Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Análisis y predicción de la cantidad de fallecidos por causa y provincia

Implementación de modelos predictivos sobre los datos públicos del Ministerio de Salud Argentino 2005-2019

AUTOR: SOLANGE R.E. FRANCO
MENTORA: NÉLIDA MÓNICA CANTONI RABOLINI

DICIEMBRE 2022

Resumen

La administración de recursos de la Nación Argentina es un aspecto vital para el correcto funcionamiento de los organismos dependientes que deben dar sustento y garantía a los derechos sociales establecidos en el marco normativo de la Argentina. Entre los deberes del estado se destaca el derecho a la salud, mencionado de manera implícita en el artículo 33 de la Constitución Nacional (CN).

Dado el presente y constante desarrollo de herramientas que facilitan la gestión y análisis de datos, se plantea una propuesta de metodologías de estudio alternativa a las que realiza actualmente el Ministerio de Salud. Se pretende aportar una mejora mediante el uso de herramientas de análisis multivariado y métodos predictivos, para que el organismo pueda obtener una visión ordenada y clara de las problemáticas que afronta.

El objetivo del presente trabajo es realizar una ponderación de las principales causas de fallecimiento en la Argentina partiendo de una base de datos recolectada e informada por el Ministerio de Salud. Mediante ello, aplicar metodologías de análisis de componentes principales y de clústeres para analizar la correlación entre las variables conformantes y brindar visibilidad sobre los datos más relevantes, su correlación y agrupación geográfica. Adicionalmente, se plantea realizar la predicción razonable de la cantidad de fallecimientos por año con visibilidad por causa, grupo etario, sexo, y a su vez por ubicación geográfica de las provincias de la Argentina. Mediante las metodologías planteadas se pretende ofrecer al Ministerio de Salud un panorama que facilite la toma de decisiones, con el propósito de que la asignación de recursos y elaboración de proyectos en torno a la salud social sea acorde a las necesidades o situaciones futuras.

Para abordar este objetivo, se propone adquirir conocimiento a través de los datos históricos reportados para el rango de períodos 2005 a 2019 por el Ministerio de Salud. En la etapa de procesamiento de datos se aplicarán técnicas de manipulación y limpieza de datos. Luego se realizará el análisis multivariado utilizando la metodología de análisis de componentes principales para reducir la dimensionalidad de los datos, posteriormente se aplicarán diferentes modelos analíticos predictivos que ofrecen las técnicas de aprendizaje automático.

Palabras clave: Causas de fallecimiento - Ministerio de Salud –Análisis de componentes principales - Modelos analíticos predictivos.

Índice

Introducción

El Ministerio de Salud es un organismo público de la República Argentina encargado de atender las cuestiones administrativas relacionadas con el servicio de salud. Donde, dentro de sus amplios deberes se encuentra el deber de entendimiento en la actualización de las estadísticas de salud y los estudios de recursos disponibles, oferta, demanda y necesidad. Así como también, el deber de dar un diagnóstico de la situación necesaria para la planificación estratégica del sector salud.

En el presente trabajo se plantearán metodologías cuantitativas de análisis de datos en el contexto del Ministerio de Salud como organización pública. La finalidad es ampliar y profundizar la materia de conocimiento actual mediante una perspectiva diferente. De modo que, se logre abordar de manera más eficiente las situaciones y necesidades que la población y el sistema afrontan.

Se propone para desarrollar el mismo, llevar a cabo una ponderación de las principales causas de fallecimiento en la Argentina. A través de la metodología de análisis de componentes principales partiendo de la base de datos reportada por el Ministerio de Salud para el período 2019. Así como también, se realizará la utilización de un mayor rango de datos históricos el cual comprenderá desde el año 2005 hasta 2019, con la finalidad de obtener una predicción futura que contemple la mayor cantidad de registros posibles, mediante la técnica de análisis de aprendizaje automático. Se pretende lograr profundidad en el análisis como también ofrecer un panorama que facilite la toma de decisiones de la organización, que podrá resultar en una oportunidad para la administración de recursos.

Metodologías y análisis estadístico actual del Ministerio de Salud.

El presente apartado introduce el tópico a través de una contextualización detallada respecto de las actuales metodologías de análisis que lleva a cabo la respectiva Dirección de Estadísticas e Información de Salud del Ministerio de Salud sobre las causas y casos de fallecimiento recabados en la Argentina para el período 2019. Se presenta el estado actual del conocimiento y se explica cómo se realizará la evaluación de los datos con los que cuenta el organismo. Así como también, se planteará la propuesta de análisis desde otra perspectiva analítica y metodológica.

La información para desarrollar dicha contextualización fue producto de la investigación de los diferentes reportes del organismo, entre ellos se destaca el reporte de publicación anual "Estadísticas Vitales-Información Básica" desarrollados por la Dirección de Estadísticas e Información de Salud del Ministerio de Salud (DEIS). Los boletines informan acerca de las estadísticas sobre hechos vitales para el total del país; nacimientos, defunciones, defunciones fetales, entre otros. Los datos que serán objeto de estudio corresponden a información recopilada por la (DEIS) con apego a las normas internacionales vigentes de la Red Latinoamericana y del Caribe para el Fortalecimiento de los Sistemas de Información de Salud (RELACSIS).

La disposición y el tipo de registros que componen la base de datos elegida, fueron considerados apropiados para poder desarrollar y aplicar las metodologías que se proponen en el presente trabajo de investigación. Las cuales corresponden al método multivariado de Análisis de Componentes Principales (ACP) y la implementación de diferentes métodos de tipo predictivo mediante las técnicas de Aprendizaje Automático para evaluar cuál es el más adecuado para el presente caso. Entre las técnicas que se utilizarán para su evaluación se encuentran distintas técnicas de conjunto (ensamble) de árboles de decisión. De las mencionadas se seleccionará la que se considere más apropiada, considerando para ello la raíz cuadrática media como la métrica de referencia.

1.1 La organización; consideraciones sobre el procesamiento y el análisis que realiza actualmente sobre de los datos.

El Ministerio de Salud (MSAL) como organismo público, posee deberes de vital importancia para brindarle a la población argentina la garantía de cumplimiento de uno de sus derechos fundamentales, que es el acceso a la salud. El MSAL lleva a cabo cuestiones administrativas relacionadas con el servicio de salud, pero a su vez, dentro de los deberes que le competen, se encuentra el entendimiento en la actualización de las estadísticas de salud y los estudios de recursos disponibles, oferta, demanda y necesidad, así como el diagnóstico de la situación necesaria para la planificación estratégica del sector salud.

Para ello el Ministerio de Salud cuenta con una dirección denominada Dirección de Estadísticas e Información de Salud (DEIS). La misma tiene un plan de publicaciones que incluye Boletines y Series que se editan ininterrumpidamente desde 1984. Entre ellos se encuentran reportes sobre consideraciones sobre el procesamiento de los datos, definiciones, conceptos e indicadores que han utilizado sobre información histórica recopilada. Respecto a los indicadores se encuentran los de natalidad, fecundidad y mortalidad infantil por jurisdicción de residencia, entre otros.

Por otra parte, el Anuario de Estadísticas Vitales, denominado “Estadísticas Vitales- Información Básica” (Guevel, 2021), y corresponde a la publicación anual de la DEIS que permite la difusión en un documento consolidado de la información estadística sobre los hechos vitales –nacimientos, defunciones, defunciones fetales y matrimonios- ocurridos en la República Argentina.

La obtención de los datos que dan lugar a la publicación de dicho anuario supone el cumplimiento de al menos tres etapas siendo a nivel local, jurisdiccional y nacional.

A nivel local, el personal de salud de los establecimientos certifica los hechos y capta los datos básicos a partir de los instrumentos de recolección de datos normalizados. Los registros civiles y sus delegaciones inscriben y registran legalmente los hechos vitales. A ellos compete, además, la recopilación y transmisión de los datos al nivel jurisdiccional.

En el nivel jurisdiccional, las unidades de Estadísticas Vitales y de Salud de las jurisdicciones provinciales y la Ciudad Autónoma de Buenos Aires realizan la recepción, el

control, la codificación, el ingreso y la elaboración de los datos, suministrando anualmente los archivos al nivel nacional.

A nivel nacional, se cuenta Dirección de Estadística e Información en Salud (DEIS), como responsable del Sistema Estadístico de Salud (SES). Ésta es la encargada de elaborar las estadísticas sobre hechos vitales para el total del país. También interviene en la normalización de todos los procesos que hacen a la producción de información. Asimismo, publica y difunde información de interés nacional, teniendo en cuenta recomendaciones internacionales.

Desde 1994 hasta la actualidad, existe sólo un año de diferencia entre la recolección de los datos en el nivel local y la publicación y difusión de estos en el nivel nacional. Para un país de organización político-administrativa federal, este lapso puede considerarse adecuado. Desde una perspectiva de análisis de calidad de datos, da cumplimiento al principio de oportunidad, es decir, que no hay un gran retraso de tiempo entre el momento en el cual se recaban los datos en las jurisdicciones y se los presenta en la base de datos unificada nacional (Guevel, 2021).

Según lo reportado, en algunas áreas de ciertas jurisdicciones, persisten problemas de cobertura y de calidad de datos para algunas variables de registro más complejo. Se observan en dichos casos variaciones importantes en el número de hechos vitales y, consecuentemente, en las tasas resultantes del análisis. Dichos aspectos vulneran el objetivo de calidad de datos denominado, específicamente el principio de consistencia, en base a los objetivos de control de grandes volúmenes de datos en el ámbito del control interno y auditoría.

En relación con lo descripto y lo expuesto en los informes que ha desarrollado y presentado el Ministerio de Salud y su Dirección, cabe destacar que los reportes son desarrollados en base al análisis descriptivo en cuanto al conteo de decesos por causa de fallecimiento en un período determinado de tiempo. Se observa la aplicación de una metodología de análisis simple, basándose en el estudio en base a la mortalidad proporcional y tasa de mortalidad (cada 100.000 habitantes). Adicionalmente, el organismo realiza distintas clasificaciones de las defunciones por grupo de edad y sexo según grupo de causas de seleccionadas para un período anual determinado, donde no se pudo identificar un estudio de mayor alcance.

Básicamente, se busca describir que aún no se han aplicado otro tipo de metodologías de análisis, como podrían ser análisis multivariante o predictivo sobre la base de datos recopilada. Como resultado, se vio la posibilidad de pretender la aplicación de dichas metodologías en el

presente trabajo. Con el fin de otorgar un enfoque distintivo y ampliar el panorama para anticiparse las jurisdicciones a las tendencias y compatibilizar de manera adecuada sus planes de acción.

1.2 Introducción a la base de datos y descripción de las variables de estudio que la componen.

La base de datos a utilizar en el presente trabajo comprende datos sobre la cantidad de fallecidos y su clasificación según la causa, correspondiendo a los datos de la población argentina durante el período 2005 al 2019.

Respecto a sus características estructurales, se dispone de datos con atributos reales y actualizados, con atributos de tipo mixtos. Los mismos corresponden a una base de datos abierta y están disponibles para su utilización en la página web del Ministerio de Salud de la Argentina. Los datos corresponden a información recopilada por la Dirección de Estadística e Información en Salud (DEIS) del Ministerio de Salud, con apego a las normas internacionales vigentes de la Red Latinoamericana y del Caribe, para el Fortalecimiento de los Sistemas de Información de Salud (RELACSIS). La misma considera actualización anual, donde en esta oportunidad la fecha de relevamiento dicta fecha del año 2019 y fue publicada el 18 de marzo de 2019 en formato de archivo csv.

La base de datos que se seleccionó para el desarrollo del presente trabajo se obtuvo de la web del Ministerio de Salud de la Nación. La misma corresponde al período de 2005 al 2019 y contaba de 1073 variables de tipo métricas intercorrelacionadas que representan las principales causas de mortalidad. De ellas se puede ver el valor porcentual de las observaciones con una apertura por provincia.

En base a las características mencionadas, se la consideró razonable que, para poder desarrollar el trabajo de manera acertada y oportuna, se optó por sólo considerar el último período que corresponde al año 2019 y aquellas variables cuyos datos representan una significatividad mayor al 0.7% y a su vez cuya información sea consistente y veraz, dado que se detectaron simultáneos registros de enfermedades con 1 solo caso fallecido respecto al total, o habían causas no definidas con valores muy reducidos en relación al total de casos. Dicha discriminación fue realizada con motivo de compatibilizar la base de datos objeto de estudio, la aplicación de la metodología de Análisis de Componentes Principales y su desarrollo en la herramienta elegida. Para poder proceder con el análisis, se utilizó el entorno de desarrollo RStudio versión 1.4, que es dedicado a la computación estadística y gráficos.

Los resultados que se proponen obtener en el presente trabajo de investigación otorgarán la posibilidad de observar la relación entre las variables y la distribución sobre los objetos que componen la base de datos, brindando una visión más amplia del panorama donde se logre identificar la existencia de relaciones considerables entre las principales causas de muerte que ocurren en el país.

Por otra parte, resulta oportuno aclarar, que, dado que no se cuenta con una base de datos actualizada al presente, en los resultados no será considerando el contexto actual vinculado a la pandemia por el COVID-19. Por ello, es probable que las predicciones y valores que se obtengan con las técnicas de Aprendizaje Automático sean disímiles al presente, dado que no están comprendidos los decesos por la variante de enfermedad en la base de datos informada por el Ministerio de Salud. De todos modos, se plantea la aplicación de la metodología, lo cual en bases de datos futuras que incluyan el COVID-19 como causa de fallecimiento, la metodología mostrará un escenario en relación con los datos de origen al momento de su aplicación.

Variables de estudio

Dada la preselección mencionada precedentemente, a continuación, se mencionan las variables bajo estudio a las cuales les será aplicada la metodología de Análisis de Componentes Principales en el siguiente apartado. En la aplicación de técnica de Aprendizaje Automático, la cantidad de variables a considerar, puede que se amplíe, dado que no posee las mismas limitaciones que ACP. Las variables se mencionan a continuación: diabetes mellitus, neumonía, lesión auto-infligida intencionalmente, insuficiencia cardíaca, diabetes mellitus no insulinodependiente, cardiomiopatía, otras enfermedades del sistema digestivo, otros trastornos respiratorios, aneurisma y disección aórticos, tumor maligno del encéfalo, otras causas mal definidas, otras arritmias cardiacas, enfermedad renal crónica, enfermedad isquémica crónica del corazón, insuficiencia respiratoria, otras sepsis, otras enfermedades pulmonares crónicas, tumor maligno de hígado y vías biliares, hemorragia intraencefálica, tumor maligno de colon, otras enfermedades cerebrovasculares, tumor maligno sitio no especificado y la variable infarto agudo de miocardio

1.3 Otra perspectiva de análisis sobre los datos de salud de la sociedad argentina, mediante el A.C.P. y técnicas de aprendizaje automático.

