9	4%	0.276	64	21%	0.028	70	44%	0.69	13	81%	0.071
Obesidad	0	0%	0.595	4	2%	0.95	4	1%	0.685	3	2%	0.869	1	6%	0.593
Paciente en Diálisis	0	0%	NaN	0	0%	NaN	0	0%	0.741	0	0%	0.936	0	0%	0.593
Púerpera	0	0%	NaN	0	0%	0.442	0	0%	NaN	0	0%	NaN	0	0%	NaN
TACO	0	0%	0	1	0%	0.442	1	0%	0.741	0	0%	0.442	2	12%	0.425
Tabaco	1	4%	0.741	12	5%	0.801	18	6%	0.436	4	3%	0.628	0	0%	NaN
Vacuna Influenza	4	14%	0.455	32	13%	0	54	17%	0	52	33%	0.002	6	38%	0.577
Vacuna Pneumoco	0	0%	0.058	0	0%	0.262	1	0%	0.741	4	3%	0.976	2	12%	0.425

Clúster 4 (6/8)

	Rango etáreo														
	0-19			20-39			40-59			60-79			80+		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**81**	**10%**	**0**	**452**	**54%**	**0**	**247**	**29%**	**0.657**	**59**	**7%**	**0**	**2**	**0%**	**0**
Confirmados	53	65%	0.022	362	80%	0	205	83%	0	44	75%	0.852	2	100%	0.878
Probables	28	35%	0.022	90	20%	0	42	17%	0	15	25%	0.852	0	0%	0.878
F	44	54%	0.813	283	63%	0.002	153	62%	0.367	46	78%	0.005	2	100%	0.754
Edad	13.7	6.4	0	29.1	5.5	0.456	47.8	5.7	0	65	3.9	0.002	82	0	0.338
Desenlace															
Hospitalizados	0	0%	0.956	2	0%	0.039	3	1%	0.001	5	8%	0.311	0	0%	0.814
Fallecimientos	0	0%	NaN	0	0%	0.534	0	0%	0.686	1	2%	0.855	0	0%	0.726
Duración Síntomas															
Promedio + Std	12.3	4.1	0	12.2	3.5	0	13	4.4	0	13.5	5.2	0.031	15	1.4	0.631
Antecedentes															
Alcohol	0	0%	0.048	10	2%	0.346	3	1%	0.814	0	0%	0.508	0	0%	NaN
Alergias	2	2%	0.851	15	3%	0.002	2	1%	0.658	2	3%	0.484	1	50%	0.001
Asma	1	1%	0.17	14	3%	0.762	8	3%	0.906	2	3%	0.74	0	0%	0.072
Cardiopatías	0	0%	0.219	0	0%	0.217	2	1%	0.548	2	3%	0.752	0	0%	0.358
Cáncer	0	0%	0.048	0	0%	0.534	1	0%	0.843	1	2%	0.635	0	0%	0.007
Diabetes	0	0%	0.723	7	2%	0.954	16	6%	0.157	13	22%	0.915	0	0%	0.98
Drogas	1	1%	0.387	2	0%	0.853	2	1%	0.385	0	0%	NaN	0	0%	NaN
Embarazada	0	0%	0.219	8	2%	0.452	0	0%	0.139	0	0%	NaN	0	0%	NaN
Enfermedad Renal	0	0%	NaN	0	0%	0.217	0	0%	0.814	1	2%	0.994	0	0%	0.136
EPOC	0	0%	NaN	0	0%	0.217	0	0%	0.686	0	0%	0.417	1	50%	0.414
Fibrosis Pulmonar	0	0%	NaN	0	0%	NaN	0	0%	NaN	0	0%	0.508	0	0%	0.001
HTA	1	1%	0.653	11	2%	0.886	31	13%	0.13	27	46%	0.671	1	50%	0.6
Obesidad	1	1%	0.805	10	2%	0.581	3	1%	0.689	2	3%	0.775	0	0%	0.007
Paciente en Diálisis	0	0%	NaN	0	0%	NaN	0	0%	0.616	0	0%	0.508	0	0%	0.007
Púerpera	0	0%	NaN	0	0%	0.762	0	0%	NaN	0	0%	NaN	0	0%	NaN
TACO	0	0%	0.048	0	0%	0.762	1	0%	0.616	1	2%	0.994	0	0%	0.202
Tabaco	1	1%	0.653	22	5%	0.65	12	5%	0.939	1	2%	0.684	0	0%	NaN
Vacuna Influenza	4	5%	0.317	38	8%	0.333	22	9%	0.564	18	31%	0.199	2	100%	0.141
Vacuna Pneumoco	0	0%	0.387	0	0%	0.534	1	0%	0.616	2	3%	0.879	0	0%	0.202

