Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

COVID-19 en México - Proyecto Final

Diplomado Ciencia de Datos - Módulo V

Autor:
André Marx Puente Arévalo

Profesor:
José Gustavo Fuentes Cabrera

1 / Septiembre / 2021

Índice

1. Introducción

En la actualidad, el mundo entero ha sido doblegado por la aparición de un nuevo virus denominado coronavirus SARS-Cov-2, que provoca una enfermedad infecciosa llamada COVID-19. Sus orígenes provienen de China, posteriormente se extendió a todos los continentes del mundo, provocando una pandemia a nivel mundial.

La mayoría de las personas infectadas con el virus COVID-19 experimentarán una enfermedad respiratoria leve a moderada y se recuperarán sin necesidad de un tratamiento especial. Las personas mayores y aquellas con problemas médicos subyacentes como enfermedades cardiovasculares, diabetes, enfermedades respiratorias crónicas y cáncer tienen más probabilidades de desarrollar enfermedades graves.

Por el momento, la mejor manera de prevenir y ralentizar la transmisión es estar bien informado sobre el virus COVID-19, la enfermedad que causa, cómo se propaga y vacunándose. El virus se transmite principalmente a través de gotitas de saliva o secreciones nasales cuando una persona infectada tose o estornuda.

Se han reportado aproximadamente 216 millones de casos confirmados alrededor del mundo, de los cuales, 4.5 millones han sido muertes confirmadas según la Organización Mundial de la Salud para finales de agosto 2021.

1.1. Acotamiento del Problema

Dado el contexto anterior, en el presente proyecto se plantea realizar limpieza y análisis estadístico de un conjunto de datos, el cual, es proporcionado por el gobierno mexicano, contiene los registros hasta el **31 de octubre del 2020** de casos diarios asociados a COVID-19 a nivel federal. Cuenta con contenido desagregado por sexo, edad, nacionalidad, padecimientos asociados, entre otros. En un principio se cuentan con 2,403,499 registros y 38 columnas.

Se seleccionó México como objeto de estudio porque es uno de los países que más afectaciones ha sufrido por el virus en muchos ámbitos como en la pérdida de vidas humanas o económicamente hablando y muchos otros, a su vez es el en país en el que radico actualmente y este virus ha llegado a cambiar nuestro estilo de vida de una forma inimaginable.

Ahora bien, resulta de interés el poder resumir la información que contienen los datos con pocas cifras o gráficas, ¿cómo cambian los datos tras su procesamiento? ¿se puede medir el porcentaje de pacientes por cada estado? En el presente proyecto se pretende dar solución a estos problemas con los datos anteriormente mencionados, al igual crear una tabla con la información necesaria y preparada para desarrollar modelos de machine learning, uno que sea capaz de predecir la mortalidad de un paciente que tiene COVID y otros padecimientos que lo hacen del sector vulnerable. Se clasificará como sobreviviente (0) o falla (1). Otro que sea capaz de obtener segmentos o grupos sobre nuestro conjunto de datos, esto para lograr categorizar a los pacientes.

2. Análisis Exploratorio Previo al Procesamiento de Datos

La finalidad del Análisis Exploratorio es examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas. Proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de datos, tratamiento y evaluación de datos ausentes, identificación de casos atípicos y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes.

2.1. Diccionario de Datos

Dentro del conjunto de datos brindado en el portal de "Datos Abiertos" se encuentran los siguientes datos que han sido tipificados entre:

- Variables Continuas (c)
- Variables Categóricas (v)
- Variables de Texto (t)
- Variables de Fecha (d)

Valriable	Tipo	Descripción
FECHA_ACTUALIZACION	Fecha	La base de datos se alimenta diariamente, esta variable permite identificar la fecha de la ultima actualización
ID_REGISTRO	Texto	Cadena identificadora del caso.
ORIGEN	Categórica	La vigilancia centinela se realiza a través del sistema de unidades de salud monitoras de enfermedades respiratorias (USMER). Las USMER incluyen unidades médicas del primer, segundo o tercer nivel de atención y también participan como USMER las unidades de tercer nivel que por sus características contribuyen a ampliar el panorama de información epidemiológica, entre ellas las que cuenten con especialidad de neumología, infectología o pediatría.
SECTOR	Categórica	Identifica el tipo de institución del Sistema Nacional de Salud que brindó la atención.
ENTIDAD_UM	Categórica	Identifica la entidad donde se ubica la unidad medica que brindó la atención.
SEXO	Categórica	Identifica al sexo del paciente.
ENTIDAD_NAC	Categórica	Identifica la entidad de nacimiento del paciente.
ENTIDAD_RES	Categórica	Identifica la entidad de residencia del paciente.
MUNICIPIO_RES	Categórica	Identifica el municipio de residencia del paciente.

TIPO_PACIENTE	Categórica	Identifica el tipo de atención que recibió el paciente en la unidad. Se denomina como ambulatorio si regresó a su casa o se denomina como hospitalizado si fue ingresado a hospitalización.
FECHA_INGRESO	Fecha	Identifica la fecha de ingreso del paciente a la unidad de atención.
FECHA_SINTOMAS	Fecha	Identifica la fecha en que inició la sintomatología del paciente.
FECHA_DEF	Fecha	Identifica la fecha en que el paciente falleció.
INTUBADO	Categórica	Identifica si el paciente requirió de intubación.
NEUMONIA	Categórica	Identifica si al paciente se le diagnosticó con neumonía.
EDAD	Continua	Identifica la edad del paciente.
NACIONALIDAD	Categórica	Identifica si el paciente es mexicano o extranjero.
EMBARAZO	Categórica	Identifica si la paciente está embarazada.
HABLA_LENGUA_INDIG	Categórica	Identifica si el paciente habla lengua índigena.
CLASIFICACION_FINAL	Categórica	Identifica si el paciente es un caso de COVID-19 según el catálogo CLASIFICACION_FINAL.

Nota: En el "Apéndice", en la sección de "Variables Restantes" se encuentran el resto de variables con su respectivo nombre, tipificación y descripción.

Es preciso mencionar que se cuenta con un total de 38 variables de las cuales 32 son de tipo categóricas, una es de tipo continua, cuatro de tipo fecha y una de tipo texto. Todas estas variables cuentan con su respectivo catálogo, ya que, en la base de datos original los registros son números (en la mayoría de los casos), pero estos, no nos dicen nada si no sabemos qué significa.

2.2. Creación de la Tabla Analítica de Datos

En el caso del conjunto de datos de casos de COVID-19 que estoy analizando, en un principio no se presentan datos faltantes o registros en blancos, ni registros duplicados, por lo que procedí a realizar los respectivos cruces con los catálogos (obtenidos de la misma fuente que el conjunto de datos) para saber qué significan los valores numéricos contenidos en el conjunto de datos.

Analizando el conjunto de datos, me percaté que para realizar el cruce con el catálogo de los municipios es necesario generar otra variable de apoyo tanto en el catálogo como en mi conjunto de datos, esto porque en el catálogo de municipios, la clave se reinicia por cada estado, entonces lo que hice fue generar una nueva variable que contenga la información de ambas columnas, es decir, CLAVE_ENTIDAD - CLAVE_MUNICIPIO. Dicha variable recibe el nombre de CLAVE_ENTIDAD_MUNICIPIO.

2.3. Estadísticas Descriptivas

La estadística descriptiva es la rama de las matemáticas que recolecta, presenta y caracteriza un conjunto de datos con el fin de describir apropiadamente las diversas características de ese conjunto.

Dada la naturaleza de nuestras variables, sólo es posible obtener las estadísticas descriptivas de la variable "c_EDAD", ya que, es la que presenta valores continuos. Sus estadísticas son:

	c_EDAD
Conteo	2,403,499
Media	41
Desviación	16
Mínimo	0
25%	29
50%	40
75%	53
Máximo	120

Cuadro 1: Estadísticas descriptivas de la variable c_EDAD

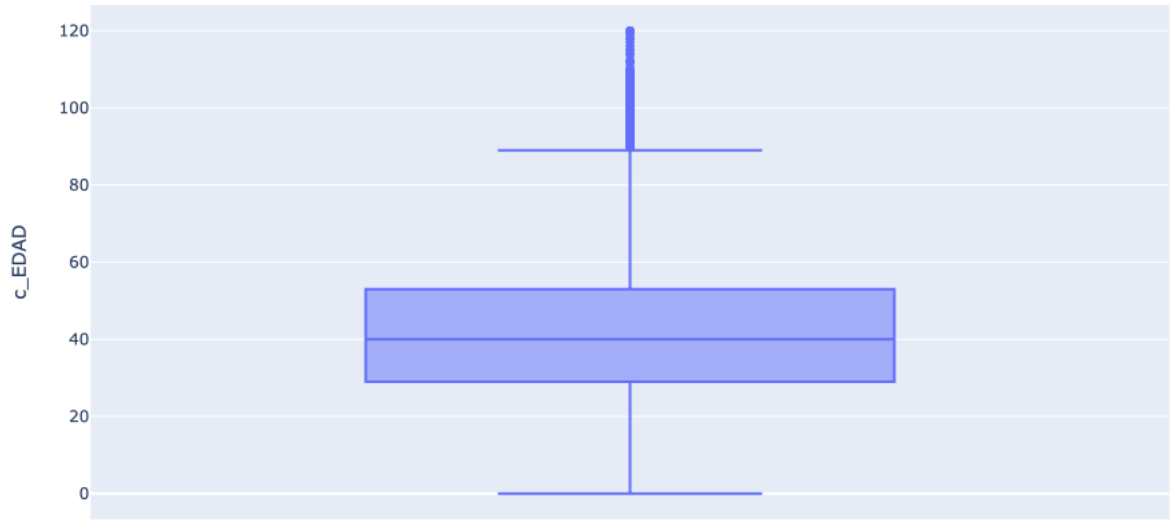


Figura 1: Box plot de la variable c_EDAD

Como se puede observar en la Gráfica 1 y Figura 1, la media de la edad es 41, esto quiere decir, que en promedio las personas que han contraído el virus tienen 41 años, más en general se aprecia que los contagios se concentran en un intervalo de edad que es de 29 a 53 años. Siendo algo "benéfico" para la población, ya que, las personas mayores a 65 años son del sector vulnerable, lo que significa que tienen más probabilidades de que se agravien sus síntomas. Analizando más a profundidad la Figura 1, nótese que presentamos unos cuantos datos atípicos en la edad, son las personas más longevas que han sido víctima del virus.

Por otro lado me resulta de interés analizar las edades respecto al sexo del paciente para intentar llegar a conclusiones más contundentes.

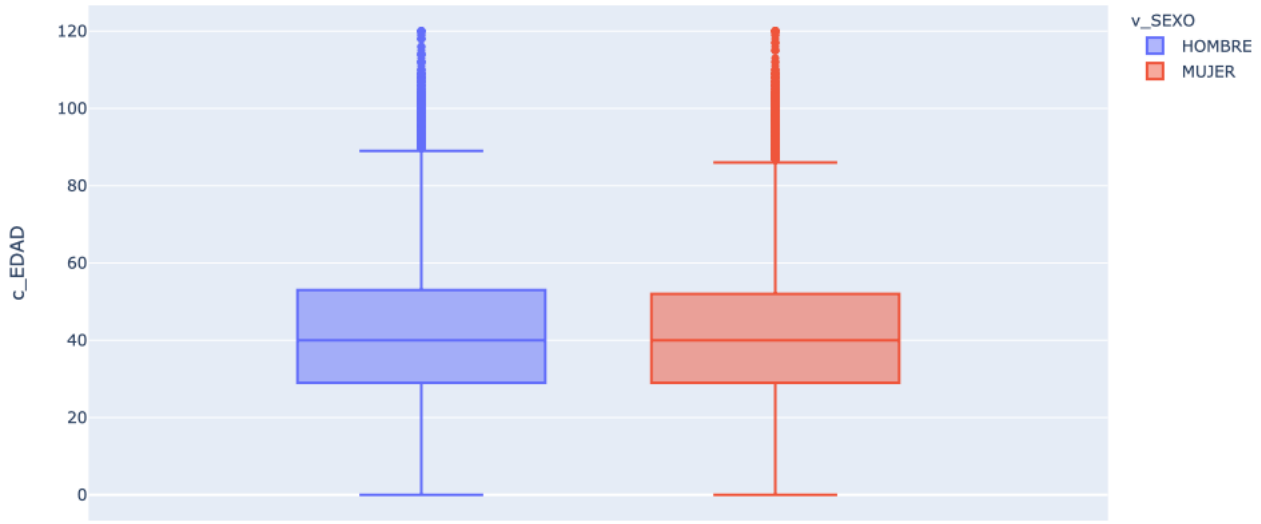
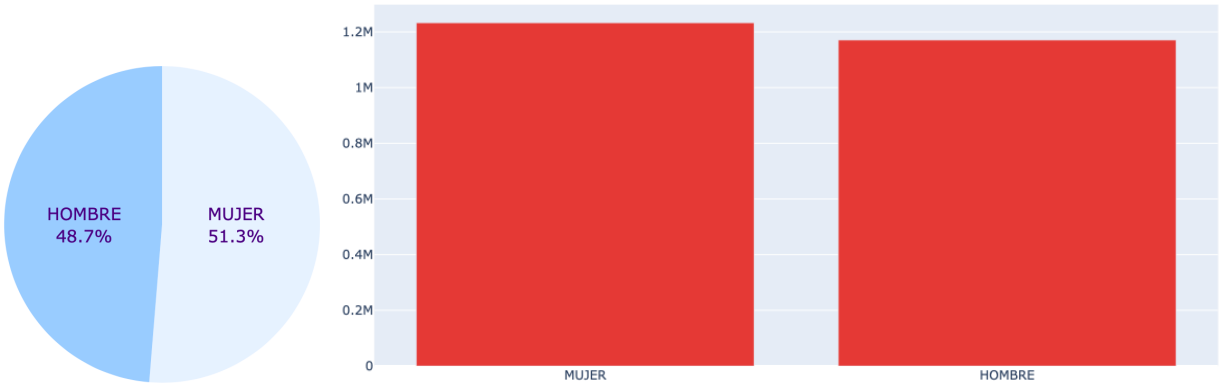


Figura 2: Box plot de la variable c_EDAD respecto a cada sexo

Tras analizar la Figura 2, nos damos cuenta que no hay mucha variación si tomamos en cuenta el sexo de cada paciente, las principales diferencias es que la edad promedio en la que más contagiados hay en ambos sexos es de 40 años y en el caso particular de las mujeres el intervalo de edad en el que se concentran las pacientes es de 29 a 52 años.

2.4. Visualización de datos

Como persona que está viviendo actualmente la pandemia, me han surgido dudas como la siguiente: ¿se han atendido más hombres que mujeres por COVID-19? para dar solución a esta pregunta analicé la variable v_SEXO.



(a) Porcentaje sexo (b) Conteo sexo

Figura 3: Variable v_SEXO

Como se puede observar en la Figura 3, la proporción de hombres y mujeres es prácticamente la misma, pero las mujeres presentan un número mayor de casos, siendo este de **1,233,012** contra **1,170,487** casos de hombres.