Dando continuación a lo detallado en el apartado anterior, la base de datos descripta fue considerada adecuada debido a su estructura y tipo de datos para realizar sobre ella el desarrollo de la metodología de Análisis de Componentes Principales (A.C.P.) la cual requiere que las variables de estudio sean de tipo métricas, con una distribución por objeto, en este caso las serán las 24 Provincias que constituyen la Nación Argentina. Como a su vez, es candidata para construir sobre ella un modelo de análisis predictivo, teniendo una variable a predecir, que sería la cantidad de fallecidos por provincia.

El presente trabajo no solo se propone aplicar metodologías de estudio, sino que se propone la aplicación de las mencionadas debido al potencial que tiene la base de datos para proporcionar el desarrollo de un enfoque diferente sobre los datos que el Ministerio de Salud presenta. Se considera como una oportunidad mediante la cual se podrá identificar, mediante el ACP, la existencia de relaciones considerables entre las principales causas de muerte que ocurren en el país. Como también, desarrollar predicciones estimativas de los futuros y posibles decesos por categoría. Se plantea ofrecer la posibilidad de un mejor entendimiento de los datos para poder hacer uso de las técnicas de aprendizaje automático, entre ellas diferentes conjuntos y tipos de árboles de decisión. Entre las ventajas de estas se destacan; la predicción de tendencias y necesidades, la ayuda a que los aspectos habituales no cubiertos en la gestión, o errores cometidos para no sean repetidos mediante su detección y el desarrollo de acciones preventivas.

En el contexto organizacional en el cual se plantea su aplicación, dichos aspectos otorgarían una más adecuada proyección de los planes y estrategias para el sistema de Salud por parte del Ministerio de Salud y las jurisdicciones de que éste dependen. Puesto que éstas tienen el deber de tomar las decisiones relacionadas con la salud y la vida de la población argentina.

Análisis multivariado de las causas de fallecimiento de la población argentina.

El presente apartado introduce brevemente sobre las técnicas cuantitativas, cuyo foco es el desarrollo descriptivo y la aplicación del método de Análisis de Componentes Principales ACP. Para ello se partió de datos que fueron preparados y estructurados, conformando la base de datos origen descripta en el apartado anterior. El objetivo del ACP es esencialmente la reducción de la dimensionalidad para proveer un mejor panorama para la interpretación de los datos o variables como producto de las componentes obtenidas. Por su parte, se adicionará al estudio la descripción y la posterior aplicación del análisis multivariante de conglomerados, el cual, servirá para una geolocalización de las componentes y su distribución por grupo de provincias.

Finalmente se estudiará la contribución de cada variable y se analizarán aquellas de carga elevada según cada componente obtenida. Posteriormente se desarrollará la interpretación y conclusión sobre los resultados. A su vez, se evaluará cómo estas metodologías posibilitan mejorar la perspectiva y visión de la situación de la población argentina. Lo cual sirve fundamentalmente para que la gestión y toma de decisiones por parte del organismo sea oportuna y adecuada.

1.4 Resumen y descripción de la técnica estadística

El Análisis de Componentes Principales (A.C.P), se trata de un método de análisis multivariado focalizado en el estudio de la interdependencia que surge entre variables de tipo métricas (Catena, Ramos y Trujillo, 2003). Resulta habitual contar con información que dispone de numerosas variables inter-correlacionadas en mayor o menor grado. El método se vale de la creación de variables ortogonales que resumen la variabilidad de las originales. Por ello, el método en referencia procura obtener un conjunto más pequeño de variables llamadas factores o componentes principales, con una pérdida mínima de información y donde se elimina la correlación preexistente. Por lo tanto, su objetivo es esencialmente la reducción de la dimensionalidad para proveer un mejor panorama para la interpretación de los datos bajo análisis.

El presente trabajo enseñará el desarrollo del análisis y los resultados que serán obtenidos en base a la aplicación de las técnicas de Análisis escogida. Durante el mismo se estudiarán variables métricas que representan la proporcionalidad de las principales causas de

fallecimiento estimadas de las personas con relación al total de la población fallecida por Provincia. La base de datos utilizada corresponde a la última reportada por el Ministerio de Salud de la Nación siendo la del período 2019. El objetivo del estudio es analizar los factores que diferencian las causas de fallecimiento entre sí, como así también determinar las que se consideran semejantes, pudiendo definir las características causantes.

Así mismo, mediante la aplicación del método de clúster, se buscará clasificar las distintas provincias en grupos, para luego, proceder con la representación gráfica de dispersión y su análisis, para lograr así, explicar de mejor manera las relaciones entre variables y objetos.

Validación de la Matriz por utilizar

Para poder llevar a cabo el método A.C.P. se precisa de determinados requisitos, como ser el tipo de variable y la cantidad de estas, pero principalmente se requiere la evaluación de la correlación entre ellas. Para ello, existen diferentes pruebas para definir si es recomendable la aplicación de la técnica o no sobre la matriz de datos escogida.

Una de las mencionadas pruebas de evaluación, es el Test de esfericidad de Bartlett. El mismo, siempre se debe de realizar previo al desarrollo del análisis de componentes principales, puesto que su fin es evaluar que la matriz de correlaciones no sea una matriz de Identidad, ya que de ser así las covarianzas serían nulas y no habría agrupación entre variables que justifiquen la existencia de una componente principal. La prueba, realiza su comprobación mediante el p-valor, el cual se define como la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta. El supuesto, por tanto, considera que el p Valor debe ser menor a 0.5, para poder rechazar la hipótesis nula y así proceder al desarrollo del análisis.

Definición del número de Componentes a considerar

Para definir la cantidad de componentes a extraer, se utilizan distintos criterios de elección. Entre los cuales se destacan:

- El Criterio del auto-valor superior a la unidad: Éste consta de retener aquellas componentes cuyo auto-valor asociado sea superior a 1.
- El Criterio del gráfico de sedimentación - Scree plot : Consiste en retener las componentes donde la línea que une a los autovalores comienza a aplanarse, es decir, tomar aquellas componentes que se encuentren hasta el “codo” o corte de la línea.
- El Método paralelo: La estrategia de Horn consiste en contrastar los autovalores obtenidos mediante un PCA Paralelo, aplicado a muestras aleatorias de variables no correlacionadas

con el mismo número de variables y observaciones que la muestra original, y que generarán los autovalores que se han ajustado para evitar el error muestral. Es una metodología para objetivar el criterio del gráfico de sedimentación, dado que hay mucha subjetividad al querer definir dónde se produce el codo de la línea (Catena, Ramos y Trujillo, 2003).

El análisis de Componentes principales se vale de la creación de variables ortogonales que resumen la variabilidad de las originales. Las variables con mayor varianza son las que poseen una mayor influencia en la generación de componentes principales. Para interpretar las componentes extraídas, es necesario observar la contribución de cada variable y analizar aquellas de carga elevada. Cuanto mayor sea ésta, más alta será la influencia que tiene esa variable en la formación de la componente, y pudiendo así dar uso a las variables más representativas y otorgar una interpretación al eje.

Una vez realizada la interpretación correspondiente, se debe de complementar éste con análisis gráficos de cargas por componente (o dimensión), como también graficar los objetos y así compararlos para lograr la interpretación visual de la relación entre los mismos.

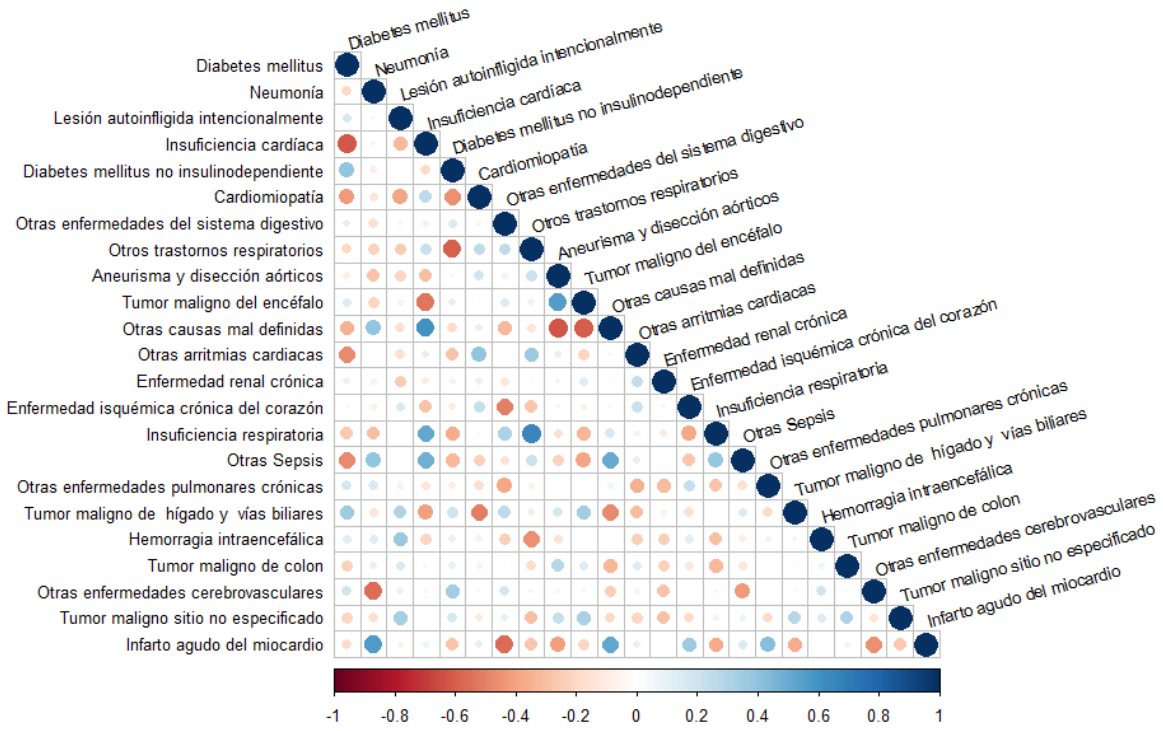
1.5 Implementación y desarrollo metodológico del modelo

Una vez definida la base de datos a utilizar, de ahora en más matriz, así como las variables y objetos que la componen. Entre los requisitos más importantes que debe cumplir la matriz de datos se encuentra el requisito fundamental de que las variables tienen que estar altamente correlacionadas. Para ello se aplica la técnica de Pearson o el coeficiente de correlación de Pearson. Este tiene el objetivo de indicar cuán asociadas se encuentran dos variables entre sí.

Por tanto, para poder definir si era correcta la aplicación del Análisis de Componentes Principales, se procedió con el método de Pearson para obtener la Matriz de Correlación (R) (Aldás y Uriel., 2017). Y posteriormente sobre ésta, se aplicó la explicada prueba de Esfericidad de Bartlett. Y así, comprobar si la Matriz a utilizar es una de identidad o no, es decir si se rechaza o no la Hipótesis Nula. En este caso el p valor resultó 0.0000001230269, lo cual indica que dicho determinante es menor a 0.5 y por ende muy próximo a cero, por lo que se rechazó la hipótesis nula y fue factible continuar con el Análisis.

Figura 1.

Gráfico de correlación entre las variables.



Nota. Gráfico elaborado mediante la utilización del software Rstudio (RStudio Team, 2021*).*

Al estudiar el gráfico de correlaciones expuesto en la figura 1, es oportuno explicar los colores que se observan en él y analizarlo a razón de las variables que involucra. Los puntos de color rojo del gráfico indican que la correlación entre las variables que se cruzan es menor a cero, lo cual significa que es negativa. Es decir, que las variables se relacionan inversamente. Cuando el valor de alguna variable es alto, el valor de la otra variable es bajo. Mientras más próximo se encuentre a -1, más clara será la covariación extrema. Si el coeficiente es igual a -1, nos referimos a una correlación negativa perfecta. En esta oportunidad, podemos ejemplificar mediante las variables “Insuficiencia cardíaca” y la variable “Diabetes mellitus” cuya relación no es muy cercana y está señalada con un punto rojo.

Por otra parte, al observar los puntos de color azul la correlación es mayor a cero e igual a +1 significa que es positiva perfecta. En este caso significa que la correlación es positiva, es decir, que las variables se correlacionan directamente. Cuando el valor de una variable es alto, el valor de la otra también lo es, sucede lo mismo cuando son bajos. Si es cercano a +1, el coeficiente será la covariación. En esta oportunidad, podemos ejemplificar mediante las

variables “Insuficiencia respiratorio” y la variable “Otros trastornos respiratorios”. Dicha relación está señalada con un punto azul y tiene un mayor sentido de correlación.

En cambio, cuando se observa ausencia de color en el gráfico, la correlación es igual o muy próxima a cero lo que significa que no es posible determinar algún sentido de covariación. Sin embargo, no significa que no exista una relación no lineal entre las variables. En este caso, se pueden mencionar la variable “Diabetes mellitus no insulinodependiente” respecto a la variable “Lesión autoinfligida intencionalmente”.

Tipificación de variables y definición de las componentes principales en R

Para el desarrollo del estudio, se decidió trabajar con las variables tipificadas (estandarizadas) para llevar a una medida homogénea todas las variables. Luego se corrió la sentencia en R para la obtención de los autovalores tal como se puede observar en la figura 2 ordenados de mayor a menor de la matriz de correlaciones R). Así mismo se calculó el autovector unitario de módulo uno correspondiente para cada autovalor. En el trabajo realizado se obtuvieron 24 autovectores, de los cuales se muestra el output de R de solo las primeras 8 componentes, que tienen mayor representación del conjunto.

Figura 2.