Clúster 5 (7/8)

	Rango etáreo														
	0-19			20-39			40-59			60-79			80+		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**177**	**15%**	**0.824**	**672**	**58%**	**0**	**236**	**20%**	**0**	**78**	**7%**	**0**	**1**	**0%**	**0**
Confirmados	93	53%	0.947	459	68%	0.248	154	65%	0.011	58	74%	0.759	1	100%	0.467
Probables	84	47%	0.947	213	32%	0.248	82	35%	0.011	20	26%	0.759	0	0%	0.467
F	101	57%	0.217	405	60%	0.013	162	69%	0.002	47	60%	0.964	1	100%	0.762
Edad	14.3	5.1	0	28.7	5.6	0.001	48.1	5.6	0	66.1	5.7	0.065	81	NaN	0.228
Desenlace															
Hospitalizados	1	1%	0.751	7	1%	0.168	6	3%	0.024	4	5%	0.035	0	0%	0.417
Fallecimientos	0	0%	NaN	0	0%	0.749	0	0%	0.712	3	4%	0.885	0	0%	0.352
Duración Síntomas															
Promedio + Std	8.2	3.5	0.073	8.6	3.2	0.074	9.4	4	0.053	10.3	4.5	0.285	15	NaN	0.632
Antecedentes															
Alcohol	1	1%	0.236	12	2%	0.766	5	2%	0.692	1	1%	0.654	0	0%	NaN
Alergias	3	2%	0.835	15	2%	0.148	4	2%	0.818	1	1%	0.731	0	0%	0
Asma	4	2%	0.097	13	2%	0.19	6	3%	0.878	3	4%	0.778	1	100%	0.007
Cardiopatías	0	0%	0.562	0	0%	0.369	0	0%	0.659	3	4%	0.885	0	0%	0.113
Cáncer	0	0%	0.236	0	0%	0.749	1	0%	0.798	0	0%	0.443	0	0%	0
Diabetes	1	1%	0.83	11	2%	0.709	21	9%	0.976	11	14%	0.123	1	100%	0.579
Drogas	0	0%	0.793	4	1%	0.85	0	0%	0.773	0	0%	NaN	0	0%	NaN
Embarazada	0	0%	0.562	11	2%	0.483	0	0%	0.127	0	0%	NaN	0	0%	NaN
Enfermedad Renal	0	0%	NaN	0	0%	0.369	0	0%	0.841	0	0%	0.844	0	0%	0.021
EPOC	0	0%	NaN	0	0%	0.369	1	0%	0.712	4	5%	0.231	0	0%	0.144
Fibrosis Pulmonar	0	0%	NaN	0	0%	NaN	0	0%	NaN	0	0%	0.654	0	0%	0
HTA	1	1%	0.704	13	2%	0.447	34	14%	0.507	27	35%	0.199	1	100%	0.892
Obesidad	1	1%	0.979	14	2%	0.62	2	1%	0.406	3	4%	0.448	0	0%	0
Paciente en Diálisis	0	0%	NaN	0	0%	NaN	0	0%	0.592	0	0%	0.654	0	0%	0
Púerpera	0	0%	NaN	0	0%	0.997	0	0%	NaN	0	0%	NaN	0	0%	NaN
TACO	0	0%	0.236	0	0%	0.997	0	0%	0.592	1	1%	0.844	0	0%	0.042
Tabaco	1	1%	0.704	36	5%	0.998	15	6%	0.295	4	5%	0.624	0	0%	NaN
Vacuna Influenza	7	4%	0.028	39	6%	0.15	24	10%	0.911	14	18%	0.362	0	0%	0.63
Vacuna Pneumoco	0	0%	0.793	0	0%	0.749	0	0%	0.592	0	0%	0.226	0	0%	0.042

Clúster 6 (8/8)

