UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

FACULTAD DE ECONOMÍA

“ANÁLISIS EXPLORATORIO DE LOS SINIESTROS HOSPITALARIOS EN LOS SEGUROS DE GASTOS MÉDICOS MAYORES PARA MÉXICO, 2019-2020”

TESIS

QUE PARA OBTENER EL TÍTULO DE

LICENCIADA EN ACTUARÍA

PRESENTA:

PAOLA LIZETH PERALTA MIRANDA

ASESOR:

M. en E.S.R y M. Martha Paola Hernández Soto

TOLUCA ESTADO DE MÉXICO

JUNIO DEL 2023

Tabla de Contenido

Introducción general

El objetivo principal de este trabajo es realizar un análisis exploratorio de datos de los siniestros hospitalarios ocurridos entre 2019 y 2020 en los seguros de gastos médicos mayores a través de la técnica clúster, la cual a través de la segregación de variables cuantitativas (edad, monto de hospitalización, montos honorarios médicos), modelará algunos conglomerados, con base en esta división, mismos que podrán separar a su vez de las variables cualitativas que contengan la base de datos, logrando así, tener todas las variables en el estudio.

Dentro de los objetivos del presente trabajo se encuentra; adquirir información necesaria sobre los seguros de gastos médicos mayores, conocer las técnicas y herramientas del análisis multivariado y, sobre todo, el análisis de conglomerados o análisis clúster, para poder encontrar relación entre los siniestros asignados a cada clúster y poder encontrar diferencias entre los años 2019 y 2020.

De esta forma, se plantea una hipótesis principal: la pandemia causada por el virus SARS-CoV-2 tuvo un impacto en el número de siniestros hospitalarios relativos a las enfermedades crónico-degenerativas, lo que habría afectado el comportamiento de los datos en los clústers y, por ende, la forma de agrupación.

Por tal motivo, es posible que la pandemia imposibilitara a los servicios de salud la capacidad de otorgar un tratamiento oportuno a enfermedades diversas, debido a la alta demanda de atención médica para los pacientes con el virus; en otras palabras, que enfermedades que en 2019 eran consideradas de “menor riesgo” o “menor costo”, en 2020 dejaron de serlo. Por lo tanto, se considera pertinente estudiar los datos disponibles y utilizar algunas herramientas estadísticas para poder identificar patrones o relación entre siniestros y poder comprender mejor el impacto de la pandemia a las enfermedades crónico-degenerativas.

Para lograr comprobar la hipótesis y alcanzar los objetivos planteados el trabajo se compone de cuatro capítulos, organizados de la siguiente forma: los primeros dos capítulos darán el marco teórico-conceptual al estudio, el seguro y el análisis multivariado, respectivamente, cuya finalidad radica en la comprensión de la importancia de ambos conceptos; por otra parte, en el capítulo tres se explicará todo lo referente al análisis metodológico, es decir, cómo aplicar el análisis clúster a los datos presentados, y finalmente el capítulo cuatro presentará los resultados obtenidos del modelo, mostrando la formación de clústers y el tipo de datos contenidos en cada uno de ellos.

Breve descripción de los capítulos

El seguro es una herramienta que surgió muchos siglos atrás y que ha prevalecido hasta nuestros días, cuyo origen puede remontarse a las civilizaciones antiguas como la Babilónica; no obstante, el desarrollo del seguro moderno inicia en Europa en el siglo XVII y XVIII. Desde entonces ha ido diversificándose y tomando nuevas formas, hasta convertirse en parte integral del sector financiero frente a un mundo cada vez más incierto. En México, por ejemplo, el sector asegurador ha crecido en las ultimas décadas, y ofrece una amplia variedad de seguros, entre ellos: de automóviles, de vida o de gastos médicos.

Es importante señalar los principales componentes de un seguro: un riesgo incierto, un contrato de seguro, una prima (obligación del asegurado con la aseguradora), indemnización al ocurrir el siniestro (respuesta de la aseguradora con el asegurado).

Entrando en materia, el seguro que se empleará para este estudio pertenece al ramo de accidentes y enfermedades, siendo uno de los de mayor participación en México, y de los que mayor interés genera en el mercado. Así también, dentro del ramo de accidentes, el seguro de mayor comercialización es el seguro de gastos médicos mayores, que se ha convertido en una necesidad para las personas, en especial aquellas que padecen o están predispuestas a ciertas enfermedades graves. Asimismo, a partir del inicio de la pandemia en 2020 y el conocimiento sobre las repercusiones de este virus en pacientes con ciertas enfermedades crónico-degenerativas, resulta fundamental saber cuáles de estas son las que más aquejan a los mexicanos, mismas que mencionaremos a continuación.

México tiene como principales causas de muerte las enfermedades del corazón, la diabetes mellitus, los tumores malignos, las enfermedades del hígado, las enfermedades cerebrovasculares, la influenza y la neumonía, la enfermedad pulmonar obstructiva crónica (EPOC) y la insuficiencia renal, las cuales no son nada económicas; en este trabajo hay un apartado que explica de manera rápida que son estas enfermedades y el costo promedio para algunos tratamientos, en los cuales incluso pueden ser cifras millonarias.

Además de revisar la literatura pertinente para definir y delimitar tanto el instrumento con el que se trabajará (el seguro de gastos médicos mayores), como las enfermedades a tratar, es importante conocer qué herramientas apoyarán el estudio, mediante el análisis multivariado, el cual es un conjunto de técnicas estadísticas que apoyan el análisis de datos complejos y permiten obtener información sobre el comportamiento de las variables, logrando incluso, en algunos casos, predecir resultados a futuro.

El análisis multivariante se compone de distintas técnicas, las cuales puedes ser divididas en explicativas (aquellas que muestran interacción entre variables) y descriptivas (muestran la interacción entre los datos). En este último grupo se encuentra el análisis clúster, el cual será utilizado para realizar el análisis exploratorio en el presente trabajo.

El análisis clúster implica la identificación de grupos, conglomerados o segmentos cuyos datos sean similares. Para poder diferenciar qué tan parecido es un dato de otro, se utiliza la medición de distancias con diferentes técnicas, la más común y base de algunas otras, la distancia euclídea; también se explican los criterios de selección óptima para el número de clústers, para finalizar con la metodología de clusterización k-Means. El principio máximo en el análisis clúster es la minimización de la distancia intraclúster y la maximización de la distancia intraclúster; esto es, que a menor distancia mayor similitud entre datos.

Ahora, si bien el marco teórico-conceptual ayuda a entender el trabajo, en el último apartado se mostrarán, además, los resultados obtenidos al implementar el análisis clúster a la base de datos de la Comisión Nacional de Seguros y Fianzas (CNSF), tratando de encontrar alguna o algunas diferencias entre los años 2019 y 2020, mismas que, de forma indirecta, están relacionadas con la emergencia sanitaria mundial.

En la parte final del presente trabajo se utilizan apoyos gráficos para entender la información contenida en cada uno de los clústers de cada año, cerrando con el apartado de conclusiones, donde se expresa el cumplimiento de los objetivos y logros de la investigación, así como un apartado con el anexo del script utilizado en el software R.

Capitulo 1 El Seguro

1.1 Introducción del Capítulo

El seguro es una herramienta financiera esencial que permite a las personas y empresas protegerse contra los riesgos y contingencias que pueden afectar su patrimonio. La historia de los seguros se remonta a la antigua Babilonia, donde los contratos se usaban para proteger el comercio marítimo. Con el tiempo, la industria de seguros se ha convertido en un sector importante de la economía global.

En este capítulo se explora los conceptos básicos de los seguros, desde la definición de riesgo y primas, hasta pólizas y reclamos. También exploraremos la industria de seguros en México, mercado en crecimiento que ofrece a los clientes una amplia gama de productos y servicios.

Es importante prestar especial atención al ramo de accidentes y enfermedades, segmento importante del mercado de seguros de salud que brinda protección financiera en caso de enfermedad o lesión. Centrando la información en el ramo de accidentes y enfermedades, el seguro de gastos médicos mayores es uno de los productos más populares en este segmento, ya que cubre el costo de los servicios de salud necesarios para tratar enfermedades y lesiones graves.

Por tal razón, se desarrolla una breve explicación de las principales enfermedades responsables de muerte en México y cuyos tratamientos mantienen un costo elevado, se explora su relevancia, su costo promedio por paciente en México y cómo inciden en la salud pública. Para tal fin, este apartado proporciona información de las principales enfermedades causantes de las defunciones en México según el Instituto Nacional de Estadística y Geografía (INEGI).

En conclusión, este capítulo brindará una descripción general de los seguros y su importancia en la vida moderna, así como un estudio detallado de la industria aseguradora mexicana, enfocado en el ramo de accidentes y enfermedades, principalmente el seguro de gastos médicos mayores, información que servirá como un preámbulo a la justificación de la selección de enfermedades hecha en esta investigación.

1.2 Breve Historia del Seguro

Desde el surgimiento de la humanidad, los seres humanos se han desarrollado en ambientes hostiles, acechados por constantes riesgos físicos, que lo pueden conducir a perder algún miembro o, incluso, a la muerte; posteriormente, conforme las sociedades humanas fueron volviéndose más complejas y surgió la propiedad, los bienes de valor considerable también corrían el riesgo de la sufrir una pérdida total o parcial.

A partir de esos primeros humanos se ha tratado de disminuir el riesgo a sufrir algún daño, reduciendo las actividades que pueden poner en riesgo su integridad u ocasionar alguna lesión o pérdida. Prueba de esta búsqueda de la seguridad es la vida en comunidad, la construcción de herramientas y el conocimiento del entorno. Estas medidas de protección se han ido sofisticando con el paso de los siglos, como se ha visto a lo largo de la historia, y se han ido implementando conocimientos empíricos para poder minimizar, ya que no eliminar, el perjuicio.

En India, Egipto, Grecia y Roma, aunque no existía como tal el concepto de seguro, existía una organización cuyo objetivo principal era la previsión de algunos riesgos, principalmente en el mar. Específicamente en Grecia y Roma existía la ley de Rodas, que señalaba “las obligaciones de los cargadores de contribuir a la indemnización de los graves daños causados en perjuicio común en caso de tempestad o rescate de buque apresado por enemigos o piratas.” (Minzoni Consorti, 2005)

Los antecedentes del seguro podrían remontarse a Babilonia, en la Edad Antigua, donde se brindaba protección contra robos, dado que cuando alguno sufría este tipo de percances, la comunidad entera, de forma solidaria, apoyaba a los perjudicados (Deance Rupit & Osorio López, 2004).

Del mismo modo, la organización por caravanas ayudaba a la dispersión del riesgo, ya que la pérdida de una de ellas no generaba mayor perjuicio, dado que solo se perdería una porción de la totalidad de mercancías.

Blanco (2001) hace notar que el “Préstamo a la Gruesa” serviría como vestigio al seguro: dicho contrato estaba efectuado entre el dueño de una embarcación y un banquero que financiara el viaje; este contrato especificaba que si el dueño de la embarcación pedía financiación a un banco y dicha embarcación se perdía durante el viaje, el préstamo se tomaría como cancelado.

A través de los años, cada civilización fue desarrollando, de acuerdo con sus necesidades, un contrato similar a lo que actualmente se conocería como seguro, derivando en diversos ramos; cada uno de ellos dentro de un contexto muy específico, pero teniendo algunas de las siguientes características: un ente que asumiera el riesgo, incertidumbre frente a un evento aleatorio e incierto que tiene posibilidades de ocurrir, dispersión del riesgo, pérdida monetaria y compensación del daño causado.

De esta manera se fue desarrollando un sector asegurador cada vez más complejo y arriesgado, pues, en la medida que hubiera más situaciones adversas a las que una persona o sus bienes estuvieran expuestos, tendría que existir una forma de medir la probabilidad de que el evento catastrófico sucediera, y a su vez estimar el impacto económico que causaría. De este modo se fueron creando empresas que se dedicaron a realizar esta serie de cálculos, y así surgieron las primeras empresas aseguradoras en Inglaterra entre los siglos XVII y XVIII (Mehr & Osler, 1994, pág. 38).

La industria aseguradora ha cambiado con los años, ha evolucionado y, sin duda, lo seguirá haciendo, pues, así lo hacen también las necesidades de cada sociedad: el ritmo de vida acelerado de los asegurados, el crecimiento de los países y los avances tecnológicos hacen que tanto el seguro como el contrato del seguro siga en cambio constante; en cierto modo, lo único que no cambiará es la incertidumbre de enfrentar un infortunio que puede causar una perdida.

Esta información nos ha ayudado a definir y delimitar qué es y para qué sirven los seguros, y aunque en la actualidad hay diversas formas de gestionar el riesgo, ninguna de ellas ofrece todavía una seguridad completa; sin embargo, el seguro es una opción eficaz (Schumacher, 2019), puesto que protege a los individuos frente a las consecuencias de los riesgos.

1.3 ¿Qué es el Seguro?

Existen muchas formas de definir el seguro: qué es, para qué sirve y los elementos que lo componen. Asimismo, puede ser desde un punto de vista jurídico, económico, estadístico o actuarial; pero en general y de forma muy concreta, podemos definirlo como un instrumento de protección ante la pérdida, destrucción, o daño de una persona, bien o propiedad.

El seguro participa en la pérdida y ayuda a crear certeza frente a un suceso catastrófico, basándose en transferir el riesgo a la aseguradora, la cual se encargará de reparar o indemnizar todo o parte del perjuicio producido por la realización del infortunio. Algunas definiciones resaltan al seguro como un servicio de solidaridad humana, la cual hace acreedora de una prestación; consistiendo en la transformación del riesgo en pagos periódicos presupuestables (Fundación MAPFRE, 2017).

Se puede abordar desde dos perspectivas el concepto de seguro; la primera, desde el punto de vista del asegurado, que puede ver al seguro no como un gasto sino más bien una previsión y ahorro, el cual se hace efectivo en caso de que un siniestro especificado ocurra. No obstante, la visión más acertada sería la de un instrumento de protección financiera ante el riesgo de perder un bien que sería difícil reponer debido a que representa un fuerte desembolso de dinero en un corto periodo de tiempo.

La segunda perspectiva corresponde a la compañía, quien se encarga de identificar, entre sus asegurados, aquellos que tenga similares o iguales necesidades de protección, uniendo de esta manera a un número grande de clientes que, a través de las primas pagadas ayudarán a la entidad aseguradora a hacer frente a la indemnización o reparación de un siniestro.

Este sistema de protección puede llevarse a cabo de diversas formas, por mencionar dos ejemplos: se puede proteger una persona en integridad física, salud o existencia, o bien, mediante seguros que cubren patrimonios o algún objeto que se considere expuesto a un riesgo que amenace con la destrucción total o parcial del ente asegurado.

> “El seguro es un contrato por el cual una de las partes en consideración a un precio que a ella se le pagó adecuado al riesgo, da seguridad a la otra parte de que ésta no sufrirá pérdidas, daño o perjuicio por el acaecimiento de los peligros especificados sobre ciertas cosas que pueden estar expuestas a tales peligros (Magee, 1947, pág. 34).”

El seguro mantiene elementos concretos que ayudan a delimitarlo como concepto. Posteriormente se enunciarán los conceptos y una breve definición de cada uno de ellos, sin estos conceptos el seguro no podría existir.

1.3.1 Elementos del Seguro

- **Asegurador/Aseguradora:** Parte que suscribe el contrato, entidad que, a través del contrato del seguro, asume las consecuencias dañosas producidas por el siniestro, volviendo la obligación de indemnizar cierta y concreta cuando este ocurre. Son entidades privadas, puede ser sociedad anónima, mutualista, mutualista cooperativa o mutualidad de previsión social.
- **Suma Asegurada**: Es el valor atribuido por el cual está protegido los bienes cubiertos por la póliza y cuyo importe es la obligación máxima que está obligado a pagar el asegurador, en caso del siniestro.
- **Asegurado**: Persona física o jurídica, la cual está expuesta a un riesgo del cual quiere protegerse, está comprometido a pagar las primas estipuladas y cuyo derecho es poder cobrar la indemnización cuando el siniestro se manifieste y tenga sus consecuencias.
- **Indemnización:** Importe que se pagará contractualmente la aseguradora en caso de producirse el siniestro, puede ser igual o menor a la suma asegurada.
- **Prima:** Aportaciones económicas periódicas que ha de satisfacer el asegurado con la aseguradora; en contraprestación por cubrir el riesgo, es el precio que pagará por mitigar la incertidumbre de la pérdida económica. La prima está calculada en función del riesgo, suma asegurada, duración del contrato y la cobertura de la póliza.
- **Interés Asegurable:** Principal objeto del seguro, aquello que está en riesgo y podría causar una pérdida económica en caso de que un siniestro se produzca; existe cierto deseo sincero de protección para que el siniestro no se realice, puesto que esto originaría un perjuicio en el patrimonio.

- **Siniestro:** Cualquier hecho que ponga en juego las garantías (bien asegurado) del contrato de seguro, es la ocurrencia de uno de los riesgos, acontecimiento inesperado y que se encuentra contemplado en la póliza de seguro.
- **Riesgo:** Probabilidad de ocurrencia de un suceso y las consecuencias (positivas o negativas) que pueden dañar a una persona o bien. Debe de tener ciertas características para ser un riesgo asegurable, incierto o aleatorio, posible, concreto, licito y fortuito. (Fundación MAPFRE)

1.4 El sector Asegurador en México

Conforme el seguro se fue expandiendo y llegando a distintas regiones del mundo, ha ido adecuándose a las necesidades de cada país. Se habla de que el sector asegurador será tan grande como la economía del país en que se esté ejerciendo. Del mismo modo que las necesidades cambian de acuerdo con la zona, región o nación, existen lugares donde sus habitantes están dispuestos a asumir ciertos riesgos puesto que no ven necesario tal o cual seguro.

Organismos internacionales como la Organización para la Cooperación y el Desarrollo Económico (*OECD* por sus siglas en inglés) realizan comparaciones en la penetración del sector asegurador en el mercado de los países. Estas comparaciones las hacen a partir de la relación primas-Producto Interno Bruto (PIB), matemáticamente formulado: (Primas/PIB), y cuyos datos son expresados en dólares[1] estadounidenses.

Bajo este análisis, en 2019 el país poseía un 2.4% de penetración en el mercado. Dado que este indicador se ve fuertemente influenciado por el ingreso y el nivel de educación financiera, México se posiciona por debajo de países que poseen características similares, como Colombia y Chile, que mantienen un 2.8% y 4.5% respectivamente, y aunque sus ingresos *per cápita* son menores o similares al de México, la cobertura promedio de la OCDE es de 8.9%, con un pago *per cápita* de 216 dólares americanos anuales en el ámbito de seguros (Comisión Nacional de Seguros y Fianzas, 2020).

Entre los años 1994 y 2017 el sector asegurador mexicano tuvo un crecimiento de un poco más del doble, pasando de 48 instituciones en 1994 a 98 en septiembre del 2017, siendo el punto más alto en 2015 con 105 aseguradoras en operación. Este incremento en compañías de seguros se debió al Tratado de Libre Comercio (TLC), el cual permitía que empresas extranjeras encontraran un nuevo mercado en México a través de empresas filiales (Romero Gatica, 2017).

[1] En este trabajo se les denominará dólares a los dólares americanos/ dólares estadounidenses

En el año 2020 el sector asegurador en México emitió $594 mil millones de pesos en primas, lo cual fue un incremento de 0.19% con respecto al año 2019, en el que fueron $593 mil millones (De la Rosa Analytics Solutions, 2021). Para el mismo periodo, en México operaban 112 instituciones de seguros y fianzas, de las cuales 102 eran aseguradoras y 10 instituciones de fianzas. De las 112 empresas, 13 tenían autorización de operar exclusivamente el ramo de vida, 63 para no vida y 36 de forma combinada vida y no vida (Comisión Nacional de Seguros y Fianzas, 2017).

Teniendo en cuenta información proporcionada por la Comisión Nacional de Seguros y Fianzas (2021), en México operan 112 instituciones pertenecientes a la industria, de las cuales 36 compañías se encuentran especializadas de la siguiente manera:

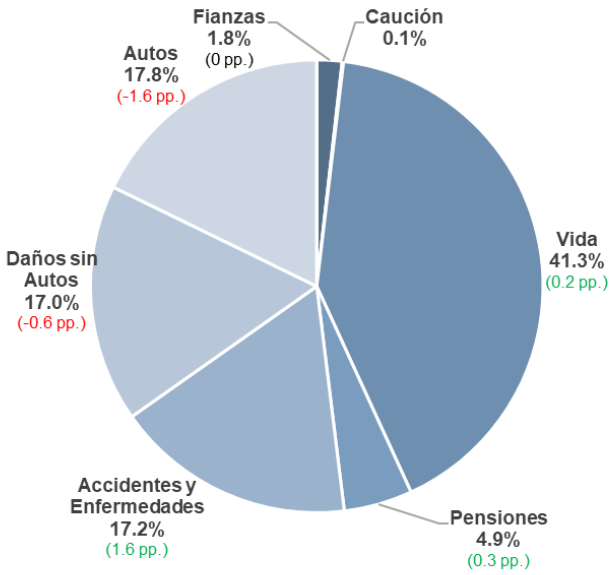
- 10 en Finanzas.
- 7 en Pensiones derivadas de las Leyes de Seguridad Social.
- 10 en el Ramo de Salud.
- 2 en el Ramo de Crédito a la Vivienda.
- 1 en el Ramo de Garantía Financiera.
- 6 en Seguros de Caución.

Para el año 2020, respecto al 2019, el sector asegurador decreció 2.3% en el ramo de vida, la operación de accidentes y enfermedades aumento 7.4%, mientras que en los seguros de pensiones derivadas de leyes de Seguridad Social aumentó 3.7%.[2]

[2] En términos reales.

El siguiente grafico muestra la composición de la cartera de seguros para finales de año del 2020.

Gráfico 1.1 Composición del Mercado Asegurador



Fuente: Elaboración propia con base a la Comisión Nacional de Seguros y Fianzas, 2017.

Como se mencionó anteriormente, se sabe que el sector asegurador va ligado a la economía del país en cuestión, pero a partir de finales del 2020 empezó una contracción en las economías mundiales, derivado del confinamiento causado por la enfermedad COVID-19, causada por el virus SARS-CoV-2; esto, a su vez, causaría el mismo efecto en el área aseguradora. No obstante, derivado del mismo confinamiento y todas las consecuencias sociales que trajo consigo, se espera que esta desaceleración en la penetración del mercado mexicano sea superada en un corto periodo de tiempo, para después tener un alza.

Esta situación resulta lógica, ya que el mundo se conmocionó frente a una nueva enfermedad que causó enormes cuentas hospitalarias, muertes y una saturación sanitaria nunca vista. La cancelación masiva de eventos, la disminución y retrasos en las transacciones mercantiles desveló un sentimiento de protección frente a estos percances, pues las pérdidas fueron inimaginables; a consecuencia de esto se prevé que los seguros se seguirán renovando y se puede adivinar qué camino seguirán, es decir, qué ramo(s) o tipo(s) de seguro tendrán más potencial para ser comercializados en México en el futuro, dejando clara las nuevas áreas de oportunidad del sector.

1.4.1 El Ramo de Accidentes y Enfermedades

El ramo de accidentes y enfermedades es de los de mayor demanda en México. Este ramo comprende 3 principales subramos, los cuales son:

- **Gastos Médicos Mayores**

  Brinda protección financiera para los gastos médicos originados por un accidente o enfermedad.

- **Accidentes Personales**

  Brinda protección financiera en caso de muerte accidental, pérdidas orgánicas, reembolsos de gastos médicos, incapacidad total o parcial.

- **Salud**

  Ayuda a prevenir enfermedades, conservar y reestablecer la salud del asegurado, ofreciendo protección financiera.

La composición del ramo es la siguiente: 92% en gastos médicos mayores, 4.8% accidentes personales y 3.1% en salud (De la Rosa Analytics Solutions, 2021). Para el 2020, el ramo de accidentes y enfermedades creció 7.4 % con respecto al año anterior, impulsado principalmente por el seguro de gastos médicos que también obtuvo su mayor alza en este periodo, esto es, del 8.1 puntos porcentuales (p.p); por el contrario, accidentes personales disminuyó 0.8 p.p y salud un incremento casi imperceptible de 0.1 p.p (Comisión Nacional de Seguros y Fianzas, 2017).

1.4.2 El Seguro de Gastos Médicos Mayores

Los seguros de gastos médicos mayores son planes de protección financiera para enfrentar gastos generados por accidentes o enfermedades, de los cuales nadie está exento, puede cubrir gastos hospitalarios, medicamentos, incluso en el seguimiento. Su finalidad es la protección económica en materia de salud, misma que genera tranquilidad.

A través de una prima, la entidad aseguradora cubrirá gastos hospitalarios, atención médica, intervenciones quirúrgicas, medicamentos, análisis clínicos, imagenología, incluso alimentos, entre otros servicios especificados en la póliza que puede o no incluir a alguno(s) dependiente(s) menores de 25 años. El monto de la prima será calculado de acuerdo con ciertos factores, como la edad, suma asegurada, cobertura(s) contratada(s) y el estado de salud del o de los asegurados; la aseguradora entrega algunos cuestionarios, los cuales deben ser contestados con sinceridad, ya que, en caso contrario, la aseguradora podrá rescindir del contrato, sin verse en la obligación de pagar o brindar el beneficio.

En los últimos cinco años, la prima de gastos médicos mayores ha crecido un 28.8%, el mayor porcentaje en este periodo; esto ha significado que los costos se hayan incrementado debido al aumento en las contrataciones de los seguros de gastos médicos mayores individuales.

De acuerdo con el Instituto Nacional de Estadística y Geográfica (INEGI), el 77% de la población mexicana cuenta con un programa de protección brindada por una Institución de Seguro Social, mientras que el otro 23% restante tiene que recurrir a su ahorros o ingresos corrientes para cubrir los gastos generados por algún accidente o enfermedad. Por otra parte, de dicha cifra poco más 7% tiene una póliza de seguro de gastos médicos (Comisión Nacional para la Protección y Defensa de los Usuarios Financieros, 2017), del mismo modo la Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros (Condusef), estima que hay 19 millones de asegurados en México y que solo el 27% es por un seguro de gastos médicos mayores.