La tabla 1 indica los 23 Autovalores (de mayor a menor). La tabla 2 enseña la carga factorial de las primeras 8 Componentes o Auto –vectores.

```
Standard deviations (1, .., p=23):
 [1] 2.137053122 1.887372830 1.642168391 1.457568920 1.341166785 1.242324649 1.197505785 1.061098263 0.953900910 0.877691306 0.777150024 0.66
[13] 0.621119843 0.525532790 0.450614915 0.416245073 0.360490291 0.310826295 0.271233920 0.224823867 0.147436136 0.111054470 0.009993299

Rotation (n x k) = (23 x 23):
                                                PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8
Diabetes mellitus                        0.28858666 -0.12767059  0.04404402  0.38640756 -0.222739006  0.10221061 -0.003605185  0.06935745
Neumonía                                -0.18775630 -0.26966548  0.11638227 -0.21615223 -0.209850606 -0.05310615 -0.228856131  0.27949967
Lesión autoinfligida intencionalmente    0.13865726 -0.13485838  0.25692442 -0.13736024  0.092957209  0.22869681  0.490352883  0.15318461
Insuficiencia cardíaca                  -0.33622483  0.15190120  0.15229543  0.02383229  0.268510972 -0.18781422 -0.148534468 -0.15640583
Diabetes mellitus no insulinodependiente  0.25949995 -0.10775130  0.13524925  0.15400035  0.047069725 -0.49930259 -0.126066527  0.04264460
Cardiomiopatía                          -0.17513887  0.12904933 -0.39491848 -0.05454322  0.189117325 -0.06494556  0.088035928 -0.20473832
Otras enfermedades del sistema digestivo  0.14226828  0.32206015  0.14452724  0.05912266 -0.064433698 -0.11127123 -0.123316925  0.27650445
Otros trastornos respiratorios          -0.14565937  0.38969593 -0.09944137  0.06836516 -0.088800989  0.37451456 -0.066022811  0.17712686
Aneurisma y disección aórticos           0.19655340  0.19162256 -0.29547250 -0.28469088 -0.009206968  0.09526328 -0.135382532 -0.14278176
Tumor maligno del encéfalo               0.29738890  0.05267888 -0.19481604 -0.22005196 -0.137491960  0.03638509 -0.182803050 -0.30221020
Otras causas mal definidas              -0.37021833 -0.16255859  0.16476865  0.13796370  0.034846538 -0.13803210 -0.018736690 -0.21481653
Otras arritmias cardiacas               -0.17827172  0.15543277 -0.28011548 -0.14838048 -0.016486513 -0.15236429  0.298036653  0.43596394
Enfermedad renal crónica                -0.03090564  0.03034177 -0.20835609  0.08252739 -0.467572790 -0.36948980  0.188547875 -0.16904728
Enfermedad isquémica crónica del corazón -0.01296784 -0.24971734 -0.32077597  0.05587464  0.084192134  0.14909459  0.391539767 -0.07279621
Insuficiencia respiratoria              -0.16812760  0.37211777  0.23853056  0.12090451  0.039054526  0.20260748  0.111853513 -0.10357514
Otras Sepsis                            -0.27459058  0.05147605  0.32530135 -0.21053954 -0.187996014  0.06540827  0.028717251 -0.05768480
Otras enfermedades pulmonares crónicas  -0.00465854 -0.27004794 -0.09443560  0.11518707  0.094607962  0.43215272 -0.423710910 -0.04738349
Tumor maligno de  hígado y  vías biliares  0.27430114  0.05929508  0.25446907 -0.12515421 -0.264002449  0.14361317  0.097798477 -0.02806835
Hemorragia intraencefálica               0.10699681 -0.24500765  0.11202195 -0.01762621  0.296256703  0.02409312  0.244863539 -0.03866483
Tumor maligno de colon                   0.10051437 -0.04407867 -0.04539584 -0.42216280  0.258090421 -0.11928792 -0.185955292  0.40143734
Otras enfermedades cerebrovasculares     0.16867013  0.11773160 -0.04397183  0.34790007  0.457564103 -0.06336025 -0.048107966  0.11421078
Tumor maligno sitio no especificado      0.15221743  0.05655881  0.21734384 -0.41627467  0.208814566 -0.07002395  0.056429964 -0.37217796
Infarto agudo del miocardio             -0.24613316 -0.36866984 -0.11370770 -0.09213232 -0.080884269  0.07167670 -0.098237848  0.06690125
```

Nota. Gráfico elaborado mediante la utilización del software Rstudio (RStudio Team, 2021*).*

Criterio de selección de las componentes principales

Para una selección más adecuada, se aplicaron tres criterios diferentes para la selección. Para luego elegir uno que defina una cantidad razonable de Componentes a utilizar. Al usar el criterio del autovalor>1, resultó una cantidad recomendada de 8 (hasta PC8). Al utilizar el criterio del gráfico de sedimentación, el codo en la gráfica indicó que el número de componentes debía ser 3 (PC3). Pero, para poder obtener una mejor interpretación visual del mismo, se aplicó el método Paralelo de Horn y se graficó el mismo en R, donde puede verse la recta horizontal mostrando el corte de la línea del scree-plot Parallel Analysis que se encuentra en el extremo derecho de la figura 3. Finalmente fueron seleccionadas 3 componentes principales, las cuales representan el 0.47 (47%) de la varianza y que explican lo suficiente para poder desarrollar el análisis.

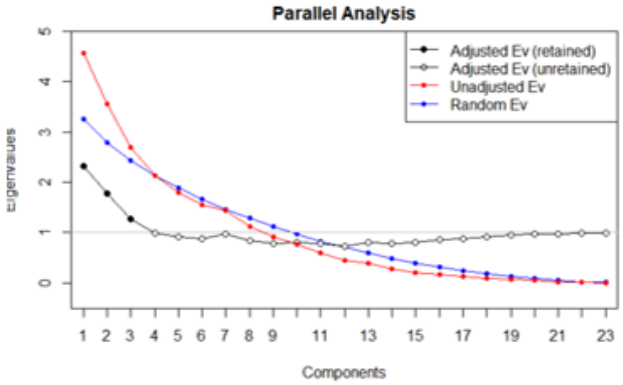
Figura 3.

Aplicación del método paralelo de Horn y su representación gráfica.

```
Results of Horn's Parallel Analysis for component retention
690 iterations, using the mean estimate

-------------------------------------------------
Component   Adjusted     Unadjusted    Estimated
            Eigenvalue   Eigenvalue    Bias
-------------------------------------------------
1           2.320004     4.566996      2.246991
2           1.779728     3.562176      1.782447
3           1.273343     2.696717      1.423373
-------------------------------------------------

Adjusted eigenvalues > 1 indicate dimensions to retain.
(3 components retained)
```



Nota. Gráfico elaborado mediante la utilización del software Rstudio (RStudio Team , 2021*).*

1.6 Evaluación de las componentes obtenidas

El factor o la componente se interpreta en función de las variables más correlacionadas a él y no se toma el valor como tal (los signos), sino la magnitud. En consecuencia, las variables con mayor ponderación en el primer factor son: las definidas como Otras causas mal definidas, Insuficiencia cardíaca, Otras Sepsis e Infarto agudo de miocardio. Por otra parte, están las otras 4 (cuatro) variables, que a su vez poseen una gran carga en la misma componente (PC1). Éstas son; Diabetes mellitus no insulinodependiente, Diabetes mellitus, Tumor maligno de hígado y vías biliares y por último Tumor maligno del encéfalo. La componente 1 (uno) podríamos

considerar como afecciones relacionadas a diabetes y problemas del corazón, así como causas tumorales cuyas causas son indefinidas en su mayoría.

Cabe destacar que las personas que tienen diabetes o prediabetes presentan un mayor riesgo de enfermedad cardíaca relacionada con esta. Las enfermedades cardíacas relacionadas con la diabetes pueden ser: enfermedad coronaria, insuficiencia cardíaca y cardiomiopatía diabética (CDC Centro para el Control y la Prevención de Enfermedades, 2021)

Por su parte, el segundo factor tiene mayor carga en las variables, Infarto agudo de miocardio

Otras enfermedades pulmonares crónicas, Neumonía, Insuficiencia respiratoria y Otros trastornos respiratorios. Éste, se encuentra vinculado a problemas de salud diferentes a los mencionados en la componente 1, y se encuentra mayormente vinculado a enfermedades de índole respiratoria. En la mayoría de los casos, la dificultad para respirar se atribuye a enfermedades del corazón o de los pulmones. El corazón y los pulmones participan en el transporte de oxígeno hacia los tejidos y en la eliminación de dióxido de carbono, y los problemas relacionados con cualquiera de estos dos procedimientos afectan la respiración (Personal de Mayo Clinic, 2020).

El tercer factor bajo análisis explica mejor las variables: Cardiomiopatía, Enfermedad isquémica crónica del corazón, Aneurisma y disección aórticos y por último otras arritmias cardiacas. Por lo descripto, queda en evidencia la preponderancia de las enfermedades vinculadas al corazón en el tercer componente bajo análisis.

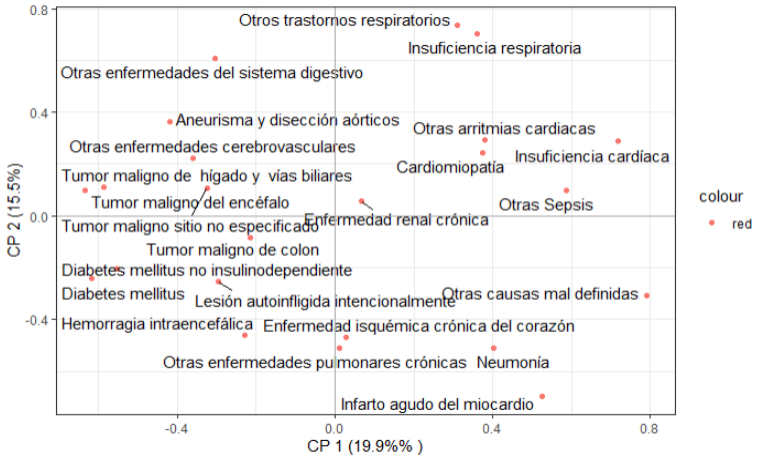
Para continuar con la línea de investigación, cabe destacar que las causas más importantes de cardiopatía son una dieta malsana, la inactividad física, el consumo de tabaco y el consumo nocivo de alcohol. Los efectos de los factores de riesgo comportamentales pueden manifestarse en las personas en forma de hipertensión arterial, hiperglucemia, hiperlipidemia y sobrepeso u obesidad (Organización Mundial de la Salud [OMS], 2017).

A continuación, se procede a describir algunos aspectos sobre un gráfico producto del estudio en dos dimensiones de las componentes principales 1 y 2 mediante el programa RStudio. Las mismas son las obtenidas en el método de análisis multivariado aplicado. Como se puede observar en el cuadro (figura 4) a continuación, se pueden notar grupos de variables, por ejemplo: otras arritmias cardíacas, cardiomiopatía, insuficiencia cardíaca y otras sepsis. En el cuadrante inferior izquierdo se puede observar otra agrupación que incluye las enfermedades denominadas diabetes mellitus no insulinodependiente y diabetes mellitus. En el límite inferior del cuadrante superior izquierdo se encuentran agrupadas las enfermedades tumorales, siendo

tumor maligno de vías biliares, tumor maligno de encéfalo y tumor maligno sitio no especificado.

Figura 4.

Gráfico de variables en 2 dimensiones (C.P.1 y C.P.2)



Nota. Gráfico elaborado mediante la utilización del software Rstudio (RStudio Team, 2021*).*

1.7 Análisis de conglomerados.

En el presente trabajo, se desarrolla brevemente otro método multivariado para poder complementar el estudio de componentes principales. El mismo es el método de análisis de conglomerados o clústeres, con la finalidad de visualizar la distribución de las variables por grupo de provincias prestando una. A continuación, se describirá la metodología brevemente para un mayor entendimiento sobre la misma.

El análisis de Clústeres es una técnica diseñada para clasificar distintas observaciones en grupos, de forma tal que cada grupo sea homogéneo, es decir, que cada observación contenida en él sea parecida a todas las que estén incluidas en el mismo grupo. Y a su vez que los grupos sean lo más distintos posible unos de otros, respecto a las variables consideradas (Amat, 2017).

Para poder distinguir a esta metodología de otras técnicas de agrupación, se debe destacar que en ésta no existe un criterio preestablecido previamente sobre la definición de grupos, es decir son desconocidos y es necesaria su derivación de las observaciones.

Para el proceso de clustering, se debe disponer de n observaciones (individuos, países, etc.) de las que tiene información sobre las observaciones de k variables. Sobre éstas se establece un indicador que explique en qué medida (distancia o similaridad) se parecen entre sí. Se forman

los grupos en base a las observaciones con mayor parecido, de acuerdo con la medida de similaridad. En este punto se debe realizar la elección del tipo de análisis de conglomerado: que puede ser jerárquico o no jerárquico, que permite establecer los grupos para luego poder describirlos. Existen diversas medidas de similaridad para las distancias métricas como ser; la distancia euclídea, euclídea al cuadrado, minkowski, city block o Manhattan. Mientras tanto, para datos binarios podemos utilizar entre otros: Jaccard, Sokal y Michener, Rogers y Tanimoto, Ochiai, coeficiente S2 de Gower y Legendre (Catena, Ramos y Trujillo., 2003).

Análisis de conglomerado Jerárquicos

Pueden ser aglomerativos (parte de los individuos y va fusionando los grupos) o des-agregativos (parte un único grupo con todo el universo y va dividiéndolo). Entre estos, podemos mencionar el método del centroide, vecino más cercano, vecino más lejano, vinculación promedio y el método de Ward que se diferencia por buscar maximizar la homogeneidad dentro de cada conglomerado, calculando los centroides resultantes de las posibles fusiones.

Estas metodologías dispondrán de un historial de la conglomeración, los sucesivos pasos y cálculos de distancia; el mismo puede analizarse gráficamente mediante un dendograma que permitirá considerar el número óptimo de grupos con los cuales finalizar la iteración.

Fundamentos para su aplicación

Utilizando la misma base de datos analizada previamente, se decidió profundizar en su análisis a través de la aplicación del método de Ward de tipo jerárquico. Con el fin de definir la cantidad óptima de clústeres o grupos de observaciones homogéneos explicados por las variables tratadas. Luego se procedió con la representación gráfica de dispersión en el programa RStudio, mediante el denominado método de k-means que corresponde al tipo de análisis no jerárquico.