	Rango etáreo														
	0-19			20-39			40-59			60-79			80+		
Variables	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**	**N**	**%**	**p valor**
Pacientes	**204**	**12%**	**0**	**695**	**41%**	**0.744**	**482**	**28%**	**0.666**	**299**	**17%**	**0**	**30**	**2%**	**0.631**
Confirmados	128	63%	0.002	498	72%	0.395	358	74%	0.381	243	81%	0.032	24	80%	0.877
Probables	76	37%	0.002	197	28%	0.395	124	26%	0.381	56	19%	0.032	6	20%	0.877
F	111	54%	0.593	382	55%	0.639	282	59%	0.833	178	60%	0.998	19	63%	0.944
Edad	12	6.6	0.004	29.4	5.6	0.676	49.5	5.9	0.515	66.9	5.4	0.231	85.5	4.6	0.752
Desenlace															
Hospitalizados	0	0%	0.427	15	2%	0.463	35	7%	0.31	39	13%	0.75	2	7%	0.117
Fallecimientos	0	0%	NaN	1	0%	0.769	2	0%	0.883	10	3%	0.782	4	13%	0.873
Duración Síntomas															
Promedio + Std	8	3	0.268	8.3	3.2	0.001	9.1	3.5	0	10.5	7.3	0.074	9.9	4.5	0.24
Antecedentes															
Alcohol	0	0%	0.287	2	0%	0.005	7	1%	0.938	0	0%	0.548	0	0%	NaN
Alergias	5	2%	0.455	8	1%	0.467	3	1%	0.211	3	1%	0.63	0	0%	0.708
Asma	13	6%	0.484	20	3%	0.944	16	3%	0.668	20	7%	0.237	1	3%	0.624
Cardiopatías	1	0%	0.637	1	0%	0.384	1	0%	0.783	10	3%	0.987	3	10%	0.893
Cáncer	0	0%	0.287	1	0%	0.769	4	1%	0.492	4	1%	0.718	0	0%	0.954
Diabetes	3	1%	0.725	9	1%	0.92	55	11%	0.074	63	21%	0.81	5	17%	0.311
Drogas	1	0%	0.876	2	0%	0.341	0	0%	0.368	0	0%	NaN	0	0%	NaN
Embarazada	0	0%	0.637	9	1%	0.875	0	0%	0.373	0	0%	NaN	0	0%	NaN
Enfermedad Renal	0	0%	NaN	0	0%	0.384	0	0%	0.426	2	1%	0.995	1	3%	0.793
EPOC	0	0%	NaN	0	0%	0.384	2	0%	0.883	8	3%	0.939	2	7%	0.928
Fibrosis Pulmonar	0	0%	NaN	0	0%	NaN	0	0%	NaN	0	0%	0.548	0	0%	0.708
HTA	3	1%	0.891	18	3%	0.854	77	16%	0.968	119	40%	0.371	15	50%	0.545
Obesidad	1	0%	0.83	10	1%	0.558	12	2%	0.231	4	1%	0.471	1	3%	0.954
Paciente en Diálisis	0	0%	NaN	0	0%	NaN	0	0%	0.998	0	0%	0.548	0	0%	0.954
Púerpera	0	0%	NaN	0	0%	0.982	0	0%	NaN	0	0%	NaN	0	0%	NaN
TACO	0	0%	0.287	1	0%	0.982	0	0%	0.998	0	0%	0.152	1	3%	0.967
Tabaco	4	2%	0.396	36	5%	0.816	17	4%	0.204	8	3%	0.489	0	0%	NaN
Vacuna Influenza	20	10%	0.592	58	8%	0.223	42	9%	0.282	76	25%	0.255	9	30%	0.998
Vacuna Pneumoco	1	0%	0.876	1	0%	0.769	0	0%	0.998	10	3%	0.718	2	7%	0.967

Anexo D. Gráficos de calor asociados a los clústeres completos. Un total de 7:

Cluster: 0 / Cantidad= 1809 / Porcentaje= 17.99%

	0_	1_	2_	3_	4_	5_	6_	7_	8_	9_	10_	11_	12_	13_	14_	15_	16_	17_	18_	19_+
Anorexia	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0	0	0	0
Anosmia	0.05	0.05	0.06	0.07	0.09	0.1	0.09	0.09	0.08	0.06	0.04	0.02	0.01	0.01	0.01	0	0	0	0	0
Calofríos	0.03	0.03	0.03	0.03	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0
Cefalea	0.73	0.73	0.71	0.64	0.55	0.44	0.35	0.27	0.21	0.15	0.09	0.06	0.04	0.02	0.01	0.01	0	0	0	0
Cianosis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. de conciencia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. estado gral. (CEG)	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0
Congestión nasal	0.13	0.14	0.13	0.13	0.13	0.12	0.11	0.1	0.08	0.06	0.03	0.03	0.01	0.01	0	0	0	0	0	0
Decaimiento	0.12	0.12	0.12	0.12	0.12	0.11	0.11	0.09	0.07	0.06	0.04	0.03	0.02	0.01	0.01	0	0	0	0	0
Diarrea	0.14	0.14	0.13	0.12	0.11	0.09	0.07	0.06	0.05	0.04	0.02	0.02	0.01	0.01	0	0	0	0	0	0
Disgeusia/ageusia	0.04	0.04	0.05	0.06	0.07	0.08	0.08	0.07	0.06	0.04	0.03	0.02	0.01	0	0	0	0	0	0	0
Disnea	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.04	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01
Dolor abdominal	0.07	0.07	0.07	0.07	0.06	0.05	0.04	0.03	0.03	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0
Dolor torácico	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0
Fatiga	0.03	0.03	0.04	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0
Fiebre >37,8°C	0.29	0.29	0.27	0.23	0.18	0.12	0.07	0.04	0.03	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0
Mialgias	0.97	0.97	0.95	0.87	0.69	0.51	0.35	0.24	0.16	0.11	0.07	0.03	0.02	0.02	0.01	0	0	0	0	0
Náuseas	0.04	0.04	0.03	0.04	0.03	0.03	0.03	0.03	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0	0
Odinofagia	0.28	0.28	0.26	0.24	0.2	0.14	0.1	0.07	0.04	0.03	0.02	0.01	0.01	0.01	0	0	0	0	0	0
Otro	0.07	0.06	0.07	0.07	0.06	0.06	0.06	0.04	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0	0
Postración	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Retracción costal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rinorrea	0.06	0.06	0.06	0.05	0.05	0.04	0.03	0.03	0.02	0.01	0.01	0	0	0	0	0	0	0	0	0
Taquipnea	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tos productiva	0.1	0.1	0.11	0.11	0.11	0.1	0.09	0.08	0.07	0.06	0.04	0.03	0.03	0.02	0.01	0.01	0.01	0	0	0
Tos seca	0.04	0.03	0.03	0.04	0.09	0.11	0.14	0.13	0.11	0.1	0.07	0.05	0.03	0.03	0.01	0.01	0.01	0.01	0.01	0.01
Vómitos	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0

Cluster: 1 / Cantidad= 2393 / Porcentaje= 23.79%

	0_	1_	2_	3_	4_	5_	6_	7_	8_	9_	10_	11_	12_	13_	14_	15_	16_	17_	18_	19_+
Anorexia	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0
Anosmia	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.05	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0	0	0	0
Calofríos	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cefalea	0.49	0.48	0.42	0.36	0.3	0.24	0.18	0.15	0.11	0.08	0.06	0.04	0.03	0.02	0.01	0.01	0.01	0	0	0
Cianosis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. de conciencia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. estado gral. (CEG)	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0
Congestión nasal	0.2	0.2	0.19	0.18	0.15	0.14	0.11	0.08	0.06	0.04	0.03	0.02	0.01	0.01	0.01	0	0	0	0	0
Decaimiento	0.09	0.09	0.09	0.08	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0	0	0	0.01
Diarrea	0.11	0.11	0.1	0.09	0.08	0.06	0.04	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0
Disgeusia/ageusia	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.05	0.04	0.03	0.02	0.01	0.01	0	0	0	0	0	0	0
Disnea	0.05	0.06	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Dolor abdominal	0.06	0.06	0.05	0.05	0.04	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0	0
Dolor torácico	0.05	0.05	0.05	0.05	0.04	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0	0
Fatiga	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0
Fiebre >37,8°C	0.2	0.19	0.16	0.12	0.09	0.06	0.04	0.02	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0	0
Mialgias	0.11	0.09	0.05	0.04	0.05	0.07	0.06	0.06	0.05	0.03	0.03	0.02	0.01	0.01	0.01	0.01	0	0	0	0
Náuseas	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0
Odinofagia	0.27	0.26	0.24	0.2	0.16	0.12	0.08	0.05	0.04	0.03	0.02	0.01	0.01	0.01	0	0	0	0	0	0
Otro	0.1	0.1	0.09	0.08	0.07	0.05	0.05	0.04	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Postración	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Retracción costal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rinorrea	0.09	0.08	0.08	0.07	0.06	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0	0
Taquipnea	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tos productiva	0.11	0.11	0.11	0.1	0.1	0.08	0.08	0.07	0.05	0.05	0.04	0.02	0.02	0.01	0.01	0.01	0.01	0	0	0.01
Tos seca	0.09	0.08	0.05	0.05	0.06	0.09	0.1	0.1	0.08	0.06	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01
Vómitos	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0