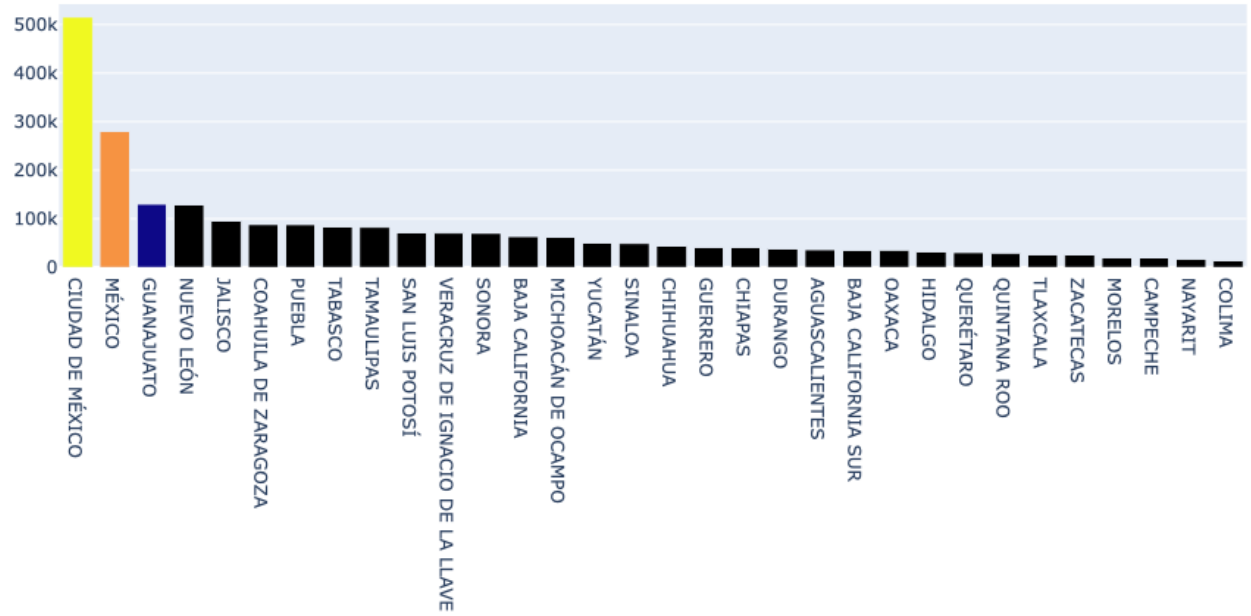


Figura 4: Número de pacientes por cada estado

En la Figura 4, se puede apreciar el número de pacientes que se han reportado desde que inició el registro de los pacientes (01/01/2020) hasta la última actualización de los datos (31/10/2020) siendo la Ciudad de México el estado de la república con más casos (esto es consecuencia de ser la ciudad más poblada del país), seguido del Estado de México y las ciudades más pobladas del país hasta el estado con menos registros de casos que es Colima.

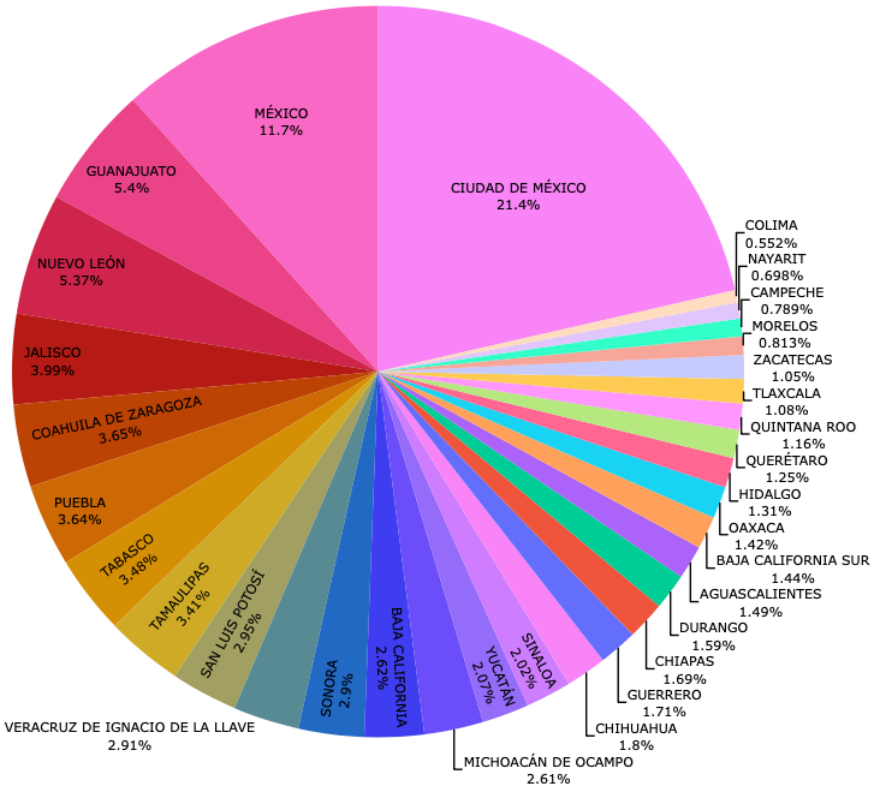


Figura 5: Porcentaje de pacientes por estado

Es preciso mencionar que los datos tienen coherencia como se muestra en la Figura 4 y en la Figura 5, ya que los que tienen mayor número de casos, son aquellos que tienen mayor número de habitantes y viceversa. Aunque, resulta bastante interesante estudiar el caso de Jalisco, ya que no se encuentra posicionado en el top 4, teniendo un mayor número de habitantes, alcanzando la cifra de **8,368,602** para 2020 según datos de la INEGI, cuando Nuevo León y Guanajuato cuenta con aproximadamente 6 millones de habitantes y tienen un mayor número de casos.

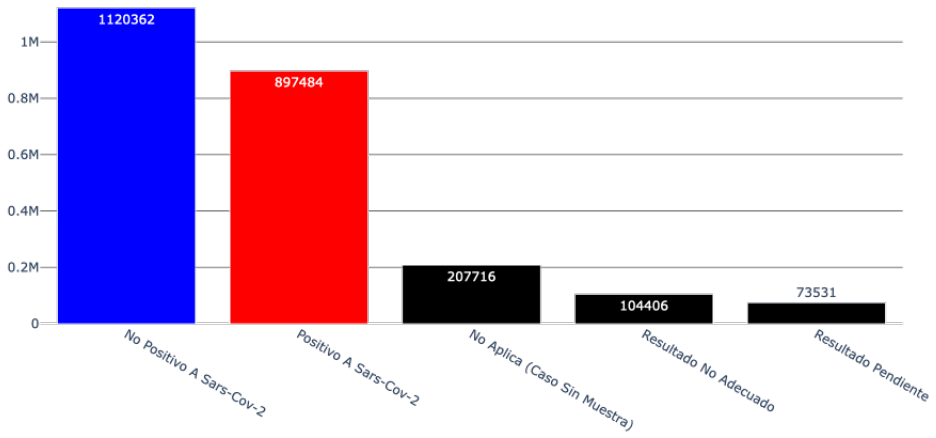


Figura 6: Número de resultados de laboratorio por cada tipo

Con la Figura 6 se pretende analizar el número de personas que se han realizado estudios para saber si tienen COVID-19. Por lo que se observa, la gran mayoría ha salido con un resultado favorable, es decir, no son positivos, pero un número considerable de personas, equivalente al **37 %** aproximadamente, han resultado positivos. Esta figura igual permite detectar que con los datos que se cuenta hasta el momento, les hace falta pasar por un proceso de limpieza y de verificación de calidad, ya que, la barra correspondiente a "No aplica" corresponde a lo que serían datos faltantes en la muestra.

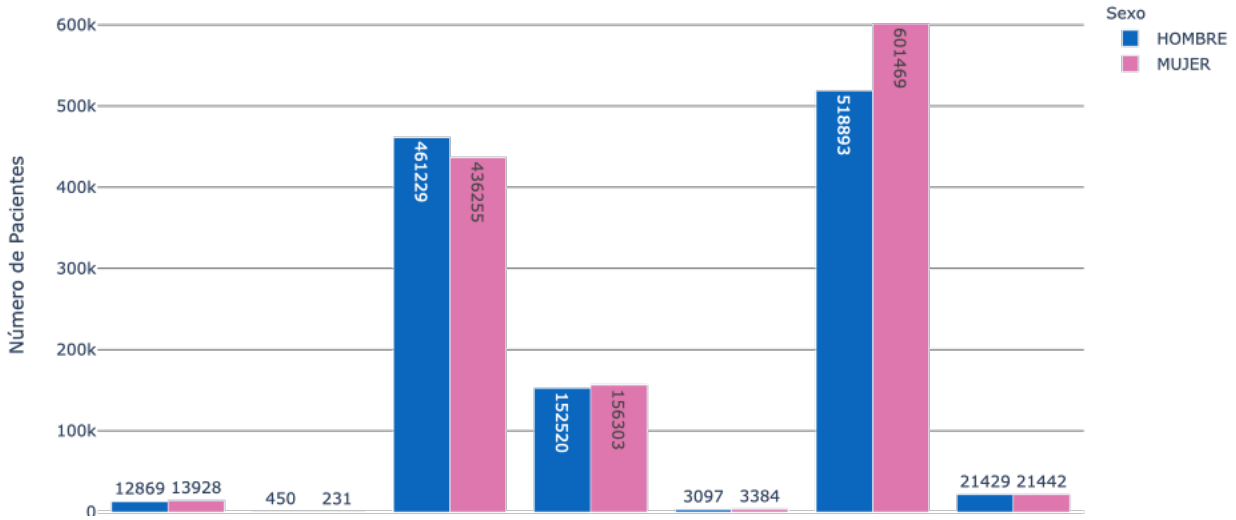


Figura 7: Total de clasificaciones finales por sexo

En la Figura 7, se pueden apreciar las clasificaciones que se tienen hasta el momento previo a la calidad de datos y el número de pacientes por cada sexo. Resultan interesantes

los siguientes datos, en los que se enlistan las clasificaciones finales en el orden que aparecen en el gráfico anterior (de izquierda a derecha):

Clasificación Final	**% Hombres**	**% Mujeres**
CASO DE COVID-19 CONFIRMADO POR ASOCIACIÓN CLÍNICA EPIDEMIOLÓGICA	1.1	1.13
CASO DE COVID-19 CONFIRMADO POR COMITÉ DE DICTAMINACIÓN	0.04	0.02
CASO DE SARS-COV-2 CONFIRMADO POR LABORATORIO	39.4	35.38
CASO SOSPECHOSO	13.03	12.68
INVÁLIDO POR LABORATORIO	0.26	0.27
NEGATIVO A SARS-COV-2 POR LABORATORIO	44.33	48.78
NO REALIZADO POR LABORATORIO	1.83	1.74

Cuadro 2: Porcentaje de clasificaciones finales por sexo

Como se puede apreciar en la Figura 7 y en el Cuadro 2, la mayoría de personas que asisten a atenciones médicas por COVID-19 resultan tener clasificaciones negativas y esto es debido a que muchas personas se sugestionan y sumado a lo anterior, los síntomas de esta enfermedad son muy parecidos a los de una gripa o resfriado común, por lo que las personas confunden la sintomatología por los de una enfermedad recurrente.

El tema de la sugestión es un delicado, ya que invade a muchas personas el pensamiento de poder tener COVID-19 y por temor realizan acciones que perjudican al resto, un ejemplo claro de esto es que al inicio de la cuarentena (al menos en el caso de México) la gente se alteró demasiado y empezó a hacer compras de pánico, agotando en muchos establecimientos el papel higiénico y haciendo que este aumentara drásticamente de precio por la escasez y la alta demanda. Por lo que el gobierno (no sólo el de México) han tomado iniciativas para evitar este tipo de acciones, ejemplo de estas fue el no promover el uso de cubrebocas al inicio de la pandemia, ya que si lo hacían se hubiera generado una alta demanda y escasez del producto, provocando que gente como los médicos y enfermeras que realmente los necesitaban en esos momentos, no tuvieran suficiente material.

Nota: En el "Apéndice", se encuentra una sección llamada "Visualización de Datos - Comparación" en donde se encuentran el resto de gráficas de las variables con su respectiva comparación (en caso de que hayan sufrido cambios).

3. Calidad de Datos

La calidad de datos le permite preparar y gestionar los datos, al tiempo que los pone a disposición de toda su organización. Los datos de alta calidad permiten a los sistemas estratégicos integrar todos los datos relacionados para proporcionar una visión completa de la organización y las interrelaciones dentro de la misma.

La calidad de los datos es una característica esencial que determina la confiabilidad de la toma de decisiones.

Se comenzó con el proceso de calidad de datos verificando que no existieran registros duplicados tanto de manera general como en la variable correspondiente al ID de identificación de paciente, se concluyó que **no** se encontraba ningún registro o paciente duplicado.

3.1. Variables Categóricas

Continuando con el proceso, se analizaron las variables dependiendo de su tipificado. Dado que la mayoría de variables eran categóricas, inicié analizando este tipo de variables. Se verificó que cada una de estas tomara valores dentro de su dominio únicamente y se encontró que muchas variables contenían categorías correspondientes a valores que deberían ser nulos, tales se muestran en la siguiente tabla:

Variables	**Categoría**
v_SECTOR, v_ENTIDAD_NAC, v_INTUBADO, v_NEUMONIA, v_MIGRANTE	no especificado
v_EMBARAZO, v_DIABETES, v_EPOC, v_ASMA, v_HIPERTENSION, v_OBESIDAD	se ignora
v_PAIS_ORIGEN, v_PAIS_NACIONALIDAD	se desconoce

Cuadro 3: Variables y su categoría correspondiente a nulos

Es importante mencionar que las variables mencionadas anteriormente sólo son un subconjunto de aquellas que contenían estos datos, ya que al final se descubrió que 22 de las 32 variables categóricas guardaban datos correspondientes a valores nulos en alguna de estas tres categorías, por lo que tras identificarlos se cambiaron a valores tipo NaN.

Por otro lado, la variable correspondiente a la entidad de residencia del paciente (v_ENTIDAD_RES) fueron modificadas sus categorías, esto pensando en la mejor manera de aprovechar la información que contiene, ya que como es una variable categórica, contiene los 32 estados que hay en el país. El problema radica en que cuando se vaya a transformar la información en datos numéricos , se generarían 31 variables nuevas, para no generar tantas variables, se agruparon en cinco regiones, tales que definió el gobierno mexicano al diseñar su estrategia de seguridad a nivel nacional. Las regiones son:

- **noroeste**: baja california, baja california sur, chihuahua, sinaloa, sonora.
- **noreste**: coahuila de zaragoza, durango, nuevo leon, san luis potosi, tamaulipas.
- **occidente**: aguascalientes, colima, guanajuato, jalisco, michoacan de ocampo, nayarit, queretaro, zacatecas.
- **centro**: ciudad de mexico, mexico, guerrero, hidalgo, morelos, puebla, tlaxcala.
- **sureste**: campeche, chiapas, oaxaca, quintana roo, tabasco, veracruz de ignacio de la llave, yucatan.

De esta manera, pasamos de tener que crear 31 variables a sólo **cuatro**, guardando la mayor cantidad de información posible.

3.2. Variable Continua

Se analizó los valores que tomaba la variable correspondiente a la edad y como se muestra anteriormente, no presenta valores nulo, pero sí contiene outliers, tales que en un proceso más adelante serán modificados.

3.3. Variable Tipo Texto

En este caso no se realizó ninguna análisis extra a la variable correspondiente al ID de indentificación de los pacientes, ya que todos su datos ya venían homologados y sin duplicados.

3.4. Variables Tipo Fecha

Para poder hacer uso de la información contenida en las variables correspondientes a fechas (de última actualización, de ingreso del paciente, en que se presentaron síntomas y de defunción) se tuvo que realizar una transformación en el tipo de dato, ya que, en un principio estas variables contenían puros datos tipo string los cuales no permitían el manejo adecuado de la información, por lo que se transformaron a valores tipo "dateTime".

La transformación realizada en cada una de las cuatro variables, provocó que la correspondiente a la fecha de defunción (d_FECHA_DEF) tuviera una enorme cantidad de valores tipo NaT, es decir, valores tipo nulos. Esto porque la mayoría de pacientes que son atendidos (correspondientes al **95%** aproximadamente del total de pacientes) no fallecen.

3.5. Completitud

El término "Completitud" se refiere a cuando todos los campos y registros están dentro del conjunto de datos, no existen espacios en blanco.

Tras haber verificado la calidad de datos en todos los tipos de variables, es importante verificar como ha cambiado su completitud, ya que, aquellas que **no** pasen el umbral de completitud del **80%** serán eliminadas, pues ya tendrían un número alto de valores nulos y si se intentan imputar (término que se explicará más adelante) podría cambiar mucho su distribución.

Analizando el Cuadro 4, es notorio que no todas las variables tienen una completitud del 100% como se creía en un principio, en realidad sólo 10 cuentan con todos sus datos. Nótese que hay dos variables que no alcanzan el umbral establecido para la completitud, pero solo se eliminó "v_MIGRANTE", dado que la variable de tipo fecha contiene información que es interesante con los pocos datos que posee.