Con base en los datos de la Condusef (2017), se sabe que el costo anual promedio de una consulta o tratamiento médico es cercano a los $3,786 pesos anuales, esto representaría el 9% del ingreso anual de una persona promedio, aunque es importante destacar que para personas con un ingreso menor podría representar hasta el 51% de su ingreso anual.

El boletín de la Comisión Nacional de Seguros y Fianzas (2021) declara que en México hay 32 empresas que están autorizadas para operar seguros de gastos médicos mayores, de las cuales 9 acaparan el 90% de participación y el índice de concentración es del 69.4%, donde las principales 5 aseguradoras son: Grupo Nacional Provincial (GNP) el 23.4%, Axa Seguros 17.9%, MetLife México 14.4%, Seguros Monterrey New York Life 9.8% y Seguros Inbursa 3.8%. De manera análoga, el crecimiento del ramo fue de 7.4%, impulsado principalmente por los seguros de gastos médicos mayores; como resultado, las empresas que más crecieron fueron: MetLife México un 2.4%, GNP 1.7%, Axa Seguros 1.4%, Mapfre México 0.6%, Seguros Monterrey New York Life 0.8%

Según el Análisis Estadístico de la CNSF (2021), seguros de gastos médicos mayores son contratados por empresas para sus trabajadores, puesto que es una prestación un tanto común en la industria privada. Esta forma de comercializar el seguro es a través de un seguro colectivo; dicho producto va dirigido a un conjunto de personas que forman parte de un grupo homogéneo o que tienen un vínculo en común, que en este caso particular consiste en ser trabajadores de la misma empresa.

No obstante, la otra forma de encontrar un seguro de gastos médicos mayores en el mercado es de forma individual o familiar, cuya finalidad es la protección del individuo y/o familia del contratante, cónyuge e hijos menores de 25 años. El costo de estos tipos de seguros dependerá de las edades, sexo, deducible, cobertura, número de integrantes y características específicas de la póliza.

Además, existe dos tipos de coberturas, la cobertura básica en la que la aseguradora se hace cargo de pagar honorarios médicos, medicamentos, servicios auxiliares para el diagnóstico de enfermedades o padecimientos, gastos hospitalarios, tratamientos, honorarios de enfermeras, servicios dentales y aparatos ortopédicos. A su vez, las coberturas adicionales son servicios opcionales que permiten crear, así, un seguro aún más completo, el cual puede incluir servicios de emergencia en el extranjero, gastos funerarios y enfermedades catastróficas. Para agregar este tipo de coberturas se debe de pagar un costo extra.

Aunque el principal objetivo del seguro de gastos médicos mayores es cubrir el riesgo económico frente a una enfermedad, ayudando a sufragar los gastos médicos que puedan poner su estabilidad económica en riesgo. El monto de los gastos a los que el asegurado puede hacer frente sin necesidad de recurrir al seguro se le conoce como deducible; si el gasto incurrido es mayor a este deducible, el asegurado pagará el costo del deducible y la aseguradora se hará cargo del resto. En la mayoría de los planes el asegurado debe cubrir el deducible, pero si además decide coparticipar con cierto porcentaje de los gastos generados en el evento, a este monto se le llama coaseguro. Entre más alto sea el deducible y el coaseguro, menor será la prima.

Los gastos médicos mayores cubren accidentes o enfermedades. Para este tipo de seguros, la Asociación Mexicana de Instituciones de Seguros (2016) define:

Accidente: Lesión sufrida involuntariamente por el titular o sus dependientes, por acción fortuita y violenta de una fuerza externa, siempre y cuando el suceso y el primer pago del gasto médico del tratamiento ocurra mientras la póliza se encuentre en vigor. Del mismo modo, las lesiones o enfermedades desarrolladas como consecuencia inmediata o directa del accidente quedaran protegidas para su tratamiento médico o quirúrgico, además, quedaran cubiertos también recaídas, complicaciones y secuelas consideradas del accidente previamente cubierto.

Enfermedad: Alteración de la salud que sufra el titular o sus dependientes, que no sean de origen congénito y amerite el tratamiento médico o quirúrgico, siempre y cuando el suceso y primer pago de gastos médicos ocurra mientras la póliza este en vigor. Las alteraciones que sean consecuencia inmediata de una enfermedad ya cubierta podrán verse beneficiadas por el seguro, en cuestión de atención y tratamiento médico o quirúrgico, así como recaídas, complicaciones y secuelas que la primera enfermedad cubierta desencadene.

Las enfermedades preexistentes pueden o no estar cubiertas. Se debe de entender como enfermedades preexistentes, aquellas cuyos síntomas o manifestaciones ocurrieron antes de la prescripción del plan, enfermedades que hayan causado un gasto para su diagnóstico señalando, un padecimiento que tuvo lugar antes de la fecha de inicio del plan y/o enfermedades que fueron aparentes a simple vista, por lo que no pudieron pasar desapercibidas.

1.5 Características de las Defunciones en México

Desde la estadística realizada por el INEGI (2020), fueron registradas 747,784 decesos en México; de este número, el 56.4% fueron hombres, 43.5% mujeres y existen 473 de sexo no especificado. La Ciudad de México es la entidad que posee el mayor número de muertes en el país, teniendo una tasa de defunciones de 70[3]. Del total de decesos, el 88.8% fueron muertes causados por enfermedades y problemas relacionados con la salud, como: Enfermedades del Corazón, Diabetes Mellitus, Tumores Malignos, Enfermedades del Hígado, Enfermedades Cerebrovasculares, Influenza y Neumonía, Enfermedad Pulmonar Obstructiva Crónica e Insuficiencia Renal.

La importancia del estudio de cada enfermedad radica en señalar el costo monetario que implica, para los afectados o para el sistema de salud, la recuperación o tratamiento de dichas enfermedades.

1.5.1 Enfermedades del Corazón

"Enfermedad del corazón" es un término general utilizado para referirse a problemas con el corazón y los vasos sanguíneos; entre las más comunes se encuentran: Enfermedad cardíaca coronaria, insuficiencia cardíaca, arritmia, enfermedad de las válvulas cardíacas, enfermedad arterial periférica, presión arterial alta (hipertensión), accidente cerebrovascular, enfermedad cardíaca congénita.

La Organización Mundial de Salud (Organizacion Mundial de la Salud, 2017) señala a las enfermedades cardiovasculares como la principal causa de muerte en todo el mundo, pues se cobra la vida de aproximadamente 17.9 millones de personas cada año. Más de las cuatro quintas partes de las muertes por enfermedades cardiovasculares se deben a enfermedades coronarias y accidentes cerebrovasculares, y un tercio de estas muertes ocurren prematuramente en personas menores de 70 años.

Una nueva investigación a cargo de la World Heart Federation (2016) muestra que el impacto económico de la principal causa de muerte en América Latina, las enfermedades cardíacas, supera los $30 mil millones; asimismo, en México el gasto es de $6,100 millones de dólares, representando el 4% del gasto destinado en salud, ya que en la población mexicana el 26% es afectada por alguna de estas enfermedades.

[3] La tasa de defunción está dada por el número de defunciones por cada 10,000 habitantes.

1.5.2 Diabetes Mellitus

La diabetes sacarina o diabetes *mellitus* es una enfermedad crónica que ocurre cuando el páncreas no produce suficiente insulina o el cuerpo no puede utilizar de manera efectiva el tipo de insulina que produce. La condición más común de la diabetes es la hiperglucemia (es decir, niveles altos de azúcar en la sangre), que, con el tiempo, daña gravemente muchos órganos y sistemas del cuerpo, especialmente los nervios y los vasos sanguíneos (Organizacion Mundial de la Salud, 2022).

En el marco del Dia Internacional de la Salud, la revista Forbes (2016) informa que, en Latinoamérica, 15.5 millones de personas tienen diabetes, mientras que en México 10.6 millones de habitantes la padecen, lo que resulta en un costo anual de $3,872 millones de dólares en atención. Del mismo modo, la Federación Mexicana de Diabetes (2019) señala que el costo anual de un paciente no controlado con complicaciones diabéticas es de $1,163,028 pesos, mientras que un paciente controlado puede incurrir en un costo de 88,024 pesos en el mismo periodo.

1.5.3 Tumores Malignos

Cáncer es un término común utilizado para referirse a un gran grupo de enfermedades que pueden afectar cualquier parte del cuerpo; también conocido como melanoma o neoplasia maligna, es la principal causa de muerte en el mundo. En el 2020, esta enfermedad causó la muerte de 10 millones de personas (Organización Mundial de la Salud, 2022).

Según estadísticas del Instituto Nacional del Cáncer (2021), el cáncer es considerado una de las principales causas de muerte a nivel mundial. Durante el 2018, se registraron 18.1 millones de casos nuevos y al año se estima que 9.5 millones de personas en todo el mundo mueren por padecimientos referentes al cáncer y los tumores malignos.

En México, 14 de cada 100 personas mueren de cáncer y la esperanza de vida de quienes padecen la enfermedad es de 63 años. Su tasa de mortalidad está aumentando rápidamente. Entre 1990 y 2019, el número de muertos se duplicó: de 41,000 a 89,000. Además, se estima que habrá costos directos e indirectos que van de los 23 a 30 mil millones de pesos por año, equivalentes a una quinta parte del presupuesto total del Instituto de Salud y Bienestar (Insabi) para el año 2020 (Flamand, 2021).

1.5.4 Enfermedades del Hígado

El hígado es el órgano más grande del cuerpo y lo ayuda a digerir los alimentos, almacenar energía y eliminar toxinas. Existen muchos tipos de enfermedades hepáticas que incluyen: enfermedades causadas por virus (hepatitis A, B o C), enfermedades causadas por drogas, venenos o toxinas, o consumo excesivo de alcohol (hígado graso y cirrosis) (American Kidney Fund, 2022), así como enfermedades genéticas (enfermedad de la hemoglobina y de Wilson).

La Fundación para el Estudio de las Hepatitis Virales (2019), a través de su revista *Journal of Hepatology*, indicó que las enfermedades hepáticas se consideran un problema de salud grave, puesto que son causantes de aproximadamente 2 millones de muertes cada año. En los países industrializados, la enfermedad más común es la grasa hígado, que resulta del sobrepeso y la obesidad. En México, según el INEGI (2021), las enfermedades relacionadas con el hígado son la sexta causa de decesos de la población mayor de 25 años.

1.5.5 Enfermedades Cerebrovasculares

Un accidente cerebrovascular ocurre cuando el flujo de sangre a una parte del cerebro se detiene o cuando un vaso sanguíneo en el cerebro se rompe, causando un sangrado dentro de la cabeza.

El INEGI (2021) relaciona 37,021 muertes en México causadas por enfermedades cerebrovasculares. Para el año 2020, no obstante, se señala que “Uno de cuatro mexicanos sufre un infarto cerebral y su incidencia representa un costo de atención de alrededor de 5.6 miles de millones de pesos” (Garduño, 2021).

1.5.6 Influenza y Neumonía

La neumonía es una infección de uno o ambos pulmones la cual puede ser causada por bacterias, virus y hongos patógenos. Inflamación de los alvéolos y llenados de líquido o pus. Las enfermedades pueden variar de casos leves hasta casos graves, según el tipo de agente causal, el estado general de salud y la edad de la persona que las padece. Alternativamente, puede desarrollarse debido a una infección viral como un resfriado o gripe.

La neumonía puede ser una complicación de la influenza "común", causada por los virus de la influenza, que con poca frecuencia causan complicaciones importantes, pero son agentes biológicos de la misma que se vuelve persistente. Con 31,081 muertes en 2019, la neumonía y la influenza son la octava causa de muerte en el país y la sexta enfermedad (Instituto Nacional de Estadística y Geografía, 2020).

La OMS (2021), estima el costo del tratamiento de la neumonía en $109 millones de dólares por año, incluido el tratamiento médico y las pruebas necesarias para diagnosticar la neumonía. Por otro lado, la Organización Panamericana de la Salud (2021) reporta que la influenza puede causar hospitalización o muerte en grupos de alto riesgo, calculando que los brotes anuales provocan de 3 a 5 millones de casos graves y de 290.000 a 650.000 muertes.

1.5.7 Enfermedad Pulmonar Obstructiva Crónica (EPOC)

Otra enfermedad que involucra el sistema respiratorio y causa el deceso de cerca de 24,000 personas es la EPOC, cifra que la convierte en la novena causa de muerte en México , y la tercera causa de muerte a nivel mundial, pues al 2019 causó 3.23 millones de muertes, 80% de ellas ocurridas en países de bajos y medianos ingresos (Organización Mundial de la Salud, 2022).

En México, el Instituto Nacional de Enfermedades Respiratorias (2017) ubica a la EPOC entre el sexto y cuarto lugar en las tablas anuales de morbilidad y mortalidad, sin mayor prevalencia en algún género. Los factores de riesgo para padecer EPOC comúnmente son el tabaquismo o la exposición al humo de la leña, aunque también puede ser causante los gases, humos y sustancias químicas inhalados en ambientes laborales.

1.5.8 Insuficiencia Renal

La insuficiencia renal es definida por la American Kidney Fund (2022), como una falla en los riñones en la que estos órganos comienzan a dejar de funcionar hasta el punto en que la persona no podría sobrevivir sin diálisis o un trasplante de riñón.

El Instituto Nacional de Salud Pública (2020), reportó 51.4 muertes por cada 100,000 habitantes en México, lo que también tuvo un impacto significativo en las finanzas institucionales y la economía doméstica. En 2014, el gasto promedio anual en salud por persona se estimó en $8,966 dólares para la Secretaría de Salud, mientras que para el IMSS el gasto promedio fue $9,091 dólares.

Como se mencionó anteriormente, la insuficiencia renal es irreversible, sin embargo, un tratamiento común es la diálisis, hemodiálisis y eventualmente algún trasplante renal, se estima que “los costos anuales promedio de un paciente con insuficiencia renal crónica tratado con hemodiálisis oscilan entre $223,183 y $257,000; el costo de por vida es de $1,198,968. El costo total para el establecimiento está entre el 1.7% y el 1.73% del presupuesto” (Villareal-Rios E., 2020).

Capitulo 2 El Análisis Multivariado

2.1 Introducción del Capítulo

El Análisis Multivariante es una técnica estadística utilizada para el estudio simultaneo de múltiples variables y la relación entre ellas. Estas técnicas se han desarrollado desde hace algún tiempo, pero su mayor auge ha sido en los últimos años al resultar de gran utilidad en diversos campos de estudio, como el *"Machine Learning",* en el que se utiliza ampliamente. En este capítulo se estudia el surgimiento del Análisis Multivariado, su definición, los objetivos, las técnicas, además de explorar a detalle la técnica especifica del Análisis Clúster.

El Análisis Multivariante surge en el siglo XIX con el desarrollo de la probabilidad y estadística, época donde la medición de relación entre dos variables estuvo "de moda", pero donde la correlación no se empleó en analizar más de dos variables simultáneamente. Entre las primeras técnicas desarrolladas están: análisis de varianza y regresión múltiple; no obstante, la complejidad de los datos que no entraban es estas técnicas innovadoras apoyó el desarrollo de nuevas técnicas más sofisticadas como: análisis factorial, discriminante y clúster; nuevas formas de estudiar y entender datos de un mundo cada vez más complejo.

La importancia del Análisis Multivariado radica en que, al ser una técnica estadística de exploración de datos, esta puede emplearse en diversas áreas de investigación; entre ellas, la psicología, economía, sociología, medicina y biología, en las que puede ayudar a identificar patrones de datos, explorar relaciones entre variables, disminuir la dimensionalidad de los datos e incluso apoyar a predecir valores (*forecast*).

Una vez explicado el origen e importancia, es relevante explicar más detalladamente el análisis clúster, definición, objetivos y métodos que pueden emplearse, con la finalidad de que el lector conozca sus distintos usos y aplicaciones, cómo debe de ser el manejo de los datos, qué técnicas existen para hacer particiones, si hacerlo a través de la metodología de jerarquización o partición y para qué sirven cada una de ellas; asimismo, se detallan las distintas formas de medir la distancia, y algunos métodos de clusterización como lo son el k-Means o el k-Modes; esto con el fin de que en capítulos posteriores se pueda hablar con soltura sobre la ejecución del análisis clúster.

2.2 Surgimiento del Análisis Multivariado

A partir de finales del siglo XIX y principios del siglo XX, ramas sociales y médicas han buscado técnicas que ayuden a comprender el comportamiento de los seres humanos. Esta búsqueda se aceleró con el avance de las computadoras y la llegada de la época informática, puesto que con la ayuda de ordenadores resultó mucho más sencillo analizar cantidades muy grandes de datos y disminuir la complejidad de sus análisis. De esta forma, al buscar técnicas para el procesamiento de datos, se creó el análisis multivariado.

En los comienzos del siglo XX, el psicólogo Charles Spearman y el matemático Karl Pearson comienzan a introducir los conceptos fundamentales de la estadística moderna. Finalmente, Hotelling, Wilks, Fisher, Mahalanobis y Bartlett terminaron sentando las bases definitivas (Bramardi, 2002).

Algunas técnicas del análisis multivariado nacen de la sociología o de la psicología, y no necesariamente de las matemáticas. El ímpetu del ser humano por querer simplificar procesos que son difíciles de ver desde una óptica matemática o cuantitativa a simple vista ayudó a la creación de diversas técnicas y procedimientos que ayudaron a simplificar la relación entre la nueva cantidad de información y la finalidad de ser explicada de mejor manera.

La Consultoría Estratégica de Investigación de Mercados (CIMEC) cataloga como fundamental al análisis multivariado para el procesamiento de datos en los estudios de mercados, pues busca determinar qué tipo de personas compra cierto(s) producto(s) (2020), pero incluso pueden verse utilizadas algunas técnicas para la agricultura; calculando la resistencia de ciertas cosechas a daños por plagas o sequias, o en la psicología, explicando la relación entre dos individuos de acuerdo a sus actitudes (Nieto Barajas, 2020).

Ahora, dado que en el siglo XXI las economías digitales constantemente generan grandes cantidades de datos a partir de nuevas estructuras, la industria se ha tenido que modernizar, y ha creado infraestructuras complejas que utilizan los datos recolectados para obtener información del comportamiento de los usuarios y hacia donde se inclinarían los mismos. Para el procesamiento de estos datos es importante el uso técnicas del análisis multivariado.

El análisis multivariado surge en el momento en que, a un mismo objeto, ya sea un individuo (personas o animales), o un concepto (amor o amistad), se le miden distintos atributos o características de interés para el investigador; dichos objetos debes de ser independientes entre sí. A cada característica o atributo se le llamara variable, las cuales pueden ser numéricas: continuas o discretas, o categóricas: ordenadas o no ordenadas.

2.3 Definición del Análisis Multivariado

El análisis multivariado es definido de diversas formas, aunque todas tienen en común que es un conjunto de técnicas que ayuda al investigador a analizar o examinar un conjunto de variables. En ocasiones, este tipo de análisis puede señalarse como el caso general del análisis bivariante y univariante. Del mismo modo, se puede definir como el conjunto de métodos estadísticos cuya finalidad es analizar simultáneamente grupos de datos que incluyes diversas variables, que describen al objeto de estudio, cuya principal finalidad es un entendimiento mayor de los fenómenos analizados (Humberto, 2013).

Esta herramienta se apoya de diversos métodos estadísticos enfocados en observar y procesar de manera simultánea diferentes variables con la finalidad de obtener nueva información. Dado que, a diferencias del análisis univariante en donde se analiza la media y varianza de una variable, o en el caso del análisis bivariado donde se describe la relación de dos variables mediante la correlación, en el análisis multivariado se trata de analizar covarianzas o correlaciones que reflejen relaciones entre más de 3 variables, estas deben ser aleatorias y tener una relación, pero que de manera individual no se puedan interpretar (Lozares Colina & López Roldán, 1991).

De este modo el Análisis Multivariado toma mayor relevancia cuando somos conscientes de que en la vida diaria y en las situaciones reales la mayoría de los estudios involucran muchas variables que interactúan de manera simultánea.

2.4 Objetivos del Análisis Multivariante

Los principales objetivos de este tipo de análisis son:

- Resumir los datos mediante un pequeño conjunto de nuevas variables.
- Encontrar y clasificar en grupos los datos.
- Proporcionar métodos para el estudio de datos que no pueden ser procesados a través de un análisis uni o bivariante.
- Auxiliar al investigador en la toma de decisiones.

2.5 Clasificación de las Técnicas del Análisis Multivariado

Como se mencionó anteriormente, el Análisis Multivariado se apoya de distintas técnicas estadísticas que se clasifican de acuerdo con su objetivo o relación de variables y, posteriormente, por el tipo de variable que utilicen. De este modo, Humberto (2013) clasifica las técnicas en dos categorías: Explicativas o de Dependencia y Descriptivas o de Interdependencia.

Como ha dicho Lozares Colina y López Roldán (1991), el analisis multivariado es la generalizacion del analisis uni y bivariado, en donde algunas técnicas pueden provenir de estos e, incluso, podrian usarse indistintamente, con ciertas variaciones.[4]

2.5.1 Técnicas Explicativas o de Dependencia

Utilizados cuando es posible diferenciar las variables entre dependientes e independientes, usualmente son útiles para investigaciones experimentales. Estas técnicas consideran a las variables independientes como potencial causa de las variables dependientes, investigando la existencia o ausencia de relación entre dependientes e independientes.

[4] Como lo seria dividir una MANOVA en tantas ANOVAS como sea necesario.

2.5.1.1 Análisis de Regresión Múltiple

Son las técnicas más utilizadas, estudian la dependencia de una variable en función de otras. Es utilizado cuando se tiene la certeza de que el valor de una variable cambia de acuerdo con el comportamiento de otras variables; su aplicación permite observar de qué manera las variables explicativas[5] pueden predecir el comportamiento de la variable explicada[6]. Los coeficientes estimados de la regresión son parámetros estadísticos descriptivos que ayudan a obtener el significado de la(s) variable(s) independiente(s) para los datos de la variable dependientes.

El modelo simplificado de los análisis de regresión múltiple es:

$$Y_1 \leftarrow (X_1, X_2, X_3, \dots, X_n)$$

Donde Y es una variable métrica y X puede ser tanto métrica como no métrica.

Sagaró del Campo y Zamora Matamoros (2020) enuncian entre los métodos más comunes de regresión los siguientes:

- **Regresión Lineal:** Utilizada cuando la variable dependiente es continua y cada coeficiente representa un cambio en la variable independiente que afectará de igual manera a la variable dependiente (lineal)**.**
- **Regresión Binomial:** Es de fácil identificación puesto que la variable respuesta es dicotómica (solo puede tomar una de dos posibles opciones).
- **Regresión Logística:** Ocurre cuando la variable dependiente es categórica (la variable puede tomar el valor de un número limitado de categorías).
- **Regresión de Poisson:** Como su nombre lo indica, la condición principal es que la variable dependiente se ajuste bien a una distribución Poisson.
- **Regresión de Cox:** También llamado Modelo de riesgos proporcionales, es comúnmente empleada para los estudios de supervivencia, puesto que la variable explicada está en función del tiempo hasta un determinado evento.

[5] Variable Explicativa y Variable Independiente son sinónimos en Estadística.
[6] Variable Explicada; Variable Dependiente y Variable Respuesta son sinónimos Estadística.

- **Regresión Log-Lineal:** Son modelos que, a diferencias de las Regresiones Lineales, permite un análisis óptimo de las variables cualitativas, categóricas, dicotómicas o politómicas. El principal objetivo es definir un método que estudie las relaciones entre las variables cualitativas o no numéricas.

2.5.1.2 Análisis Discriminante

Esta técnica trata de encontrar una función lineal de varias variables que permita clasificar individuos que pertenecen a un mismo grupo. El objetivo es encontrar una función que ayude a clasificar a los nuevos individuos a partir de los valores discriminantes previamente establecidos. Las diferencias entre grupos se pueden describir con características similares o iguales.

En el método discriminante se debe de conocer previamente a qué grupos pertenecen ciertos individuos; los grupos ya son conocidos y lo que se busca es saber en qué forma las variables nos ayudan a discriminar los individuos creando dichos grupos (Martinez de Lejarza & Martinez de Lejarza, 1995).

2.5.1.3 Correlación Canónica

Para esta técnica se toma un grupo de variables y se trata de predecir los valores en función de otro grupo de variables. Dicho de otra manera, se busca la existencia de una relación entre variables independientes y un conjunto de variables dependientes. El objetivo radica en calcular la combinación lineal de cada conjunto de variables que maximice la correlación entre dos grupos de variables.

Puede tomarse como una extensión de las regresiones, dado que el procedimiento implica obtener ponderaciones para el conjunto de variables dependientes e independientes que proporcione la máxima correlación entre los conjuntos.

El modelo simplificado de los análisis de regresión múltiple es:

$$(Y_1, Y_2, Y_3, \dots, Y_N) \leftarrow (X_1, X_2, X_3, \dots, X_m)$$

Donde Y puede ser variable métrica o no métrica y X puede ser tanto métrica como no métrica (Humberto, 2013).

2.5.1.4 Análisis de Varianzas y Covarianzas

En este estudio se busca que la muestra total sea dividida en varios grupos basados en una o varias variables independientes, cuyo objetivo es encontrar si hay diferencias significativas entre estos grupos en las variables dependientes.

Aunque hay autores como Hair Jr (1995, pág. 13) que solo contemplan a la MANOVA y MANCOVA, otros autores como Lozares Colina y López Roldán (1991) tambien consideran a la ANOVA y ANCOVA; posteriormente se definirán dichas técnicas, puesto que en estas ultimas se basan la MANOVA y MANCOVA:

ANOVA: Proviene del ingles ***An****alysis* ***of*** ***Va****riance* y es una prueba donde existen 3 o más grupos pertenecientes a la variable de respuesta continua. El factor es la variable que distingue la pertenencia de este grupo: siempre que haya un factor se llamara ANOVA unidireccional (más recurrente), y cuando existan 2 factores se le llamara ANOVA bidireccional.

ANCOVA: A diferencias de la ANOVA, la ANCOVA contempla la covarianza y posee solo una variable de respuesta continua que se basa en los principios de ANOVA y regresión (M.H. & Castillo J. & Wong, 2008).

