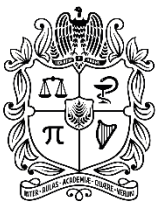


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Un modelo de segmentación sociodemográfica para los consumidores de un portafolio de productos y servicios enfocado en marketing digital

Gustavo Adolfo González Hernández

Universidad Nacional de Colombia

Facultad de Minas, Área Curricular de Ingeniería de Sistemas e Informática

Medellín, Colombia

2023

Un modelo de segmentación sociodemográfica para los consumidores de un portafolio de productos y servicios enfocado en marketing digital

Gustavo Adolfo González Hernández

Trabajo de profundización presentado como requisito parcial para optar al título de:

Magister en Ingeniería – Ingeniería Analítica
Profundización

Director (a):

Ph.D. Fernán Alonso Villa Garzón

Universidad Nacional de Colombia

Facultad de Minas, Área Curricular de Ingeniería de Sistemas e Informática

Medellín, Colombia

2023

Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

_____

GUSTAVO ADOLFO GONZÁLEZ HERNÁNDEZ

Fecha: 31/01/2023

Resumen

La pandemia del COVID-19, declarada como tal a inicios del año 2020, tuvo serios impactos en la vida cotidiana. La población mundial tuvo que adaptarse a nuevas condiciones, entre ellas, al uso de diferentes elementos físicos y tecnológicos con el fin de contener el contagio. Así, el tapabocas se convirtió en el accesorio más usado a nivel mundial y, empresas como las agencias publicitarias, pueden conocer qué personas lo usan a diario por medio de la aplicación de Machine Learning a sus estrategias de Marketing Digital. Para resaltar esta relación, esta tesis propone un modelo de segmentación sociodemográfica para los consumidores de una campaña de marketing digital de la agencia de publicidad “A” que ofrece al público un portafolio de tapabocas quirúrgicos desechables. Los datos utilizados son los aportados por el Instituto Nacional de Salud y el Ministerio de Salud sobre los casos positivos de Covid-19 en Colombia desde el 2 de marzo de 2020 hasta el 21 de diciembre de 2022. Metodológicamente, se realiza una clusterización bajo el método k-means. Como resultado, se obtuvieron 5 clústeres a partir de la programación por aprendizaje no supervisado de Machine Learning. Se concluye que la aplicación de modelos de Machine Learning L al Marketing Digital resulta ser efectiva para la clasificación de posibles grupos de usuarios de productos y servicios que se puedan ofrecer por estas plataformas como Facebook e Instagram.

Palabras clave: Machine Learning, Aprendizaje No Supervisado, Clusterización, Marketing digital, Pandemia.

Abstract

A sociodemographic segmentation model for consumers of a portfolio of products and services focused on digital marketing

The COVID-19 pandemic, declared as such at the beginning of 2020, had serious impacts on everyday life. The world population had to adapt to new conditions, including the use of different physical and technological elements to contain the virus. Thus, the mask became the most used accessory worldwide and, companies such as advertising agencies, can know which people use it daily through the application of Machine Learning to their Digital Marketing strategies. To highlight this relationship, this thesis proposes a sociodemographic segmentation model for the consumers of a digital marketing campaign of the advertising agency "A" that offers the public a portfolio of disposable surgical masks. The data used are those provided by the National Institute of Health and the Ministry of Health on positive cases of Covid-19 in Colombia from March 2, 2020 to December 21, 2022. Methodologically, a clustering is performed under the k-means method. As a result, five clusters were obtained from Machine Learning unsupervised learning programming. It is concluded that the application of Machine Learning L models to Digital Marketing is effective for the classification of groups of users of products and services that can be offered by these platforms such as Facebook and Instagram.

Keywords: Machine Learning, Unsupervised Learning, Clustering, Digital Marketing, Pandemic.

Contenido

Pág.

Lista de figuras

Pág.

Lista de tablas

Pág.

Lista de Símbolos y abreviaturas

ML:	Machine Learning
SML:	Supervised Machine Learning
USML:	Unsupervised Machine Learning
IA:	Inteligencia Artificial
INS:	Instituto Nacional de Salud

Introducción

La globalización, que podemos interpretar como la estrecha relación entre puntos geográficos y sus distintas dinámicas sociales, políticas y económicas, hace que sin lugar a duda, tejamos nuestras relaciones sociales en un mundo interconectado (Held, 2002; Santos, 1993). Lo que pasa en un país al otro lado del mundo, tiene directas repercusiones en lo que podemos hacer (o no) dónde nos ubicamos. Esta hiperlaxitud en nuestras redes de comunicación como geográficas, económicas y políticas, permiten que la información sea un recurso altamente valioso para ejecutar cualquier acción. Una muestra de ello fueron las acciones tomadas por los gobiernos de distintos países para contener la pandemia por el virus COVID-19.

La pandemia del COVID-19 fue declarada como tal a inicios del año 2020 y tuvo serios impactos en la economía mundial, en la acción gubernamental y en cómo se debería enfrentar un desafío en términos de salud pública (Benton et al., 2021). Por unanimidad, las fronteras geográficas de cada uno de los países del mundo estuvieron cerradas por aproximadamente un año, permitiendo el acceso solamente a aquellos connacionales atrapados en otros países y que deseaban regresar a sus hogares. No sólo las fronteras nacionales se cerraron: nuestras casas se convirtieron a su vez en una frontera restringida, en la cual tuvimos que acostumbrarnos a continuar con el curso de nuestras actividades matutinas y nuestra movilidad por fuera de ella se limitó al abastecimiento de alimentos y productos de primera necesidad, así como en la asistencia a centros médicos en caso de contagio.

Así las cosas, toda la población mundial tuvo que adaptarse a las nuevas condiciones a las que este virus nos llevó no sólo en términos de movilidad, de libertad de locomoción o de restricción al acceso a diferentes equipamientos como colegios, edificios, museos, etc., sino que también nos empujó al uso cotidiano de diferentes elementos digitales con el fin de contener el contagio. Un ejemplo de esto fue el desarrollo e implementación, por parte

de los gobiernos, de aplicaciones móviles para rastrear el virus mediante la información aportada por los ciudadanos sobre su estado de salud y así, tomar decisiones de política pública que pudiesen contener o eliminar el Covid-19 con éxito. Algunas de estas aplicaciones fueron “Cuidar Covid-19” en Argentina, “Coronavirus SUS” en Brasil y “CoronApp” en Chile y Colombia (Vargas, 2021).

Adicional a los desarrollos digitales, también se hizo necesaria la adquisición de productos y elementos de bioseguridad como geles antibacteriales para las manos, sprays desinfectantes, ropa impermeable de protección, guantes, protectores faciales y, finalmente, el tapabocas. El tapabocas se convirtió en el accesorio más usado a nivel mundial, debido a la obligatoriedad de su uso en exteriores por parte de los gobiernos y como una de las tácticas más efectivas, según la Organización Mundial para la Salud, para la contención del virus Covid-19 (World Health Organization, 2021, 2022a). Se estima que entre 2020 y 2021 se fabricaron 4,5 billones de tapabocas N95 en el mundo (Fernandez, 2021) y muchas de las industrias textiles, severamente golpeadas por el cierre de las fábricas y los centros de trabajo en medio de los confinamientos estrictos por el virus, sustituyeron la producción de sus habituales productos por la de tapabocas.

Con la llegada de las vacunas y su distribución acelerada a nivel mundial[1] (Organización Mundial de la Salud, 2021, 2022), la mortalidad del virus disminuyó y con ella, también se relajaron las medidas preventivas para su contención. Así las cosas, el uso de tapabocas se convirtió en una opción. Pese a los anuncios de la circulación de nuevas cepas[2] de COVID-19, muchas de ellas agresivas y altamente contagiosas, medidas como el confinamiento estricto y el uso de tapabocas obligatorio sólo sea vuelto a registrar en países muy específicos como China (Deutsche Welle, 2022). En Colombia, por ejemplo, ante el anuncio de 2022 del Ministerio de Salud del uso obligatorio de tapabocas en las calles, se presentó un rechazo por gran parte de la población ante la medida, por lo que esta entidad tuvo que retractarse de la misma (León, 2022).

[1] COVAX fue uno de los mecanismos que se establecieron para la distribución equitativa y rápida de las vacunas contra el Covid-19 en países en vías de desarrollo, impulsado por la ONU y OMS, entre otras organizaciones internacionales. Colombia se convirtió en el primer país beneficiado por este mecanismo, recibiendo un lote de 117.000 dosis el 1 de marzo de 2021 (Organización Panamericana de la Salud, 2021).

[2] En 2022, “El perro del infierno” fue la última cepa en circulación, con un alto potencial de contagio (World Health Organization, 2022b)

Pese al aparente rechazo unánime por parte de la población colombiana a retornar al uso obligatorio del tapabocas en las calles, es posible afirmar que el uso voluntario del mismo no se ha descartado, ya sea por costumbre o considerar que su uso es muy importante no sólo para la contención del virus del Covid-19, sino también para cuidar su salud ante otros posibles virus u otras afecciones respiratorias por cuenta de la polución.

En este sentido, las agencias publicitarias juegan un rol importante con relación entre la globalización y el manejo por parte de los gobiernos de la pandemia del virus por Covid-19. La globalización ha hecho posible que los datos de todos los habitantes del planeta que se encuentran en línea, principalmente en las redes sociales, puedan ser utilizados para develar las preferencias de las personas con respecto a sus propias técnicas de contención del virus. Así las cosas, las agencias publicitarias pueden conocer, por ejemplo, qué grupos específicos de personas están dispuestas aún a usar un tapabocas en su cotidianidad.

Sin embargo, esta tarea no puede ser realizada rudimentariamente. Captar y procesar miles de millones de datos es una tarea titánica y, por lo tanto, es necesario emplear herramientas analíticas certeras. Es aquí donde el ML juega un rol fundamental: es a través de éste que, gracias a algoritmos, se puede programar el aprendizaje de una computadora para que procese los datos que sean requeridos, sin importar su volumen.

Con el fin de resaltar esta relación entre el marketing y la publicidad con el ML, esta tesis propone un modelo de segmentación sociodemográfica para los consumidores de un portafolio de productos y servicios enfocado en marketing digital. Para lograr este propósito, la investigación se sitúa en el contexto de la pandemia por el Covid-19 y el uso del tapabocas como elemento de protección y bioseguridad cotidiano. Así las cosas, se toma una campaña de marketing digital de la agencia de publicidad “A”[3] que debe ofrecer al público un portafolio de tapabocas quirúrgicos desechables para realizar la segmentación ya propuesta. Los datos en los que se sustenta esta investigación son los aportados por el Instituto Nacional de Salud y el Ministerio de Salud sobre los casos

[3] Para los efectos de esta investigación, el nombre de la agencia de publicidad conserva el principio de confidencialidad y, por tanto, su nombre es sustituido en esta tesis.

positivos de Covid-19 en Colombia desde el 2 de marzo de 2020 hasta el 21 de diciembre de 2022.

Metodológicamente, ese trabajo realiza una clusterización bajo el método k-means, donde se procesaron 6’345.115 registros en 23 variables de la base de datos original para transformar este mismo número de registros en 65 variables dummy, qué replicaban la información de 10 variables seleccionadas como fundamentales y explicativas del modelo de segmentación sociodemográfica. Como resultado, se obtuvieron 5 clústeres que, a partir de la programación por aprendizaje por machine learning de aprendizaje no supervisado, agruparon los sujetos y sus características bajo un patrón. este modelo de segmentación se aplicó a una campaña de marketing digital Facebook y e Instagram para promover y uso de tapabocas desechables como método de prevención ante el contagio por covid-19 y otras afecciones respiratorias.

En consecuencia, esta tesis se divide en 4 partes. Primero, se realiza un diseño de investigación que incluye el planteamiento del problema, la revisión de literatura, el marco teórico y un marco conceptual sobre el ML y su aplicación en el marketing digital; seguido por un apartado en el cual se describen los datos del dataset original y se describe la metodología. El tercer apartado está constituido por los asuntos procedimentales, donde se evidencia el análisis descriptivo de las variables y el proceso de clusterización, además de la implementación del modelo de segmentación sociodemográfica a la campaña de marketing digital y sus resultados. Esta tesis finaliza con un apartado de conclusiones sobre la investigación.

1. Diseño de investigación

1.1 Planteamiento del problema

Las agencias de publicidad, habitualmente, construyen sus campañas a partir de “prácticas de ensayo y error humano”, donde sus colaboradores, con base en un presupuesto, construyen audiencias y toman decisiones según sus conocimientos empíricos.

Los clientes de las agencias solicitan procesos o alternativas que les permita optimizar los recursos de sus campañas, predecir el comportamiento de los consumidores e identificar patrones con el fin tener una mayor conexión con la audiencia en cuanto a marca, posicionamiento y precio, para que sea transformado en una compra, recordación o una conversación.

Sin embargo, es imprescindible anotar que las campañas de publicidad, además de ajustarse a las condiciones que demandan sus clientes, también deben considerar las coyunturas sociales, políticas y económicas en las que éstas se desarrollan. Un hecho como la pandemia por Covid-19, donde la afectación de la vida de las personas se dio a escala global en todas las áreas de la cotidianidad, sin duda representó un giro de 180° en cómo las campañas publicitarias se llevan a cabo y de cómo la información puede ser captada para determinar cuáles son las necesidades de los potenciales consumidores.

En ese sentido, la aplicación del ML y de los diversos recursos de la IA disponibles en la actualidad son fundamentales para que las metas de las

campañas publicitarias se cumplan y tengan el impacto esperado por los clientes. Entonces, la optimización de recursos y el enfoque más certero de las campañas hacen que este tipo de herramientas transformen la forma de ofrecer y consumir publicidad.

Con base en lo previamente establecido, esta tesis se propone responder cómo, a través de un modelo de segmentación sociodemográfica para los consumidores de un portafolio de productos y servicios enfocado en marketing digital, se potencia el uso de ML y, con ello, se incrementa el rendimiento y se optimizan los recursos de la campaña liderada por la agencia de publicidad “A” respecto al uso del tapabocas en Colombia en el escenario de la pandemia y postpandemia por Covid-19.

Éste se constituye como un problema real dados los eventos recientes en los que el Ministerio de Salud de Colombia propuso el retorno del uso obligatorio del tapabocas en los espacios públicos abiertos y cerrados (León, 2022; Portafolio, 2022). Pese al revés de la medida, es de interés para la agencia de publicidad “A” identificar, con el método de clusterización del ML, cuáles grupos de la población colombiana podrían ser potenciales consumidores de un producto como el tapabocas quirúrgico desechable, todo a través de una campaña de marketing digital en redes sociales.

Este problema está, entonces, estrechamente relacionado con la salud pública en Colombia, con el manejo del virus Covid-19 tanto en el contexto de pandemia más estricto (2020-2021) como el de postpandemia (2022) y el ML como optimizador de las campañas de marketing digital en redes sociales.

1.2 Objetivos

1.2.1 Objetivo General

Proponer un modelo de segmentación sociodemográfica para los consumidores de un portafolio de productos y servicios enfocado en marketing digital.

1.2.2 Objetivos específicos

a) Caracterizar las variables requeridas para la segmentación sociodemográfica para los consumidores de un portafolio de productos y servicios enfocado en marketing digital.
b) Implementar un modelo para la segmentación sociodemográfica de los consumidores de un portafolio de productos y servicios enfocado en marketing digital.
c) Validar el modelo implementado de segmentación sociodemográfica para los consumidores de un portafolio de productos y servicios enfocado en marketing digital.

1.3 Revisión de trabajos previos

La bibliografía sobre los estudios de Machine Learning aplicados al marketing es amplia, especialmente en la academia norteamericana. Esta tesis seleccionó once trabajos académicos que abordan la materia, los cuales se presentan brevemente a continuación.

La importancia de la aplicación de la IA en el marketing a través del ML es abordada por Ma y Sun (2020), cuyo estudio comprueba el gran potencial que tiene el ML para entender el comportamiento de managers y consumidores, así como aumentar la capacidad de toma

de decisiones automatizadas que mejoren el desempeño de los negocios con base en una alta precisión predictiva. Sin embargo, los investigadores destacan los retos que implica la implementación del ML en el marketing en términos de interpretación, debido a su estructura de modelos flexibles (Ma & Sun, 2020).

Bayoude et al. (2018b) destacan la función del ML en el marketing a través de los factores de oferta (tecnología, situaciones ligadas al sector financiero) y demanda (rentabilidad, competitividad y regulación), así como las ventajas de su aplicación en cuanto a su alta predictibilidad de comportamiento de consumidores (por medio de análisis predictivo y modelos de propensión), alta predictibilidad del mercado, habilitación de uso de chatbots y despliegue de publicidad.

Lo anterior también es abordado por Ngai y Wu (2022), ahondando en la aplicación del ML en el marketing en aspectos como la personalización del producto (precio, recomendaciones, manejo de marcas y predicción de decisiones de compra), su promoción (publicidad, chatbots y predicciones de demanda) y en la caracterización del consumidor (targeting y engagement). Por otra parte, Hair Jr. y Sarstedt (2021), retomando la importancia de la aplicación del ML en el marketing, recomiendan combinar la capacidad predictiva del ML con métodos de investigación causales-explicatorios, con el fin de que la inferencia causal destaque el diseño experimental aplicado al marketing.