Aspectos previos al desarrollo de la metodología

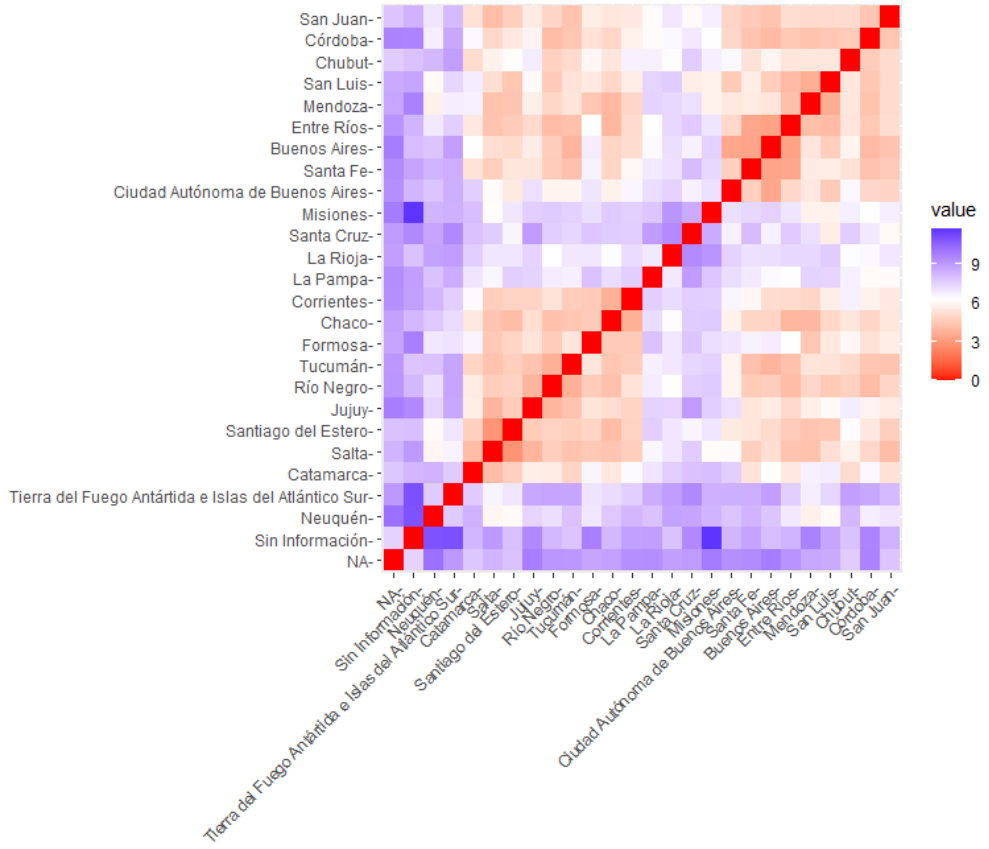
Previo a la aplicación del método, se estandarizaron los valores de las variables, y se realizó una evaluación sobre la tendencia a la agrupación de datos. La misma se realizó mediante el cálculo de la matriz de distancias (partiendo de las distancias euclídea y euclídea al cuadrado) y su representación en forma gráfica con un mapa de calor que se puede observar en la figura 5. El nivel de color es proporcional al valor de la disimilitud entre las observaciones: rojo

intenso si tiende a cero y azul intenso corresponde al valor más alto de la distancia euclidiana calculada.

Adicionalmente se realizó la prueba de Hopkins, la misma se utilizó para comprobar que la probabilidad de que los datos sean generados por una distribución aleatoria uniforme. Esta suele utilizarse como medida sustitutiva de la tendencia al agrupamiento. Cuando los valores del estadístico se aproximan a 0,5 se indica un alto grado de aleatoriedad espacial. Exitosamente, el valor resultante de la ejecución en R fue de H = 0.41.

Figura 5

Matriz de distancias, mapa de calor



Nota. Gráfico elaborado mediante la utilización del software Rstudio (RStudio Team ,2021*).*

Aplicación del análisis de Clústeres por el método de Ward

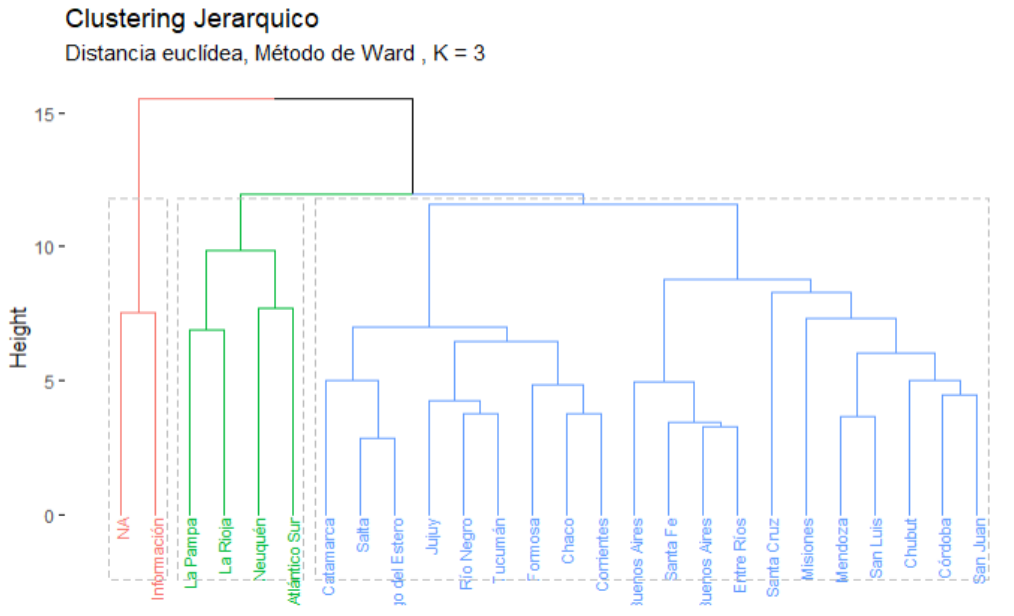
Para el estudio de la base de datos bajo análisis, se optó con el método jerárquico de Ward o bien Método de varianza mínima. Dado que este tipo de agrupamiento no requiere establecer

previamente el número de clústeres que van a generarse. Se decidió desarrollar su aplicación mediante el uso de sentencias en el programa RStudio. Donde se calculó en primera instancia la distancia euclídea, la aplicación del método de Ward y la representación del dendograma. Con esta metodología se persigue la minimización de la varianza intragrupal y la maximización la homogeneidad dentro de los grupos. La misma calcula los centroides de los grupos resultantes de las posibles fusiones y luego obtiene la distancia euclídea al cuadrado al centroide de todas las observaciones del grupo.

Luego de la aplicación del método jerárquico de Ward en el programa RStudio, se obtuvo como resultante que el número de grupos o clústeres óptimo es de 3 tal como se puede observar en el gráfico que se presenta en la figura 6. La definición de cantidad de clústeres se identifica mediante las casillas superiores que comprenden las inferiores. Las mismas se encuentran referenciadas con 3 colores distintos, siendo rojo, verde y azul.

Figura 6.

Dendograma - método de Ward



Nota. Gráfico elaborado mediante la utilización del software Rstudio (RStudio Team, 2021*).*

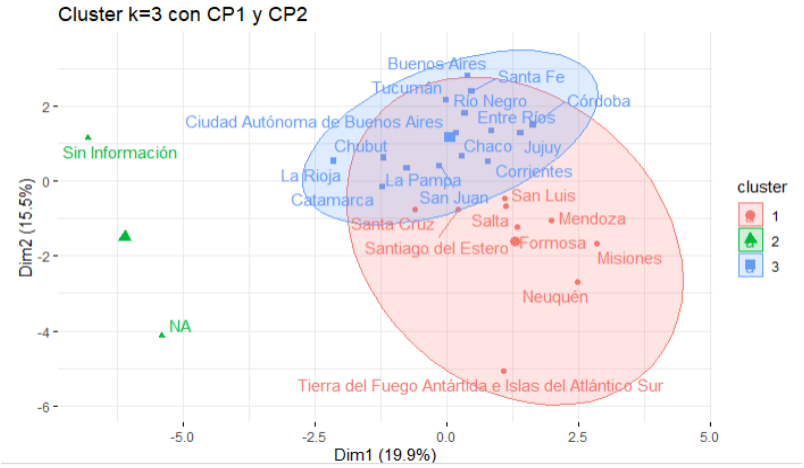
1.8 Análisis de los clústeres obtenidos y su relación con las componentes principales

Mediante la metodología de Ward y los índices se pudo definir la cantidad de agrupaciones óptima. Mediante el dendograma y la representación gráfica de clústeres en 2 dimensiones siendo la componente principal 1(uno) y 2 (dos) dispuesto en la figura 7, se puede visualizar claramente la composición de cada uno de ellos. Partiendo del análisis del clúster 1, podemos describir las provincias que lo componen: Tierra del Fuego e Islas del Atlántico Sur, Neuquén, Misiones, Formosa, Santiago del Estero, Salta, Mendoza, San Luis y Santa Cruz. Al tener mayor grado de correlación a la dimensión 1 (componente 1), se puede asumir que las mencionadas provincias poseen un mayor número de causas de fallecimiento vinculado a las afecciones relacionadas a diabetes y problemas del corazón.

En cuanto al tercer grupo de provincias o clúster 3, las mismas son Bs.As., Tucumán, Santa Fe, Rio Negro, Córdoba, Chaco, Jujuy, CABA, La Rioja, Catamarca, La Pampa, San Juan, Corrientes y Jujuy. Por lo que se pudo obtener al estudiar e interpretar el gráfico de distribución de Clústeres en las componentes 1 y 2 como en del de las componentes 2 y 3, el mencionado clúster tiene una mayor proximidad en cuanto a la componente número dos, y por lo tanto se interpreta que dichas provincias poseen una mayor cantidad de fallecidos en el año por causa de enfermedades de índole respiratoria. El clúster 2 queda excluido del análisis debido que la base de datos cuenta con conteo de fallecidos y su causa, pero no fue definida la provincia a la cual pertenecieron los mismos, figurando como NA y Sin información.

Figura 7.

Gráfico de Clústeres en 2 dimensiones, entre las componentes principales 1 y 2.



Nota. Gráfico elaborado mediante la utilización del software Rstudio (RStudio Team, 2021*).*

1.9 Resultados del análisis multivariado y su contribución a la gestión organizacional presente.

Los resultados obtenidos, analizándolos desde la perspectiva sanitaria, las tres componentes principales poseen aspectos en común sobre determinados grupos de provincias, que pueden darle un indicio o visión considerable a la Dirección de Estadísticas e Información de Salud respecto a la tendencia de salud de la sociedad argentina. Como bien se mencionó al describir las componentes, la diabetes está fuertemente vinculada a tendencias de afecciones cardíacas, las cuales a su vez están relacionadas a los problemas respiratorios. Al buscar información al respecto, se halló un informe de la OPS, el cual ha demostrado que las medidas sencillas de estilo de vida son eficaces para prevenir o retrasar la aparición de la diabetes tipo 2. Para ayudar a prevenir la misma y sus complicaciones, las personas deben de llevar un estilo de vida más saludable. Para ello, se debe lograr y mantener un peso corporal saludable, es decir, se requiere más actividad física para controlar el peso, seguir una dieta saludable, evitando el azúcar y las grasas saturadas. Así como también evitar el consumo de tabaco, ya que fumar aumenta el riesgo de diabetes y enfermedades cardiovasculares según informa la OPS (Organización Panamericana de la Salud [OPS], 2021).

Sobre el estudio desarrollado, las principales causas de fallecimiento mencionadas presentan una fuerte correlación respecto al estilo de vida. Por lo tanto, se procede a dar continuación a la problematización planteada en el presente trabajo, considerando las tareas y deberes del Ministerio de Salud de la Nación (M.S.N.). Según el informe publicado por la OMS, las políticas sanitarias son fundamentales para crear entornos propicios que aseguren la asequibilidad y la disponibilidad de opciones saludables para motivar a las personas para que adopten y mantengan comportamientos sanos. Y, por ende, el MSN en base a las correlaciones analizadas y explicadas, podría servirse de ellas como herramientas para evaluar el panorama y realizar un plan estratégico por área geográfica o grupo de provincias que se vieron más afectadas por las determinadas principales causas de fallecimiento. Dando lugar, a crear dichos entornos propicios para instruir y ayudar a la sociedad argentina a prevenir las mencionadas enfermedades, consideradas las principales causantes de fallecimiento del país.

Si bien dicho análisis puede ser una oportunidad para una más clara visibilidad de los datos y registros con los que cuenta el Ministerio. Se pretende un aporte adicional y novedoso en cuanto a las metodologías cuantitativas para el análisis y gestión de datos, mediante la técnica de

Aprendizaje automático como método predictivo a través de los registros de fallecidos por provincia de la nación y las enfermedades causantes.

Análisis predictivo aplicado a los datos reportados por el Ministerio de Salud

El presente apartado, busca introducir al usuario sobre la técnica de aprendizaje automático, focalizándose en la apreciación de la técnica como objeto generador de valor agregado para la gestión de datos en contextos organizacionales. En la actualidad, las tecnologías son consideradas parte de los capitales necesarios de las organizaciones dado que permiten tomar mejores decisiones basadas en información de valor. La implementación de este tipo de técnicas proporciona la posibilidad de generar predicción de resultados futuros, detección de errores o anomalías y la visualización de patrones de comportamiento. Para ello es necesario proceder con una serie de pasos esenciales, iniciando con la recolección de datos o elección de la base de datos, la preparación de los datos, la elección del modelo acorde a la variable que se pretende predecir, realizar un entrenamiento de la máquina o modelo, testeo de este, configuración de parámetros y finalmente realizar la evaluación de los resultados de la predicción o inferencia obtenida.

El presente estudio pretende proponer la aplicación de dicha metodología para el análisis y gestión de datos que posee el Ministerio de Salud. El foco de la problemática es generar un modelo que ayude a la previsualización o estimación anual de la cantidad de fallecidos con una visibilidad por tipo de causa y ocurrencia en las provincias de la república argentina. Incluyendo como atributos adicionales las variables rango etario y sexo.

Para dar lugar a dicha propuesta se procedió a la selección de la base de datos de Defunciones ocurridas y registradas en la República Argentina que ofrece el Ministerio de Salud y la Dirección de Estadística e Información en Salud (Dirección de Estadística e Información en Salud [DEIS], 2019). Sobre la misma, se debe de realizar la serie de pasos mencionados. Iniciando con el proceso de limpieza, transformación y modelado de información. Donde luego se utilizará sobre ésta las técnicas conjunto (ensamble) de árboles de decisión (“Random Forest”) y la Regresión del árbol de decisión potenciado (“Boosted Decision Tree”). De las mencionadas se seleccionará la que se considere más apropiada, en base a los valores de referencia de la raíz del error cuadrático medio, también conocido como RMSE (“Root Mean Squared Error”), la cual es la métrica para evaluar la eficiencia del

modelo. También se considerará el coeficiente de determinación, o bien conocido como R2, para evaluar la eficiencia de clasificación.

Luego de evaluar los resultados obtenidos, se concluirá con un desarrollo sobre la interpretación de los valores y su aporte como medio para dar amplitud a la visión y estudio de datos por parte del Ministerio de Salud.

1.10 Resumen y presentación del desarrollo metodológico

El problema de predicción a enfrentar en el presente estudio y la fuente de datos elegida corresponde un modelo predictivo de regresión lineal, el cual pretende la predicción de la cantidad de fallecidos por causa y provincia de la república argentina, como también por rango etario y sexo. Como se mencionó en los apartados previos, en el presente se utilizó una fuente de datos que comprende registros de fallecidos desde el año 2005 hasta 2019. Para poder desarrollar un buen modelo predictivo se considera oportuno obtener la mayor cantidad de registros históricos posibles, para que le puedan otorgar al mismo una mejor estructura y patrones de comportamiento para su entrenamiento. Otro aspecto elemental del aprendizaje automático es que éste permite la disminución de errores. A partir de un error cometido, esta metodología registra las variables y en un futuro permite que no se repita. La solidez del sistema dependerá en gran medida del tiempo que lleve integrado al proceso de la organización, de los ajustes y del seguimiento de cada modelo aplicado por parte de los analistas de datos

El aspecto más desafiante del presente estudio residió en el diseño y planificación de la limpieza, creación y unificación de variables en un conjunto de datos definitivo, acorde al problema organizacional seleccionado. Requiriendo de la creación de atributos que corresponden a agrupaciones de enfermedades, para que permitiera realizar predicciones más próximas y acordes, en algunos casos además fue una condición necesaria para poder continuar con la investigación. En la siguiente sección se expone en detalle la etapa de limpieza, transformación de datos y la creación unificación de toda la data en una estructura tabular definitiva y la ejecución de predicciones por categoría de enfermedades establecidas.

1.11 Limpieza, transformación de los datos y selección de atributos

La base originalmente contaba con once (11) campos cuya denominación y tipo de dato eran los siguientes: *anio* (numérico), *jurisdiccion_de_residencia_id* (numérico), *jurisdicion_residencia_nombre* (varchar), *cie10_casusa_id* (varchar), *lista_clasificacion* (varchar), *sexo_id* (numérico) , *Sexo* (varchar), *grupo_edad* (varchar), *muerte_materna_id* (varchar), *muerte_materna_clasificacion* (varchar) y la variable a predecir *cantidad* (numérico).

Es de suma importancia la evaluación de la disposición y calidad de los datos, la cantidad de registros y de campos (atributos), ya que pueden llevar a un modelo a ser óptimo como también a ser subóptimo. En algunos casos es usual que se genere el fenómeno overfitting, mejor conocido como "sobre ajustes", los cuales afectan la performance del modelo en trato. Tal como menciona Zach (Zach , 2020), este ocurre cuando un modelo es reducido y demasiado cercano a los datos de entrenamiento y, por lo tanto, se termina construyendo un modelo que no es útil para hacer predicciones sobre nuevos datos. Esto afecta la calidad del modelo, obteniendo valores óptimos en las métricas que miden el rendimiento de la regresión durante el entrenamiento , pero luego al aplicarse el mismo para la generación de predicciones, las mismas indiquen resultados totalmente opuestos y no satisfactorios.

La preparación y limpieza de datos son aspectos fundamentales del proceso que otorgan el grado de eficiencia del modelo. Como bien indica Logicalis (Logicalis, 2015), el objetivo a través de dicho tratamiento es encontrar buenos subconjuntos de predictores o variables explicativas, es decir, hallar los que mayor utilidad aportan y los que mejor se ajustan a los datos. Por dicho motivo, se procedió con una serie de pasos de transformación y depuración de la base de datos para generar la calidad necesaria en los atributos. Con el fin de obtener una mejor noción de los puntos clave y las técnicas de limpieza, se tomó como inspiración adicional la lista de recomendaciones de ajustes de datos para modelos predictivos, del artículo publicado por Zita (Zita, 2022). Éste principalmente dio foco a la selección óptima de atributos como también la creación de nuevos. Este paso dependió en gran medida de la evaluación que se realizó de los resultados que se obtuvieron durante cada proceso de entrenamiento. Entre cada uno de ellos, y mediante la inspección de los valores predichos y los coeficientes, se ajustó en muchas oportunidades el modelo como también se procedió a la eliminación y/o generación de atributos.

A continuación, se listan los pasos de depuración de datos y selección de atributos realizados:

a. Fueron editados los campos *jurisdicion_residencia_nombre, lista_clasificacion* y *grupo_edad*. En los registros se mantuvieron todos los valores en mayúscula y se quitaron de las celdas caracteres especiales tales como tildes(´), comillas dobles (“”) y las comas (,), ya que estas generaban en la lectura de columnas desplazamientos de lugar al guardar como CSV (archivo delimitado por comas). También fue reemplazada la letra *ñ* por *ni* en todos los registros para que no generase inconvenientes en el sistema al momento de leer y procesar el archivo de datos. Al atributo *lista_clasificacion* se le quitaron los números debido que sobre una clasificación existían números disímiles cuando en realidad correspondían a la misma. Lo mismo se realizó con los registros del campo denominado *grupo_edad.*
b. El atributo *jurisdicion_residencia_nombre* que representa el nombre de la provincia donde ocurrió el deceso poseía 27 registros en NA, lo cual no era de utilidad para el estudio que se pretende y si se conservaban podía afectarlo, por tanto, se procedió a su eliminación ya que representaba el 0.00392% respecto al total de registros, y se consideró que no eran valores significativos.
c. El campo Sexo contenía cuatro clasificaciones, las cuales eran *masculino, femenino, desconocido e indeterminado*. En dicha oportunidad se optó por unificar las últimas dos como *desconocido,* puesto que era la que poseía más cantidad de registros que con el concepto *indeterminado*.
d. Las columnas *jurisdiccion_de_residencia_id* y *sexo_id* fueron excluídas, dado que, al ser campos numéricos, el modelo podía interpretar los id de mayor valor como un número que representa mayor ponderación. Y se decidió conservar los atributos descriptivos que representaban a los mismos siendo éstos *jurisdicion_residencia_nombre* y *Sexo.*
e. Se optó por la exclusión de los atributos *muerte_materna_id* y *muerte_materna_clasificacion*, debido que en su mayoría estaban con dato NA (No Aplica) y estos representaban un 99,4% del total de registros en dichos campos. También se observó que no generaba aportes representativos ya que la mayor parte de los fallecidos no correspondía a muertes maternas.
f. Se procedió con la eliminación del campo denominado *cie10_casusa_id*. Si bien éste suponía que representaba el Id de la causa de fallecimiento, en cuanto se analizó la

correlación con la causa, ésta no era directa y tampoco poseía una relación directa con ninguno de los otros campos. Esto afectaba al modelo tanto en su entrenamiento como en las predicciones.

g. En resumen, se excluyeron los cinco campos mencionados dado que no contribuían al modelo y se crearon 4 nuevos atributos que fueron: *Jurisdiccion_causa_muerte, Region, Etapas_vida* y *Categoria_enfermedad,*

h. Dada la pretensión de pronosticar la cantidad de casos por provincia y por causa, se creó el campo *Jurisdiccion_causa_muerte.* El objetivo fue crear un label o etiqueta única que contemple la combinación del nombre de la provincia donde ocurrió el deceso *(jurisdicion_residencia_nombre)* y la causa de la muerte *(lista_clasificacion).*

i. El atributo *Region* corresponde a la agrupación de las provincias en las ocho regiones geográficas a la que corresponden las mismas. Las regiones son: Buenos Aires, Cuyo, Extremo Austral, Litoral, Noroeste, Pampas, Patagonia y Sierras.

j. *Etapas_vida*, dados los rangos etarios que se poseían como dato en la base de origen, se consideró como oportunidad de mejora generar un atributo que corresponda a las etapas de la vida en la cual se encontraban las personas que fallecieron. Reduciendo de este modo, el universo que analiza el modelo.

k. Por otra parte, se creó el atributo Categoria_enfermedad, donde las causas de muerte, pertenecientes al campo lista_clasificacion, fueron clasificadas en once categorías según el tipo de afección. Las once mencionadas se listan a continuación: *accidentes y otros, afecciones digestivas, afecciones respiratorias, cancer y tumores malignos, enfermedad bacterial, enfermedades cerebrovasculares, enfermedades congenitas, enfermedades por virus, otros, problemas cardiacos y circulatorios* y por último *resto de enfermedades del sistema genitourinario*. El motivo de la creación de la categoría fue crucial, debido que dependiendo de la causa de fallecimiento tenían cantidad de casos muy diferentes. En algunas causas se poseía solo un registro por año y en otros miles. Esto afectaba al modelo predictivo, generando inconsistencias en la predicción de cantidad de casos al obtener valores negativos, lo cual resultaba erróneo e inaceptable.

l. A continuación, se listan los atributos que fueron utilizados en el modelo final de predicción aplicado; anio , Jurisdiccion_causa_muerte, jurisdicion_residencia_nombre, Region, Causa, Categoria_enfermedad, Sexo, grupo_edad y Etapas_vida. La variable para predecir fue el campo denominado como Cantidad_total_casos.

Como sección complementaria se puede consultar el apartado de Apéndices, donde se encuentra un cuadro elaborado que resume los valores de uno de los atributos contemplados.

Para arribar al conjunto de datos definitivo, el proceso de transformación de datos fue prolongado y el de mayor desafío en el presente trabajo de investigación. Puesto que cada modificación sobre los mentados afectaba en gran medida los resultados que se obtenían durante cada ejecución de los modelos.

1.12 Componentes del proceso del modelo de aprendizaje automático

Para la aplicación y desarrollo de los modelos predictivos se empleó la herramienta de minería de datos Microsoft Machine Learning Studio (Microsoft, 2022). El motivo corresponde a su fácil implementación y entendimiento para el usuario, además no requiere de una computadora con elevadas capacidades de procesamiento ya que no se ejecuta de manera local, lo cual genera que se procesen rápidamente los datos y sin costos vinculados al su uso ni a una suscripción.

El modelo de datos o diagrama de proceso de entrenamiento de datos fue planteado con siete pasos. Los mismos se explican a continuación:

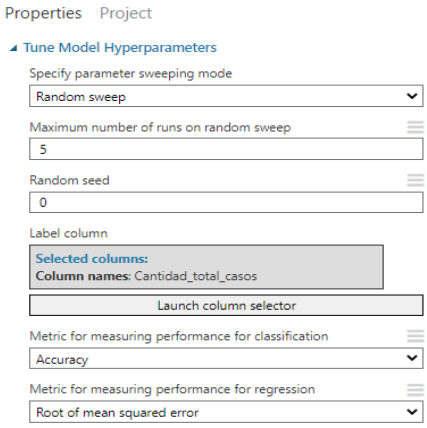
a. El primero de ellos corresponde al dataset, producto del input de la base de datos previamente depurada y correspondiente al rango de años 2005 al 2018 inclusive con el cual se pretendía el entrenamiento del modelo. La carga en sistema fue realizada a través del archivo plano que la contenía o bien conocido como archivo delimitado por comas.
b. El segundo paso implicó una transformación de datos mediante el lenguaje de consulta estructurado SQL (“Structure Query Languaje”). La misma constó de la modificación de los campos y registros no numéricos definiéndolos en letra mayúscula. Esto fue establecido debido que algunos sistemas o programas poseen *upper case sensitive words*, es decir, que reconocen como diferentes palabras a todas aquellas que solo difieren entre si al contener mayúsculas o minúsculas.
c. Continuando con el flujo del proceso, el siguiente punto pertenece al componente división de datos (“Split data”). El mismo produjo el particionamiento aleatorio de la base de datos en dos secciones. Donde una de ellas correspondía a los datos que son utilizados para entrenar al modelo predictivo. Y la parte restante, que no conformaba el entrenamiento, fue la que el sistema utilizó para testear el modelo instruido. Con dichos datos, el modelo

ejecutado predice la variable elegida y evalua si los resultados se aproximan a los valores reales. Esto es conocido como proceso de validación cruzada de retención.

d. A partir del punto previo, el esquema se conectó a dos actividades que correspondían al conjunto de parámetros conocidos como hiperparámetros (“Tune model Hiperparameters”). Estos fueron aplicados a cada modelo de regresión elegidos y ejecutaron un número limitado de combinaciones posibles para evaluar y obtener de las mismas la mejor versión de cada uno de ellos. En dicha sección se definió la variable a predecir, que correspondía a los valores de la columna Cantidad_total_casos. También fue definido el número máximo de quince (15) ejecuciones en barrido aleatorio, el cual comprende la iteración de todas las combinaciones posibles de hiperparámetros. La métrica elegida para medir el rendimiento de la regresión fue la raíz del error cuadrático medio (RMSE).

Figura 8.

Determinación y configuración de hiperparámetros para el modelo Regresión de bosque de decisión.



Nota. Modelo elaborado para la implementación de los modelos predictivos mediante Microsoft Machine Learning Studio (Microsoft, 2022).

e. Los modelos aplicados fueron dos, por una parte, el modelo de Regresión de bosque de decisión (“Decision forest regression”). Este fue configurado con; el método de remuestreo Bagging, con un entrenamiento de parámetro singular el cual se especificó con mil (1000) árboles de decisión cuya profundidad máxima por árbol fue de cien (100). El número de divisiones aleatorias por nodo elegido fue de veinte (20) cuyo mínimo de muestras por nodo hoja fue de diez (10). El segundo fue Regresión del árbol de decisión potenciado (“Boosted decision tree regression”) determinado con un entrenamiento de parámetro singular

definido con mil (1000) árboles de decisión y un máximo de veinte (20) hojas por árbol, cuya mínima cantidad de muestras por hoja nodo fue de diez (10).

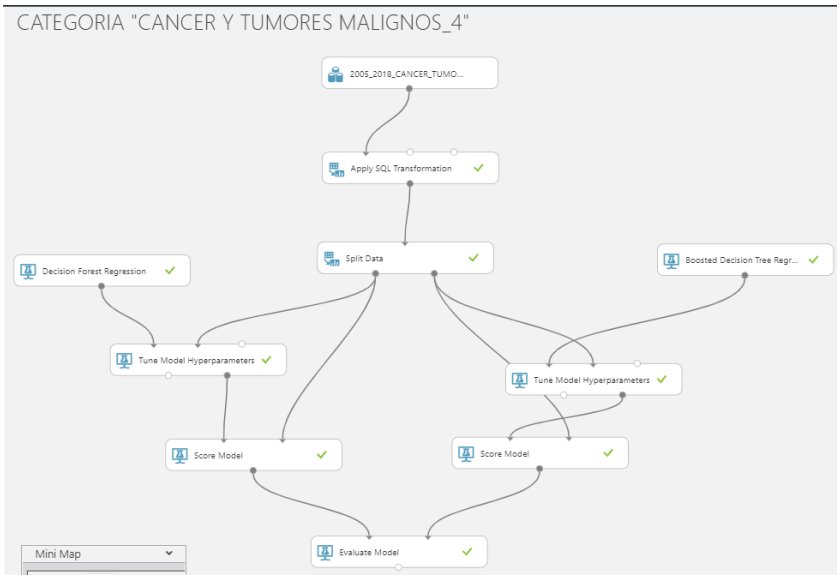
f. Luego de procesar cada modelo con las configuraciones y los hiperparámetros mencionados, se procedió a la visualización de la predicción resultante de ellos mediante la actividad denominada Score Model. Esta permite observar tanto el campo con los valores reales como de los valores pronosticados para que el científico de datos pueda tomar noción de la disparidad o congruencia entre los mismos.

g. Finalmente, el proceso concluyó con el paso de evaluación de modelo ("Evaluate model"). Este permitió tener visibilidad de los valores de la raíz del error cuadrático medio y el coeficiente de determinación y la comparabilidad de éstos sobre cada modelo.

h. Los valores seleccionados en los parámetros, informados precedentemente, fueron aquellos que proporcionaron un mejor resultado en los modelos luego de transcurrid una gran cantidad corridas, evaluación y edición del proceso. Este flujo de tareas se ejecutó por cada categoría o grupo de enfermedades y se evaluó el RMSE y R2 de cada uno de ellos para identificar su eficiencia. En la *figura 9* se puede observar la estructura del flujo de procesos previamente descripto.

Figura 9.

Modelo de aprendizaje automático para la categoría cáncer y tumores malignos.



Nota. Modelo elaborado para la implementación de los modelos predictivos mediante Microsoft Machine Learning Studio (Microsoft, 2022).

1.13 Descripción e introducción a los modelos predictivos a implementar

Los modelos de aprendizaje automático seleccionados para este estudio debían de ser apropiados para un modelo de regresión lineal, puesto que se pretendía en base a la relación entre las variables bajo análisis predecir su comportamiento y la cantidad de casos. Dada la diversidad de algoritmos posibles, opté por la implementación de dos modelos basados en árboles de decisión. La elección fue en principio orientada en gran medida por las recomendaciones otorgadas y explicadas por el artículo de selección de algoritmos para Azure Machine Learning de Microsoft (Microsoft, 2022). De todos modos, la decisión final fue producto de la realización de múltiples y variadas pruebas y ajustes en los modelos.

Uno de los modelos utilizados fue el denominado Regresión de bosque de decisión ("Decision forest regression"). La técnica de árboles de decisión es un modelo de ensamble en el que diferentes clasificadores generan un mejor modelo como un conjunto, que, al generar distintos sets de entrenamiento para el mismo modelo, éste logra reducir la volatilidad y mejorar la performance predictiva. Para esto, genera un sistema de votos en el cual cada árbol presenta sus reglas, y aquellas que aparezcan en mayor cantidad son las escogidas para el modelo final. De esta manera, reduce la probabilidad de overfitting y la varianza entre los resultados. Una de las desventajas, sin embargo, es el mayor tiempo relativo a otros modelos que no precisan de ensambles. Al modelo Regresión de bosque de decisión se le aplico "Bagging" como método de remuestreo, debido que éste genera diferentes subconjuntos del conjunto de datos de entrenamiento, seleccionando puntos de datos al azar que evitan el overfitting no deseado tal como fue descrito y es recomendado por Breiman (Breiman , 2001).

Como contrapartida, en el presente estudio se utilizó el modelo de Regresión del árbol de decisión potenciado ("Boosted decision tree regression") como método comparativo al mencionado anteriormente. Los árboles de regresión potenciados es una de varias técnicas que tienen como objetivo mejorar el rendimiento de un modelo mediante el ajuste muchos modelos y luego combinándolos para una óptima predicción (Hastie, 2008). Éste es un método de conjunto similar al "Bagging", sin embargo, en lugar de construir diversos árboles en paralelo, construye árboles secuencialmente. Lo que hace es utilizar los resultados del árbol de la secuencia anterior para identificar sus errores y construir un nuevo árbol en base a las correcciones del anterior.

Para el desarrollo del modelo de aprendizaje automático, no sólo es relevante la elección de la metodología o modelo a aplicar, sino que también resulta crucial la variación de los

parámetros y configuraciones que se aplican a cada uno de ellos, así como la repetición del proceso de evaluación y testeo.

1.14 Evaluación de los modelos predictivos aplicados

Uno de los puntos más relevantes a la hora de definir cuál es el mejor modelo, es aquel que estima correctamente la raíz del error cuadrático medio (RMSE) sobre los datos de entrenamiento. Tal como menciona Sanahuja (Sanahuja, 2021), el RMSE tiene la ventaja de tener las mismas unidades que la variable predicha, por lo que es más fácil de interpretar directamente. A su vez, esta métrica es más sensible a los valores atípicos, por lo que es más útil cuando los errores grandes son particularmente indeseables. Por otra parte, se estableció el Coeficiente de determinación (R-cuadrado). El coeficiente de determinación es la proporción de la varianza total de la variable explicada por la regresión. En otras palabras, dicho término trata de explicar la bondad del ajuste de un modelo a la variable que pretende analizar. Este coeficiente puede ofrecer valores entre 0 y 1. Si el resultado es cercano a 1, se puede indicar que el modelo y la variable que se pretende explicar se ajustan mucho. Por el contrario, si se acerca más a 0, el modelo se ajustará menos y por ende es menos fiable (Méndez ,2019).

Los modelos de aprendizaje automático que se seleccionaron contienen a su vez un conjunto de parámetros conocidos como hiperparámetros, (Obi, 2019) presenta una introducción al tema. Malik (Malik, 2020) expone que la forma más común de encontrar los valores óptimos de éstos es a través de técnicas de búsquedas de cuadrícula. Los valores para seleccionar en los hiperparámetros corresponden a: especificar el modo de barrido de parámetros, la definición del número máximo de ejecuciones en barrido aleatorio, la selección de la columna o campo a predecir, la métrica para medir el desempeño para la clasificación y la métrica para medir el rendimiento de la regresión.

Las combinaciones de parámetros fueron ajustadas en cada ejecución y evaluación del proceso, hasta arribar a la combinación de hiperparámetros con mejor rendimiento según la métrica de error RMSE.

En base a las métricas de evaluación mencionadas, fue seleccionado como mejor modelo el de *Regresión de bosque de decisión*. Y si bien, desde la perspectiva del análisis conceptual de los modelos mencionados con antelación, se esperaba obtener mejores resultados por parte del modelo de *Regresión del árbol de decisión potenciado*, no fue así a la hora de realizar

pruebas de predicción. Esto pudo ser corroborado no solo con el RMSE resultante del modelo, sino mediante validaciones adicionales.

Mediante el entrenamiento del modelo se aplicó la metodología de validación cruzada de retención ("Holdout cross-validation") sobre el rango de períodos 2005 a 2018. Kumar (Kumar, 2020) expone que la misma consiste en la división aleatoria del conjunto de datos en dos tipos, uno de entrenamiento y otro de prueba o testeo. El conjunto de datos de entrenamiento representó el 85% y finalmente se puso a prueba al modelo con la partición faltante que representa el 15%. El RMSE producto de esta primera validación del modelo, indicó que la metodología de *Regresión del árbol de decisión potenciado* era el más recomendable indicando predicción de valores muy próximos a los reales y con un R2 muy próximo a 1. Pero en base al criterio técnico, es muy usual que ocurra que el Modelo durante el entrenamiento de cierta manera "memorice" los valores e indique que es un buen modelo cuando en términos prácticos y de aplicación no cumplen con las expectativas. Por ello, adicionalmente a la validación cruzada de retención, durante el entrenamiento de las once bases de datos del rango de años 2005 al 2018. Se planteó la necesidad de hacer un control adicional mediante la metodología validación cruzada en series temporales ("Time Series cross-validation") (Shrivastava ,2020). En este estudio se tomó el modelo ya entrenado y preparado, y se aplicó el mismo para predecir los valores del período 2019, sobre el cual ya se contaba con valores reales. El objetivo fue evaluar la proximidad y certeza del modelo de aprendizaje automático en un contexto de datos reales.

Durante la implementación y pruebas de los modelos y versiones de bases de datos, producto de los ajustes y limpieza de datos mencionados anteriormente, se obtuvieron valores negativos en las cantidades de fallecidos que el modelo predecía lo cual resultaba erróneo e inaceptable. La problemática impulsó a la conclusión de que los entrenamientos y predicciones se debían de particionar en once (11) bases de datos, puesto que la base original efectivamente se encontraba desbalanceada. Esto ocurría debido que dependiendo de la causa de fallecimiento había registros de cantidades de casos muy disímiles. En determinadas categorías, tales como; "enfermedades por virus" cuyas causas específicas como enfermedad de Chagas o hepatitis viral, se poseía menos de diez casos por año. Y por otra parte otros como la categoría denominada "problemas cardíacos" o la de "afecciones respiratorias" donde las causas especificas eran insuficiencia cardíaca y neumonía poseían miles registros anuales. Sobre dicho aspecto fue basado el Método de conjunto equilibrado ("Balanced Ensemble Methods"), una

recomendación de Charu Makhijani (Makhijani, 2020). La partición fue basada en el atributo de las once categorías que se había creado como producto de la agrupación de las enfermedades. Finalmente se entrenaron las diferentes bases de datos y mediante dicho proceso se obtuvo el resultado final donde no resultó ninguna predicción con números negativos.

Los valores finales fueron evaluados en cada una de las predicciones por categoría mediante las métricas de error mencionadas. También se analizaron las cantidades resultantes de manera conjunta, es decir agrupando nuevamente y analizando las predicciones de todas las categorías como un todo. Como resultado de esa segunda validación y tal como se mencionó previamente, los mejores resultados de RMSE y R2 resultaron ser los provenientes del modelo Regresión de bosque de decisión. En la figura 10, se pueden observar en la parte superior los valores obtenidos por categoría de RMSE y R2 mediante la metodología validación cruzada de retención del periodo 2005 a 2018 en color naranja. Y correspondiente a los resultados de la evaluación en series temporales del período 2019.

Figura 10.

Resultados de la metodología validación cruzada de retención y validación cruzada en series temporales

Orden	Categoría	Entrenamiento 85% y 15%		Resultado predictivo 2019	
		RMSE	R2	RMSE	R2
1	ACCIDENTES OTROS	4	96%	6	90%
2	AFECCIONES DIGESTIVAS	9	96%	6	98%
3	AFECCIONES RESPIRATORIAS	8	96%	31	98%
4	CANCER Y TUMORES MALIGNOS	8	97%	10	95%
5	ENFERMEDAD BACTERIAL	9	99%	11	100%
6	ENFERMEDADES CEREBROVASCULARES	12	99%	20	99%
7	ENFERMEDADES CONGENITAS	20	98%	25	95%
8	ENFERMEDADES POR VIRUS	2	97%	3	91%
9	OTROS	5	95%	9	95%
10	PROBLEMAS CARDIACOS Y CIRCULATORIOS	35	98%	39	97%
11	RESTO DE ENFERMEDADES DEL SISTEMA GENITOURINARIO	7	99%	17	97%
	Modelo completo			19	98%

Nota. Tabla elaborada mediante la utilización de Excel y la obtención de datos mediante Microsoft Machine Learning Studio (Microsoft, 2022).

1.15 Resultados de la metodología y su utilidad para el organismo

Desde una perspectiva de tipo organizacional, al estudiar los resultados obtenidos del método predictivo, se considera que puede brindarle al Ministerio de Salud un útil panorama de tendencias y comportamiento de las variables sanitarias de la sociedad. Si bien, los mismos nunca serán exactos, ya que pueden ocurrir casos aislados como lo fue el COVID-19 como principal causa de muerte en el año 2020, el modelo planteado analiza y predice, sobre los registros históricos, la cantidad de fallecimientos por causa, por grupo etario, por sexo, y a su vez por las provincias del país al que pertenecen. Adicional a la predicción de casos mencionada, el presente estudio puede fomentar a las organizaciones públicas a vincularse con diferentes tecnologías y metodologías de análisis.

Cabe destacar que, en la realidad, la métrica de evaluación de modelos va a deber tener en cuenta la variable temporal para poder generar la predicción del período anual siguiente. Dado que el presente trabajo se realizó mediante una base de datos pública con apertura anual y una limitada cantidad de atributos, se considera que probablemente la Dirección de Estadística e Información en Salud, cuente con mayor grado de información sobre las personas fallecidas, pudiendo así proveer datos representativos a la base de datos e incrementar la eficiencia de los modelos predictivos que mejor se ajusten las necesidades de la administración.

Conclusión

El estudio desarrollado y descripto en el presente informe dio inicio sobre la base de análisis de los informes estadísticos de la mortalidad de los ciudadanos de la Argentina para un período anual determinado. Estos fueron y son reportados anualmente por la Dirección de Estadística e Información en Salud perteneciente al Ministerio de Salud argentino. Desde la perspectiva de un analista de datos, los informes comentados fueron desarrollados mediante la utilización de metodologías de análisis de tipo descriptivas y simples, donde no se pudo identificar un estudio de mayor alcance. Como resultado de ello, se vio la oportunidad desarrollar la aplicación de las metodologías de análisis multivariado y de aprendizaje automático, sobre la base de datos pública que comprende el registro anual de los fallecidos en la Argentina, proveída por el Ministerio de Salud.

Este estudio fue una gran oportunidad para profundizar el conocimiento de la metodología de análisis e implementar distintos enfoques de estudio en una problemática social como lo es la ocurrencia y cantidad de fallecimientos por provincia y las causas que las producen.

La aplicación de la metodología de análisis de componentes principales permitió visualizar la correlación de las enfermedades de índole respiratoria, la diabetes mellitus y las afecciones cardíacas. Mediante la técnica de clústeres, se pudo observar no solo la correlación de las principales causas de fallecimiento y las enfermedades, sino que también proporcionó un análisis respecto a la distribución geográfica y su relación con las 24 jurisdicciones o provincias de la Argentina. Las componentes principales representadas por diabetes y afecciones cardíacas fueron en su mayoría explicadas por la región sur y norte del país. Por otra parte, al analizar la agrupación de la región central del país, principalmente; la Ciudad Autónoma de Buenos Aires, la Pampa, Buenos Aires, se observó que poseen una fuerte correlación y distribución con las enfermedades de índole respiratoria. Para continuar con la línea de análisis sobre lo mencionado, cabe destacar que las causas más importantes de cardiopatía son una dieta malsana, la inactividad física, el consumo de tabaco y el consumo nocivo de alcohol. Los efectos de los factores de riesgo comportamentales pueden manifestarse en las personas en forma de hipertensión arterial, hiperglucemia, hiperlipidemia y sobrepeso u obesidad. Respecto a afecciones respiratorias, podría considerarse el factor de polución, tala de árboles y aglomeración, como factor común de las grandes ciudades. Los indicadores mencionados, pueden contribuirle al Ministerio de Salud y la administración que la acompaña, un parámetro de alerta ya que podría evidenciar puntos clave para atender en la sociedad argentina.

En referencia a la metodología de aprendizaje automático y los resultados obtenidos a partir del modelo Regresión de bosque de decisión aplicado, se toman como gratos los resultados de dicha metodología en el presente estudio y se considera que da respuesta a la problemática planteada. La eficiencia del modelo fue evaluada mediante los valores de la raíz del error cuadrático medio y el coeficiente de correlación obtenidos a partir del entrenamiento y utilización del modelo predictivo para cada una de las once categorías por afección causante del deceso. Mediante la proximidad observada, entre los valores estimados para el período 2019 y los valores reales del mismo año, se cree que esta metodología efectivamente puede brindarle al Ministerio de Salud un útil panorama de tendencias y comportamiento de las variables sanitarias de la sociedad. Al poder predecir y contar con cifras de mortalidad posibles, el organismo tendrá visibilidad para establecer planes estratégicos que lo ayuden atender las necesidades de la sociedad en cada grupo etario, según el sexo y región del país y establecer medidas preventivas. Un ejemplo de ello podría ser que la administración del Ministerio focalice el estudio en las principales causas de fallecimiento obtenidas a partir del modelo predictivo explicado, e informe a la sociedad sobre los síntomas previos de éstas, para que puedan prestar la atención médica necesaria con antelación.

Si bien, los valores predichos nunca serán exactos ni otorgarán soluciones “mágicas”, ya que pueden ocurrir casos aislados como lo fue el COVID-19, el modelo planteado predijo una cantidad razonable de fallecimientos por causa, por grupo etario, por sexo, y a su vez por las provincias para el año 2019 muy próximas a las reales para el período.

La aplicación de algoritmos de aprendizaje automático y metodologías de análisis multivariado en entornos organizacionales, brindaron la oportunidad de obtener una nueva perspectiva y una mayor comprensión del funcionamiento interno de los modelos algorítmicos y su influencia sobre el análisis de los datos. Las bases conceptuales y los fundamentos teóricos fueron imprescindibles para responder interrogantes referentes a los diversos modelos y metodologías, principalmente para definir para qué tipo de problemas resultó más conveniente la implementación de estos. Sin embargo, muchos modelos tienen componentes matemáticos y estadísticos que requirieron de la complementación de la aplicación práctica para poder comprenderlos. El fin del estudio, fue otorgar un enfoque distintivo y ampliar el panorama del organismo para anticiparse las jurisdicciones a las tendencias de decesos y compatibilizar de manera adecuada sus planes de acción.

A continuación, se presentan algunas propuestas para futuras líneas de investigación relacionadas con el trabajo de investigación, así como también posibles mejoras o enfoques complementarios. Dado que el presente trabajo se realizó mediante una base de datos pública con apertura anual y una limitada cantidad de atributos, se considera que probablemente la DEIS, correspondiente al Ministerio de Salud argentino, cuente con mayor grado de información sobre las personas fallecidas. El organismo público quizás al contar con experiencia en la materia y un nivel granular de información, este podría mejorar ampliamente la metodología propuesta y arribar a resultados de mayor precisión. Desde la perspectiva de gestor de datos, se proponen algunas sugerencias adicionales para incluir a la base de datos. Algunos posibles aspectos tales como la inclusión de datos sobre antecedentes de enfermedad familiares, nivel de ingresos o de educación como ejemplos, quizás podrían resultar de utilidad para realizar atributos representativos vinculados a las causas de los decesos registrados. También se plantea la aplicación de un enfoque similar al inferencial, pero con el objetivo de evaluar e interpretar los modelos de aprendizaje automático, observando la contribución y relación de las variables dentro de los modelos al momento de generar predicciones.

Referencias bibliográficas

Aldás, J. y Uriel E. (2017). *Análisis multivariante aplicado con R (2da edición).* Ediciones Paraninfo.

Breiman (enero, 2001). *RANDOM FORESTS*. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Guevel, C. (abril ,2021). *DEIS. Boletín estadísticas vitales – Argentina Año 2019.* https://www.argentina.gob.ar/sites/default/files/serie5numero63.pdf

Catena A, Ramos M, y Trujillo H. (2003). *Análisis multivariado, un manual para investigadores.* Editorial biblioteca nueva.

Makhijani, C. (29 de septiembre de 2020). *Ensemble Learning Techniques*. Towards data science. https://towardsdatascience.com/ensemble-learning-techniques-6346db0c6ef8

Dirección de Estadística e Información en Salud DEIS (18 de marzo de 2019). *Defunciones ocurridas y registradas en la República Argentina*. Ministerio de Salud. https://datos.gob.ar/dataset/salud-defunciones-ocurridas-registradas-republica-argentina

División de Diabetes Aplicada (abril de 2021). *La diabetes y su corazón*. CDC Centro para el Control y la Prevención de Enfermedades. https://www.cdc.gov/diabetes/spanish/resources/features/diabetes-and-heart.html

Amat, R. (septiembre de 2017). *Gráficos de R. Clustering y heatmaps: aprendizaje no supervisado.* Ciencia de datos. https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps

Kumar, S. (septiembre de 2020*). Understanding 8 types of Cross-Validation*. Towards Data Science. https://towardsdatascience.com/understanding-8-types-of-cross-validation-80c935a4976d

Logicalis, Architects of change (mayo de 2015). *Modelos predictivos: reforzando el valor de una buena decisión*. Logicalis, Architects of change. https://blog.es.logicalis.com/analytics/modelos-predictivos-reforzando-el-valor-de-una-buena-decision#:~:text=El%20verdadero%20desaf%C3%ADo%20para%20los%20modelos%20predictivos%20es,predecir%20los%20datos%20que%20todav%C3%ADa%20no%20se%20tienen.

Malik, F. (18 de febrero de 2020). *What is Grid Search?. Medium*. FinTechExplained. https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a

Méndez, D. (01 de abril de 2019). *Definición de Coeficiente de determinación*. Economía Simple .net. https://www.economiasimple.net/glosario/coeficiente-de-determinacion

Microsoft (12 de agosto de 2022). *Microsoft Machine Learning Studio (classic).* Microsoft. Recuperado el 11 de noviembre de 2022 de https://learn.microsoft.com/en-us/azure/machine-learning/migrate-overview

Obi, B. (30 de octubre de 2019). *Model Parameters and Hyperparameters in Machine Learning - What is the difference?*. Towards Data Science. https://towardsdatascience.com/model-parameters-and-hyperparameters-inmachine-learning-what-is-the-difference-702d30970f6

OMS Organización Mundial de la Salud (11 de junio de 2021). *Cardiovascular diseases (CVDs).* Centro de prensa. Enfermedades. Recuperado el 10 de noviembre de 2021 de https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

OPS, Organización Panamericana de la Salud (mayo de 2021). *Diabetes prevención y tratamiento*. OPS. Recuperado el 02 de noviembre de 2021 de https://www.paho.org/es/temas/diabetes

Personal de Mayo Clinic (13 de junio de 2020). *Dificultad para respirar*. Mayo Clinic. https://www.mayoclinic.org/es-es/symptoms/shortness-of-breath/basics/causes/sym-20050890

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Recuperado el 10 de noviembre de 2021 de https://www.R-project.org/

RStudio Team (2021). *RStudio: Integrated Development for R*. RStudio. Recuperado el 10 de noviembre de 2021 de http://www.rstudio.com/

Sanahuja, P. (5 de enero de 2021*). Métricas de evaluación de rendimiento para predicciones de series temporales*. Pol Martí Sanahuja. https://polmartisanahuja.com/metricas-de-evaluacion-de-rendimiento-para-predicciones-de-series-temporales/

Shrivastava, S. (14 de enero de 2020). *Cross Validation in Time Series*. Medium.com. https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4#:~:text=Cross%20Validation%20on%20Time%20Series,for%20the%20forecasted%20data%20points.

Zach (04 de noviembre de 2020). *What is Overfitting in Machine Learning?*.Statology https://www.statology.org/overfitting-machine-learning/

Zita (07 de marzo de 2022). *5 Effective Ways to Improve the Accuracy of Your Machine Learning Models*. Towards Data Science. https://towardsdatascience.com/5-effective-ways-to-improve-the-accuracy-of-your-machine-learning-models-f1ea1f2b5d65

Anexos/ apéndices

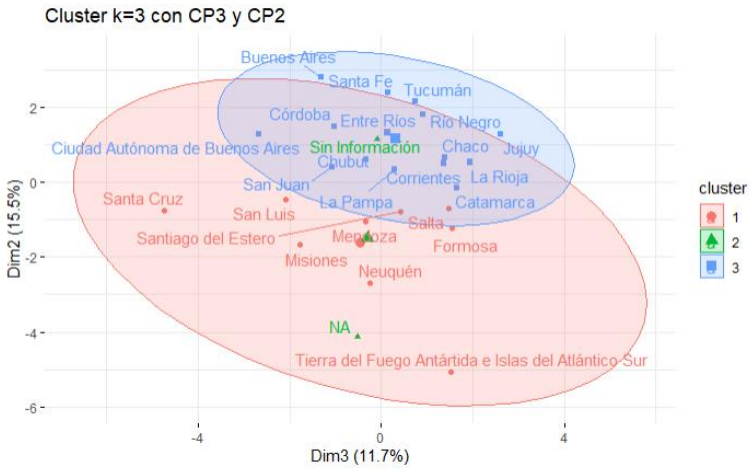
Anexos

Apéndices

A continuación, se añaden gráficos de producción propia no incluidos en el cuerpo del trabajo del segundo apartado, debido a la limitación en la extensión de este.

Figura 1.

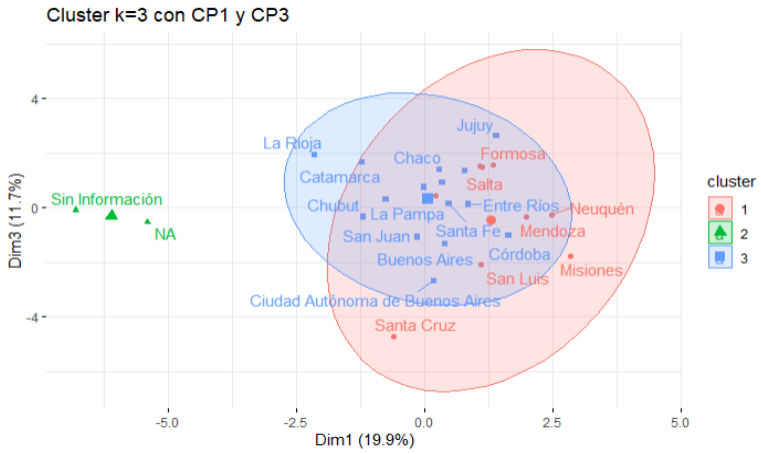
Gráfico de clústeres, dispuesto en 2 dimensiones, siendo Dim2 la componente principal 2 y Dim3 la componente principal 3.



Nota. Gráfico elaborado mediante la utilización del software Rstudio (RStudio Team, 2021*).*

Figura 2.

Gráfico de clústeres, dispuesto en 2 dimensiones, siendo Dim2 la componente principal 2 y Dim3 la componente principal 3.



Nota. Gráfico elaborado mediante la utilización del software Rstudio (RStudio Team, 2021*).*

A continuación, se observan diferentes cuadros de producción propia que resumen los valores de cada atributo y algunos casos las agrupaciones de estos como producto de nuevos atributos. El mismo corresponde a la implementación de Metodologías de aprendizaje automático desarrollado en el tercer apartado del presente.

Figura 3.

Atributo año, el contempla rango de años del 2005 al 2019. Donde el 2019 fue excluido durante el entrenamiento del modelo, para poder luego hacer la predicción de este y evaluar que el modelo no esté generando "overfitting" durante el entrenamiento.

Anio
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

Nota. Tabla elaborada mediante la utilización de la herramienta Excel, del paquete Office de Microsoft.

Figura 4.

Atributos región y jurisdicion_residencia_nombre . La primera fue producto de indicar la región a la cual pertenecen las 24 jurisdicciones o provincias de la Argentina.

Región	jurisdicion_residencia_nombre
BUENOS AIRES	Ciudad Autonoma de Buenos Aires
CUYO	Mendoza
	San Juan
	San Luis
EXTREMO AUSTRAL	Tierra del Fuego Antartida e Islas del Atlantico Sur
LITORAL	Chaco
	Corrientes
	Entre Rios
	Formosa
	Misiones
	Santa Fe
NOROESTE	Catamarca
	Jujuy
	La Rioja
	Salta
	Santiago del Estero
	Tucuman
PAMPAS	Buenos Aires
	La Pampa
PATAGONIA	Chubut
	Neuquen
	Rio Negro
	Santa Cruz
SIERRAS	Cordoba

Nota. Tabla elaborada mediante la utilización de la herramienta Excel, del paquete Office de Microsoft.

Figura 5.

Lista de Categoría_enfermedad y causa. La primera fue producto de agrupar las causas según el tipo de enfermedad o afección causante.

Categoria_enfermedad	Causa
ACCIDENTES Y OTROS	ACCIDENTES DE TRANSPORTE
ACCIDENTES Y OTROS	ACCIDENTES POR EXPLOSIONES EXPOSICION A CORRIENTE ELECTRICA HUMO FUEGO Y LLAMAS
ACCIDENTES Y OTROS	AGRESIONES
ACCIDENTES Y OTROS	AHOGAMIENTO ACCIDENTAL
ACCIDENTES Y OTROS	CAIDAS ACCIDENTALES
ACCIDENTES Y OTROS	ENVENENAMIENTO ACCIDENTAL
ACCIDENTES Y OTROS	OTROS ACCIDENTES
ACCIDENTES Y OTROS	SECUELAS DE CAUSAS EXTERNAS
ACCIDENTES Y OTROS	SUICIDIO
ACCIDENTES Y OTROS	TRASTORNOS MENTALES Y DEL COMPORTAMIENTO DEBIDOS AL USO DE ALCOHOL
AFECCIONES DIGESTIVAS	DESNUTRICION
AFECCIONES DIGESTIVAS	DIABETES MELLITUS
AFECCIONES DIGESTIVAS	DIARREA Y GASTROENTERITIS DE PRESUNTO ORIGEN INFECCIOSO
AFECCIONES DIGESTIVAS	ENFERMEDAD ALCOHOLICA DEL HIGADO
AFECCIONES DIGESTIVAS	ENTERITIS Y COLITIS NO INFECCIOSAS
AFECCIONES DIGESTIVAS	HERNIA ILEO PARALITICO Y OBSTRUCCION INTESTINAL SIN HERNIA
AFECCIONES DIGESTIVAS	OTRAS ENFERMEDADES DEL HIGADO
AFECCIONES DIGESTIVAS	OTRAS ENFERMEDADES INFECCIOSAS INTESTINALES E INTOXICACIONES ALIMENTARIAS
AFECCIONES DIGESTIVAS	OTRAS ENFERMEDADES INFECCIOSAS Y PARASITARIAS
AFECCIONES DIGESTIVAS	RESTO CIERTAS AFECCIONES ORIGINADAS EN EL PERIODO PERINATAL
AFECCIONES DIGESTIVAS	RESTO ENFERMEDADES DEL SISTEMA DIGESTIVO
AFECCIONES DIGESTIVAS	ULCERA PEPTICA
AFECCIONES RESPIRATORIAS	ASMA
AFECCIONES RESPIRATORIAS	BRONQUITIS Y OTRAS ENFERMEDADES PULMONARES OBSTRUCTIVAS CRONICAS
AFECCIONES RESPIRATORIAS	ENFERMEDAD CARDIOPULMONAR Y DE LA CIRCULACION PULMONAR
AFECCIONES RESPIRATORIAS	ENFISEMA
AFECCIONES RESPIRATORIAS	NEUMONIA
AFECCIONES RESPIRATORIAS	RESTO DE ENFERMEDADES DEL SISTEMA RESPIRATORIO
AFECCIONES RESPIRATORIAS	TRASTORNOS RESPIRATORIOS Y CARDIOVASCULARES
AFECCIONES RESPIRATORIAS	TUBERCULOSIS OTRAS FORMAS
AFECCIONES RESPIRATORIAS	TUBERCULOSIS RESPIRATORIA
CANCER Y TUMORES MALIGNOS	ENFERMEDAD DE HODGKIN
CANCER Y TUMORES MALIGNOS	LEUCEMIA
CANCER Y TUMORES MALIGNOS	MELANOMA MALIGNO DE LA PIEL
CANCER Y TUMORES MALIGNOS	OTROS TUMORES MALIGNOS DEL TEJIDO LINFATICO ORGANOS HEMATOPOYETICOS Y TEJIDOS AFINES
CANCER Y TUMORES MALIGNOS	RESTO DE TUMORES (IN SITU BENIGNOS DE COMPORTAMIENTO INCIERTO O DESCONOCIDO)
CANCER Y TUMORES MALIGNOS	RESTO DE TUMORES MALIGNOS
CANCER Y TUMORES MALIGNOS	TUMOR BENIGNO DEL ENCEFALO Y DE OTRAS PARTES DEL SISTEMA NERVIOSO CENTRAL
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DE LA LARINGE
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DE LA MAMA
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DE LA PROSTATA
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DE LA TRAQUEA DE LOS BRONQUIOS Y DEL PULMON
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DE LA VEJIGA URINARIA
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL HIGADO Y DE LAS VIAS BILIARES INTRAHEPATICAS
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL COLON
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL CUELLO DEL UTERO
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL CUERPO DEL UTERO
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL ENCEFALO Y SISTEMA NERVIOSO CENTRAL
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL ESOFAGO
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL ESTOMAGO
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL LA UNION RECTOSIGMOIDEA RECTO ANO Y CONDUCTO ANAL
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL LABIO DE LA CAVIDAD BUCAL Y DE LA FARINGE
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL OVARIO
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL PANCREAS
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL RINION Y PELVIS RENAL
CANCER Y TUMORES MALIGNOS	TUMOR MALIGNO DEL UTERO PARTE NO ESPECIFICADA

Categoria_enfermedad	Causa
ENFERMEDAD BACTERIAL	ENFERMEDADES ESTREPTOCOCICAS
ENFERMEDAD BACTERIAL	INFECCION MENINGOCOCICA
ENFERMEDAD BACTERIAL	MENINGITIS
ENFERMEDAD BACTERIAL	OTRAS ENFERMEDADES BACTERIANAS
ENFERMEDAD BACTERIAL	SEPTICEMIAS
ENFERMEDAD BACTERIAL	TETANOS
ENFERMEDAD BACTERIAL	TETANOS NEONATAL
ENFERMEDAD BACTERIAL	TOS FERINA
ENFERMEDADES CEREBROVASCULARES	ENFERMEDADES CEREBROVASCULARES
ENFERMEDADES CONGENITAS	ARTROPATIAS
ENFERMEDADES CONGENITAS	ENFERMEDAD DE ALZHEIMER
ENFERMEDADES CONGENITAS	HIDROCEFALIA CONGENITA Y ESPINA BIFIDA
ENFERMEDADES CONGENITAS	MALFORMACIONES CONGENITAS DEL SISTEMA CIRCULATORIO
ENFERMEDADES CONGENITAS	OTROS SINTOMAS SIGNOS Y HALLAZGOS ANORMALES CLINICOS Y DE LABORATORIO NO CLASIFICADOS EN OTRA PARTE
ENFERMEDADES CONGENITAS	RESTO MALFORMACIONES CONGENITAS DEFORMACIONES Y ANOMALIAS CROMOSOMICAS
ENFERMEDADES CONGENITAS	TRASTORNOS RELACIONADOS CON GESTACION CORTA Y BAJO PESO AL NACER
ENFERMEDADES POR VIRUS	ENFERMEDAD DE CHAGAS
ENFERMEDADES POR VIRUS	ENFERMEDAD POR VIRUS DE INMUNODEFICIENCIA HUMANA
ENFERMEDADES POR VIRUS	FIEBRES VIRALES TRANSMITIDAS POR ARTROPODOS
ENFERMEDADES POR VIRUS	HEPATITIS VIRAL
ENFERMEDADES POR VIRUS	INFECCION DEL SISTEMA NERVIOSO CENTRAL POR VIRUS LENTO
ENFERMEDADES POR VIRUS	INFLUENZA
ENFERMEDADES POR VIRUS	MALARIA
ENFERMEDADES POR VIRUS	OTRAS ENFERMEDADES VIRALES
OTROS	ABORTO
OTROS	ANEMIAS
OTROS	CAUSAS OBSTETRICAS DIRECTAS
OTROS	CAUSAS OBSTETRICAS INDIRECTAS
OTROS	CAUSAS POST-OBSTETRICAS
OTROS	COMPLICACIONES DE LA ATENCION MEDICA Y QUIRURGICA
OTROS	ENFERMEDADES DE LA PIEL Y DEL TEJIDO SUBCUTANEO
OTROS	ENFERMEDADES DEL OIDO Y DE LA APOFISIS MASTOIDES
OTROS	ENFERMEDADES DEL OJO Y SUS ANEXOS
OTROS	EVENTOS DE INTENCION NO DETERMINADA
OTROS	HIPERPLASIA DE LA PROSTATA
OTROS	INFECCION CON MODO DE TRANSMISION PREDOMINANTEMENTE SEXUAL
OTROS	INTERVENCION LEGAL Y OPERACIONES DE GUERRA
OTROS	OTRAS ENFERMEDADES INFLAMATORIAS DEL SISTEMA NERVIOSO CENTRAL
OTROS	RESTO DE ENFERMEDADES DEL SISTEMA MUSCULAR Y DEL TEJIDO CONJUNTIVO
OTROS	RESTO DE ENFERMEDADES DEL SISTEMA NERVIOSO
OTROS	RESTO DE ENFERMEDADES ENDOCRINAS NUTRICIONALES Y METABOLICAS
OTROS	RESTO DE TRASTORNOS MENTALES Y DEL COMPORTAMIENTO
OTROS	RESTO ENF. DE LA SANGRE Y ORG. HEMATOPOYETICOS Y CIERTOS TRASTORNOS MECANISMO DE INMUNIDAD
OTROS	SINDROME DE LA MUERTE SUBITA INFANTIL
PROBLEMAS CARDIACOS Y CIRCULATORIOS	ENFERMEDADES DE LAS ARTERIAS ARTERIOLAS Y VASOS CAPILARES
PROBLEMAS CARDIACOS Y CIRCULATORIOS	ENFERMEDADES HIPERTENSIVAS
PROBLEMAS CARDIACOS Y CIRCULATORIOS	FIEBRE REUMATICA AGUDA Y ENFERMEDADES CARDIACAS REUMATICAS CRONICAS
PROBLEMAS CARDIACOS Y CIRCULATORIOS	FLEBITIS EMBOLIAS Y TROMBOSIS VENOSAS
PROBLEMAS CARDIACOS Y CIRCULATORIOS	INFARTO AGUDO DEL MIOCARDIO
PROBLEMAS CARDIACOS Y CIRCULATORIOS	INSUFICIENCIA CARDIACA
PROBLEMAS CARDIACOS Y CIRCULATORIOS	OTRAS ENFERMEDADES ISQUEMICAS DEL CORAZON
PROBLEMAS CARDIACOS Y CIRCULATORIOS	OTRAS FORMAS DE ENFERMEDADES DEL CORAZON
PROBLEMAS CARDIACOS Y CIRCULATORIOS	RESTO DE ENFERMEDADES DEL SISTEMA CIRCULATORIO
RESTO DE ENFERMEDADES DEL SISTEMA GENITOURINARIO	ENFERMEDADES RENALES Y DEL URETER
RESTO DE ENFERMEDADES DEL SISTEMA GENITOURINARIO	RESTO DE ENFERMEDADES DEL SISTEMA GENITOURINARIO

Nota. Tabla elaborada mediante la utilización de la herramienta Excel, del paquete Office de Microsoft.

Figura 6.

Atributos Etapas_vida y grupo_edad. El primero corresponde a la definición de cada etapa o momento de la vida de los humanos. La mentada fue creada de acuerdo con los rangos o grupos de edades registrados en cada caso de ciudadano fallecido en la Argentina.

Etapas_vida	grupo_edad
Adolescencia	15 a 24
Adultez	25 a 34
	35 a 44
	45 a 54
	55 a 64
Ancianidad	65 a 74
	75 y mas
Infancia	1 a 4
	Menores de 1 anio
Niniez	5 a 14
NA	Sin especificar

Nota. Tabla elaborada mediante la utilización de la herramienta Excel, del paquete Office de Microsoft.

Figura 7.

El atributo Sexo; contempla los valores: femenino, masculino y desconocido.

grupo_edad
desconocido
femenino
masculino

Nota. Tabla elaborada mediante la utilización de la herramienta Excel, del paquete Office de Microsoft.

Reporte del Mentor sobre el Trabajo Final de Especialización de Solange R.E. Franco:

"ANÁLISIS Y PREDICCIÓN DE LA CANTIDAD DE FALLECIDOS POR CAUSA Y PROVINCIA

Implementación de modelos predictivos sobre los datos públicos del Ministerio de Salud Argentino 2005-2019"

Mentora: Nélida Mónica Cantoni Rabolini

Este trabajo plantea el problema de determinar las principales causas de fallecimiento en la Argentina y su predicción. Considero que dicho problema es muy relevante en el contexto de una organización pública como lo es el Ministerio de Salud de Argentina.

Se observa que el problema está identificado y definido correctamente. Se realiza una descripción de la situación del problema y qué se pretende resolver con este trabajo. El interrogante es consistente con la problemática planteada.

El objetivo general del trabajo es realizar una ponderación de las principales causas de fallecimiento en Argentina utilizando métodos multivariados y métodos analíticos predictivos.

El planteo del problema y el objetivo se encuentran articulados, presentan coherencia interna y corresponden con los contenidos de la especialización que está realizando.

El tema elegido resulta interesante y relevante para aplicar dentro del contexto de la salud de un país. Se trabajó con datos reales y actualizados, que corresponden a una base de datos abierta y están disponibles para su utilización.

En el desarrollo del trabajo se observa la fundamentación del problema, el procesamiento de datos, la aplicación de las diferentes metodologías, los resultados obtenidos, la conclusión articulada con el objetivo del trabajo y la bibliografía actualizada.

Un aspecto importante para la realización de este trabajo es la etapa de preprocesamiento de datos que incluye el proceso de limpieza, selección y transformación de atributos para poder aplicar los métodos de aprendizaje automático. Se presenta un detalle de todos los procedimientos realizados en el trabajo que son consistentes con los contenidos académicos desarrollados en la especialización. Asimismo, se presenta la aplicación de las metodologías abordadas en las diferentes asignaturas utilizando herramientas informáticas adecuadas a cada tema.

El trabajo resulta muy interesante y se puede transferir para que sea consultado por organizaciones tanto públicas como privadas del ámbito de la salud.