Cluster: 2 / Cantidad= 1381 / Porcentaje= 13.73%

	0_	1_	2_	3_	4_	5_	6_	7_	8_	9_	10_	11_	12_	13_	14_	15_	16_	17_	18_	19_+
Anorexia	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0
Anosmia	0.04	0.04	0.05	0.06	0.08	0.09	0.09	0.09	0.07	0.05	0.03	0.02	0.01	0.01	0.01	0	0	0	0	0
Calofríos	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0
Cefalea	0.05	0.04	0.03	0.05	0.09	0.12	0.12	0.11	0.09	0.07	0.05	0.03	0.02	0.01	0.01	0.01	0	0	0	0.01
Cianosis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. de conciencia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. estado gral. (CEG)	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0
Congestión nasal	0.17	0.17	0.17	0.17	0.17	0.14	0.12	0.09	0.07	0.06	0.04	0.02	0.02	0.01	0.01	0.01	0.01	0	0	0
Decaimiento	0.1	0.1	0.1	0.1	0.1	0.1	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.01	0	0	0	0	0
Diarrea	0.09	0.09	0.09	0.08	0.08	0.08	0.06	0.05	0.03	0.03	0.02	0.01	0.01	0	0	0	0	0	0	0
Disgeusia/ageusia	0.03	0.03	0.04	0.05	0.05	0.06	0.07	0.06	0.05	0.03	0.02	0.01	0.01	0.01	0	0	0	0	0	0
Disnea	0.07	0.07	0.07	0.07	0.08	0.07	0.07	0.06	0.07	0.06	0.05	0.04	0.02	0.03	0.02	0.02	0.01	0.01	0.01	0.01
Dolor abdominal	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0	0	0	0	0	0	0	0	0	0
Dolor torácico	0.06	0.06	0.06	0.06	0.06	0.05	0.04	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0
Fatiga	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0.01
Fiebre >37,8°C	0.24	0.24	0.22	0.19	0.14	0.1	0.06	0.04	0.03	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0
Mialgias	0.38	0.37	0.35	0.31	0.27	0.22	0.17	0.12	0.09	0.07	0.05	0.04	0.02	0.01	0.01	0	0	0	0	0.01
Náuseas	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0
Odinofagia	0.25	0.25	0.23	0.2	0.16	0.12	0.09	0.06	0.05	0.03	0.02	0.01	0.01	0.01	0.01	0	0	0	0	0
Otro	0.08	0.08	0.08	0.08	0.08	0.06	0.05	0.06	0.04	0.04	0.03	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Postración	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Retracción costal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rinorrea	0.09	0.09	0.08	0.08	0.06	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0
Taquipnea	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tos productiva	0.03	0.03	0.05	0.06	0.09	0.1	0.11	0.1	0.1	0.08	0.06	0.05	0.04	0.02	0.01	0.01	0.01	0.01	0	0.01
Tos seca	0.96	0.96	0.92	0.83	0.7	0.57	0.41	0.35	0.28	0.23	0.15	0.11	0.08	0.06	0.04	0.03	0.03	0.02	0.02	0.02
Vómitos	0.02	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0	0