A través de la aplicación de USML, Ebrahimi et al. (2022) prueban cinco diferentes dimensiones del marketing en redes sociales (‘entretenimiento’, ‘personalización’, ‘interacción’, ‘poder del voz a voz’ y ‘tendencias de compra’) y cómo influencian a los consumidores en sus decisiones de compra. Este estudio analiza el caso de consumidores de la plataforma Marketplace de Facebook en Hungría aplicando dos métodos de clusterización: el Análisis Jerárquico de Clúster (HCA) y K-Medios. Las conclusiones de este trabajo son: 1) los cinco factores analizados sí influencian la decisión de compra de los consumidores de la plataforma; y 2) la clusterización de los consumidores, acorde con sus características demográficas, son determinantes para responder a la pregunta sobre el comportamiento de compra de bienes y servicios. El estudio resalta, a su vez, la necesidad de que este enfoque de clusterización sea adoptado por parte de los managers de marketing, con el fin que la segmentación del mercado y la focalización en un conjunto determinado de consumidores incrementen la eficiencia de sus negocios en línea.

El trabajo de Chen et al. (2022) realiza un marco conceptual sobre la relación entre inteligencia artificial y el marketing business to business. Ese trabajo determina que tradicionalmente, la inteligencia artificial se utiliza para las actividades de marketing y presión externa que sean impuestas por la informatización. En ese sentido, se identificó que la mejora en la eficiencia, la mejora en la precisión, la mejor toma de decisiones, el mejoramiento de la relación con el consumidor, el incremento en las ventas, las reducciones en los costos y en los riesgos, es producida a través de la adopción de la inteligencia artificial en el marketing business to business. Adicionalmente, De Mauro, Sestino y Bacconi (2022) resaltan en la misma vía de Chen et al. (2022) las grandes ventajas que tiene el uso de la inteligencia artificial en los procesos de marketing digital. Sin embargo, a diferencia de los anteriores autores, De Mauro, Sestino y Bacconi sí destacan las técnicas de machine learning en esta actividad coma de la cual proponen una taxonomía en casos específicos de uso en marketing digital. entonces, rescatan las áreas fundamentales de aplicación del machine learning en el marketing digital: impacto financiero, toma de decisiones, experiencia del consumidor y fundamentos de la compra.

El trabajo de Guarda et al. (2012) reitera la necesidad de que la administración de los negocios relacionados con marketing digital se actualicen, con el fin de que puedan responder a las necesidades de los clientes en la misma medida en la que la información circula actualmente. Entonces, plantea la importancia de la aplicación del “Business Intelligence” y del “Marketing intelligence” para mejorar la toma de decisiones en el campo por medio de un proceso de cinco etapas: planificación, recolección de datos, análisis de datos, representación y proyección.

Hagen et al. (2020) complementa la anterior idea, exponiendo cómo el ML es un auxiliar indispensable de la investigación sobre el comportamiento aplicada al marketing. Así las cosas, explica de qué se trata el ML, cómo funciona en términos de la investigación comportamental en marketing, cuáles son sus diferentes tipos (aprendizaje supervisado, no supervisado y semi supervisado) e ilustra casos reales en los cuales éstos funcionan, como la estimación de efectos de tratamiento heterogéneos, el muestreo de estímulos y la evaluación de dinámicas comportamentales.

Por su parte, Kaličanin et al. (2019) hace una aproximación, a manera introductoria, de cómo se relacionan la IA, el ML y el Deep Learning. Asimismo, se refieren a los algoritmos del ML (regresión logística, árbol de decisión, vector de apoyo, bosque aleatorio y perceptrón multicapa) y a las partes del "Marketing Intelligence" (inteligencia del competidor y del producto, entendimiento del mercado y del consumidor).

Finalmente, Ullal et al. (2021) ejemplifica la aplicación del ML al marketing digital mediante el análisis de las respuestas de los consumidores de la India, clasificados en varios perfiles demográficos, a una serie de máquinas que les vendían productos, Las variables que determinaron la eficiencia del ML en esta investigación fueron la "actitud" y el "comportamiento" de los clientes. Más allá del proceso de decodificacón que se expone en este trabajo, se hace énfasis en la comprensión de categorías teóricas básicas en el área, como el análisis de la actitud del consumidor, el análisis comportamental y el análisis de la decisión.

1.4 Marco conceptual

1.4.1 Marketing

El marketing se define como un "conjunto de actividades destinadas a satisfacer las necesidades y deseos de los mercados meta a cambio de una utilidad o beneficio para las empresas u organizaciones que la ponen en práctica" (Viteri Luque et al., 2017, p. 975)

1.4.2 Publicidad

La publicidad consiste en el acto de "informar a una o varias personas sobre un producto o servicio por medio de un anuncio pagado, con la intención de conseguir un objetivo" (Erickson, 2010, p. 15)

1.4.3 Campaña publicitaria

La campaña publicitaria es “un plan de publicidad amplio para una serie de anuncios diferentes, pero relacionados, que aparecen en diversos medios durante un periodo específico” (Guzmán Elisea, 2003, p. 9). De esta forma, este plan se constituye como un instrumento estratégico de corto plazo para alcanzar un objetivo o hallar la solución a un problema (Guzmán Elisea, 2003).

1.4.4 Posicionamiento de marca

El posicionamiento de marca es el “Instrumento fundamental tanto para la propuesta de valor como para la estrategia de comunicación seleccionadas por las organizaciones, con el propósito de crear y mantener ventaja competitiva” (Olivar Urbina, 2021, p. 56). Asimismo, el posicionamiento de marca relacionan “al individuo y al mercado; se refiere a la ubicación concreta y definitiva que logra un producto, una marca o una organización en la mente de las personas a quienes va dirigido” (Olivar Urbina, 2021, p. 56).

1.4.5 Redes sociales

Desde una perspectiva amplia, las redes sociales se definen como los “sistemas de vínculos entre entidades sociales” (Luna, 2004, p. 59). Estos sistemas de vínculos dan cuenta, a su vez, del comportamiento de los individuos de una sociedad y las interacciones que ocurren entre sí (McPherson et al., 1992, p. 9545).

Por otro lado, desde una perspectiva más acotada, esta tesis adopta la definición de redes sociales como aquellas “plataformas digitales formadas por comunidades de individuos con intereses, actividades o relaciones en común […] [, que] permiten el contacto entre personas y funcionan como un medio para comunicarse e intercambiar información” (Concepto, n.d.).

1.4.6 Indicador sociodemográfico

Los indicadores sociodemográficos son datos que describen condiciones, y con ello, tendencias en la realidad social en un lapso de tiempo determinado (Departamento

Nacional de Planeación, 2017). Datos como los grupos según edad y género, la distribución territorial, las condiciones de vivienda y del hogar, educación, economía (índices de pobreza, población económicamente activa, inactiva o en edad de trabajar), salud, entre otros, se consideran indicadores demográficos (Instituto Nacional de Estadística e Informática de la República del Perú, 2018; Jara, 2015)

1.4.7 Inteligencia artificial

La inteligencia artificial se define como “la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano” (Petteri, 2018, p. 17). Asimismo, se habla de un agente operado por IA cuando “tiene la posibilidad de asistir a las personas y actuar en nombre propio” (Kaličanin et al., 2019, p. 473)

1.5 Marco teórico

1.5.1 Machine Learning

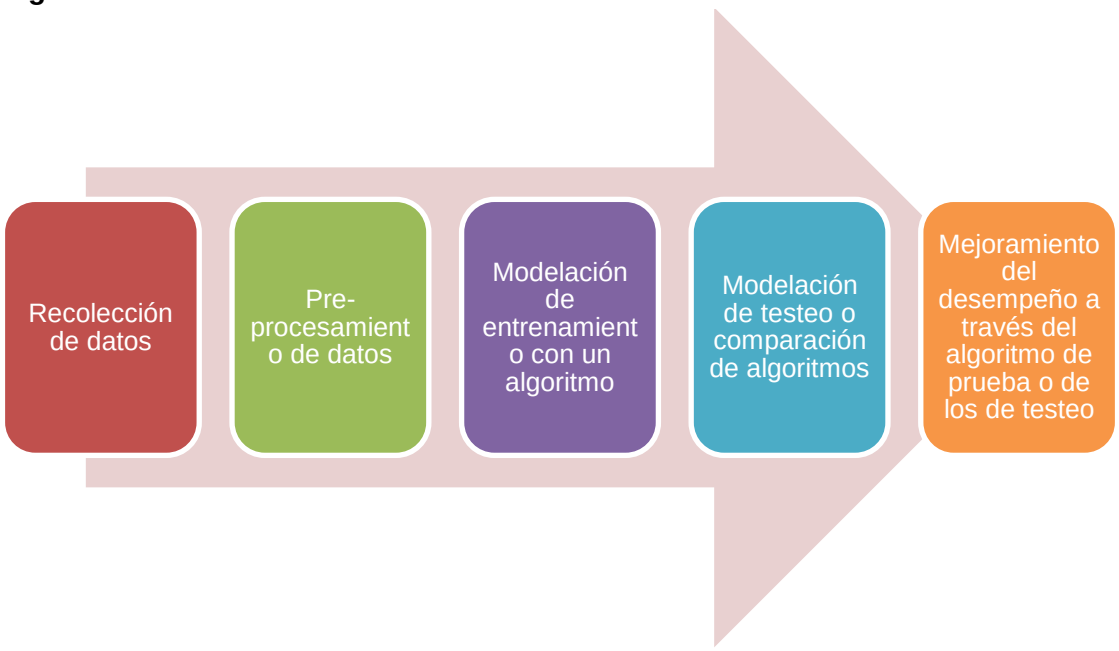
El ML es “un aspecto de la informática en el que los ordenadores o las máquinas tienen la capacidad de aprender sin estar programados para ello” (Petteri, 2018, p. 19). Este aprendizaje automático se vale del uso de algoritmos predictivos complejos para aprender datos y sus patrones específicos, con el fin de generar modelos matemáticos precisos (Bayoude et al., 2018b; Petteri, 2018). Este componente de aprendizaje automatizado es el que diferencia al ML de la IA, aunque ambos cumplan el mismo objetivo (Chinnamgari, 2019).

Los primeros desarrollos de la IA y del ML se llevaron a cabo en la década de los cincuenta, aunque dieron un salto significativo a lo que conocemos hoy en la década de 1990

(Kaličanin et al., 2019). En ese sentido, el ML hace parte de la Ciencia de Datos, de la Minería de Datos y del Big Data (Chinnamgari, 2019).

El ML tiene un proceso de cinco etapas en las que se desarrolla el manejo de los datos (Figura 1-1): i) recolección de datos, ii) pre-procesamiento de datos o transformación para la modelación, iii) modelación de entrenamiento con un algoritmo, iv) modelación de testeo o comparación de algoritmos, v) mejoramiento del desempeño a través del algoritmo de prueba o de los de testeo (Bayoude et al., 2018a).

Figura 1–1: Proceso de ML.

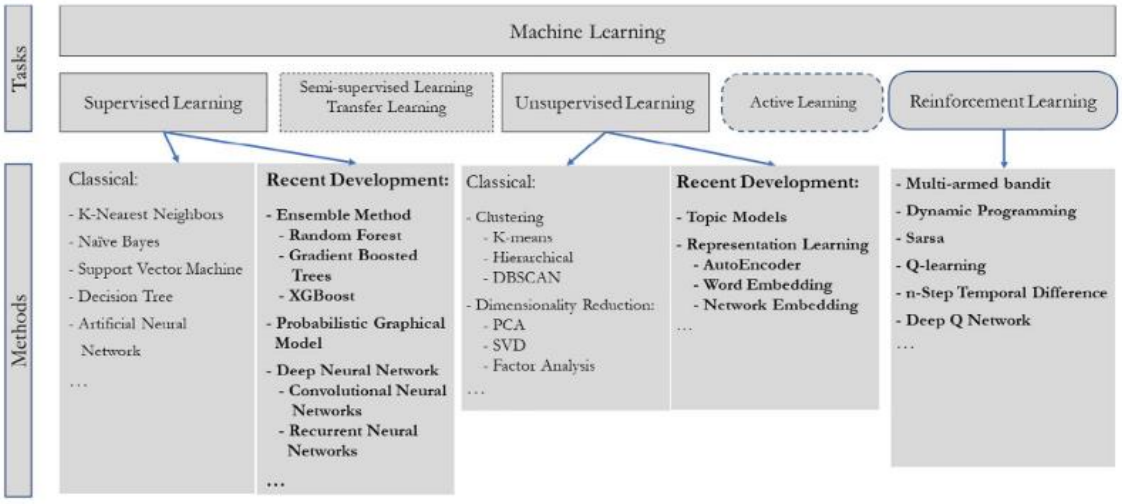


Fuente: Bayoude et al., 2018, p. 374. Elaboración propia.

Entre las fortalezas del ML están su flexibilidad y su capacidad de procesar datos desestructurados o con estructuras complejas, de gran extensión y en diversos formatos (Ma & Sun, 2020).

Finalmente, el ML se clasifica en Aprendizaje supervisado (SML), Aprendizaje Semi-Supervisado o Aprendizaje de Transferencia, Aprendizaje No Supervisado (USML) y Aprendizaje Reforzado (RML) (Figura 1-2).

Figura 1–2: Clasificación de ML.



Fuente: Ma y Sun, 2020, p. 483.

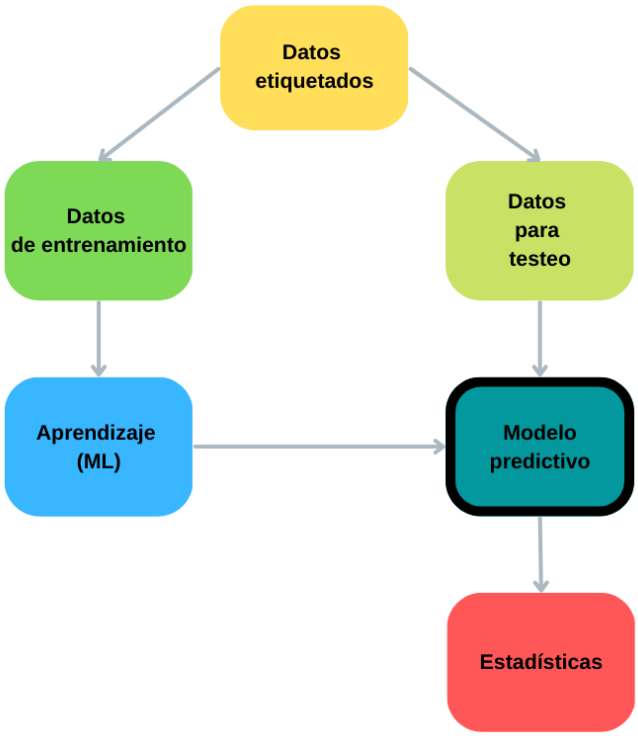
1.5.2 Aprendizaje supervisado (SML)

El aprendizaje supervisado es el tipo de ML en el que se usan algoritmos con “datos que ya han sido etiquetados u organizados previamente para indicar cómo tendría que ser categorizada la nueva información” (Petteri, 2018, p. 21). Debido a la naturaleza de los datos, se requiere de la intervención humana para que exista retroalimentación.

Cabe resaltar que el SML es aplicado en aquellos casos donde el resultado del procesamiento es claro, pero se requiere conocer cómo las relaciones entre los datos repercute en dicho resultado (Chinnamgari, 2019). Así las cosas, el procesamiento de los datos parte, primero, del entrenamiento con los datos etiquetados, que luego se transforma en un modelo predictivo producto del entrenamiento. Sin embargo, una porción de los

datos se conserva para testear el modelo, obtener retroalimentación y determinar su desempeño (Figura 1-3).

Figura 1–3: Procesamiento de datos con SML.



Fuente: Chinnamgari, 2019, p. 14. Elaboración propia.

El SML, a su vez, se categoriza en clasificación y regresión: la primera, ayuda en la predicción de etiquetas de tipo nominal, mientras que la segunda hace posible la predictibilidad de los valores numéricos (Bayoude et al., 2018a; Chinnamgari, 2019; Ma & Sun, 2020).

Para concluir, es importante resaltar que los algoritmos más comunes del SML son los árboles de clasificación y regresión (CART), regresión logística, regresión lineal,

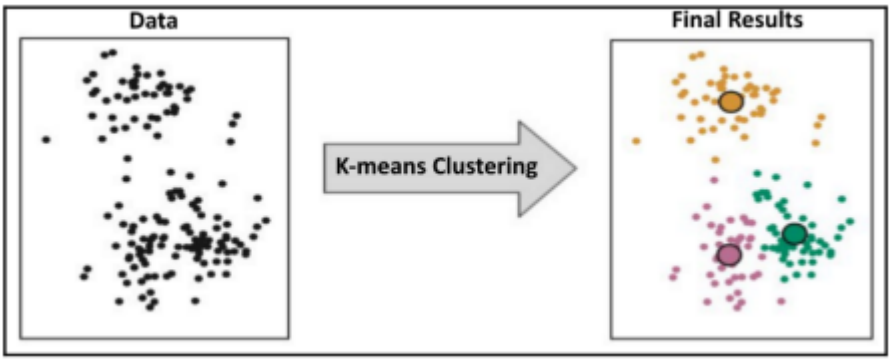
clasificador bayesiano o Naive Bayes, redes neuronales, K-nearest neighbors (KNN) y máquinas de vectores de soporte (SVM) (Bayoude et al., 2018a; Chinnamgari, 2019).