MANOVA: Del inglés Multivariate Analysis of Variance, es la extensión de la ANOVA y es utilizada de manera simultánea para explorar las relaciones entre categorías de variables independientes (tratamientos), y dos o más variables métricas dependientes. Puede ser unidireccional (2 o más variables dependientes y con 1 sólo factor) o bidireccional (2 o más variables dependientes y 2 o más factores), (Humberto, 2013).

MANCOVA: Es la abreviatura del analisis multivariado de covarianza, puede verse como la combinacion de una ANCOVA y MANOVA, pues cuenta con 2 o más variables de respuesta, pero sus variables independientes no son factores ya que agregan 1 o mas covariables.

2.5.1.5 Ecuaciones Estructurales

Son útiles para flexibilizar los modelos de regresión, pues son menos estrictos, dado que incluyen los errores de medida en las variables dependientes e independientes; a su vez, aunque suelen ser más complejos de estimar, ayudan a saber si existe relación entre 2 conjuntos de variables (variables observadas, medidas de manera directa y las variables latentes medidas a través de otras variables observadas) (Sagaró del Campo & Zamora Matamoros, 2020).

La estructura general según Humberto, (2013), es:

$$Y_1 \leftarrow (X_{11}, X_{12}, X_{13}, \dots, X_{1m})$$

$$Y_2 \leftarrow (X_{21}, X2_2, X_{23}, \dots, X_{2m})$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$Y_n \leftarrow (X_{n1}, X_{n2}, X_{n3}, \dots, X_{nm})$$

Donde Y es una variable métrica y X puede ser tanto métrica como no métrica.

2.5.2 Técnicas Descriptivas o de Interdependencia

Estos métodos no distinguen entre variables dependientes e independientes, su objetivo es identificar qué variables están relacionadas, por qué y cómo es dicha relación; normalmente se aplica para contextos no experimentales; su principal objetivo es el resumir los datos para un manejo más sencillo y que se adecue al objetivo de la investigación.

2.5.2.1 Análisis de Componentes Principales.

Es de las principales técnicas de reducción de datos, cuyo principal objetivo es construir una combinación lineal de las variables originales que muestran la mayor parte de información. Cada combinación lineal se construye de la forma en que las variables están intercorrelacionadas, para descartar variables que cada vez aportan menos información. Dicho de otra forma, cada combinación lineal aporta menos y menos información, descartándola para utilizar solo las que puedan describir los datos de mejor manera. Esto ayuda a disminuir la dimensionalidad de los datos, dado que quita atributos irrelevantes a la hora de tomar decisiones. De esta forma, siempre existirán menos componentes principales que variables estudiadas (Moreno Madueño, 2016), que puede expresarse de la siguiente manera: “Si existieran **n** variables, existirán **p** componentes principales que cumpla **p<n**”.

2.5.2.2 Análisis Factorial

Aunque a veces puede resultar parecido al método anterior, el análisis factorial no solo busca reducir el número de variables, sino explicar, en términos de factores ocultos, la variable original. El objetivo es establecer las variables latentes (factores), que podrían causar correlación entre las variables; las correlaciones entre las variables de cada grupo serán superiores a las correlaciones de las variables entre grupos.

Humberto (2013) define las diferencias entre las dos ténicas de la siguiente manera:

> "Ambas técnicas se usan para analizar interrelaciones entre un número elevado de variables métricas, explicando dichas interrelaciones en términos de un número menor de variables denominadas factores si son inobservables o componentes principales si son observables."

2.5.2.3 Análisis de Correspondencias

Aunque parecido al análisis factorial, este método utiliza solo variables categóricas para representar en un espacio multidimensional la relación entre las categorías de dos variables no métricas, reduciéndolas. El mapa que se crea obtiene las distancias entre los diferentes niveles de las dos variables, por lo que el análisis de correspondencias ayuda a visualizar las tablas de contingencia y realiza una transformación de los datos no métricos, para obtener datos métricos

2.5.2.4 Escalamiento Dimensional

También es una técnica de reducción de datos, cuyo objetivo principal es representar objetos en un espacio dimensional reducido, de forma que la distorsión causada por la reducción afecte a los datos lo menos posible; dicho de otro modo, trata de que las distancias entre los objetivos representados en el espacio dimensional nuevo sean parecidas a la distancia en el espacio dimensional original.

2.5.2.5 Análisis de Conglomerados o Clúster

Es una técnica que trata de identificar grupos entre las observaciones de acuerdo con los valores de las variables; las observaciones se asignan gráficamente a grupos individuales. Está diseñada para clasificar las observaciones en grupos. La diferencia con el análisis factorial es que mientras el análisis clúster agrupa los objetos, aquel las variables; por otra parte, con respecto al análisis discriminante es que no se conoce el número de grupos ni la composición de estos.

En el análisis clúster se busca que los grupos sean heterogéneos e internamente sean homogéneos, apoyándose de diversos índices de similitud (distancia). Los clústers pueden ser jerárquicos (ascendentes o descendentes), cuando no existe un numero de grupos a priori, o no jerárquicos, si se parte de un número fijo de clústers, ya sea por hipótesis o por cálculos previos.

2.6 El Análisis Clúster

La importancia de la clasificación de datos siempre ha sido una gran herramienta para el análisis de datos, pues ha apoyado el desarrollo de las ciencias al ser una operación básica y sencilla para delimitar y ordenar objetos. La clasificación nace de la necesidad de entender, identificar y, además, simplificar el manejo de la complejidad del mundo, pues busca patrones con la finalidad de entenderlo de mejor manera. Algunos procesos de estudios estadísticos ocupan una etapa de recolección y clasificación de datos para poder tener un buen proceso de análisis de datos.

2.6.1 El Análisis Clúster

Técnica ocupada para desarrollar subgrupos significativos de individuos u objetos; el objetivo de clasificar una muestra de entidades en un número pequeño de grupos mutuamente excluyentes basados en similitudes entre dichos objetos.

Hair Jr (1995) divide en 2 etapas la realización de un análisis clúster:

1. Buscar una medida, forma de similitud o asociación entre las entidades para determinar cuántos grupos existen en realidad en la muestra.
2. Describir las personas o variables para determinar su composición.

Los agrupamientos por clústers están basados en la idea de distancias o similitudes ente observaciones; mientras que la obtención de los clústers, el número de clústers dependerá de lo que se considere similar. Normalmente esta técnica es utilizada para grandes cantidades de datos, después de ser agrupados en sus respectivos clústers, se ocupan para la comparación de características específicas y descripción de datos.

En el análisis clúster al encontrar los grupos homogéneos, el investigador podría lograr algunos objetivos, entre los que se encuentran (Prieto Guerra, 2006):

- **Descripción de una taxonomía:** permite obtener una clasificación de los objetos o individuos (recurrentemente individuos), la cual podría ser comparada con una tipología (una clasificación teórica) propuesta.
- **Simplificación de datos:** las observaciones ya agrupadas pueden apoyar para estudios posteriores, puesto que todos los miembros estarán en un conglomerado de características generales similares.
- **Identificación de relaciones:** debido a que todos los datos están definidos y estructurados en grupos, el investigador podrá identificar relaciones entre los individuos que a simple vista sería muy difícil detectar en las observaciones individuales.

Como se mencionó, el número de clúster puede variar de acuerdo con lo que el investigador considere. De la Fuente Fernández (2011) enuncia que las soluciones no son únicas, y en medida en que la pertenencia al conglomerado se dé, cualquier numero de soluciones se vera completamente influenciado por los elementos y procedimientos que sean elegidos por el investigador.

2.6.2 Previo al Análisis Cluster

Si bien en la mayoría de los estudios y análisis el investigador define su(s) objetivo(s) e igualmente selecciona las variables de estudio, en el caso del análisis clúster se consideran 3 etapas antes de la iniciar el proceso de partición; estas etapas son:

1) Detección de valores atípicos.

2) Selección y transformación de variables a utilizar (estandarización).

3) Selección del concepto de distancia o similitud, asimismo la medición de estas.

2.6.2.1 Detección de Valores Atípicos

Los valores atípicos son valores muy diferentes a las observaciones del mismo grupo de datos; normalmente pueden ser ocasionados por: a) errores de procedimiento, b) acontecimientos extraordinarios, c) valores extremos, d) causas desconocidas. Los valores atípicos pueden provocar una mala representación del conjunto de datos.

Existen diversas maneras de detectar datos atípicos, una de las principales es examinar la distribución de observaciones para cada variable, y los valores atípicos serán aquellos que queden fuera de los rangos de la distribución. El investigador designará un umbral para la designación de un caso atípico. Otros métodos comunes son los métodos gráficos, como histogramas o diagramas como el boxplot, o por pruebas numéricas como la prueba Tuckey. (Ocaña Peinado, 2020)

2.6.2.2 Estandarización de Datos

En el método clúster existe el problema al medir las distancias en el uso de datos, ya que sin una estandarización existirá un gran número de inconsistencias, específicamente cuando la escala de las variables cambia.

Galbiati R. (2006), enuncia las siguiente formula para la estandarizacion de datos:

$$z_{ic} = \frac{x_{ic} - m_c}{s_c}$$

Donde m_c es la media y s_c es la desviación estándar, denotadas:

$$m_c = \frac{\sum_{i=1}^{n} x_{ic}}{n} \quad ; \quad s_c = \sqrt{\frac{\sum_{i=1}^{n} (x_{ic} - m_c)}{n}}$$

2.6.2.3 Medidas de Asociación (Distancia y Similaridad)

Son técnicas numéricas que ayudan a unir objetos o individuos que caractericen las relaciones entre las variables o los individuos. Cada medida de asociación es particular y el investigador deberá de escoger la que considere mejor para el análisis de sus datos (De la Fuente Fernández, 2011).

Por otra parte, la medida de asociación puede ser de distancia o de similaridad:

- **Método de distancia**: los grupos están formados de manera que la distancia entre individuos sea la más pequeña posible. Teniendo la siguiente clasificación:
  - × Distancia Euclídea

$$\sqrt{\sum_{j=1}^{p} (X_{rj} - X_{sj})^2}$$

- Distancia Euclídea Cuadrada

$$\sum_{j=1}^{p}(X_{rj} - X_{sj})^2$$

- Distancia Minkowski

$$\sqrt[q]{\sum_{j=1}^{p}(X_{rj} - X_{sj})^q}$$

La fórmula Minkowski es la generalización de la Distancia Euclídea y Distancia Euclídea Cuadrada, solo que con q=2 y q=1 respectivamente.

- Distancia Manhattan d_1 o Ciudad (City Block)

$$\sum_{j=1}^{p}|X_{rj} - X_{sj}|$$

- Distancia de Tchebychev o del Máximo

$$max_{\ j}|X_{rj} - X_{sj}|$$

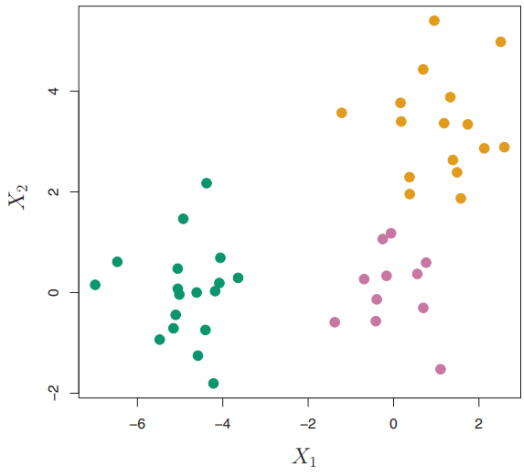
- **Método de similaridad**: los grupos tendrán elementos que posean semejanza entre ellos. Clasificados en dos grupos:
  - Coseno del Ángulo de dos Vectores
  - Coeficientes de Correlación (Pearson, Kendall, Spearman).

2.6.3 Técnicas Clúster

Existen dos tipos de formacion de conglomerados (De la Fuente Fernández, 2011):

- **Algoritmos de Partición o No Jerárquicos:** Utilizado para dividir el conjunto de observaciones en *k* conglomerados, donde *k* es un número definido previamente por el investigador. Empieza con una partición de objetos en *k* grupos y posteriormente intercambia objetos entre los clústers de forma que, al finalizar, se tenga el mejor acomodo según los requerimientos y condiciones del investigador.

Gráfico 2.1 Ejemplo del Algoritmo de Partición como Técnica Clúster.

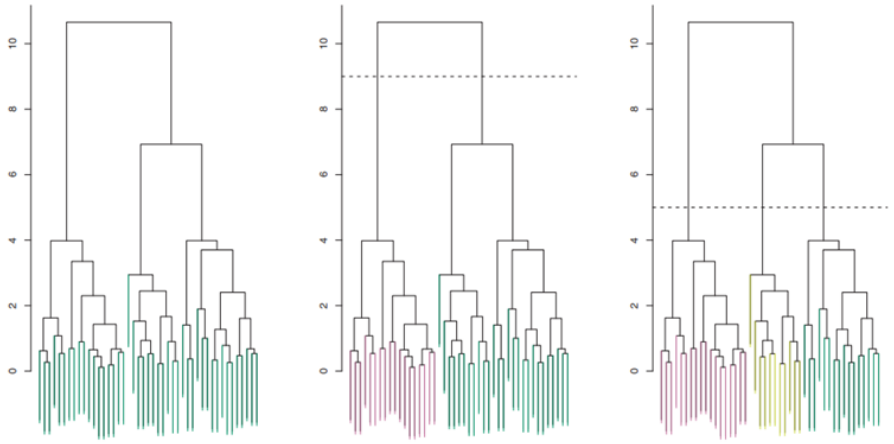


Fuente: An Introduction to Statistical Learning. (James, Gareth, et al., 2013)

En el grafico 2.1 se ejemplifican 45 puntos que después de ser sometidos a un proceso de Clústering Particional se obtienen 3 clústers señalados con diferentes colores.

- **Algoritmos Jerárquicos:** Estos métodos entregan una jerarquía de divisiones en conjuntos de elementos en conglomerados. Estos algoritmos sirven para formar un nuevo clúster o separarlo, de forma que ese nuevo(s) clúster(s) maximice su similaridad o minimice su distancia, de acuerdo con lo que el investigador desee. A los diagramas resultantes del clústering jerárquico se le conoce como dendograma y, de acuerdo con el corte en dicho gráfico, se obtendrá en número de clústers para el grupo de datos.

Gráfico 2.2 Ejemplo del Algoritmo Jerárquico como Técnica Clúster.



Fuente : James, Gareth, et al., 2013.

En el *gráfico 2.2* se ejemplifican 45 puntos que después de ser sometidos a un proceso de Clústering Jerárquico de izquierda a derecha, se tiene un clúster, dos y tres clústeres respectivamente.

A su vez los metodos jerarquicos se subdividen en dos, que son:

× **Métodos jerárquicos aglomerativos o asociativos:** se parte de que cada observación forma un conglomerado y a lo largo de los pasos se van uniendo las observaciones, hasta que, finalmente, todas están en un conglomerado único. Es el método más utilizado, empieza con un clúster por objeto mientras va uniendo distintos elementos de características similares, distancia más pequeñas).

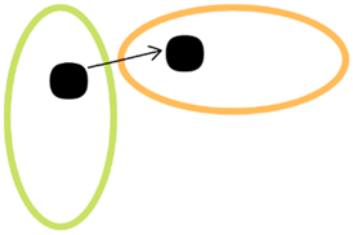
- **Métodos jerárquicos disociativos:** de manera contraria al método anterior, parte de un gran conglomerado y en los subsecuentes pasos se irá dividiendo hasta que cada observación quede en un conglomerado diferente. Normalmente es un proceso más complejo, ya que va dividiendo elementos que no se consideran tan similares al resto, según se encuentren más alejados.

Una vez realizado el clústering es necesario poder medir la distancia entre clústers, siguiendo así con el principio básico del Análisis Clúster; ***"maximizar la distancia interclúster y a su vez minimizar la distancia intraclúster"***.

Entre todos los criterios comunes para medir la distancia entre los congomerados A y B se encuentran (Galbiati R., 2006):

Gráfico 2.3 Distancia entre Conglomerados A y B "Vecino más cercano"

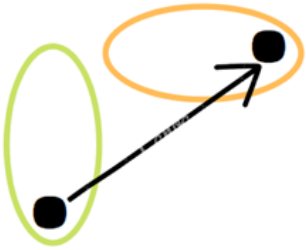
$$d(A,B) = min_{i\in A\ y\ j\in B}\ d(i,j)$$



Fuente: Elaboración propia con base en Galbiati R., 2006.

Gráfico 2.4 Distancia entre Conglomerados A y B "Vecino más lejano"

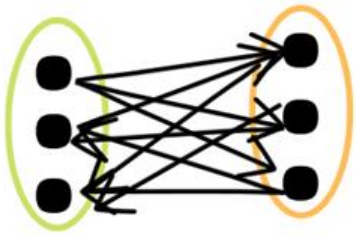
$$d(A,B) = máx_{i\in A\ y\ j\in B}\ d(i,j)$$



Fuente: Elaboración propia con base en Galbiati R., 2006.

Gráfico 2.5 Distancia entre Conglomerados A y B “Promedio de grupos”

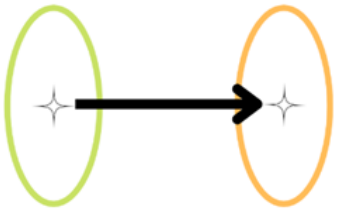
$$d(A,B) = \frac{1}{n_A n_b} \sum_{i \in A \ y \ j \in B} d(i,j)$$



Fuente: Elaboración propia con base en Galbiati R., 2006.

Gráfico 2.6 Distancia entre Conglomerados A y B “centroide”

$$d(A,B) = d(\bar{x}_A, \bar{x}_B)$$



Donde $\bar{x}_A, \bar{x}_B$ son los centroides de los clústers A y B respectivamente.

Fuente: Elaboración propia con base en Galbiati R., 2006.

Capitulo 3 Metodología del Análisis Clúster.

3.1 Introducción del Capítulo

El análisis de clúster o análisis por conglomerados es un método de análisis estadístico que permite agrupar observaciones o individuos que posean características o atributos similares en grupos, conglomerados o clústers. Este método es ampliamente utilizado en diversas disciplinas como la biología, la psicología, la economía y el *marketing*. El análisis de conglomerados es una herramienta poderosa para identificar patrones en grandes conjuntos de datos y puede utilizar con fines descriptivos y predictivos.

Al principio de este capítulo se explica el planteamiento y objetivo del análisis clúster, para posteriormente explicar los pasos necesarios para un análisis de conglomerados del estudio. Desde la primera parte se explicarán los pasos a realizar previo a la ejecución del modelo; es decir, la selección de base a trabajar, filtrado de datos, diagnósticos de interés y, finalmente, la selección de las variables; esto último para evitar problemas al momento de las iteraciones para la clusterización.

La segunda parte consiste en cómo realizar el análisis por conglomerados; específicamente la técnica del presente proyecto, la distancia empleada (distancia euclídea), criterios para la selección de número óptimo de clústers (codo, rodilla o silueta), método y algoritmo de clusterización (k-Means). Se emplean apoyos visuales (imágenes, diagramas y tablas) para poder discernir entre clústers.

Además, se mostrarán las herramientas que se ocuparon para el estudio presentado, la selección de variables, el filtrado de los datos, qué sub-diagnósticos se utilizaron para complementar el diagnóstico de las enfermedades; esto es, aunque una de las enfermedades seleccionadas fue los tumores malignos, existe una gran gama de tipos de tumores. Finalmente, se explica qué tipo de distancia se utiliza, el método de clusterización y el algoritmo para la ejecución del análisis.

En conclusión, el capítulo proporciona la metodología completa y detallada sobre técnicas el estudio, dando un acercamiento a la aplicación de la base seleccionada; mientras que en el siguiente capítulo se explicará la ejecución del este algoritmo por medio de softwares y se presentarán los resultados obtenidos.

3.2 Planteamiento y Objetivo

Se utilizará un análisis clúster como parte de un análisis exploratorio de los datos de la Seguros de Gastos Médicos Mayores para el periodo 2019-2020, cuya finalidad será encontrar agrupaciones para las enfermedades más relevantes de México en 2019, y aún más, poder identificar si la pandemia de COVID-19 causó un impacto de manera indirecta a la los SGMM para el año 2020, puesto que entre las enfermedades seleccionadas no se encuentra la enfermedad causada por SARS-CoV-2.

De la Fuente Fernández (2011) enuncia como punto de partida de un una muestra **X** con **m** individuos de estudio que cuentan con **p** variables, las cuales pueden ser ordenados en una matriz numerica de la siguiente forma:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{12} & \cdots & x_{1p} \\ \vdots & & & \ddots & \vdots \\ x_{m1} & & x_{m2} & \cdots & x_{mp} \end{pmatrix}$$

Donde:

x_{11}: Valor que presente el primer individuo en la primera variable.

x_{1p}: Valor que presente el primer individuo en la última variable (p).

x_{m1}: Valor que presente el último individuo (m) en la primera variable.

x_{mp}: Valor que presente el último individuo (m) en la última variable (p).

3.3 Base de Datos

3.3.1 Elección de Base

Como se describió en el capítulo anterior, existe una serie de procesos a realizar antes de la ejecución del análisis clúster. Para este trabajo se utilizará la información estadística de Salud proporcionada por la CNSF, la cual cuenta con información del sistema Estadístico del Seguro de Salud, con fundamento en la Disposición 38.1.9, numeral V de la Circular Única de Seguros y Fianzas.

La CNSF (2021) muestra los siguientes apartados de información estadística: Índice, Emisión, Siniestros No Hospitalarios y Siniestros Hospitalarios, siendo este apartado la sección de interés para nuestro estudio, dado que estos normalmente tienen un costo elevado siendo los de mayor interés para la aseguradora. Dicho conjunto de datos posee atributos relevantes y específicos para el tema tratado: *Edad* (0-99 años)*, Género, Entidad de Residencia, Actividad Económica, Subtipo de Seguro, Diagnóstico, Tipo de Evento Hospitalario* (Internamiento o Ambulatorio)*, Procedencia de Ingreso* (Consulta Externa, Urgencia, Referencia de Otra Unidad), *Motivo de Egreso* (Curación, Mejoría)[7]*, Número De Reclamaciones, Monto De Hospitalización, Monto Honorarios Médicos, Monto De Deducible o Copago, Monto Coaseguro*.

3.3.2 Filtrado de Datos

Al año, según el INEGI (2021), se registraron alrededor de un millón de muertes, de las cuales, el 90% son causadas por alguna enfermedad. Si tenemos esto en cuenta, lo importante es saber qué enfermedades son, y el costo en el que incurren los pacientes de estas en su tratamiento; es decir, lo importante que resulta el SGMM para personas que sufren estas enfermedades, puesto que la adecuada atención podría ayudar a una mejor calidad de vida e, incluso, a prolongar la esperanza de vida de las personas con dicho padecimiento.

Por obvias razones, se dejarán fuera del estudio todos aquellos datos que las variables tengan como resultado “No Disponible”, con la finalidad de que los datos puedan tener la mayor cantidad de información para describir los clústers.

[7] Para el año 2020 se agregaron las opciones de *otra* y *defunción.*

Del mismo modo, el INEGI (2020), con una tendencia retrospectiva de 5 años, explica qué enfermedades son causantes del mayor número de decesos, la cantidad y el género, en el año 2019. De este modo es de vital importancia encontrar patrones en los datos para los SGMM.

En la Tabla 3.1 se muestran las enfermedades causantes del mayor número de decesos para México por sexo; se omiten los accidentes y actos violentos.

Tabla 3.1 Principales Causas de Muerte en México

Total	**Hombres**	**Mujeres**
Enfermedades del Corazón 156 041	Enfermedades del Corazón 83 258	Enfermedades del Corazón 72 768
Diabetes Mellitus 104354	Diabetes Mellitus 51 711	Diabetes Mellitus 52 643
Tumores Malignos 88 680	Tumores Malignos 43 296	Tumores Malignos 45384
Enfermedades del Hígado 40 578	Enfermedades del Hígado 20 602	Enfermedades Cerebrovasculares 17 659
Enfermedades Cerebrovasculares 35 303	Enfermedades Cerebrovasculares 17 644	Influenza y Neumonía 14 046
Influenza y Neumonía 31 081	Influenza y Neumonía 17 034	Enfermedad Pulmonar Obstructiva Crónica 11 269
Enfermedad Pulmonar Obstructiva Crónica 23 768	Enfermedad Pulmonar Obstructiva Crónica 12 499	Enfermedades del Hígado 10 879
Insuficiencia Renal 14630	Insuficiencia Renal 8 271	Insuficiencia Renal 6 359

Fuente: Elaboración propia con información de Instituto Nacional de Estadística y Geografía, 2020.

Bajo esta referencia, después de seleccionar la base de datos proporcionada por la CNSF, a la par que se tiene como referente al INEGI para conocer padecimientos comunes en consecuencia de mayor mortalidad para los mexicanos, los datos se sometieron a un primer filtrado. Dado que se cuentan con más de 100 distintos diagnósticos y no todos son de interés para la línea de investigación de este documento, las enfermedades elegidas son aquellas que el investigador, y a criterios médicos previamente investigados, considera directamente relacionados a alguna enfermedad enunciada por el INEGI.

3.3.3 Diagnósticos Seleccionados

Entre los diagnósticos seleccionados, se encuentras otras enfermedades que se están asociadas a la enfermedad principal adelante se enuncian

3.3.3.1 Enfermedades del Corazón:

- Enfermedad aterosclerótica del corazón
- Enfermedad isquémica crónica del corazón (especificada y no especificada)
- Lesiones contiguas del corazón, mediastino y la pleura
- Otras enfermedades isquémicas agudas del corazón

3.3.3.2 Diabetes Mellitus:

Dentro de los diagnósticos de Diabetes Mellitus se encuentra una gran variedad de sub-diagnósticos entre los que destacan:

- Diabetes mellitus en el embarazo
- Diabetes mellitus insulinodependiente, con y sin complicaciones
- Diabetes mellitus no especificada
- Diabetes mellitus no insulinodependiente, con y sin complicaciones
- Otras diabetes mellitus especificadas, con y sin complicaciones

3.3.3.3 Tumores Malignos:

Al igual que la diabetes y por metaanálisis, es conocido que los tumores malignos pueden aparecer y ser diagnosticados en cualquier parte u órgano del cuerpo por esta razón la selección correspondiente es:

- Tumor maligno de algún miembro u órgano
- Tumor maligno de tejido
- Tumor maligno secundario
- Mielomas múltiples.