Cluster: 3 / Cantidad= 759 / Porcentaje= 7.55%

	0_	1_	2_	3_	4_	5_	6_	7_	8_	9_	10_	11_	12_	13_	14_	15_	16_	17_	18_	19_+
Anorexia	0.04	0.04	0.04	0.04	0.04	0.04	0.06	0.05	0.06	0.06	0.04	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01
Anosmia	0.07	0.07	0.08	0.09	0.1	0.12	0.16	0.18	0.18	0.19	0.18	0.18	0.16	0.13	0.1	0.09	0.07	0.06	0.06	0.06
Calofríos	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0	0.01
Cefalea	0.75	0.75	0.73	0.73	0.72	0.69	0.65	0.64	0.62	0.61	0.57	0.54	0.5	0.45	0.38	0.31	0.27	0.23	0.2	0.21
Cianosis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. de conciencia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. estado gral. (CEG)	0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.03
Congestión nasal	0.11	0.11	0.1	0.1	0.11	0.13	0.13	0.14	0.13	0.13	0.12	0.11	0.11	0.1	0.09	0.08	0.08	0.07	0.07	0.07
Decaimiento	0.16	0.16	0.17	0.18	0.19	0.19	0.22	0.23	0.25	0.25	0.21	0.2	0.19	0.15	0.13	0.1	0.08	0.07	0.07	0.08
Diarrea	0.14	0.15	0.15	0.16	0.16	0.15	0.16	0.16	0.16	0.15	0.14	0.13	0.12	0.1	0.09	0.08	0.07	0.05	0.05	0.07
Disgeusia/ageusia	0.07	0.07	0.08	0.09	0.1	0.11	0.14	0.16	0.17	0.17	0.16	0.16	0.13	0.11	0.1	0.09	0.07	0.06	0.06	0.06
Disnea	0.14	0.14	0.15	0.16	0.17	0.21	0.22	0.24	0.27	0.28	0.31	0.32	0.28	0.26	0.24	0.21	0.18	0.16	0.14	0.16
Dolor abdominal	0.06	0.06	0.06	0.07	0.07	0.06	0.06	0.07	0.07	0.08	0.06	0.07	0.07	0.06	0.04	0.02	0.03	0.02	0.02	0.03
Dolor torácico	0.1	0.1	0.11	0.11	0.11	0.11	0.13	0.15	0.17	0.17	0.18	0.16	0.14	0.13	0.12	0.1	0.08	0.07	0.07	0.08
Fatiga	0.05	0.05	0.06	0.05	0.06	0.08	0.09	0.1	0.11	0.1	0.11	0.1	0.09	0.08	0.09	0.07	0.07	0.06	0.06	0.08
Fiebre >37,8°C	0.29	0.29	0.28	0.27	0.26	0.24	0.19	0.18	0.17	0.15	0.1	0.08	0.06	0.05	0.04	0.03	0.02	0.02	0.02	0.02
Mialgias	0.73	0.73	0.73	0.73	0.71	0.68	0.64	0.64	0.6	0.56	0.52	0.44	0.38	0.32	0.27	0.22	0.17	0.14	0.11	0.12
Náuseas	0.06	0.06	0.06	0.07	0.08	0.08	0.08	0.09	0.09	0.1	0.08	0.08	0.05	0.05	0.04	0.04	0.04	0.03	0.02	0.03
Odinofagia	0.34	0.34	0.33	0.31	0.29	0.26	0.22	0.21	0.19	0.17	0.15	0.13	0.12	0.1	0.09	0.07	0.06	0.06	0.06	0.08
Otro	0.11	0.1	0.1	0.11	0.1	0.1	0.1	0.1	0.1	0.09	0.09	0.08	0.08	0.08	0.07	0.06	0.06	0.06	0.07	0.08
Postración	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Retracción costal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rinorrea	0.11	0.11	0.11	0.1	0.1	0.08	0.08	0.08	0.07	0.05	0.05	0.04	0.05	0.04	0.03	0.03	0.03	0.02	0.03	0.04
Taquipnea	0	0	0	0	0	0	0	0	0.01	0	0	0.01	0	0	0	0	0	0	0	0
Tos productiva	0.08	0.08	0.09	0.1	0.11	0.13	0.13	0.15	0.17	0.16	0.16	0.17	0.16	0.13	0.13	0.11	0.1	0.09	0.08	0.09
Tos seca	0.58	0.58	0.57	0.57	0.57	0.55	0.53	0.53	0.53	0.55	0.54	0.53	0.48	0.43	0.38	0.33	0.31	0.26	0.21	0.21
Vómitos	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.05	0.05	0.04	0.03	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01