1.5.3 Aprendizaje no supervisado (USML)

El aprendizaje no supervisado es un tipo de ML en el que “los algoritmos no usan ningún dato etiquetado u organizado previamente para indicar cómo tendría que ser categorizada la nueva información, sino que tienen que encontrar la manera de clasificarlas ellos mismos” (Petteri, 2018, p. 21). En consecuencia, el USML no requiere la supervisión humana para el etiquetado de los datos.

En el USML, la información desestructurada se constituye como el input del algoritmo, el cual indaga por patrones en los datos y por los clústers que agrupen atributos con características similares (Figura 1-4). No obstante, sí se hace necesaria una verificación por parte del investigador para determinar si la agrupación es acertada (Chinnamgari, 2019; De Mauro et al., 2022).

Figura 1–4: Clusterización por K-medios a través de USML.



Fuente: Chinnamgari, 2019, p. 16.

Dentro de los algoritmos más comunes de USML se encuentran los de clusterización como K-medias, K-modos y jerárquico; el algoritmo de minería por regla de asociación como la transformación de equivalencia de clase (Eclat) y el patrón de frecuencia de crecimiento

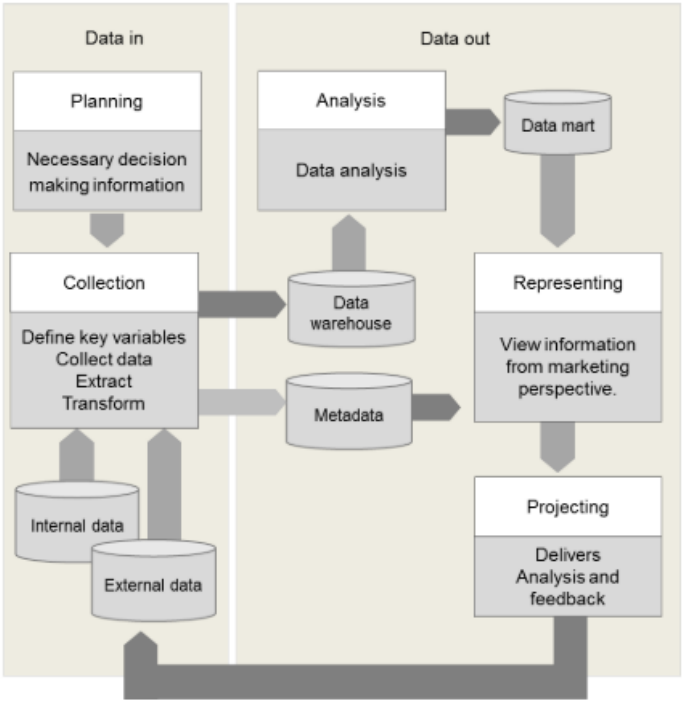
(FPG); y la detección de anomalías (Bayoude et al., 2018a; Chinnamgari, 2019; De Mauro et al., 2022; Ma & Sun, 2020).

1.5.4 Toma de decisiones aplicado al marketing con ML

El ML se relaciona directamente con el marketing en el proceso de toma de decisiones de una campaña de publicidad en una agencia. Dicho proceso se sustenta en el input/output de datos y en los métodos de recolección, análisis y representación de la información (Guarda et al., 2012; Olivar Urbina, 2021).

Con base en lo anterior, la toma de decisiones aplicadas al marketing con ML constan de cinco pasos, según lo referenciado por Guarda et al. (2012), como se muestra a continuación (Figura 1-5):

Figura 1–5: Proceso de toma de decisiones aplicadas al marketing con ML.



Fuente: Guarda et al., 2012, p. 458

En la planeación, se identifica cuál es la información requerida para tomar las decisiones y se sustenta, por tanto, en un input de datos. La recolección de los datos es la etapa en la que se definen las variables de estudio y, con base en ello, se recolecta la información, se extrae y procesa en aras de su transformación partiendo de la información interna y externa que tiene la agencia, como estudios del mercado objetivo. El análisis de los datos comprende los procedimientos de detección de patrones, organización y codificación según lo planteado en la etapa anterior. En la representación, se aplican los modelos de datos pertinentes. Por último, en la proyección, se muestran los resultados para su respectiva evaluación y retroalimentación, que cierra el ciclo apoyado en el output de datos (Guarda et al., 2012).

2. Datos

2.1 Descripción de variables

La base de datos empleada para la proposición del modelo de segmentación sociodemográfica para los consumidores de un portafolio de productos y servicios enfocado en marketing digital de esta tesis es un documento público titulado "Casos positivos de COVID-19 en Colombia"[4] del INS, publicada en el portal Datos Abiertos, con fecha de actualización del 4 de enero de 2023. Esta base cuenta con 6'345.115 registros. Las variables y su tipología consideradas por el INS en la base de datos se muestran en la Tabla 2-1:

Tabla 2-1: Variables de la base de datos general.

#	Variable	Tipo	Categorías
1	Fecha reporte web	Fecha	Sin categorías
2	ID de caso	Numérica	Sin categorías
3	Fecha de notificación	Fecha	Sin categorías
4	Código DIVIPOLA departamento	Numérica	Sin categorías
5	Nombre departamento	Categórica	Sin categorías
6	Código DIVIPOLA municipio	Categórica	Sin categorías
7	Nombre municipio	Categórica	Sin categorías

[4] Disponible en https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr

Tabla 2-1: (Continuación)

#	Variable	Tipo	Categorías
8	Edad	Numérica	Sin categorías
9	Unidad de medida de edad	Categórica	Años
			Meses
			Días
10	Sexo	Categórica	Femenino
			Masculino
11	Tipo de contagio	Categórica	Relacionado
			Importado
			En estudio
			Comunitario
12	Ubicación del caso	Categórica	Casa
			Hospital
			Hospital UCI
			Fallecido
			No aplica
13	Estado	Categórica	Leve
			Moderado
			Grave
			Fallecido
			No aplica
14	Código ISO del país	Categórica	Sin categorías
15	Nombre del país	Categórica	Sin categorías
16	Recuperado	Categórica	Recuperado
			Fallecido
			No aplica
			(Vacío)
17	Fecha de inicio de síntomas	Fecha	Sin categorías
18	Fecha de muerte	Fecha	Sin categorías
19	Fecha de diagnóstico	Fecha	Sin categorías
20	Fecha de recuperación	Fecha	Sin categorías
21	Tipo de recuperación	Categórica	PCR
			Tiempo clínico
22	Pertenencia étnica	Numérica	Indígena
			ROM
			Raizal
			Palenquero
			Negro
			Otro
23	Nombre del grupo étnico	Categórica	Sin categorías

Fuente: Instituto Nacional de Salud, 2020. Elaboración propia.

2.2 Metodología

Con el fin de proponer el modelo de segmentación demográfica objetivo de esta tesis, que aplica el USML, se realizó un preprocesamiento de los datos de la base descrita en el apartado anterior, que implicó la elección de diez variables consideradas relevantes para la campaña de marketing digital según los criterios especificados en la Tabla 2-2.

Tabla 2-2: Nombre, tipo y descripción de las variables seleccionadas para el preprocesamiento de los datos.

#	Variable	Tipo y descripción
1	ID de caso	Nominal. Enuncia el número de identificación del sujeto que reportó tener contagio por COVID-19. Esta variable confirma que los sujetos son reales.
2	Nombre departamento	Nominal. Enuncia el nombre del departamento donde el sujeto reportó el contagio por COVID-19.
3	Edad	Nominal. Enuncia la edad del sujeto que reportó el contagio por COVID-19.
4	Sexo	Nominal. Enuncia el sexo del sujeto que reportó el contagio por COVID-19.
5	Tipo de contagio	Nominal. Enuncia el entorno en el que el sujeto fue contagiado por COVID-19
6	Ubicación del caso	Nominal. Enuncia el lugar donde el sujeto contagiado por COVID-19 llevó a cabo su proceso de recuperación o si falleció.
7	Estado	Nominal. Enuncia el nivel de afectación en la salud del sujeto contagiado por COVID-19 y la complejidad de los síntomas.
8	Recuperado	Nominal. Enuncia si el sujeto contagiado por COVID-19 se recuperó o si falleció.
9	Tipo de recuperación	Nominal. Refiere si el sujeto contagiado por COVID-19 comprobó su recuperación mediante la toma de prueba PCR con resultado negativo o si en el tiempo clínico (21 días) notó mejoría o desaparición de sintomatología relacionada con el virus.
10	Pertenencia étnica	Nominal. Enuncia la pertenencia étnica del sujeto que reportó el contagio por COVID-19, basada en: i) el registro de esta categoría en las EPS/IPS del país; ii) el autorreconocimiento; iii) el censo departamental

Fuente: Información del Instituto Nacional de Salud (2020). Elaboración propia.

En cuanto al procedimiento, el análisis descriptivo de las variables se llevó a cabo a través de R y para la clusterización, se empleó Phyton. Sin embargo, como se pudo demostrar en la Tabla 2-2, todas las variables seleccionadas son nominales, por lo que debieron ser codificadas bajo la técnica One-hot para realizar la clusterización en Phyton y que el modelo de segmentación demográfica aplicado a ML, planteado en esta tesis, pueda ejecutarse.

En consecuencia, la codificación por One-hot de las variables seleccionadas se transforman en variables “dummies”, cuya asignación es “1” o “0”, dependiendo de su pertenencia a la categoría. Se aplica para el caso, entonces, K-1 variables de carácter binario, siendo K la totalidad de las categorías. Esto quiere decir que, después de la decodificación por One-hot, se pasó de tener 10 variables a 65, que, siendo redundantes, contienen la misma data.

La recodificación de las variables, como en este escenario, es un paso necesario para hacer uso de los métodos de clasificación tradicionales. Los códigos utilizados están incluidos en los Anexos 5.1 y 5.2.

3. Procedimiento

3.1 Análisis estadístico de las variables

En concordancia con el capítulo anterior, se eligieron y preprocesaron diez variables para realizar los análisis correspondientes. A continuación, se muestra en la figura 3-1 un resumen general de las variables procesadas en R.

Figura 3–1: Dataframe de las variables seleccionadas para esta tesis. Elaboración propia.

```
'data.frame':   6345115 obs. of  10 variables:
 $ ID.de.caso          : chr  "2,265,685" "2,265,686" "2,265,687" "2,265,688" ...
 $ Nombre.departamento : Factor w/ 36 levels "amazonas","antioquia",..: 6 6 6 6 6 6 24 24 24 24 ...
 $ Edad                : int  49 49 51 51 51 52 24 24 33 35 ...
 $ Sexo                : Factor w/ 2 levels "f","m": 2 2 1 1 1 1 2 2 2 1 ...
 $ Tipo.de.contagio    : Factor w/ 3 levels "comunitaria",..: 1 3 1 3 1 3 3 1 3 1 ...
 $ Ubicación.del.caso  : Factor w/ 5 levels "casa","fallecido",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Estado              : Factor w/ 5 levels "fallecido","grave",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Recuperado          : Factor w/ 4 levels "activo","fallecido",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ Tipo.de.recuperación: Factor w/ 3 levels "","pcr","tiempo": 3 3 3 3 3 3 3 3 3 3 ...
 $ Etnia               : Factor w/ 5 levels "indigena","negro",..: 3 3 3 3 3 3 1 1 1 3 ...
```

Como muestra la figura 3-2, de las variables seleccionadas, ocho son variables tipo factor, una es numérica y la variable restante es tipo carácter (indica el ID del caso), que fue descartada de este análisis por no contener información relevante para la modelación. En esta ilustración se evidencia que existe mayor número de casos positivos por Covid-19 en Colombia en mujeres, con una media de edad de 40 años, con un contagio predominantemente comunitario y ubicados en su mayoría en las ciudades de Bogotá, Medellín y Cali.

Figura 3–2: Dataframe detallado de las variables seleccionadas. Elaboración propia.

```
 ID.de.caso          Nombre.departamento       Edad        Sexo          Tipo.de.contagio
Length:6345115     bogota      :1868554   Min.   :  1.00   f:3390429   comunitaria:4403144
Class :character   antioquia   : 948916   1st Qu.: 27.00   m:2954686   importado  :   3699
Mode  :character   valle       : 567982   Median : 38.00               relacionado:1938272
                   cundinamarca: 329530   Mean   : 39.95
                   santander   : 295491   3rd Qu.: 53.00
                   barranquilla: 276896   Max.   :114.00
                   (Other)     :2057746
   Ubicación.del.caso        Estado            Recuperado      Tipo.de.recuperación      Etnia
casa        :6165940   fallecido: 142179   activo    :   7988         : 180822       indigena:  83706
fallecido   : 142179   grave    :    202   fallecido : 142179   pcr    : 933420       negro    : 137919
hospital    :   1410   leve     :6165940   n/a       :  30655   tiempo:5230873       otro     :6122947
hospital uci:    202   moderado :   1410   recuperado:6164293                        raizal   :    405
n/a         :  35384   n/a      :  35384                                             rom      :    138
```

Figura 3–3: Cajas de bigotes. Gráfico A: Grupos etarios de los registros de la base de datos. Gráfico B: Relación entre edad y sexo de los registros de la base de datos. Gráfico C: Relación entre edad y etnia de los registros de la base de datos. Elaboración propia

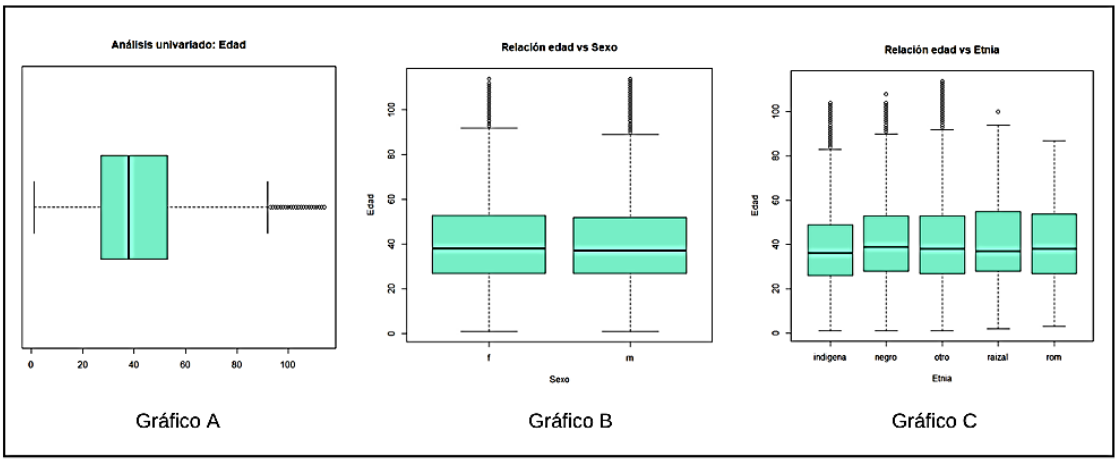


Gráfico A Gráfico B Gráfico C

La figura 3-3 representa tres cajas de bigotes que indican la distribución etaria de los contagiados por Covid-19 en Colombia según la etnia y el sexo. El gráfico A expone que el rango de edad de mayor contagio está entre 39 y 53 años, con un patrón similar entre hombres y mujeres (Gráfico B) y mayor prevalencia en grupos poblacionales diferentes a las comunidades étnicas identificadas (Gráfico C). Adicionalmente, se evidencia una media menor respecto a la edad en las comunidades indígena y raizal.

A continuación, se presentan algunas distribuciones porcentuales de los datos:

Figura 3–4: Relación porcentual de personas contagiadas por Covid-19 en Colombia según sexo. Elaboración propia.

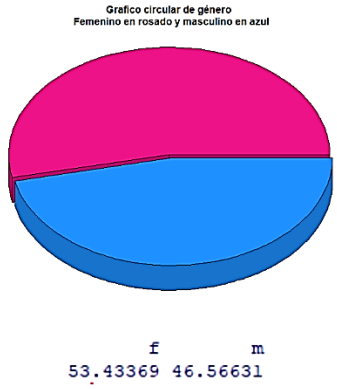


Figura 3–5: Relación porcentual de personas contagiadas por Covid-19 en Colombia según sexo y tipo de contagio. Elaboración propia.