3.3.3.4 Enfermedades del Hígado:

- Absceso del hígado
- Degeneración grasa del hígado
- Enfermedad del hígado, especificada y no especificada
- Todos los tipos de hepatitis
- Otras enfermedades inflamatorias del hígado
- Enfermedades funcionales del hígado

3.3.3.5 Enfermedades Cerebrovasculares

- Enfermedades cerebrovasculares y sus secuelas
- Otras enfermedades cerebrovasculares y sus secuelas
- Trastornos de accidentes cerebrovasculares

3.3.3.6 Influenza y Neumonía

- Influenza con neumonías debida a virus, especificada y no especificada
- Influenza con otras manifestaciones respiratorias, con y sin virus identificado
- Influenza debido a virus, identificado y no identificado
- Influenza con otras manifestaciones
- Neumonía bacteriana, clasificada y no especificada.
- Neumonía debido a bacteria especificada

- Neumonía por otros microrganismos infecciosos y no infecciosos
- Neumonía debido a virus especificado
- Otros tipos de neumonías bacterianas y de microrganismos

3.3.3.7 Enfermedad Pulmonar Obstructiva Crónica

- Enfermedad pulmonar obstructiva crónica con exacerbación aguda
- Enfermedad pulmonar obstructiva crónica con infección aguda
- Enfermedad pulmonar obstructiva crónica

3.3.3.8 Insuficiencia Renal

- Enfermedad renal hipertensiva con insuficiencia renal
- Insuficiencia renal aguda, crónica y crónica no especificada
- Insuficiencia renal terminal

3.3.4 Elección de Variables

Al comenzar el análisis de los datos, la base elegida para 2019 cuenta con 23 749 datos; del mismo modo, para 2020, el concentrado es de 17,726 datos,; en ambos casos se consideran los 32 estados de la República Mexicana con sus respectivas características: edad, género, entidad de residencia, actividad económica, subtipo de seguro, diagnóstico, tipo de evento hospitalario, procedencia de ingreso, motivo de egreso, número de reclamaciones, monto de hospitalización, montos honorarios médicos, monto de deducible o copago, monto coaseguro. (Comisión Nacional de Seguros y Fianzas, 2022)

Las variables relevantes para el análisis principalmente fueron:

- Edad
- Género
- Entidad de Residencia
- Diagnóstico
- Tipo de Evento Hospitalario
- Procedencia de Ingreso
- Motivo de Egreso

- Número de Reclamaciones
- Monto de Hospitalización
- Monto de Honorarios Médicos
- Monto de Deducible o Copago
- Monto de Coaseguro

Descartando actividad económica y subtipo de seguro ya que la primera tenía una gran cantidad de actividades, lo cual dificultaría el manejo y la segunda porque la mayoría de los datos son otros o no disponible, lo cual no aporta más información relevante.

Aunque para el análisis clúster solo se pueden utilizar variables numéricas como:

- Edad
- Número de Reclamaciones
- Monto de Hospitalización
- Monto de Honorarios Médicos
- Monto de Deducible o Copago
- Monto de Coaseguro

Sin embargo, por cuestiones de sesgo, se ha decidido eliminar las variables monto de deducible o copago y monto de coaseguro, ya que tenían una gran cantidad de ceros, por otra parte, el número de reclamaciones no es un dato que la aseguradora tenga contemplado, pues este estadístico se crea a partir de que el siniestro ocurra; quedando como variables óptimas para el análisis a ejecutar, las siguientes:

- Edad
- Monto de Hospitalización
- Montos Honorarios Médicos

Estas variables son cruciales para los SGMM y su segmentación al momento de los diagnósticos, puesto que algunos ocurren con mayor frecuencia a ciertas edades y los montos de hospitalización y de honorarios médicos son fundamentales para el SGMM.

Para el análisis clúster es difícil manejar un gran número de variables, pues tardaría mucho en converger; por este motivo, al tener 3 variables, cuya importancia es sustantiva, se considera viable el estudio.

3.4 El Método Multivariado por Utilizar

El análisis clúster es importante para encontrar estructuras en la información y la metodología, ayuda a sistematizar y cuantificar las estructuras, está basado en el Machine Learning no supervisado (las clases no están previamente definidas) encontrando agrupaciones donde aparentemente no hay.

Como principio en el análisis clúster se tiene que la distancia es proporcional a la similaridad, en otras palabras, entre menor distancia exista entre dos elementos, más similares serán, este tipo de técnica o método descriptivo es comúnmente utilizado para preprocesar e interpretar datos, los cuales posteriormente serán sometidos a otras técnicas o análisis para el descubrimiento de nuevo conocimiento.

La metodología clúster es una técnica virtuosa, no obstante, puede tener sus inconvenientes lo cuales deberán ser tratados debidamente; uno de ellos seria encontrar variables altamente correlacionadas, ocasionando subgrupos (clúster) poco relevantes y/o carentes de información, usualmente este método puede ser influenciado en gran medida por la teoría o hipótesis que respalde al análisis realizado, dado que es fácil prestarse a que el investigador ”encuentre” conexiones o significados en fenómenos no relacionados (Schiaffino, 2018).

El manejo de *outliers* debe ser tratado con cautela, dado que en ocasiones el método puede crear un clúster con ese único dato. Por tal razón, para este trabajo es imperativo identificarlos, pero no eliminarlos, puesto que pueden ser valores muy grandes cruciales para la aseguradora.

3.4.1 Distancia Euclídea

Para el método clúster se debe de elegir una medida de similitud y como antes se ha mencionado, un elemento será tan similar a otro como la distancia que los separe. Partiendo de esta idea, para comenzar un análisis clúster se tendría que elegir que procedimiento ocupar para medir dichas distancias, en la estadística existen distintos métodos de medición, pero comúnmente se utiliza la medición de **distancia euclídea** la cual es el caso particular de la distancia Minkowski.

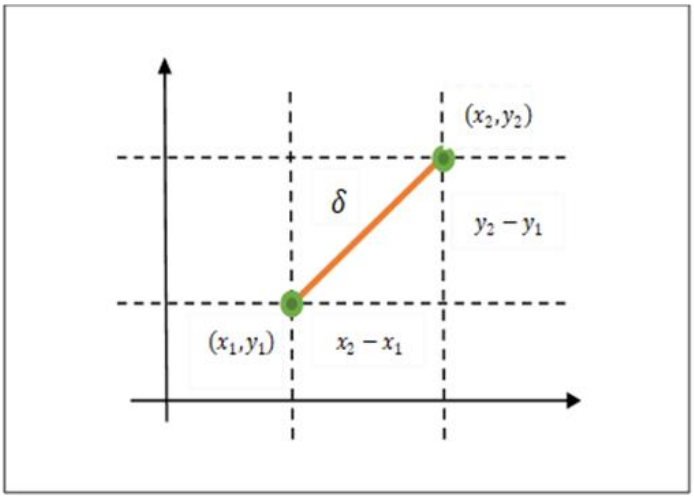
García Hernández y González Segura citan que para verificar un método de distancia se requiere que (2015):

Sea *x,y,z* elementos de una matriz de datos y la distancia sea denotada por $\delta(x, y)$

- $\delta(x, y) > 0$
  La distancia entre dos elementos siempre será real y positiva.
- $\delta(x, y) = 0 \leftrightarrow x = y$
  La distancia entre dos elementos será cero si y solo si se está hablando del mismo elemento.
- $\delta(x, y) = \delta(y, x)$
  La distancia debe ser simétrica, la distancia de *x* a *y* debe ser la misma que dé *y* a *x*.
- $\delta(x, y) \leq \delta(x, z) + \delta(z, y)$
  Propiedad de la desigualdad triangular, la distancia de x a y es menor o igual que la suma de *x* a *z* más la distancia de *z* a *y*. La distancia euclídea siempre será dada por dos elementos dentro de un espacio euclídeo, está basada en el teorema de Pitágoras.

En un plano de dos dimensiones así se visualiza:

Gráfico 3.1 Distancia Euclídea en Dos Dimensiones.



Fuente: Elaboración propia con base en García Hernández & González Segura, 2015.

Para el caso anterior se tienen dos elementos A=(x_1, y_1) y B=(x_2, y_2)por lo tanto se define de la siguiente manera:

$$\delta(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

De forma general se tienen dos puntos (*P* y *Q)* en un plano n-dimensional

$$\delta(P, Q) = \sqrt{(Q_1 - P_1)^2 + (Q_2 - P_2)^2 + \cdots + (Q_n - P_n)^2}$$

3.4.2 Selección del Número de Clústers

Como antes se había mencionado, el análisis clúster forma parte del aprendizaje no supervisado, puesto que no hay una etiqueta preconocida que pueda servir como guía para el proceso o verificar su validez. En la metodología de clúster, el siguiente paso sería encontrar el número adecuado de clústers necesario para una partición adecuada, ya que todos los métodos de partición necesitan conocer cuántas agrupaciones se realizarán.

Existen diversos métodos para poder conocer el número óptimo de clústers para los datos, así como para calcular el numero ***K*** entre los más comunes:

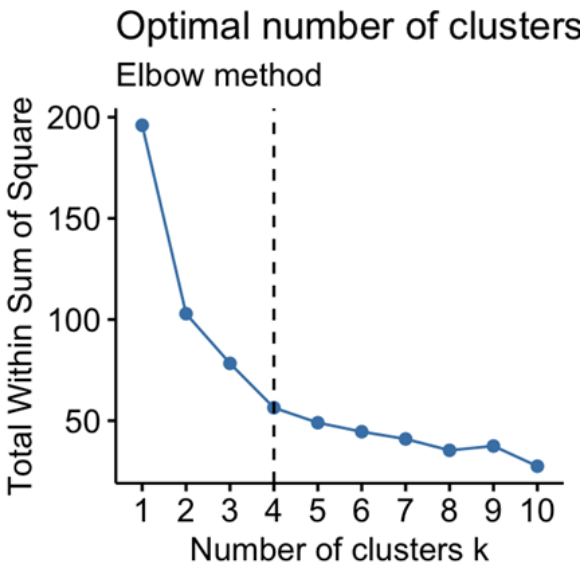
3.4.2.1 Elbow Methods - Método del Codo

Es una técnica visual que puede ayudar a detectar el número de conglomerados que podría usarse para analiza un conjunto de datos. Este tipo de gráfico resulta del cálculo de la suma del error cuadrado (SSE *por sus siglas en ingles*) (Humaira & Rasyidah, 2018). Dicho de otra manera, el método del codo calculará la "inercia", que es la distancia al cuadrado de cada objeto del clúster, al centroide de este. la fórmula para calcular en SSE o "inercia será:

$$SSE = \sum_{i=0}^{N} ||x_i - \mu||^2$$

Este método está basado en la distancia intraclúster, en el que el algoritmo calcula los SSE para diversos valores de ***K***, de este modo al graficar dichos valores se observará un cambio brusco (el punto codo) que indicará qué número de clúster sería el más adecuado para esos datos.

Gráfico 3.2 Ejemplo Grafico de Método del Codo (Elbow Method) con optimo K=4.



Fuente: Tomada de Cluster Validation Essentials (Kassambara, 2017)**.**

Este cambio en el grafico se debe a que en medida que incrementa el número **K,** la distancia media intraclúster no mejora; incluso, se mantiene, y, aunque el descenso se mantiene, la caída se vuelve menos dramática.

3.4.2.2 Silhouette Method – Método de la Silueta

Al igual que el método anterior , este es un método gráfico que estudia la distancia de separación entre los grupos resultantes. Es decir, muestra cuán cerca está cada punto en un clúster de los puntos en los clústers vecinos. La relación matemática será definida como (Rousseeuw, 1986):

$$S(x) = \frac{b(x) - a(x)}{\max\left[a(x), b(x)\right]}$$

Donde:

a(x)= Distancia promedio de x a todos los demás puntos del mismo clúster

b(x)= Distancia promedio de x a todos los demás puntos del clúster más cercano.

El coeficiente de la silueta cumplirá $-1 \leq S(x) \leq 1$ donde la interpretación es:

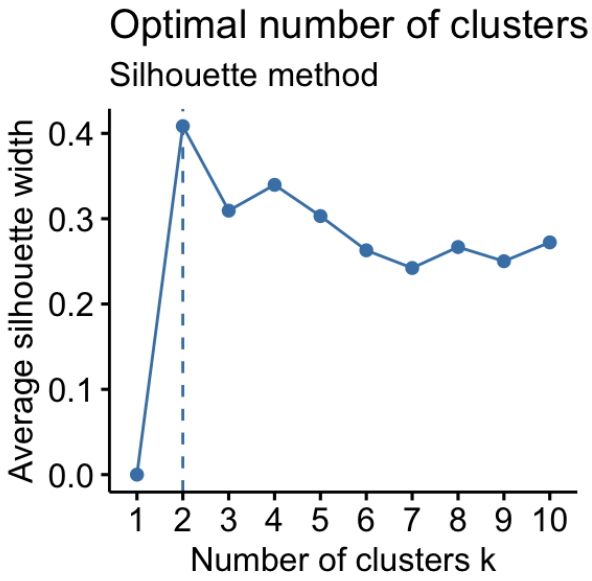
- -1 si es un mal agrupamiento; lo que significa que los puntos están más cerca de los clústeres vecinos que del mismo clúster.
- 0 si es un agrupamiento indiferente.
- 1 si es un buen agrupamiento; puesto que los puntos están muy cerca de su propio clúster y muy lejos de los clústeres vecinos.

  El coeficiente del clúster completo será:

$$SC = \frac{1}{N}\sum_{i=1}^{N} S(x)$$

Es uno de los métodos más significativos y claros para determinar el número de clúster; sin embargo, su complejidad radica en que puede ser una técnica muy larga y tediosa, dado que el cálculo se debe de realizar para cada elemento de estudio en los diferentes valores de ***K*** (Wang, Franco, & Kelleher, 2017)**.**

Gráfico 3.3 Ejemplo Grafico de Método de la Silueta (Silhouette Method) con óptimo K=2.



Fuente: Tomada de Cluster Validation Essentials (Kassambara, 2017)

Se trazará una curva con la silueta promedio para cada número de clústers y el numero adecuado para el análisis será donde se encuentre el máximo global de la gráfica. Para el método de la silueta es deseable que haya un puntaje alto. (Gráfico 3.3)

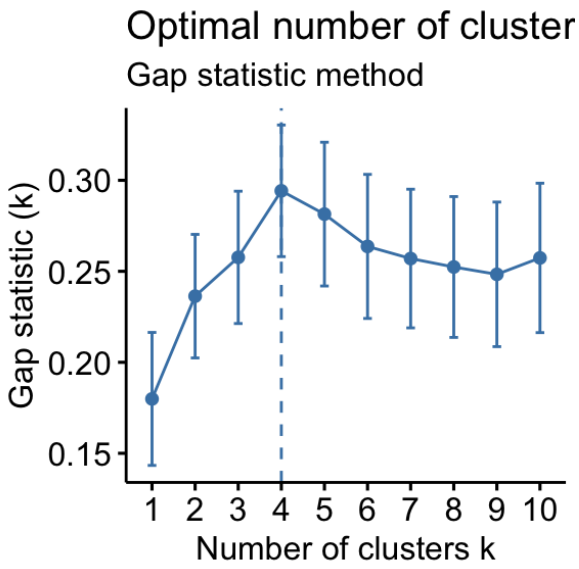
3.4.2.3 Gab Static Method – Método de la Brecha Estadística

Este método compara la variación intraclúster con los diferentes valores que puede tomar ***K*** contra los valores esperados dentro de una distribución de referencia de datos. El número de clúster optimo será un valor que maximice la brecha, cuyo significado será que la estructura clúster estará lejos de la estructura de una distribución uniforme aleatoria (Mohajer, Englmeier, & Schmid, 2010).

La forma matemática de poder calcular la brecha es:

$$Gap_n(k) = E_n^*[\log(W_k)] - \log(W_k)$$

Gráfico 3.4 Ejemplo Grafico de Método de la Brecha Estadística (Gab Static Method) con óptimo K=4.



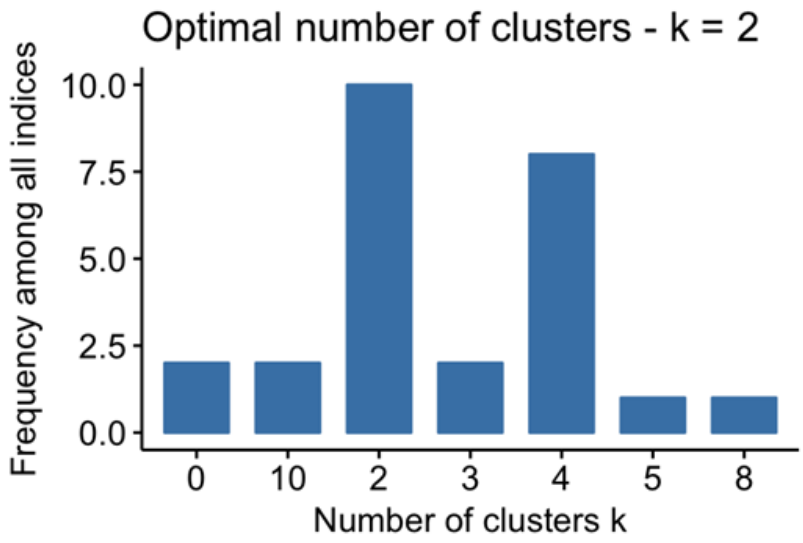
Fuente: Tomada de Cluster Validation Essentials (Kassambara, 2017)**.**

Al igual que en el método se la silueta, en la brecha estadística gráficamente se busca el valor máximo.

Estos son los métodos más comunes. Sin embargo, aunque existen una gran cantidad de algoritmos que nos pueden ayudar a obtener el valor ***K***, es importante señalar que ninguno tiene una efectividad total, puesto que en ocasiones dependerá del criterio del investigador y la forma de abordar el tema. Asimismo, cabe señalar que es común que el análisis clúster sea realizado a través de softwares estadísticos, con lo que se pueden calcular simultáneamente varios métodos para encontrar ***K***.

En este caso particular, R Studio cuenta con la función ***NbClust***, la cual calcula varios métodos e indica cuántos proponen ***K*** números de clústers, lo que permite al investigador, a través de la regla de la mayoría y/o estudio del caso, decidir la ***K*** óptima para su análisis apoyándose de un gráfico de frecuencia (R Documentation, 2022).

Gráfico 3.5 Frecuencia de Número de Clúster Según Índices



Fuente: Tomada de Cluster Validation Essentials *(Kassambara, 2017).*

3.4.3 Método K-Means

Para finalizar con nuestra metodología, es importante saber qué algoritmo de clústering será utilizado. Entre los más comunes destacan: *K-Means*, *K-Modes y K-Medoids*, cuya diferencia radica en el tipo de variable que se utiliza. Entre estos, el método más utilizado es *K-Means*.

Trueba Espinosa (2017) define al Método *K-Means* como un método de clústering particional, el cual requiere conocer el valor **K** que define los números de clústers en los que se partirán los datos. A menudo se utiliza la distancia euclidiana para medir la disimilitud, cuya función se define como:

$$J = \sum_{i=1}^{c} \sum_{x_k \epsilon G_i} ||x_k - c_i||^2$$

Los grupos se definirán por una matriz $c \times n$ de pertenencia ***U***, donde el elemento $u_{i,j}$ tomará los valores:

$$\begin{cases} 1, & ||x_j - c_i||^2 \leq \left||x_j - c_k|\right|^2, \quad \forall k \neq i \\ 0, & en\ otro\ caso \end{cases}$$

Esto se explica, ya que un punto solo puede pertenecer a un grupo; además, la matriz de pertenencia ***U*** tiene la siguiente propiedad:

$$\sum_{i=1}^{c} u_{i,j} = 1; \ \ \forall j = 1, \dots, n$$

Donde además cumple:

$$\sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j} = n$$

Si $u_{i,j}$ es fijo, el centro optimo c_i que minimiza *J* es la media de los vectores del grupo i. Esta media del conjunto de datos es importando puesto que aproxima a encontrar el centroide.

$$c_i = \frac{1}{|G_i|} \sum_{x_k \in G_i} x_k$$

Donde $|G_i| = \sum_{j=1}^{n} u_{i,j}$ es el tamaño de G_i.

El algoritmo de *K-Means* es presentado con un conjunto de datos x_i, $i = 1, \dots, n$.

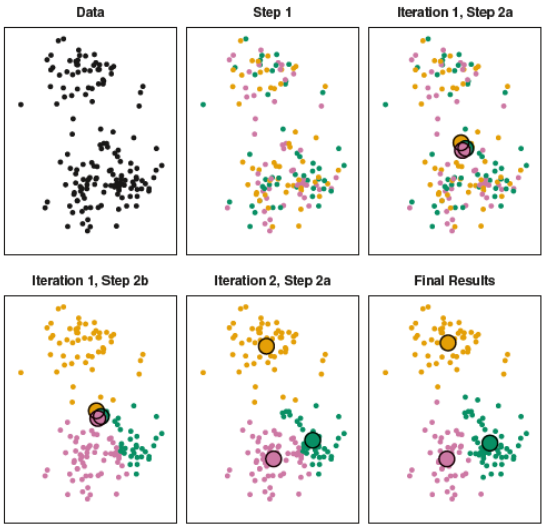
3.4.4 Algoritmo K-Means

Rendón Lara y Abúndez Barrera (2016) mencionan, como pasos a seguir para realizar el algoritmo de *K-Means*, los siguientes :

1. Iniciar los centros de los grupos $c_i, i = 1, \dots c$. Comúnmente se seleccionan aleatoriamente c puntos de conjunto de datos.
2. Determinar la matriz ***U*** con la función de pertenencia.
3. Calcular la función *J*. El algoritmo se detiene hasta alcanzar el valor de tolerancia marcado. Es decir, minimizar el valor de *J* hasta encontrar el valor que el investigador considere conveniente.
4. Actualizar los centros y comenzar el algoritmo de nuevo.

Gráficamente el algoritmo pasaría por estas formas:

Gráfico 3.6 Representación gráfica de iteraciones en los clústers con K=3.



Fuente: An Introduction to Statistical Learning. (James, Gareth, et al., 2013)

Para poder tener un método *K-Means* correcto es importante realizar múltiples iteraciones, pues el algoritmo convergerá. La estandarización de datos es una buena herramienta para que la partición sea justa y no exista ponderación en las variables (a menos que así lo requiera el investigador); comenzar con un valor ***K*** lo bastante grande para poder disminuirlo poco a poco eliminando clústers pequeños que posean poca o nula información.

Capitulo 4 Análisis Exploratorio de Datos

4.1 Introducción de Capítulo

En este capítulo se presentarán los resultados del análisis clúster aplicado a los Seguros de Gastos Médicos Mayores, donde el objetivo principal es identificar patrones de comportamiento entre los siniestros ocurridos en el periodo 2019-2020.

Para comenzar, se explicará la aplicación del método clúster, es decir, cómo se maneja la base a través del software R Studio, librerías a utilizar, así como la forma de cargar de la base de datos, filtrar la base (variables y datos), y así, poder encontrar el número de clúster para ambas bases (2019 y 2020).

Una vez que se tiene la propuesta del número de clúster, podemos implementarlos en las respectivas bases de datos y analizar si ese número es óptimo para el estudio que se trabaja y encontrar si existe algún error; de este modo, cuando el número de Clúster sea el pertinente, se puede realizar un análisis exploratorio por cada clúster; cuya finalidad es encontrar qué tipo de datos están en cada conglomerado para poder encontrar un patrón de división entre los años estudiados.

4.2 Aplicación del Método Clúster

Las bases obtenidas de la CNSF (2022) oscilan entre los 17 y 24 mil datos para 2020 y 2019, respectivamente. En términos prácticos, la aplicación y manejo de los datos se empleó del mismo modo para los dos años. Por tal razón, se explicará primero para el año 2019.

Como previo a emplear el software estadístico R se debe de conocer qué librerías se utilizan para hacer un análisis clúster. La siguiente tabla sintetiza las herramientas utilizadas.

Ilustración 4.1 Librerías de R a Utilizar

```
library(readxl)
library(dplyr)
library(stringr)
library(cluster)
library(NbClust)
library(ggplot2)
library(factoextra)
library(ClusterR)
library(psych)
```

Fuente: Elaboración propia.

4.2.1 Implementación Método Clúster

Una vez que se seleccionaron las enfermedades que participarían en el estudio y se eliminaron todos aquellos datos que tenían información "No Disponible", el tamaño de la base se reduce considerablemente, pasando de 23,749 a 1,534 datos en 2019, y de 17,726 a 1,326 en 2020, con todos sus atributos (Edad, Genero, Entidad de Residencia, Subtipo de Seguro, Diagnostico, Tipo de Evento Hospitalario, Procedencia de Ingreso, Motivo de Egreso, Numero de Reclamaciones, Monto de Honorarios Médicos, Monto de Deducible o Copago y Monto de Coaseguro), los cuales serán sometidos al estudio en el software estadístico R Studio.

Como primer paso a realizar se mandará a llamar la base a través del comando *read_excel* y se aplicará la función *na.omit* para que automáticamente se eliminen los valores faltantes, tal como se muestra a continuación:

Ilustración 4.2 Llamado de Base de Datos 2019

```
Salud_2019 <- read_excel("C:/Users/paola lizeth/Desktop/TESIS/Enfermedades de Mayor Muerte en Mexico.xlsx",
                         sheet = "2019")
Salud_2019<-na.omit(Salud_2019)
```

Fuente: Elaboración propia.

Este proceso se debe hacer para cada una de las bases de datos; por consiguiente, para el año 2020 el comando sería el siguiente:

Ilustración 4.3 Llamado de Base de Datos 2020

```
Salud_2020 <- read_excel("C:/Users/paola lizeth/Desktop/TESIS/Enfermedades de Mayor Muerte en Mexico.xlsx",
                         sheet = "2020")
Salud_2020<-na.omit(Salud_2020)
```

Fuente: Elaboración propia.