Cluster: 4 / Cantidad= 841 / Porcentaje= 8.36%

	0_	1_	2_	3_	4_	5_	6_	7_	8_	9_	10_	11_	12_	13_	14_	15_	16_	17_	18_	19_+
Anorexia	0.02	0.01	0.02	0.02	0.03	0.03	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.01	0	0	0	0	0	0
Anosmia	0.28	0.28	0.32	0.41	0.55	0.7	0.81	0.86	0.84	0.79	0.58	0.44	0.34	0.24	0.17	0.12	0.09	0.05	0.04	0.04
Calofríos	0.03	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0
Cefalea	0.74	0.73	0.72	0.67	0.59	0.51	0.45	0.4	0.37	0.31	0.26	0.19	0.15	0.1	0.08	0.07	0.05	0.04	0.03	0.03
Cianosis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. de conciencia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. estado gral. (CEG)	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0	0.01	0.01	0	0.01	0	0	0	0	0	0	0
Congestión nasal	0.2	0.2	0.2	0.22	0.24	0.25	0.2	0.19	0.16	0.13	0.11	0.07	0.07	0.05	0.04	0.02	0.01	0.01	0.01	0.02
Decaimiento	0.11	0.11	0.12	0.12	0.13	0.12	0.11	0.1	0.09	0.06	0.05	0.04	0.03	0.02	0.01	0	0.01	0.01	0.01	0.01
Diarrea	0.14	0.14	0.14	0.15	0.16	0.13	0.12	0.12	0.1	0.09	0.07	0.05	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.01
Disgeusia/ageusia	0.24	0.24	0.28	0.36	0.48	0.64	0.75	0.82	0.81	0.75	0.55	0.4	0.31	0.22	0.16	0.12	0.09	0.06	0.04	0.03
Disnea	0.07	0.07	0.07	0.07	0.08	0.08	0.1	0.09	0.1	0.1	0.09	0.06	0.05	0.03	0.03	0.02	0.02	0.02	0.01	0.01
Dolor abdominal	0.05	0.05	0.05	0.04	0.05	0.04	0.04	0.05	0.05	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0	0.01	0	0
Dolor torácico	0.05	0.05	0.05	0.05	0.06	0.07	0.06	0.06	0.06	0.06	0.05	0.04	0.04	0.02	0.01	0.01	0.01	0.01	0	0.01
Fatiga	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Fiebre >37,8°C	0.29	0.28	0.25	0.21	0.16	0.11	0.07	0.05	0.03	0.03	0.03	0.02	0.01	0.01	0.01	0	0	0	0	0
Mialgias	0.74	0.74	0.71	0.65	0.58	0.47	0.41	0.34	0.26	0.19	0.14	0.1	0.07	0.05	0.04	0.03	0.02	0.02	0.01	0.01
Náuseas	0.04	0.04	0.04	0.04	0.04	0.05	0.05	0.06	0.06	0.05	0.04	0.03	0.02	0.01	0.01	0	0	0	0	0
Odinofagia	0.41	0.41	0.39	0.35	0.29	0.23	0.17	0.14	0.11	0.08	0.07	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0	0.01
Otro	0.07	0.07	0.07	0.07	0.07	0.06	0.07	0.06	0.06	0.06	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0	0
Postración	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Retracción costal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rinorrea	0.11	0.11	0.11	0.11	0.11	0.08	0.08	0.07	0.06	0.05	0.05	0.04	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01
Taquipnea	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tos productiva	0.07	0.07	0.08	0.09	0.12	0.11	0.12	0.13	0.12	0.11	0.09	0.07	0.05	0.04	0.03	0.03	0.02	0.02	0.01	0.01
Tos seca	0.55	0.55	0.53	0.49	0.45	0.38	0.36	0.33	0.31	0.31	0.26	0.2	0.15	0.12	0.09	0.06	0.04	0.03	0.03	0.03
Vómitos	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0	0

Cluster: 5 / Cantidad= 1164 / Porcentaje= 11.57%

	0_	1_	2_	3_	4_	5_	6_	7_	8_	9_	10_	11_	12_	13_	14_	15_	16_	17_	18_	19_+
Anorexia	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0
Anosmia	0.89	0.9	0.92	0.91	0.85	0.74	0.6	0.51	0.41	0.3	0.18	0.1	0.07	0.04	0.03	0.02	0.02	0.01	0.01	0.01
Calofríos	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cefalea	0.44	0.44	0.41	0.38	0.35	0.28	0.2	0.17	0.13	0.1	0.07	0.04	0.03	0.02	0.02	0.01	0.01	0	0	0
Cianosis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. de conciencia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. estado gral. (CEG)	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0
Congestión nasal	0.23	0.23	0.23	0.21	0.19	0.17	0.14	0.11	0.09	0.07	0.04	0.03	0.02	0.01	0.01	0	0	0	0	0
Decaimiento	0.1	0.1	0.1	0.1	0.08	0.06	0.05	0.05	0.04	0.03	0.03	0.02	0.01	0.01	0.01	0	0	0	0	0
Diarrea	0.09	0.09	0.09	0.08	0.07	0.07	0.06	0.04	0.04	0.03	0.03	0.01	0.01	0.01	0	0	0	0	0	0
Disgeusia/ageusia	0.8	0.81	0.84	0.84	0.76	0.66	0.51	0.43	0.34	0.25	0.14	0.08	0.06	0.04	0.02	0.02	0.01	0.01	0.01	0.01
Disnea	0.05	0.06	0.06	0.06	0.05	0.05	0.04	0.04	0.04	0.03	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0	0
Dolor abdominal	0.04	0.04	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0
Dolor torácico	0.04	0.05	0.05	0.05	0.04	0.04	0.03	0.03	0.02	0.01	0.01	0.01	0.01	0	0.01	0	0	0	0	0
Fatiga	0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0
Fiebre >37,8°C	0.14	0.14	0.12	0.11	0.09	0.06	0.03	0.02	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0
Mialgias	0.34	0.34	0.32	0.29	0.24	0.19	0.12	0.09	0.07	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0	0	0.01	0
Náuseas	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0	0.01	0	0	0	0	0	0	0	0
Odinofagia	0.21	0.2	0.2	0.18	0.15	0.11	0.06	0.04	0.03	0.02	0.02	0.01	0.01	0	0.01	0	0	0	0	0
Otro	0.04	0.04	0.03	0.04	0.04	0.04	0.05	0.05	0.04	0.03	0.02	0.01	0.01	0.01	0	0	0	0	0	0
Postración	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Retracción costal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rinorrea	0.09	0.08	0.08	0.07	0.06	0.05	0.03	0.02	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0	0
Taquipnea	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tos productiva	0.07	0.07	0.07	0.08	0.08	0.08	0.06	0.05	0.05	0.04	0.04	0.03	0.02	0.01	0.01	0	0	0	0	0
Tos seca	0.27	0.28	0.27	0.25	0.23	0.2	0.15	0.13	0.11	0.1	0.07	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01
Vómitos	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0	0.01	0	0	0	0	0	0	0	0	0	0