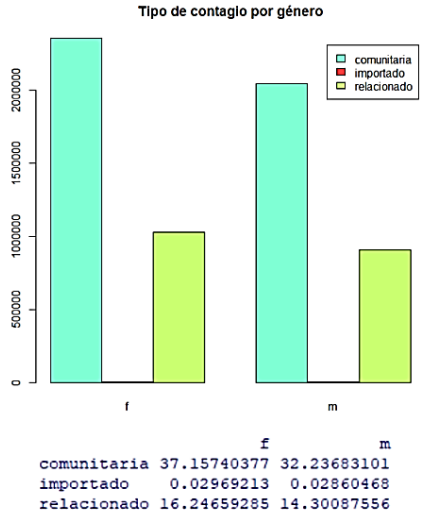
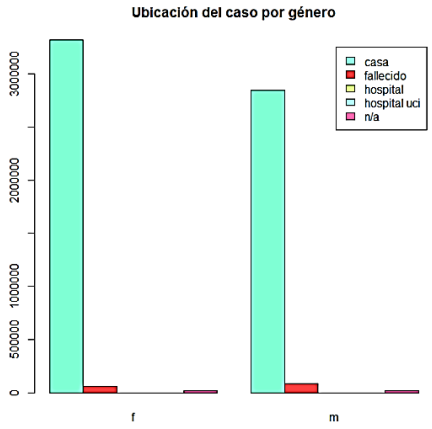
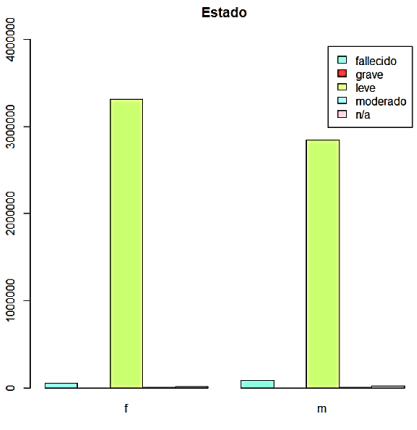


Figura 3–6: Relación porcentual de personas contagiadas por Covid-19 en Colombia según sexo y la ubicación del caso. Elaboración propia.



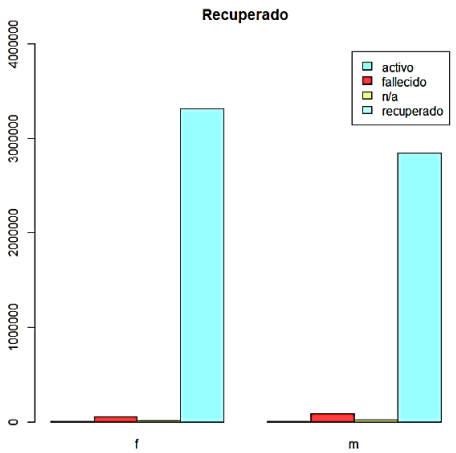
	f	m
casa	3317806	2848134
fallecido	55953	86226
hospital	694	716
hospital uci	96	106
n/a	15880	19504

Figura 3–7: Relación porcentual de personas contagiadas por Covid-19 en Colombia según sexo y estado de salud durante el contagio. Elaboración propia.



	f	m
fallecido	0.881827989	1.358935181
grave	0.001512975	0.001670576
leve	52.289138968	44.887035144
moderado	0.010937548	0.011284271
n/a	0.250271272	0.307386076

Figura 3–8: Relación porcentual de personas contagiadas por Covid-19 en Colombia según sexo y recuperación. Elaboración propia.



	f	m
activo	0.07052670	0.05536543
fallecido	0.88182799	1.35893518
n/a	0.21698582	0.26614175
recuperado	52.26434824	44.88586889

En proporción, es mayor el contagio de Covid-19 en el género femenino, con un 53,43% (Figura 3-4) De este porcentaje, 37,15% reportó contagio comunitario y el 16,24% lo hizo por contagio en relación con algún conocido o familiar. De estas afectaciones, se remitieron a casa el 52,28% de los casos, 52,26% tuvo una recuperación exitosa y 0.88% presentaron mortalidad (Figuras 3-6 y 3-8).

Por otro lado, el género masculino representa el 46,56% del total de contagiados de los cuales 32,23% reportó contagio comunitario y el 14,30% de contagiados por relación con algún conocido o familiar. De estos cuadros clínicos, se remitieron a casa el 48,88% de los casos, 44,88% tuvo una recuperación exitosa y el 1,35% de los contagiados masculinos falleció (Figuras 3-5, 3-6, 3-7 y 3-8).

Sobre el estado de salud de las personas contagiadas, se reportó que el 52,28% de las mujeres y el 44,88% de los hombres presentaron síntomas leves por el virus, datos que parecen concordantes con los índices de remisión para recuperación en casa y de recuperación exitosa (Figuras 3-5, 3-6, 3-7 y 3-8).

En conclusión, el género masculino presentó mayor índice de mortalidad y menor porcentaje de recuperación exitosa, aunque el contagio por Covid-19 en mujeres fue mayor.

A continuación, se muestran algunos gráficos descriptivos univariados y bivariados de la data.

Figura 3–9: Gráficos descriptivos univariados. Gráfico A: Edad. Gráfico B: Etnia. Gráfico C: Tipo de recuperación. Gráfico D: Ubicación de caso. Gráfico E: Departamentos. Gráfico F: Recuperados. Gráfico G: Estado. Gráfico H: Tipo de contagio. Elaboración propia.

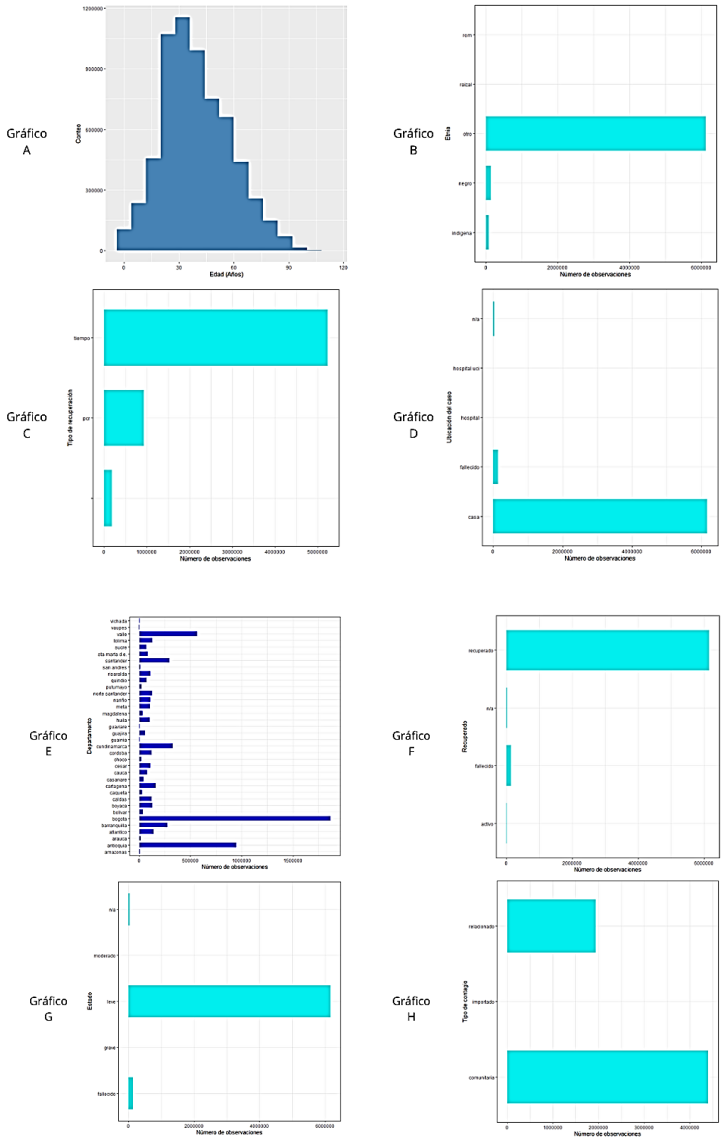
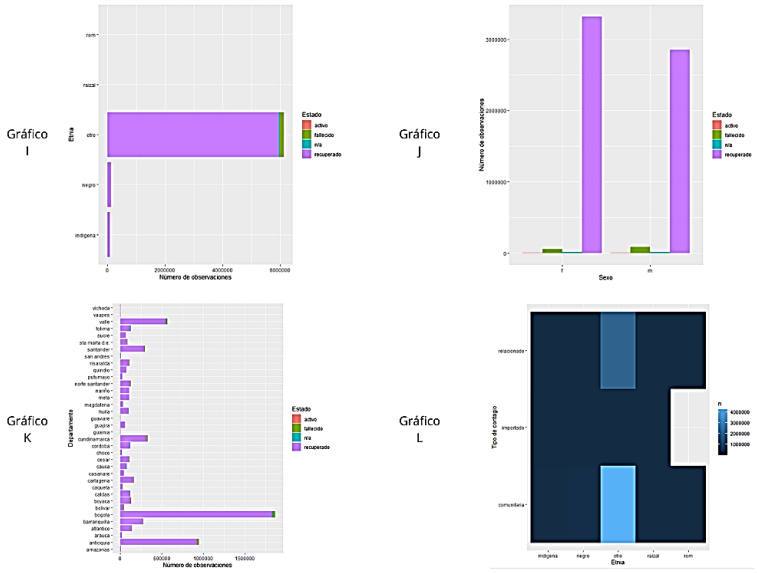


Figura 3–10: Gráficos descriptivos bivariados. Gráfico I: Relación Estado-Etnia. Gráfico J: Relación Estado-Sexo. Gráfico K: Relación Estado-Departamento. Gráfico L: Relación Etnia-Tipo de contagio. Elaboración propia.



Las figuras 3-9 y 3-10 muestran cada una de las variables seleccionadas de forma detallada. Este es un insumo fundamental para el proceso de clusterización, ya que identifica una tendencia de personas que reportaron contagio por Covid-19 con un rango de edad entre los 25 y 57 años, ubicados en los departamentos de Bogotá D.C.[5], Antioquia y Valle del Cauca y cuyos contagios se dieron en contextos comunitarios, como espacio público, aglomeraciones, transporte público, entre otros.

Estos gráficos también evidencian que en la mayoría de los casos, las personas contagiadas por Covid-19 fueron remitidas a su casa debido a la sintomatología leve del virus, lo que conllevó a una recuperación exitosa.

En relación con las personas fallecidas, el género masculino reportó un mayor número de fallecidos con relación al número de infectados. Por otro lado, el género femenino reportó un número menor de casos activos y una proporción mayor de recuperadas.

[5] Bogotá, debido a su estatus de Distrito Capital, es considerado en la base de datos como un departamento.

3.2 Modelación

En concordancia con el apartado 2.2. de esta tesis, se preprocesaron las 10 variables seleccionadas. Para la transformación de variables nominales a variables dummies, se utilizó la librería Pandas *get_dummies*. Además, con el fin de mejorar la normalidad de los datos numéricos, se aplicó la función *PowerTransformer* de la librería *sklearn*.

Figura 3–11: Muestra de los registros del dataset original. Elaboración propia.

index	Departamento	Edad	Sexo	Tipocontagio	Ubicacion	Estado	Recuperado	Tiporecuperacion	Etnia
0	BOGOTA	49	M	COMUNITARIA	CASA	LEVE	RECUPERADO	TIEMPO	OTRO
1	BOGOTA	49	M	RELACIONADO	CASA	LEVE	RECUPERADO	TIEMPO	OTRO
2	BOGOTA	51	F	COMUNITARIA	CASA	LEVE	RECUPERADO	TIEMPO	OTRO

Figura 3–12: Muestra de la recodificación del dataset por One-hot. Elaboración propia.

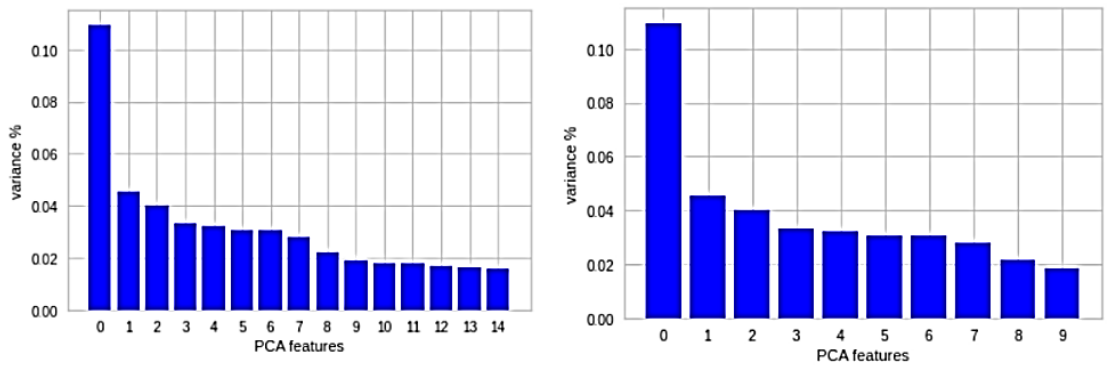
index	Edad	Departamento_AMAZONAS	Departamento_ANTIOQUIA	Departamento_ARAUCA	Departamento_ATLANTICO	Departamento_BARRANQUILLA	Departamento_BOGOTA
0	0.5419308591324175	0	0	0	0	0	1
1	0.5419308591324175	0	0	0	0	0	1
2	0.641267152796299	0	0	0	0	0	1
3	0.641267152796299	0	0	0	0	0	1
4	0.641267152796299	0	0	0	0	0	1
5	0.6904836026095628	0	0	0	0	0	1
6	-0.8367437126636557	0	0	0	0	0	0
7	-0.8367437126636557	0	0	0	0	0	0
8	-0.30491084076222197	0	0	0	0	0	0
9	-0.1930862986140327	0	0	0	0	0	0
10	-0.13790941258230902	0	0	0	0	0	0
11	-0.30491084076222197	0	0	0	0	0	0
12	-0.4766181583324581	0	0	0	0	0	0
13	-0.028940652132666323	0	0	0	0	0	0
14	-0.08319956122536153	0	0	0	0	0	0
15	-0.899290777307675	0	0	0	0	0	0
16	-1.1582323442676716	0	0	0	0	0	0
17	-0.653641042060886	0	0	0	0	0	0
18	-0.36159628753077283	0	0	0	0	0	0
19	-0.41882458804380224	0	0	0	0	0	0
20	1.487602183266923	0	0	0	0	0	0
21	0.7880500466972029	0	0	0	0	0	0
22	-0.24874735122854003	0	1	0	0	0	0
23	-1.0920783656582456	0	1	0	0	0	0
24	0.3905466109053029 6	0	1	0	0	0	0

Para ilustrar a grandes rasgos el proceso, las figuras 3-11 y 3-12 muestran, respectivamente, 3 registros del dataset original y la transformación de los datos mediante la técnica One-hot, previo a la clusterización. En consecuencia, se recodificaron 8 variables categóricas nominales y se mejoró la normalidad de la variable numérica *Edad*. La variable

“ID de caso” se descartó por no presentar información relevante para el modelo de segmentación.

Anterior a la agrupación con K-Means, se realizó un análisis de componentes principales con el objetivo de condensar un nuevo marco de datos con sus características más relevantes. Se limitó el número máximo de componentes a 15, con el objetivo de determinar si la varianza total se puede explicar con 15 o menos componentes.

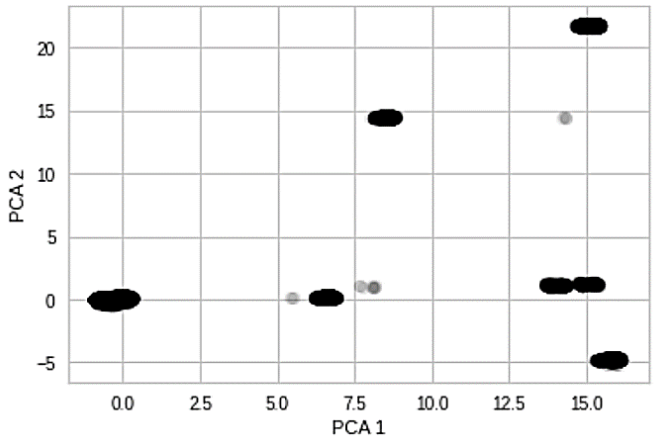
Figura 3–13: Gráficos PCA. Elaboración propia.



El gráfico PCA (Figura 3-13) muestra que la contribución de los componentes a la varianza general se desvanece a menos del 2% cada uno después de 8 características. Entonces, hasta 8 componentes parecen reflejar adecuadamente los patrones del marco de datos de origen con sus 65 columnas, muchas de las cuales son sólo variables ficticias binarias.

Con el objetivo de discernir de grupos distintivos, se graficaron los dos primeros componentes:

Figura 3–14: Gráfico PCA. Elaboración propia.



El gráfico PCA (Figura 3-14) mostró entre 5 y 6 grupos significativos, de los cuales 4 se encuentran flotando en la parte inferior y 2 en la parte superior derecha. Esto quiere decir que, envés de hacer uso de la base de datos inicial, se utilizó el marco de datos reducido con sus 8 componentes, que son fiel reflejo del dataset inicial.

Por consiguiente, se procedió a aplicar el método de clasificación K-Means para números alternativos de conglomerados. Se seleccionaron entre dos y diez, por lo que se recopilaron las puntuaciones de inercia resultantes.

Figura 3–15: Aplicación de método de clasificación K-means. Elaboración propia.