Una vez que la base esté en R, podemos preparar los datos y seleccionar solo las variables numéricas a utilizar: Edad, Monto de Hospitalización y Monto de Honorarios Médicos. Dichas variables tendrán que ser mayores o igual a cero, puesto que no tendrían lógica algún negativo en esas variables. Asimismo, los datos resultantes tendrán que ser sintetizados de la forma *tibble*, similar a un *data.frame* cuya dimensión será 1,414 × 3 para el año 2019, (véase Ilustración 4.4) y de tamaño 1,275 × 3 en el caso 2020, (Ilustración 4.5). La disminución en los datos se debe a la función *na.omit*, la cual omitirá todos aquellos datos que no puedan ser manejados por el software, como los que poseen la característica de "No Disponible", o los que no cumplan el criterio de positividad en los Montos de Hospitalización y Monto de Honorarios Médicos.

Ilustración 4.4 Selección de Datos Mayores o Iguales a Cero 2019

```
base<-as_tibble(Salud_2019) %>% mutate(EDAD=as.numeric(EDAD)) %>%
  filter(`MONTO DE HOSPITALIZACION`>=0,
         `MONTO HONORARIOS MEDICOS`>=0)
```

Fuente: Elaboración propia.

Ilustración 4.5 Selección de Datos Mayores o Iguales a Cero 2020

```
base<-as_tibble(Salud_2020) %>% mutate(EDAD=as.numeric(EDAD)) %>%
  filter(`MONTO DE HOSPITALIZACION`>=0,
         `MONTO HONORARIOS MEDICOS`>=0)
```

Fuente: Elaboración propia.

Ahora, se renombrarán las variables para que su uso sea más adecuado al lenguaje de programación que emplea el software. Este procedimiento se realiza de la misma manera para ambos años.

Ilustración 4.6 Renombre de Variables

```
numericas<-base %>% select(EDAD,`MONTO DE HOSPITALIZACION`,`MONTO HONORARIOS MEDICOS`,
                           `MONTO DE DEDUCIBLE O COPAGO`, `MONTO COASEGURO`)
colnames(numericas)<-c("Edad","Monto_Hosp","Monto_Hon","Monto_Ded","Monto_Coa")
```

Fuente: Elaboración propia.

Como se mencionó en el capítulo tres, una parte fundamental previa a llevar a cabo un análisis clúster es la estandarización de datos, pues sus unidades de medición varían; por ejemplo, para la variable Edad (años), el tipo de dato es entero positivo de máximo 99, mientras que para Monto de Hospitalización y Monto de Honorarios Médicos (pesos), el tipo de dato incluye decimales y el dato puede llegar a millones.

Ilustración 4.7 Estandarización de Datos

```
datos_cluster<-numericas %>% mutate(Edad=scale(Edad),
                                    Monto_Hosp=scale(Monto_Hosp),
                                    Monto_Hon=scale(Monto_Hon))
```

Fuente: Elaboración propia.

La función *NbClust* es recurrentemente empleada, pues ayudará a encontrar el número de clústers óptimo a utilizar de acuerdo con diversos índices. A dicha función se le debe indicar qué tipo de distancia *("euclídea")* y qué método de clusterización *("k Means")* se utilizará, así como los índices, que queda expresado de la siguiente manera:

Ilustración 4.8 Utilización del Comando NbClust

```
num_cluster<-NbClust(data=datos_cluster,distance = "euclidean",
                     method = "kmeans", index = "all")
```

Fuente: Elaboración propia.

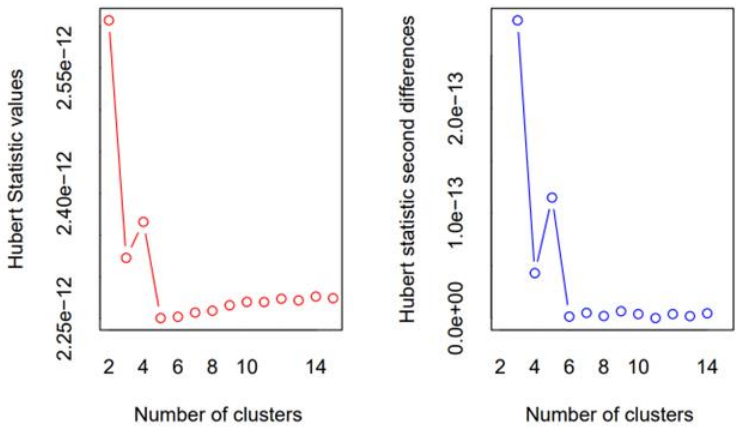
Por tal motivo, a partir de este momento los resultados cambian para cada base; por tanto, se denominan los siguientes apartados.

4.2.2 Encontrar el Número de Clústers para 2019

Una vez utilizado el comando que ayuda a encontrar el número de clúster más adecuado, con base a los gráficos, podemos darnos una idea de cuál ***k*** utilizar. El primer grafico a entender es el Índice de Hubert.

El *Índice de Hubert* es un método gráfico que ayuda a determinar el número de conglomerados, buscando una *"rodilla"*, el cual corresponderá a un aumento significativo.

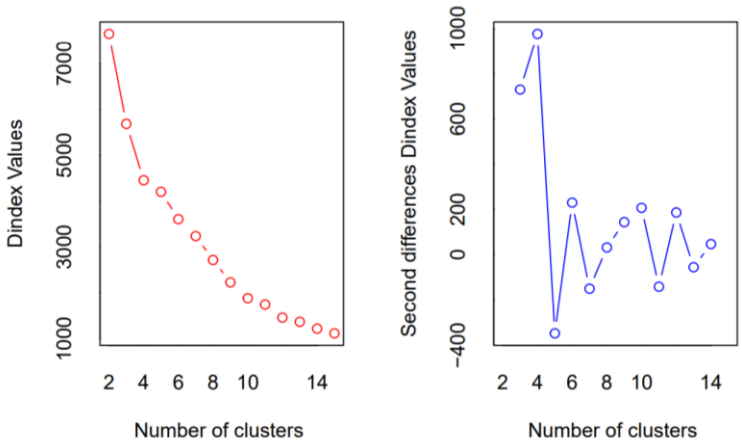
Gráfico 4.1 Índice de Hubert para Clústers 2019



Fuente: Elaboración propia.

Del mismo modo, R Studio arroja como resultado el gráfico del Índice D, el cual, al igual que en el gráfico de Hubert, se busca un *"codo"*, un cambio significativo a partir de las segundas diferencias.

Gráfico 4.2 Índice D para Clústers 2019



Fuente: Elaboración propia.

Los demás índices no se representan de modo grafico sino como un resumen en una lista; según este resumen el investigador determinará que valor le dará a ***k*** para dividir sus datos en ***k*** número de clústers.

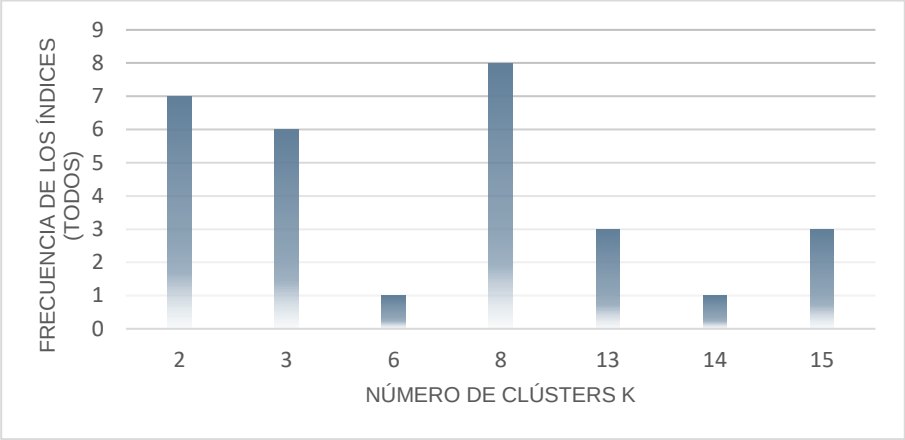
Ilustración 4.9 Listado de Índices, con sus Frecuencias 2019

```
*****************************************************************
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 6 proposed 3 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 2 proposed 8 as the best number of clusters
* 3 proposed 13 as the best number of clusters
* 1 proposed 14 as the best number of clusters
* 3 proposed 15 as the best number of clusters

                 ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

Fuente: Elaboración propia.

A su vez el grafico de frecuencias quedaría de la siguiente forma:

Gráfico 4.3 Frecuencia en Numero de Clúster 2019



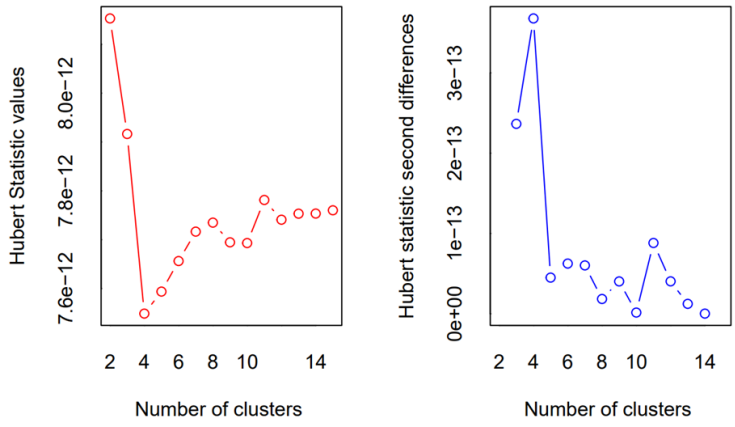
Fuente: Elaboración propia.

Por tanto, de acuerdo con la regla de la mayoría, podemos determinar que un **número adecuado de clústers para los datos del 2019 son dos clústers.**

4.2.3 Encontrar el Número de Clúster para 2020

En cuanto al año 2020, al utilizar la función *NbClust*, encontramos diferencias notables entre los índices gráficos. Es decir, el Índice Hubert no se ve tan armónico, parece tener más saltos que lo observado en 2019.

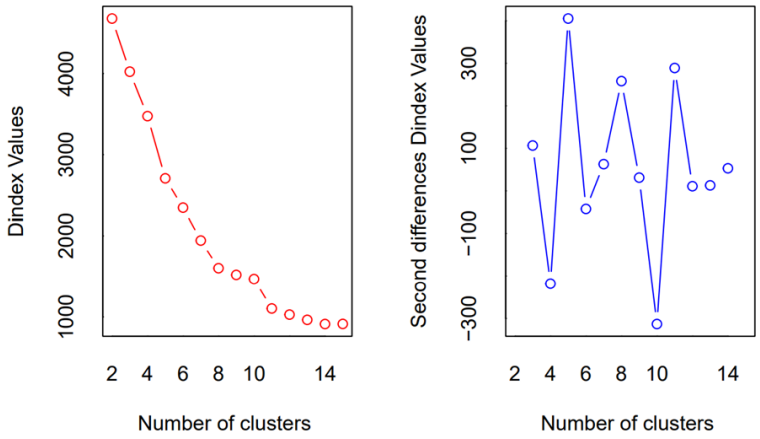
Gráfico 4.4 Índice de Hubert para Clústers 2020



Fuente: Elaboración propia.

Por tanto, el Índice D sufrirá la misma suerte,

Gráfico 4.5 Índice de D para Clústers 2020



Fuente: Elaboración propia.

Con estos dos gráficos podemos ver que el manejo de datos para el 2020 será un poco más complejo que para 2019, siendo aquel un año atípico a nivel global. Por tal motivo, el listado de índices por propuesta de ***k*** clústers es el siguiente:

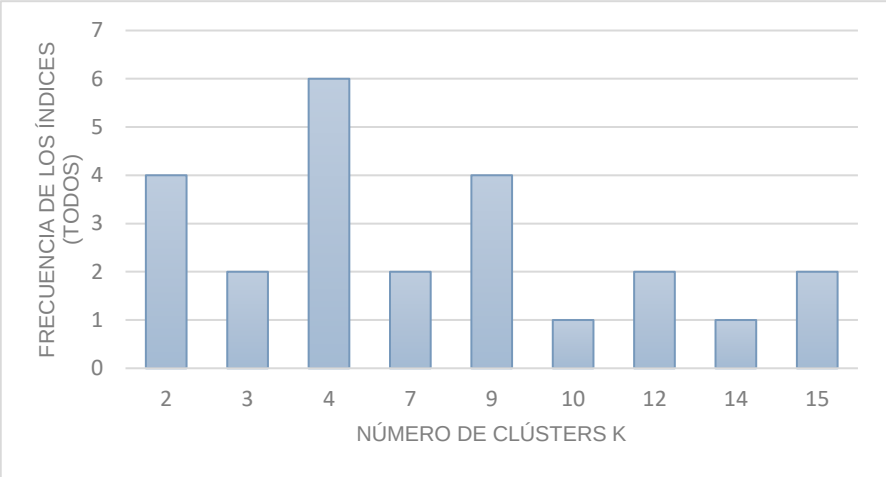
Ilustración 4.10 Listado de Índices, con sus Frecuencias 2020

```
*******************************************************************
* Among all indices:
* 4 proposed 2 as the best number of clusters
* 2 proposed 3 as the best number of clusters
* 6 proposed 4 as the best number of clusters
* 2 proposed 7 as the best number of clusters
* 4 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters
* 2 proposed 12 as the best number of clusters
* 1 proposed 14 as the best number of clusters
* 2 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  4
```

Fuente: Elaboración propia.

El histograma de frecuencias nos muestra más opciones de particionar los datos:

Gráfico 4.6 Frecuencia en Numero de Clúster 2020



Fuente: Elaboración propia.

La información y apoyos estadísticos nos sugieren que una partición **optima seria en cuatros clústers**, pero, aplicando el principio de parsimonia, el cual apela a una explicación sencilla: el año 2020 es un año atípico, pues presentó un alza en una enfermedad no vista anteriormente y, con el resultado para 2019, podría determinarse que **dos clústers es un numero adecuado para el estudio de los datos 2020.**

4.3 División de los Datos Según Clústers (Resultados)

De acuerdo con los cálculos, se determinó que la forma óptima de particionar los datos es de dos y cuatro para 2019 y 2020, respectivamente. Una vez teniendo esa información, se procede a la clusterización, la cual se mostrará de forma gráfica.

4.3.1 Propuesta de Clústers 2019

Entre los clústeres propuestos, los que más se repiten son dos y tres particiones; siete índices proponían dos clústeres. Por consiguiente, se harán dos modelos, el **Modelo 1** contará con dos centroides, mientras que el **Modelo 2** tendrá tres centroides, esto con la finalidad de encontrar la mejor opción entre ellos.

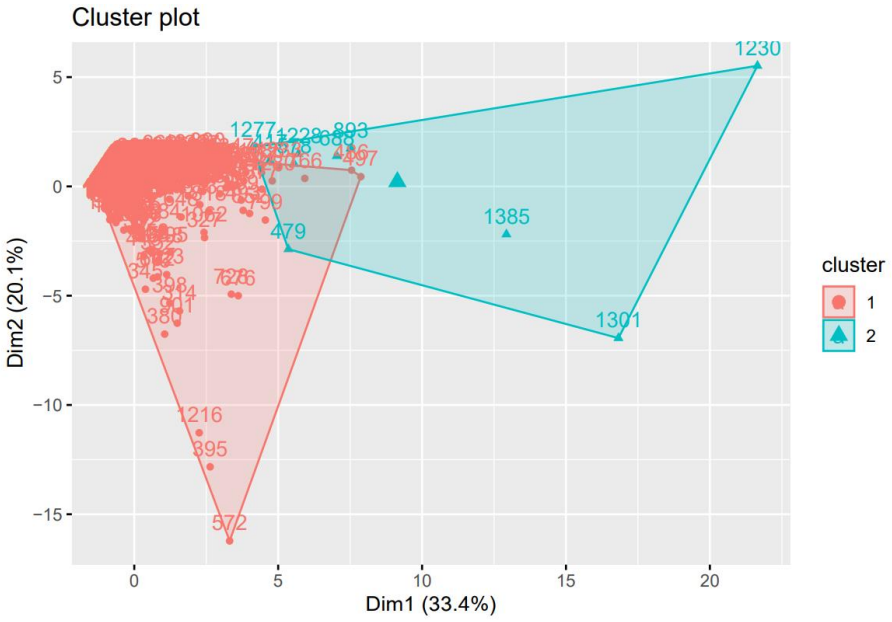
Ilustración 4.11 Creación del Modelo 1 y Modelo 2

```
Modelo1<- kmeans(x=datos_cluster,centers = 2)
base$Modelo1<-Modelo1$cluster
fviz_cluster(object = Modelo1,data = datos_cluster)
Modelo2<- kmeans(x=datos_cluster,centers = 3)
base$Modelo2<-Modelo2$cluster
fviz_cluster(object = Modelo2,data = datos_cluster)
```

Fuente: Elaboración propia.

La función *fviz_cluster* ayudará a graficar los datos ya agrupados; así que, si a los datos estandarizados los separamos en dos, se expresarían gráficamente de la siguiente forma:

Gráfico 4.7 Propuesta de Dos Clústers Año 2019

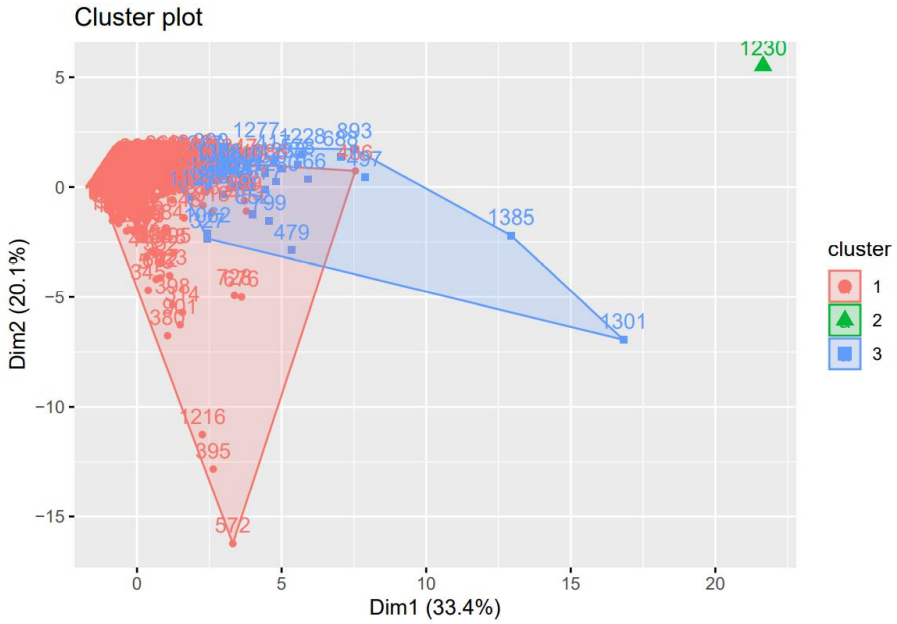


Fuente: Elaboración propia.

Este gráfico representa la mayor parte de variabilidades de los datos, puesto que se concentra en 53.5% de esta. A simple vista, se puede ver que los clústers están bien definidos, aunque existe un sesgo de datos en el clúster número uno.

Cuando se utilizó la función *NbClust*, la segunda mayoría era tres, puesto que seis índices consideraban “óptima” una división en 3 centroides. Por lo tanto, al graficar dicha división, nos encontramos con el siguiente grafico:

Gráfico 4.8 Propuesta de Tres Clústers Año 2019



Fuente: Elaboración propia.

Se puede observar que tres clústeres, pese a ser el segundo ***k*** más apoyado por los índices, resultan no ser “óptimos”, ya que el conglomerado número dos no posee información relevante, pues está conformada solo por un elemento que, dado sus características, podría tratarse de un outlier, del cual se conoce que el Monto de Hospitalización fue de 1,493,100 pesos, con nueve reclamaciones y de 64 de edad.

Naturalmente, por los principios previamente planteados en la metodología se determina que una división de dos clústers sería más favorable para el estudio, pues aportaría más información y podría dar una idea general para “clasificar” nuevos casos.

4.3.2 Propuesta de Clústers 2020

En el apartado anterior se proporcionó evidencia de que lo óptimo para el análisis fue emplear dos clústers. Para el año 2020 la situación es diferente, puesto que se tiene como primera opción dividir los datos en cuatro conglomerados y, como segunda opción, una división bipartida.

Se volverá a ejemplificar dos modelos: el Modelo 1 será aquel que divida los datos en dos clústers, mientras que el Modelo 2 es en donde se opta por la opción de 4 centroides (véase Ilustración 4.12).

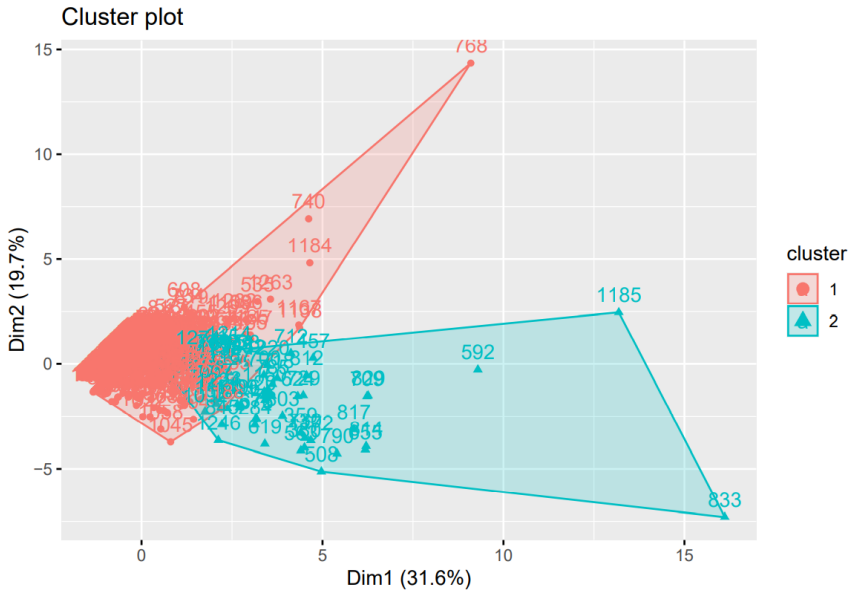
Ilustración 4.12 Creación del Modelo 1 y Modelo 2