Variables	Completitud
v_MIGRANTE	0.397 %
v_HABLA_LENGUA_INDIG	96.287 %
d_FECHA_DEF	5.44 %
v_OTRO_CASO	87.941 %
v_NEUMONIA	99.238 %
v_ENTIDAD_NAC	99.525 %
v_EMBARAZO	99.617 %
v_DIABETES	99.663 %
v_TABAQUISMO	99.673 %
v_HIPERTENSION	99.684 %

Cuadro 4: Completitud de las variables más afectadas

3.6. Generación de la Variable Objetivo

La variable objetivo es la dependiente, es decir, es la variable que se intentará predecir o modelar en función del resto de variables independientes.

Debido a que no se cuenta con una variable que proporcione de manera explícita la información donde indique si el paciente sobrevivió o falló, se construyó esta variable. Utilizando la variable "FECHA_DEF" la cual si tiene la fecha "9999-99-99" indica que el paciente sobrevivió y si cuenta con una fecha distinta indica que el paciente falló. Con la información anterior se creó la variable "**tgt_mortalidad**", la cual toma los siguientes valores:

- **Sobrevivió**: 0
- **Falló**: 1

De esta forma se tiene explícitamente la información del objetivo del análisis.

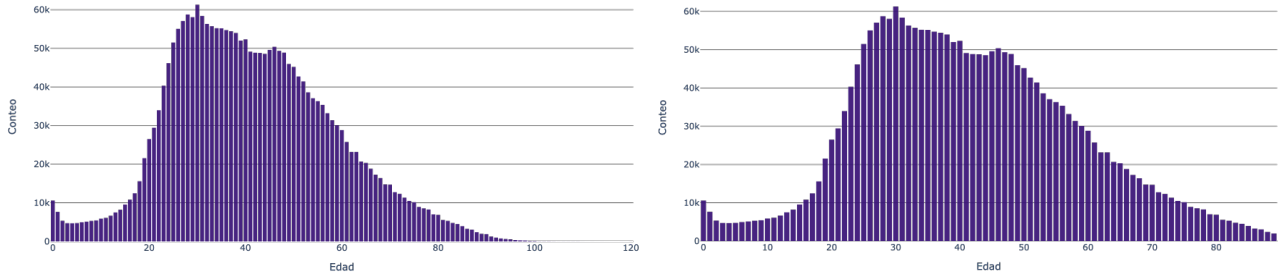
4. Datos Anómalos

Los datos anómalos u outliers (en inglés) son aquellos que tienen características diferentes de la multitud, por lo que resulta importante tratarlos, ya que si no, al modelar con los datos podrían provocar resultados erróneos.

Como en el proceso de calidad de datos ya se había revisado que las variables categóricas y las de tipo fecha tuvieran datos dentro de la naturaleza de su domino, la única que resta por detectar outliers es en la continua (c_EDAD), por lo que se le aplicó una función que mediante tres métodos de detección de datos atípicos los cuales son:

- Inter Quantile Range (IQR)
- Z-Score
- Percentiles

La función encuentra los índices de los datos atípicos que al menos hayan sido detectados por dos de los métodos anteriores. En el caso particular de la variable correspondiente a



(a) Distribución con outliers (b) Distribución sin outliers

Figura 8: Comparación de la variable c_EDAD

la edad se detectaron **7,684**, lo que representa al **0.32 %** del total de los registros, por lo que se procedió a eliminar estas observaciones anómalas.

La Figura 8, muestra el cambio que sufrió la variable edad al detectar y eliminar sus outliers, en un principio, se contaba con una edad máxima de **120**, tras el tratamiento se cuenta con una edad máxima de **89**.

5. Análisis Descriptivo Post-Procesamiento de Datos

5.1. Visualización de datos

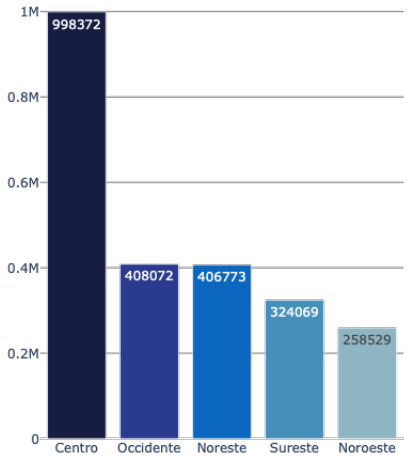


Figura 9: Número de pacientes por regiones

La Figura 9 es el contraste de la Figura 4, porque muestra las nuevas categorías que se crearon para la variable v_ENTIDAD_RES, en la que como se puede apreciar, no se perdió demasiada información, dado que si bien no es tan específica como en un principio, nos sigue permitiendo dimensionar la cantidad de pacientes que hay en los estados. Analizando la gráfica se llega a las siguientes conclusiones:

- El **42 %** de los pacientes atendidos por COVID-19 provienen de la región centro.
- El número de pacientes de la región occidente y noreste es equivalente al **17 %** cada una.

- La región sureste representa el **13 %** de los pacientes totales.
- La región noroeste representa el **11 %**.

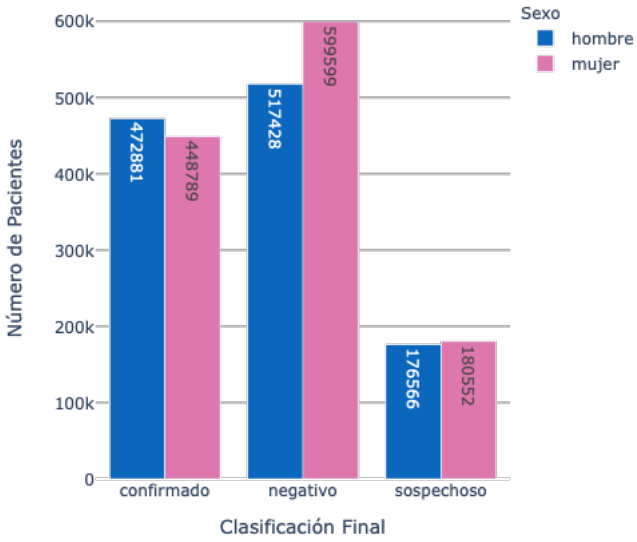


Figura 10: Clasificaciones finales por sexo

Mediante la Figura 10 se puede contrastar la Figura 7, ya que muestra los cambios que se realizaron a la variable objetivo para reducir el número de categorías que eran redundantes. Analizando la gráfica se tiene que el **36 %** de las pacientes ha sido clasificada como caso confirmado, el **49 %** como negativo y el **15 %** como caso sospechoso. En cuanto a los pacientes el **40 %** son casos confirmados, **44 %** casos negativos y **16 %** casos sospechosos.

A partir de lo que se lleva del tratamiento de los datos, es posible generar gráficas utilizando las variables de tiempo, pues en procesos anteriores, fueron tipificadas de manera correcta para realizar estos análisis.

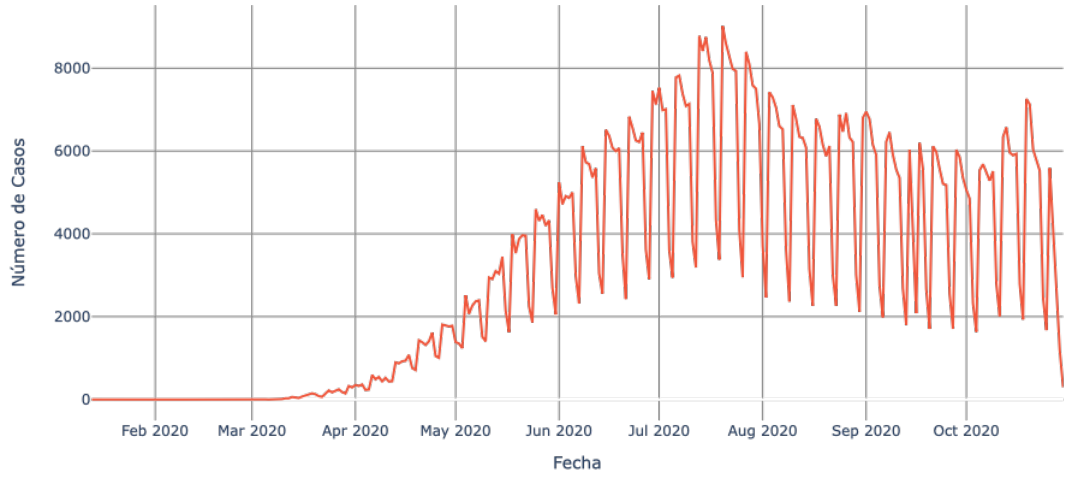


Figura 11: Número de casos confirmados diario

En la Figura 11, se puede visualizar cómo se va comportando el número de casos confirmados durante el paso de los días. Es importante destacar que los datos comienzan desde el 1º de enero hasta el 31 de octubre del 2020. Se logran llegar a las siguientes conclusiones:

- El 20 de julio es el pico más alto con **9,051** casos confirmados.

- El 31 de octubre se alcanza el pico más bajo con **296** casos confirmados.
- En promedio al día se tienen **3,168** nuevos casos confirmados.

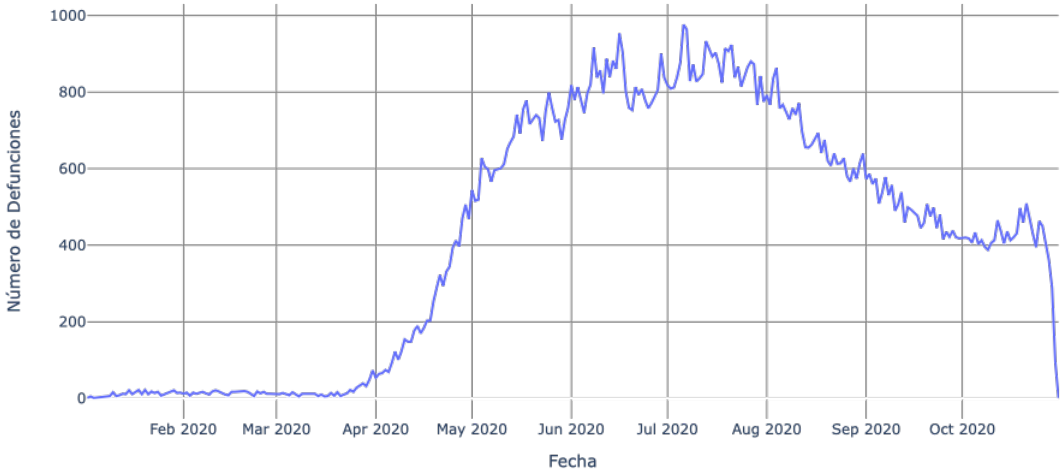


Figura 12: Número de fallecimientos diarios

Por otra parte, en la Figura 11 se puede apreciar el comportamiento del número de fallecimientos de pacientes que fueron atendidos por COVID-19 con el paso de los días y se pueden resaltar los siguientes aspectos:

- El 6 de julio se alcanzó el pico más alto con **985** defunciones.
- El 31 de octubre se alcanza el pico más bajo con **1** defunción.
- En promedio al día se tienen **431** defunciones.

6. Imputación de Valores Faltantes

La imputación de valores faltantes consiste en métodos que permiten aumentar la completitud de las variables, logrando que alcancen el 100 %.

Antes de proceder a imputar valores, es necesario conocer los datos para detectar el método a utilizar, en el caso presente se trabaja con una serie de tiempo pero dado que se imputaron valores a variables categóricas, no es necesario verificar si presenta tendencia o no. Partiendo de lo anterior se decidió imputar los valores haciendo uso de la moda, es decir, tomando el valor o categoría que más se repite.

A partir de este punto, se dividió el conjunto de datos en dos subconjuntos: prueba y entrenamiento. El primero contiene el **80 %** y el segundo el **20 %** restante. Es importante aclarar que los registros de cada subconjunto son seleccionados por orden, es decir, para el caso del conjunto de entrenamiento se seleccionan los registros considerando el orden de la fecha de ingreso del paciente y hasta el registro en el que se acumule el 80 % de la tabla original. Se realiza de esta manera, ya que la moda que se obtiene es sobre el conjunto del entrenamiento y es valor que se le imputa a ambos conjuntos.

Resultan importantes estos subconjuntos porque cuando se vaya a modelar con los datos, se entrenará al modelo con el conjunto de entrenamiento y se evaluará el desempeño de

predicción del modelo con el conjunto de prueba.

Variables	Moda
v_SECTOR	ssa
v_ENTIDAD_NAC	ciudad de mexico
v_INTUBADO	no aplica
v_NEUMONIA	no
v_EMBARAZO	no
v_HABLA_LENGUA_INDIG	no
v_DIABETES	no
v_EPOC	no
v_ASMA	no
v_INMUSUPR	no
v_HIPERTENSION	no

Cuadro 5: Variables y su moda

Nota: El resto de variables y su respectiva moda se encenta en el ”Apéndice” en la sección de ”Imputación de Valores Faltantes - Restantes”.

El Cuadro 7 muestra un subconjunto de las variables que fueron completadas y el valor que les fue imputado.

7. Ingeniería de Variables

La ingeniería de variables consiste en generar nuevas variables con las que ya se cuenta en un principio con el propósito de conservar la mayor cantidad de información posible o extraer nueva información.

7.1. Variables Categóricas

Este procesos consistió en transformar todas las variables categóricas, tales que contienen valores tipo ”string”, a variables que sólo contengan valores de tipo **numérico**. Para realizar el objetivo anterior, se crearon **variables dummines** (variables que sólo contienen 1 si ocurre el evento que describen ó 0 si no ocurre el evento que describen) de casi todas las variables categóricas menos de: v_ENTIDAD_UM, v_ENTIDAD_NAC, v_RESULTADO_LAB, v_PAIS_ORIGEN, v_MUNICIPIO_RES, v_PAIS_NACIONALIDAD. Debido a que estas seis tendrán un proceso diferente que será explicado más adelante.

En resumen, se pasó de tener 37 variables a terminar con un total de **53**, de las cuales 10 no fueron modificadas (las seis mencionadas anteriormente y las cuatro tipo fecha), las 26 restantes fueron reemplazadas por 41 variables nuevas que contienen la información de las variables en forma numérica.

8. Reducción de Dimensiones

La reducción de dimensiones tiene como objetivo que mediante la proyección de los datos a un subespacio de menor dimensión se plantea captar la "**esencia**" de los datos. En otras palabras consiste en reducir el número de variables independientes conservando la mayor cantidad de información de los datos.

Tras haber creado variables dummies, el conjunto de datos aumentó significativamente en cuanto al número de variables, por lo que se aplicaron diferentes métodos que ayudaran a medir su desempeño en conjunto e individual.

Recuérdese que para este punto sólo hemos eliminado la variable v_MIGRANTE porque tras aplicarle el método de **relación de valor perdido**, no alcanzó el umbral de completitud.

Se eliminaron las variables v_ENTIDAD_UM, v_ENTIDAD_NAC, v_PAIS_ORIGEN, v_PAIS_NACIONALIDAD y v_MUNICIPIO_RES debido a que no aportan mucha información que sea útil para la variable objetivo, y a su vez, la información relevante de estas variables se encuentra almacenada en las variables v_ENTIDAD_RES y v_NACIONALIDAD.

Por otro lado, no aporta mayor información la variable v_RESULTADO_LAB por lo que fue eliminada, ya que la variable objetivo contiene prácticamente la misma información y esta es la importante. A su vez, la variable de texto que contiene el ID que identifica a cada paciente (t_ID_REGISTRO) fue eliminada porque no aporta información que sea de utilidad para la variable objetivo. En cuanto a las variables de tipo fecha, hay una que no aporta información a la variable dependiente y esta es d_FECHA_ACTUALIZACION que indica el último día que se actualizaron los datos (en este caso es en el día en el que se descargaron) y es constante para todos los registros.

Ahora se aplicará un método llamado **filtro de alta correlación**, el cual consiste en eliminar las variables que resulten redundantes, pues contienen información que otras variables guardan.

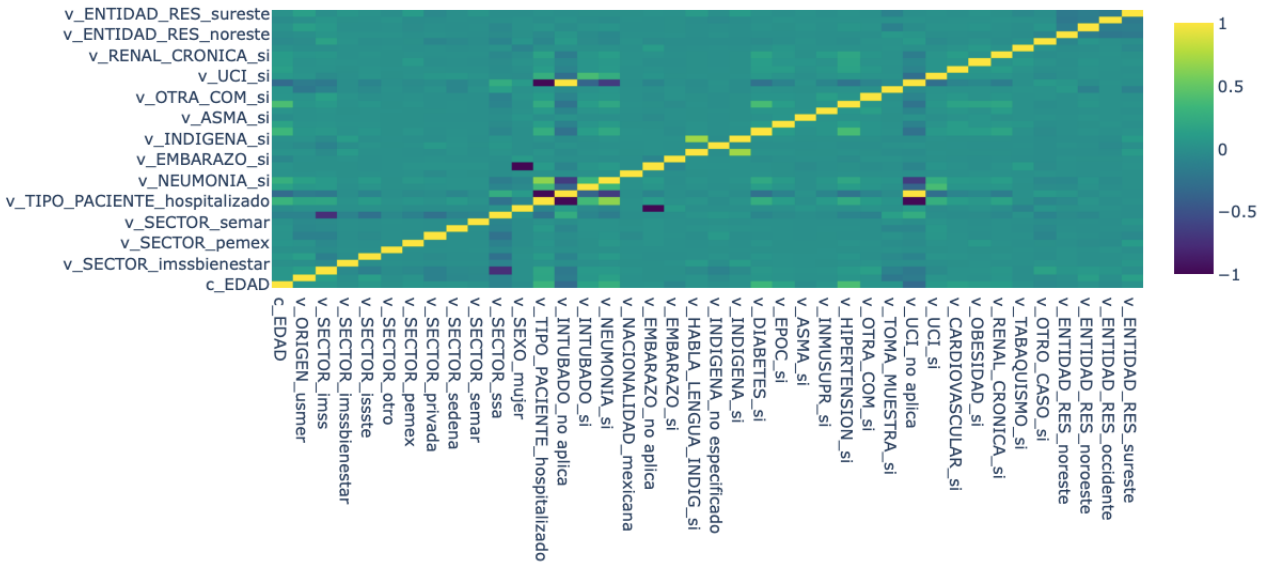


Figura 13: Correlaciones entre variable independientes

Analizando la Figura 13, se eliminaron **seis** variables que tienen su correlación mayor a 0.6 o menos a -0.6 con alguna otra variable.

- Se eliminó la variable v_SECTOR_ssa pues presentaba una alta correlación negativa con la variable v_SECTOR_imss, esto quiere decir, al aumentar el número de pacientes del imss, disminuyen los pacientes de la Secretaria de Salubridad y Asistencia de manera proporcional.
- Se eliminó la variable v_EMBARAZO_no aplica, dado que tiene una alta correlación negativa con la variable v_SEXO_mujer.
- Se eliminaron las variables v_INTUBADO_no aplica y v_UCI_no aplica por tener una correlación alta negativa con la variable v_TIPO_PACIENTE_hospitalizado. Por otro lado, también presenta una alta correlación con esta última la variable v_NEUMONIA_si, es decir, si aumentan los pacientes de tipo hospitalizados igual aumentará el número de pacientes que tiene neumonía.
- Se eliminó la variable v_HABLA_LENGUA_INDIG_si dado que tiene una alta correlación con la variable v_INDIGENA_si.

Con el método anterior, se medió el desempeño de las variables independientes por parejas, así que, se pretende analizar ahora su desempeño de todas en conjunto, por lo que se procedió a medir su nivel de **multicolinealidad**, para esto se usa una métrica llamada Factor de Inflación de la Varianza (por sus siglas en inglés VIF).

Variable	VIF
v_NACIONALIDAD_mexicana	22.211
v_TOMA_MUESTRA_si	10.821

Cuadro 6: Variables y su respectivo VIF

Para este método se consideró un umbral para el VIF de 10, es decir, aquellas variables que obtuvieran un VIF mayor que el umbral establecido serían eliminadas, ya que la información que almacenan se puede obtener mediante la combinación lineal del resto de variables.

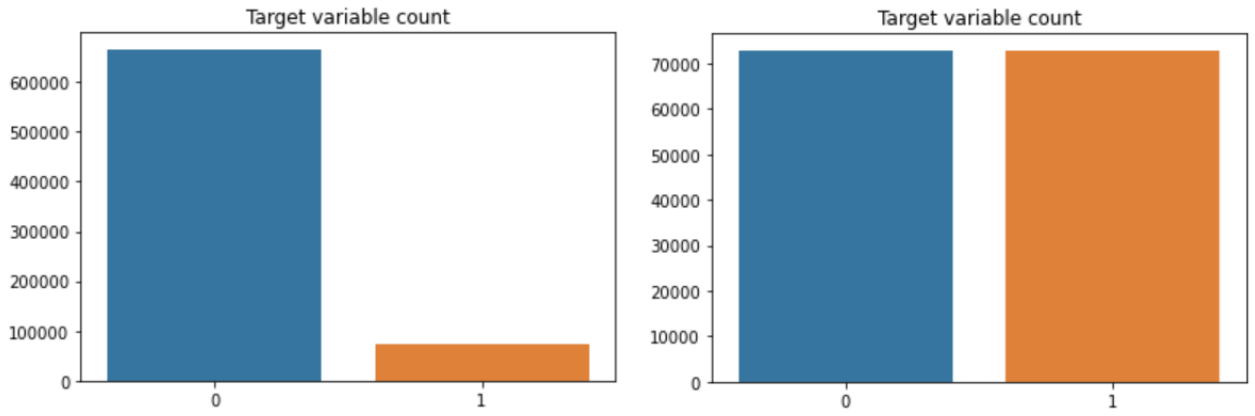
En resumen de esta sección, se logró reducir de 51 variables a **35**, es decir, se eliminó un total de 17 (considerando la que fue eliminada tras revisar la completitud).

9. Modelación Supervisada

El nombre de aprendizaje ”supervisado” se origina de la idea de que entrenar este tipo de algoritmos es como tener un profesor supervisando todo el proceso. El término ”supervisado” se refiere a un conjunto de ejemplos de entrenamiento (datos de entrada) donde ya se conocen los datos de salida deseados (etiquetas). En el caso del problema planteado en este proyecto, es un problema de tipo clasificación, pues la variable objetivo es categórica.

Nótese que la variable objetivo está desbalanceada, por lo que se probaron varios modelos con dos conjuntos de datos diferentes, uno con el desbalance (conjunto original) y para el

otro se utilizó una técnica de balanceo llamada ”undersample”, la cual consiste en eliminar observaciones de las que más dominio tienen en la variable objetivo hasta equilibrarla.



(a) Sin balanceo (b) Undersample

Figura 14: Conjuntos de entrenamiento

Analizando la Figura 14, se tiene que la muestra sin balanceo tiene 90% de registros pertenecientes a casos de pacientes que sobrevivieron y 10% de casos de pacientes que fallaron, mientras que la muestra con undersample se alcanza un balance del 50%. Es preciso mencionar que se seleccionó esta técnica de balanceo porque se cuenta con un conjunto de datos que contiene muchos registros y tras eliminar observaciones se sigue conservando un número alto para poder entrenar los modelos.

No está de más destacar que la mayoría de los modelos que se mostrarán en las secciones siguientes son el resultado de una hiperparametrización, es decir, se probaron esos mismos modelos con diferentes parámetros y se conservaron aquellos que generaban el mejor rendimiento.

9.1. Sin Balanceo

A continuación se muestran modelos de clasificación que se entrenaron para intentar predecir la variable objetivo utilizando el conjunto de datos original, es decir, sin haber balanceado la muestra. Nótese que los ”scores” mostrados son tras probar el modelo con el conjunto de prueba, lo que significa que, se midió el rendimiento de los modelos con una muestra desconocida para el modelo.

Modelo	AUC	ACC	F1
Árbol de Decisión	0.4994	0.9009	0
XGBoost	0.4963	0.9009	0
Regresión Logística	0.4985	0.9009	0
Naive Bayes	0.4982	0.8861	0.0299
Red Neuronal	0.4993	0.9009	0

Cuadro 7: Modelos sin balanceo y sus scores

Observando el Cuadro 7, es notable que se tienen ACC’s altos, esto debido a que la muestra está desbalanceada, por eso es que el F1 score es prácticamente cero en todos los casos.

9.2. Undersample

A continuación se muestra el desempeño de los mismos modelos entrenados anteriormente, pero ahora con el conjunto de datos balanceado. De igual manera los scores obtenidos son con el conjunto de prueba.

Modelo	AUC	ACC	F1
Árbol de Decisión	0.4994	0.6543	0.149
XGBoost	0.5012	0.5194	0.1644
Regresión Logística	0.4994	0.4952	0.1654
Naive Bayes	0.5011	0.2019	0.1786
Red Neuronal	0.4951	0.4341	0.1677

Cuadro 8: Modelos con balanceo y sus scores

Tras analizar los scores mostrados en el Cuadro 8, resulta destacable que el performance se logró mejorar, pues el F1 score tuvo un aumento notable en todos los modelos, pero no se alcanza un rendimiento satisfactorio. Tras observar que con la muestra balanceada los modelos lograban predecir ”mejor”, se hicieron dos pruebas más y las métricas obtenidas son:

Modelo	AUC	ACC	F1
Ensamble Votante	0.4999	0.4053	0.171
Naive Bayes - K Best	0.5001	0.1584	0.1791

Cuadro 9: Modelos extras con balanceo y sus scores

En el caso del modelo llamado ”Ensamble Votante” se utilizaron los tres modelos que anteriormente dieron los mejores rendimientos, los seleccionados fueron: Naive Bayes, Red Neuronal y Regresión Logística, los cuales, mediante un ensamble trabajaron en conjunto para predecir la variable objetivo, pero no se obtuvo un rendimiento destacable. Por otro lado, se escogió el modelo con mejor rendimiento, el cual fue Navie Bayes y se entrenó utilizando un método de selección de variables, con el cual, se logró obtener el mejor F1 score de todos los modelos presentados y en lugar de entrenar el modelo utilizando 35 variables, se entrenó con las 10 mejores variables, lo que quiere decir que no solamente fue el mejor modelo en cuanto a performance, sino que también fue el modelo menos complejo.

10. Modelación No Supervisada

En contraste con la sección anterior, el aprendizaje no supervisado se utiliza cuando no se cuenta con una variable objetivo, es decir, no se conocen las salidas deseadas. Por lo que estos modelos toman mucha relevancia para lograr extraer información valiosa de los conjuntos de datos de covariables obteniendo como salida grupos (clusters) o centroides representativos de cada grupo. En el presente proyecto, dado que se cuenta con un gran número de covariables que guardan información de cada uno de los pacientes, se realizó modelaje no supervisado para obtención de clusters.

10.1. Visualización

Dado que se cuenta con **35 variables** resulta imposible visualizarlas sin un procesamiento previo, por lo que se aplicó un escalamiento de variables y el método llamado "Análisis de Componentes Principales" (o por sus siglas en inglés **PCA**), logrando reducir a tres dimensiones.

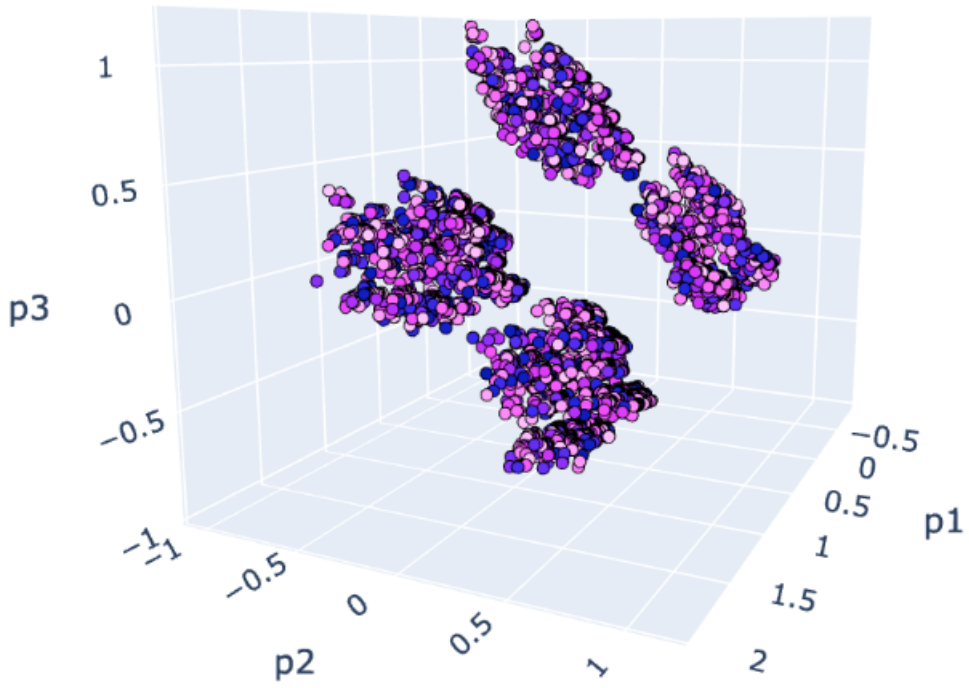


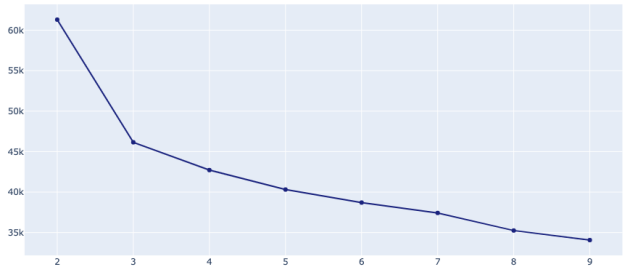
Figura 15: PCA en 3D

En la Figura 15, se observa la representación en tres dimensiones del conjuntos de covariables, se logran apreciar aparentemente cuatro posibles grupos. Cabe mencionar que esta visualización almacena el **32 %** de la explicabilidad del conjunto de datos.

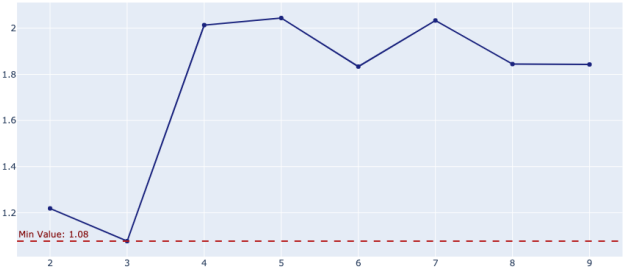
10.2. Número Óptimo de Clusters

Dado que no es muy confiable el número de clusters observados en la Figura 15, se utilizaron tres diferentes métodos que permitieran la detección de este número que represente de la mejor manera a las covariables. Es importante mencionar que se aplicaron sobre el conjunto ya escalado.

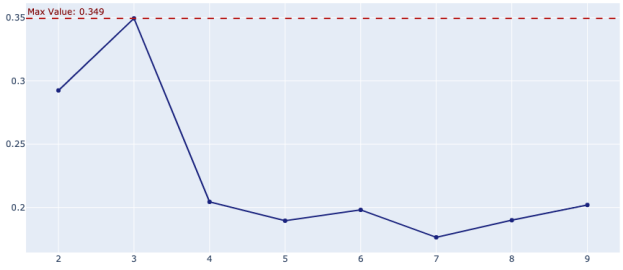
La manera correcta de interpretar la Figura 16 es la siguiente: para el caso del método del codo (a) se selecciona el número de clusters en el que se genera el menor ángulo en la gráfica, el de Davies-Bouldin (b) nos indica el número óptimo de segmentos en el mínimo y finalmente, el Silhouette (c) nos indica el número óptimo de grupos en el máximo. Tras analizar dicha figura, se concluyó que el número adecuado de clusters son 3.



(a) Gráfica del Codo



(b) Gráfica de Davies-Bouldin



(c) Gráfica de Silhouette

Figura 16: Número de clusters óptimo

10.3. Generación de Clusters

Se probaron cuatro diferentes algoritmos de aprendizaje no supervisado, los cuales fueron:

- **K - medioides**: Parte de la premisa que el mejor representante de cada grupo es la mediana y calcula los puntos más cercanos a ella mediante la distancia euclidiana.
- Clustering espacial basado en densidad de aplicaciones con ruido (**DSCAN**): Es un método de clustering de densidad, el cual, mediante las distancias separa los grupos densos del ruido disperso.
- **Mezclas Gaussianas**: Otro método de clustering de densidad, el cual, mediante la densidad probabilística, es decir, mide la probabilidad de que una muestra haya sido generada por una mezcla gaussiana.
- **Clustering Jerarquico Aconglomerativo**: Parte de la premisa que cada observación es un cluster y mediante el cómputo iterativo de las similitudes entre observaciones se van produciendo grupos.

Con K - medioides no se obtuvieron resultados decentes, ya que, sólo logró generar dos clusters y uno abarcando el 99% de la muestra aproximadamente. DBSCAN generó 5 grupos pero uno de ellos era ruido y predominaba fuertemente respecto al resto. Mediante el Clustering Jerarquico Aconglomerativo se obtuvieron tres grupos y la visualización en PCA parecía bastante buena, pero al intentar realizar el perfilamiento, no se producía una buena interpretación de los clusters. Finalmente el modelo de **Mezclas Gaussianas** fue el seleccionado, dado que brinda una buena visualización sobre PCA en 3 dimensiones y también permite generar un perfilamiento suficientemente entendible.

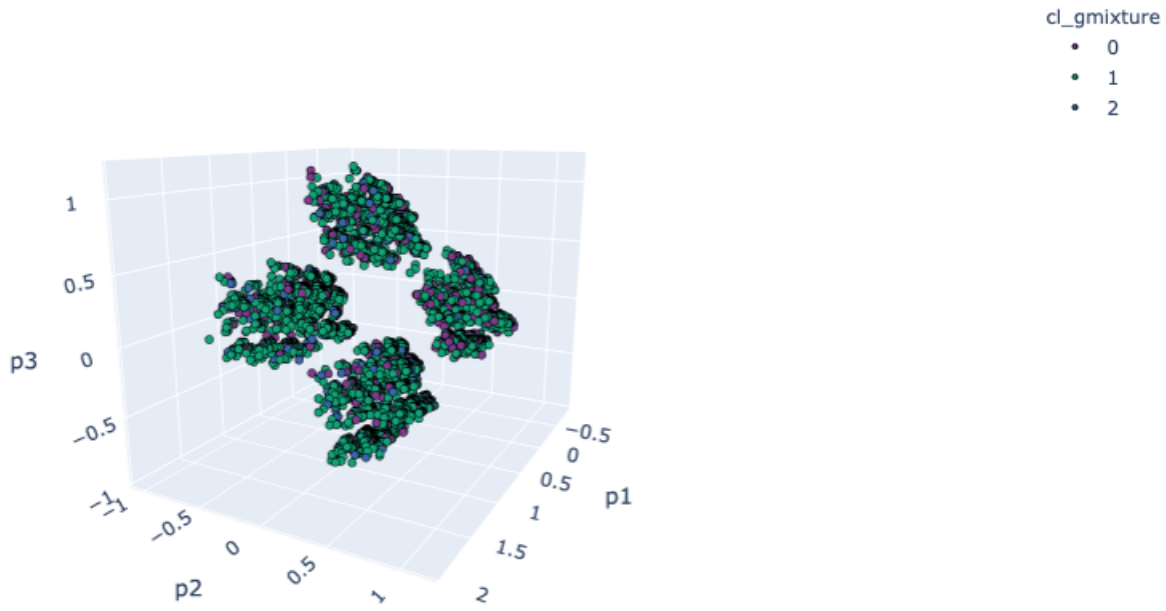


Figura 17: Mezclas Gaussianas sobre PCA en 3D

10.4. Perfilamiento

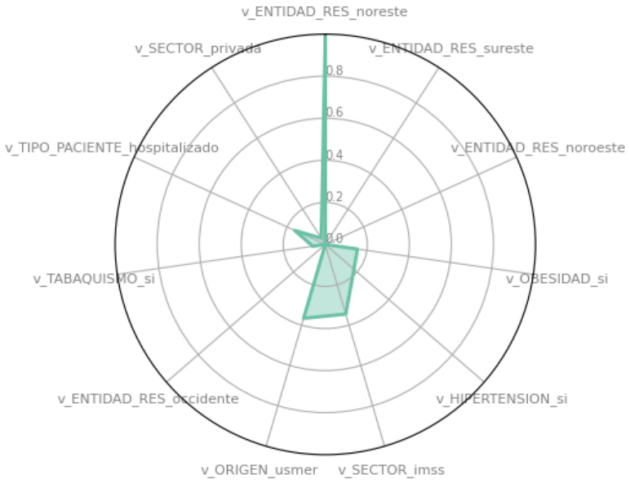
Previo a realizar el perfilamiento, se optó por ajustar un modelo de aprendizaje supervisado, "árbol de decisión", utilizando como variable objetivo los clusters obtenidos con Mezclas Gaussianas, se dejó que el método sobreajustara, ya que, para el único propósito que se utilizó fue para obtener la importancia de las covariables. Obteniendo los siguientes resultados:

Variable	Importancia
v_ENTIDAD_RES_noreste	0.6035
v_ENTIDAD_RES_sureste	0.04
v_ENTIDAD_RES_noroeste	0.0359
v_OBESIDAD_si	0.0342
v_HIPERTENSION_si	0.0327
v_SECTOR_imss	0.0326
v_ORIGEN_usmer	0.031
v_ENTIDAD_RES_occidente	0.0309
v_TABAQUISMO_si	0.0307
v_TIPO_PACIENTE_hospitalizado	0.0293
v_SECTOR_privada	0.0218

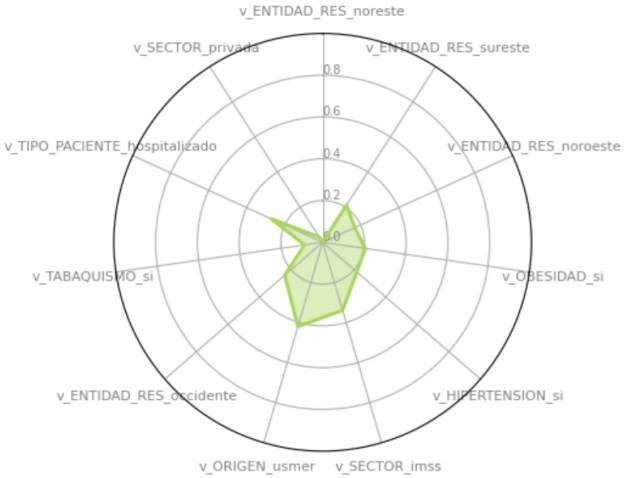
Cuadro 10: Importancia de variables según un árbol de decisión

Nota: En el "Apéndice", en la sección de "Importancia de Variables" se encuentran el resto de variables con su respectivo nombre e importancia.

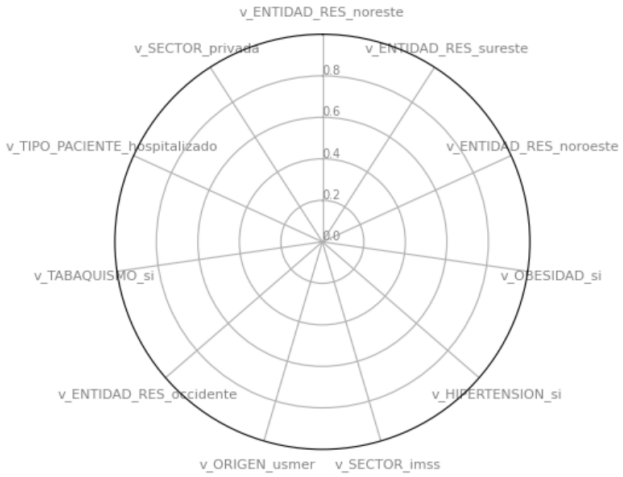
Considerando el tratamiento anterior se tienen los siguientes perfiles que se muestran en la Figura 18 en la que se están considerando las 11 mejores variables.



(a) Cluster 0



(b) Cluster 1



(c) Cluster 2

Figura 18: Gráficas de radar

- **Los Intermedios** (Cluster 0): Cerca del 40 % son trabajadores del sector privado, su estado de salud no es el mejor pero tampoco es muy malo, muy pocas hospitalizaciones.
- **Los Vulnerables** (Cluster 1): Aquellos pacientes que en su mayoría son trabajadores del sector privado que tienen malos hábitos de ejercicio y alimenticios, lo cual, los ha llevado a padecer de enfermedades como obesidad e hipertensión. Sumado a lo anterior también son fumadores, por lo que se han enfermado en gravedad y han sido hospitalizados.
- **Los Saludables** (Cluster 2): Son los pacientes que se encuentran un estado óptimo de salud, no tienen malos hábitos como el fumar y no fueron hospitalizados.

10.5. Prueba de Kruskal - Wallis

El test de Kruskal-Wallis contrasta si las diferentes muestras están equi distribuidas y que por lo tanto pertenecen a una misma distribución (población), por lo que las hipótesis a compara son:

- H_0: Todas las muestras provienen de la misma población (distribución).
- H_a: Al menos una muestra proviene de una población con una distribución distinta.

Variable	P - Value
v_ENTIDAD_RES_noreste	0
v_ENTIDAD_RES_sureste	0
v_ENTIDAD_RES_noroeste	0
v_OBESIDAD_si	0
v_HIPERTENSION_si	0
v_SECTOR_imss	0
v_ORIGEN_usmer	0
v_ENTIDAD_RES_occidente	0
v_TABAQUISMO_si	0
v_TIPO_PACIENTE_hospitalizado	0
v_SECTOR_privada	0

Cuadro 11: P - Value tras aplicar el test

Como se observa en el Cuadro 11, el P-Value de todas las variables es menor que 0.05, es decir, no se cuenta con información suficiente de que las muestras provienen de la misma población. Por lo tanto, los clusters logran segmentar de manera correcta.

10.6. Prueba de estabilidad

Dada la naturaleza del conjunto de datos que se está trabajando, se cuenta con la variable del tiempo asociada por lo que se generaron dos periodos de cinco meses cada uno, Enero a Mayo y Junio a Octubre del 2020, y se realizó un contraste a través del tiempo (o Backtesting) para verificar la distribución de los clusters en los diferentes periodos.

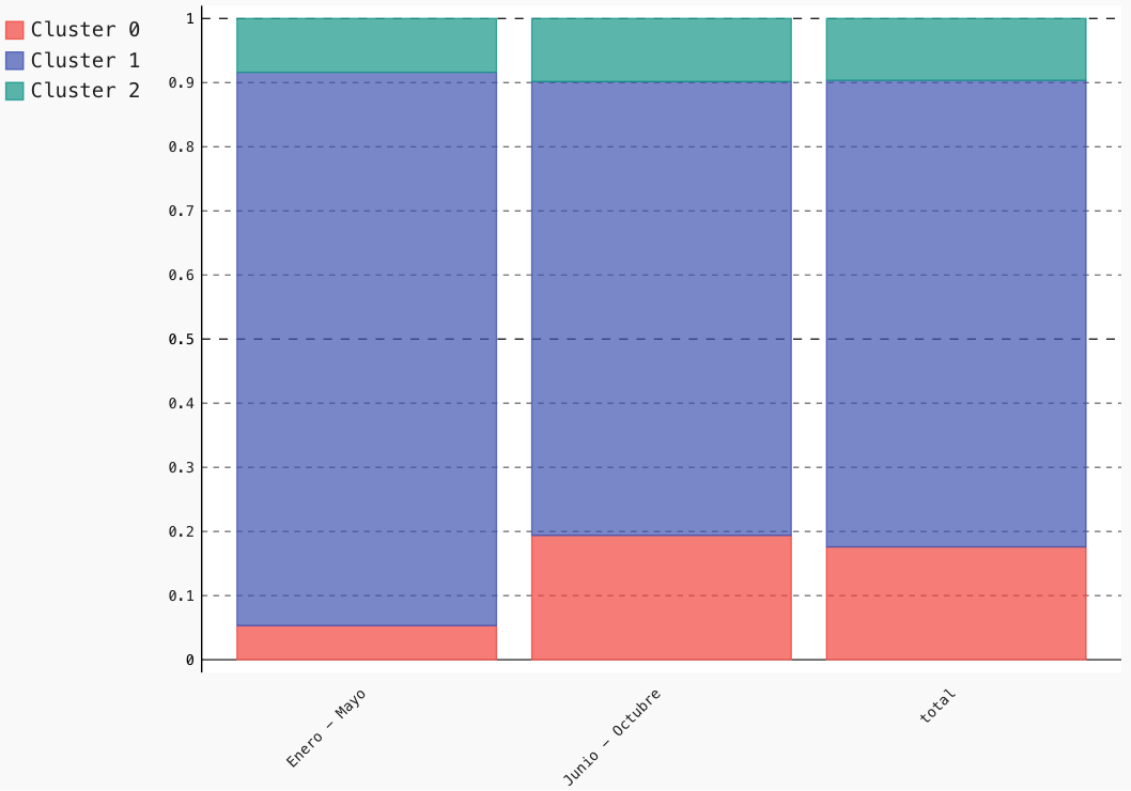


Figura 19: Backtesting

Analizando la Figura 19, es claro que la distribución de los clusters se conserva con el tiempo, para el periodo Junio - Octubre se ve un aumento significativo en el cluster 0 pero esto es explicado por que en ese periodo se alcanzó un pico en los contagios, por lo que hubo un incremento en el número de pacientes como se mostró en el análisis exploratorio.

Finalmente, cabe mencionar que el predominio del grupo de ”Los Vulnerables” es por los malos hábitos alimenticios y de deporte que se tienen en la cultura mexicana, siendo este país de los que encabezan listas con mayor población que padece de problemas relacionados a estos.

11. Aprendizaje profundo

El aprendizaje profundo es una subcategoría del aprendizaje de máquina (o en inglés machine learning), el cual, trata del uso de **redes neuronales** para mejorar cosas tales como el reconocimiento de voz, la visión por ordenador, el procesamiento del lenguaje natural, entre muchas otras aplicaciones.
Por otro lado, se entiende como red neuronal a los programas y estructuras de datos que emulan el funcionamiento del cerebro humano.

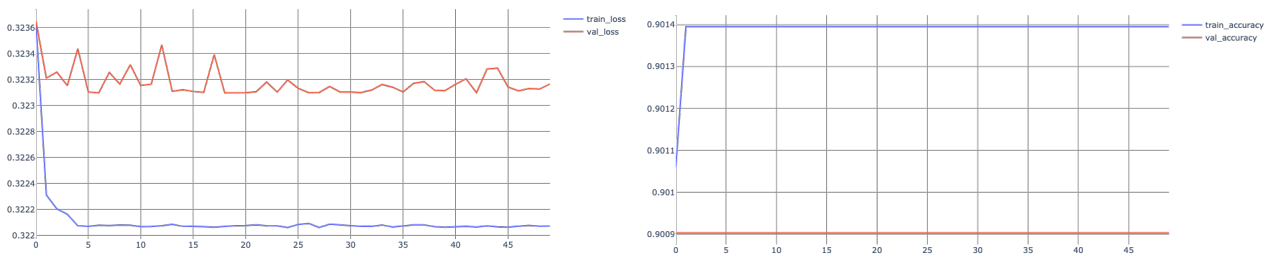
Se recurrió al uso del aprendizaje profundo para modelar el objetivo planteado para la sección de ”modelación supervisada”, el cual, fue predecir la mortalidad de los pacientes con COVID 19 en función de sus padecimientos. Como se muestra en la sección mencionada, no se obtuvieron modelos que predijeran con buena precisión por lo que se optó por diseñar una red neuronal para predecir el mismo fenómeno.

11.1. Arquitectura

La arquitectura que se empleó para la red neuronal es la siguiente:

- Capa de entrada: Recibe 32 variables.
- Primera capa oculta: Tiene 10 neuronas.
- Segunda capa oculta: Tiene 5 neuronas.
- Capa de Salida: Tiene 1 neuronas.

Para las capas ocultas se utilizó la función de activación ”Relu” y para la capa de salida la función ”Sigmoide”, obteniendo los siguientes resultados:



(a) Función de Costo (b) Accuracy Score

Figura 20: Resultados del modelado

Como se aprecia en la Figura 20, con la arquitectura planteada se logró obtener muy buenos resultados de predicción, porque tanto la función de costo, como el accuracy score no se alejan entre las muestras de entrenamiento y validación. Tampoco se cayó en un sobreajuste porque no se está obteniendo scores excesivamente elevados.

12. Scoring

La selección inicial de variables predictivas es un paso muy importante, ya que, tenemos que asegurar que el modelo sea ”parsimonioso”, es decir, que tenga el mayor poder predictivo posible con el menor número de variables.

Tras haber probado diferentes modelos de aprendizaje supervisado y sólo haber obtenido uno con un accuracy score ”aceptable”. Se optó por realizar una ingeniería de datos diferente, de tal manera que se preparan los datos para un modelo de **Scoring**.

Se partió del procesamiento de datos, explicado anteriormente, tras haber imputado los valores faltantes, posterior a ello se discretizó la variable continua ”c_EDAD”, obteniendo cinco variables, las cuales era la edad discretizada en dos bukets, en tres y así sucesivamente hasta tener máximo seis bukets.

En cuanto a las variables categóricas, se seleccionaron aquellas que brindaran información acerca del estado de salud actual del paciente previo a ser atendido en la unidad médica, la entidad de la república y el sector en el que se le brindó el servicio. Utilizando la variable ” CLASIFICACION_FINAL” se filtraron a todos aquellos que fueron confirmados por COVID19.

Finalmente, la variable objetivo ha sido la misma que anteriores modelos, indica la mortalidad en el paciente, 0 en caso que haya sobrevivido y 1 en caso de fallo.

12.1. Medición del poder predictivo

Haciendo uso de un análisis bivariado se obtuvo el peso de la evidencia que indica el poder predictivo de una variable independiente en relación con la variable dependiente mediante el método Weigth of Evidence (WOE).

El **WOE** mide la fuerza predictiva de cada una de las categorías de las variables en la discriminación entre buenos y malos, es equivalente a la probabilidad que una persona en cada categoría sobreviva o falle.

$$WOE = ln(\frac{\%Evento}{\%NoEvento})$$

Una vez se ha calculado el WOE para cada categoría de la variable, se determina el poder predictivo total de la variable, es decir, la capacidad de discriminar sobrevivientes de no sobrevivientes, para esto se calcula el Information Value (IV), que es la suma ponderada de los WOE’s en cada variable.

$$IV = \sum(\%Evento - \%NoEvento) * WOE$$

Variable	IV
d_c_EDAD_6	1.71
v_DIABETES	0.47
v_HIPERTENSION	0.44
v_OTRO_CASO	0.32
v_SECTOR	0.26
v_RENAL_CRONICA	0.12

Cuadro 12: IV de las variables

Si el IV < 1 se considera como un predictor débil y si es IV $> .5$ se considera demasiado bueno para ser cierto, pero recordemos que la selección de las variables queda a criterio del problema o negocio, por lo que se conservó la variable de edad a pesar de tener un IV alto. Por lo tanto, las variables presentadas en el Cuadro 12 son las seleccionadas para entrenar el modelo.

12.2. Modelado

Tras haber sustituido cada característica de las variables anteriores por su respectivo WOE, se entrenó un modelo de regresión logística, el cual, es bastante beneficiado por las transformaciones realizadas anteriormente.

Considérese que se tiene una variable objetivo bastante desbalanceada, por lo que se realizó el modelo con la variable sin balanceo y con balanceo mediante la técnica de undersample (mostrada anteriormente).

Modelo	AUC	ACC	F1
Regresión Logística - Sin Balanceo - Entrenamiento	0.861	0.932	0.108
Regresión Logística - Sin Balanceo - Validación	0.869	0.931	0.107
Regresión Logística - Undersample - Entrenamiento	0.859	0.786	0.786
Regresión Logística - Undersample - Validación	0.869	0.789	0.445

Cuadro 13: Performance de Modelos y conjuntos de datos

Como se aprecia en el Cuadro 13, el AUC Score es similar entre los modelos, mientras que el accuracy es mayor para el modelo entrenado con la muestra desbalanceada, lo cual, hace sentido, dado que a pesar de que se equivoca como hay bastante más sobrevivientes, no le afecta. Pero el F1 Score es el que ayuda a seleccionar el modelo con undersample, ya que, si bien tiene menos accuracy, podemos garantizar que se equivocará menos, ya que, el F1 es mayor.

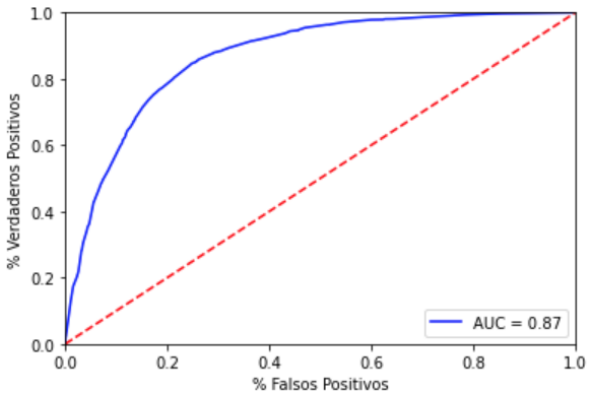


Figura 21: Curva Roc - Conjunto de Validación

12.3. Score Card

La gran ventaja de haber utilizado un modelo de scoring es su interpretabilidad, por lo que mediante 6 preguntas es posible obtener la probabilidad de fallar de los pacientes y esto queda explicado por la score card.

Pregunta	Respuesta	Puntos
¿Qué edad tienes?	[0 - 26]	279
	[27 - 34]	257
	[35 - 41]	182
	[42 - 49]	116
	[50 - 59]	48
	[60 - 89]	-52
¿A qué sector acudes por servicios médicos?	Estatal	43
	IMSS	39
	IMSS Bienestar	68
	ISSSTE	13
	Otro	111
	PEMEX	18
	Privada	94
	SEDENA	4
	SEMAR	82
	SSA	98
¿Padeces de diabetes?	No	83
	Sí	43
¿Tienes hipertensión?	No	81
	Sí	43
¿Has estado en contacto con otra persona que tiene COVID?	No	131
	Sí	35
¿Padeces de insuficiencia renal crónica?	No	75
	Sí	-46

Ahora bien, una vez sumados los puntos de cada pregunta mostrada en el Cuadro anterior, se tienen las siguientes probabilidades de sobrevivencia o fallo:

Score	% Sobrevivencia	% Fallo
[1, 203)	57.40%	42.60%
[203, 404)	80.99%	19.01%
[404, 606)	96.73%	3.27%
[606, 807)	99.50%	0.50%

Cuadro 14: Probabilidades de sobreviviencia y fallo según el score

Tras analizar los últimos dos cuadros, es posible darse cuenta que tu riesgo incrementa ante menor edad y más padecimientos de salud. Esto tiene mucho sentido, ya que, está segmentando a los grupos que son más vulnerables.

13. Conclusión

En conclusión, se logró cumplir con el objetivo de crear una tabla analítica de datos, donde su contenido ya estuviera preparado para poder realizar modelos estadísticos sobre ellos. Se cumplió tras haber transformado a valores numéricos todos los registros y conservar variables que fueran de utilidad para predecir la variable objetivo. Al igual que se separó de manera adecuada la información en conjunto de entrenamiento y de prueba.

Por otra parte, durante este proceso se logró extraer y presentar información que resultara de interés para el tema, tales como identificación del estado con más pacientes, día en el que se presentaron más casos confirmados y día en el que se presentó el mayor número de defunciones, entre muchos otros.

También se logró visualizar los cambios realizados durante el procesamiento de datos y se explicó la razón y el medio por el que se realizaron.

Después de probar varios modelos de aprendizaje supervisado, se logró cumplir con el objetivo principal del proyecto, el cual, era construir un modelo de clasificación capaz de predecir la mortalidad de un paciente con COVID19 y otros padecimientos, obteniendo un modelo de scoring totalmente interpretable y con muy buen performance.

Finalmente, utilizando herramientas de la modelación no supervisada, se obtuvieron resultados desalentadores, porque si bien se generaron los grupos y se hicieron pruebas con las variables y pasaron, los hallazgos no son de utilidad para el problema planteado, ya que, está dividiendo los datos principalmente por el estado y claramente se tendría grupos sin necesidad de hacer un modelado de machine learning. Esto es propiciado por la falta de variables continuas, dado que, se modelo prácticamente con variables dummies, culminando en que no se pudiera cumplir uno de los objetivos planteados inicialmente.

14. Apéndice

14.1. Variables Restantes

En esta sección encontramos el resto de variables no mencionadas anteriormente pero que resulta también de interés su análisis para tener información y resultados completos.

Valriable	Tipo	Descripción
INDIGENA	Categórica	Identifica si el paciente se autoidentifica como una persona indígena.
DIABETES	Categórica	Identifica si el paciente tiene un diagnóstico de diabetes.
EPOC	Categórica	Identifica si el paciente tiene un diagnóstico de EPOC.
ASMA	Categórica	Identifica si el paciente tiene un diagnóstico de asma.
INMUSUPR	Categórica	Identifica si el paciente presenta inmunosupresión.
HIPERTENSION	Categórica	Identifica si el paciente tiene un diagnóstico de hipertensión.
OTRAS_COM	Categórica	Identifica si el paciente tiene diagnóstico de otras enfermedades.
CARDIOVASCULAR	Categórica	Identifica si el paciente tiene un diagnóstico de enfermedades cardiovasculares.
OBESIDAD	Categórica	Identifica si el paciente tiene diagnóstico de obesidad.
RENAL_CRONICA	Categórica	Identifica si el paciente tiene diagnóstico de insuficiencia renal crónica.
TABAQUISMO	Categórica	Identifica si el paciente tiene hábito de tabaquismo.
OTRO_CASO	Categórica	Identifica si el paciente tuvo contacto con algún otro caso diagnosticado con SARS CoV-2.
TOMA_MUESTRA	Categórica	Identifica si al paciente se le tomó muestra.
MIGRANTE	Categórica	Identifica si el paciente es una persona migrante.
UCI	Categórica	Identifica si el paciente requirió ingresar a una Unidad de Cuidados Intensivos.
RESULTADO_LAB	Categórica	Identifica el resultado del análisis de la muestra reportado por el laboratorio de la Red Nacional de Laboratorios de Vigilancia Epidemiológica (INDRE, LESP y LAVE) y laboratorios privados avalados por el INDRE cuyos resultados son registrados en SISVER.
PAIS_NACIONALIDAD	Categórica	Identifica la nacionalidad del paciente.
PAIS_ORIGEN	Categórica	Identifica el país del que partió el paciente rumbo a México.

14.2. Imputación de Valores Faltantes - Restantes

Variables	Moda
v_OTRA_COM	no
v_UCI	no aplica
v_CARDIOVASCULAR	no
v_OBESIDAD	no
v_RENAL_CRONICA	no
v_TABAQUISMO	no
v_OTRO_CASO	si
v_PAIS_ORIGEN	no aplica
v_PAIS_NACIONALIDAD	maxico
v_MUNICIPIO_RES	iztapalapa

14.3. Importancia de Variables

Variable	Importancia
v_DIABETES_si	0.0207
v_INDIGENA_no especificado	0.0188
v_ASMA_si	0.0077
v_SECTOR_semar	0.0062
v_RENAL_CRONICA_si	0.0036
v_INDIGENA_si	0.0035
v_SECTOR_issste	0.0035
v_SECTOR_sedena	0.0033
v_CARDIOVASCULAR_si	0.0024
v_EMBARAZO_si	0.0022
v_INMUSUPR_si	0.0015
v_OTRA_COM_si	0.0012
v_SECTOR_pemex	0.001
v_EPOC_si	0.0005
v_SECTOR_otro	0.0005
v_SECTOR_imssbienestar	0.0004
c_EDAD	0.0003
v_INTUBADO_si	0
v_OTRO_CASO_si	0
v_SEXO_mujer	0
v_UCI_si	0

14.4. Visualización de Datos - Comparación

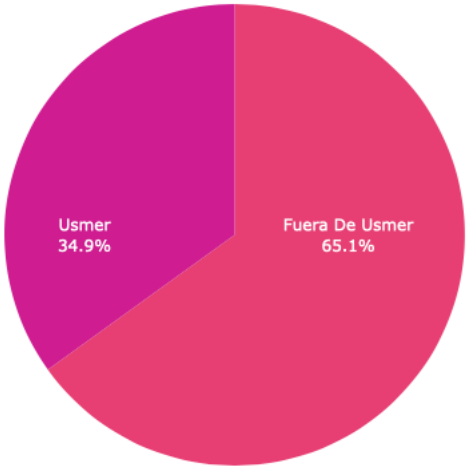
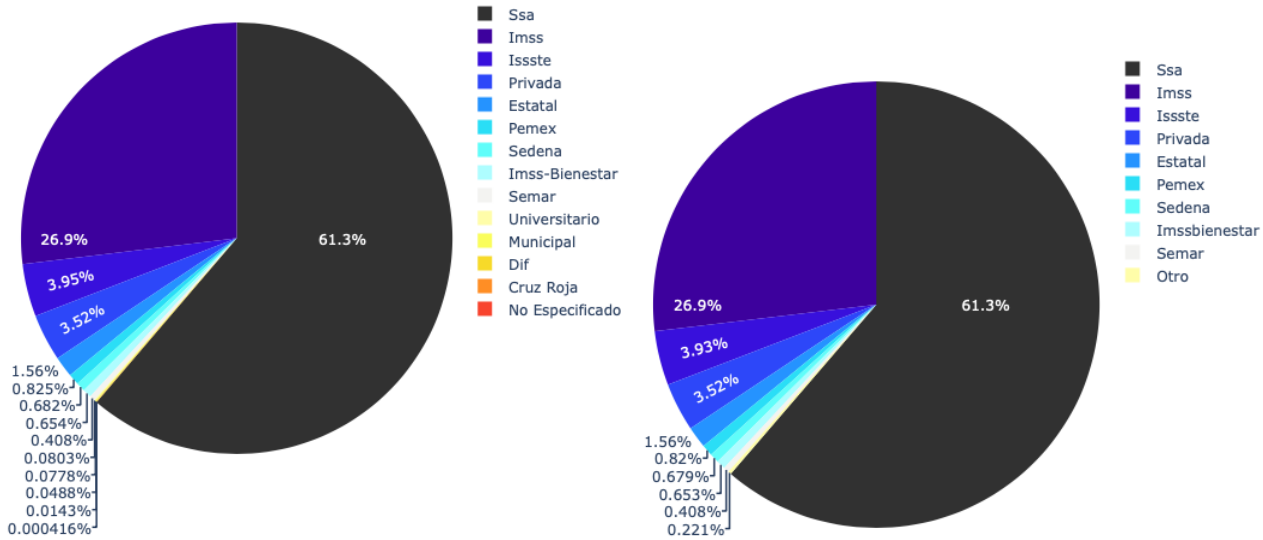


Figura 22: Porcentaje del origen de los pacientes



(a) Sin Procesamiento (b) Con Procesamiento

Figura 23: Comparación de la variable v_SECTOR

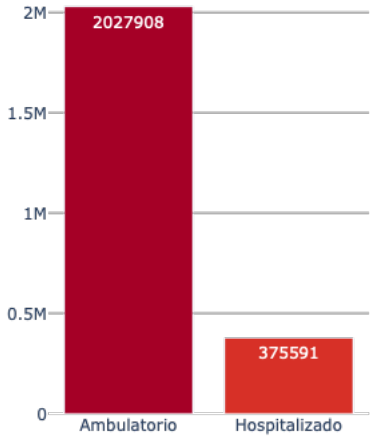


Figura 24: Número de pacientes por cada tipo

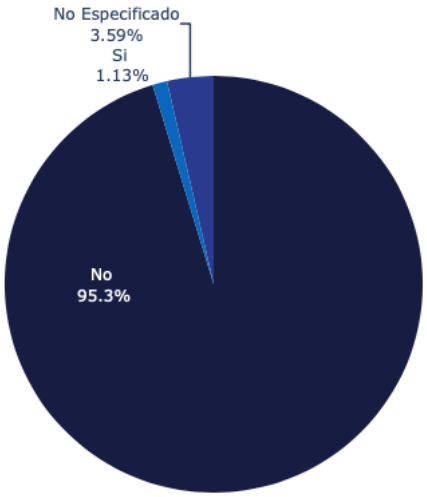
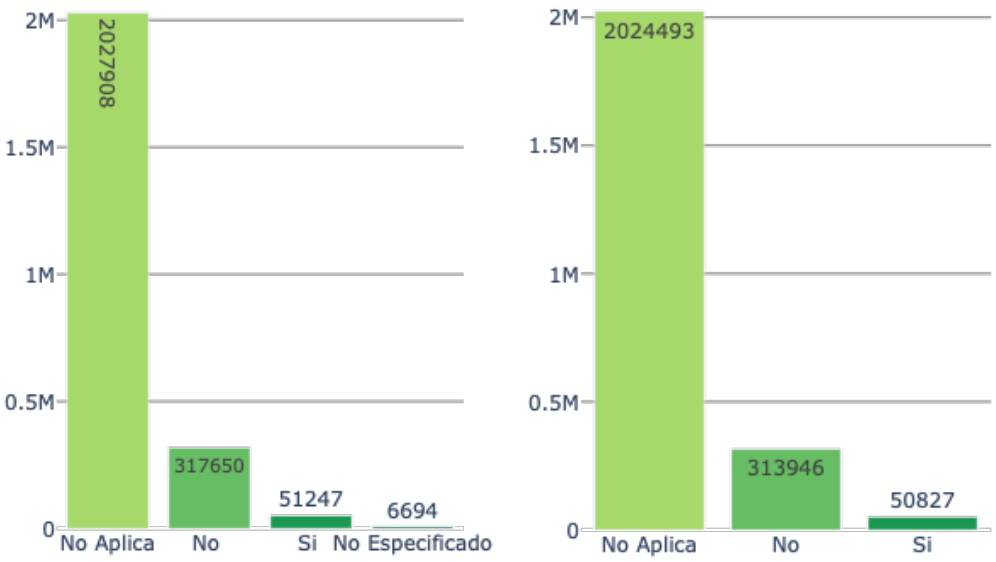
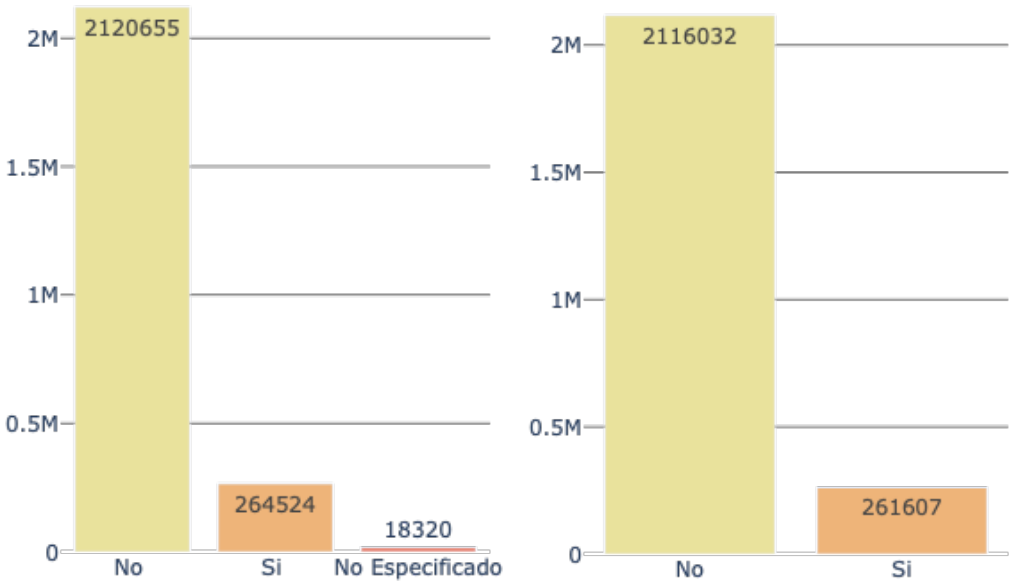


Figura 25: Porcentaje de pacientes que son indígenas



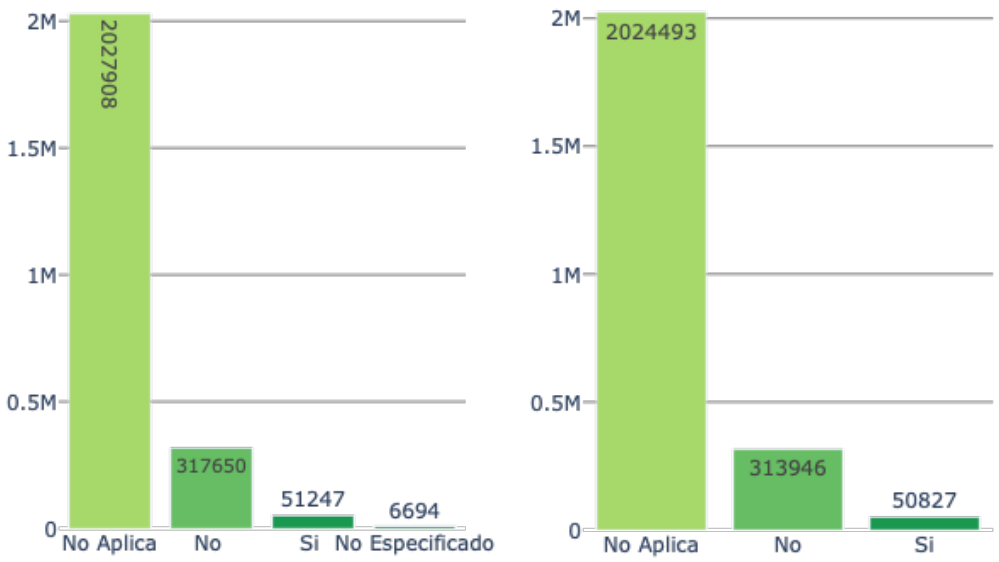
(a) Sin Procesamiento (b) Con Procesamiento

Figura 26: Comparación de la variable v_INTUBADO



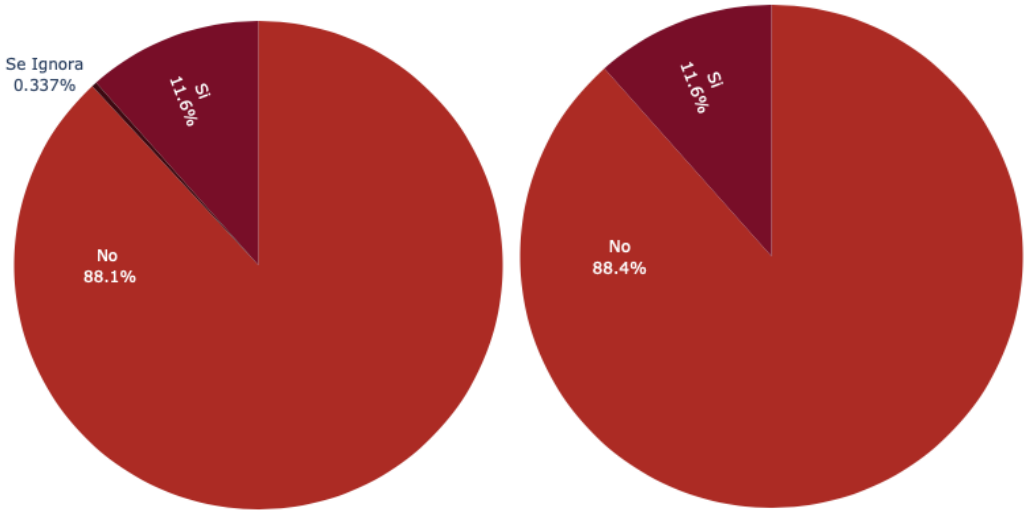
(a) Sin Procesamiento (b) Con Procesamiento

Figura 27: Comparación de la variable v_NEUMONIA



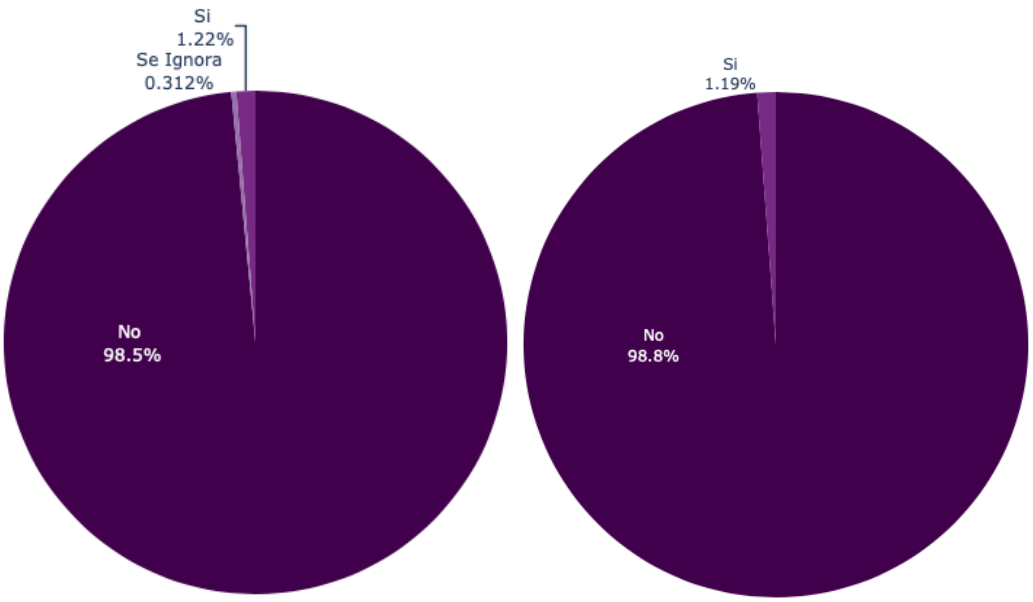
(a) Sin Procesamiento (b) Con Procesamiento

Figura 28: Comparación de la variable v_INTUBADO



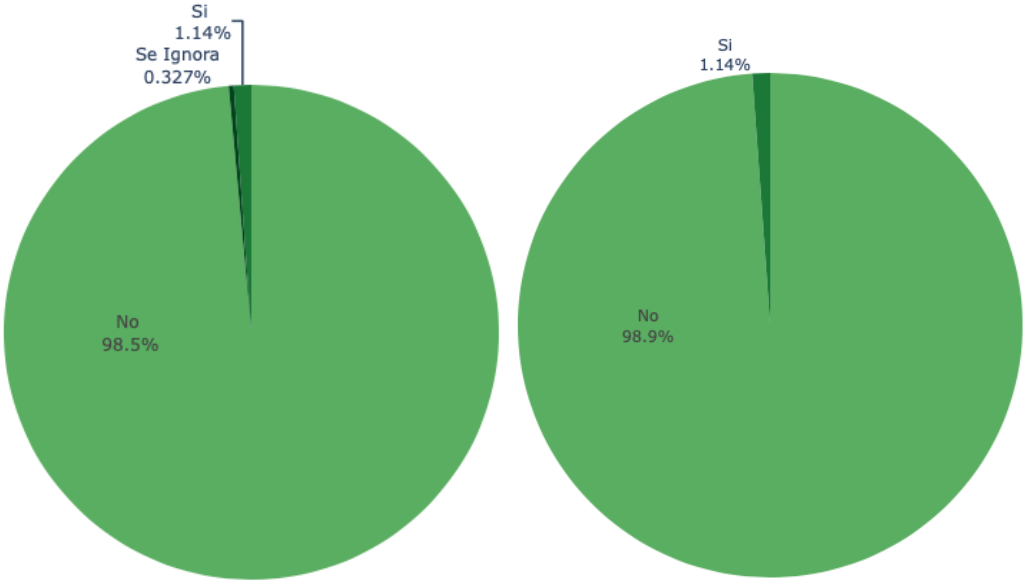
(a) Sin Procesamiento (b) Con Procesamiento

Figura 29: Comparación de la variable v_DIABETES



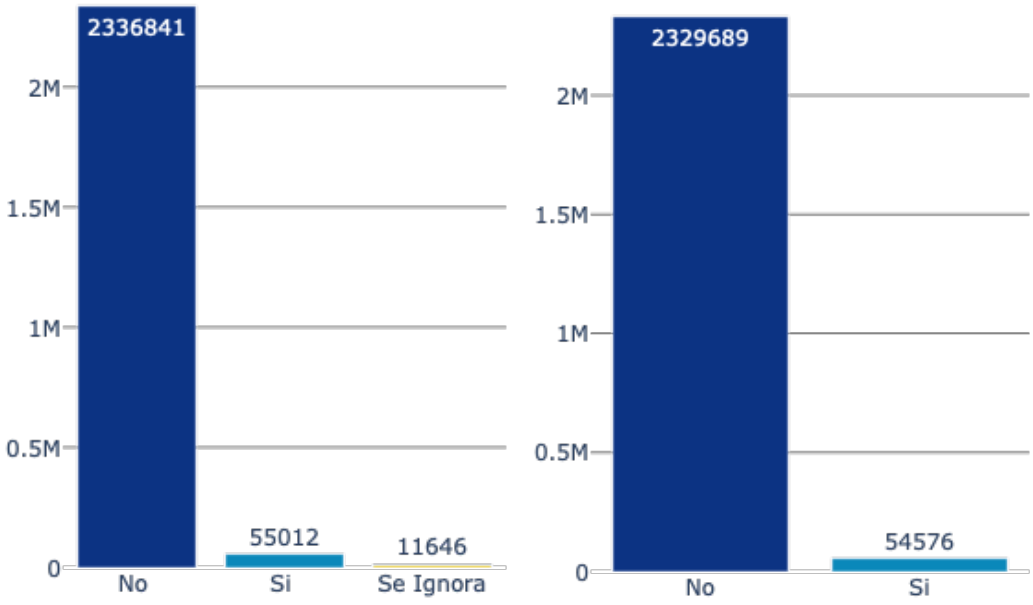
(a) Sin Procesamiento (b) Con Procesamiento

Figura 30: Comparación de la variable v_EPOC



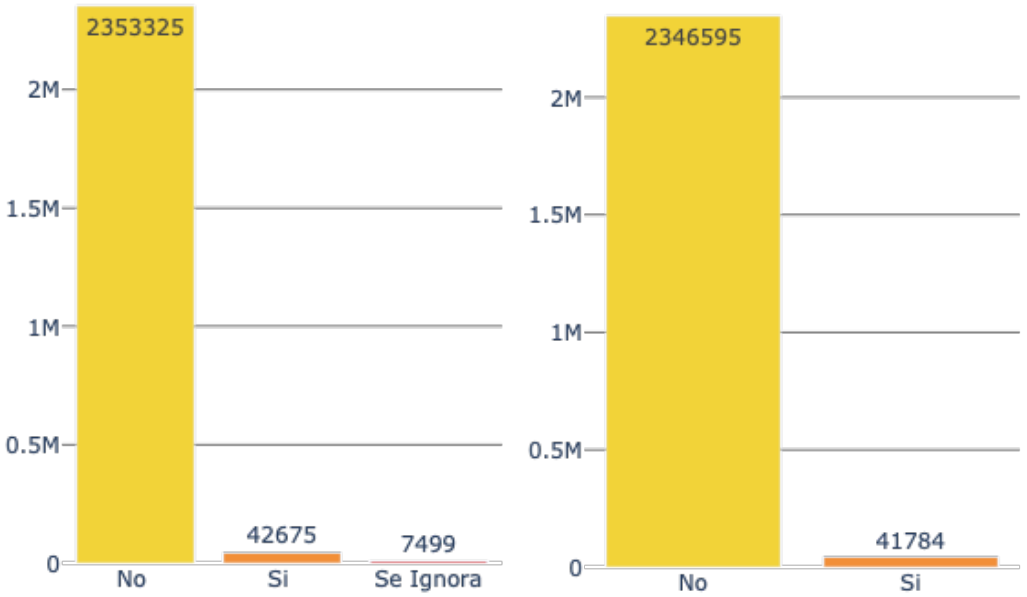
(a) Sin Procesamiento (b) Con Procesamiento

Figura 31: Comparación de la variable v_INMUSUPR



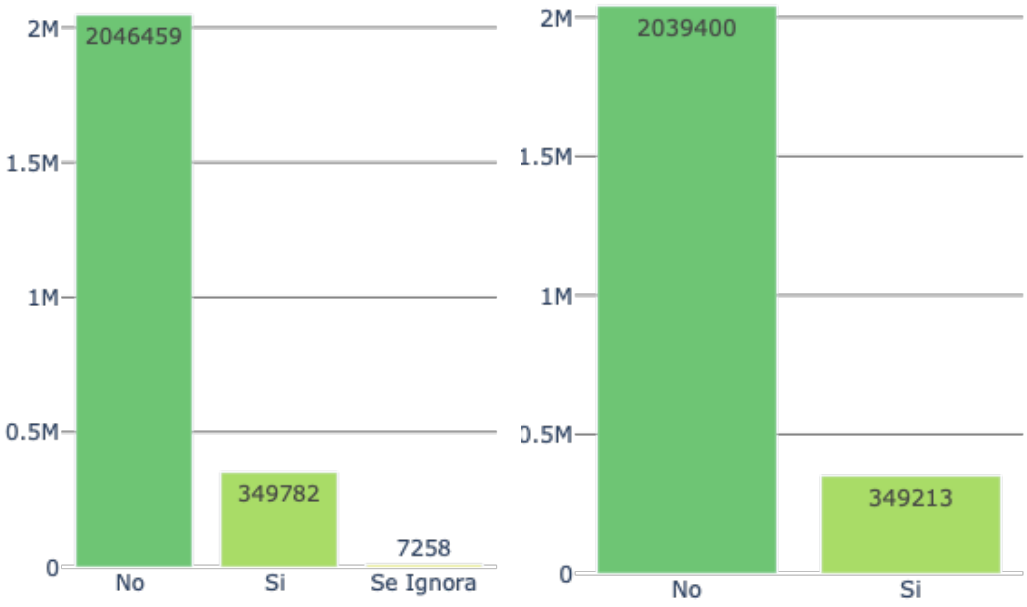
(a) Sin Procesamiento (b) Con Procesamiento

Figura 32: Comparación de la variable v_OTRA_COM



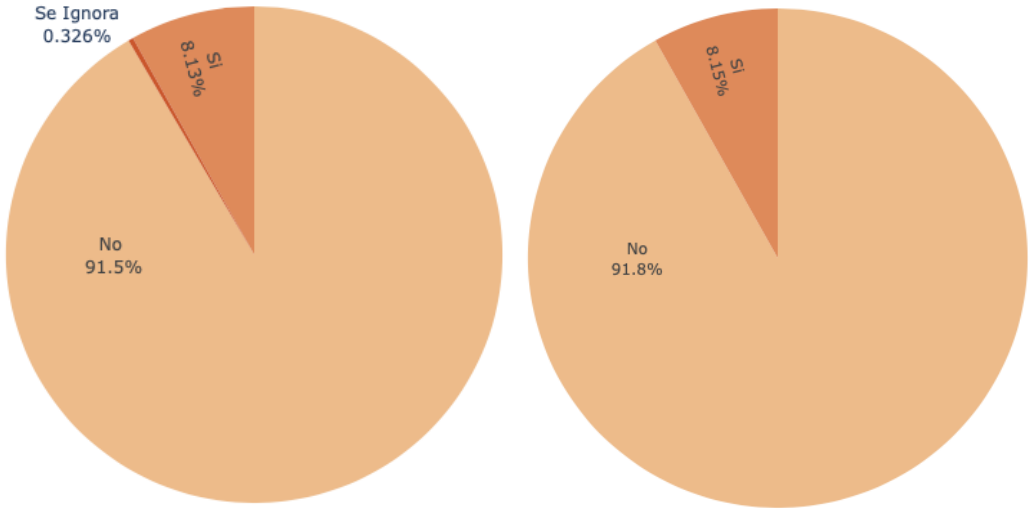
(a) Sin Procesamiento (b) Con Procesamiento

Figura 33: Comparación de la variable v_CARDIOVASCULAR



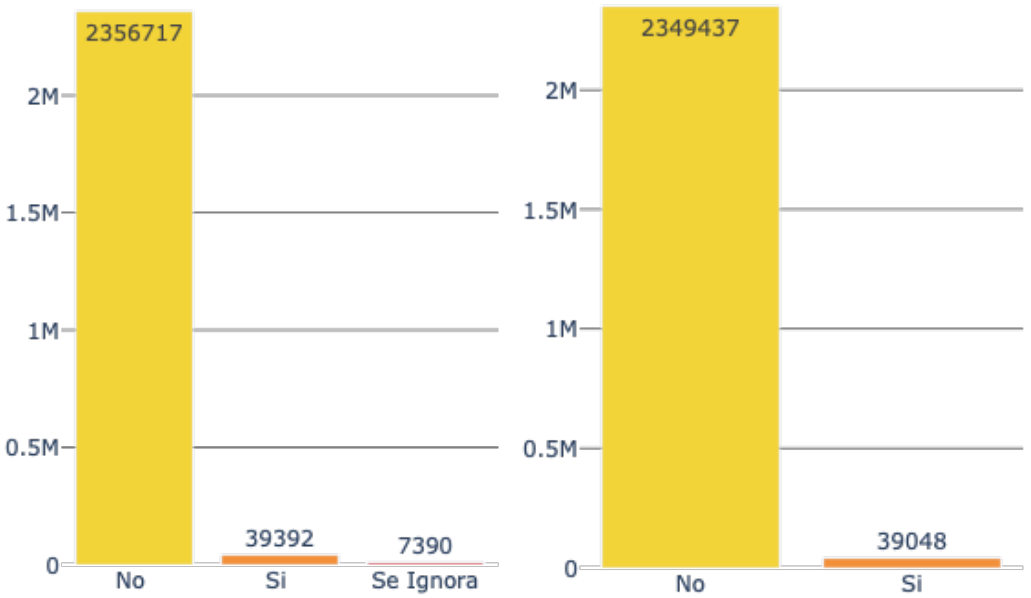
(a) Sin Procesamiento (b) Con Procesamiento

Figura 34: Comparación de la variable v_OBESIDAD



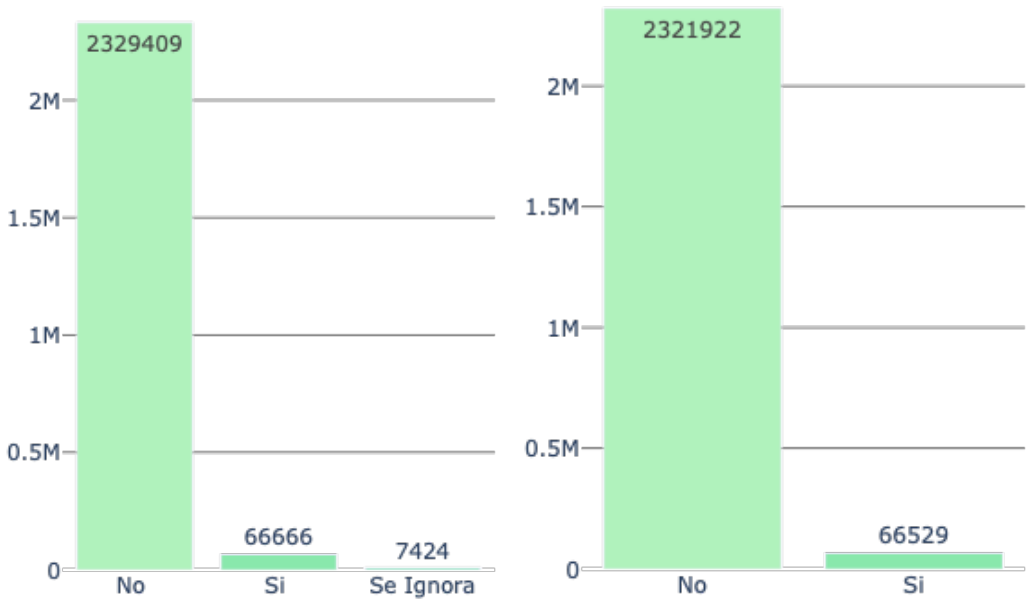
(a) Sin Procesamiento (b) Con Procesamiento

Figura 35: Comparación de la variable v_TABAQUISMO



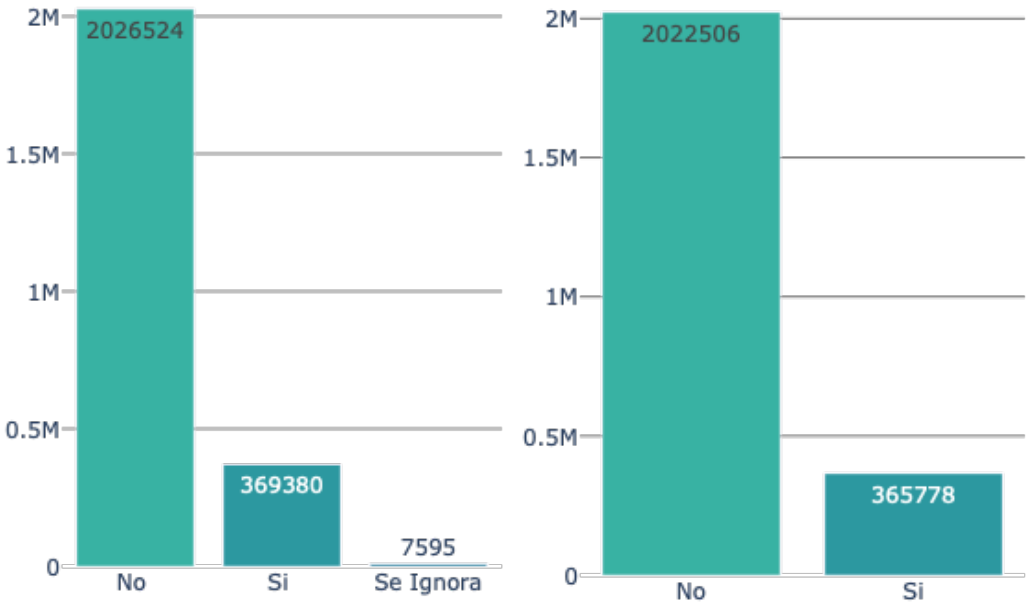
(a) Sin Procesamiento (b) Con Procesamiento

Figura 36: Comparación de la variable v_RENAL_CRONICA



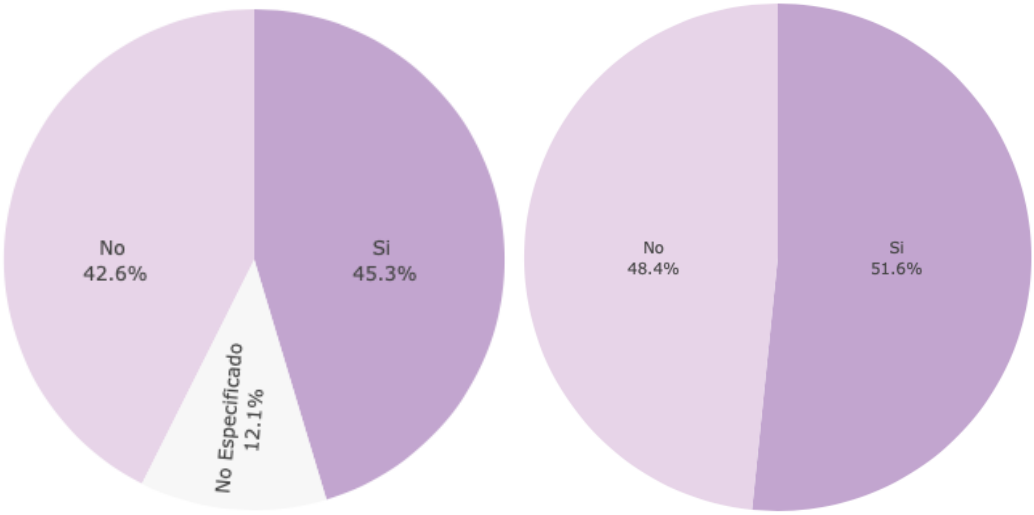
(a) Sin Procesamiento (b) Con Procesamiento

Figura 37: Comparación de la variable v_ASMA



(a) Sin Procesamiento (b) Con Procesamiento

Figura 38: Comparación de la variable v_HIPERTENSION



(a) Sin Procesamiento (b) Con Procesamiento

Figura 39: Comparación de la variable v_OTRO_CASO

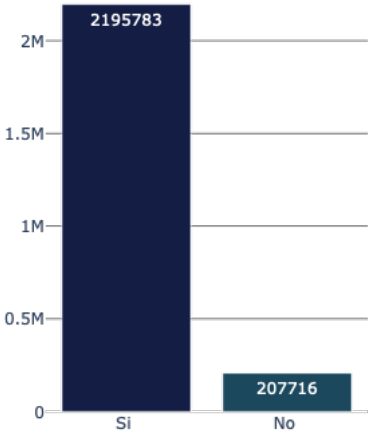


Figura 40: Número de pacientes con que se les tomó muestra

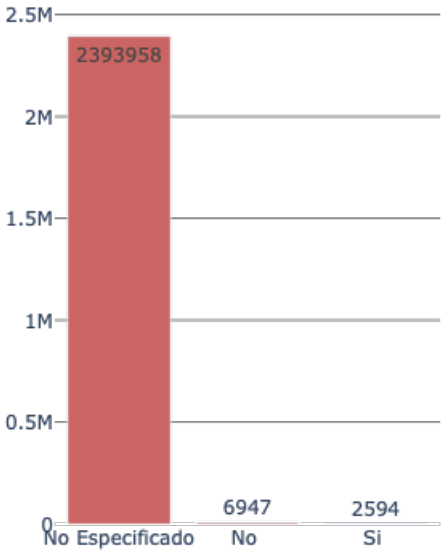
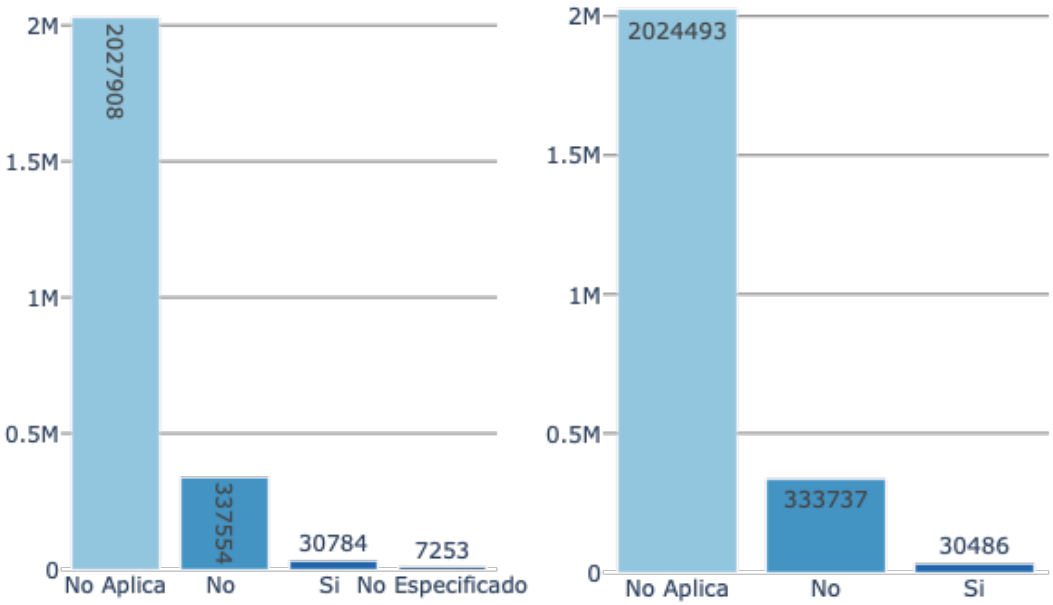


Figura 41: Número de pacientes que son migrantes



(a) Sin Procesamiento (b) Con Procesamiento

Figura 42: Comparación de la variable v_UCI