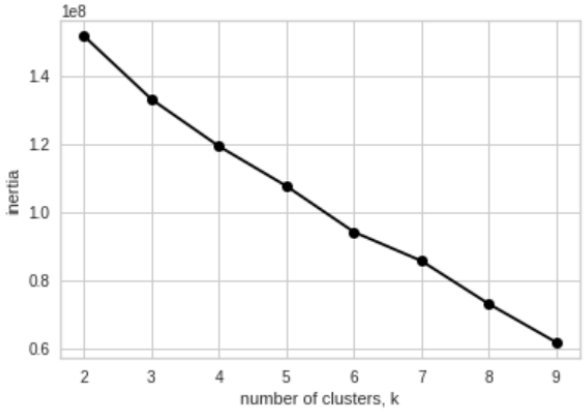


Figura 3–16: Aplicación del método Kneelocator. Elaboración propia.

```
from kneed import KneeLocator
inertia_knee_b3 = KneeLocator(
        range(2,10),
        inertias_pca,
        S=0.1, curve="convex", direction="decreasing")

K_inertia_b3 = inertia_knee_b3.elbow
print("elbow at k =", f'{K_inertia_b3:.0f} clusters')
```

```
elbow at k = 5 clusters
```

En la figura 3-15, es posible evidenciar el punto del codo en K = 5. Por este motivo, se empleó el paquete *Kneed* y su función *Kneelocator* para identificar, numéricamente, el punto del codo. Este último proceso confirmó el resultado del gráfico de codo, por lo tanto, el mejor número de grupos es 5 (Figura 3-16). La clusterización resultante se resume en la Tabla 3-1.

Tabla 3-1: Tabla resumen de la clusterización final. Elaboración propia.

Variable/Clúster	0	1	2	3	4
Departamento	Bogotá, Nariño, Risaralda, N. Santander Antioquia	Cartagena	Antioquia, Caquetá, Valle, Cundinamarca, Santa Marta D. E	Caquetá, Barranquilla, Antioquia, Valle, Cundinamarca	Bogotá, Nariño, Santander, Boyacá, N. Santander
Sexo	Masculino, Femenino	Femenino, Masculino	Femenino, Masculino	Femenino, Masculino	Femenino
Etnia	Otro, Indígena, Negro	Otro, Negro, Indígena	Otro, Negro, Indígena	Otro, Negro, indígena	Otro, indígena, Negro
Tipo de contagio	Comunitaria, Relacionado, Importado	Comunitaria, Relacionado, Importado	Comunitaria, Relacionado, Importado	Comunitaria, Relacionado, Importado	Comunitaria, Importado
Estado	Leve	Leve	Fallecido, Moderado, Grave	No aplica	Leve
Ubicación	Casa	Casa	Fallecido, Hospital, Hospital UCI	No aplica	Casa
Tipo de recuperación	Tiempo, PCR	Tiempo, PCR	No aplica, Tiempo	No aplica, PCR	Tiempo, PCR
Recuperado	Recuperado, Activo	Recuperado, Activo	Fallecido, Activo, Recuperado	No aplica, Recuperado, Activo	Recuperado, Activo

Finalmente, conformados los grupos, se insertó una nueva columna en la base de datos original con las etiquetas de cada clúster (Figura 3-17).

Figura 3–17: Muestra de la base de datos con clusterización aplicada. Elaboración propia.

index	Cluster	Departamento	Edad	Sexo	Tipocontagio	Ubicacion	Estado	Recuperado	Tiporecuperacion	Etnia
6345113	1	META	11	M	COMUNITARIA	CASA	LEVE	RECUPERADO	TIEMPO	OTRO
6345114	3	META	34	F	COMUNITARIA	CASA	LEVE	RECUPERADO	TIEMPO	OTRO

3.3 Implementación

Una vez finalizado el proceso de clusterización que dio como resultado el modelo de segmentación sociodemográfica que sirve como insumo principal para la campaña de marketing digital que promueve la adquisición de tapabocas, se preparó un tablero para las directivas de la agencia de publicidad A y otras personas interesadas en la campaña[6], que expusiera los atributos más importantes de cada uno de los clústeres, como se ilustra en las figuras 3-18 y 3-19.

Figura 3–18: Código utilizado para la creación del tablero, aplicando la función *cluster_profile.* Elaboración propia.

```
[ ]  # helper function: medians and modes for each cluster
     def cluster_profile(df):
         dfc = df.groupby("Cluster").agg({
             "Edad": "mean",
             "Departamento": lambda x: x.value_counts().index[0],
             "Sexo": lambda x: x.value_counts().index[0],
             "Tipocontagio": lambda x: x.value_counts().index[0],
             "Ubicacion": lambda x: x.value_counts().index[0],
             "Estado": lambda x: x.value_counts().index[0],
             "Recuperado": lambda x: x.value_counts().index[0],
             "Tiporecuperacion": lambda x: x.value_counts().index[0],
             "Etnia": lambda x: x.value_counts().index[0]
                      })    #.sort_values(by=["MonthlyCharges"], ascending=False)

         cluster_pies(df)
         return dfc
```

[6] Presentación llevada a cabo el 11 de enero de 2023.

Figura 3–19: Tablero de perfiles de los clústeres. Elaboración propia.

100%|████████| 5/5 [00:04<00:00, 1.05it/s]

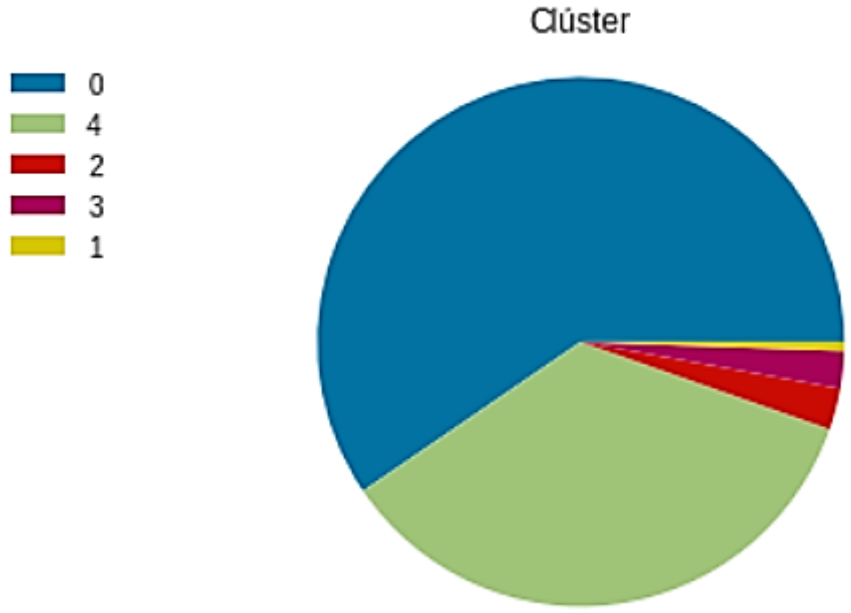
Cluster	0	1	2	3	4
Edad	38.561052	38.607281	68.17805	68.31161	40.121103
Departamento	BOGOTA	CARTAGENA	BOGOTA	BOGOTA	BOGOTA
Sexo	M	F	M	M	F
Tipocontagio	COMUNITARIA	COMUNITARIA	COMUNITARIA	COMUNITARIA	COMUNITARIA
Ubicacion	CASA	CASA	FALLECIDO		CASA
Estado	LEVE	LEVE	FALLECIDO		LEVE
Recuperado	RECUPERADO	RECUPERADO	FALLECIDO		RECUPERADO
Tiporecuperacion	TIEMPO	TIEMPO			TIEMPO
Etnia	OTRO	OTRO	OTRO	OTRO	OTRO

En términos de visualización general, la Tabla 3-2 reporta el valor más frecuente de las variables categóricas para cada clúster y la media de las variables numéricas (Edad).

Tabla 3-2: Resumen de la clusterización en número de individuos, edad media y porcentaje. Elaboración propia.

Clúster	**Individuos**	**Edad media**	**Porcentaje**
0	3.776.886	38.56	59.521%
1	160.168	38.61	2.524%
2	143.791	68.18	2.266%
3	35.384	68.31	0.557%
4	2.228.886	40.12	35.132%

Figura 3–20: Proporción de los clústeres. Elaboración propia.



El grupo 0 y el grupo 4 fueron los más representativos con 59.52% y 35,13%, respectivamente (Figura 3-20), de las personas contagiadas, con una edad media entre los 38 y 41 años.

Sin embargo, cuando se da una mirada detallada de cada uno de los grupos (Figura 3-19), se puede determinar además que:

1. El grupo 0, de mayor proporción, abarca mayoritariamente la población masculina mestiza, con una edad media de 38 años, habitante de tres departamentos: Bogotá, Nariño y Risaralda. Adicionalmente, reporta una prelación de casos de contagio de tipos comunitario y relacionado, con sintomatología leve y con una recuperación exitosa en casa después de 21 días[7].
2. El grupo 1 abarca un pequeño número de personas que, en su mayoría, son mujeres con una edad media de 38 años que se identifican como mestizas o negras, residen en Cartagena y que reportaron contagio comunitario con sintomatología leve, con una recuperación exitosa en casa después de 21 días.

[7] Tiempo de recuperación de virus Covid-19.

3. El grupo 2 también se corresponde con un número pequeño de personas que, en mayor proporción, responden al perfil de mujeres mestizas que fallecieron producto del contagio comunitario de Covid-19, con una edad media de 68 años, residían en Antioquia, Caquetá y Valle del Cauca.
4. El grupo 3 está compuesto mayoritariamente por mujeres habitantes de Caquetá, Barranquilla y Antioquia, con una edad media de 68 años, que reportaron contagio comunitario, y cuyo estado o sintomatología asociada al virus, así como su proceso de recuperación son desconocidas, pues están en la categoría “No Aplica”, que es de carácter transitorio.
5. El grupo 4 es el segundo más representativo, abarcando en mayor medida mujeres mestizas con una edad media de 40 años, residentes de Bogotá, Nariño y Santander, que reportaron contagio comunitario con sintomatología leve, con una recuperación exitosa en casa después de 21 días.

Con el propósito de privilegiar la presentación de la información en un formato visual, a continuación, se presenta el conjunto de gráficos circulares (Figuras 3-21, 3-22, 3-23, 3-24 y 3-25) que ilustran los 8 atributos característicos de cada uno de los grupos arrojados por el modelo de segmentación demográfica, objeto de esta tesis, y que fueron presentados a las directivas de la Agencia de publicidad A.

Figura 3–21: Clúster 0.

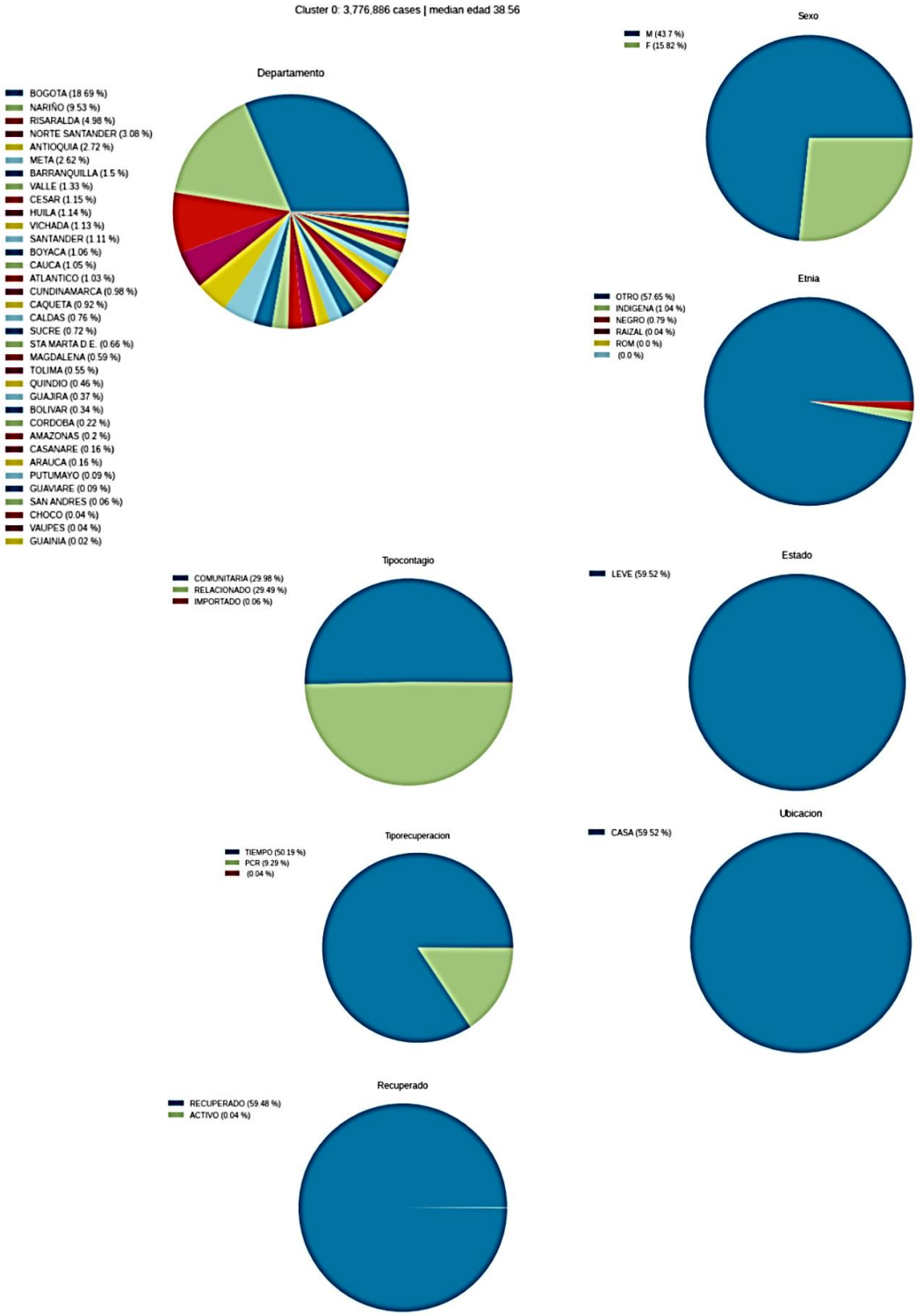


Figura 3–22: Clúster 1.

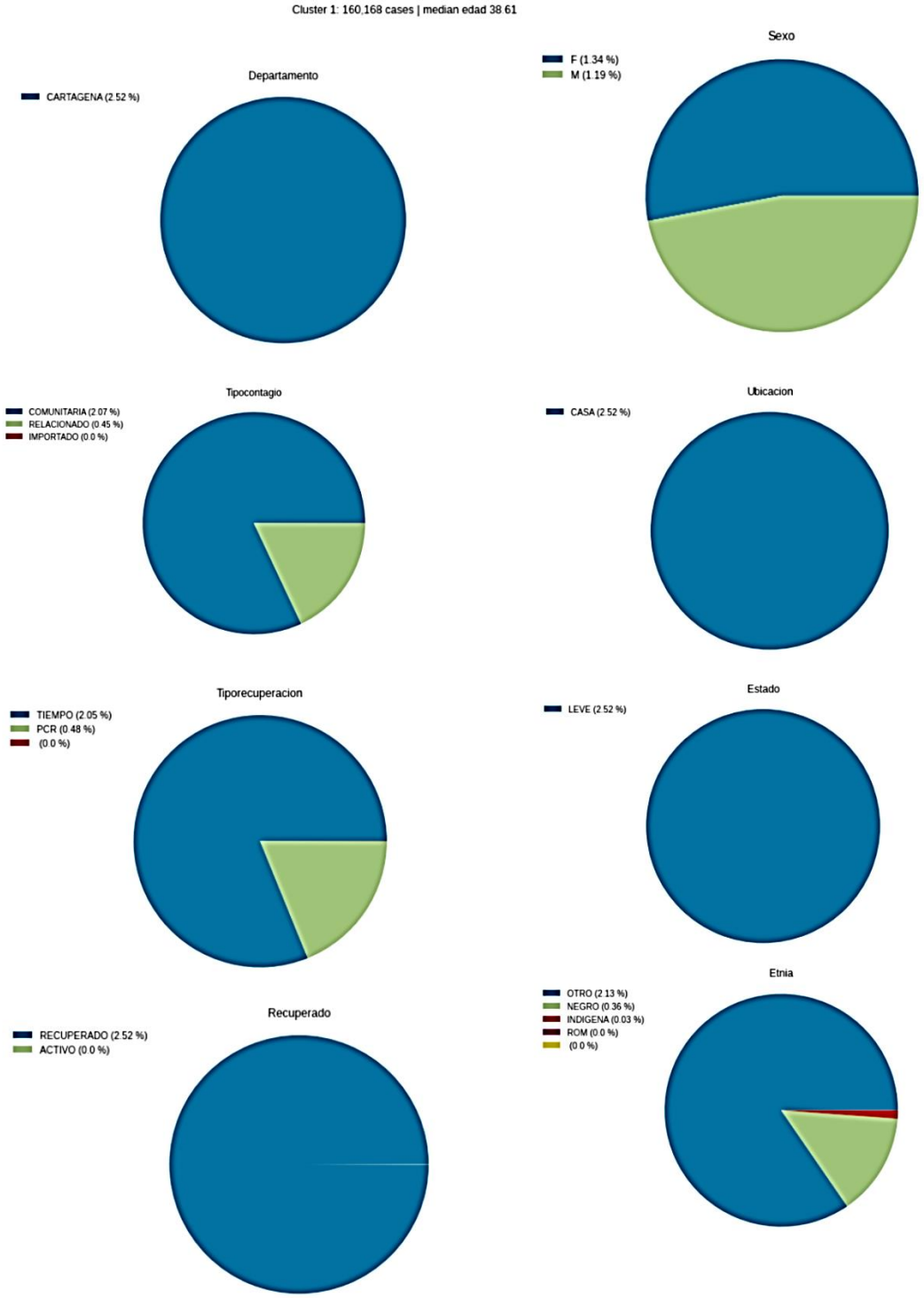


Figura 3–23: Clúster 2.

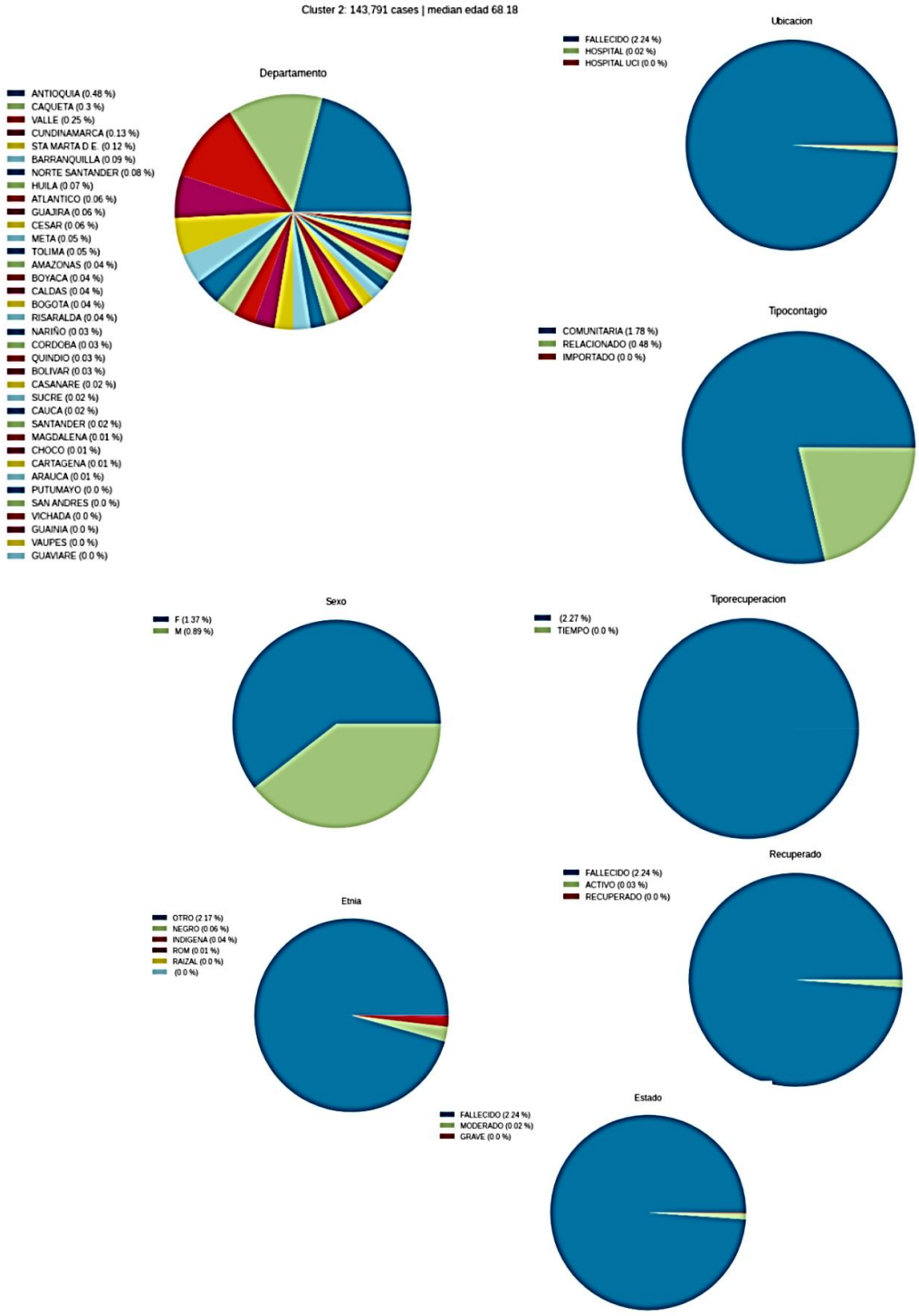


Figura 3–24: Clúster 3

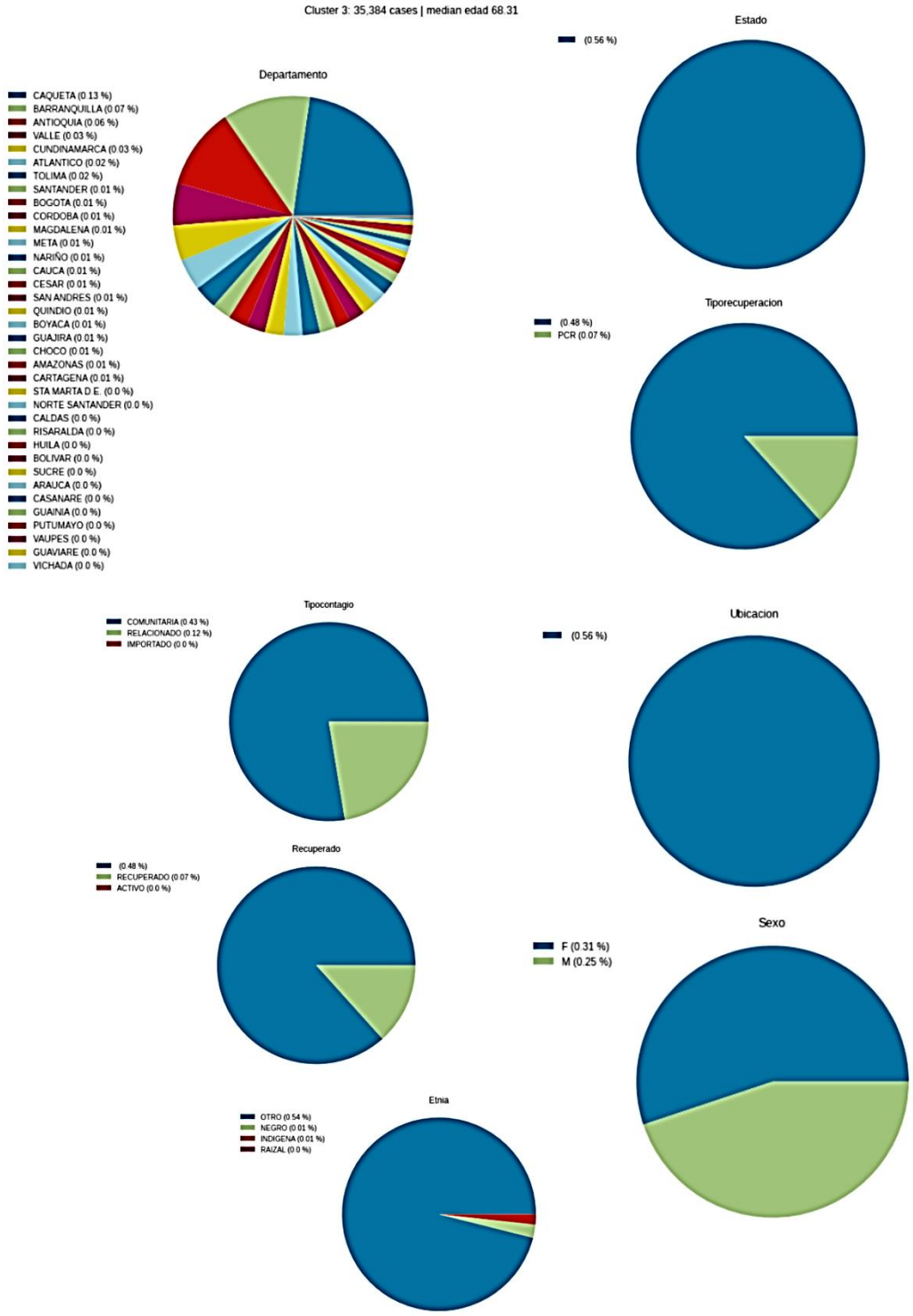
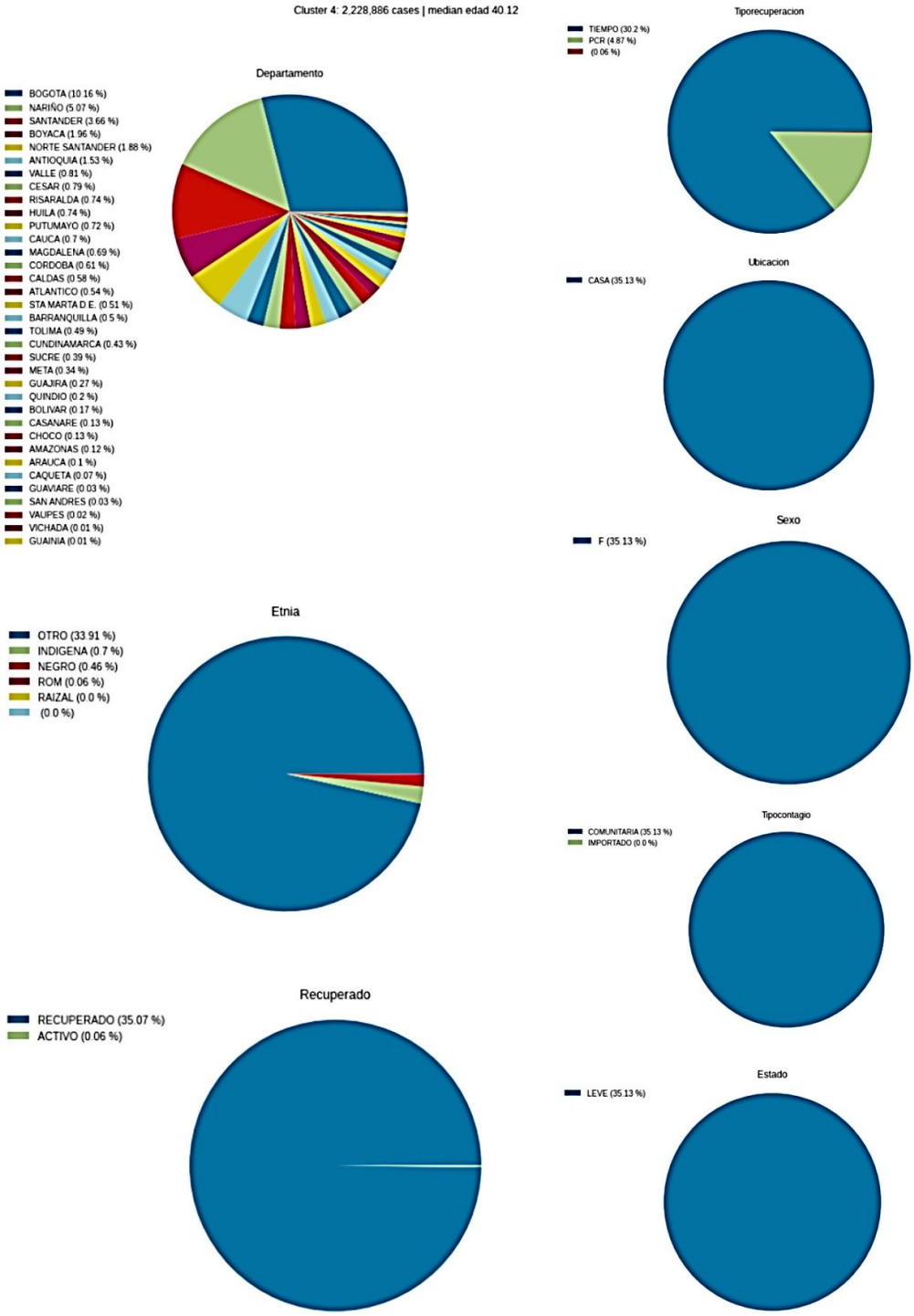


Figura 3–25: Clúster 4.



Ahora bien, sobre la implementación del modelo de segmentación demográfica con aplicación de USML, es imperativo mencionar que fue parte central de las etapas de planificación y recolección de datos del proceso de toma de decisiones aplicadas al marketing y el ML, a las que hacen alusión Guarda et al. (2012). Contar con la base de datos oficial del gobierno colombiano sobre los reportes de contagio a nivel nacional permitió tener un amplio conocimiento sobre la dimensión de la pandemia en el país, así como posibilitó aislar las variables consideradas necesarias para el cliente (productor y comercializador de tapabocas) y la Agencia de publicidad A en la campaña de marketing digital en redes sociales, principalmente en Facebook e Instagram.

Finalmente, en la tercera etapa del proceso de toma de decisiones, referente al análisis de los datos, para incentivar el uso del tapabocas en poblaciones que en este dataset se presentan como compradores potenciales, la Agencia de Publicidad A determinó:

1. Los clústeres 0 y 4 comparten una serie de atributos atractivos para la segmentación de la campaña de marketing digital en redes sociales. El género, en este sentido, se constituye como una estructura de oportunidad para que se focalice la publicidad que los potenciales consumidores verán. Los departamentos en los cuales se segmentan también resultan importantes para la campaña, debido a que el cliente productor y comercializador de tapabocas presenta una actividad comercial predominante en esos lugares.
2. El clúster 1 no se incluye en la campaña, ya que los últimos reportes de contagio por Covid-19 en Cartagena son cercanos a cero[8] y las medidas preventivas allí son prácticamente inexistentes. Por lo tanto, este contexto no es propicio para promocionar la compra y uso del tapabocas desechable.
3. El clúster 2 queda descartado de la inclusión en la campaña de marketing digital, debido a que su población mayoritaria corresponde a personas que se reportaron como fallecidas.
4. El clúster 3 también queda descartado de la inclusión en la campaña, con el fin de minimizar riesgos y aumentar la eficiencia y eficacia de esta, dado que el estado o sintomatología asociada al virus, así como el proceso de recuperación de la población mayoritaria del grupo son desconocidas.

[8] Datos tomados de Google Statistics: https://www.google.com/search?q=covid+en+cartagena&oq=covid+en+cartagena&aqs=chrome.0.69i59j0i512l2j0i22i30l4j69i61.1637j0j7&sourceid=chrome&ie=UTF-8

3.4 Resultados

Como resultado, se aplicó la cuarta etapa del proceso de toma de decisiones o representación (Guarda et al., 2012), con base en el análisis del modelo de segmentación propuesto en esta tesis. Así las cosas, se estableció la creación de dos audiencias en Facebook e Instagram, con el fin de crear *Brand Awareness*[9] e incentivar la compra de tapabocas desechables a través de las redes sociales, como se muestra en la Tabla 3-3.

Tabla 3-3: Audiencias creadas para la campaña de marketing digital, basada en la clusterización planteada en esta tesis.

Características / Audiencia	Audiencia 1 (Clúster 0)	Audiencia 2 (Clúster 4)
Sexo-género	Hombres	Mujeres
Edad	33-43 años	35-45 años
Departamento	Bogotá, Nariño, Risaralda	Bogotá, Nariño, Santander
Intereses	Salud, salud pública, higiene, estado físico	Salud, salud pública, higiene, estado físico

A partir de la aplicación de los clústeres del modelo de segmentación sociodemográfica, estas audiencias representan los intereses de la agencia de publicidad A para la promoción y comercialización del tapabocas desechable, bajo los requerimientos del cliente. No obstante, no es posible determinar la efectividad de esta clusterización final en 2 audiencias, dado que la campaña de marketing digital tiene lugar entre el 15 de enero de 2023 y el 15 de marzo de 2023, como parte de un plan piloto para la implementación de modelos de ML en las campañas de marketing digital de la agencia de publicidad A. Así las cosas, la presentación de esta tesis coincide temporalmente con la puesta en marcha de la campaña de marketing digital ya citada.

[9] Recordación de la marca

4. Conclusiones

- La aplicación de modelos de ML al Marketing Digital resulta ser efectiva para la clasificación de posibles grupos de usuarios de productos y servicios que se puedan ofrecer por estas plataformas. El uso generalizado de las redes sociales por gran parte de la población se convierte en un insumo importante para que estas campañas sean ejecutadas y que los modelos de clusterización sean replicados mediante las herramientas de segmentación, como están disponibles, por ejemplo, en Facebook Ads.

- Si bien, la utilización de USML arroja una clusterización eficiente, es necesaria una corroboración *ex post* para que, según los propósitos, los grupos cumplan con los objetivos de la campaña de marketing digital.

- El modelo de segmentación demográfica planteado en esta tesis tiene aplicaciones no sólo para la venta de tapabocas, como se enunció en este caso, sino también para la adquisición de otros productos y servicios relacionados con salud, como desinfectantes tópicos, productos destinados a la prevención de enfermedades respiratorias, servicios terapéuticos, entre otros. Por tanto, se constituye como un elemento fundamental para la planeación de las campañas de marketing futuras de la Agencia de Publicidad A que se relacionen con este tema.

- Sobre el tema, es imprescindible abordar otras aristas del fenómeno de la pandemia por Covid-19 y que pueden representar una oportunidad para las empresas para ofrecer sus catálogos de productos y servicios. Uno de los casos,

por ejemplo, son los efectos particulares del contagio y la vacunación por Covid-19 en la salud de las mujeres. Es así como este tema puede extenderse a otros estudios.

5. Anexos

5.1 Código de R empleado para el análisis descriptivo de las variables

```
1.
2. options(download.file.method = "wininet")
3.
4. install.packages("tidyr","readxl")
5. install.packages("ggpubr")
6. install.packages("lessR")
7. install.packages("plotrix")
8. install.packages("tidyverse")
9. install.packages("stringr")
10.    install.packages("forcats")
11.    install.packages("lubridate")
12.    install.packages("magrittr")
13.    install.packages("broom")
14.    install.packages("datasets")
15.
16.    library(tidyverse)
17.    library(stringr)
18.    library(forcats)
19.    library(lubridate)
20.    library(magrittr)
21.    library(broom)
22.    library(datasets)
23.    library(dplyr)
24.    library(tidyr)
25.    library(readxl)
26.    library(ggpubr)
27.    library(lessR)
28.    library(plotrix)
29.
30.
31.    dir()
32.    datos<-read.csv("Datos.csv", head=TRUE)
33.    head(datos)
34.    summary(datos)
```

```
35.
36.
37.    dim(datos)
38.
39.    muestra<-head(datos, n=50)
40.    View(muestra)
41.
42.
43.    str(datos)
44.
45.
46.    Demo <- select(datos,ID.de.caso, Nombre.departa
  mento, Edad, Sexo, Tipo.de.contagio, Ubicación.del.
  caso, Estado, Recuperado, Tipo.de.recuperación, Per
  tenencia.étnica)
47.
48.    Demo <- Demo %>%
49.       mutate(Etnia = case_when(Pertenencia.étnica
  %in% c(1) ~ "Indigena",
50.                                Pertenencia.étnica %
  in% c(2) ~ "ROM",
51.                                Pertenencia.étnica %
  in% c(3) ~ "Raizal",
52.                                Pertenencia.étnica %
  in% c(4) ~ "Palenquero",
53.                                Pertenencia.étnica %
  in% c(5) ~ "Negro",
54.                                Pertenencia.étnica %
  in% c(6) ~ "Otro",
55.                                TRUE ~ "Otro"))
56.    Demo1 <- select(Demo,ID.de.caso, Nombre.departa
  mento, Edad, Sexo, Tipo.de.contagio, Ubicación.del.
  caso, Estado, Recuperado, Tipo.de.recuperación, Etn
  ia)
57.    muestra2<-head(Demo1, n=50)
58.    View(muestra2)
59.    str(Demo1)
60.
61.
62.    table(Demo1$Sexo
63.    table(Demo1$Nombre.departamento)
```

```
64.    table(Demo1$Edad)
65.    table(Demo1$Tipo.de.contagio)
66.    table(Demo1$Ubicación.del.caso)
67.    table(Demo1$Estado)
68.    table(Demo1$Recuperado)
69.    table(Demo1$Tipo.de.recuperación)
70.    table(Demo1$Etnia)
71.
72.
73.    Demo2 <- mutate_if(Demo1, is.character, tolower
  )
74.    muestra3<-head(Demo2, n=50)
75.    View(muestra3)
76.    str(Demo2)
77.
78.    table(Demo2$Sexo)
79.    table(Demo2$Nombre.departamento)
80.    table(Demo2$Edad)
81.    table(Demo2$Tipo.de.contagio)
82.    table(Demo2$Ubicación.del.caso)
83.    table(Demo2$Estado)
84.    table(Demo2$Recuperado)
85.    table(Demo2$Tipo.de.recuperación)
86.    table(Demo2$Etnia)
87.
88.    head(Demo2)
89.    dim(Demo2)
90.
91.
92.    Demo2$Nombre.departamento<-
  as.factor(Demo2$Nombre.departamento)
93.    Demo2$Sexo<-as.factor(Demo2$Sexo)
94.    Demo2$Tipo.de.contagio<-
  as.factor(Demo2$Tipo.de.contagio)
95.    Demo2$Ubicación.del.caso<-
  as.factor(Demo2$Ubicación.del.caso)
96.    Demo2$Estado<-as.factor(Demo2$Estado)
97.    Demo2$Recuperado<-as.factor(Demo2$Recuperado)
98.    Demo2$Tipo.de.recuperación<-
  as.factor(Demo2$Tipo.de.recuperación)
99.    Demo2$Etnia<-as.factor(Demo2$Etnia)
```

```
100.   str(Demo2)
101.
102.
103.   summary(Demo2)
104.
105.   boxplot(Demo2$Edad, horizontal= TRUE, col="aqua
  marine2", main="Análisis univariado: Edad") #Edad
106.   boxplot(Edad ~
  Etnia, data = Demo2, col="aquamarine2", main="Relac
  ión edad vs Etnia") # Equivalente
107.   boxplot(Edad ~
  Sexo, data = Demo2, col="aquamarine2", main="Relaci
  ón edad vs Sexo") # Equivalente
108.   tabla1 <- table(Demo2$Estado, Demo2$Sexo)
109.   tabla4 <- table(Demo2$Sexo)
110.   tablaplot(tabla1, col = c("red", "blue"), main
  = "Estado paciente vs genero")
111.
112.
113.   tabla2<-table(Demo2$Sexo)
114.
115.   library(plotrix)
116.   pielabels<-c("Hombres","Mujeres")
117.   pie3D(tabla2, border=par("fg"), labels= pielabe
  ls, theta=pi/3, explode = 0.1, radius=1.3, main="Gr
  afico circular de género \n Femenino en rosado y
  masculino en
  azul",col=c("deeppink2","#1E90FF"),labelpos=lp, the
  tha = 0.9)
118.   help(pie3D)
119.
120.
121.   summary(Demo2)
122.   tabla<-table(Demo2$Estado,Demo2$Sexo)
123.   tabla
124.   medidas<-(prop.table(tabla)*100)
125.   medidas
126.   barplot(tabla, ylim =c(0,4000000), beside= T, l
  egend = T, main = "Estado", col=c("#7FFFD4", "#FF40
  40", "#CAFF70","darkslategray1","Pink"))
127.
```

```
128.   )
129.
130.   genero <- data.frame(gen = Demo2$Sexo)
131.   genero
132.   ggplot(genero, aes(x = gen)) +
133.   geom_bar(fill = "blue", width = 0.7) +
134.   coord_flip() +
135.   xlab("Genero") + ylab("Número de
  observaciones")+
136.   theme_bw()
137.
138.
139.   Edad1 <- data.frame(Edad = Demo2$Edad)
140.
141.   ggplot(Edad1, aes(x = Edad)) +
142.       geom_histogram(binwidth = 8, fill = "steelb
  lue") +
143.     xlab("Edad (Años)") + ylab("Conteo")
144.
145.
146.   departamento <- data.frame(dep = Demo2$Nombre.d
  epartamento)
147.
148.   ggplot(departamento, aes(x = dep)) +
149.   geom_bar(fill = "blue", width = 0.7) +
150.   coord_flip() +
151.   xlab("Departamento") + ylab("Número de
  observaciones")+
152.   theme_bw()
153.
154.
155.   Contagio <- data.frame(con = Demo2$Tipo.de.cont
  agio)
156.
157.   ggplot(Contagio, aes(x = con)) +
158.   geom_bar(fill = "cyan2", width = 0.7) +
159.   coord_flip() +
160.   xlab("Tipo de contagio") + ylab("Número de
  observaciones")+
161.   theme_bw()
162.
```

```
163.
164.   Ubicacion <- data.frame(ubi = Demo2$Ubicación.d
  el.caso)
165.
166.   ggplot(Ubicacion, aes(x = ubi)) +
167.   geom_bar(fill = "cyan2", width = 0.7) +
168.   coord_flip() +
169.   xlab("Ubicación del caso") + ylab("Número de
  observaciones")+
170.   theme_bw()
171.
172.
173.   Estado <- data.frame(Esta = Demo2$Estado)
174.
175.   ggplot(Estado, aes(x = Esta)) +
176.   geom_bar(fill = "cyan2", width = 0.7) +
177.   coord_flip() +
178.   xlab("Estado") + ylab("Número de
  observaciones")+
179.   theme_bw()
180.
181.
182.   Recuperado <- data.frame(Recu = Demo2$Recuperad
  o)
183.
184.   ggplot(Recuperado, aes(x = Recu)) +
185.   geom_bar(fill = "cyan2", width = 0.7) +
186.   coord_flip() +
187.   xlab("Recuperado") + ylab("Número de
  observaciones")+
188.   theme_bw()
189.
190.
191.   Tiporecu <- data.frame(tipo = Demo2$Tipo.de.rec
  uperación)
192.
193.   ggplot(Tiporecu, aes(x = tipo)) +
194.   geom_bar(fill = "cyan2", width = 0.7) +
195.   coord_flip() +
196.   xlab("Tipo de recuperación") + ylab("Número de
  observaciones")+
```

```
197.   theme_bw()
198.
199.
200.   Etnia <- data.frame(Etn = Demo2$Etnia)
201.
202.   ggplot(Etnia, aes(x = Etn)) +
203.   geom_bar(fill = "cyan2", width = 0.7) +
204.   coord_flip() +
205.   xlab("Etnia") + ylab("Número de
  observaciones")+
206.   theme_bw()
207.
208.
209.   Cruce <- data.frame(cru = Demo2$Sexo, cri = Dem
  o2$Recuperado)
210.
211.   ggplot(Cruce, aes(x = cru, fill = cri)) +
212.     geom_bar(position = "dodge") +
213.     labs(x = "Sexo", y = "Número de
  observaciones", fill = "Estado")
214.
215.
216.   Cruce2 <- data.frame(cru = Demo2$Nombre.departa
  mento, Estado = Demo2$Recuperado)
217.
218.   ggplot(Cruce2, aes(x = cru, fill = Estado)) +
219.     geom_bar() +
220.     coord_flip() +
221.     xlab("Departamento") + ylab("Número de
  observaciones")
222.
223.
224.   Cruce3 <- data.frame(cru = Demo2$Etnia, Estado
  = Demo2$Recuperado)
225.
226.   ggplot(Cruce3, aes(x = cru, fill = Estado)) +
227.     geom_bar() +
228.     coord_flip() +
229.     xlab("Etnia") + ylab("Número de
  observaciones")
230.
```

```
231.   Cruce4 <- data.frame(cru = Demo2$Etnia, cri = D
  emo2$Tipo.de.contagio)
232.
233.   Cruce4 %>%
234.       count(cru,cri) %>%
235.       ggplot(mapping = aes(x = cru, y = cri)) +
236.           geom_tile(mapping = aes(fill = n)) +
237.           labs(x = "Etnia", y = "Tipo de
  contagio", fill = "n")
```

5.2 Código de Phyton empleado para la clusterización

```
1. Original file is located at
2.     https://colab.research.google.com/drive/10ApMD2
  bFWpWfuqXEmyLw17Eivhau6Z6H
3. """
4.
5. !pip install kmodes
6. !pip install kneed
7. import numpy as np
8. import pandas as pd
9. import matplotlib.pyplot as plt
10.    import seaborn as sns
11.    import os
12.    RNDN = 42
13.
14.    from sklearn.cluster import KMeans, MeanShift,
  estimate_bandwidth
15.    from sklearn.preprocessing import
  StandardScaler, PowerTransformer
16.    from sklearn.metrics import silhouette_score
17.    from sklearn.model_selection import
  train_test_split
18.
19.    from kmodes.kprototypes import KPrototypes
20.
```

```
21.    from yellowbrick.cluster import
  KElbowVisualizer, SilhouetteVisualizer,
  InterclusterDistance
22.    from kneed import KneeLocator
23.
24.    from sklearn.decomposition import PCA
25.
26.
27.    from tqdm import tqdm
28.    import sys
29.    import warnings
30.    warnings.filterwarnings("ignore")
31.
32.    from google.colab import files
33.
34.    from google.colab import drive
35.
36.    drive.mount('/content/drive')
37.
38.    df =
  pd.read_csv('/content/drive/MyDrive/Datos.csv')
39.
40.    df.shape
41.
42.    df.columns
43.
44.    rm = df.drop (columns = ['ID de caso','fecha
  reporte web', 'Fecha de notificación',
45.          'Código DIVIPOLA departamento',
46.          'Código DIVIPOLA municipio', 'Nombre
  municipio',
47.          'Unidad de medida de edad',
48.          'Código ISO del país',
49.          'Nombre del país', 'Fecha de inicio de
  síntomas',
50.          'Fecha de muerte', 'Fecha de
  diagnóstico', 'Fecha de recuperación',
51.          'Nombre del grupo étnico'])
52.
53.    rm.dtypes
54.
```

```
55.     rm.head(3)
56.
57.     rm.describe()
58.
59.     rm.info()
60.
61.     rm.columns
62.
63.     rm.columns =
  ['Departamento','Edad','Sexo','Tipocontagio','Ubica
  cion','Estado','Recuperado','Tiporecuperacion','Etn
  ica']
64.     rm.columns
65.
66.     rm.loc[rm.Etnica == 1, "Etnia"] = "Indigena"
67.     rm.loc[rm.Etnica == 2, "Etnia"] = "ROM"
68.     rm.loc[rm.Etnica == 3, "Etnia"] = "Raizal"
69.     rm.loc[rm.Etnica == 4, "Etnia"] = "Palenquero"
70.     rm.loc[rm.Etnica == 5, "Etnia"] = "Negro"
71.     rm.loc[rm.Etnica == 6, "Etnia"] = "Otro"
72.
73.     rm.head(3)
74.
75.     rm['Departamento'] =
  rm['Departamento'].str.upper()
76.     rm['Sexo'] = rm['Sexo'].str.upper()
77.     rm['Tipocontagio'] =
  rm['Tipocontagio'].str.upper()
78.     rm['Ubicacion'] = rm['Ubicacion'].str.upper()
79.     rm['Estado'] = rm['Estado'].str.upper()
80.     rm['Recuperado'] = rm['Recuperado'].str.upper()
81.     rm['Tiporecuperacion'] =
  rm['Tiporecuperacion'].str.upper()
82.     rm['Etnia'] = rm['Etnia'].str.upper()
83.
84.     rm.head(3)
85.
86.     tm = rm.drop (columns = ['Etnica'])
87.
88.     tm.head(3)
89.
```

```
90.    tm.isnull().any()
91.
92.    tk = tm.fillna('', inplace=True)
93.
94.    tm.isnull().any()
95.
96.    Factores = tm["Ubicacion"].value_counts()
97.
98.    Factores
99.
100.   general = labels=tm["Ubicacion"].unique()
101.
102.   general
103.
104.   tm.isnull().values.any()
105.
106.   tm.head
107.
108.   tm.isnull().any()
109.
110.   tm.isnull().sum()
111.
112.   pd.value_counts(tm['Sexo'])
113.
114.   tm.Departamento.value_counts()
115.
116.   tm1 = tm.copy()
117.   tm1.tail(2)
118.
119.   df_num = tm1.select_dtypes(exclude='object')
120.   df_cat = tm1.select_dtypes(include='object')
121.
122.   for c in df_num.columns:
123.      pt = PowerTransformer()
124.      df_num.loc[:, c] =
  pt.fit_transform(np.array(df_num[c]).reshape(-1,
  1))
125.    df_cat = pd.get_dummies(df_cat)
126.   df_cat
127.    dfb2 = pd.concat([df_num, df_cat], axis=1)
128.   dfb2
```

```
129.    from sklearn.preprocessing import
  StandardScaler
130.   tm1 = dfb2.copy()
131.   tm1 = StandardScaler().fit_transform(tm1)
132.
133.   from sklearn.decomposition import PCA
134.   import matplotlib.pyplot as plt
135.
136.   pca = PCA(n_components=15)
137.   res_pca = pca.fit_transform(tm1)
138.    features = range(pca.n_components_)
139.   plt.bar(features,
  pca.explained_variance_ratio_, color='blue')
140.   plt.xlabel('PCA features')
141.   plt.ylabel('variance %')
142.   plt.xticks(features)
143.
144.   df_pca = pd.DataFrame(res_pca)
145.
146.   from sklearn.decomposition import PCA
147.   import matplotlib.pyplot as plt
148.
149.   pca = PCA(n_components=10)
150.   res_pca = pca.fit_transform(tm1)
151.
152.   features = range(pca.n_components_)
153.   plt.bar(features,
  pca.explained_variance_ratio_, color='blue')
154.   plt.xlabel('PCA features')
155.   plt.ylabel('variance %')
156.   plt.xticks(features)
157.
158.   df_pca = pd.DataFrame(res_pca)
159.    plt.scatter(df_pca[0], df_pca[1], alpha=.1,
  color='black')
160.   plt.xlabel('PCA 1')
161.   plt.ylabel('PCA 2')
162.
163.   from sklearn.cluster import KMeans
164.   import matplotlib.pyplot as plt
165.
```

```
166.    ks = range(2, 10)
167.    inertias_pca = []
168.    for k in ks:
169.
170.        model = KMeans(n_clusters=k)
171.
172.
173.        model.fit(df_pca.iloc[:,:])
174.
175.
176.        inertias_pca.append(model.inertia_)
177.
178.    plt.plot(ks, inertias_pca, '-o', color='black')
179.    plt.xlabel('number of clusters, k')
180.    plt.ylabel('inertia')
181.    plt.xticks(ks)
182.    plt.show()
183.
184.    !pip install --upgrade kneed
185.
186.    from kneed import KneeLocator
187.    inertia_knee_b3 = KneeLocator(
188.            range(2,10),
189.            inertias_pca,
190.            S=0.1, curve="convex", direction="decre
  asing")
191.
192.    K_inertia_b3 = inertia_knee_b3.elbow
193.    print("elbow at k =", f'{K_inertia_b3:.0f}
  clusters')
194.
195.    from scipy.sparse import random
196.    model1 = KMeans(n_clusters=K_inertia_b3, random
  _state=42)
197.    clusters = model1.fit_predict(tm1)
198.    tm.insert(0, "Cluster", clusters)
199.    tm.tail(2)
200.
201.    tm.tail(2)
202.
203.    tm.info()
```

```
204.
205.   pd.crosstab(tm["Cluster"],
206.               tm["Ubicacion"],
207.               values=tm["Ubicacion"],
208.               aggfunc="count",
209.               normalize=False)
210.
211.   def cluster_pies(df):
212.
213.      c = len(df.select_dtypes("object").nunique()
  )
214.
215.      K = df["Cluster"].nunique()
216.
217.       for k in tqdm(range(K)):
218.           dfc = df[df["Cluster"]==k]
219.           eda = dfc["Edad"].mean()
220.           cases = dfc.shape[0]
221.
222.           fig = plt.figure(figsize=(50, 12))
223.           fig.suptitle("Cluster " + str(k) + ":
  " + \
224.               f'{cases:,.0f}' + " cases | " + \
225.               "median edad " + f'{eda:.2f}')
226.
227.           ax1 = plt.subplot2grid((2,c),(0,1))
228.           plt.figure
229.           leyenda = []
230.           for Sexo, conteo in
  zip(dfc["Sexo"].unique(),round((((dfc["Sexo"].valu
  e_counts())/6345115))*100),ndigits=2)): leyenda.app
  end(Sexo + ' (' + str(conteo) + ' %)')
231.           plt.pie(dfc["Sexo"].value_counts())
232.           plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
233.           plt.title("Sexo")
234.           plt.axis('equal')
235.           plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
236.           plt.show()
237.
```

```
238.            ax2 = plt.subplot2grid((2,c),(0,2))
239.            plt.figure
240.            leyenda = []
241.            for Departamento, conteo in
  zip(dfc["Departamento"].unique(),round(((((dfc["Dep
  artamento"].value_counts())/6345115))*100),ndigits=
  2)): leyenda.append(Departamento + '
  (' + str(conteo) + ' %)')
242.            plt.pie(dfc["Departamento"].value_count
  s())
243.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
244.            plt.title("Departamento")
245.            plt.axis('equal')
246.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
247.            plt.show()
248.
249.            ax3 = plt.subplot2grid((2,c),(0,3))
250.            plt.figure
251.            leyenda = []
252.            for Tipocontagio, conteo in
  zip(dfc["Tipocontagio"].unique(),round(((((dfc["Tip
  ocontagio"].value_counts())/6345115))*100),ndigits=
  2)): leyenda.append(Tipocontagio + '
  (' + str(conteo) + ' %)')
253.            plt.pie(dfc["Tipocontagio"].value_count
  s())
254.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
255.            plt.title("Tipocontagio")
256.            plt.axis('equal')
257.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
258.            plt.show()
259.
260.            ax4 = plt.subplot2grid((2,c),(0,4))
261.            plt.figure
262.            leyenda = []
263.            for Ubicacion, conteo in
  zip(dfc["Ubicacion"].unique(),round(((((dfc["Ubicac
```

```
  ion"].value_counts())/6345115))*100),ndigits=2)): l
  eyenda.append(Ubicacion + ' (' + str(conteo) + '
  %)')
            plt.pie(dfc["Ubicacion"].value_counts()
  )
            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
            plt.title("Ubicacion")
            plt.axis('equal')
            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
            plt.show()

    def cluster_piess(df):


        c = len(df.select_dtypes("object").nunique()
  )

        K = df["Cluster"].nunique()

         for k in tqdm(range(K)):
            dfc = df[df["Cluster"]==k]
            eda = dfc["Edad"].mean()
            cases = dfc.shape[0]

            fig = plt.figure(figsize=(50, 12))
            fig.suptitle("Cluster " + str(k) + ":
  " + \
                f'{cases:,.0f}' + " cases | " + \
                "median edad " + f'{eda:.2f}')

            ax5 = plt.subplot2grid((2,c),(0,1))
            plt.figure
            leyenda = []
            for Estado, conteo in
  zip(dfc["Estado"].unique(),round(((((dfc["Estado"].
  value_counts())/6345115))*100),ndigits=2)): leyenda
  .append(Estado + ' (' + str(conteo) + ' %)')
            plt.pie(dfc["Estado"].value_counts())
```

```
293.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
294.            plt.title("Estado")
295.            plt.axis('equal')
296.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
297.            plt.show()
298.
299.            ax6 = plt.subplot2grid((2,c),(0,2))
300.            plt.figure
301.            leyenda = []
302.            for Recuperado, conteo in
  zip(dfc["Recuperado"].unique(),round((((dfc["Recup
  erado"].value_counts())/6345115))*100),ndigits=2)):
   leyenda.append(Recuperado + ' (' + str(conteo) + '
  %)')
303.            plt.pie(dfc["Recuperado"].value_counts(
  ))
304.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
305.            plt.title("Recuperado")
306.            plt.axis('equal')
307.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
308.            plt.show()
309.
310.            ax7 = plt.subplot2grid((2,c),(0,3))
311.            plt.figure
312.            leyenda = []
313.            for Tiporecuperacion, conteo in
  zip(dfc["Tiporecuperacion"].unique(),round((((dfc[
  "Tiporecuperacion"].value_counts())/6345115))*100),
  ndigits=2)): leyenda.append(Tiporecuperacion + '
  (' + str(conteo) + ' %)')
314.            plt.pie(dfc["Tiporecuperacion"].value_c
  ounts())
315.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
316.            plt.title("Tiporecuperacion")
317.            plt.axis('equal')
```

```
318.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
319.            plt.show()
320.
321.            ax8 = plt.subplot2grid((2,c),(0,4))
322.            plt.figure
323.            leyenda = []
324.            for Etnia, conteo in
  zip(dfc["Etnia"].unique(),round((((dfc["Etnia"].va
  lue_counts())/6345115))*100),ndigits=2)): leyenda.a
  ppend(Etnia + ' (' + str(conteo) + ' %)')
325.            plt.pie(dfc["Etnia"].value_counts())
326.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
327.            plt.title("Etnia")
328.            plt.axis('equal')
329.            plt.legend(leyenda,bbox_to_anchor=(0, 1
  ))
330.            plt.show()
331.
332.    def cluster_profile(df):
333.        dfc = df.groupby("Cluster").agg({
334.            "Edad": "mean",
335.            "Departamento": lambda
  x: x.value_counts().index[0],
336.            "Sexo": lambda
  x: x.value_counts().index[0],
337.            "Tipocontagio": lambda
  x: x.value_counts().index[0],
338.            "Ubicacion": lambda
  x: x.value_counts().index[0],
339.            "Estado": lambda
  x: x.value_counts().index[0],
340.            "Recuperado": lambda
  x: x.value_counts().index[0],
341.            "Tiporecuperacion": lambda
  x: x.value_counts().index[0],
342.            "Etnia": lambda
  x: x.value_counts().index[0]
343.                                        })
344.
```

```
345.         cluster_pies(df)
346.         return dfc
347.
348.   cluster_profile(tm).T
349.
```

Bibliografía

Bayoude, K., Ouassit, Y., Ardchir, S., & Azouazi, M. (2018a). How machine learning potentials are transforming the practice of digital marketing: State of the art. *Periodicals of Engineering and Natural Sciences*, *6*(2), 373–379. https://doi.org/10.21533/pen.v6i2.526

Bayoude, K., Ouassit, Y., Ardchir, S., & Azouazi, M. (2018b). How machine learning potentials are transforming the practice of digital marketing: State of the art. *Periodicals of Engineering and Natural Sciences*, *6*(2), 373–379. https://doi.org/10.21533/PEN.V6I2.526

Benton, M., Batalova, J., Davidoff-Gore, S., & Schmidt, T. (2021). *COVID-19 and the State of Global Mobility in 2020*. Migration Policy Institute & International Organization for Migration. https://publications.iom.int/books/covid-19-and-state-global-mobility-2020

Chinnamgari, S. K. (2019). *R Machine Learning Projects. Implement supervised, unsupervised, and reinforcement learning techniques using R 3.5*. Packt Publishing. https://books.google.com.co/books?hl=es&lr=&id=4dKDDwAAQBAJ&oi=fnd&pg=PP1&dq=machine+learning+supervised+unsupervised&ots=pxmI3Mf0tG&sig=9GPNZQuvrIpn9bTYP8nvqIih0_o&redir_esc=y#v=onepage&q=machine learning supervised unsupervised&f=true

Concepto. (n.d.). *Redes Sociales - Qué son, tipos, ejemplos, ventajas y riesgos*. Concepto. Retrieved December 2, 2022, from https://concepto.de/redes-sociales/

De Mauro, A., Sestino, A., & Bacconi, A. (2022). Machine learning and artificial intelligence use in marketing: a general taxonomy. *Italian Journal of Marketing*, 1–19. https://doi.org/10.1007/S43039-022-00057-W

Departamento Nacional de Planeación. (2017). *Indicadores sociodemográficos*. Departamento Nacional de Planeación. Observatorio de Familia. https://observatoriodefamilia.dnp.gov.co/Sistema-de-monitoreo/Indicadores-sociodemográficos/Tipologias-de-familias/Paginas/Indicadores-

sociodemográficos.aspx

Deutsche Welle. (2022, November 21). China: nuevos confinamientos tras repunte de COVID. *Deutsche Welle*. https://www.dw.com/es/china-nuevos-confinamientos-tras-repunte-de-covid/av-63835652

Ebrahimi, P., Basirat, M., Yousefi, A., Nekmahmud, M., Gholampour, A., & Fekete-Farkas, M. (2022). Social Networks Marketing and Consumer Purchase Behavior: The Combination of SEM and Unsupervised Machine Learning Approaches. *Big Data and Cognitive Computing*, *6*(35), 1–18. https://doi.org/10.3390/bdcc6020035

Erickson, B. F. (2010). *La publicidad*. Firmas Press. https://books.google.com.co/books?id=zHTpDwAAQBAJ&dq=que+es+publicidad+&lr=&hl=es&source=gbs_navlinks_s

Fernandez, L. (2021). *3M's N95 mask production worldwide 2019-2021*. Statista. https://www.statista.com/statistics/1232566/global-n95-mask-production-of-3m/

Guarda, T., Santos, M., Pinto, F., Silva, C., & Lourenço, J. (2012). A Conceptual Framework for Marketing Intelligence. *International Journal of E-Education, e-Business, e-Management and e-Learning*, *2*(6), 455–459. https://doi.org/10.7763/ijeeee.2012.v2.163

Guzmán Elisea, J. (2003). *Desarrollo de Campaña Publicitaria* [Universidad Autónoma de Nuevo León]. http://eprints.uanl.mx/5347/1/1020149150.PDF

Hagen, L., Uetake, K., Yang, N., Bollinger, B., Chaney, A. J. B., Dzyabura, D., Etkin, J., Goldfarb, A., Liu, L., Sudhir, K., Wang, Y., Wright, J. R., & Zhu, Y. (2020). How can machine learning aid behavioral marketing research? *Marketing Letters*, *31*(4), 361–370. https://doi.org/10.1007/S11002-020-09535-7/TABLES/1

Hair Jr., J. F., & Sarstedt, M. (2021). Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. *Journal of Marketing Theory and Practice*, *29*(1), 65–77. https://doi.org/10.1080/10696679.2020.1860683

Held, D. (2002). Introducción. In *Transformaciones globales : Política, economía y cultura*.

Oxford University Press.

Instituto Nacional de Estadística e Informática de la República del Perú. (2018). *Definición de indicadores sociodemográficos*. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1753/definiciones.pdf

Instituto Nacional de Salud. (2020). Sobre el dataset de casos de COVID-19 en Colombia. In *วารสารวิชาการมหาวิทยาลัยอีสเทิร์นเอเชีย* (Vol. 4, Issue 1). https://www.ins.gov.co/BibliotecaDigital/dataset-casos.pdf

Jara, L. (2015, August 8). Indicadores Sociodemográficos. *Observatorio Económico Social UNR*. https://observatorio.unr.edu.ar/indicadores-sociodemograficos/

Kaličanin, K., Čolović, M., Njeguš, A., & Mitić, V. (2019). Benefits of Artificial Intelligence and Machine Learning in Marketing. *Sinteza 2019 - International Scientific Conference on Information Technology and Data Related Research*, 472–477. https://doi.org/10.15308/sinteza-2019-472-477

León, A. (2022, December 12). Tapabocas en Colombia: Minsalud informa que no volverá a ser obligatorio. *El Tiempo*. https://www.eltiempo.com/salud/tapabocas-en-colombia-minsalud-informa-que-no-volvera-a-ser-obligatorio-725252

Luna, M. (2004). Redes sociales. *Revista Mexicana de Sociología*, *66*, 59. https://doi.org/10.2307/3541443

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing – Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481–504. https://doi.org/10.1016/j.ijresmar.2020.04.005

McPherson, J. M., Popielarz, P. A., & Drobnic, S. (1992). Social Networks and Organizational Dynamics. *American Sociological Review*, *57*(2), 153. https://doi.org/10.2307/2096202

Ngai, E., & Wu, Y. (2022). Machine learning in marketing: A literature review, conceptual framework, and research agenda. *Journal of Business Research*, *145*, 35–48.

https://doi.org/10.1016/j.jbusres.2022.02.049

Olivar Urbina, N. (2021). El proceso de posicionamiento en el marketing: pasos y etapas. *Academia & Negocios*, *7*(1), 55–64. https://revistas.udec.cl/index.php/ran/article/view/3066/3179

Organización Mundial de la Salud. (2021). *Enfermedad por coronavirus (COVID-19): Accesibilidad y asignación de las vacunas*. Organización Mundial de La Salud. https://www.who.int/es/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-vaccine-access-and-allocation

Organización Mundial de la Salud. (2022). *Enfermedad por el coronavirus (COVID-19): Vacunas*. Organización Mundial de La Salud.

Organización Panamericana de la Salud. (2021). *Colombia recibe las primeras vacunas que llegan a las Américas a través del Mecanismo COVAX*. Organización Panamericana de La Salud. https://www.paho.org/es/noticias/1-3-2021-colombia-recibe-primeras-vacunas-que-llegan-americas-traves-mecanismo-covax

Petteri, L. (2018). *Inteligencia artificial. 101 cosas que debes saber hoy sobre nuestro futuro*. Editorial Planeta. https://static0planetadelibroscom.cdnstatics.com/libros_contenido_extra/40/39308_Inteligencia_artificial.pdf

Portafolio. (2022, December 2). ¿Volverá uso obligatorio de tapabocas en Colombia? Esto dice Minsalud. *Portafolio*. https://www.portafolio.co/economia/gobierno/volvera-uso-obligatorio-de-tapabocas-en-espacios-cerrados-575030

Santos, M. (1993). Los espacios de la globalización. *Anales de Geografía de La Universidad Complutense*, *13*, 69–77. https://revistas.ucm.es/index.php/AGUC/article/download/AGUC9393110069A/31671

Ullal, M. S., Hawaldar, I. T., Soni, R., & Nadeem, M. (2021). The Role of Machine Learning in Digital Marketing. *SAGE Open*, *11*(4), 1–12. https://doi.org/10.1177/21582440211050394/FORMAT/EPUB

Vargas, P. (2021). Más de 50 países han desarrollado aplicaciones móviles para rastrear el nuevo coronavirus. *La República*. https://www.larepublica.co/internet-economy/mas-de-50-paises-desarrollaron-aplicaciones-para-rastrear-el-nuevo-coronavirus-3124948

Viteri Luque, F. E., Herrera Lozano, L. A., & Bazurto Quiroz, A. F. (2017). Las Tendencias del Marketing: Cuáles son y definiciones. *Recimundo*, *1*(5), 974–988. https://doi.org/10.26820/recimundo/1.5.2017.974-988

World Health Organization. (2021). *Coronavirus disease (COVID-19) advice for the public: When and how to use masks*. World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/when-and-how-to-use-masks

World Health Organization. (2022a). *Coronavirus disease (COVID-19): Masks*. World Health Organization. https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-masks

World Health Organization. (2022b). *Tracking SARS-CoV-2 variants*. World Health Organization. https://www.who.int/activities/tracking-SARS-CoV-2-variants