```
Modelo1<- kmeans(x=datos_cluster,centers = 2)
base$Modelo1<-Modelo1$cluster
fviz_cluster(object = Modelo1,data = datos_cluster)
Modelo2<- kmeans(x=datos_cluster,centers = 4)
base$Modelo2<-Modelo2$cluster
fviz_cluster(object = Modelo2,data = datos_cluster)
```

Fuente: Elaboración propia.

Una vez construidos los modelos, se obtiene del Modelo 1 una variabilidad de 51.3% menor que la obtenida para el 2019, y se encuentra un sesgo en donde hay una gran cantidad de datos que son repartidos entre los dos clústers.

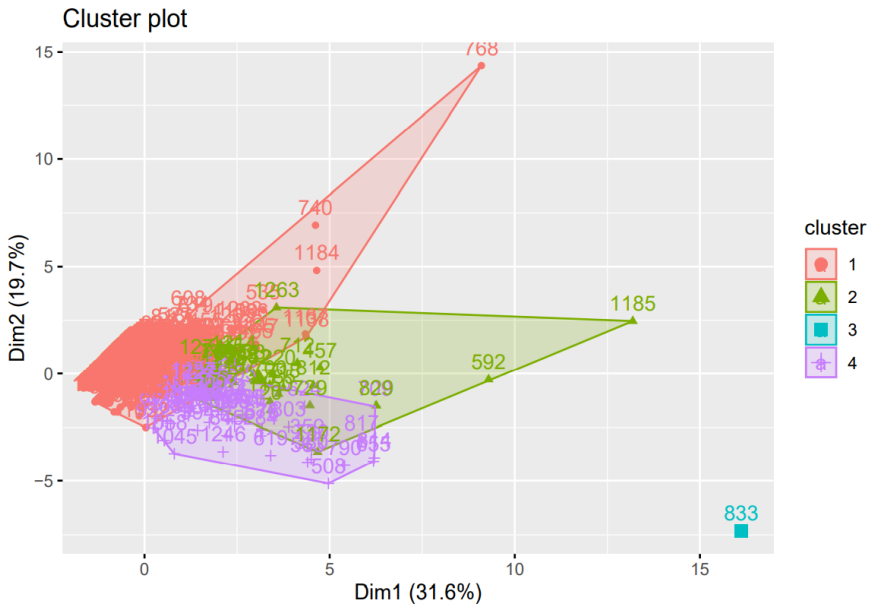
Gráfico 4.9 Propuesta de Dos Clústers Año 2020



Fuente: Elaboración propia.

Finalmente, en el Modelo 2 se aprecia un comportamiento similar a los tres clústers del 2019, es decir existe un clúster formado por un elemento (véase Gráfico 4.10).

Gráfico 4.10 Propuesta de Cuatro Clústers Año 2020



Fuente:

Fuente: Elaboración propia.

Esta partición no es óptima, pues el clúster numero 3 solo posee un dato: 17 reclamaciones con un monto de hospitalización en 848,352 pesos y una edad de 35 años. Al igual que en el 2019, este tipo de partición no aportara más información relevante.

Como consecuencia, se puede decir que cualquier otra partición mayor a dos clústers resultaría en clústers con un solo elemento. Con todo esto, podemos determinar que lo más conveniente sería particionar ambas bases de datos (2019 y 2020) en dos clústers cada uno, lo cual podría dar una mejor explicación a los datos.

4.4 Descripción de los Clústers del Año 2019

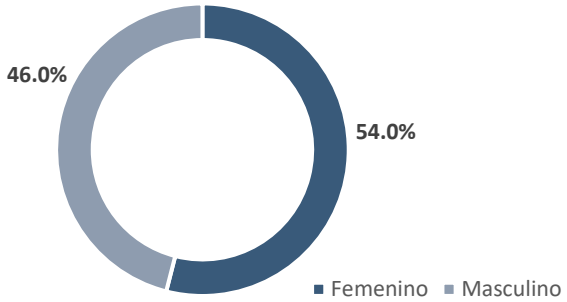
Como parte del Análisis Exploratorio de Datos (EDA, por sus siglas en inglés), definimos la partición de la base de datos del 2019 y 2020 en dos clústers, los cuales procederemos a describir estadísticamente.

4.4.1 Clúster 1 Año 2019 (C119)

Para el año 2019, el estudio se redujo a 1,414 casos de estudio, donde el Cluster 119 posee 1404 datos, cantidad que equivale a 99% de los datos que conforman el estudio para el año 2019. De este modo, el 56% de los pacientes requirieron internamiento, el 54% fueron femeninas, la referencia de otra unidad resulta ser la principal causa de reclamaciones, mientras que casi el 97% de los pacientes fueron dados de alta por mejoría.

Los datos completos son:

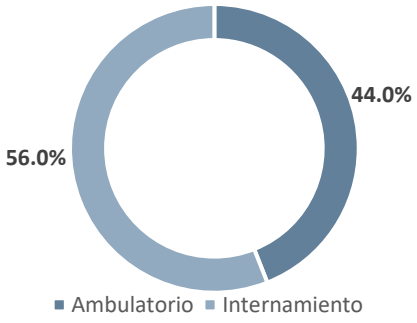
Gráfico 4.11 Datos de Género C119



Género	
Femenino	Masculino
758	646

Fuente: Elaboración propia.

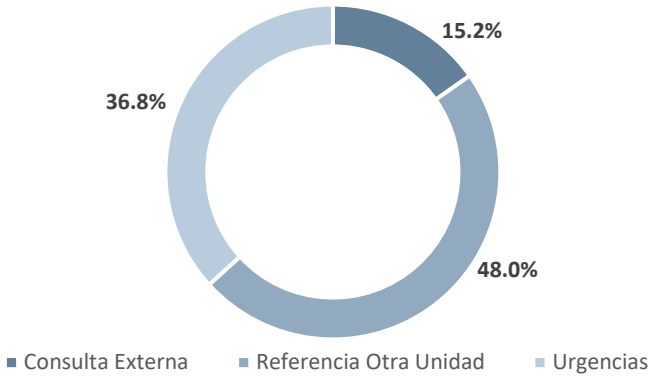
Gráfico 4.12 Datos de Tipo de Internamiento C119



Tipo de Internamiento	
Ambulatorio	Internamiento
618	786

Fuente: Elaboración propia.

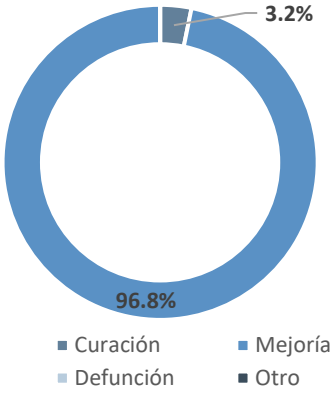
Gráfico 4.13 Datos de Procedencia de Ingreso C119



Procedencia de Ingreso		
Consulta Externa	Referencia Otra Unidad	Urgencias
214	674	516

Fuente: Elaboración propia.

Gráfico 4.14 Datos de Procedencia de Egreso C119

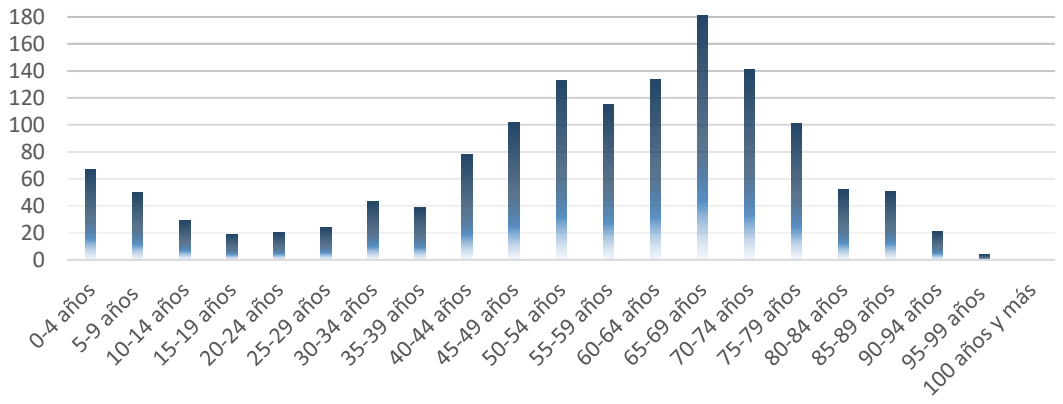


Motivo de Egreso			
Curación	Mejoría	Defunción	Otro
45	1359	0	0

Fuente: Elaboración propia.

Los datos están divididos de acuerdo con los grupos quinquenales determinados por el INEGI. La distribución de la edad para C119 se muestra en la siguiente gráfica:

Gráfico 4.15 Distribución de Edades C119



Fuente: Elaboración propia.

Los grupos quinquenales de 65-69 años y 70-74 años son los que mantienen un alto número de casos registrados, 181 y 141 respectivamente, generando la mayor ponderación: 12.9% y 10.0%.

Otras estadísticas que podemos definir sobre la edad serian su mínimo, mediana, media y máxima, concentradas en la siguiente tabla.

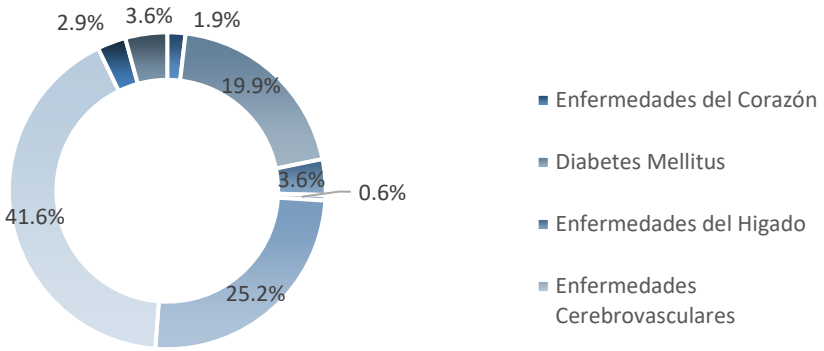
Tabla 4.1 Estadísticas de la Edad para C119

Edad		
Mínimo	**1Q**	**Mediana**
0	44	59
Media	**3Q**	**Máximo**
54.1	70	98

Fuente: Elaboración propia.

Por otro lado, al analizar las enfermedades, podemos determinar que la enfermedad que posee una mayor cantidad de datos es la de tumores malignos, seguido de la influenza y neumonía, siendo la diabetes mellitus la tercera enfermedad de más reclamaciones presenta para el C119.

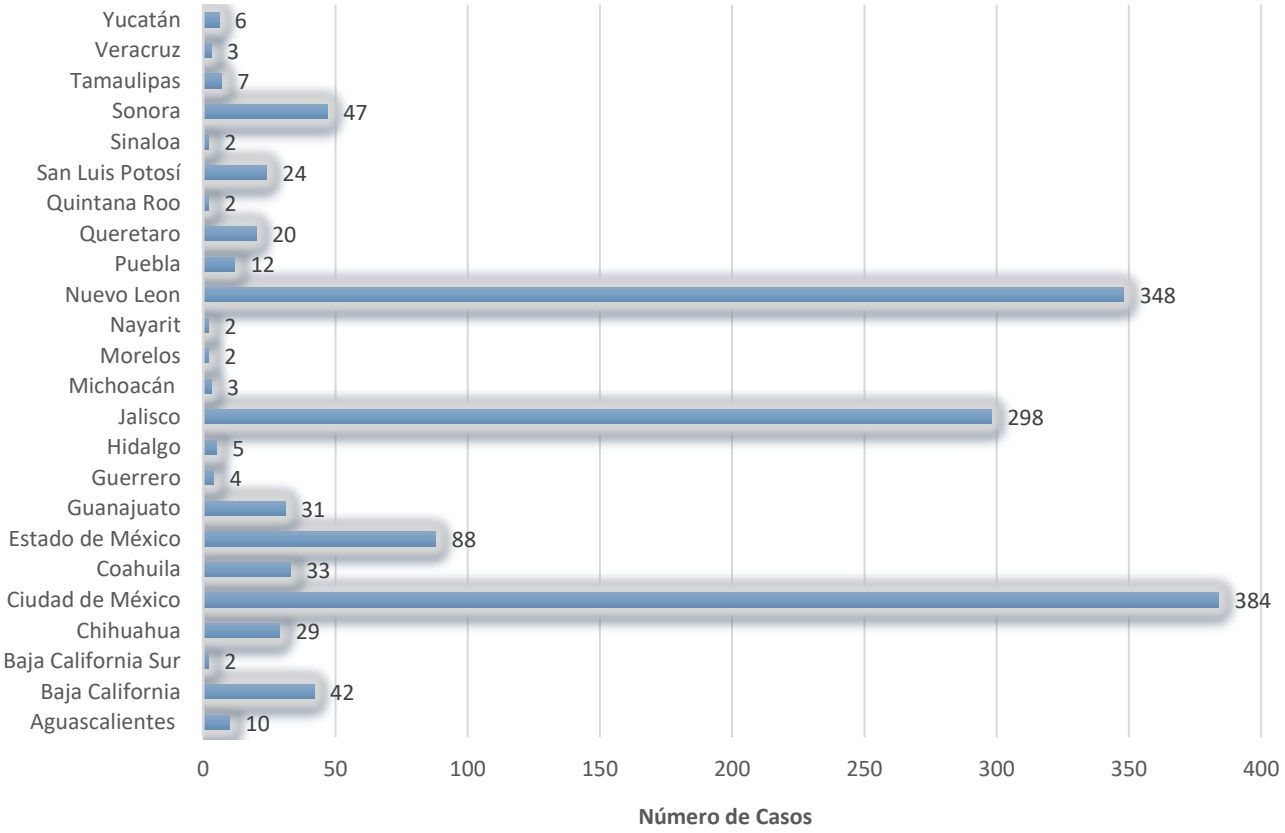
Gráfico 4.16 Porcentajes de Casos Según Enfermedades en C119



Fuente: Elaboración propia.

Los casos por entidad federativa se verían de la siguiente manera, omitiendo los estados que tienen cero casos reportados.

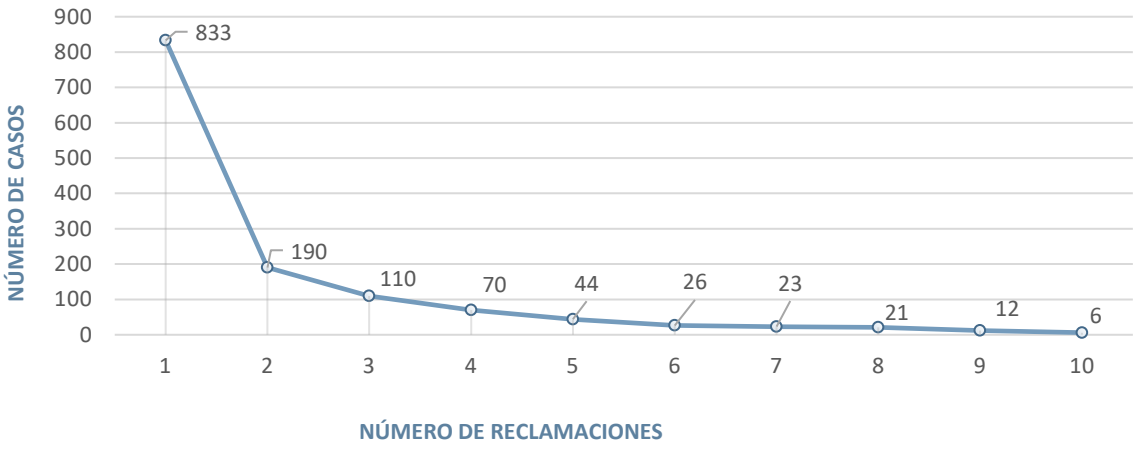
Gráfico 4.17 Estados donde se reportaron reclamaciones de C119



Fuente: Elaboración propia.

Los datos prevalecen con características definidas como un número de reclamaciones pequeño; es decir, mayor número de casos en un menor número de reclamaciones; de tal manera, se pue ver el grafico de la siguiente manera *(véase gráficos 4.18 y 4.19)*

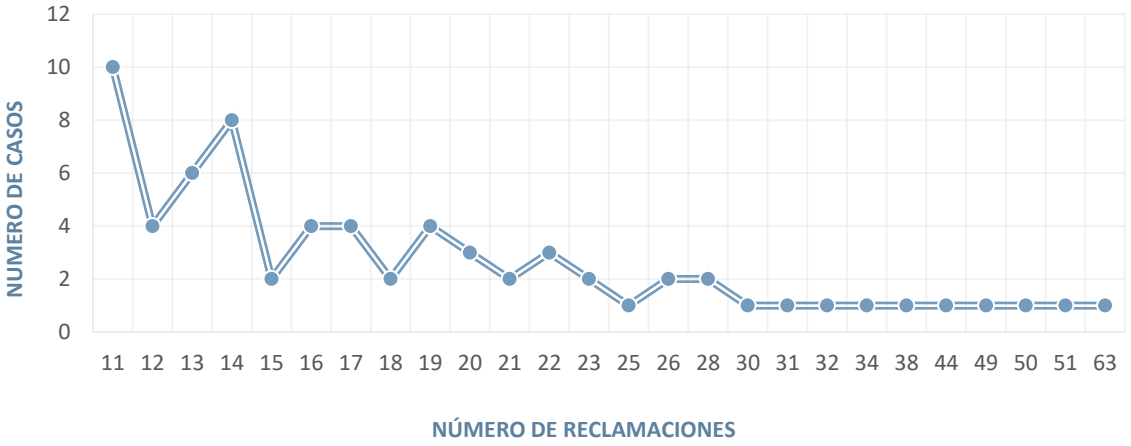
Gráfico 4.18 Número de Casos por Primeras 10 Reclamaciones para C119



Fuente: Elaboración propia.

La mayoría de los casos contenidos en el C119 posee solo una reclamación, el 59.3% de los datos disminuyen estrepitosamente hasta tener un caso entre 63 reclamaciones. Al cambiar la escala para una mejor visualización, podremos observar lo siguiente:

Gráfico 4.19 Frecuencia del Resto de Reclamaciones en C119



Fuente: Elaboración propia.

Pasando al lado económico y analizar los montos de hospitalización, honorarios médicos, coaseguro y deducible concentrado de la siguiente forma. El clúster C119 contiene el monto de hospitalización y de honorarios médicos más altos de toda la base.

Tabla 4.2 Información Económica de los Datos de C119

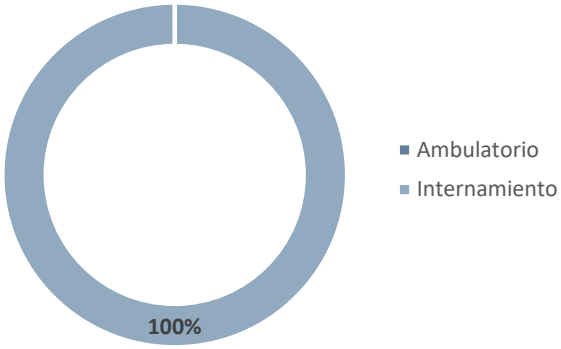
Monto de Hospitalización		
Mínimo	**1Q**	**Mediana**
$0	$1,882	$11,030
Media	**3Q**	**Máximo**
$72,587	$55,731	$2,634,391
Monto de Honorarios Médicos		
Mínimo	**1Q**	**Mediana**
$0	$0	$0
Media	**3Q**	**Máximo**
$1,717	$0	$206,360
Monto de Coaseguro		
Mínimo	**1Q**	**Mediana**
$0	$0	$0
Media	**3Q**	**Máximo**
$3,192	$264	$88,948
Monto de Deducible		
Mínimo	**1Q**	**Mediana**
$0	$0	$0
Media	**3Q**	**Máximo**
$2,001	$0	$73,718

Fuente: Elaboración propia.

4.4.2 Clúster 2 Año 2019 (C219)

Entre tanto podemos analizar el clúster número dos, el cual contiene muy pocos datos, solo 10, los cuales representan 0.7%. Los detalles más importantes resultan ser:

Gráfico 4.20 Datos de Género C219

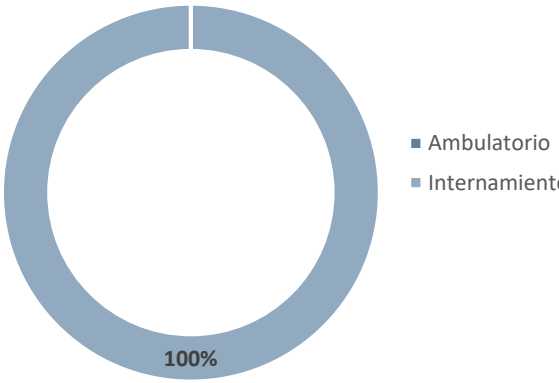


Género	
Femenino	Masculino
5	5

Fuente: Elaboración propia.

La distribución de los datos por género es muy equitativa, pero al momento de pasar a analizar el tipo de internamiento se obtiene que todos los datos del C219 son internamientos, referenciados de otra unidad y además fueron egresados por mejoría *(véase gráficos 4.21, 4.22, 4.23)*.

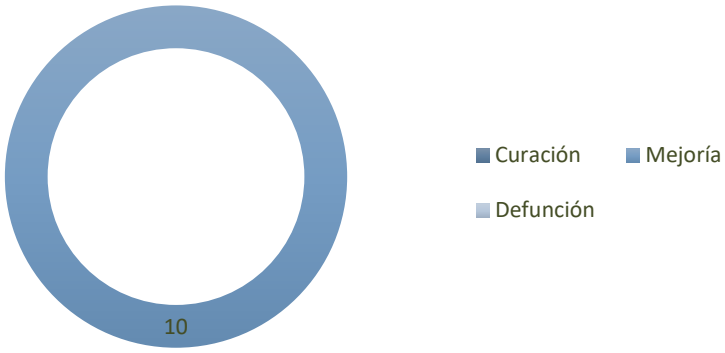
Gráfico 4.21 Datos de Tipo de Internamiento C219



Tipo de Internamiento	
Ambulatorio	Internamiento
0	10

Fuente: Elaboración propia.

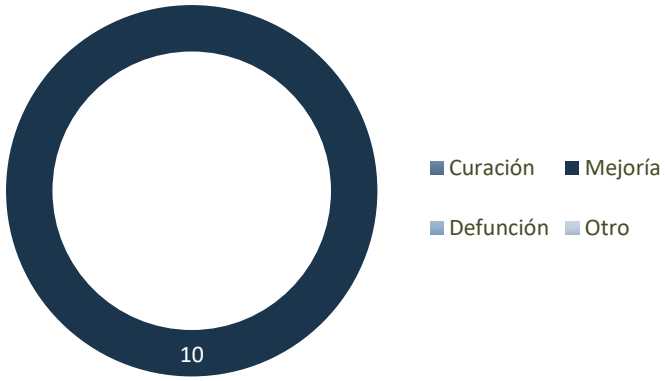
Gráfico 4.22 Datos de Procedencia de Ingreso C219



Procedencia de Egreso		
Consulta Externa	Referencia Otra Unidad	Urgencias
0	10	0

Fuente: Elaboración propia.

Gráfico 4.23 Datos de Procedencia de Egreso C219

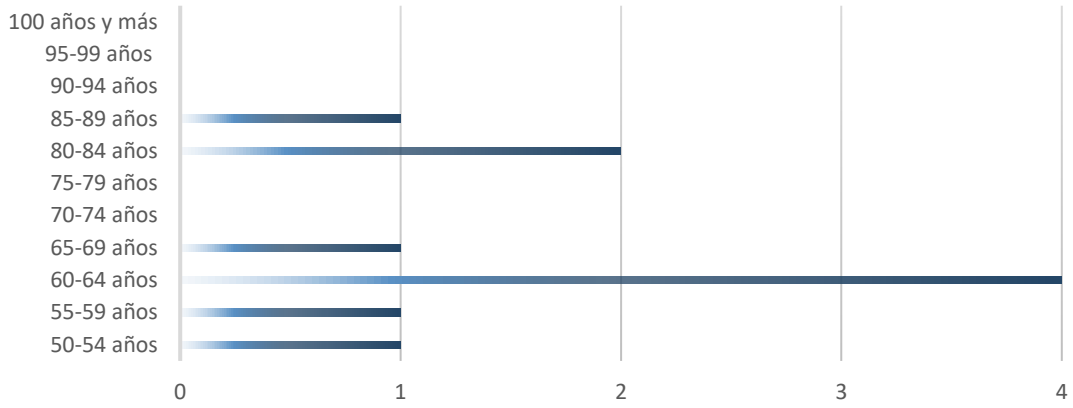


Procedencia de Egreso			
Curación	Mejoría	Defunción	Otro
0	10	0	0

Fuente: Elaboración propia.

La distribución de edades es un tanto más uniforme, principalmente en edades más avanzadas, “adultos mayores”, dado que comprenden edades de 50 a 87 años, con una media de 62 años.

Gráfico 4.24 Distribución de Edades C219



Fuente: Elaboración propia.

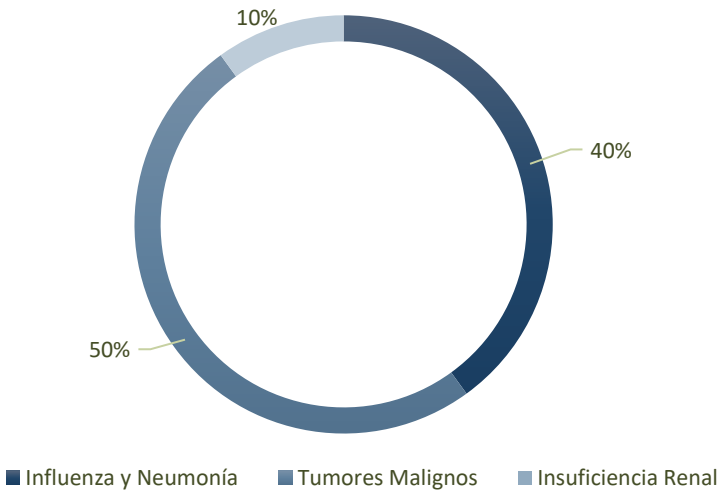
Tabla 4.3 Estadísticas de la Edad para C219

Edad		
Mínimo	**1Q**	**Mediana**
50	62.25	64
Media	**3Q**	**Máximo**
67.7	78.5	87

Fuente: Elaboración propia.

Un dato muy importante es qué tipos de enfermedades causaron estos internamientos, los cuales, en su mayoría, resultan ser tumores malignos; esto tendría relación con que fueran referenciados de otras unidas médicas, dado que esta enfermedad es normalmente atendida en hospitales de alta especialidad, y se da el alta por mejoría en la mayoría de los casos.

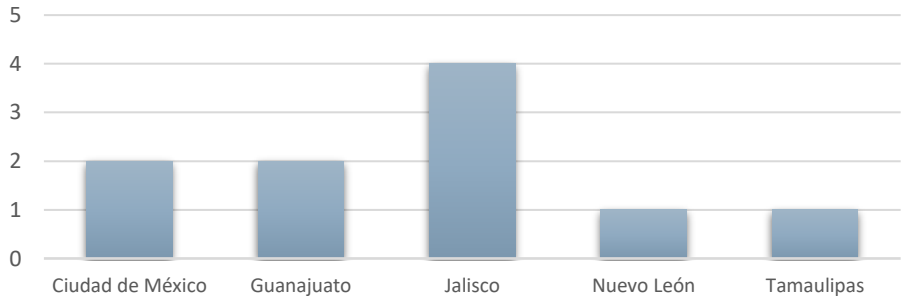
Gráfico 4.25 Porcentajes de Casos Según Enfermedades de C219



Fuente: Elaboración propia.

Al analizar los estados donde se registraron la mayoría de los casos de estas enfermedades, esto es, la Ciudad de México, Nuevo León, Jalisco, Tamaulipas y Guanajuato, nos damos cuenta de que los primeros dos son estados de altos ingresos per cápita.

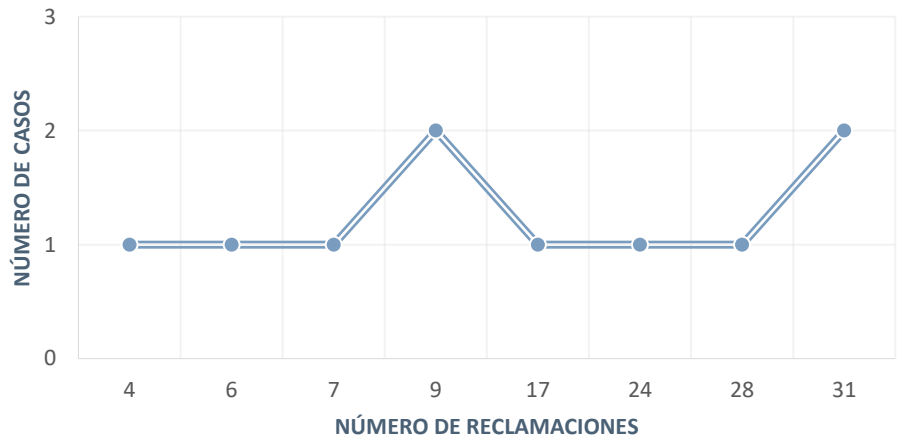
Gráfico 4.26 Estados Donde se Reportan Reclamaciones de C219



Fuente: Elaboración propia.

Los 10 casos registraron una alta demanda en reclamaciones, dado que todos ellos fueron mayores a cuatro reclamaciones, teniendo un pico en 9 y 31 reclamaciones, dando indicios de que dichas enfermedades son recurrentes.

Gráfico 4.27 Número de Casos por Reclamaciones C219



Fuente: Elaboración propia.

Los montos económicos registrados para el segundo clúster del año 2019 representan los datos. A diferencia del clúster C119, aquí se alcanza a ver que el coaseguro y el deducible son más altos, siendo estos los montos más grandes dentro de toda la base de datos.

Tabla 4.4 Información Económica de los Datos C219

Monto de Hospitalización		
Mínimo	**1Q**	**Mediana**
$275,102	$389,350	$579,630
Media	**3Q**	**Máximo**
$750,589	$882,885	$1,657,696
Monto de Honorarios Médicos		
Mínimo	**1Q**	**Mediana**
$0	$0	$0
Media	**3Q**	**Máximo**
$12,391	$2,609	$74,169
Monto de Coaseguro		
Mínimo	**1Q**	**Mediana**
$35,511	$102,594	$133,803
Media	**3Q**	**Máximo**
$175,413	$156,704	$671,296
Monto de Deducible		
Mínimo	**1Q**	**Mediana**
$0	$16,080	$23,157
Media	**3Q**	**Máximo**
$61,624	$32,592	$252,865

Fuente: Elaboración propia.

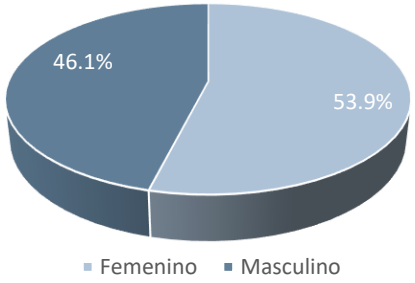
4.5 Descripción de los Clústers del Año 2020

Al pasar al año 2020 y comenzar el EDA, en los dos clústers se muestra todo el conjunto de datos, que suman un total de 1,275, que al dividirse se obtiene una partición de 94.6% de los datos en un solo clúster, mismo que a continuación se describen a detalle.

4.5.1 Clúster 1 Año 2020 (C120)

El clúster llamado C120 cuenta con 1,207 datos que, siendo divididos, tienen una proporción por género del 54% para mujeres y 46% para hombres, cumpliendo casi una paridad y siendo la misma proporción que en 2019. Casi el 80% de los padecimientos fueron internamientos ambulatorios; a su vez, la referencia de otra unidad compone la mayoría de los datos, seguida por el ingreso a urgencias y que la mayoría de los egresos son por mejoría, seguido de algún otro motivo (el cual generalmente es por alta voluntaria):

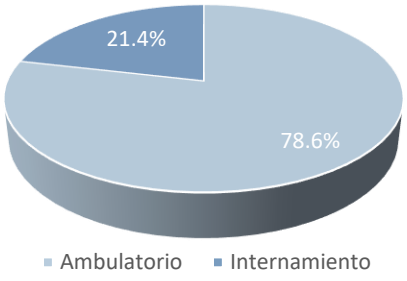
Gráfico 4.28 Datos de Género C120



Género	
Femenino	Masculino
651	556

Fuente: Elaboración propia.

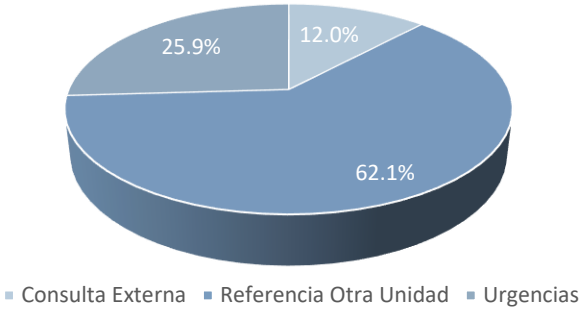
Gráfico 4.29 Datos de Tipo de Internamiento C120



Tipo de Internamiento	
Ambulatorio	Internamiento
949	258

Fuente: Elaboración propia.

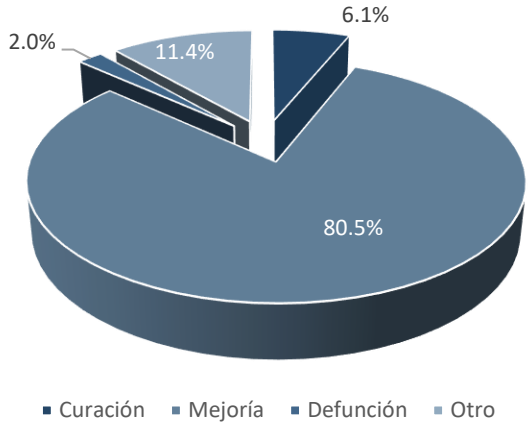
Gráfico 4.30 Datos de Procedencia de Ingreso para C120



Procedencia de Ingreso		
Consulta Externa	Referencia Otra Unidad	Urgencias
145	749	313

Fuente: Elaboración propia.

Gráfico 4.31 Datos de Procedencia de Egreso C120

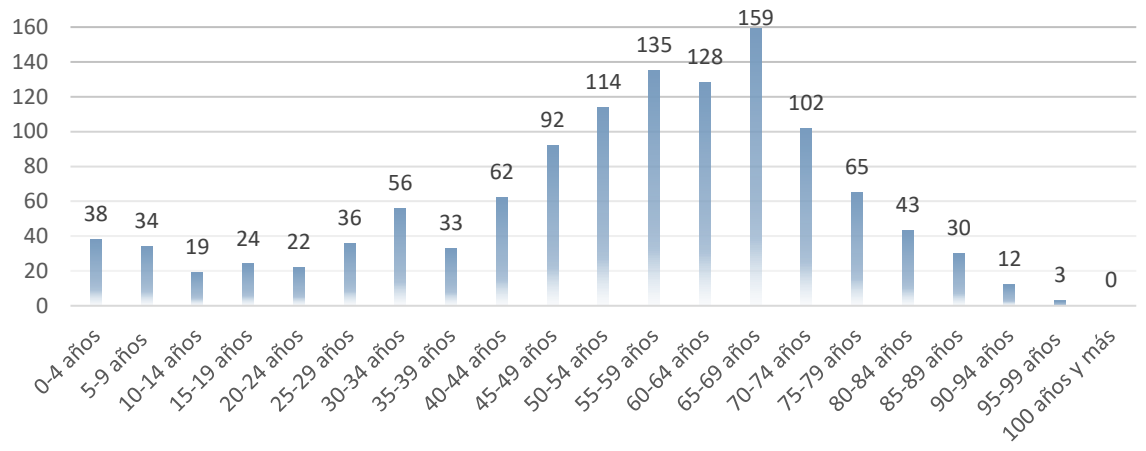


Motivo de Egreso			
Curación	Mejoría	Defunción	Otro
74	972	24	137

Fuente: Elaboración propia.

Al realizar la distribución de edades para este grupo de datos, siguiendo con la división por grupos quinquenales del INEGI, el grafico se verá:

Gráfico 4.32 Distribución de Edades C120



Fuente: Elaboración propia.

Es fácil observar que las edades más afectadas con padecimientos comienzan en el grupo de edades de 40-44 años y llegan hasta los 75-79 años, donde comienza un descenso esperado, dado que la esperanza de vida aproximadamente es de 75-76 años.

Interpretado de otra manera, podemos ver que la media de edad se encuentra en 52.32 años y un máximo de 95 años, lo que nos permite saber que los 3 datos que se encuentran en el grupo quinquenal de 95-99 años, solo están representados por una sola edad.

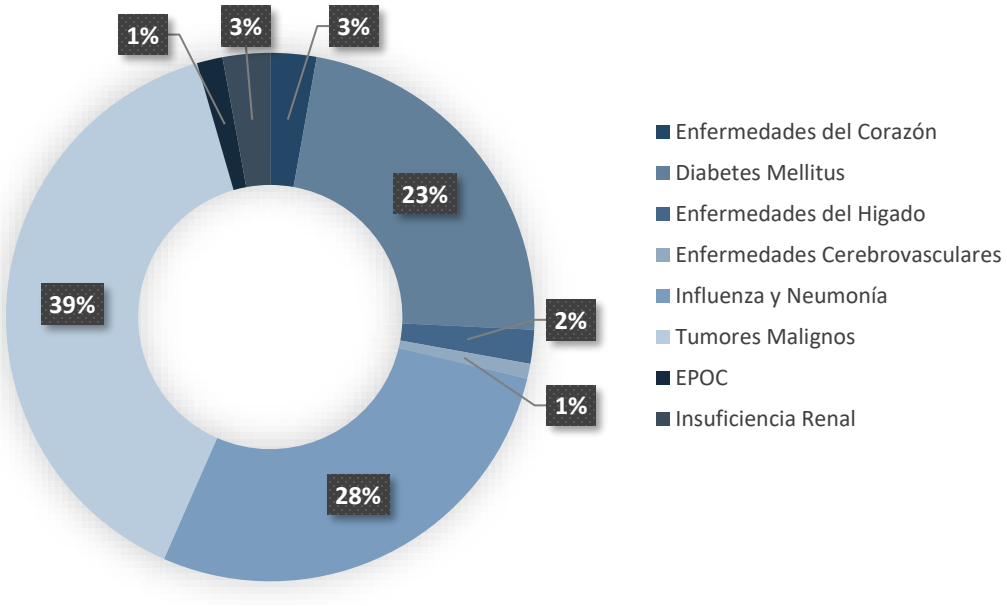
Tabla 4.5 Estadísticas de la Edad para C120

Edad		
Mínimo	**1Q**	**Mediana**
0	43	57
Media	**3Q**	**Máximo**
52.32	68	95

Fuente: Elaboración propia.

Todas las enfermedades están representadas en este conglomerado, pues posee varios datos, los porcentajes se ven de la siguiente manera.

Gráfico 4.33 Porcentajes de Casos Según Enfermedades en C120

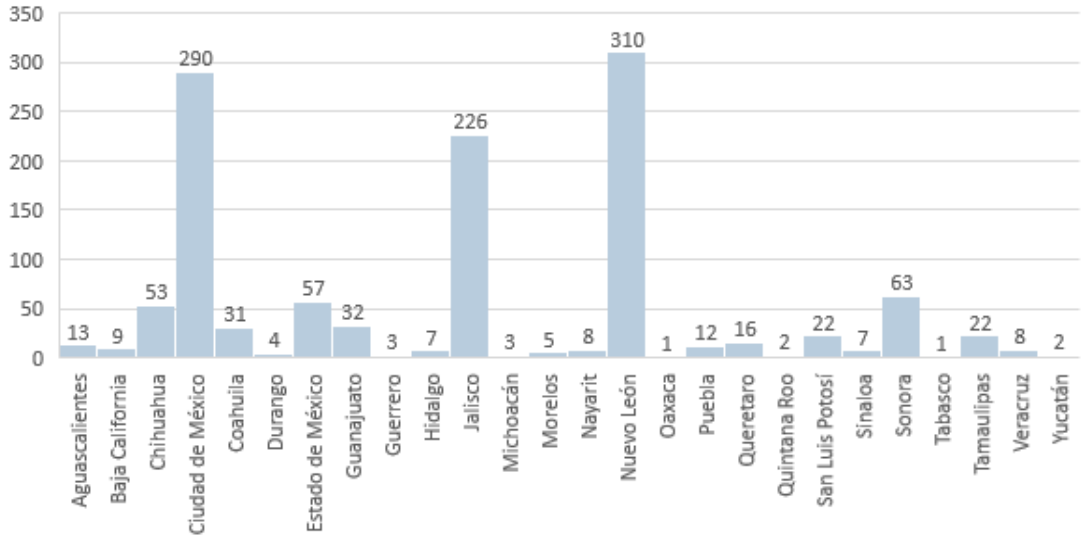


Fuente: Elaboración propia.

Aquí podemos ver que las enfermedades dominantes fueron tumores malignos, influenza y neumonía, seguido de la diabetes mellitus; por otro lado, las enfermedades cerebro vasculares representan apenas el uno por ciento.

Los estados que representan las mayores reclamaciones son Ciudad de México, Jalisco y Nuevo León; Tabasco, por otra parte, es el estado que menos reclamaciones.

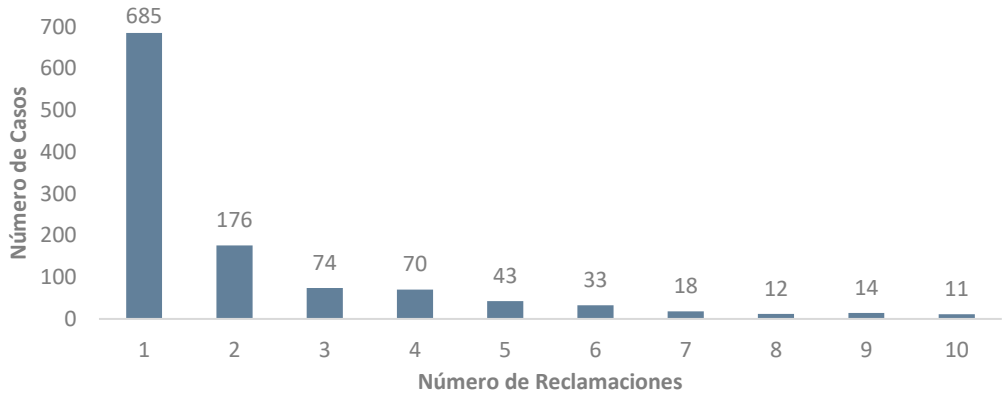
Gráfico 4.34 Estados Donde se Reportan Reclamaciones de C120



Fuente: Elaboración propia.

El número de reclamaciones será dividido para mejor la visualización, representando sólo los primeros diez datos, aquellos que tienen una mayor cantidad de datos .

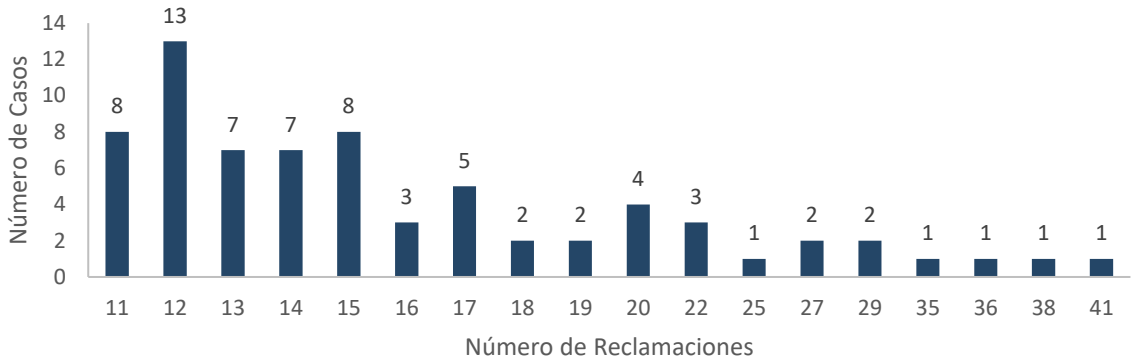
Gráfico 4.35 Número de Casos por Reclamación, Primeros Datos en C120



Fuente: Elaboración propia.

Los datos del clúster C120 tienen concentrados el 56.8% de los casos, con una sola reclamación, mismos que suben hasta 94.1% en las primeras diez reclamaciones.

Gráfico 4.36 Resto de Número de Casos por Reclamación C120



Fuente: Elaboración propia.

A medida que sube el número de reclamaciones, los casos decaen hasta llegar a un caso con 41 reclamaciones.

Pasando a los montos económicos registrados, se observa que lo Montos de Honorarios Médicos, Coaseguro y Deducible están altamente sesgados con cero, y los Montos de Hospitalización también comienzan en cero.

Gráfico 4.37 Información Económica de los Datos de C120

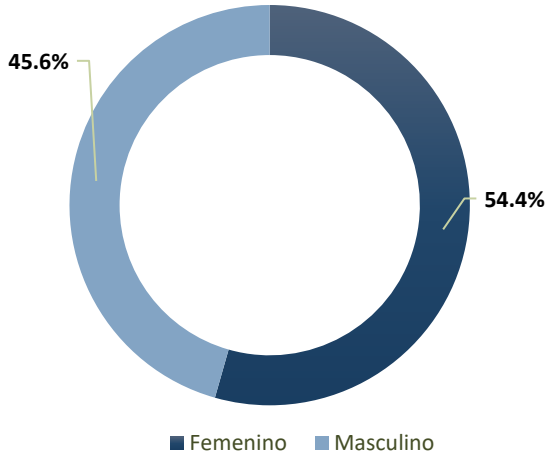
Monto de Hospitalización		
Mínimo	**1Q**	**Mediana**
$18	$3,511	$21,401
Media	**3Q**	**Máximo**
$113,360	$93,128	$3,469,384
Monto de Honorarios Médicos		
Mínimo	**1Q**	**Mediana**
$0	$0	$0
Media	**3Q**	**Máximo**
$6,593	$0	$1,040,328
Monto de Coaseguro		
Mínimo	**1Q**	**Mediana**
$0	$0	$0
Media	**3Q**	**Máximo**
$1,322	$0	$22,234
Monto de Deducible		
Mínimo	**1Q**	**Mediana**
$0	$0	$0
Media	**3Q**	**Máximo**
$1,074	$0	$35,291

Fuente: Elaboración propia.

4.5.2 Clúster 2 Año 2020 (C220)

Finalmente, para terminar de describir los clústers estudiados, el segundo conglomerado del año 2020 consta de 68 datos, equivalentes al 5.3% de la información del año, donde el 54.4% fueron mujeres. Para el tipo de internamiento, las cifras fueron equitativas, siendo marginalmente mayor los casos ambulatorios, y casi todos fueron referenciados de otra unidad médica (98.5%). Por otra parte, el 98.5% de los casos egresaron por mejoría o defunción, siendo esta última la menor de las causas.

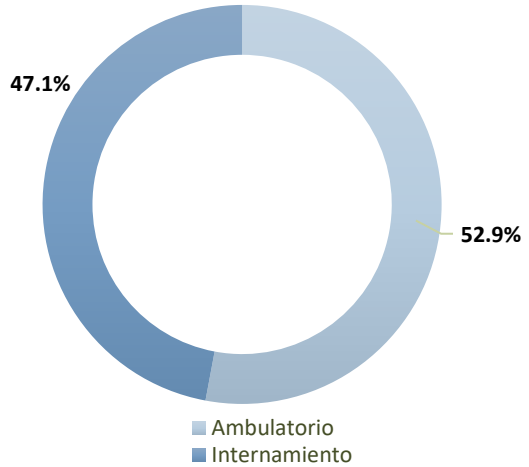
Gráfico 4.38 Datos de Género C220



Género	
Femenino	Masculino
37	31

Fuente: Elaboración propia.

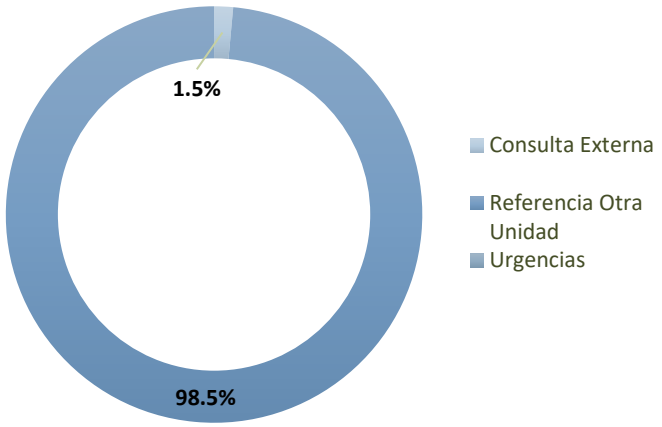
Gráfico 4.39 Datos de Tipo de Internamiento C220



Tipo de Internamiento	
Ambulatorio	Internamiento
36	32

Fuente: Elaboración propia.

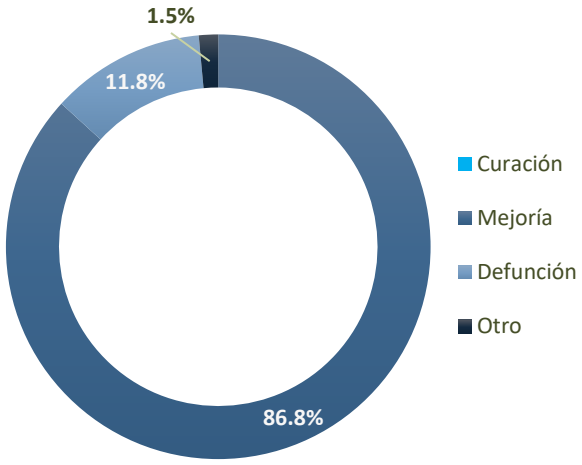
Gráfico 4.40 Datos de Procedencia de Ingreso C220



Procedencia de Ingreso		
Consulta Externa	Referencia Otra Unidad	Urgencias
1	67	0

Fuente: Elaboración propia.

Gráfico 4.41 Datos de Tipo Procedencia de Egreso C220

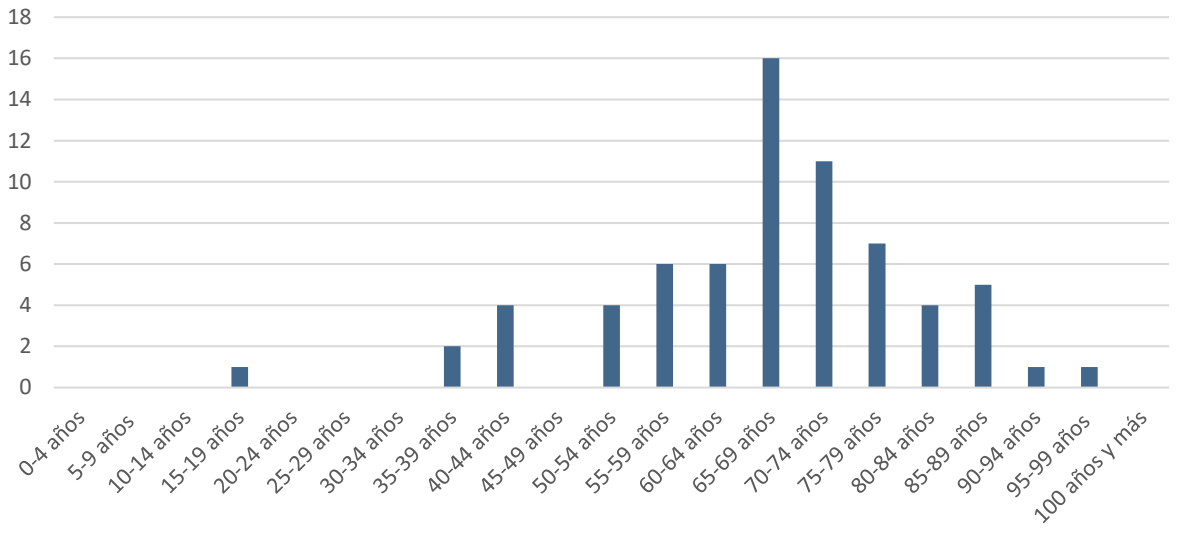


Motivo de Egreso			
Curación	Mejoría	Defunción	Otro
0	59	8	1

Fuente: Elaboración propia.

Al analizar las edades referidas en el clúster, se observa un comportamiento sesgado más hacia las edades avanzadas, ya que de los 50 años en adelante está el 89.7% de la información.

Gráfico 4.42 Distribución de Edades C220



Fuente: Elaboración propia.

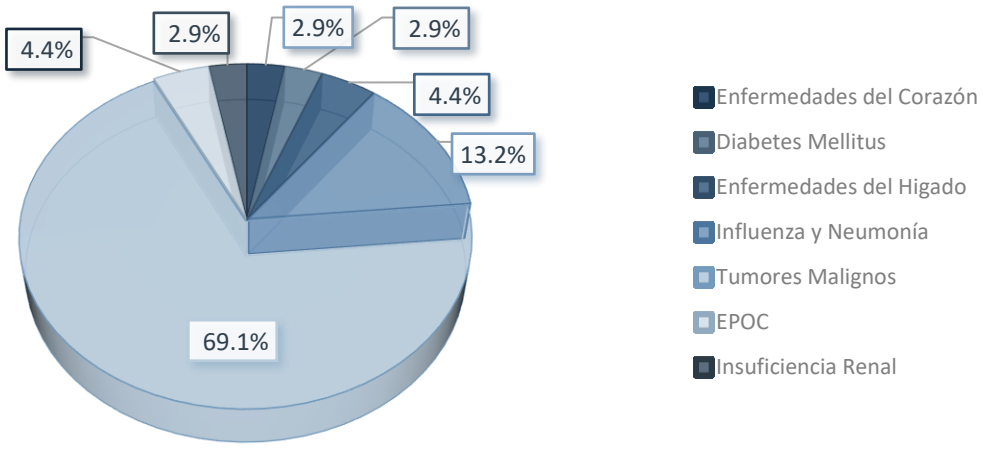
Tabla 4.6 Estadísticas de la Edad para C220

Edad		
Mínimo	**1Q**	**Mediana**
19	59.75	68
Media	**3Q**	**Máximo**
66.65	75	95

Fuente: Elaboración propia.

La tendencia del grafico 4.42 podría ser similar a una normal, a excepción del caso de 15-19 años, que coinciden con la media y mediana en el rango que más casos posee, que es el de 65 a 69 años, y que desciende gradualmente hasta el caso registrado de 95 años. Del mismo modo, la media de edad se sitúa en 66.65 años.

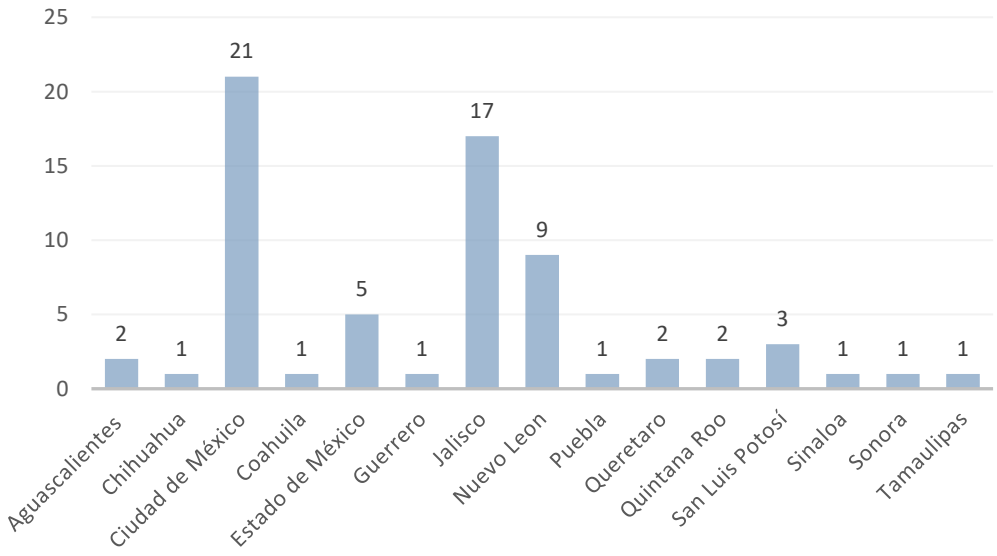
Gráfico 4.43 Porcentaje de Casos Según Enfermedades



Fuente: Elaboración propia.

La mayoría de las enfermedades en este clúster son tumores malignos con 69.1%, seguido de influencia y neumonía (13.2%) y EPOC (4.4%), enfermedades relacionadas con el tracto respiratorio. Contrario a lo observado en C219, donde solo había presencia de 3 enfermedades, en este clúster podemos visualizar todas las enfermedades excepto accidentes cerebrovasculares.

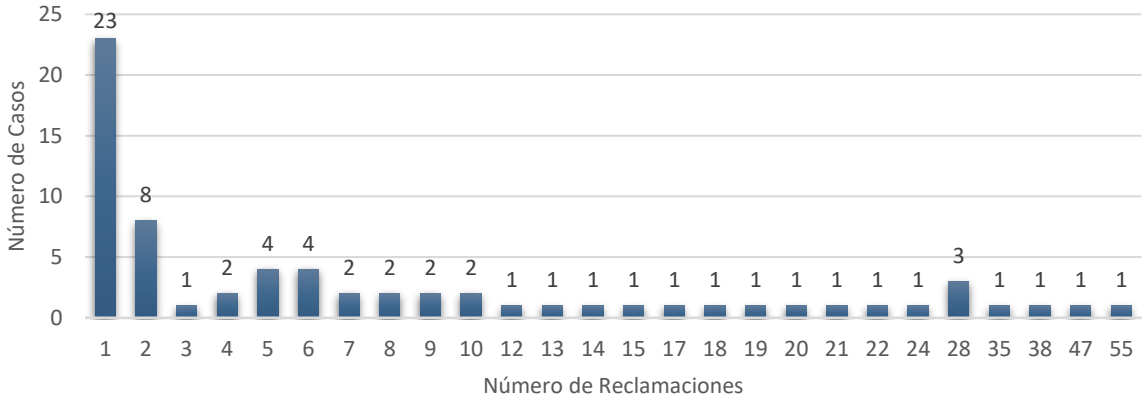
Gráfico 4.44 Estados Donde se Reportan Reclamaciones de C220



Fuente: Elaboración propia.

El gráfico anterior muestra los estados que presentaron reclamaciones este año, donde aún son los principales la Ciudad de México, Jalisco y Nuevo León, lo cual es un comportamiento esperado, dado que son los estados que contienen mayor cantidad de seguros contratados, además de tener mayor población.

Gráfico 4.45 Número de Casos por Reclamaciones



Fuente: Elaboración propia.

La mitad de los casos presentaron entre una y cuatro reclamaciones, y descienden hasta tres casos con 28 reclamaciones y un caso que recurrió a 55 reclamaciones. De este modo, también se observa que entre mayor sea el número de reclamaciones el número de casos que se registra es menor.

Gráfico 4.46 Información Económica de los Datos de C220

Monto de Hospitalización		
Mínimo	**1Q**	**Mediana**
$104,410	$245,520	$405,729
Media	**3Q**	**Máximo**
$624,364	$697,187	$7,981,520
Monto de Honorarios Médicos		
Mínimo	**1Q**	**Mediana**
$0	$0	$0
Media	**3Q**	**Máximo**
$18,561	$29,727	$156,777
Monto de Coaseguro		
Mínimo	**1Q**	**Mediana**
$4,295	$18,077	$30,063
Media	**3Q**	**Máximo**
$37,098	$42,039	$293,502
Monto de Deducible		
Mínimo	**1Q**	**Mediana**
$0	$0	$8,857
Media	**3Q**	**Máximo**
$19,040	$37,340	$77,279

Fuente: : Elaboración propia.

Conclusiones

El estudio realizado para los años 2019 y 2020 representa los resultados de cómo segmentar los Seguros de Gastos Médicos Mayores; pese a la desigualdad del año 2020, se encuentran patrones similares a los observados en 2019. El objetivo principal del trabajo presentado es encontrar la presencia de cambios en la segmentación del número de siniestros hospitalarios registrados en el periodo considerado.

Se presenta la recolección de información sobre los SGMM, su composición y su paso a través de la historia, cómo fue evolucionando de acuerdo con las necesidades de la época hasta alcanzar al siglo XXI; su llegada a México, importancia, comercialización, además se muestra la estructura del mercado.

Partiendo del ámbito general, se encuentra información sintetizada del Análisis Multivariado qué lo compone: dónde nace, quiénes son sus principales impulsores y su estrecha relación con las ciencias biológicas; además de ahondar en las técnicas más importantes y explicar en qué tipo de casos se utiliza cada uno de ellos.

El Análisis Clúster o Análisis de Conglomerados es una herramienta ampliamente utilizada en los análisis exploratorios, de manera particular, en el ámbito asegurador es recurrentemente utilizado como parte de segmentación de clientes al momento de tarifaciones. Incluso, estos son empleados en la mercadotecnia del sector asegurador, donde su importancia radica en encontrar qué tipo de clientes se adecua a algún tipo de seguro. Este tipo de análisis puede encontrar y entender patrones, tendencias del mercado e incluso desafíos; más específicamente, en los SGMM la principal utilidad es encontrar grupos de riesgo a enfermedades similares.

Una vez practicada la metodología, se observó homogeneidad; es decir, existe un clúster que acaparaba la mayor cantidad de información. No obstante, el segundo clúster (que contiene menor cantidad de datos) es aquel que posee información importante para la aseguradora, dado que contiene los incidentes con mayor Monto de Hospitalización. Cabe mencionar que este fenómeno no solo fue observado en 2019, sino que es replicado con mayor intensidad en 2020.

Es importante resaltar el incremento de datos en el clúster número dos entre un año y otro: de manera estadística, para 2019 se tiene 99.7% de la información en el clúster uno, ya que 0.3% de los datos recaían en accidentes más costosos; por otra parte, en lo concerniente al año 2020, los incidentes más costosos se incrementaron al 5.3% de la base de datos, lo que significó un aumento entre 2019 y 2020 del 754.1%, a pesar de que la base 2019 contiene 139 observaciones más que la de 2020. A partir de estas cifras, se puede entender que en 2020 hubo mayor prevalencia en enfermedades costosas.

De tal forma, podemos determinar que no hay evidencia suficiente para rechazar la hipótesis nula (H0); dicho de otro modo, se ha podido comprobar la hipótesis planteada en la investigación del presente trabajo, la cual esperaba que la pandemia causada por el SARS-CoV-2 tuviera un impacto en la composición del clústers, entre un año (2019) y otro (2020).

En la ciencia actuarial, el análisis clúster comúnmente es empleado para el marketing de los productos actuariales y/o elaboración de tarifación, pero, yendo más allá, encontramos que, al hacer una segregación dentro de cada clúster, podemos hallar estructuras de dependencia; es decir, las compañías de seguros normalmente buscan encontrar dependencias de riesgos y los efectos que tienen en aquellos eventos de la pérdida múltiple, esto para poder evaluar riesgos simultáneamente.

En este sentido, el análisis clúster en el sector asegurador puede estudiar a aquellos asegurados que están más propensos a realizar mayor número de reclamaciones, por más “pequeñas” que estas sean (por montos menores), y diferenciarlos de los asegurados que representan riesgos más altos en reclamaciones. Entonces, se sabe que el riesgo es dependiente, heterogéneo multifactorial y varía de acuerdo con el asegurado. En el caso particular de este trabajo, el objeto de estudio fue el seguro de gastos médicos mayores, el cual presentó heterogeneidad de riesgo porque el asegurado fue susceptible a ciertos padecimientos, incluidos los empleados en este estudio).

Así también, los seguros se agrupan en riesgos similares, de forma que es posible identificar homogeneidad en el riesgo; además, mediante el análisis clúster, es posible llegar a formar grupos con riesgos homogéneos que a simple vista no son perceptibles; y más aún, existe la posibilidad de que eventualmente sea posible realizar una “predicción”, puesto que la próxima vez que un actuario se enfrente a determinados casos, podrá diferenciar a qué grupo (clúster) de riesgo pertenece.

Futuras áreas de investigación

Una línea de investigación derivada de este trabajo correspondería a un análisis más profundo de cada clúster determinado y, más aún, poder expandir este trabajo hasta el final de la emergencia y encontrar el comportamiento de los SGMM con el paso de la enfermedad del COVID-19.

Con respecto a la elección de variables y base de datos, se puede mencionar que no existe, hasta el corte de esta investigación, alguna institución o entidad que posea para consulta libre datos similares a los empleados; sin embargo, dentro de alguna compañía podría replicarse el estudio y contrastar los resultados obtenidos aquí.

Otro aspecto que podría contemplarse para futuros estudios sería un mejor filtrado de los datos; es decir, tratar de que aquellos datos que son igual a cero puedan ser sustituidos o eliminados del estudio porque representan dificultad al momento de la clusterización, ya que estos tienden a emigrar entre clústers dada la combinación lineal empleada y, sin importar el número de iteraciones, estos siguen móviles.

Finalmente, sería interesante emplear este método o alguno similar de conglomerados en diversas ramas de los seguros. No obstante, podría tomarse una línea de acción muy importante dentro de cada clúster para poder llegar a emplear un método de *forecasting* que, a largo plazo y con varias pruebas, pudiese ser empleado en el sector asegurador.

Anexos

```
library(readxl)
library(dplyr)
library(stringr)
library(cluster)
library(NbClust)
library(ggplot2)
library(factoextra)
library(ClusterR)
library(psych)

Salud_2019 <- read_excel("C:/Users/paola lizeth/Desktop/TESIS/Enfermedades de Mayor Muerte en Mexico.xlsx",
                         sheet = "2019")
Salud_2019<-na.omit(Salud_2019)

base<-as_tibble(Salud_2019) %>% mutate(EDAD=as.numeric(EDAD)) %>%
  filter(`MONTO DE HOSPITALIZACION`>=0,
    `MONTO HONORARIOS MEDICOS`>=0)
base<-na.omit(base)

numericas<-base %>% select(EDAD,`MONTO DE HOSPITALIZACION`,`MONTO HONORARIOS MEDICOS`,
                           `MONTO DE DEDUCIBLE O COPAGO`, `MONTO COASEGURO`)
colnames(numericas)<-c("Edad","Monto_Hosp","Monto_Hon","Monto_Ded","Monto_Coa")

datos_cluster<-numericas %>% mutate(Edad=scale(Edad),
                                    Monto_Hosp=scale(Monto_Hosp),
                                    Monto_Hon=scale(Monto_Hon))

num_cluster<-NbClust(data=datos_cluster,distance = "euclidean",
                     method = "kmeans", index = "all")


Modelo1<- kmeans(x=datos_cluster,centers = 2)
Modelo2<- kmeans(x=datos_cluster,centers = 3)
```

```
library(readxl)
library(dplyr)
library(stringr)
library(cluster)
library(NbClust)
library(ggplot2)
library(factoextra)
library(ClusterR)
library(psych)

Salud_2020 <- read_excel("C:/Users/paola lizeth/Desktop/TESIS/Enfermedades de Mayor Muerte en Mexico.xlsx",
                         sheet = "2020")
Salud_2020<-na.omit(Salud_2020)

base<-as_tibble(Salud_2020) %>% mutate(EDAD=as.numeric(EDAD)) %>%
  filter(`MONTO DE HOSPITALIZACION`>=0,
    `MONTO HONORARIOS MEDICOS`>=0)
base<-na.omit(base)

numericas<-base %>% select(EDAD,`MONTO DE HOSPITALIZACION`,`MONTO HONORARIOS MEDICOS`,
                           `MONTO DE DEDUCIBLE O COPAGO`, `MONTO COASEGURO`)
colnames(numericas)<-c("Edad","Monto_Hosp","Monto_Hon","Monto_Ded","Monto_Coa")

datos_cluster<-numericas %>% mutate(Edad=scale(Edad),
                                    Monto_Hosp=scale(Monto_Hosp),
                                    Monto_Hon=scale(Monto_Hon))

num_cluster<-NbClust(data=datos_cluster,distance = "euclidean",
                     method = "kmeans", index = "all")


Modelo1<- kmeans(x=datos_cluster,centers = 2)
Modelo2<- kmeans(x=datos_cluster,centers = 4)
```

Bibliografía

American Kidney Fund. (2022). *La Falla Renal o Insuficiencia Renal Terminal (IRT)*. Obtenido de Kidney Fund: https://www.kidneyfund.org/es/todo-sobre-los-rinones/la-falla-renal-o-enfermedad-renal-terminal-ert#qu-es-la-falla-renal

Blanco García, S., & Carvajal Molina, P. (2001). *Antecedentes del Seguro a Prima.* Obtenido de https://eprints.ucm.es/id/eprint/6728/1/0104.pdf

Bramardi, S. J. (2002). *Análisis Multivariado. Su Aplicación en la Caracterización de Recursos Genéticos.* Argentina: Universidad Canahue.

Comisión Nacional de Seguros y Fianzas. (2017). *Presenta CONDUSEF Simulador de Gastos Médicos Mayores.* Obtenido de https://www.condusef.gob.mx/?p=contenido&idc=544&idcat=1#:~:text=De%20acuerdo%20con%20datos%20del,una%20p%C3%B3liza%20de%20gastos%20m%C3%A9dicos

Comisión Nacional de Seguros y Fianzas. (2020). *Penetración del Seguro en México.* Obtenido de Secretaria de Hacienda y Credito Público: https://www.gob.mx/cms/uploads/attachment/file/656301/Comparativo_Penetracion_Seguros_M_xico__jul_2021.pdf

Comisión Nacional de Seguros y Fianzas. (2021). *Información Estadística de Salud*. Obtenido de CNSF: http://www.cnsf.gob.mx/EntidadesSupervisadas/InstitucionesSociedadesMutualistas/Paginas/Salud.aspx

Comisión Nacional de Seguros y Fianzas. (2022). *Información Estadística de Salud*. Obtenido de CNSF: https://www.cnsf.gob.mx/EntidadesSupervisadas/InstitucionesSociedadesMutualistas/Paginas/Salud.aspx

Comisión Nacional para la Protección y Defensa de los Usuarios Financieros. (2017). *Presenta CONDUSEF Simulador de Gastos Médicos Mayores*. Obtenido de CONDUSEF: https://www.gob.mx/cms/uploads/attachment/file/652863/Analisis_Estadistico_de_Gastos_Medicos_2020.pdf

Consultoría Estratégica de Investigación de Mercados. (2020). *Análisis Multivariable. Enfoque Práctico en Investigación Cuantitativa*. Obtenido de CIMEC: https://www.cimec.es/analisis-multivariable-investigacion-cuantitativa/

De la Fuente Fernández, S. (2011). *Análisis Conglomerados.* Obtenido de https://www.academia.edu/32046069/An%C3%A1lisis_Conglomerados_Santiago_de_la_Fuente_Fern%C3%A1ndez

De la Rosa Analytics Solutions. (2021). *Resultado del Sector Asegurador Mexicano 2020.* Obtenido de Asociación Mexicana de Agentes de Seguros y Fianzas: https://amasfac.info/boletines/anexos/ANALISIS%20Resultados%20SectorAseguradorMexicano%202020%20(DRASolutions,Abril2021)V5.pdf

Deance Rupit, V. M., & Osorio López, H. A. (2004). *Modelación de Portafolios de Seguros de Vida Individual.* Obtenido de Universidad de las Américas Puebla: http://catarina.udlap.mx/u_dl_a/tales/documentos/lat/deance_r_vm/capitulo_2.html

Federacion Mexicana de Diabetes. (2019). *Los Costos de la Diabetes*. Obtenido de https://fmdiabetes.org/los-costos-la-diabetes/

Flamand, L. M. (2021). *Cáncer y Desigualdades Sociales en México.* Obtenido de Colegio de México: https://desigualdades.colmex.mx/cancer/informe-cancer-desigualdades-2020.pdf

Forbes Staff. (2016). *Diabetes en México cuesta 3,872 mdd*. Obtenido de https://www.forbes.com.mx/diabetes-mexico-cuesta-3872-mdd/amp/

Fundación MAPFRE. (2017). *Diccionario MAPFRE de Seguros*. Obtenido de Fundación MAPFRE: https://www.fundacionmapfre.org/publicaciones/diccionario-mapfre-seguros/seguro/

Fundación para el Estudio de las Hepatitis Virales. (2019). *Importancia de las Enfermedades del Hígado en el Mundo*. Obtenido de FEHV: https://fehv.org/enfermedades-higado-en-el-mundo/

Galbiati R., J. (2006). *Analisis de Conglomerados.* Obtenido de Jorge Galbiati: https://www.jorgegalbiati.cl/mayo_06/Conglomerados.pdf

García Hernandez, G., & González Segura, |. (2015). *Reconocimiento de Imágenes Dactilares por Medio de Transformada Wavelet Daubechies y Distancias Euclidianas.* Obtenido de Tesis IPN: https://tesis.ipn.mx/bitstream/handle/123456789/22119/tesis_GGH_GGS.pdf?sequence=1&isAllowed=y

Garcia Lopes, H. E., & de Selviha Gosling, M. (2020). *Cluster Analysis in Practice: Dealing with Outliers in Managerial Research.* Obtenido de Redalyc: https://www.redalyc.org/journal/840/84064925007/html/

Garduño, M. (2021). *1 de Cada 4 Mexicanos Sufre un Infarto Cerebral, ¿Cómo Identificarlo y Prevenirlo?* Obtenido de https://www.forbes.com.mx/noticias-1-de-cada-4-mexicanos-infarto-cerebral-como-identificarlo-y-prevenirlo/amp/

Hair Jr., J. B. (1995). Análisis Cluster. En J. B. Hair Jr., *Análisis Multivariante* (págs. 491-519). Madrid: Prentice Hall Iberia.

Humaira, H., & Rasyidah, R. (2018). *Determining The Appropiate Cluster Number Using Elbow.* Obtenido de ResearchGate: https://www.researchgate.net/publication/339670247_Determining_The_Appropiate_Cluster_Number_Using_Elbow_Method_for_K-Means_Algorithm

Humberto, C. A. (2013). *Análisis Multivariante, Conceptos y Aplicaciones en Psicología Educativa y Psicometría.* Obtenido de Redalyc: https://www.redalyc.org/pdf/259/25930006005.pdf

Instituto Nacional de Cáncer. (2021). *Estadísticas del Cáncer*. Obtenido de NIH: https://www.cancer.gov/espanol/cancer/naturaleza/estadisticas#:~:text=El%20c%C3%A1ncer%20es%20una%20de,c%C3%A1ncer%20a%2016%2C4%20millones.

Instituto Nacional de Enfermedades Respiratorias. (2017). *Clínica EPOC*. Obtenido de INER: http://www.iner.salud.gob.mx/interna/tabaquismo-clinEPOC.html#:~:text=En%20M%C3%A9xico%2C%20tan%20solo%20en,igual%20entre%20hombres%20y%20mujeres.

Instituto Nacional de Estadística y Geografía. (2020). *Características de las Defunciones Registradas en México Durante 2019.* Obtenido de INEGI: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2020/EstSociodemo/DefuncionesRegistradas2019.pdf

Instituto Nacional de Estadística y Geografía. (2021). *Características de las Defunciones Registradas en México Durante 2020*. Obtenido de INEGI: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/EstSociodemo/DefuncionesRegistradas2020preliminar.pdf

Instituto Nacional de Salud Pública. (2020). *La Enfermedad Renal Crónia en México*. Obtenido de INSP: https://www.insp.mx/avisos/5296-enfermedad-renal-cronica-mexico.html

James, Gareth, et al. (2013). *An Introduction to Statistical Learning.* Obtenido de Static : https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf

Kassambara, A. (2017). *Cluster Validation Essentials*. Obtenido de Determining The Optimal Number Of Clusters: 3 Must Know Methods: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

Lozares Colina, C., & López Roldán, P. (1991). *El Análisis Multivariado: Definición, Criterios y Clasificación.* Obtenido de Academia, Departamento de Sociologia: https://www.academia.edu/7218997/Analisis_Multivariado

M.H., B., & Castillo J. & Wong, A. (2008). *Uso de Análisis de Covarianza (ANCOVA) en Investigación Cientifica.* Obtenido de Innovaciones de Negocios: http://eprints.uanl.mx/12489/1/A3.pdf

Magee, J. H. (1947). *Seguros Generales.* México: Unión Tipografica: Hispano-Americana.

Martinez de Lejarza, I., & Martinez de Lejarza, J. (1995). *Introducción al Análisis Clúster*. Obtenido de Universidad de Valencia: https://www.uv.es/ceaces/multivari/cluster/CLUSTER2.htm#:~:text=La%20diferencia%20fundamental%20entre%20el,en%20qu%C3%A9%20medida%20las%20variables

Mehr, R. I., & Osler, R. W. (1994). *Modern Life Insurance.* California: The Macmillan Company.

Minzoni Consorti, A. (2005). *Crónicas de Dos Siglos del Seguro en México.* Obtenido de Comisión Nacional de Seguros y Fianzas: https://www.gob.mx/cms/uploads/attachment/file/74253/Cronica_de_dos_siglos_del_seguro_en_Mexico.pdf

Mohajer, M., Englmeier, K., & Schmid, V. (2010). *A comparison of Gap Statistic Definitions With and With-Out Logarithm Function.* Obtenido de Core: https://core.ac.uk/download/pdf/12172514.pdf

Moreno Madueño, J. (2016). *Introducción al Análisis Multivariado.* Obtenido de Universidad de Granada: https://www.academia.edu/20421132/5466678_INTRODUCCION_AL_ANALISIS_MULTIVARIADO

Nieto Barajas, L. E. (2020). *Análisis Multivariado.* Obtenido de Instituto Tecnológico Autónomo de México: http://allman.rhon.itam.mx/~lnieto/index_archivos/Modulo61.pdf

Ocaña Peinado, F. M. (2020). *Técnicas Estadísticas en Nutrición y Salud. Tratamiento Estadístico de Outliers y Datos Faltantes.* Obtenido de Universidad de Granada: https://www.ugr.es/~fmocan/MATERIALES%20DOCTORADO/Tratamiento%20de%20outliers%20y%20missing.pdf

Organizacion Mundial de la Salud. (2017). *Enfermedades Cardiovasculares*. Obtenido de OMS: https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Organización Mundial de la Salud. (2021). *Neumonía*. Obtenido de OMS: https://www.who.int/es/news-room/fact-sheets/detail/pneumonia

Organización Mundial de la Salud. (2022). *Cáncer*. Obtenido de OMS: https://www.who.int/es/news-room/fact-sheets/detail/cancer#:~:text=%C2%ABC%C3%A1ncer%C2%BB%20es%20un%20t%C3%A9rmino%20gen%C3%A9rico,%C2%BB%20o%20%C2%ABneoplasias%20malignas%C2%BB.

Organizacion Mundial de la Salud. (2022). *Diabetes*. Obtenido de OMS: https://www.who.int/es/news-room/fact-sheets/detail/diabetes

Organización Mundial de la Salud. (2022). *Enfermedad Pulmonar Obstructiva Crónica (EPOC)*. Obtenido de OMS: https://www.who.int/es/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

Organización Panamericana de la Salud. (2021). *Influenza y Otros Virus Respiratorios*. Obtenido de OPS: https://www.paho.org/es/temas/influenza-otros-virus-respiratorios

Prieto Guerra, R. E. (2006). *Técnicas Estadísticas de la Clasificación, un Ejemplo de Análisis Clúster*. Obtenido de http://dgsa.uaeh.edu.mx:8080/bibliotecadigital/handle/231104/1580

R Documentation. (2022). *NbClust: NbClust Package for Determining The Best Number of Clusters*. Obtenido de R Documentation: https://www.rdocumentation.org/packages/NbClust/versions/3.0.1

Rendon Lara, E., & Abundez Barrera, I. M. (2016). *RENTOL: Un algoritmo de Agrupamiento Basado en K-Means.* Obtenido de Research in Computing Science IPN: https://rcs.cic.ipn.mx/2016_128/RENTOL_%20Un%20algoritmo%20de%20agrupamiento%20basado%20en%20K-means.pdf

Romero Gatica, H. (2017). *Análisis descriptivo de la estructura del sector.* Obtenido de Comisión Nacional de Seguros y Fianzas: https://www.gob.mx/cms/uploads/attachment/file/293680/166._Ana_lisis_estructura_del_sector_asegurador_1994_2017.pdf

Rousseeuw, P. J. (1986). *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.* Obtenido de Kuleuven: https://wis.kuleuven.be/stat/robust/papers/publications-1987/rousseeuw-silhouettes-jcam-sciencedirectopenarchiv.pdf

Sagaró del Campo, N. M., & Zamora Matamoros, L. (2020). *Técnicas Estadísticas Multivariadas para el Estudio de la Causalidad en Medicina*. Obtenido de Revista de Ciencias Médicas de Pinar del Río:

http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1561-31942020000200287

Schiaffino, S. (2018). *Clustering.* Obtenido de Unicen: https://users.exa.unicen.edu.ar/catedras/optia/public_html/clustering.pdf

Schumacher, P. (2019). *Los Seguros Como Herrramienta de la Gestión de Riesgos*. Obtenido de Risk Prevention de México: https://riskp.com/los-seguros-como-herramienta-de-la-gestion-de-riesgos/

Trueba Espinosa, A. (2017). *Procesos de Análisis No Supervisado.* Obtenido de UAEMéx: http://ri.uaemex.mx/bitstream/handle/20.500.11799/70383/secme-16823_1.pdf?sequence=1

Villareal-Rios E., P. A. (2020). *Costo Institucional del Paciente con Enfermedad Renal Crónica Manejada con Hemodiálisis.* Obtenido de IMSS: https://www.redalyc.org/journal/4577/457769357009/html/

Wang, F., Franco, H., & Kelleher, J. &. (2017). *An Analysis of the Application of Simplified Silhouette to the Evaluation of K-Means.* Obtenido de Research Gate: https://www.researchgate.net/publication/318109824_An_Analysis_of_the_Application_of_Simplified_Silhouette_to_the_Evaluation_of_k-means_Clustering_Validity

World Heart Federation. (2016). *El Costo de las Enfermedades Cardiacas en América Latina.* Obtenido de World Congress of Cardiology & Cardiovascular Health: https://world-heart-federation.org/wp-content/uploads/2017/05/spanish-press-release.pdf