Cluster: 6 / Cantidad= 1710 / Porcentaje= 17.0%

	0_	1_	2_	3_	4_	5_	6_	7_	8_	9_	10_	11_	12_	13_	14_	15_	16_	17_	18_	19_+
Anorexia	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0
Anosmia	0.07	0.07	0.08	0.09	0.11	0.11	0.11	0.1	0.08	0.06	0.04	0.02	0.01	0.01	0.01	0	0	0	0	0
Calofríos	0.02	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0
Cefalea	0.98	0.98	0.96	0.88	0.72	0.57	0.39	0.31	0.23	0.15	0.09	0.06	0.03	0.01	0.01	0.01	0	0	0	0
Cianosis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. de conciencia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Comp. estado gral. (CEG)	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0
Congestión nasal	0.15	0.16	0.16	0.17	0.17	0.16	0.13	0.1	0.08	0.06	0.04	0.03	0.01	0	0	0	0	0	0	0
Decaimiento	0.08	0.08	0.09	0.1	0.1	0.1	0.1	0.09	0.07	0.05	0.03	0.03	0.01	0.01	0.01	0	0	0	0	0
Diarrea	0.11	0.11	0.11	0.1	0.1	0.07	0.06	0.05	0.05	0.04	0.02	0.01	0.01	0.01	0.01	0	0	0	0	0
Disgeusia/ageusia	0.05	0.05	0.05	0.07	0.08	0.08	0.08	0.08	0.06	0.04	0.02	0.01	0.01	0	0	0	0	0	0	0
Disnea	0.08	0.08	0.08	0.08	0.08	0.08	0.07	0.06	0.06	0.05	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Dolor abdominal	0.06	0.06	0.06	0.05	0.05	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0	0
Dolor torácico	0.08	0.08	0.08	0.07	0.07	0.06	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0
Fatiga	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0
Fiebre >37,8°C	0.25	0.24	0.23	0.2	0.15	0.11	0.06	0.04	0.03	0.02	0.02	0.01	0	0	0	0	0	0	0	0
Mialgias	0.65	0.64	0.62	0.55	0.45	0.35	0.26	0.19	0.14	0.09	0.05	0.02	0.01	0.01	0.01	0	0	0	0	0
Náuseas	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.02	0.01	0.01	0.01	0.01	0	0	0	0	0	0
Odinofagia	0.44	0.44	0.43	0.39	0.3	0.22	0.15	0.1	0.06	0.05	0.02	0.02	0.01	0.01	0	0	0	0	0	0
Otro	0.07	0.07	0.07	0.07	0.08	0.07	0.06	0.06	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0	0	0.01
Postración	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Retracción costal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rinorrea	0.13	0.13	0.12	0.11	0.09	0.08	0.06	0.04	0.03	0.02	0.02	0.01	0	0	0	0	0	0	0	0
Taquipnea	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tos productiva	0.04	0.04	0.04	0.06	0.09	0.1	0.12	0.11	0.1	0.08	0.06	0.04	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0
Tos seca	0.98	0.98	0.95	0.86	0.71	0.56	0.41	0.33	0.26	0.19	0.14	0.09	0.05	0.03	0.02	0.01	0.01	0.01	0.01	0.01
Vómitos	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0	0

Anexo E. Interfaz desarrollada con Streamlit: