EVALUACION DE MODELO DE DATOS EN PACIENTES CON COMORBILIDADES CONTAGIADOS DE COVID-19 EN COLOMBIA

DIANA MILENA BELTRÁN BONILLA

MAESTRIA EN INGENIERIA Y ANALITICA DE DATOS (MIAD)
FACULTAD DE INGENIERIA
UNIVERSIDAD JORGE TADEO LOZANO
BOGOTA D.C.
2021

EVALUACION DE MODELO DE DATOS EN PACIENTES CON COMORBILIDADES CONTAGIADOS DE COVID-19 EN COLOMBIA

DIANA MILENA BELTRÁN BONILLA

Trabajo de grado presentado como requisito para optar al título de MASTER EN INGENIERIA Y ANALITICA DE DATOS

Director(es):
PhD. Olmer García Bedoya[1]
MAG. Fredy Guillermo Rodríguez Páez [2]

MAESTRIA EN INGENIERIA Y ANALITICA DE DATOS (MIAD)
FACULTAD DE INGENIERIA

UNIVERSIDAD JORGE TADEO LOZANO
BOGOTA D.C.
2021

[1] Docente de tiempo completo email olmer.garciab@utadeo.edu.co

[2] Docente de tiempo completo email fredyg.rodriguezp@utadeo.edu.co

Tabla de contenido

Tabla de contenido 3
CONTENIDO GRAFICAS 6
AGRADECIMIENTOS 9
1. INTRODUCCION 10
2. MARCO TEÓRICO 11
2.1. MODELADO DE DATOS 11
2.2. MACHINE LEARNING. 11
2.2.1. APRENDIZAJE SUPERVISADO 11
2.2.2. APRENDIZAJE NO SUPERVISADO 11
2.2.3. APRENDIZAJE POR REFUERZO 12
2.3. CLASIFICACION INTERNACIONAL DE ENFERMEDADES 13
2.3.1. ENFERMEDAD NO TRANSMISIBLE 14
2.3.2. ENFERMEDAD TRANSMISIBLE 14
2.3.2.1. COVID-19 14
2.4. ENFERMEDAD CRONICA 14
2.5. BASES DE DATOS 14
2.5.1. TASA DE MORTALIDAD 15
2.5.2. DANE 15
2.5.3. SISPRO 15
2.5.4. DATOS ABIERTOS 15
2.5.5. INSTITUTO NACIONAL DE SALUD 15
3. ESTADO DEL ARTE 16
4. PLANTEAMIENTO DEL PROBLEMA 19
5. OBJETIVOS 20
5.1. OBJETIVO GENERAL: 20
5.2. OBJETIVOS ESPECIFICOS: 20
6. METODOLOGIA 21
6.1. FASE DE ANALISIS 21
6.2. ANALISIS INICIAL 21
6.2.1. ANALISIS COMPARATIVO DE LAS COMORBILIDADES 21
6.2.2. ANALISIS COMPARATIVO DE FALLECIDOS POR EDAD 22
6.3. CONVERSIÓN DE LOS DATOS EN VARIABLES, DE LOS CASOS POSITIVOS COVID-19 EN COLOMBIA. 23
6.3.1. PREPROCESAMIENTO DE LOS DATOS 23

6.3.1.1. ANALISIS DE VARIABLES. 23
6.3.1.2. Agrupamiento de Conjuntos 25
6.4. PROCESAMIENTO DE DATOS 26
6.4.1. Agrupamiento interno de variable 26
6.4.2. Binarización de variables categóricas 26
6.4.3. Agrupamiento de datos 27
6.4 Analítica Descriptiva del contagio DEL COVID-19 EN COLOMBIA 28
6.5. ESTUDIO DE LAS SERIES TEMPORALES 34
6.6. ESTUDIO DE LA DIFUSION VIRAL 35
6.7. DEFINICION DE SERIES TEMPORALES PARA PRONOSTICOS 37
6.8. MODELADO DE SERIES TEMPORALES 39
6.9. VISUALIZACION DE DATOS RELACIONALES 42
7. MODELAMIENTO DE LAS MUERTES 43
7.1. MODELAMIENTO DE LAS MUERTES POR EDAD PARA CORRELACIONES 44
7.2. MODELAMIENTO DE LAS MUERTES POR SEXO PARA CORRELACIONES 45
7.3. MODELAMIENTO ESTADISTICO 46
7.3.1. AJUSTES DEL MODELO 47
7.3.2. LINEAMIENTOS PARA CREAR EL MODELO. 47
7.3.2.1. COEFICIENTE DE CORRELACION R^2 47
7.3.2.2. COEFICIENTE DE CORRELACION $R^2_{ajustado}$ 47
7.3.3. SIGNIFICANCIA DEL MODELO F-Test 47
7.3.4. VALOR ESTADISTICO T 48
7.3.5. PRINCIPIO DE NO COLINEALIDAD 49
7.3.6. PRINCIPIO DE NO MULTICOLINEALIDAD 49
7.3.7. PRINCIPIO DE HOMOCEDASTICIDAD 49
7.3.8. PRINCIPIO DE NO HETEROCEDASTICIDAD 49
7.4. PRUEBAS INICIALES DE CORRELACION 49
7.3. PROCEDIMIENTO PARA EVALUAR LAS CORRELACIONES CON EL DATASET DE CASOS POSITIVOS COLOMBIA 52
7.3.1. MODELAMIENTO DE VARIABLE SEXO PARA CORRELACION 52
7.3.2. MODELO DE CORRELACION ENTRE SEXO FEMENINO Y MUERTE POR COVID 53
7.3.3. MODELO DE CORRELACION ENTRE SEXO MASCULINO Y FALLECIMIENTO POR COVID 55
7.4. PRUEBAS DE REGRESION 56
7.4.1. REGRESION LINEAL ENTRE LA EDAD Y LA MUERTE POR COVID 57
7.4.2. MODELO ESTADISTICO DE LAS COMORBILIDADES DEL COVID 58

7.4.2.1. ENTRENAMIENTO DE LA MAQUINA PARA LA PREDICCIÓN DE LAS REGRESIONES ML. 61
7.8.1. PRONOSTICO DE LA MORTADAD DE COVID-19 ASOCIADO A LAS COMORBILIDADES 64
8. PRONOSTICO DE LAS MUERTES POR COVID-19 ASOCIADAS A LAS COMORBILIDADES.. 70
9. DESPLIEGUE 71
10. CONCLUSIONES 72
APENDICE A 73
A.1. CORRELACION MULTIPLE ENTRE LAS VARIABLES INDEPENDIENTES Y LA VARIABLE DEPENDIENTE. 73
APENDICE B 74
CLASIFICACION AUTOMATICA DE LOS GRUPOS DE ENFERMEDADES DATASET USA. 74
9. BIBLIOGRAFIA 78

CONTENIDO GRAFICAS

Grafica 1. Relación porcentual comparativa por comorbilidades, modelado en Power BI, entre los fallecidos por comorbilidades como causa de muerte, y los fallecidos por covid-19 con comorbilidades, en el periodo 2020-2021. _____ *22*
*Grafica 2. Relación porcentual comparativa por edades decenal, modelado en Power BI, entre los fallecidos por comorbilidades como causa de muerte, y los fallecidos por covid-19 con comorbilidades, en el periodo 2020-2021. fuentes: DANE e INS.*_____ *22*
Grafica 3. Histograma de Edades. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *25*
*Grafica 4. Relaciona porcentual de probabilidad de supervivencia o de muerte con el virus. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.*_____ *29*
Grafica 5. Contagios por Edad. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS _____ *29*
Grafica 6. Recuperados por Edad. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *30*
Grafica 7. Porcentaje de Muertes por Edad. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *30*
Grafica 8. Total Hombres Contagiados. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *31*
Grafica 9. Total Mujeres Contagiadas. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *31*
Grafica 10. Georreferenciación de casos en Colombia por departamento. Fuente: Imagen generada desde Power BI tomando el set de datos descargado de INS y el DANE _____ *36*
Grafica 11. Top 10 de incremento de Casos por Departamento. Fuente: Imagen generada desde Power BI tomando el set de datos descargado de INS. _____ *36*
Grafica 12. Series temporales y su pertinencia. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *38*
Grafica 13. Contagios por semana. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *40*
*Grafica 14. Validación pertinencia de la Cuarentena en la trasmisión del Virus. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.*_____ *40*
Grafica 15. Contagios por Mes. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *41*
Grafica 16. Correlación entre Sexo y Edad respecto al contagio del Virus. Fuente: Casos de Contagios NO reportados. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. ____ *42*
Grafica 17. Total de casos de enfermos de neumonía e influenza en Estados unidos. Fuente: DatasetConditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020-2021, Imagen generada desde Python. _____ *50*
Grafica 18. Regresión por Edad y Sexo. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *56*
Grafica 19. Prueba de los datos reales vs la regresión. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *56*
Grafica 20. Fallecidos por Edad. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS. _____ *57*
Grafica 21. Diagrama estimativo, porcentual izado y totalizado de Pacientes Muertos por covid-19 en Colombia, Diseñado en Python, fuente: Dataset Descargado INS. _____ *58*
*Grafica 22. Distribución de Muertes por comorbilidad Cerebrovascular contra el total de fallecidos por el rango de edad respectivo. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.*_____ *59*
Grafica 23. Modelamiento de OLS. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS. _____ *60*
Grafica 24. Visualización de los límites. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS. _____ *63*

Grafica 25. Pronóstico de Muertes por COVID-19 por las edades más afectadas con Comorbilidades, estimación por cada mil habitantes. Fuente: Imagen generada de Power BI, modelado en Python con Facebook Prophet, en la plataforma Google Colab, tomando el set de datos _____ *67*
Grafica 26. Tendencia de Muertes Diabetes Mellitus. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS. _____ *67*
Grafica 27. Error Cuadrático Medio MSE. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS. _____ *69*
Grafica 28. Error Absoluto MAE. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS. _____ *69*
Grafica 29. Tendencia de muertes Covid-19 por comorbilidades, estimación por cada 1000 habitantes, creada En Power Bi, Generada en Python, con la plataforma Google Colab, Fuente: Set De Datos Descargado de la INS. _____ *70*

Dedicatoria...

A Mi Hijo:

Juan José Contreras Beltrán

A mis Padres:

Luz Stella Bonilla Castiblanco y Pascual Beltrán.

A mis Hermanas:

Paola Beltrán Bonilla

Bibiana Beltrán Bonilla

Por tener la paciencia y comprender el poco tiempo para dedicarles en los dos últimos años.

"La gratitud siempre tiene cabida en nuestra vida. Estudios demuestran que la gente agradecida es más feliz porque en vez de preocuparse por las cosas que le faltan, agradece lo que tiene."

Dan Buettner

AGRADECIMIENTOS

Agradezco a Dios por bendecir mi vida, por guiarme a lo largo de mi existencia, por ser el apoyo y la fortaleza en todos aquellos momentos de dificultad y de debilidad. Gracias a mis padres, Luz Stella Bonilla Castiblanco y Pascual Beltrán Rincón, a mi hijo Juan José Conteras Beltrán, a mis hermanas Paola Beltrán Bonilla y Bibiana Beltrán Bonilla. Por ser los principales promotores de mis sueños, por confiar y creer en mis expectativas, por los consejos, valores y principios que me han inculcado durante toda mi vida. Agradezco a los docentes de la Universidad Jorge Tadeo Lozano, por haber compartido sus conocimientos a lo largo de la preparación de la Maestría, de manera especial, al PhD Olmer García Bedoya y MAG. Fredy Guillermo Rodríguez Páez, tutores de mi proyecto de investigación y quienes me han guiado con su paciencia, rectitud y profesionalismo.

1. INTRODUCCION

La estructuración del proyecto tiene como fin la evaluación del impacto que ha tenido la reciente pandemia del **SARS-CoV-2 (COVID-19)**, en los servicios hospitalarios y en la tasa de muertes para pacientes con enfermedades crónicas como la **DIABETES MELLITUS**, **ENFERMEDADES RESPIRATORIAS**, **OBESIDAD**, entre otras. Para llegar a los resultados esperados los estudios se realizarán bajo los datos estadísticos de entidades como El Departamento Administrativo Nacional de Estadística (DANE), **JOHNS HOPKINS UNIVERSITY**, la **ORGANIZACIÓN MUNDIAL DE LA SALUD (OMS), DATOS ABIERTOS, INSTITUTO NACIONAL DE SALUD (INS).**

Para contextualizar sobre el impacto de enfermedades clasificadas como comorbilidad, enunciamos al caso de la **DIABETES MELLITUS** en el mundo, nos basamos en los datos de la **OMS,** que en su reporte del año **2014** indica para ese tiempo se alcanzaba los **422 millones** de personas adultas con Diabetes Mellitus a nivel mundial, siendo el **8,5%** de participación comparada con las otras enfermedades clasificadas como crónicas en la población adulta, adicional que es una de las con mayor crecimiento si es comparada con el año **1980** que tan solo alcanzaba los **108 millones** de personas con la enfermedad, y con una participación del **4,7%** sobre el total de las enfermedades existentes mundialmente para ese año[1].

Si retornamos a la actualidad, en donde nos enfrentamos a una pandemia como el **COVID-19,** y que con el tiempo está dejando retos al sector salud, mismo que a la fecha viene con un gran déficit de personal especialista y calificado, que a su vez viene siendo desmejorado y descuidado estatalmente por carencia de presupuesto y de condiciones laborales, además el cual en emergencias sanitarias como las que vivimos en la actualidad, debe tener el mismo nivel de respuesta para sus pacientes con es importante evaluarlo para tal vez bajo estos resultados evaluar mejoras y planes de acción preventivos.

En cuanto al servicio que se presta por las entidades de salud, puede que se evidencien algunas falencias como los son la falta unificación de Subredes y planeación de las organizaciones, sin embargo logran ser omitidas por altos directivos del sector y dar reportes de estabilidad e informes con servicios de calidad, tal como se indica en la **Boletín de Prensa No 848 de 2020**, donde el Ministro de Salud **Fernando Ruiz Gómez** donde enuncia lo siguiente, *"Si no hubiéramos tenido un sistema de salud mixto público-privado integrado, un sistema con una capacidad de respuesta de aseguramiento y con una cobertura universal superior al 95% no hubiéramos podido responder"*, es decir , el sector salud esta con todo en orden[2].

Sin embargo, se puede aclarar que los datos que se arrojan por estas entidades son tardíos y en muchas ocasiones se generan para ser más estudios correctivos y no preventivos. Debido a la falta de integración de las entidades públicas y privadas dueñas de la información y que complican los estudios bajo este sector primordial.

Por lo expuesto anteriormente, en este proyecto se realizó un estudio donde se estimó cuantificar y calcular, las posibles alteraciones en el servicio prestado a pacientes diagnosticados con **COMORBILIDADES** e infectados con el virus del **COVID-19,** durante el primer año tras la llegada de la pandemia a nuestro país y pretende evaluar, si toda la crisis hospitalaria que ha generado la pandemia, ha ocasionado desmejoras en los procesos habituales para estos pacientes, en la asignación de medicamentos y si de forma directa ha afectado en los resultados de la letalidad, todo esto bajo modelos de aprendizaje No supervisados y supervisados de *Machine Learning* **(ML)**.

2. MARCO TEÓRICO

En esta sección se encuentran los diferentes componentes para el desarrollo de este proyecto tanto de las especificaciones médicas, como los conceptos técnicos y teóricos a utilizar para la realización del Estudio transversal.

2.1. MODELADO DE DATOS

Es la forma de estructura, organizar y clasificar los datos para que puedan ser usado fácilmente por las bases de datos. Lo que se busca en el modelamiento de datos es dejar disponible la información para la correcta lectura y que los diferentes procesos se faciliten tanto su integración como lectura. Existen tres tipos de modelamientos de datos, **Modelos de datos conceptuales,** que se encargan de un conocimiento previo de la información, basándose en las estructuras y finalmente precursores de los **Modelos de datos lógicos** (MDL), este modelo representa los tipos de entidades lógicas, se denominan tipos de entidades, los atributos de datos son los que detallan esas entidades y las relaciones entre entidades. Finalmente, **Modelos de datos físicos (MDF)**, que son usados para el diseño interno de las bases de datos y la relación entre datos[3].

2.2. MACHINE LEARNING.

En español también es conocido como APRENDIZAJE AUTOMATICO o APRENDIZAJE DE MAQUINA, es una rama científica de la inteligencia Artificial (IA), la cual facilita a los sistemas la adquisición de conocimiento progresivo, con mejoras a las tareas y análisis de los datos continuo, sin algún tipo de programación. El *machine Learning (ML),* toma como base la información conseguida en los datos analizados por los sistemas, de alguna manera imitando el comportamiento humano; Con esto ya se pueden crear modelos predictivos, para las posteriores tomas de decisiones, con gran nivel de eficiencia en los resultados[4]. Existen varios tipos de ML, como son el aprendizaje supervisado, aprendizaje no supervisado, aprendizaje profundo.

2.2.1. APRENDIZAJE SUPERVISADO

El aprendizaje del tipo supervisado hace referencia a un Modelo Especifico de Aprendizaje Automático, y se basa en la generación de conocimiento basada en ejemplos o data con los cuales le enseñamos a la maquina como realizar clasificaciones, Con los resultados obtenidos el modelo sugiere un tipo de ajuste en la parametrización interior para siempre tener adaptación al ingreso de nueva información. Con estos modelos se logra realizar predicciones pertinentes de las conductas de los datos, por ello es muy utilizando en aplicaciones tecnológicas como detectores de imágenes, de Spam en los correos o reconocimientos de voz[4].

2.2.2. APRENDIZAJE NO SUPERVISADO

Este tipo de aprendizaje, se incluyen los grupos de datos sin etiqueta, de los que se desconoce su estructura, debido a que en este tipo de aprendizaje se busca obtener información sin tener conocimiento del resultado, estimando conseguir variables claves

o relevantes para el modelo. En Aprendizaje no supervisado encontramos dos categorías que son: **Reducción Dimensional** y ***Agrupamiento de Variables***. La Reducción dimensional, abarca datos de alta complejidad, contemplando la correlación de estos mismos y eliminando la redundancia de información, analizando con mayor eficiencia los datos considerados relevantes para el modelo. En cambio, el *Agrupamiento de variables*, es un proceso exploratorio, que clasifica los datos por grupos aun sin conocer su estructura, todo esto lo hace bajo características de datos similares. Estas tipologías son comúnmente usadas en proyectos de mercadeo, donde se busca segmentación de clientes, nichos y demás variables específicas[4].
En el aprendizaje no supervisado buscamos que la maquina aprenda a clasificar de acuerdo con los datos que obtiene y genere el modelo por su propia cuenta, esperando que se arrojen resultado esperados.

2.2.3. APRENDIZAJE POR REFUERZO

Este modelo es se asemeja a los mencionados anteriormente, pues en este modelo se hace el uso de *Deep Learning* (DL), este tiene como finalidad el mejoramiento del rendimiento , tomando como base los resultados o las interacciones realizadas, es un sistema de recompensa, en conclusión es un modelo de entrenamiento constante que busca programar todas las combinaciones posibles, con el fin de tener mejores resultados y predicciones aún más acertadas[5].

2.2.4. CORRELACION LINEAL

Es un método para cuantificar la relación que existe entre dos variables, en su método matemático se busca hallar la covarianza entre ambas, es decir que tanto varían

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(Xi - \bar{x})(Yi - \bar{y})}{N-1}$$

Donde:
Xi= Observación de X en el punto i
Yi= Observación de Y en el punto i
$\bar{x} = Media\ de\ la\ variable\ X$
$\bar{y} = Media\ de\ la\ variable\ y$

Todos estos valores varían entre -1 y 1 con lo cual debemos tener presente la fuerza de la asociación:

Correlación nula: 0
Correlación pequeña: 0.1
Correlación mediana: 0.3
Correlación moderada: 0.5
Correlación Alta: 0.7
Correlación muy alta: 0.9
Correlación perfecta: 1

Debemos tener presente que la correlación perfecta genera problemas de homocedasticidad, por lo que es necesario que estos valores se encuentren entre 0.1 y 0.99, para poder realizar la regresión lineal con la cual podamos aceptar o rechazar la hipótesis [6]**.**

2.2.5. REGRESION LINEAL

La regresión lineal es un modelo de correlación lineal, en la cual se busca predecir el valor de una variable dependiente por medio de una variable o varias variables independientes, es decir puede ser simple o múltiple, en la regresión se genera una hipótesis que debe ser aceptada o rechazada de acuerdo a la capacidad del modelo para predecirlas[7] , en términos matemáticos tenemos que:

$$yi=\beta0+\beta1xi1+\beta2xi2+...+\beta pxip+\epsilon i$$

Donde:
yi: Valor de la observación y en el punto i
β0: valor promedio de y cuando los predictores Xj = 0
βj: coeficientes parciales de predicción.
Xj: variables predictoras.

2.2.6. LIBRERIAS DE REGRESION

En este caso utilizaremos las librerías StatsModel con la cual realizaremos la regresión lineal simple, y Scikit-learn con la cual realizaremos una predicción de esta, esto con el objetivo de poder definir que tanto aprende la máquina, en su proceso de inteligencia artificial, también incluiremos el método de Pearson con lo cual lograremos definir cuáles son las variables que más se ajustan para poder realizar el pronóstico, del comportamiento de estas.

2.2.7. METODO DE MODELAMIENTO.

Para esta investigación utilizaremos el método conocido como Ordinary Least Squared (OLS), También conocido como método de Mínimos Cuadrados.
Para el caso de la regresión múltiple utilizaremos el método de correlación de Pearson, el cual nos permite realizar la una múltiple correlación entre las variables de forma lineal.

2.3. CLASIFICACION INTERNACIONAL DE ENFERMEDADES

La Clasificación Internacional de Enfermedades y problemas relacionados con la Salud (CIE), es la encargada de codificar y clasificar las estadísticas de morbilidades y mortalidad en el mundo desde 1994 por la OMS[8], en la actualidad Colombia se rige bajo la versión CIE-10, dividiéndose en los siguientes factores :

- Enfermedades no transmisibles
- Enfermedades transmisibles y nutricionales
- Causas externas de morbilidad y mortalidad- Lesiones
- Causas mal definidas
- Condiciones maternas, perinatales

La clasificación se realizó bajo la Lista del Estudio de Carga Mundial de la Enfermedad (Global Burden of Disease), y a continuación se definirán únicamente, las que son objeto del desarrollo de este proyecto, aclarando que las codificaciones de las enfermedades se tomaron a 4 dígitos, es decir, Una letra y 3 números:

2.3.1. ENFERMEDAD NO TRANSMISIBLE

Las Enfermedades No transmisibles (ENT), son aquellas de larga duración debido a factores, ambientales, genéticos, fisiológicos y conductuales, vienen siendo la mayor causal tanto de muertes como discapacidades en el mundo, se les da ese término de (ENT) porque no son adquiridas de manera transmisible y son susceptibles a duraciones extensas, evolutivas y a ser propensas de tratamiento, con cuidados a mediano y largo plazo. Entre ellas podemos encontrar enfermedades, cardiovasculares, pulmonares crónicas, cáncer, diabetes, trastornos mentales, entre otras[9].

2.3.2. ENFERMEDAD TRANSMISIBLE

Las Enfermedades transmisibles (ET), son aquellas enfermedades que son contagiosas, infecciosas o que el paciente es susceptible por sus productos tóxicos, según la OPS las poblaciones más vulnerables a este tipo de enfermedades son las de bajos recursos pues en están sujetas a factores externos como agua potable, cambio climático, inequidades de Sexo, factores socioculturales, económicos entre otros. Algunas de estas enfermedades son: Sarampión, VIH SIDA, Tuberculosis, Malaria[10] y la sujeta a nuestro caso de estudio que es el **COVID-19**.

2.3.2.1. COVID-19

El virus **COVID-19**, tiene como origen es en Wuhan (China) desde el 31 de diciembre de 2019[13], la enfermedad causada por el nuevo coronavirus conocido como **SARS-CoV-2** y la que se estimaba en sus inicios como una “neumonía vírica”[11].

2.4. ENFERMEDAD CRONICA

Una enfermedad Crónica es aquella que su duración es prolongada y con progresión lenta y continua. Entre ella podemos encontrar el cáncer, enfermedades de tipo cardiaco, Artritis, asma, diabetes, etc. Este tipo de enfermedades son las causales del aproximadamente **63%** de todas las muertes en el mundo, sin distinción de Sexo y teniendo para el año **2008**, un aproximado de 36 millones en personas menos de 60 años[12].

2.5. BASES DE DATOS

Conocida también con el termino en ingles de *database*, se refiere a Información que se relaciona entre sí, almacenada y organizada para facilitar su búsqueda, utilización, preservación y uso. Durante los años las bases de datos han evolucionado de manera exponencial y han migrado de sistemas analógicos a digitales aumentando su capacidad[13].

A continuación, se definen las bases de datos utilizadas para nuestro estudio:

- Conditions_Contributing_to_COVID-19_Deaths__by_State_and_Age__Provisional_2020-2021, del Center for disease control and prevention, de los Estados Unidos.
- Casos positivos de COVID-19 en Colombia, del instituto nacional de salud.

- Defunciones no fetales del 2015-2021pr del DANE
- Defunciones COVID por comorbilidad del instituto nacional de salud.

2.5.1. TASA DE MORTALIDAD

Básicamente la tasa de mortalidad o mortandad es la proporción de decesos registrados respecto al total de individuos pertenecientes a un país, región, en un determinado rango de tiempo, este tipo de dato estadístico, se puede representar porcentualmente dependiendo la necesidad y en otras ocasiones se representa por medio de un K que puede ser 100, 1.000, 100.000 o 1 millón según la escala que permita facilitar su entendimiento, además, también es correcto estimar la cifra numérica de los mismos. La importancia de este cálculo es estimar las causas de las muertes y tomar medidas correctivas o preventivas frente a los eventos[14].

2.5.2. DANE

El Departamento Administrativo Nacional de Estadística (DANE), es la entidad encargada del levantamiento, procesamiento, análisis y divulgación de las estadísticas oficiales de Colombia. El DANE ofrece al no solo a las personas del país y si no al mundo alrededor de 30 investigaciones anualmente de los sectores más relevantes del país como son la economía, industria, población, sector agropecuario y calidad de vida, entre otras[15].

2.5.3. SISPRO

El Sistema Integral de Información de la Protección Social – SISPRO, es una plataforma que integra la información varias instituciones, básicamente es usada para el monitoreo, para la validación de servicios prestados y también sirve de base para la toma de decisiones políticas. El SISPRO es un sistema que integra información de varias fuentes, para la protección social, que buscan garantizar el aseguramiento y la asistencia social[16].

2.5.4. DATOS ABIERTOS

Información dispuesta en una página web para su uso y utilización con licencia abierta y sin restricciones legales, según la ley 1712 del año 2014, sobre la transparencia y Acceso a la información pública nacional, *"todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos"*[17].

2.5.5. INSTITUTO NACIONAL DE SALUD

El Instituto Nacional de Salud (INS), es una entidad nacional que busca contribuir mediante la gestión del conocimiento en la salud pública, el seguimiento al estado de salud de la población y los servicios prestados por la entidad[18].

3. ESTADO DEL ARTE

En el siguiente cuadro se expondrán algunos de los estudios relacionados previos a esta investigación, el ideal es indicar estudios su enfoque e identificar el valor diferencial de este proyecto como se relaciona en la Tabla 1.

TITULO	OBJETIVO Y METODOLOGIA	AÑO
Smart Healthcare for Diabetes During COVID-19[19]	Estudio del riesgo que poseen los pacientes con Diabetes en adquirir el Virus del **SARS-CoV-2**, se analizar tanto los riesgos y las recomendaciones que se dan para manejar los perfiles glucémicos y el impacto que tendría en la reducción de contagios.	2020
Social determinates of health and COVID-19 mortality rates at the county level[20]	Este artículo analiza la importancia de la disponibilidad de resultados del sector salud de forma detallada y precisa, en el condado de evaluación. Explica cómo estos datos pueden ser significativos para los análisis, sin poner en peligro la privacidad del paciente, todo esto basado en modelos de regresión Lineal.	2020
Medicine Allotment for COVID-19 Patients by Statistical Data Analysis[21]	En este proyecto, se evalúa un método de asignación de medicamentos a pacientes diagnosticados con COVID-19, en los cuales se estimen factores de riesgo, como lo son la presión arterial, sufrir de diabetes, alcoholismo y cáncer, el estudio se ejecutó con algoritmos de agrupamiento.	2020
La Salud de las Personas Adultas Mayores durante la Pandemia de COVID-19[22]	Por medio de revisión al sistema de Salud en México, se evaluaron los impactos que la enfermedad COVID-19, tiene sobre pacientes con enfermedades crónicas como lo son Hipertensión, diabetes, cáncer y en personas adultas.	2020
Ultrasound Imaging: A Silent Hero in COVID-19 and Lung Diagnostics[23]	Estudia el impacto y la relevancia que tiene un estudio temprano en la salud de pacientes con y el gran aporte que este estudio genera para reducir procesos más críticos y prevenir complicaciones, mediante desarrollo algorítmico orientados a los ultrasonidos.	2020
Predictions of COVID-19 Infection Serverity Based on Co-associations between the SNPs of Co-morbid Diseases and COVID-19 throungh Machine Learning of Genetic Data[24]	Estudia por medio de *Machine Learning,* las susceptibilidades que presentan pacientes con enfermedades crónicas, como lo son el Cáncer, Diabetes Mellitus, Obesidad, entre otras. Detectar de forma temprana las vulnerabilidades y estima evaluar la efectividad de la predicción.	2020
Comorbidity Impact on COVID-19[25]	Realiza un análisis estadístico de los pacientes con comorbilidad y el impacto que tiene en la atención de para ellos, por requerir tratamiento diferenciado a los demás infectados por el virus en Brasil.	2020
COVID-19 y su relación con poblaciones vulnerables[26]	Encontró la relación entre el contagio con COVID-19 y población más vulnerable socioeconómicamente, mediante revisión bibliográfica de artículos científicos.	2020
Data Associated with Epigenetic Changes Brought by SARS-Cov-2[27]	Se efectúa un estudio del impacto que tiene exacerbar la comorbilidad a través de datos epigenéticos (cambio en los genes).	2020

Tabla 1. Estado del Arte - Fuente Propia

Smart Healthcare for Diabetes During COVID-19. Los resultados obtenidos en este Estudio revelan que las personas diagnosticadas con diabetes tienen un riesgo mayor de 2 o 3 veces que las personas que no lo padecen, adicional que una sana alimentación mitiga la posibilidad de contagio[19], este estudio abarca tanto la mortalidad de los contagios y habitualmente en los pacientes que padecen en todo el territorio colombiano.

Social determinates of health and COVID-19 mortality rates at the county level. Se encontró que no aumento la tasa de mortalidad por causas como obesidad, tabaquismo o diabetes, al contrario, si aumento 7.6% en personas con edad mayor y era directamente dependiente de los confinamientos[20] . Se realiza la evaluación en un segmento distinto en este caso Colombia, con una metodología diferenciada a la de este proyecto, pues hacemos uso de lenguaje no supervisado y supervisado que nos aporte mayor eficacia en los resultados a obtener.

Medicine Allotment for COVID-19 Patients by Statistical Data Analysis. La investigación es basada en algoritmos de Agrupamiento, este procedimiento presenta muchos factores positivos, entre los cuales pueden estar la segmentación de pacientes para la asignación de medicamentos, entre otras virtudes encontradas, es la actualización constante del algoritmo, que permite incluir variables de otras enfermedades y a si calcular la asignación de medicamentos, de acuerdo a su condición e información registrada[21] , la diferencia entre los estudios es que este proyecto se realiza en pacientes diagnosticados y busca evaluar entre muchos factores la mortalidad de este tipo de pacientes, por ende este proyecto puede servir de base para encontrar mecanismos de servicio en los medicamentos y priorizarlos para que la afectación del COVID-19 sea menor.

La Salud de las Personas Adultas Mayores durante la Pandemia de COVID-19. Se encuentra que las personas con mayor factor de riesgo en México son las personas adultas con edades superiores a los 56 años, seguidas por las personas con (Enfermedades Cardiacas, diabetes y Enfermedad respiratoria), adicionalmente se pudo encontrar un factor adicional que es su situación económica y social que posean los pacientes [22]. Este estudio sobre pacientes con comorbilidades contagiados de COVID-19 en Colombia y se enfoca en la letalidad de las personas de todo el país con comorbilidades desde los 0 años en adelante, de todas las regiones del país y de todos los estratos sociales.

Ultrasound Imaging: A Silent Hero in COVID-19 and Lung Diagnostics. Es una investigación en curso que tiene como objetivo promover el uso de robótica y de elementos de inteligencia artificial (IA), para la creación de algoritmos de detección de enfermedades respiratorias crónicas, en exámenes como es el diagnostico de pulmón y los ultrasonidos, los modelos de **Ultrasound Imaging**[23] ; La investigación abarca líneas de prevención y de estudios minuciosos, que son viables para un diagnóstico detallado en pacientes con enfermedades respiratorias, sin embargo este estudio de pacientes con comorbilidades contagiados con COVID-19 en Colombia, se enfoca en todas las enfermedades no transmisibles y diagnosticadas como comorbilidad en la letalidad de las personas que padecen estas enfermedades y fueron contagiadas por el virus, y así determinar cuál es el impacto que al mes de Enero 2021 ha tenido la emergencia sanitaria en nuestro país.

Predictions of COVID-19 Infection Serverity Based on Co-associations between the SNPs of Co-morbid Diseases and COVID-19 througn Machine Learning of Genetic Data. Esta predicción se obtuvo como resultado de incluir las patologías y

sintomatologías de los pacientes, los resultados logrados con las correlaciones son altamente confiables y podrían estimar el margen de contagio que se tiene según cuadro clínico en cada individuo[24]. Los modelos se realizan respecto a los cuadros o las historias clínicas, de un grupo de pacientes, sin embargo, este proyecto de pacientes con comorbilidades contagiados con COVID-19 en Colombia, no estima abarcar ese nivel de detalle, sino estudiar de manera general por Región, edad, sexo, entre otros factores. Las posibles causas que el virus tuvo en pacientes con comorbilidades contagiados por el virus, si se vieron alteradas sus cifras históricas de muertes y que otros indicadores en la actualidad se ven perjudicados.

Comorbidity Impact on COVID-19. En este proyecto se evidencia en sus resultados, que los pacientes con comorbilidades representan un mayor riesgo de contagio para el COVID-19, se estudiaron múltiples factores, como los son algunos de los componentes en los medicamentos que se suministran en tratamientos de enfermedades crónicas y medicamentos de control en edades avanzadas, con el fin de determinar si estos se atribuyen un factor de riesgo adicional al ya existente por su prescripción o si por el contrario al ser consumidos reducen el riesgo de contagio[25]. Este estudio de *Comorbidity Impact on COVID-19*, se realiza bajo condiciones médicas de pacientes con comorbilidad, es muy similar a este estudio, sin embargo, este proyecto de pacientes con comorbilidades contagiados con COVID-19 en Colombia, se enfoca en la tasa de mortalidad y en el impacto socioeconómico de los pacientes con COMORBILIDADES infectados por el virus, no en las prescripciones de los medicamentos y sus secuelas.

COVID-19 y su relación con poblaciones vulnerables. El proyecto concluye que la pandemia es una problemática social, que refleja condiciones diferenciadas entre la población, que hace que las personas con bajos recursos y que padecen enfermedades crónicas tienden a tener mayor vulnerabilidad que personas con condiciones económicas poco superiores, por lo cual propone plantear un sistema de salud más incluyente y menos selectivo a la hora de los tratamientos [26] . La mayor discrepancia entre los proyectos es que realmente se estima cuantificar el impacto por Sexo, edad y ubicación, entre otros. Y a su vez realizar una evaluación más detallada del impacto social de patrones de datos en pacientes con COMORBILIDADES infectados por COVID-19.

Data Associated with Epigenetic Changes Brought by SARS-Cov-2, Tras toda la ejecución del proyecto se concluye que los pacientes con dos afecciones simultáneas, es decir, alguna comorbilidad y el Virus del COVID-19. Son más propensos a padecer complicaciones dentro del proceso de la incubación del virus, lo que hace que se proponga dentro como recomendación, incluir a pacientes con estas enfermedades crónicas, en plan de seguimiento preventivo y así reducir la cantidad de muertes por la adquisición del virus[27]. Uno de los diferenciadores del estudio, es que no está orientado a un nivel genético, sin embargo, este tipo de estudios aportan de forma médica, los estudios estadísticos y generar valor en la consecución de patologías que deban estimarse para la ejecución de este proyecto.

Podemos verificar que el problema de pacientes que padecen COMORBILIDADES es un problema muy directo con el virus del COVID-19, su aparición no solo afecta su atención regular, si no que según estudios agudiza e incrementa la vulnerabilidad a la adquisición del virus, y no solo eso, sino que a su vez acelera en ocasiones la sintomatología y letalidad de dichos pacientes.

4. PLANTEAMIENTO DEL PROBLEMA

Una de las mayores problemáticas actuales en Colombia, es el sector salud no solo por el déficit que se presenta de personal de la salud, su alta demanda de usuarios, el poco tiempo de atención con el que cuentan los profesionales de la salud, para evaluar el estado real de los pacientes, la infraestructura deficiente, la alineación de subredes, si no que a todo lo anterior se suma la crisis sanitaria generada por la pandemia del **COVID-19,** situación que al 30 de Abril de 2021, lo posiciona en el ranking mundial de contagios de Número **12** y a nivel de Latinoamérica en el tercer lugar[28].

Si bien es un problema de carácter mundial, este ha centrado la atención tanto médicos generales, como de especialistas en la vigilancia de pacientes que resultan infectados por el virus. Se estima que, de una manera directa, influye en la atención, programación de los profesionales y tener menor capacidad instalada, para la atención enfermedades crónicas clasificada como **COMORBILIDADES** entre las que encontramos la Diabetes Mellitus, la Obesidad, enfermedades respiratorias, entre otras. Enfermedades que tiene una tasa de mortalidad significativa a nivel mundial y la cual viene por años aumentando de forma exponencial[29].

Por esta causa, es importante estudiar cual es el impacto que estas personas han sufrido desde el inicio de la pandemia en el país, y evidenciar si es percepción de decadencia de los servicios o si en realidad la pandemia ha desmejorado las condiciones clínicas de mencionados pacientes aumentando su probabilidad de contagio y por ende acrecentando la letalidad[12].

Otra de las falencias detectadas son los efectos colaterales en la salud y en los tratamientos e incapacidades, que se requieren para mantener las enfermedades crónicas como una alerta latente, siendo las que ocasionan la mayor cantidad de muertes a nivel mundial y Colombia no es la excepción, y quien lo afirma es la **OPS** *" Los servicios de prevención y tratamiento de las enfermedades no transmisibles (ENT) se han visto gravemente afectados desde el comienzo de la pandemia de COVID-19 en la región de las Américas, según una encuesta de la Organización Panamericana de la Salud/ Organización Mundial de la Salud (OPS/OMS)"*[30]*.*

En este proyecto se realizó una evaluación de Modelos de datos, de la población que padece **COMORBILIDADES y fueron** contagiados con **COVID-19**, según edad, sexo, región, y demás factores que se encuentren en unos datos, difícilmente localizables y con complejidad de integración, debido a la falta de unificación en las bases de datos a nivel nacional.

5. OBJETIVOS

El proyecto de investigación cuenta con tres objetivos específicos, que permitirán la consecución del objetivo general y la medición de metas parciales basada en la metodología **CRISP-DM,** para así lograr la obtención de los resultados esperados. A continuación, se enuncian cada uno de ellos:

5.1. OBJETIVO GENERAL:

Generar un Modelo de datos de diferentes fuentes de Información para la evaluación de la mortalidad de Pacientes con comorbilidades CONTAGIADOS DE COVID-19 en COLOMBIA.

5.2. OBJETIVOS ESPECIFICOS:

- Identificar las bases de datos pertinentes para la creación de Modelos de datos.
- Probar diferentes modelos de Aprendizaje automático no supervisado y supervisado que permita encontrar patrones de comportamiento del virus estimando diferentes variables.
- Establecer una predicción del comportamiento de las muertes ocasionadas por el **COVID-19** en la salud de personas con comorbilidades**.**

6. METODOLOGIA

Dentro del progreso del proyecto es sustancial a la definición de actividades, a su vez se especifica el cómo se irán ejecutando cada una de ellas, para el este proyecto se estimaron 6 etapas con la Metodología **CRISP-DM**[31]**,** detalladas a continuación:

6.1. FASE DE ANALISIS

En esta etapa del proyecto se realizó el análisis de los posibles portales que tuvieran información pertinente, veraz y a su vez a la fecha corte requerida para este estudio, la cual se determinó que era hasta **enero 2021**, esta fecha se estima porque desde el mes de **febrero** se inició el plan de Vacunación y esto afectaría los indicadores y estudios a realizar, pues los comportamientos no serían bajo las mismas condiciones de los pacientes, encontrando las siguientes fuentes:

* Datos.gov.co[3]
* data.cdc.gov[4]
* Dane.gov.co[5]
* Ins.gov.co[6]
* rodillo.org[7]

6.2. ANALISIS INICIAL

Para realizar los análisis iniciales, partimos de los datos de las estadísticas vitales del de Fallecidos no Fetales de los años 2018-2021 del DANE.
La cual se puede obtener en la página www.dane.gov.co [8]:

realizando la comparación contra los datos de los fallecidos por covid-19 con sus comorbilidades del Instituto Nacional de Salud, el cual se puede visualizar en la página https://www.ins.gov.co/[9]:

6.2.1. ANALISIS COMPARATIVO DE LAS COMORBILIDADES

Iniciamos con la comparación de las comorbilidades que está estudiando la INS, comparando su comportamiento normal, de acuerdo con las estadísticas vitales de los fallecidos no fetales del DANE, con la de aquellos pacientes que han sido declarados muertos por covid-19 y en su historia médica, se presenta la relación con alguna de las comorbilidades en la Grafica 1.

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

[4] https://data.cdc.gov/NCHS/Conditions-Contributing-to-COVID-19-Deaths-by-Stat/hk9y-quqm

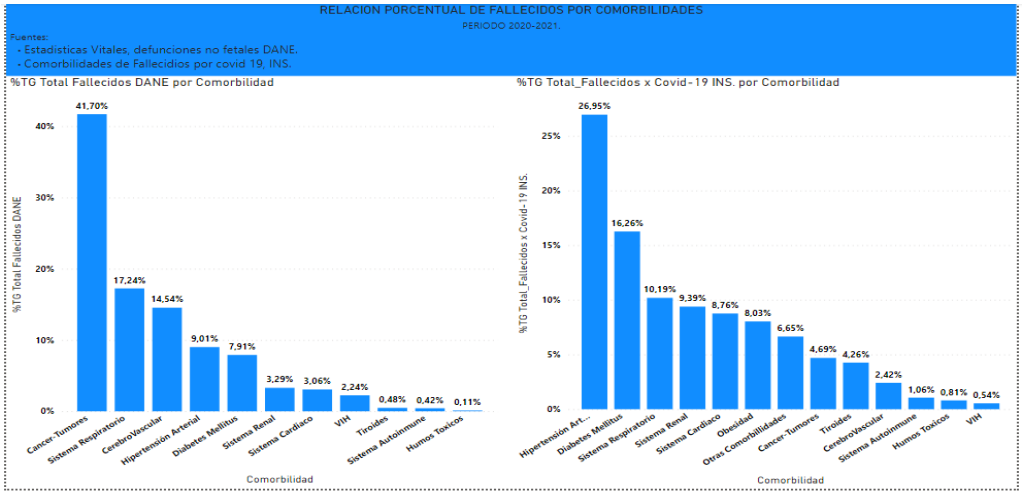
[5] https://www.dane.gov.co/

[6] https://www.ins.gov.co/Noticias/Paginas/Coronaviruss.aspx

[7] https://rodillo.org/nuestra-data/

[8] https://www.dane.gov.co/index.php/estadisticas-por-tema/salud/nacimientos-y-defunciones/defunciones-no-fetales

[9] https://www.ins.gov.co/Noticias/Paginas/Coronaviruss.aspx

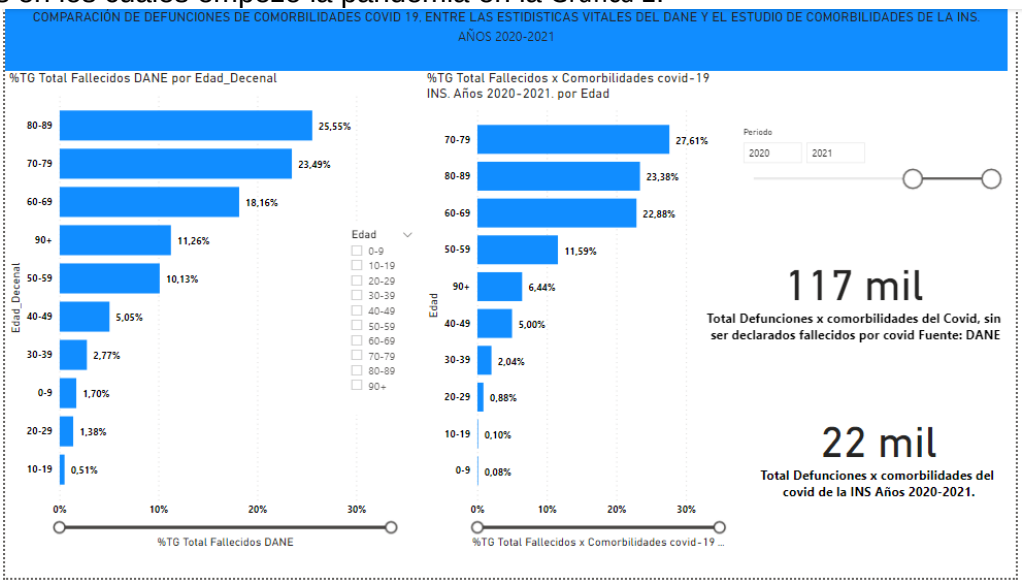


Grafica 1. Relación porcentual comparativa por comorbilidades, modelado en Power BI, entre los fallecidos por comorbilidades como causa de muerte, y los fallecidos por covid-19 con comorbilidades, en el periodo 2020-2021.

Como podemos apreciar en el gráfico de las estadísticas vitales del DANE, en términos generales, entre las que de aquí en adelante se llamarán “las comorbilidades del covid-19”, normalmente son los pacientes con canceres y tumores, seguidos de aquellos con problemas del sistema respiratorio los que más fallecen, mientras que las personas que más fallecen por el COVID-19 son aquellos con problemas de Hipertensión Arterial y Diabetes Mellitus, seguido de los pacientes con problemas en el sistema respiratorio, por lo que la distribución de fallecidos por COVID-19, presenta ciertas diferencias porcentuales, frente aquellos que se infectan y logran sobrevivir al virus.

6.2.2. ANALISIS COMPARATIVO DE FALLECIDOS POR EDAD

Otra de las comparaciones necesarias es la edad de fallecidos por las comorbilidades, esto con el objetivo de saber cuáles son los rangos de edad, en las que más fallecen los pacientes, tanto con covid-19, como sin el virus, en los años 2020-2021, que son los años en los cuales empezó la pandemia en la Grafica 2.



Grafica 2. Relación porcentual comparativa por edades decenal, modelado en Power BI, entre los fallecidos por comorbilidades como causa de muerte, y los fallecidos por covid-19 con comorbilidades, en el periodo 2020-2021. fuentes: DANE e INS.

Como podemos apreciar en ambas gráficas, son las personas de la tercera edad las que más fallecen, así no tengan COVID-19, por lo que tiene lógica, tener un mayor cuidado de las personas de la tercera edad, pues el virus en interacción con sus comorbilidades puede reducir su expectativa de vida.

6.3. CONVERSIÓN DE LOS DATOS EN VARIABLES, DE LOS CASOS POSITIVOS COVID-19 EN COLOMBIA.

Dado que las estadísticas vitales de los fallecidos no fetales del DANE, están diseñadas para generar datos históricos sobre las enfermedades de las que fallecen los colombianos, y al ser el COVID-19 una pandemia reciente, no se encuentra incluida en sus registros, por esta razón nos remitiremos al *dataset* de casos positivos COVID-19, que se encuentra en el portal de datos abiertos, en este caso dado que es un *dataset* extremadamente grande y pesado, lo descargamos filtrado, entre los ID del caso, desde el 1 hasta el 1.719.771, que son los casos reportados hasta el 7 de enero de 2021, fecha en la cual ya se ha de terminar de incluir toda la información de los casos del año 2020.

6.3.1. PREPROCESAMIENTO DE LOS DATOS

6.3.1.1. ANALISIS DE VARIABLES.

Al igual que cuando vemos una ecuación nos preguntamos qué tipo y cuantas variables contiene, en la programación también es indispensable, analizar el *dataset* inicial, cuantas variables, de qué tamaño y de qué tipo son como se evidencian en Tabla 2.

fecha reporte web	object
ID de caso	object
Fecha de notificación	object
Código DIVIPOLA departamento	object
Nombre departamento	object
Código DIVIPOLA municipio	object
Nombre municipio	object
Edad	int32
Unidad de medida de edad	object
Sexo	object
Tipo de contagio	object
Ubicación del caso	object
Estado	object
Código ISO del país	object
Nombre del país	object
Recuperado	object
Fecha de inicio de síntomas	object
Fecha de muerte	object
Fecha de diagnóstico	object
Fecha de recuperación	object
Tipo de recuperación	object
Pertenencia étnica	object
Nombre del grupo étnico	object
dtype: object	

Tabla 2. Muestra set de datos. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Esto con la intención de saber cómo las podemos trabajar, y si por algún motivo el programa no las está leyendo como esperamos poder indicarle de que tipo son, tal cual como nos sucedió con el *dataset* de casos positivos de **COVID-19** en Colombia Imagen 1

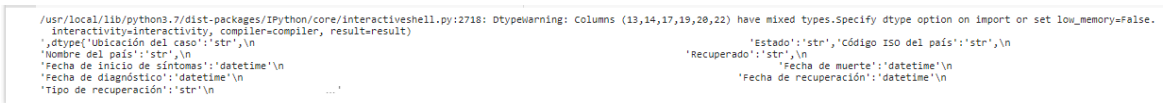


Imagen 1. Validación de tipo de Dato. Fuente: Imagen generada desde Python tomado el set de datos descargado de INS

Puesto que el *dataset* proviene de cerca de los 1102 municipios del país cada uno con sus propias fuentes de información, como lo son EPS, IPS y entidades municipales, con lo que el sistema manifiesta que se presentan diferentes tipos de configuración de variables en una misma columna, es decir algunos datos vienen de tipo numérico y otros de tipo texto, por eso lo que inicialmente debemos entrar a definirle el tipo de variables por columna.

	fecha reporte web	ID de caso	Fecha de notificación	Código DIVIPOLA departamento	Nombre departamento	Código DIVIPOLA municipio	Nombre municipio	Edad	Unidad de medida de edad	Sexo	...	Código ISO del país	Nombre del país	Recuperado	Fecha de inicio de síntomas	Fecha de muerte	Fecha de diagnóstico	Fecha de recuperación	recup
0	6/3/2020 0:00:00	1	2/3/2020 0:00:00	11	BOGOTA	11,001	BOGOTA	19	1	F	...	380	ITALIA	Recuperado	27/2/2020 0:00:00	NaN	6/3/2020 0:00:00	13/3/2020 0:00:00	
1	9/3/2020 0:00:00	2	6/3/2020 0:00:00	76	VALLE	76,111	BUGA	34	1	M	...	724	ESPAÑA	Recuperado	4/3/2020 0:00:00	NaN	9/3/2020 0:00:00	19/3/2020 0:00:00	
2	9/3/2020 0:00:00	3	7/3/2020 0:00:00	5	ANTIOQUIA	5,001	MEDELLIN	50	1	F	...	724	ESPAÑA	Recuperado	29/2/2020 0:00:00	NaN	9/3/2020 0:00:00	15/3/2020 0:00:00	
3	11/3/2020 0:00:00	4	9/3/2020 0:00:00	5	ANTIOQUIA	5,001	MEDELLIN	55	1	M	...	NaN	NaN	Recuperado	6/3/2020 0:00:00	NaN	11/3/2020 0:00:00	26/3/2020 0:00:00	
4	11/3/2020 0:00:00	5	9/3/2020 0:00:00	5	ANTIOQUIA	5,001	MEDELLIN	25	1	M	...	NaN	NaN	Recuperado	8/3/2020 0:00:00	NaN	11/3/2020 0:00:00	23/3/2020 0:00:00	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1719766	6/1/2021 0:00:00	1,719,807	23/12/2020 0:00:00	25	CUNDINAMARCA	25,126	CAJICA	19	1	M	...	NaN	NaN	Activo	19/12/2020 0:00:00	NaN	3/1/2021 0:00:00	NaN	
1719767	6/1/2021 0:00:00	1,719,808	22/12/2020 0:00:00	25	CUNDINAMARCA	25,175	CHIA	19	1	M	...	NaN	NaN	Activo	18/12/2020 0:00:00	NaN	2/1/2021 0:00:00	NaN	
1719768	6/1/2021 0:00:00	1,719,809	22/12/2020 0:00:00	25	CUNDINAMARCA	25,175	CHIA	21	1	M	...	NaN	NaN	Activo	18/12/2020 0:00:00	NaN	2/1/2021 0:00:00	NaN	
1719769	6/1/2021 0:00:00	1,719,810	20/12/2020 0:00:00	85	CASANARE	85,001	YOPAL	19	1	M	...	NaN	NaN	Activo	14/12/2020 0:00:00	NaN	31/12/2020 0:00:00	NaN	
1719770	6/1/2021 0:00:00	1,719,811	22/12/2020 0:00:00	25	CUNDINAMARCA	25,175	CHIA	19	1	M	...	NaN	NaN	Activo	18/12/2020 0:00:00	NaN	2/1/2021 0:00:00	NaN	

1719771 rows × 23 columns

Tabla 3. Set de datos total. Fuente: Fuente: Imagen generada desde Python tomando el set de datos descargado de INS

Como podemos visualizar en Tabla 3, el programa nos dice que el *dataset* está compuesto de 23 columnas y 1719771 filas, por lo que no podemos abrirlo en cualquier tipo de programa, por eso lo recomendable es mantenerlo en formato .csv, y abrirlo con programas especiales para la lectura. Otra parte importante es visualizar las variables de tipo texto, para poder conocer la cantidad de muestras que contiene la Tabla 4.

0	BOGOTA
1	VALLE
2	ANTIOQUIA
6	CARTAGENA
11	HUILA
14	META
27	RISARALDA
29	NORTE SANTANDER
30	CALDAS
44	CUNDINAMARCA
58	BARRANQUILLA
59	SANTANDER
75	QUINDIO
87	TOLIMA
120	CAUCA
154	STA MARTA D.E.
212	CESAR
221	SAN ANDRES
228	CASANARE
313	NARIÑO
420	ATLANTICO
444	BOYACA
546	CORDOBA
605	BOLIVAR
668	SUCRE

show more (open the raw output data in a text editor) .

11796	ARAUCA
12265	VAUPES
25313	GUAINIA
25659	VICHADA
18038	GUAVIARE

Tabla 4. Muestra set de datos tipo Texto. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

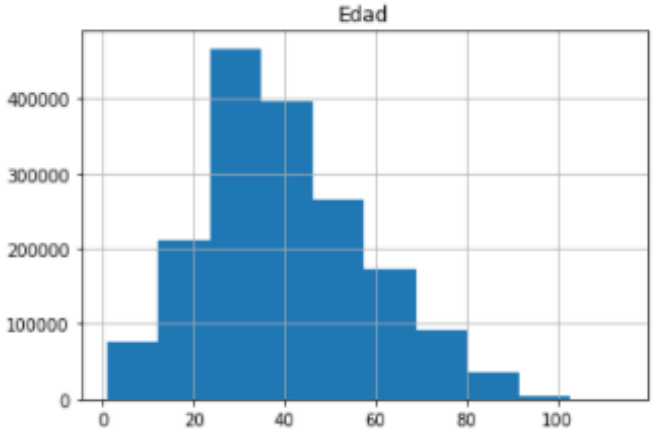
En el caso de encontrar variables con *outliers*, filas en blanco o inconsistentes, tenerlas presente al momento de realizar la limpieza.

```
0         Recuperado
151        Fallecido
790              NaN
12441      fallecido
186067        Activo
Name: Recuperado, dtype: object
```

Tabla 5. Muestra Datos con inconsistencias del dataset. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

En este caso Tabla 5, la columna recuperados nos muestra 2 problemas, tiene valores nulos o en blanco, que son los que muestra como NaN, y esta una misma palabra está escrita de dos formas “Fallecido” y “fallecido”, esto en el momento de realizar algún trabajo, el programa no los tomara como una sola variable.

También es importante analizar las variables que la herramienta nos muestra como numéricas, esto con la intención de saber qué cantidades son y sus rangos, como en este caso la variable edad que es de tipo numérico.



Grafica 3. Histograma de Edades. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

En este caso el histograma Grafica 3, nos muestra que su rango es de 0 a un poco más de 100 años y que hay alrededor de 450 muestras de las edades entre 20 y 40 años.

6.3.1.2. Agrupamiento de Conjuntos

Es indispensable que antes de trabajar nos preguntemos ¿Que Variables deseamos evaluar?, ¿Cuáles nos sirven de índice? ¿Cuáles de pivote? Debemos tener presente que una sola columna no nos creara la información que necesitamos por si sola, por lo que es indispensable saber con qué otras las agruparemos, en un nuevo set de datos o *dataframe*.

Para este primer caso, utilizaremos esas variables que generalmente son utilizadas en los estudios poblacionales, como lo son la edad y el Sexo, para saber cuántos se han recuperado, cuantos han fallecido y cuantos siguen como casos activos, provenientes de la columna Recuperado, con lo que estaremos agrupando los conjuntos por Sexo y edad.

	Sexo	Edad	Recuperado
0	F	19	Recuperado
1	M	34	Recuperado
2	F	50	Recuperado
3	M	55	Recuperado
4	M	25	Recuperado
...	...	...	...
1719766	M	19	Activo
1719767	M	19	Activo
1719768	M	21	Activo
1719769	M	19	Activo
1719770	M	19	Activo

1715215 rows × 3 columns

Tabla 6. Muestra set de datos por Sexo y Edad. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Como podemos observar en la Tabla 6, este nuevo *dataframe* que hemos creado con las 3 variables, y ha sido reducido de 1. 719. 771 datos a 1. 715. 215, por lo que, al crearlo, le hemos eliminado algunas de las inconsistencias del set de datos inicial.

6.4. PROCESAMIENTO DE DATOS

Otra de las observaciones que pudimos realizar en este nuevo *dataframe* del preprocesamiento, es que solamente contamos con una variable numérica, que va desde 0 hasta un poco más de 100 años, por lo que es muy larga y difícil de agrupar para el análisis, igualmente con las variables categóricas como lo son el sexo y el estado del paciente, por lo que es necesario ajustarlas.

6.4.1. Agrupamiento interno de variable

Debido a lo mencionado en el punto anterior, lo que procederemos a realizar es convertir las edades en conjuntos de 10 años, para poder reducir su tamaño, en una nueva columna.

6.4.2. Binarización de variables categóricas

En las variables categóricas, también es necesario realizar la conversión a numéricas, para poder contabilizar cada una de sus variables, es decir procederemos a crear una nueva columna por cada una de las categorías que se encuentran dentro de la columna, por lo tanto crearemos una nueva para el Sexo femenino, otra para el Sexo masculino, e igualmente haremos para las 3 variables de la columna Recuperado, teniendo presente que debemos agrupar en una sola las que vienen escritas como “Fallecido” y “fallecido” debido a que el lenguaje las tomara como datos independientes, estas nuevas columnas por medio de un bucle o ciclo repetitivo, las poblaremos con 1 si pertenece a ese conjunto y 0 en caso contrario Tabla 7.

	Rango_Edad	Edad	Sexo	Femenino	Masculino	Recuperado	Alentados	Activos	Fallecidos
0	11-19	19	F	1	0	Recuperado	1	0	0
1	30-39	34	M	0	1	Recuperado	1	0	0
2	50-59	50	F	1	0	Recuperado	1	0	0
3	50-59	55	M	0	1	Recuperado	1	0	0
4	20-29	25	M	0	1	Recuperado	1	0	0
...	...	...	...	...	...	...	...	...	...
1719766	11-19	19	M	0	1	Activo	0	1	0
1719767	11-19	19	M	0	1	Activo	0	1	0
1719768	20-29	21	M	0	1	Activo	0	1	0
1719769	11-19	19	M	0	1	Activo	0	1	0
1719770	11-19	19	M	0	1	Activo	0	1	0

1715215 rows × 9 columns

Tabla 7. Muestra set de datos Agrupados por Sexo y Estado. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Podemos hacer una prueba rápida y detallamos que la edad del paciente hace parte de la columna Rango Edad, por lo que nos quedó bien agrupado el rango de la edad, en el caso de la columna sexo, vemos que tenemos en la columna Femenino, 1 si es F y 0 si no lo es, igualmente con la Masculino, también ocurre con la columna Recuperado.

Posiblemente a este nuevo *dataframe* no le encontramos razón de existir, sin embargo, es indispensable para la siguiente parte.

6.4.3. Agrupamiento de datos

Para lograr agrupar el anterior dataframe, primero debemos crear otro nuevo y tener presente, la columna con la cual deseamos agrupar para convertirla en la columna índice, y cuales datos deseamos tener como pivote, con los cuales podemos realizar diferentes operaciones como sumarlos, obtener su media, su máximo, su mínimo, entre otras, como se muestra en la Tabla 8.

	Rango_Edad	Edad_Prom	Total	Femenino	Masculino	Alentados	Activos	Fallecidos
0	0-9	5.0	53057	25894	27163	49760	3235	62
1	10-19	15.0	111928	56094	55834	105716	6147	65
2	20-29	25.0	378716	197833	180883	359348	18890	478
3	30-39	34.0	399541	199082	200459	377593	20800	1148
4	40-49	44.0	280253	141831	138422	261773	15943	2537
5	50-59	54.0	231635	119404	112231	210827	15153	5655
6	60-69	64.0	141164	71105	70059	119992	10886	10286
7	70-79	74.0	74353	36510	37843	56182	6261	11910
8	80-89	84.0	36860	18970	17890	23920	3076	9864
9	90+	93.0	7708	4326	3382	4467	523	2718

Tabla 8. Set de datos Agrupados por Rango de Edad, Sexo y Estado del paciente. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS

En este caso definimos como la columna índice el rango de edad y las de pivote las otras numéricas y también hemos reducido 1. 715. 215 filas en tan solo 10 ver Tabla 8.

6.4 Analítica Descriptiva del contagio DEL COVID-19 EN COLOMBIA

Con el *dataset* anteriormente generado Tabla 8, podemos realizar las estadísticas básicas para los modelos propuestos como lo son *Regresión lineal* y *Pronósticos*, sin embargo, le realizaremos algunos ajustes para poder visualizarla y lograr trabajarla como una tabla plana.

En la cual podemos visualizar en los diferentes rangos de edad, el total de contagiados, cuántos de esos son de Sexo femenino, cuantos son de Sexo masculino, cuantos de ese total de contagiados con respecto a su edad se han alentado, cuantos siguen siendo casos activos y cuantos han fallecido.

Y es precisamente aquí donde podemos iniciar a ejecutar, los diferentes análisis simples que mencionaron inicialmente, en este caso, analizar el impacto del virus **COVID-19,** en las personas en los rangos de edad media y las cuales reportan ser las de mayor de contagio, además que a medida que aumenta su rango de edad también aumenta su probabilidad de fallecer, por lo que la población de mayor edad es la más vulnerable a fallecer.

Sin embargo, podemos estandarizar porcentualmente para poder visualizar mejor los datos, esto lo hacemos por medio de las mismas instrucciones que hemos realizado anteriormente, pero ahora cambiando la ecuación.

Con lo cual obtenemos el siguiente output, otro *dataframe* con nuevas columnas, en el cual le borramos los datos atípicos, y generamos una medición porcentual por cada uno de los grupos

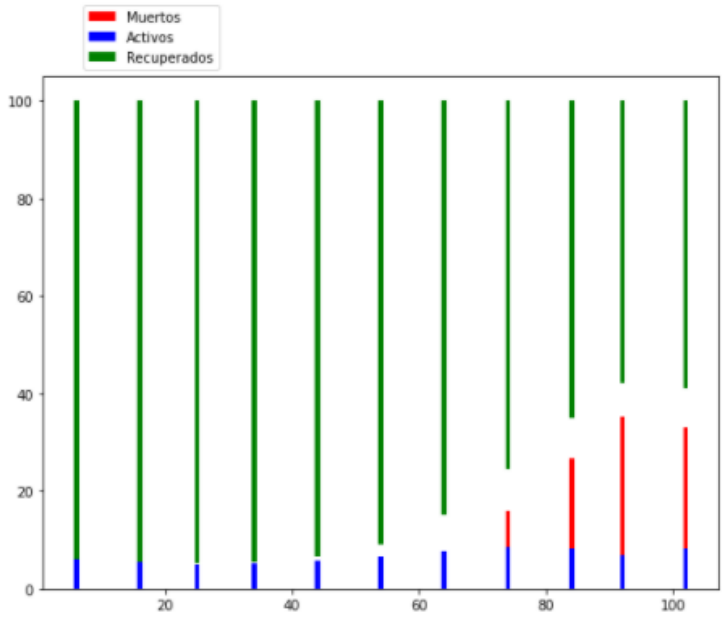
	Rango_Edad	Edad_Prom	Total	Femenino	Masculino	Alentados	Activos	Fallecidos	Tasa_%-Recuperados	Tasa_%-Activos	Tasa_%-Decesos
0	0-9	5.0	53057	25894	27163	49760	3235	62	93.79	6.10	0.12
1	10-19	15.0	111928	56094	55834	105716	6147	65	94.45	5.49	0.06
2	20-29	25.0	378716	197833	180883	359348	18890	478	94.89	4.99	0.13
3	30-39	34.0	399541	199082	200459	377593	20800	1148	94.51	5.21	0.29
4	40-49	44.0	280253	141831	138422	261773	15943	2537	93.41	5.69	0.91
5	50-59	54.0	231635	119404	112231	210827	15153	5655	91.02	6.54	2.44
6	60-69	64.0	141164	71105	70059	119992	10886	10286	85.00	7.71	7.29
7	70-79	74.0	74353	36510	37843	56182	6261	11910	75.56	8.42	16.02
8	80-89	84.0	36860	18970	17890	23920	3076	9864	64.89	8.35	26.76
9	90+	93.0	7708	4326	3382	4467	523	2718	57.95	6.79	35.26

Tabla 9. Estandarización Porcentual de datos Agrupados. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

El que nos evidencia la muestra Tabla 9, cual es la edad media máxima de cada rango de edad, el total de casos, cuantas son mujeres, cuantos hombres, cuantos, alentados, cuantos activos, cuantos, fallecidos, y las tasas porcentuales por cada edad.

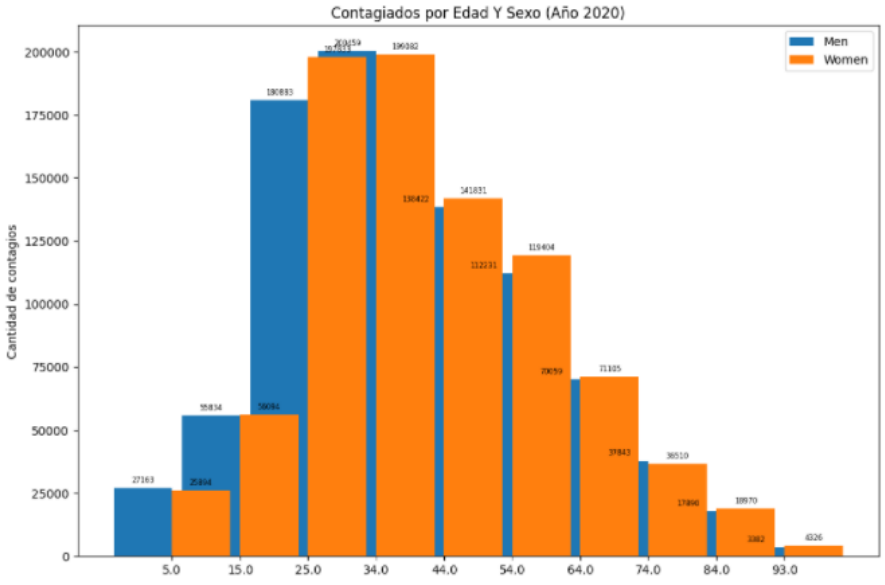
Sin embargo, también podemos realizar diferentes graficas en las cuales podamos relacionar las variables, para este caso las tasas porcentuales por edad.

Con lo que obtenemos la siguiente Grafica 4, que nos relaciona porcentualmente la probabilidad de supervivencia o de muerte con el virus, dependiendo la edad.



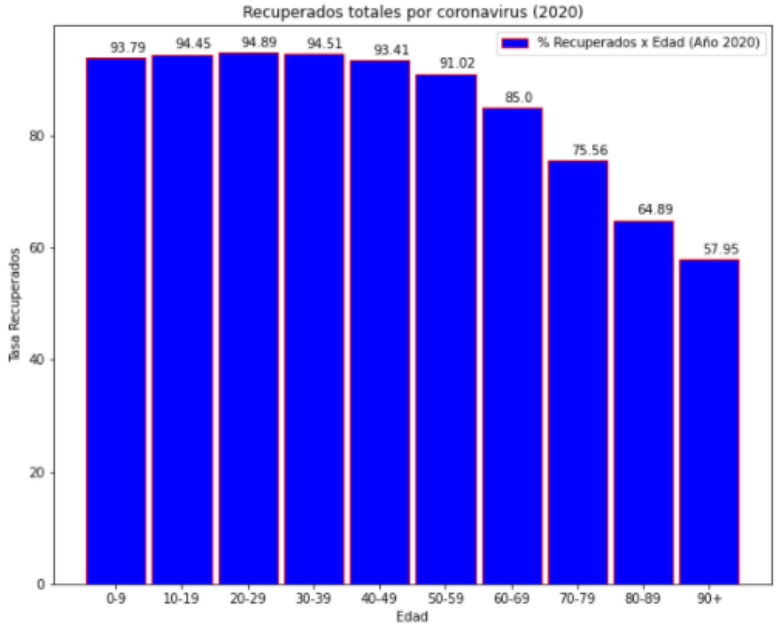
Grafica 4. Relaciona porcentual de probabilidad de supervivencia o de muerte con el virus. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Sin embargo, es mucha la información que podemos obtener de un mismo dataset, y también su visualización como en la siguiente grafica que veremos la cantidad de contagios por edad, la cual está agrupada por el promedio de la edad de cada grupo, es decir, el promedio de la edad de los que están entre 0 y 10 años son 6 años y así sucesivamente, esto porque el programa en el cual se generan los datos no recibe datos categóricos para realizar los *plots* en la Grafica 5.

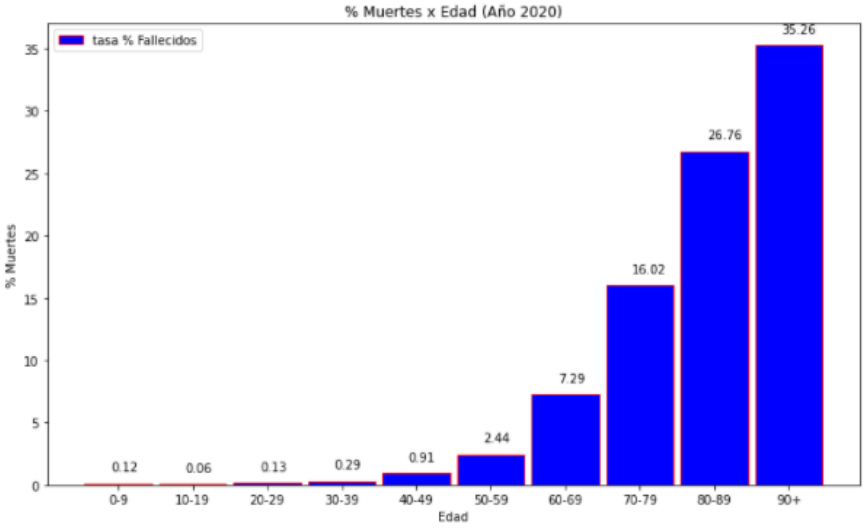


Grafica 5. Contagios por Edad. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS

Es importante que las gráficas que generemos sean lo más específicas posible en este caso, queremos visualizar las diferentes variables, como la de los que se alientan y fallecen con respecto a su edad en las Grafica 6 y Grafica 7 respectivamente.



Grafica 6. Recuperados por Edad. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.



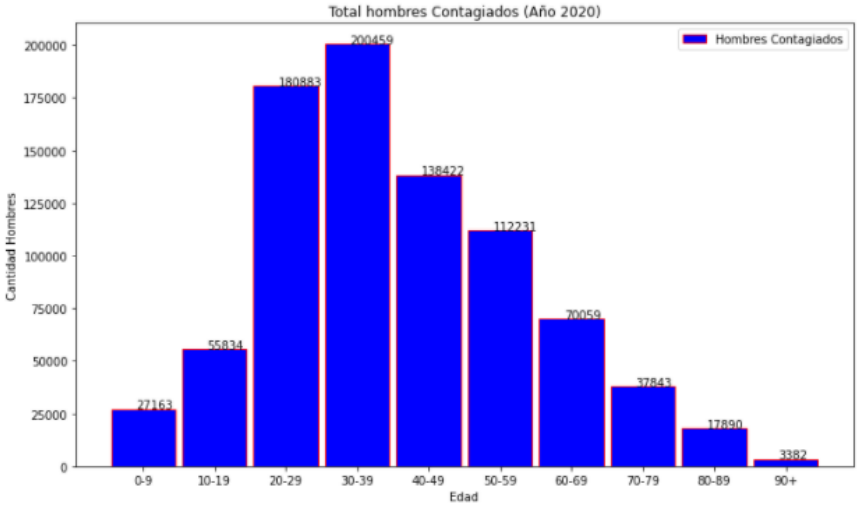
Grafica 7. Porcentaje de Muertes por Edad. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Igualmente, como la primera parte del estudio de virología podemos hacer el estudio de la cantidad de contagiados por Sexo, como ya lo hicimos por la edad, generando gráficos de barras en los que nos muestra la edad promedio de cada Sexo, para ubicarnos en la Grafica 5, como variable numérica y obtener resultados de otras variables, como el número de contagios por edad en cada uno de los dos Sexos, en este caso podemos ver son los hombres entre los 20 y los 50 los que más se contagian

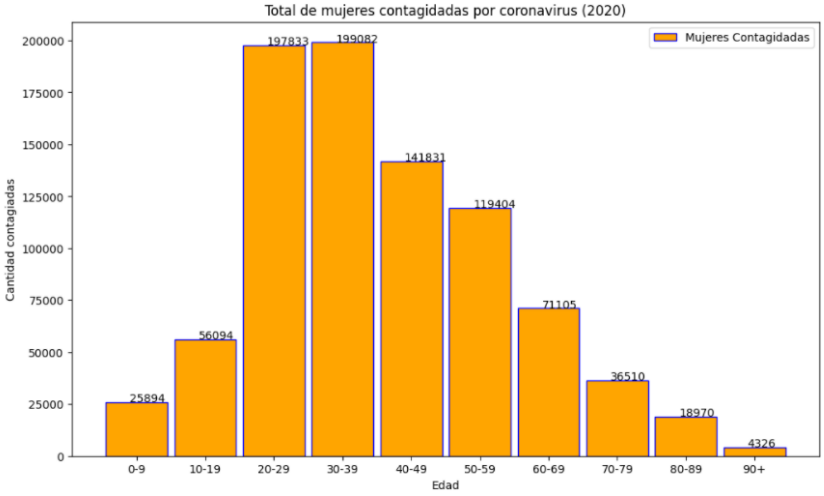
del virus, también sucede lo mismo con las mujeres, por lo que el Sexo no es el tipo de variable que no aporte para realizar clasificaciones que generen regresiones lineales, que puedan ser regresivas en el momento de convertirlas en variables de estudio.

Debido a la incidencia del virus que va más allá del Sexo, es decir pese a que en términos generales sean más las mujeres contagiadas, o más los hombre de la edad media, el Sexo no es una variable que nos ayude a definir la probabilidad de muerte, como si lo es en otras enfermedades tales como la hemofilia, que solamente ataca a los hombres y que puede ser transmitida por la madre, sin ella enfermar de esta morbilidad, o las enfermedades comunes que se presentan de acuerdo al Sexo debido a sus cromosomas o sistemas reproductivos.

Sin embargo, como parte de la investigación incluiremos el Sexo en la investigación de relaciones en el *Statsmodel*, para conocer su grado de correlación entre las variables dependientes e independientes como lo demuestran las Grafica 8 y Grafica 9.



Grafica 8. Total Hombres Contagiados. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.



Grafica 9. Total Mujeres Contagiadas. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Otra de las variables que necesitamos estudiar como parte de la difusión del virus, es el tipo de contagio, y su incidencia en la afectación de la salud del paciente. Por lo que repetiremos los mismos pasos anteriores y generaremos un nuevo *dataframe* en el cual estudiaremos las variables que hacen parte del tipo de contagio.

```
Tipo de contagio
EN ESTUDIO          484
En Estudio       503299
En estudio     62624009
Importado         92619
RELACIONADO        1830
Relacionado     5530207
Name: Edad, dtype: int32
```

Tabla 10. Tipo de Contagios en el dataset. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS

Tal como podemos visualizar en la Tabla 10, la columna tipo de contagio presenta múltiples formas de una misma variable, por lo que aquí incluiremos un paso adicional en la cual, con una condicional, agrupamos las diferentes formas en las que está escrita la misma variable, esto con el fin de consolidar toda la información en una sola.

Y con esto ahora si podemos visualizar como quedo nuestro *dataframe* antes de consolidar los datos, mostrándonos que los importados, los relacionados y los que están en estudio si quedaron en su columna respectiva Tabla 11.

	Rango_Edad	Edad	Sexo	Femenino	Masculino	Tipo de contagio	En_Estudio	Importados	Relacionados
0	11-19	19	F	1	0	Importado	0	1	0
1	30-39	34	M	0	1	Importado	0	1	0
2	50-59	50	F	1	0	Importado	0	1	0
3	50-59	55	M	0	1	Relacionado	0	0	1
4	20-29	25	M	0	1	Relacionado	0	0	1
...	...	...	...	...	...	...	...	...	...
1719766	11-19	19	M	0	1	En estudio	1	0	0
1719767	11-19	19	M	0	1	En estudio	1	0	0
1719768	20-29	21	M	0	1	En estudio	1	0	0
1719769	11-19	19	M	0	1	En estudio	1	0	0
1719770	11-19	19	M	0	1	En estudio	1	0	0

1719771 rows × 9 columns

Tabla 11. Clasificación Tipo de Contagio en el dataset. Fuente: Tipo de Contagios en el dataset. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS

Sin embargo, nos sigue mostrando que son 1719771 filas, alrededor de 4 libros de Excel, por lo que es pertinente consolidar los datos, con lo que ya obtenemos nuestro *dataframe* agrupado Tabla 12. A partir de esto podemos ver la fuerza de transmisión del virus, es decir, como con unos cuantos casos importados, va incrementando a un ritmo acelerado el número de contagiados por el virus.

Igualmente, como ya lo mencionamos con anterioridad al convertir las cantidades en tasas porcentuales, podemos ver mejor la distribución de los casos Tabla 12 .

Rango_Edad	Edad_Prom	Total	Femenino	Masculino	En_Estudio	Importados	Relacionados	Tasa_%-D_Estudio	Tasa_%-Importados	Tasa_%-Relacionados
0-9	5.0	53125	25927	27198	47949	30	5146	90.26	0.06	9.69
10-19	15.0	112002	56124	55878	100404	72	11526	89.64	0.06	10.29
20-29	25.0	378909	197898	181011	344147	482	34280	90.83	0.13	9.05
30-39	34.0	399734	199169	200565	362042	576	37116	90.57	0.14	9.29
40-49	44.0	280533	141951	138582	256889	370	23274	91.57	0.13	8.30
50-59	54.0	232164	119648	112516	214966	363	16835	92.59	0.16	7.25
60-69	64.0	142063	71470	70593	132748	228	9087	93.44	0.16	6.40
70-79	74.0	75388	36912	38476	70489	83	4816	93.50	0.11	6.39
80-89	84.0	37838	19396	18442	34856	31	2951	92.12	0.08	7.80
90+	93.0	8015	4489	3526	7182	1	832	89.61	0.01	10.38

Tabla 12. Agrupación de dataframe Por tipo de Contagio y tasas porcentuales. Fuente: Tipo de Contagios en el dataset. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS

Un dato importante para tener presente es la ubicación del caso, pues de aquí se puede obtener información relevante respecto a la recuperación del paciente, por lo que procedemos a realizar otro *dataframe* repitiendo la metodología y obtener el consolidado de los casos Tabla 13.

	Rango_Edad	Edad_Prom	Total	Femenino	Masculino	Casa	Fallecidos	Ucis	%_Casa	%_Fallecidos	%_Ucis
0	0-9	5.0	53043	25886	27157	51729	62	1252	97.52	0.12	2.36
1	10-19	15.0	111912	56088	55824	111007	65	840	99.19	0.06	0.75
2	20-29	25.0	378664	197810	180854	375855	478	2331	99.26	0.13	0.62
3	30-39	34.0	399476	199063	200413	395224	1148	3104	98.94	0.29	0.78
4	40-49	44.0	280170	141802	138368	274093	2537	3540	97.83	0.91	1.26
5	50-59	54.0	231477	119336	112141	220658	5655	5164	95.33	2.44	2.23
6	60-69	64.0	140932	71001	69931	125330	10286	5316	88.93	7.30	3.77
7	70-79	74.0	74112	36405	37707	58193	11910	4009	78.52	16.07	5.41
8	80-89	84.0	36636	18859	17777	24650	9864	2122	67.28	26.92	5.79
9	90+	93.0	7637	4284	3353	4593	2718	326	60.14	35.59	4.27

Tabla 13. Ubicación de detección de Casos COVID-19. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

En este caso podemos observar que las personas de 50 años en adelante son los que más llegan a las Ucis pese a que porcentualmente son pocos, si llegan las cantidades suficientes para congestionar las camas ucis disponibles Tabla 13.

Otra de las variables de estudio con respecto al dataset de casos en colombia es el estado del paciente, pues nos ayuda a medir la fuerza con la que ataca el virus en una poblacion especifica, por lo que realizamos tambien otro dataframe con las variables de estudio.

Rango_Edad	Edad_Prom	Total	Femenino	Masculino	Leves	Moderados	Graves	Fallecidos	(%)casos-Leves	(%)casos-Graves	(%)casos-Moderados	(%)casos-Fallecidos
0-9	5.0	53043	25886	27157	51729	1169	83	62	97.52	0.16	2.20	0.12
10-19	15.0	111912	56088	55824	111007	768	72	65	99.19	0.06	0.69	0.06
20-29	25.0	378664	197810	180854	375855	2186	145	478	99.26	0.04	0.58	0.13
30-39	34.0	399476	199063	200413	395224	2796	308	1148	98.94	0.08	0.70	0.29
40-49	44.0	280170	141802	138368	274093	3151	389	2537	97.83	0.14	1.12	0.91
50-59	54.0	231477	119336	112141	220658	4468	696	5655	95.33	0.30	1.93	2.44
60-69	64.0	140932	71001	69931	125330	4517	799	10286	88.93	0.57	3.21	7.30
70-79	74.0	74112	36405	37707	58193	3429	580	11910	78.52	0.78	4.63	16.07
80-89	84.0	36636	18859	17777	24650	1913	209	9864	67.28	0.57	5.22	26.92
90+	93.0	7637	4284	3353	4593	315	11	2718	60.14	0.14	4.12	35.59

Tabla 14. Ubicación de detección de Casos COVID-19 porcentual. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS

En este caso Tabla 14, podemos percibir que de acuerdo a la edad, la mayoria de casos son leves, los casos graves no llegan al 1% y los moderados estan entre el 1% y el 6% , sin embargo si son graves los casos de fallecidos con respecto a su edad.

Una forma de medir la capacidad del sistema de salud, es con el tiempo de procesamiento de muestras, para validar si el virus ya no es transmisible por el paciente, con lo que se realiza el estudio del tipo de recuperación, junto al tiempo de respuesta del mismo, por ende realizaremos otro *dataframe* en el cual estudiaremos esta variable.

	Rango_Edad	Edad_Prom	Total	Femenino	Masculino	Prueba_PCR	Recuperado-Tiempo	%-PCR	%-Rec_tiempo
0	0-9	5.0	49760	24361	25399	4803	44957	9.65	90.35
1	10-19	15.0	105716	52930	52786	9815	95901	9.28	90.72
2	20-29	25.0	359348	187362	171986	39464	319884	10.98	89.02
3	30-39	34.0	377593	187891	189702	44845	332748	11.88	88.12
4	40-49	44.0	261773	133014	128759	30740	231033	11.74	88.26
5	50-59	54.0	210827	109848	100979	23462	187365	11.13	88.87
6	60-69	64.0	119992	62274	57718	12821	107171	10.68	89.32
7	70-79	74.0	56182	29229	26953	5808	50374	10.34	89.66
8	80-89	83.0	23920	13369	10551	2601	21319	10.87	89.13
9	90+	93.0	4467	2795	1672	494	3973	11.06	88.94

Tabla 15. Capacidad de Diagnostico del Sistema de Salud. Fuente: Ubicación de detección de Casos COVID-19. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Como podemos notar la mayoria de los casos en la Tabla 15, se declara recuperado por tiempo, mas que por la realizacion de la segunda prueba, la cual va mas dirigida a las personas de la tercera edad.

6.5. ESTUDIO DE LAS SERIES TEMPORALES

Los tiempos son otra variable importante en el estudio, pues nos ayudan a medir y realizar los pronosticos, por esta razón es pertinente realizar otros *dataframe* en los cuales tengamos en cuenta el tiempo.

Incialmente realizaremos el dataframe que nos enseñe cada cuanto se reportan nuevos casos de contagio por departamento, esto con la intencion de poder filtrar y ver con

respecto al tiempo en cada lugar, cuantos nuevos casos se reportan, es decir su difusion viral Tabla 16.

	Fecha de diagnóstico	Nombre departamento	Total
0	2020-01-04	ANTIOQUIA	6
1	2020-01-04	BARRANQUILLA	3
2	2020-01-04	BOGOTA	81
3	2020-01-04	CASANARE	1
4	2020-01-04	CAUCA	3
...	...	...	...
8550	2021-05-01	TOLIMA	390
8551	2021-05-01	VALLE	272
8552	2021-05-01	VAUPES	1
8553	2021-06-01	BOYACA	1
8554	2021-06-01	VALLE	7

8555 rows × 3 columns

Tabla 16. Reportes de Caso por Departamento. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Por lo que a este *dataframe* y los posteriores, le convertimos las fechas en un tipo especial de dato de fecha, es decir lo convertimos en un *datetime* o dato temporal, que es el que vemos en la imagen, de la fecha en el que se dio el inicio de los sintomas Tabla 17.

	Fecha de inicio de síntomas	Nombre departamento	Total
0	2020-01-03	BOGOTA	3
1	2020-01-03	CARTAGENA	1
2	2020-01-03	VALLE	1
3	2020-01-04	ANTIOQUIA	12
4	2020-01-04	ATLANTICO	2
...	...	...	...
9220	2021-05-01	HUILA	1
9221	2021-05-01	NARIÑO	1
9222	2021-05-01	NORTE SANTANDER	1
9223	2021-05-01	PUTUMAYO	1
9224	2021-05-01	VALLE	8

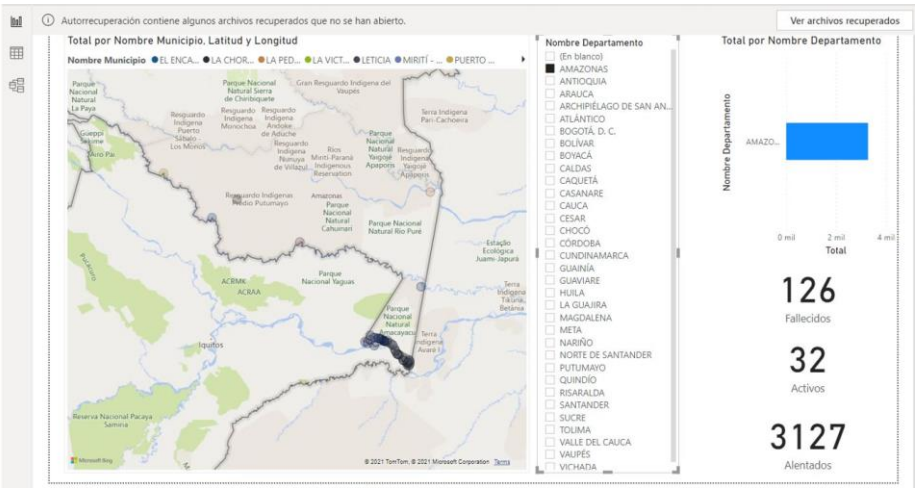
9225 rows × 3 columns

Tabla 17. Dataframe con Fecha de Contagio por Departamento. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

6.6. ESTUDIO DE LA DIFUSION VIRAL

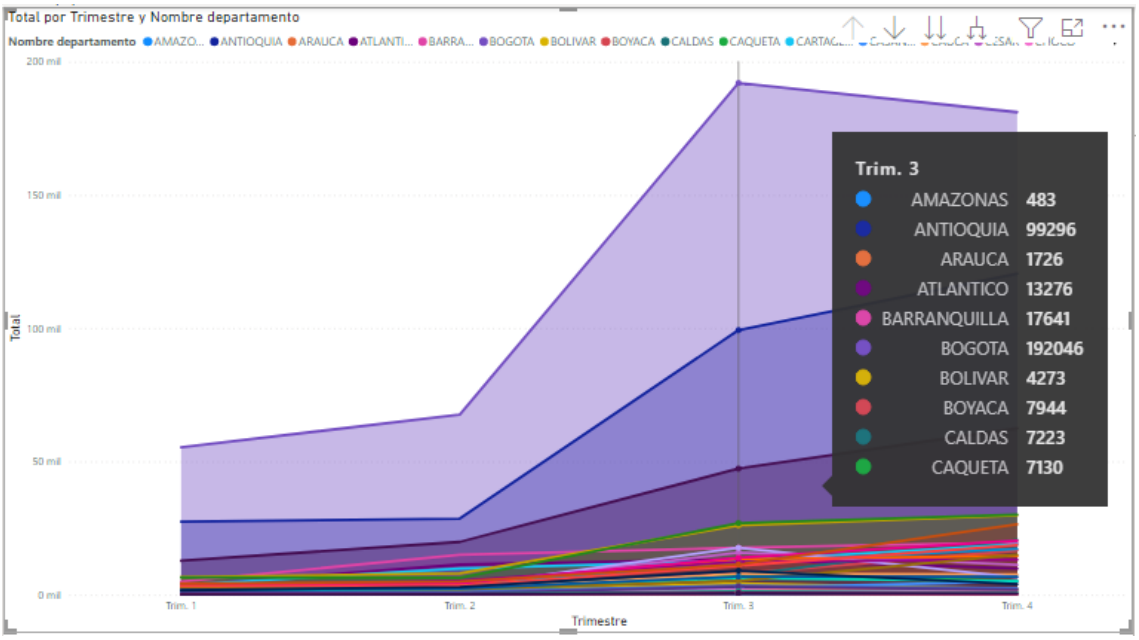
Es importante realizar este estudio de difusion viral, pues con el mismo podemos medir la facilidad de tranmisión del virus, y tener una serie temporal totalizada por dias, es un tipo ideal de calendario que pueda ser compatible con varios lenguajes, y que sean serializables es decir en diferentes periodos, tanto semanales, mensuales y anuales.

Una vez agrupados y serializados los datos, tenemos un dataframe que logramos utilizar y desplegar en Power BI, para generar la visualizacion del comportamiento de contagio regional, adicionandole los diferentes puntos georeferenciados de todo el país Grafica 10. Estos datos fueron enlazados por medio de los municipios o codigo divipola del municipio, que se encuentra en el dataset de casos COVID-19.



Grafica 10. Georreferenciación de casos en Colombia por departamento. Fuente: Imagen generada desde Power BI tomando el set de datos descargado de INS y el DANE

Tambien podemos realizar informes que sean interpretables, graficando los totales que ya hemos obtenido anteriormente, como en este caso, el top 10 de las ciudades por trimestre que mas generaron casos positivos de COVID-19 Grafica 11.



Grafica 11. Top 10 de incremento de Casos por Departamento. Fuente: Imagen generada desde Power BI tomando el set de datos descargado de INS.

Tambien realizamos el mismo tratamiento con las otras fechas, para mantener el estandar y poder realizar los diferentes analisis que necesitamos para la investigacion Tabla 18.

	Fecha de inicio de síntomas	Fecha de diagnóstico	Tiempo-Diagnost	Day-Week	Week	Day-Month	Month-Year	Year
0	2020-02-27	2020-06-03	97.0	3	23	3	6	2020
1	2020-04-03	2020-09-03	153.0	4	36	3	9	2020
2	2020-02-29	2020-09-03	187.0	4	36	3	9	2020
3	2020-06-03	2020-11-03	153.0	2	45	3	11	2020
4	2020-08-03	2020-11-03	92.0	2	45	3	11	2020
...	...	...	...	...	...	...	...	...
1719766	2020-12-19	2021-03-01	72.0	1	9	1	3	2021
1719767	2020-12-18	2021-02-01	45.0	1	5	1	2	2021
1719768	2020-12-18	2021-02-01	45.0	1	5	1	2	2021
1719769	2020-12-14	2020-12-31	17.0	4	53	31	12	2020
1719770	2020-12-18	2021-02-01	45.0	1	5	1	2	2021

1470991 rows × 8 columns

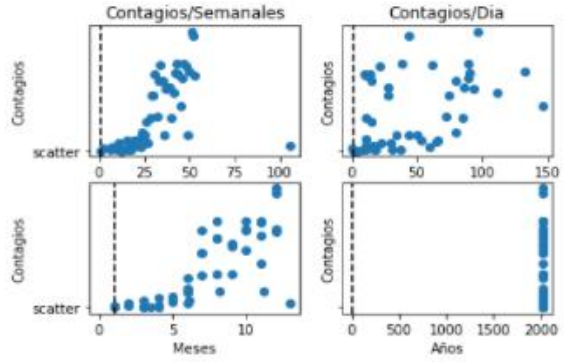
Tabla 18. Formateo de Fechas del dataset. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Como ejemplo, en este caso medimos la capacidad de respuesta entre el inicio de los sintomas y la fecha de diagnostico, para saber en promedio cuantos dias demora en ser diagnosticado un paciente, con lo que podemos evidenciar mejoria en el sistema de salud respecto al diagnostico, pues respecto a las primeras pruebas que fueron toma das su respuesta estuvo entre 97 y 180 dias en diagnosticarse, y ahora el tiempo estimado es 2 dias, este es el tipo de analisis son utiles y son pertinentes de realizar antes de iniciar a modelar los pronosticos.

Tambien podemos visualizar que exactamente, fue a partir del día 3 de la semana 23 del año 2020, cuando empezaron las autoridades de salud, a tener claridad y certeza sobre los casos de contagio, de los primeros casos que iniciaron en la ultima semana de febrero de 2020, por lo que hubo 97 dias entre ambas fechas.

6.7. DEFINICION DE SERIES TEMPORALES PARA PRONOSTICOS

Otra parte importante es tener presente que para realizar otros estudios que veremos posteriormente, es necesario estandarizar los tiempos en semanas o meses, teniendo como referencia un dia inicial de semana o mes, para realizar los estudios con el FbProphet, con el cual podremos medir las series temporales, por lo que añadiremos a la matriz estas variables. Teniendo presente las diferencias que se presentan entre las fechas de inicio de sintomas y el diagnostico, podemos ver que el dia por ser muy disperso y el año por ser muy corto no son un buen referente para poder realizar la medicion, por lo que es preferible realizarlas con la semana o el mes Grafica 12.



Grafica 12. Series temporales y su pertinencia. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Dado que tanto la transmision del virus como la de los datos demora como minimo una semana, procederemos a estandarizar la medicion a semanas y en ocasiones en meses, lo segundo se realizara unicamente en las ocasiones cuando la semana sea insuficiente Tabla 19.

	Fecha de diagnóstico	Tiempo-Diagnost	Month-Year
0	2020-06-03	97.0	6
1	2020-09-03	153.0	9
2	2020-09-03	187.0	9
3	2020-11-03	153.0	11
4	2020-11-03	92.0	11
...	...	...	...
1719727	2020-12-30	14.0	12
1719736	2020-12-31	15.0	12
1719741	2020-12-24	73.0	12
1719742	2020-12-24	73.0	12
1719769	2020-12-31	17.0	12

1067358 rows × 4 columns

Tabla 19. Estandarización de Medición en Semanas y Meses. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Sin embargo en el caso de las semanas podemos estandarizar la semana del año en curso, como tambien el numero de semanas transcurridas, tal cual vemos en la siguiente Tabla 20, en la que podemos visualizar al final de la misma que la penultima columna corresponde a la semana del año en curso, y la ultima corresponde a la semana periodo es decir, desde la semana de inicio de la toma de datos hasta la ultima semana de la cual se descargo el dataset, realizando este procedimiento podemos generar infromacion de seguimiento en el programa, de modo que solamente será cuestion de descargar un nuevo *dataset* y hacerlo correr en el programa para que actualice toda la informacion con la que se cuenta.

	Fecha de diagnóstico	Tiempo-Diagnost	Week	Week-Period
0	2020-06-03	97.0	23	23
1	2020-09-03	153.0	36	36
2	2020-09-03	187.0	36	36
3	2020-11-03	153.0	45	45
4	2020-11-03	92.0	45	45
...	...	...	...	...
1719764	2021-02-01	44.0	5	58
1719767	2021-02-01	45.0	5	58
1719768	2021-02-01	45.0	5	58
1719769	2020-12-31	17.0	53	53
1719770	2021-02-01	45.0	5	58

1071040 rows × 4 columns

Tabla 20. Estandarización de Semanas por trascurrida o en Curso. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

En este proceso por medio de un ciclo repetitivo procedemos a recorrer las columnas que ya habiamos realizado, para crear el periodo semanal y el periodo mensual. Con estos datos, ya podemos crear un nuevo dataframe que nos muestre por periodo semanal cuantos fueron los nuevos contagios y el tiempo transcurrido en Tabla 21.

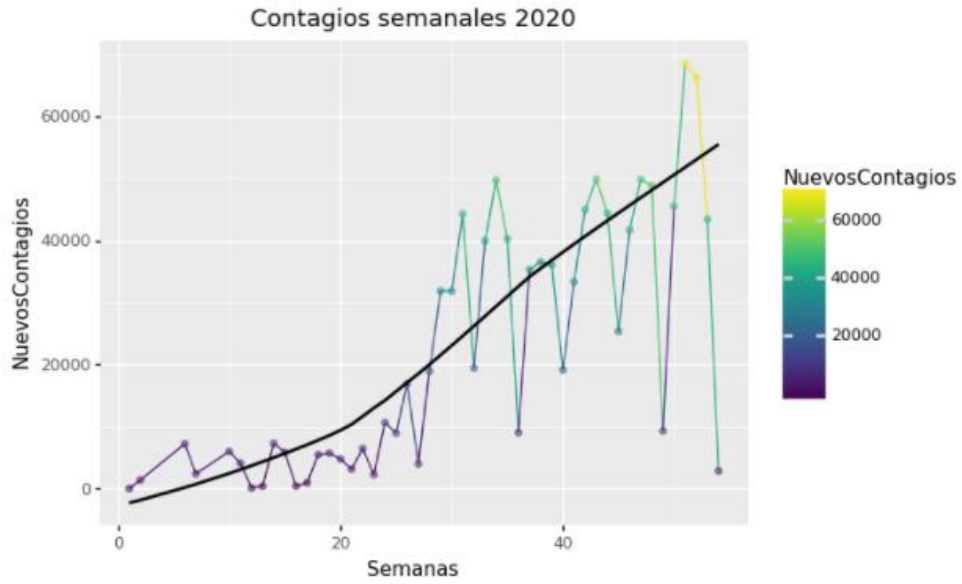
Week-Period	Semanas	NuevosContagios	DiasTranscurridos
1	1.0	6	1.00
2	2.0	1375	0.03
6	6.0	688	18.27
7	7.0	2399	12.48
10	10.0	531	37.20
11	11.0	4097	30.33
12	12.0	121	5.14
13	13.0	389	8.50
14	14.0	335	17.26
15	15.0	5845	53.22
16	16.0	419	31.46
17	17.0	942	15.73
18	18.0	747	11.44
19	19.0	5709	66.10
20	20.0	4765	64.47
21	21.0	3153	15.80
22	22.0	6463	11.03
23	23.0	2271	59.66
24	24.0	10656	79.43
25	25.0	8953	43.20
26	26.0	16921	15.29
27	27.0	3985	22.37
28	28.0	18932	85.26
29	29.0	31865	74.05
30	30.0	31838	28.34

Tabla 21. Casos de Contagio vs semanas - días trascurridos. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

6.8. MODELADO DE SERIES TEMPORALES

A partir de este modelo podemos medir la fuerza que tiene el virus para transmitirse entre personas, que como podemos observar desde la semana 1 hasta la 40 es

significativa la transmision del virus, y se visualiza la tendencia de manera exponencial Grafica 13.



Grafica 13. Contagios por semana. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Otra estadistica que podemos mencionar de acuerdo al anterior Grafica 13, y a los errores que ya hemos mencionado, es la eficiencia de la cuarentena, pues fue durante la misma en la que empezó a tener ese comportamiento exponencial Grafica 14.

Grafica 14. Validación pertinencia de la Cuarentena en la trasmisión del Virus. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Igualmente se debe tener en cuenta, es que los datos son recolectados via encuesta telefonica a los pacientes y sus familiares, por lo que es necesario eliminar estos periodos inconsistentes, en los que los encuestados no recuerdan bien las fechas que

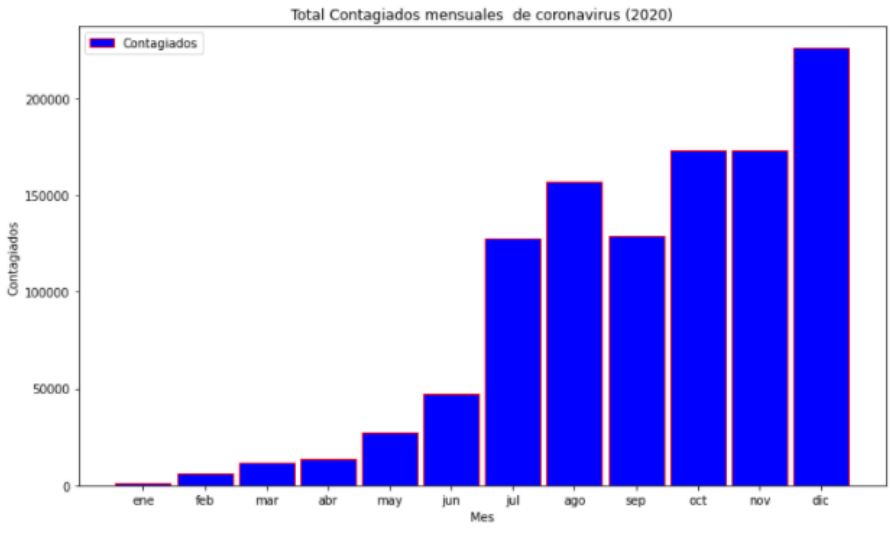
se les estan solicitando, y terminan dando fechas posteriores al evento solicitado, pues ademas de generar series temporales negativas, generan ruido en el estudio.

Por otro lado con la serie temporal mensual, una forma de comprobar que nos ha quedado filtrado de forma correcta, luego de haber realizado la resta entre la fecha final y la inicial y nos resulta misma semana del año con la semana del periodo de forma positiva Tabla 22.

Month-Period	Mes	NuevosContagios	DiasTranscurridos
1	1	1381	0.03
2	2	3087	13.77
3	3	5321	28.19
4	4	8105	43.35
5	5	20090	40.10
6	6	42204	39.42
7	7	127514	43.47
8	8	156693	53.18
9	9	128505	61.01
10	10	172795	58.72
11	11	173159	75.08
12	12	225654	72.79
13	13	2850	11.13

Tabla 22. Validación efectividad segmentación de semanas. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Indistintamente podemos ver una medicion mas ajustada utilizando el mes como periodo de observacion, por lo que a partir del mes 12 del 2020, vemos un descenso significativo, pero no es reduccion de los casos, sino porque en el mes 13 todavia se estan construyendo los datos, para el mismo Grafica 15.



Grafica 15. Contagios por Mes. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Otro tema a tener en cuenta y que realmente no es muy perceptible a la vista, es el hecho de la generacion de casos que nunca fueron diagnosticados .

	ID de caso	Fecha de inicio de síntomas	Time-Diagnost	Time-Notif	Time-Defunc	Time-Report
151	152	2020-10-03	-197 days	-199 days	-152 days	-197 days
152	153	2020-03-18	2 days	0 days	78 days	2 days
156	157	2020-12-03	-258 days	-258 days	-255 days	-258 days
187	188	2020-08-03	-135 days	-139 days	-134 days	-135 days
196	197	2020-06-03	-79 days	-82 days	-79 days	-79 days

Tabla 23. Casos de Contagios NO reportados. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Procedemos a borrar los datos que no tienen prueba oficial de diagnostico, para quedarnos con aquellos que si hayan sido reportados quedando 1067358 cantidad de registros.

Debemos tener en cuenta que si queremos realizar estudios solamente de tiempos temporales, sin otras variables, podemos crear un dataframe de solo series temporales, teniendo presente como en los casos anteriores borrar los datos errados Tabla 23 y se evidencia en la Tabla 24.

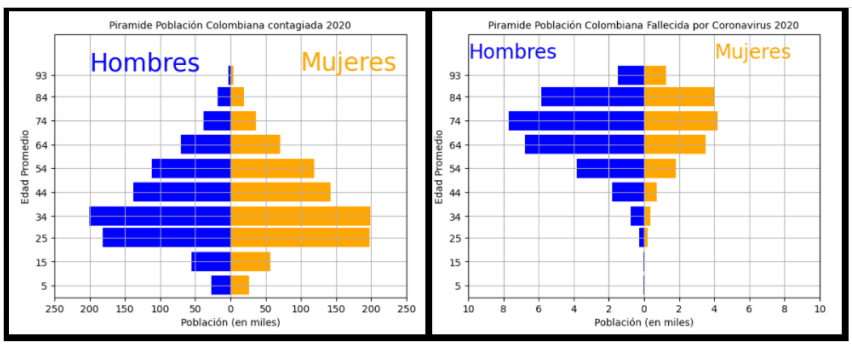
```
ID de caso                        object
Fecha de inicio de síntomas    datetime64[ns]
Fecha de diagnóstico           datetime64[ns]
Fecha de notificación          datetime64[ns]
Fecha de muerte                datetime64[ns]
fecha reporte web              datetime64[ns]
dtype: object
```

Tabla 24. Creación Dataframe con Series Temporales. Fuente: Casos de Contagios NO reportados. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

6.9. VISUALIZACION DE DATOS RELACIONALES

Una de las aplicaciones en la ingenieria es la simulacion, por medio de esta y haciendo uso del Sexo y la edad podemos visualizar el comportamiento de correlacion de ambas variables respecto al contagio del virus, en este caso mediante la piramide poblacional, podemos visualizar por sexo y edad como fue el comportamiento y la variacion tanto del contagio como de los fallecimientos en miles de personas por covid en el año 2020.

ado que la edad y sexo de los contagiados son los estudios mas tempranos que sea han generado estas son las primeras variables a las cuales debemos estudiar, pues como podemos apreciar hay un cierto numero de personas que por su edad y su sexo se contagian mas, sin embargo es otro segmento de edad y sexo las que mas fallecen por lo mismo debemos estudiar su correlacion en la muerte por el virus.



Grafica 16. Correlación entre Sexo y Edad respecto al contagio del Virus. Fuente: Casos de Contagios NO reportados. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Como tambien podemos realizar comparaciones, en este caso podemos visualizar Grafica 16, que sin importar el sexo, son las personas entre los 20 y 40 años las que mas se contagian del virus, pues son las que mas estan expuestas al mismo, por estar en su etapa mas productiva de vida, siendo el caso contrario entre las personas que se encuentran entre la primera y la ultima edad, pese a que las personas de la primera infancia sean las que menos mueren por causas asociadas al virus, o en su efecto son mas resistentes o cuentan con menos para enfrentar al virus en su sistema inmunologico, como lo podemos apreciar en la Tabla 25.

Y pese a que la variable Edad nos genera una linea perfectamente lineal con la que podriamos realizar las respectivas regresiones, debemos tener presente que la edad y Sexo, son variables, mucho mas grandes que las defunciones, por lo que se presentan problemas de correlación, ya que la variable dependiente es mas pequeña que las independientes.

	Rango_Edad	Edad_Prom	Total	Femenino	Masculino	Alentados	Activos	Fallecidos
0	0-9	5.0	53057	25894	27163	49760	3235	62
1	10-19	15.0	111928	56094	55834	105716	6147	65
2	20-29	25.0	378716	197833	180883	359348	18890	478
3	30-39	34.0	399541	199082	200459	377593	20800	1148
4	40-49	44.0	280253	141831	138422	261773	15943	2537
5	50-59	54.0	231635	119404	112231	210827	15153	5655
6	60-69	64.0	141164	71105	70059	119992	10886	10286
7	70-79	74.0	74353	36510	37843	56182	6261	11910
8	80-89	84.0	36860	18970	17890	23920	3076	9864
9	90+	93.0	7708	4326	3382	4467	523	2718

Tabla 25. Correlación del Sexo respecto al contagio del Virus. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Encontramos que su comportamiento Grafica 16, no es lineal sino normal, por lo que sirve para otros estudios, en la cual necesitemos hacer otro tipo de predicciones, sin embargo su funcion principal al igual que la de sexo se encuentra en la de agrupar las demas variables como podemos ver aquí en la tabla 25 agrupamos por la cantidad total de mujeres y hombres que se contagiaron del virus de acuerdo a sus respectivas edades,como tambien obtuvimos la cantidad de fallecidos de cada una. Finalmente, siguiendo nuestro compromiso de aportar a la comunidad estudiantil, cientifica e investigativa, procedemos a convertir nuestros dataframes de archivos (.csv) y los consolidados, para que puedan ser estudiados por otros investigadores.

7. MODELAMIENTO DE LAS MUERTES

Los contagios por COVID-19 analizados en la sección anterior no existen datos para relacionarlos con las comorbolidades. El analisis descriptivo mezclando edad y sexo se puede hacer tanto en las correlaciones como con las series temporales como se relaciona en la Grafica 16.

7.1. MODELAMIENTO DE LAS MUERTES POR EDAD PARA CORRELACIONES

Dependiendo de la necesidad, tenemos dos formas de hacerlo, relacionando las muertes totales por Covid del *dataset* de casos positivos covid Colombia, en la cual relacionamos el rango de la edad, la edad promedio y la cantidad de fallecidos, con este agrupamiento ya convertirmos en columnas la cantidad de hombres y mujeres fallecidas Tabla 26.

	Rango_Edad	Sexo	Edad_Prom	Fallecidos	MujeresFallecidas	HombresFallecidos
0	0-9	F	5.0	28	28	0
1	0-9	M	5.0	34	0	34
2	10-19	F	15.0	28	28	0
3	10-19	M	15.0	37	0	37
4	20-29	F	25.0	199	199	0
5	20-29	M	25.0	279	0	279
6	30-39	F	34.0	370	370	0
7	30-39	M	34.0	778	0	778
8	40-49	F	44.0	718	718	0
9	40-49	M	44.0	1819	0	1819
10	50-59	F	54.0	1810	1810	0
11	50-59	M	54.0	3845	0	3845
12	60-69	F	64.0	3510	3510	0
13	60-69	M	64.0	6776	0	6776
14	70-79	F	74.0	4201	4201	0
15	70-79	M	74.0	7709	0	7709
16	80-89	F	84.0	4028	4028	0
17	80-89	M	84.0	5836	0	5836
18	90+	F	93.0	1242	1242	0
19	90+	M	93.0	1476	0	1476

Tabla 26. Muertes totales por Covid-19 por rangos de edad y Sexo. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Sin embargo para realizar correlaciones de edad con respecto a los fallecidos necestiamos ampliar el numero de regresores en X, esto para evitar problemas de colinealidad junto a otros problemas matemáticos, y probabilísticos que exige la librería statsmodel que se arreglen como el caso del *test curtosis*, el cual necesita alrededor de 20 observaciones para poderse ejecutar, por lo que en este caso es pertinente entrar a dividir en 20 rangos la edad, para poder evitar los problemas anteriormente descritos Tabla 27.

Rango_Edad	Edad_Prom	TotalContagiados	Femenino	Masculino	Alentados	Activos	Fallecidos
0-5	3.0	27969	13468	14501	26118	1812	39
6-10	8.0	32268	15926	16342	30460	1783	25
11-15	13.0	43473	21966	21507	41013	2441	19
16-20	18.0	87003	43590	43413	82221	4712	70
21-25	23.0	176965	93431	83534	167843	8926	196
26-30	28.0	220812	114038	106774	209699	10777	336
31-35	33.0	205368	102528	102840	194286	10607	475
36-40	38.0	184868	91683	93185	174210	9886	772
41-45	43.0	145606	73682	71924	136104	8359	1143
46-50	48.0	124314	63413	60901	115394	7315	1605
51-55	53.0	120244	62351	57893	110201	7553	2490
56-60	58.0	105067	53598	51469	94055	7375	3637
61-65	63.0	76102	38674	37428	65403	5887	4812
66-70	68.0	55577	27602	27975	45308	4441	5828
71-75	73.0	40426	19656	20770	31014	3383	6029
76-80	78.0	29523	14625	14898	20950	2551	6022
81-85	83.0	20830	10808	10022	13693	1725	5412
86-90	88.0	11839	6178	5661	7134	1000	3705
91-95	92.0	4578	2571	2007	2615	307	1656
96-100	97.0	1042	634	408	585	62	395
100+	103.0	185	112	73	117	11	57

Tabla 27. Muertes totales por Covid-19 por rangos de edad y Sexo. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

De esta forma obtenenos dos formas de evaluar la edad, como lo son la cantidad de personas que pertenecen al mismo grupo de edad, como las edades promedio de cada grupo, con lo que evitamos problemas de duplicidad en las edades, sin embargo se debe hacer la claridad, que en el caso de las comorbilidades , lo que se correlaciona son las cantidades de personas con la comorbilidad por edad, por lo que en este caso es pertinente realizar la misma apreciacion para lograr unificar las hipotesis.

7.2. MODELAMIENTO DE LAS MUERTES POR SEXO PARA CORRELACIONES

Dado que no podemos entrar a valorar el sexo como un numero estadístico, porque solamente son dos como se evidencia en la Tabla 28.

	Sexo	Femenino	Masculino	Fallecidos	Tasa%MujeresFallecen	Tasa%HombresFallecen
0	F	870534	0	16134	1.85	0.00
1	M	0	843525	28589	0.00	3.39

Tabla 28. Valor de la Variable Sexo como numero estadístico. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Debemos entrar a correlacionar las cantidades de cada uno normalizando la cantidad de hombres y mujeres como apreciamos en la siguiente Tabla 29, esto con el objetivo de evitar los inconvenientes que ya fueron anteriormente.

	Rango_Edad	Edad_Prom	TotalContagiados	Femenino	Masculino	Alentados	Activos	Fallecidos	Femenino_Normalizado	Masculino_Normalizado
0	0-5	3.0	27969	13468	14501	26118	1812	39	0.117234	0.135219
12	6-10	8.0	32268	15926	16342	30460	1783	25	0.138809	0.152473
2	11-15	13.0	43473	21966	21507	41013	2441	19	0.191826	0.200879
3	16-20	18.0	87003	43590	43413	82221	4712	70	0.381634	0.406182
4	21-25	23.0	176965	93431	83534	167843	8926	196	0.819119	0.782195
5	26-30	28.0	220812	114038	106774	209699	10777	336	1.000000	1.000000
6	31-35	33.0	205368	102528	102840	194286	10607	475	0.898970	0.963131
7	36-40	38.0	184868	91683	93185	174210	9886	772	0.803776	0.872644
8	41-45	43.0	145606	73682	71924	136104	8359	1143	0.645770	0.673386
9	46-50	48.0	124314	63413	60901	115394	7315	1605	0.555633	0.570079
10	51-55	53.0	120244	62351	57893	110201	7553	2490	0.546311	0.541888
11	56-60	58.0	105067	53598	51469	94055	7375	3637	0.469480	0.481682
13	61-65	63.0	76102	38674	37428	65403	5887	4812	0.338483	0.350090
14	66-70	68.0	55577	27602	27975	45308	4441	5828	0.241297	0.261497
15	71-75	73.0	40426	19656	20770	31014	3383	6029	0.171550	0.193972
16	76-80	78.0	29523	14625	14898	20950	2551	6022	0.127390	0.138940
17	81-85	83.0	20830	10808	10022	13693	1725	5412	0.093886	0.093242
18	86-90	88.0	11839	6178	5661	7134	1000	3705	0.053245	0.052371
19	91-95	92.0	4578	2571	2007	2615	307	1656	0.021584	0.018125
20	96-100	97.0	1042	634	408	585	62	395	0.004582	0.003140
1	100+	103.0	185	112	73	117	11	57	0.000000	0.000000

Tabla 29. Normalización de datos por Sexo. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

7.3. MODELAMIENTO ESTADISTICO

La regresión lineal es una técnica que se utiliza para predecir el valor de una variable dependiente de acuerdo a una o varias independientes, es decir, busca hallar la correlación entre ambas, para este caso utilizaremos el método de mínimos cuadrados del modelo estadístico *StatsModel* utilizando, el modelo de regresión lineal ordinaria más conocida como *Ordinary Least Regression* (OLS), el cual además de proporcionarnos el método de mínimos cuadrados, nos ofrece otra gama de mediciones, con las cuales nos guiamos y nos aporta para identificar las variables que se correlacionan mejor, y así lograr convertirlas en los próximos pronósticos, también utilizaremos *Scikit Learn* para realizar las predicciones de la correlación entre las variables, esto con la intención de simular el comportamiento antes de convertirlo en pronósticos, sin embargo el modelo tiene una reglas específicas para modelarlo, que son las mencionaremos a continuación [32].

El modelo se basa en la siguiente ecuación:

$yi=\beta0+\beta1xi1+\beta2xi2+...+\beta pxip+\epsilon i$

Donde:

yi Valor de la correlación entre la variable dependiente y la independiente→

β0 Valor promedio del origen yi→

βpxip→ Valor que va tomando la observación predictora, es decir son los coeficientes de regresión.

Sin embargo, para el caso de la maquina en el modelo Patsy lo determina como:

y= Variable independiente (y1)

X= Variable independiente (βpxip)

con lo que la correlación que se estudia es entre (y, X) de la forma y ~ X

Por lo que:

La Hipotesis Nula Ho, corresponde a una relación entre la variable Independiente y la dependiente.

La Hipotesis Alternativa Ha, corresponde a la inexistencia de la relación entre la variable Independiente o es demasiado baja.

7.3.1. AJUSTES DEL MODELO

Para esta investigación se realizará principalmente la correlación lineal simple, que consta de una sola variable independiente X y una variable dependiente Y, por lo que sus coeficientes de determinacion son R^2 y $R^2_{ajustado}$[32].

7.3.2. LINEAMIENTOS PARA CREAR EL MODELO.

Por medio de los siguientes lineamientos y principios buscaremos que la ejecución del modelo sea lo más preciso posible, es decir que no se presenten problemas numericos como R^2 =1 y que la misma maquina nos indique una ejecución perfecta, en la cual no se presente problemas de colinealidad, para que los errores estandar queden perfectamente correlacionados en la matriz de covarianza[32].

7.3.2.1. COEFICIENTE DE CORRELACION R^2

Este coeficiente mide la calidad del ajuste, es decir, de entre las infinitas lineas que pueden aparecer, logra estimar que tan bien se ajusta esta linea frente a las observaciones, su valor va en el rango de [0-1] donde 0 es que no hay ningun tipo de covarianza y 1 que existe una correlacion perfecta, sin embargo esta relacion perfecta, estaria generando errores numericos, por lo que tampoco podria ser posible medir esa covarianza[32].

7.3.2.2. COEFICIENTE DE CORRELACION $R^2_{ajustado}$

Este coeficiente ajusta el coeficiente de correlación R^2 reduciendo el error que se presenta en cuando R^2 tiene predictores que no son representativos, por lo que se puede conseguir un mejor modelo explicando la mejor variabilidad de la Dependiente, con menos predictores[32].

7.3.3. SIGNIFICANCIA DEL MODELO F-Test

Este test mide el grado de significancia del modelo, es decir que tan útil es el modelo, para describir la correlación, puesto que no todos los predictores son necesario al interior del modelo, se describe de la forma:

H0: $\beta 1 = ... = \beta p-1=0$

Ha: al menos un βi≠0

Donde:

- Ho=Hipótesis nula.
- Ha= Hipótesis alternativa
- β= coeficiente de regresión.
- p= variable predictora.

Por lo que en la regresión lineal simple la descripción de las hipótesis es la siguiente:

- La hipótesis nula Ho: se describe como el predictor p_j que no contribuye al modelo cuando (β_j=0) en presencia del resto de predictores, es decir que no existe una correlación lineal entre ambas variables, porque la pendiente del modelo, (β_j=0).
- La hipótesis alternativa Ha: se describe como el predictor p_j, si contribuye al modelo cuando (βj≠0), en presencia del resto de predictores, es decir que si existe una correlación lineal entre ambas variables[32].

7.3.4. VALOR ESTADISTICO T

Para llegar a la conclusión anterior es necesario hallar el valor estadístico t, también conocido como una prueba *t student*, la cual es una distribución de probabilidad que mide que tan significativas son las diferencias entre grupos de datos, es decir mide que tantas veces es posible repetir el experimento y que los resultados sean por casualidad, lo cual quiere decir que entre mayor sea la puntuación del valor t, habrá mayor diferencia entre los datos, por lo que obteniendo un *p-value* bajo que sea inferior a la prueba t, significa que los resultados no son producto de la casualidad En este cálculo se divide el coeficiente de correlación esperado, β^j sobre el mismo multiplicado por su desviación estándar [32], es decir:

$t=\frac{\beta j}{se(\beta j)}$

donde:

$$SE(\beta j)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(xji - \tilde{x})^2}$$

descripción:

- σ^2 = error cuadrático de la varianza.
- n= número de predictores
- $xij = predictor\ en\ el\ punto\ ij$
- $\tilde{x}$ = valor promedio de la variable predictora
- $i = valor\ inicial\ de\ n$

Sin embargo, la varianza del error σ^2 se estima con respecto a los valores y con la formula con el error estándar residual RSE:

$$\sqrt{\frac{1}{n-2\ \sum_{i=1}^{n}(y-\overline{yi})}}$$

Donde:

- n= número de observaciones.
- y=valor a predecir
- $\bar{y}$ = valor promedio de la variable y.
- $i = valor\ inicial\ de\ n$
- n-2= grados de libertad

Dada esta relación se busca que los predictores contribuyan al modelo siendo estos predictores -> p≠0, por lo que los p-value = P(|t|> valor calculado de t) = prob (F-statistic) de tal modo que el p-value es significativo siempre y cuando sea diferente de 0.000, o en su efecto que el prob (F-statistic) no se mida en la escala de $a^{-n} = \frac{1}{a^n}$ porque ya será un valor demasiado cercano a 0 por lo que deberemos entrar a validarlo bajo otros parámetros[32].

7.3.5. PRINCIPIO DE NO COLINEALIDAD

La colinealidad ocurre cuando los predictores están linealmente relacionados, con uno o varios predictores, esto impide medir la significancia estadística, es decir si su R^2 =1 [32].

7.3.6. PRINCIPIO DE NO MULTICOLINEALIDAD

La multicolinealidad se presenta cuando colinealidad es muy común entre todos los predictores, se presenta ante todo en modelos en los que se estudia la correlación múltiple, esta parte también será analizada en la matriz que crearemos para la relación múltiple, con el método de Pearson[32].

7.3.7. PRINCIPIO DE HOMOCEDASTICIDAD

En el principio de homocedasticidad se busca que la varianza de la variable de respuesta, de respuesta sea constante, por lo que necesitamos que la mayor parte de los residuos que presentan los estimadores en el dataframe sean constantes, por lo que es mucho más confiable[32].

7.3.8. PRINCIPIO DE NO HETEROCEDASTICIDAD

Corresponde al caso contrario de la homocedasticidad, es decir el comportamiento de las variables no presentan un comportamiento homogéneo, por lo que debemos evitar que los errores sean inconstantes, para que se pueda medir la varianza[32].

7.4. PRUEBAS INICIALES DE CORRELACION

Dependiendo de la correlacion entre las variables procederemos a desechar la hipotesis o a afirmarla, para posteriormente realizar el pronostico de las mismas. Inicialmente utilizaremos como *dataset* de prueba el de los casos reportados en los Estados Unidos, la cual se puede descargar en la pagina data.cdc.gov[10] :

[10] https://data.cdc.gov/NCHS/Conditions-Contributing-to-COVID-19-Deaths-by-Stat/hk9y-quqm

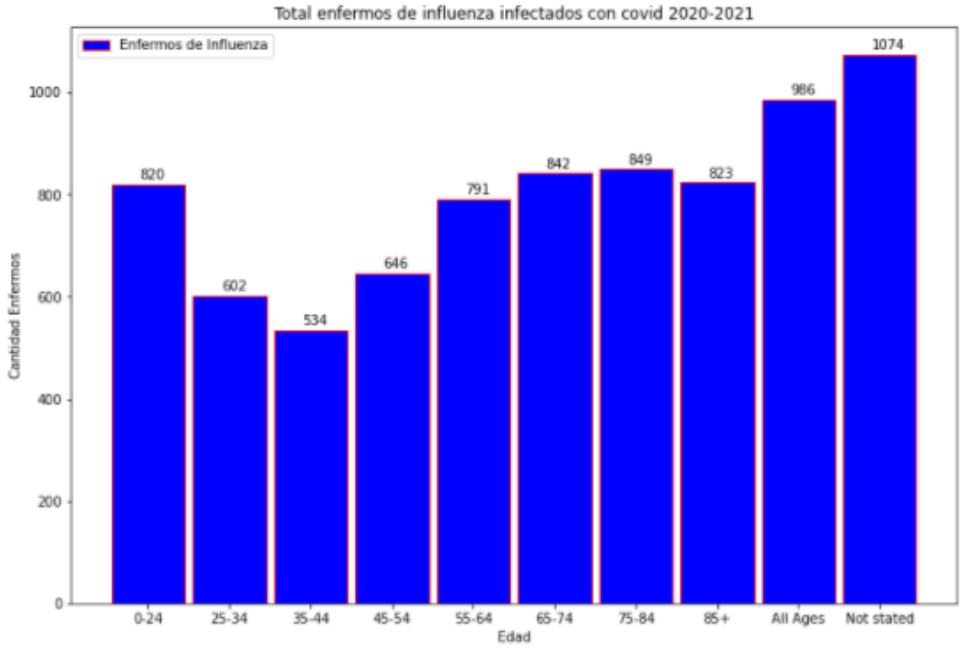
dado que este dataset nos presenta informacion con la que no contamos en Colombia, bien sea porque las autoridades se encuentran en la investigacion o bien sea porque nuestras leyes como la de *Habeas Data*, convierte en la historia clinica en un dato sensible, y dependiendo de los logros obtenidos con este dataset, procederemos a modelar el de Colombia Tabla 30.

	Data As Of	Start Date	End Date	Group	Year	Month	State	Condition Group	Condition	ICD10_codes	Age Group	COVID-19 Deaths	Number of Mentions
37260	08/29/2021	01/01/2020	01/31/2020	By Month	2020.0	1.0	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	0.0	0.0
37261	08/29/2021	02/01/2020	02/29/2020	By Month	2020.0	2.0	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	0.0	0.0
37262	08/29/2021	03/01/2020	03/31/2020	By Month	2020.0	3.0	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	9.0	9.0
37263	08/29/2021	04/01/2020	04/30/2020	By Month	2020.0	4.0	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	27.0	30.0
37264	08/29/2021	05/01/2020	05/31/2020	By Month	2020.0	5.0	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	19.0	19.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
285655	08/29/2021	04/01/2021	04/30/2021	By Month	2021.0	4.0	Puerto Rico	COVID-19	COVID-19	U071	All Ages	197.0	197.0
285656	08/29/2021	05/01/2021	05/31/2021	By Month	2021.0	5.0	Puerto Rico	COVID-19	COVID-19	U071	All Ages	184.0	184.0
285657	08/29/2021	06/01/2021	06/30/2021	By Month	2021.0	6.0	Puerto Rico	COVID-19	COVID-19	U071	All Ages	35.0	35.0
285658	08/29/2021	07/01/2021	07/31/2021	By Month	2021.0	7.0	Puerto Rico	COVID-19	COVID-19	U071	All Ages	29.0	29.0
285659	08/29/2021	08/01/2021	08/28/2021	By Month	2021.0	8.0	Puerto Rico	COVID-19	COVID-19	U071	All Ages	190.0	190.0

177842 rows × 13 columns

Tabla 30. Correlación con datos de Estados Unidos. Fuente: Imagen generada desde Python tomando el set de datos descargado de CDC.

En esta fase iniciamos con una varible que por tener una gran cantidad de casos de todas las edades prendió las alarmas en los Estados Unidos, como lo es la Influenza y neumonia.



Grafica 17. Total de casos de enfermos de neumonía e influenza en Estados unidos. Fuente: DatasetConditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020-2021, Imagen generada desde Python.

En esta parte aprendimos que debemos generar un modelo que necesita una cantidad de datos superior a 20 pruebas, esto para evitar problemas de colinealidad, de modo que en el preprocesamiento de los datos debemos agrupar por una variable que proporcione mas de 20 filas, es decir necesitamos generar un modelo con un grado de libertad amplio, para que realice el test de cortes, con el cual se evalua si es una distribucion normal, tambien debemos tener presente, que la variable regresora es decir la dependiente, es necesariamente debe normalizarse entre 0 y 1, o en su efecto, tomar

la media de la misma para poder incluirla en el modelo, esto con la intencion de evitar colinealidad perfecta de tipo $R^2=1$ Tabla 31, y lograr que aún las probabilidades sean las más acertadas posible, como en el caso de la que podemos visualizar en la cual la neumonia nos puede ayudar a correlacionar el 20.9 % del total de las muertes en casos por covid, adicional a esto, podemos ver que la probabilidad f-stactistic, el P-value y la probabilidad JB se encuentran entre 0.184 y 0.187, lo cual en terminos probabilisticos quiere decir que la relacion presentada es muy probable en el mundo real.

Tambien podemos ver al final de la tabla que no se presentaron problemas numericos y que pese a tener solamente 10 muestras, la regresion se realizó sin problemas numericos con lo que podemos afirmar de ser el caso que la influencia y neumonia, si son factores reales que causa la muerte por COVID-19.

OLS Regression Results

Dep. Variable:	Cov19Deaths	R-squared:	0.209
Model:	OLS	Adj. R-squared:	0.110
Method:	Least Squares	F-statistic:	2.115
Date:	Sun, 31 Oct 2021	Prob (F-statistic):	0.184
Time:	15:23:43	Log-Likelihood:	-154.29
No. Observations:	10	AIC:	312.6
Df Residuals:	8	BIC:	313.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.112e+05	8.35e+05	-0.133	0.897	-2.04e+06	1.81e+06
InfluPneu	2.141e+06	1.47e+06	1.454	0.184	-1.25e+06	5.53e+06

Omnibus:	8.659	Durbin-Watson:	2.365
Prob(Omnibus):	0.013	Jarque-Bera (JB):	3.352
Skew:	1.073	Prob(JB):	0.187
Kurtosis:	4.854	Cond. No.	4.30

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Tabla 31. Resultados regresión OLS. Fuente: Imagen generada desde Python tomando el set de datos descargado de CNC.

Tambien aprendimos que pese a que en la documentacion se encuentra, que es posible realizar el modelo con variables categoricas, para incrementar el nivel de filas del modelo, esto hace que el modelo se vuelva altamente colineal Tabla 32.

OLS Regression Results

Dep. Variable:	Cov19Deaths	R-squared:	0.980
Model:	OLS	Adj. R-squared:	0.980
Method:	Least Squares	F-statistic:	7.870e+05
Date:	Sat, 30 Oct 2021	Prob (F-statistic):	0.00
Time:	01:42:29	Log-Likelihood:	-1.0626e+06
No. Observations:	177842	AIC:	2.125e+06
Df Residuals:	177830	BIC:	2.125e+06
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.1102	0.640	-0.172	0.863	-1.365	1.144
AgeGroup[T.25-34]	-0.0205	0.932	-0.022	0.982	-1.847	1.806
AgeGroup[T.35-44]	-0.0305	0.968	-0.031	0.975	-1.928	1.867
AgeGroup[T.45-54]	-0.0218	1.005	-0.022	0.983	-1.992	1.949
AgeGroup[T.55-64]	-0.1007	1.013	-0.099	0.921	-2.087	1.885
AgeGroup[T.65-74]	-0.0813	1.007	-0.081	0.936	-2.055	1.893
AgeGroup[T.75-84]	0.2590	1.001	0.259	0.796	-1.703	2.221
AgeGroup[T.85+]	0.4951	0.999	0.496	0.620	-1.462	2.452
AgeGroup[T.All Ages]	0.1826	0.947	0.193	0.847	-1.673	2.038
AgeGroup[T.Not stated]	0.0311	0.879	0.035	0.972	-1.692	1.754
VascularAndUnspecifiedDementia	1.8142	1.037	1.750	0.080	-0.218	3.846

show more (open the raw output data in a text editor) ...

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.78e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Tabla 32. Validación de Variables categóricas OLS. Fuente: Imagen generada desde Python tomando el set de datos descargado de CNC.

7.3. PROCEDIMIENTO PARA EVALUAR LAS CORRELACIONES CON EL DATASET DE CASOS POSITIVOS COLOMBIA

Para nuestro caso retomando los datos obtenidos con el dataset de casos positivos covid en Colombia, y utilizaremos como variable dependiente la muerte por COVID-19, y como variables independientes utilizaremos:

- Sexo
- Edad
- Comorbilidad

7.3.1. MODELAMIENTO DE VARIABLE SEXO PARA CORRELACION

Debido a las indicaciones anteriores para el caso colombiano, procedimos a crear el nuevo *dataframe* que cuenta con la edad como índice, esto con la intención de superar las 20 filas que nos exige el modelo, también que no haya valores repetidos, y que la cantidad de contagiados por sexo sea la suma del total, también creamos la columna *Femenino_Normalizado* y *Masculino_Normalizado*, que son las columnas femenino y masculino pero normalizadas entre 0 y 1, esto con la intención de eliminar problemas de colinealidad y nula probabilidad de que sea posible la correlación Tabla 33.

Rango_Edad	Edad_Prom	TotalContagiados	Femenino	Masculino	Alentados	Activos	Fallecidos	Femenino_Normalizado	Masculino_Normalizado
0-5	3.0	27969	13468	14501	26118	1812	39	0.117234	0.135219
6-10	8.0	32268	15926	16342	30460	1783	25	0.138809	0.152473
11-15	13.0	43473	21966	21507	41013	2441	19	0.191826	0.200879
16-20	18.0	87003	43590	43413	82221	4712	70	0.381634	0.406182
21-25	23.0	176965	93431	83534	167843	8926	196	0.819119	0.782195
26-30	28.0	220812	114038	106774	209699	10777	336	1.000000	1.000000
31-35	33.0	205368	102528	102840	194286	10607	475	0.898970	0.963131
36-40	38.0	184868	91683	93185	174210	9886	772	0.803776	0.872644
41-45	43.0	145606	73682	71924	136104	8359	1143	0.645770	0.673386
46-50	48.0	124314	63413	60901	115394	7315	1605	0.555633	0.570079
51-55	53.0	120244	62351	57893	110201	7553	2490	0.546311	0.541888
56-60	58.0	105067	53598	51469	94055	7375	3637	0.469480	0.481682
61-65	63.0	76102	38674	37428	65403	5887	4812	0.338483	0.350090
66-70	68.0	55577	27602	27975	45308	4441	5828	0.241297	0.261497
71-75	73.0	40426	19656	20770	31014	3383	6029	0.171550	0.193972
76-80	78.0	29523	14625	14898	20950	2551	6022	0.127390	0.138940
81-85	83.0	20830	10808	10022	13693	1725	5412	0.093886	0.093242
86-90	88.0	11839	6178	5661	7134	1000	3705	0.053245	0.052371
91-95	92.0	4578	2571	2007	2615	307	1656	0.021584	0.018125
96-100	97.0	1042	634	408	585	62	395	0.004582	0.003140
100+	103.0	185	112	73	117	11	57	0.000000	0.000000

Tabla 33. Inclusión Variable Departamento. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

7.3.2. MODELO DE CORRELACION ENTRE SEXO FEMENINO Y MUERTE POR COVID

A partir de aquí ya podemos ingresar las variables dentro de la matriz Patsy, la cual es la que utiliza la librería statsmodel para realizar las correlaciones, para este caso visualizamos la variable Dependiente Y= Fallecidos, relacionada con la variable independiente X= Femenino_Normalizado, y el punto de intersección que le asigna la librería para que inicie la gráfica, como lo podemos observar en la Tabla 34.

	Fallecidos
0	39.0
12	25.0
2	19.0
3	70.0
4	196.0
5	336.0
6	475.0
7	772.0
8	1143.0
9	1605.0
10	2490.0
11	3637.0
13	4812.0
14	5828.0
15	6029.0
16	6022.0
17	5412.0
18	3705.0
19	1656.0
20	395.0
1	57.0

	Intercept	Femenino_Normalizado
0	1.0	0.117234
12	1.0	0.138809
2	1.0	0.191826
3	1.0	0.381634
4	1.0	0.819119
5	1.0	1.000000
6	1.0	0.898970
7	1.0	0.803776
8	1.0	0.645770
9	1.0	0.555633
10	1.0	0.546311
11	1.0	0.469480
13	1.0	0.338483
14	1.0	0.241297
15	1.0	0.171550
16	1.0	0.127390
17	1.0	0.093886
18	1.0	0.053245
19	1.0	0.021584
20	1.0	0.004582
1	1.0	0.000000

Tabla 34. Relación de la variable dependiente y la independiente Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Luego de realizar esta correlación, procedemos a ejecutar el modelo, la respuesta fue mucho mejor de lo que esperábamos, puesto que logramos eliminar los problemas que se presentan tanto de:

Colinealidad: Estos se eliminaron al normalizar la variable independiente

Probabilísticos: se logró que Probabilidad (F-statistic) = P-Value (P>|t|) = 0.189 por lo que es una medición que se encuentra cercana a 1 sin ser 1.

Correlación imperfecta: dado que R^2= 0.089 la cual es un R^2<> 1

Errores estándar de covarianza mal especificados, este es un problema que se presenta cuando la covarianza muestra una correlación inversa, es decir que la recta, va en sentido contrario a los puntos.

Dado el éxito del primer experimento, podemos asumir que los datos de salida son correctos, sin embargo, en el tema de la correlación tener R^2=0.089 es una correlación demasiado baja, como para ser aprobada la hipótesis nula, de modo que, en este caso, para el sexo femenino, es necesario desechar la correlación.

OLS Regression Results

Dep. Variable:	Fallecidos	R-squared:	0.089
Model:	OLS	Adj. R-squared:	0.041
Method:	Least Squares	F-statistic:	1.855
Date:	Sun, 21 Nov 2021	Prob (F-statistic):	0.189
Time:	22:16:58	Log-Likelihood:	-190.75
No. Observations:	21	AIC:	385.5
Df Residuals:	19	BIC:	387.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2903.0158	749.322	3.874	0.001	1334.666	4471.366
Femenino_Normalizado	-2131.1155	1564.745	-1.362	0.189	-5406.164	1143.933

Omnibus:	3.478	Durbin-Watson:	0.156
Prob(Omnibus):	0.176	Jarque-Bera (JB):	1.609
Skew:	0.328	Prob(JB):	0.447
Kurtosis:	1.813	Cond. No.	3.66

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Tabla 35. Correlación entre Sexo Femenino y Fallecidos por Covid-19. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Para iniciar la Tabla 35, comenzamos con una de las variables que encontramos en el dataset de casos Colombia, con la variable muy esperanzadora dado su alta efectividad para agrupar, el sexo pese a que se divide en 2 nos brinda la posibilidad de segmentar, con lo cual buscamos encontrar la correlación entre el sexo y la muerte por COVID, en

este caso iniciamos con el femenino, pero normalizando la variable dependiente para que sea más pequeña que la variable dependiente los fallecidos por COVID-19, sin embargo como ya lo mencionamos, su grado de correlación R^2 = 0.089, lo cual en nuestra escala de apreciación de la correlación es significativamente baja, por otro lado podemos ver que la probabilidad

7.3.3. MODELO DE CORRELACION ENTRE SEXO MASCULINO Y FALLECIMIENTO POR COVID

Por consiguiente, realizamos el mismo estudio con la variable Sexo masculino normalizada, como podemos apreciar en la Tabla 35 sin normalizar que son las que podemos visualizar en la Tabla 36, teniendo como Hipótesis nula, la correlación entre el Sexo masculino y la muerte por COVID-19, con lo que también logramos obtener los mismos resultados que obtuvimos con la relación entre sexo femenino y la muerte por COVID.

Para este caso obtuvimos un R^2= 0.087, igualmente unas mediciones de probabilidades iguales Probabilidad (F-statistic) = P-Value (P>|t|) = 0.195, por lo que al igual que en el caso de la relación entre sexo femenino y la muerte por COVID, debemos rechazar la hipótesis nula, y quedarnos con la hipótesis alternativa, que para este caso significa que no hay relación fuerte entre el sexo y la muerte por COVID.

OLS Regression Results

Dep. Variable:	Fallecidos	R-squared:	0.087
Model:	OLS	Adj. R-squared:	0.039
Method:	Least Squares	F-statistic:	1.803
Date:	Sun, 21 Nov 2021	Prob (F-statistic):	0.195
Time:	22:17:03	Log-Likelihood:	-190.78
No. Observations:	21	AIC:	385.6
Df Residuals:	19	BIC:	387.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2902.2927	755.486	3.842	0.001	1321.043	4483.542
Masculino_Normalizado	-2056.1233	1531.294	-1.343	0.195	-5261.158	1148.911

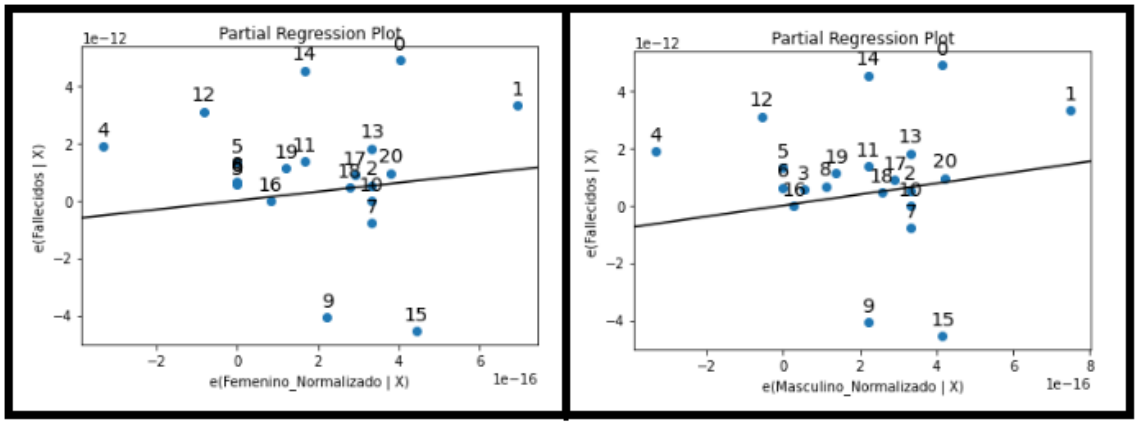
Omnibus:	3.432	Durbin-Watson:	0.154
Prob(Omnibus):	0.180	Jarque-Bera (JB):	1.624
Skew:	0.342	Prob(JB):	0.444
Kurtosis:	1.821	Cond. No.	3.61

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Tabla 36. Relación entre Sexo masculino y la muerte por COVID-19. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Para la gráfica del modelo, debemos tener presente que son las que genera la librería, por lo que no es posible generar cambios, es necesario utilizar la edad como variable categórica, con la que podamos ubicar el modelo Grafica 18, en la que podemos apreciar las variables independientes normalizadas Femenino_Normalizada y Masculino_ Normalizada, además de su variable dependiente que en ambos casos es la de fallecidos.

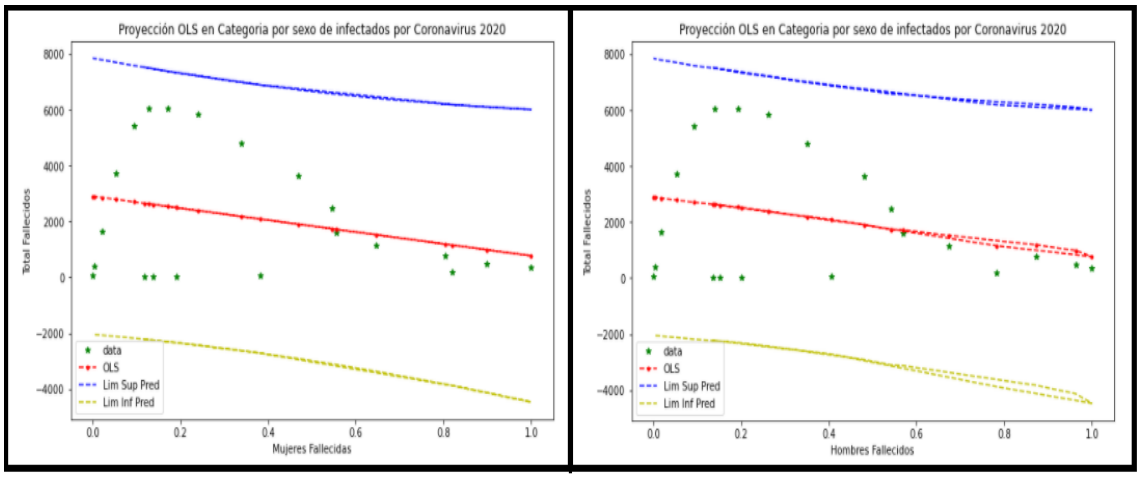
Y dado que ambas tienen la misma edad, podemos visualizar que la recta es la misma para ambos sexos, sin embargo, al haber tenido diferentes R^2, podemos llegar a la conclusión que hemos logrado la mejor recta posible para estas dos variables.



Grafica 18. Regresión por Edad y Sexo. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

7.4. PRUEBAS DE REGRESION

Una de las formas de saber, las razones por las cuales, si son concluyentes, es realizando las pruebas a la regresión, es decir hacemos la comparación entre los datos reales y la regresión OLS, generando los intervalos de predicción, para saber que sucedió Grafica 19.



Grafica 19. Prueba de los datos reales vs la regresión. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

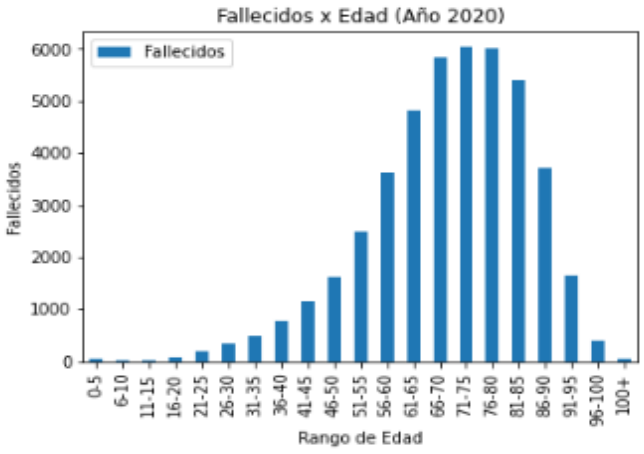
Para este caso, podemos ver que el comportamiento de los datos reales, no son completamente lineales, por lo que la estimación OLS, no se aproxima completamente al comportamiento de los datos reales, en esta Grafica 19, junto a los datos obtenidos,

se confirma que ni el sexo femenino ni el sexo masculino, tienen una relación fuerte de linealidad.

7.4.1. REGRESION LINEAL ENTRE LA EDAD Y LA MUERTE POR COVID

Una vez terminadas las pruebas de la relación entre sexo y muerte por COVID procedimos a realizar las pruebas con la variable edad en librería *Statsmodel* teniendo como hipótesis nula H0 la correlación entre el fallecimiento por COVID-19 y la edad del paciente, y como variable alternativa a la edad, y realizamos todo el proceso de regresión OLS con estas dos variables, tal cual lo dice la literatura encontrada en la página de *Statsmodel*.

Puesto que los anteriores análisis habíamos encontrado una alta probabilidad de que la edad estuviera asociada a la muerte por COVID-19, Inclusive en la Grafica 20, así mismo nos lo demostraban, intentamos reducir el error que se presenta aumentando el número de agrupaciones de la edad y reduciendo, la cantidad de datos atípicos, como lo es el de aquellos que son superiores a los 100 años, esto con la intención de obtener relaciones más fuertes, entre la edad y la cantidad de defunciones por COVID.



Grafica 20. Fallecidos por Edad. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Con el mismo dataframe que utilizamos para la correlación entre sexo y muerte por COVID, procedimos a realizar las pruebas para la variable edad. Obteniendo mejores resultados con respecto a la edad, que la obtenida con la variable sexo Tabla 37.

OLS Regression Results

Dep. Variable:	Fallecidos	R-squared:	0.278
Model:	OLS	Adj. R-squared:	0.240
Method:	Least Squares	F-statistic:	7.305
Date:	Sun, 21 Nov 2021	Prob (F-statistic):	0.0141
Time:	22:17:11	Log-Likelihood:	-188.31
No. Observations:	21	AIC:	380.6
Df Residuals:	19	BIC:	382.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	64.5157	879.393	0.073	0.942	-1776.075	1905.107
Edad	39.0353	14.442	2.703	0.014	8.807	69.263

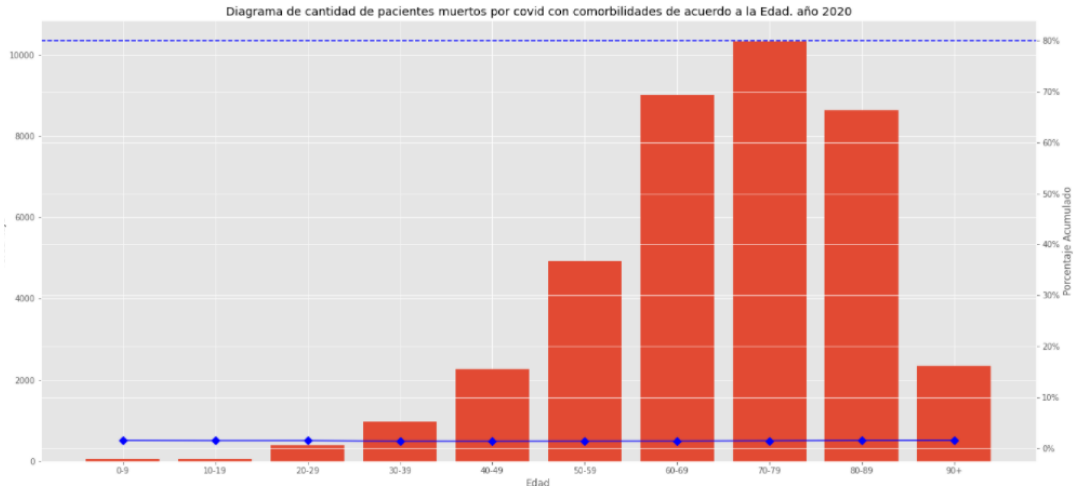
Omnibus:	0.055	Durbin-Watson:	0.197
Prob(Omnibus):	0.973	Jarque-Bera (JB):	0.095
Skew:	-0.062	Prob(JB):	0.954
Kurtosis:	2.695	Cond. No.	123.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Tabla 37. Validación de Pertinencia del modelo para Sexo y Edad. Fuente: Imagen generada desde Python tomando el set de datos descargado de INS.

Sin embargo, su f estadístico se redujo sustancialmente y su probabilidad prob (f-stactics: 0.00568) entró en los niveles de confianza 0.05 y 0.95, frente a la anterior que dio una posibilidad casi imposible de ocurrencia de 4.33e-27, siendo la que más se aproxima a la realidad. Por lo que hasta este punto concluimos basado en los resultados de la Tabla 37, que el Sexo, no es factor determinante de la muerte por COVID-19.

En el caso de la edad, tiene una correlación baja pero que sirve probar la relación, que se presenta en el *dataset* de comorbilidades, donde vemos que en la medida que aumenta la edad, mayor probabilidad de tener comorbilidades, que generen la muerte por COVID-19, como lo podemos apreciar al visualizar en el Grafica 21, de comorbilidades por edad.



Grafica 21. Diagrama estimativo, porcentual izado y totalizado de Pacientes Muertos por covid-19 en Colombia, Diseñado en Python, fuente: Dataset Descargado INS.

Teniendo presente este problema procedimos a estudiar las asociadas al paciente.

7.4.2. MODELO ESTADISTICO DE LAS COMORBILIDADES DEL COVID

Dado que obtuvimos muy buenos datos luego de realizar el modelo Estadístico del *dataset* de comorbilidades de los Estados Unidos, procedemos a realizar el proceso con las variables que está estudiando en Colombia la INS, por lo que tomando los datos de su página oficial podemos copiarlos y convertirlos en un *dataset* para poder realizar la investigación de las comorbilidades Colombia, y es la que podemos apreciar en la Tabla 38.

	Edad	CerebroVascular	HTA	DM	Renal	Tiroides	Obesidad	Fumar	Cardiaca	Respiratoria	Cancer	Autoinmune	Vih	Otros	Ninguno	En_estudio	Total_Fallecidos
0	0-9	0	1	1	0	1	1	0	7	1	3	1	1	12	3	17	49
1	10-19	0	2	2	5	3	3	0	0	4	2	0	1	8	0	18	48
2	20-29	5	22	25	15	6	43	2	12	14	29	14	11	36	4	154	392
3	30-39	13	59	67	40	18	119	1	27	35	45	13	19	38	10	460	964
4	40-49	9	214	231	99	42	304	5	55	58	47	26	28	82	10	1065	2275
5	50-59	38	640	536	290	108	439	27	138	136	158	53	30	160	14	2155	4922
6	60-69	94	1398	1110	554	248	530	59	394	402	255	53	23	297	18	3574	9009
7	70-79	174	1946	1066	635	279	322	59	650	683	296	57	11	380	13	3761	10332
8	80-89	194	1684	710	483	246	141	37	626	837	239	30	5	448	8	2946	8634
9	90+	52	494	150	129	71	23	3	190	272	51	7	0	133	3	764	2342

Tabla 38. Rango de edad y Fallecimientos por Comorbilidad. Fuente: Imagen generada desde Python tomando el set de datos copiados de INS.

Con lo que procedemos a estudiar las variables numéricas, esto con la intención de saber cuál es incidencia en el total de los datos Tabla 39.

	CerebroVascular	HTA	DM	Renal	Tiroides	Obesidad	Fumar	Cardiaca	Respiratoria	Cancer	Autoinmune	Vih	Otros	Ninguno	En_estudio	Total_Fallecidos
count	10.000000	10.000000	10.00000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.00000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
mean	57.900000	646.000000	389.80000	225.000000	102.200000	192.500000	19.300000	209.900000	244.200000	112.50000	25.400000	12.900000	159.400000	8.300000	1491.400000	3896.700000
std	72.760566	753.319911	437.33938	247.110412	112.670414	193.100003	24.436084	255.630571	303.609325	113.45704	22.126907	11.445038	160.620464	5.755191	1488.946249	4035.810398
min	0.000000	1.000000	1.00000	0.000000	1.000000	1.000000	0.000000	0.000000	1.000000	2.00000	0.000000	0.000000	8.000000	0.000000	17.000000	48.000000
25%	6.000000	31.250000	35.50000	21.250000	9.000000	28.000000	1.250000	15.750000	19.250000	33.00000	8.500000	2.000000	36.500000	3.250000	230.500000	535.000000
50%	25.500000	354.000000	190.50000	114.000000	56.500000	130.000000	4.000000	96.500000	97.000000	49.00000	20.000000	11.000000	107.500000	9.000000	914.500000	2308.500000
75%	83.500000	1208.500000	666.50000	434.750000	211.500000	317.500000	34.500000	343.000000	369.500000	218.75000	47.250000	22.000000	262.750000	12.250000	2748.250000	7706.000000
max	194.000000	1946.000000	1110.00000	635.000000	279.000000	530.000000	59.000000	650.000000	837.000000	296.00000	57.000000	30.000000	448.000000	18.000000	3761.000000	10332.000000

Tabla 39. Estadística de las comorbilidades para el año 2020. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

Ahora procedemos a visualizar el comportamiento de la primera variable que vamos a realizarle la regresión, la de las personas enfermas que murieron por comorbilidades asociadas a los problemas cerebrovasculares Grafica 22.

Grafica 22. Distribución de Muertes por comorbilidad Cerebrovascular contra el total de fallecidos por el rango de edad respectivo. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

En este caso, sin el grafico de regresión ya podemos identificar que la distribución corresponde a una recta que inicia en 0 y va aumentando, de acuerdo con los datos que van apareciendo de nuevos casos de enfermos con problemas cerebrovasculares, por lo que a primera vista tenemos un buen regresor.

Una de las herramientas que incluiremos en este estudio, es el de correlación de Pearson que nos indica como la media de la variable de respuesta y, se relaciona de forma lineal con las variables regresoras, es decir, la relación entre la cantidad de muertos con respecto a las que presentan esta comorbilidad, con lo cual buscamos obtener la medición de esa línea de regresión de la población que estamos estudiando.

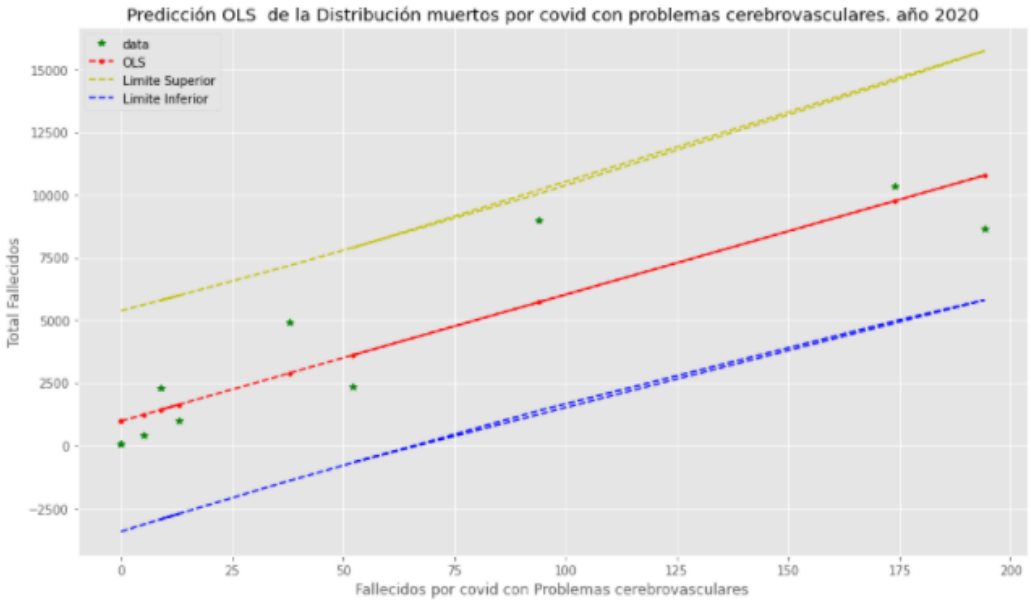
Dado a que necesitamos comparar la medición del modelo, en esta parte del estudio incluimos la correlación simple de Pearson, que nos mide también la correlación, con lo que podemos comprobar de dos formas que la correlación evidentemente es alta, por lo que nos muestra un coeficiente de correlación muy cercana al 91.08% para el caso cerebrovascular, en la Tabla 40.

```
...                            OLS Regression Results
==============================================================================
Dep. Variable:       Total_Fallecidos   R-squared:                       0.830
Model:                            OLS   Adj. R-squared:                  0.808
Method:                 Least Squares   F-statistic:                     38.96
Date:                Mon, 01 Nov 2021   Prob (F-statistic):           0.000248
Time:                        00:15:37   Log-Likelihood:                -87.843
No. Observations:                  10   AIC:                             179.7
Df Residuals:                       8   BIC:                             180.3
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        971.4871    729.251      1.332      0.220    -710.168    2653.143
CerebroVascular   50.5218      8.094      6.242      0.000      31.856      69.187
==============================================================================
Omnibus:                        1.960   Durbin-Watson:                   0.833
Prob(Omnibus):                  0.375   Jarque-Bera (JB):                1.075
Skew:                           0.773   Prob(JB):                        0.584
Kurtosis:                       2.567   Cond. No.                         118.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Tabla 40. Validación de Colinealidad de los datos. Fuente: Imagen generada desde Python tomando el set de datos descargado de DANE.

Dados los buenos resultados tanto en la regresión de Pearson como en los mínimos cuadrados, podemos proceder a realiza la gráfica de la regresión OLS. Como podemos visualizar en la Grafica 23.



Grafica 23. Modelamiento de OLS. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

Aun cuando la estimación OLS demuestra que los datos no presentan una distribución lineal, si es lineal en parámetros, por lo que no tenemos ningún problema para generar los valores de la predicción OLS, por medio de los intervalos de predicción que se obtienen con la siguiente ecuación, en la cual se relaciona el valor esperado en y, por medio del intervalo promedio.

$$\hat{y} \pm t_{\alpha/2,n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

Posteriormente solicitamos al modelo de *Statsmodel* que nos presente los valores predichos en el intervalo, es decir que nos enseñe los parámetros que se deben tener en cuenta, tales como el punto de *intercept*, que es donde debe iniciar , los errores estándar, y los valores predichos en la estimación, con los mismos podemos ver que el error estándar no es tan grande, y que los valores predichos inician en el *intercept* y comienza a predecir, como lo podemos apreciar en la Tabla 41.

```
Parameters:        Intercept          971.487115
CerebroVascular     50.521811
dtype: float64
Standard errors:   Intercept          729.250836
CerebroVascular      8.094248
dtype: float64
Predicted values:  [  971.48711459   971.48711459  1224.09617205  1628.27066399
  1426.18341802  2891.3159513   5720.53739488  9762.28231427
 10772.71854412  3598.62131219]
```

Tabla 41. Linealidad Paramétrica OLS. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

7.4.2.1. ENTRENAMIENTO DE LA MAQUINA PARA LA PREDICCIÓN DE LAS REGRESIONES ML.

Por medio del paquete especializado para predecir valores le pediremos a la máquina que oculte el 20% de los datos de las dos columnas, y que entrene con el 80%, para que con el mismo realice una predicción del comportamiento de la estimación OLS.

Por consiguiente, creamos el modelo y le agregamos una variable constante al set de entrenamiento para que pueda crear el *intercept* β0 con el cual inicie la ecuación.

Luego le declaramos cuales son las variables

endógena => y (Muertos por Covid-19)

exógena=> X (Comorbilidad Cerebrovascular)

y recreamos el modelo con otro *dataframe* al cual le llamamos mod11 para recordar que proviene del mod1, y lo imprimimos.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.743
Model:	OLS	Adj. R-squared:	0.701
Method:	Least Squares	F-statistic:	17.37
Date:	Sun, 31 Oct 2021	Prob (F-statistic):	0.00589
Time:	23:58:34	Log-Likelihood:	-70.936
No. Observations:	8	AIC:	145.9
Df Residuals:	6	BIC:	146.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1179.3562	899.530	1.311	0.238	-1021.715	3380.428
x1	47.0204	11.281	4.168	0.006	19.418	74.623

Omnibus:	2.393	Durbin-Watson:	1.166
Prob(Omnibus):	0.302	Jarque-Bera (JB):	1.264
Skew:	0.920	Prob(JB):	0.532
Kurtosis:	2.364	Cond. No.	102.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Tabla 42. Definición de Variables Entrenamiento. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

Como podemos ver Tabla 42, una vez entrenada la máquina, con las variables bien definidas, no se generan problemas de colinealidad, y solamente se desvió la predicción, de la real un 8.7%, por lo que la maquina con un 20% menos de datos nos está realizando una muy buena estimación de la correlación real.

Asimismo, cuando hacemos el modelo normal en *StatsModel* procedemos a solicitarle el fit del modelo y que nos imprima el resumen Tabla 43.

```
Parameters:        Intercept          971.487115
CerebroVascular     50.521811
dtype: float64
Standard errors:   Intercept          729.250836
CerebroVascular      8.094248
dtype: float64
Predicted values:  [  971.48711459   971.48711459  1224.09617205  1628.27066399
  1426.18341802  2891.3159513   5720.53739488  9762.28231427
 10772.71854412  3598.62131219]
```

Tabla 43. Resumen de resultados Comparativos StatsModel Sickit Learn, Fuente: generada desde Python Tomando Set de Datos Copiado del INS.

En este caso, podemos ver que la maquina empieza a realizar la predicción aproximándose al valor de la intercesión 971.487115, la cual aproxima a 971.4811459 y así sucesivamente con los otros puntos.

Otra parte importante son los intervalos de confianza para los coeficientes, el cual se realiza por medio de una prueba ***t-student***, para saber cuáles de los residuos del modelo son confiables, le definimos al modelo cuales son los puntos que se encuentran más cerca de la recta, en este caso definimos esos intervalos sean entre 0.05 de distancia entre la recta y los puntos, para que los mismos puedan ser confiables.

Una vez definidos cuales son los intervalos de confianza para los coeficientes, procedemos a definir los intervalos de confianza para las predicciones, la cual es la diferencia entre los intervalos de confianza para los coeficientes, es decir en términos porcentuales es el 95%, si definimos el 5% para los intervalos de los coeficientes Tabla 44.

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	3624.415752	701.097253	1908.892576	5339.938927	-1519.784590	8768.616093
1	1179.356212	899.530390	-1021.715357	3380.427780	-4146.476441	6505.188864
2	1179.356212	899.530390	-1021.715357	3380.427780	-4146.476441	6505.188864
3	10301.309111	1769.095065	5972.489435	14630.128787	3800.658591	16801.959630
4	1602.539593	839.605061	-451.899979	3656.979166	-3664.385444	6869.464631
5	2966.130491	713.689320	1219.795637	4712.465344	-2188.426989	8120.687971
6	5599.271534	858.711575	3498.080007	7700.463061	313.934910	10884.608157
7	1790.621097	815.619067	-205.126864	3786.369057	-3453.688897	7034.931090

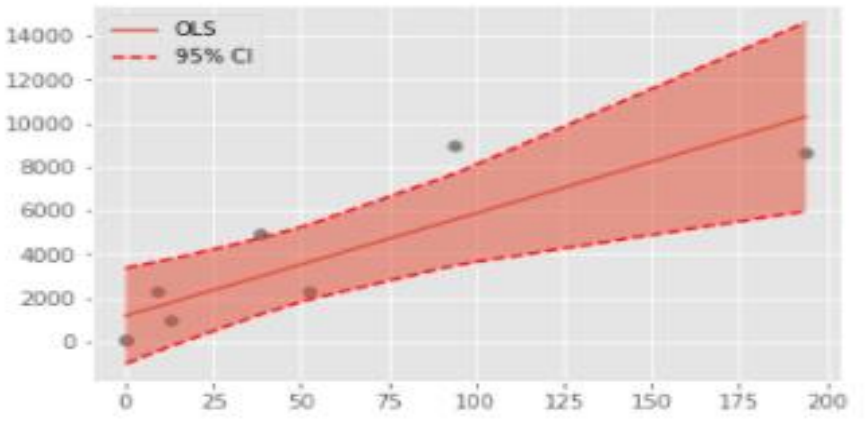
Tabla 44. Definición de Intervalos de Confianza para la Predicción. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

Del mismo modo como definimos los intervalos de confianza, procedemos a generar los limites superior e inferior de la recta para que todos aquellos puntos que se encuentran por fuera de la muestra no sean estimados Tabla 45.

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper	x	y
1	1179.356212	899.530390	-1021.715357	3380.427780	-4146.476441	6505.188864	0.0	48
2	1179.356212	899.530390	-1021.715357	3380.427780	-4146.476441	6505.188864	0.0	49
4	1602.539593	839.605061	-451.899979	3656.979166	-3664.385444	6869.464631	9.0	2275
7	1790.621097	815.619067	-205.126864	3786.369057	-3453.688897	7034.931090	13.0	964
5	2966.130491	713.689320	1219.795637	4712.465344	-2188.426989	8120.687971	38.0	4922
0	3624.415752	701.097253	1908.892576	5339.938927	-1519.784590	8768.616093	52.0	2342
6	5599.271534	858.711575	3498.080007	7700.463061	313.934910	10884.608157	94.0	9009
3	10301.309111	1769.095065	5972.489435	14630.128787	3800.658591	16801.959630	194.0	8634

Tabla 45. Definición de Limites Superior e Inferior de la Recta. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

Como podemos visualizar estos límites generan cortes entre los puntos que se generaron en los intervalos de confianza, sin embargo, podremos ver mejor lo que sucede con los puntos por medio de la Grafica 24.



Grafica 24. Visualización de los límites. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

Dado a que solamente 997 unidades de la predicción que se alejan de los datos reales, frente a los cerca de 37000 muestras totales, podemos definir que hemos logrado una muy buena predicción, de la regresión OLS.

Luego de haber realizado la regresión OLS simple y la que entrenamos con la máquina, ya podemos consolidar los resultados obtenidos Tabla 46.

Dep. Variable	Ind. Variable	R^2 OLS StatsModel	R^2 OLS Predicción ML(Scikit Learn)	Diferencia
Total_Fallecidos	Cerebrovascular	0.83	0.743	0.087
Total_Fallecidos	Diabetes Mellitus	0.958	0.935	0.023
Total_Fallecidos	Sistema Renal	0.995	0.993	0.002
Total_Fallecidos	Tiroides	0.99	0.984	0.006
Total_Fallecidos	Obesidad	0.479	0.452	0.027
Total_Fallecidos	Humos Toxicos	0.942	0.915	0.027
Total_Fallecidos	Problemas Cardiacos	0.892	0.832	0.06
Total_Fallecidos	Problemas Sistema Respiratorio	0.801	0.714	0.087
Total_Fallecidos	Cancer y Tumor	0.987	0.983	0.004
Total_Fallecidos	VIH/SIDA	0.053	0.098	-0.045
Total_Fallecidos	Autoinmunidad	0.737	0.652	0.085
Total_Fallecidos	Otras	0.92	0.896	0.024

Tabla 46. Consolidado de Predicción StatsModels y scikit Learn. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

Con lo que hemos logrado muy buenos datos luego de analizar todas las comorbilidades, y tenemos presente, que las comorbilidades son un factor real de muerte por COVID-19.

7.8.1. PRONOSTICO DE LA MORTADAD DE COVID-19 ASOCIADO A LAS COMORBILIDADES

Como ya lo mencionamos anteriormente, no contamos con un *dataset* en forma de serie temporal especifico de las comorbilidades asociadas al COVID-19, sin embargo, contamos con la información de las comorbilidades que están estudiando las autoridades de salud de Colombia y el consolidado de los casos en www.ins.gov.co[11].

[11] https://www.ins.gov.co/Noticias/Paginas/Coronaviruss.aspx

De modo que con el mismo *dataset* que creamos a partir de las gráficas estadísticas de comorbilidad de la INS, podemos ver la mayoría de los casos se encuentran en estudio, sin embargo, ya hay una tendencia en la hipertensión Arterial y la Diabetes Mellitus, como las comorbilidades que más presentan los fallecidos por COVID-19. A partir de aquí ya podemos crear nuestro *dataset*, y omitir todos pasos que anteriormente realizamos. Sin embargo, si es necesario para continuar con el estudio poder obtener las tasas que representa cada comorbilidad, con respecto al total de casos por edad, pues con estas será que sabremos, qué cantidad de las defunciones generales de COVID-19 del *dataset* de defunciones y podemos asociar cada comorbilidad, visualizando el *dataset* de comorbilidades solamente cuenta con 10 filas y 28 columnas, sin embargo imprimiremos el resumen con su llave primaria que es la edad, la cual asociaremos en los pronósticos y de las tasas porcentuales, con lo que podremos filtrar por la edad y comorbilidad asociada para entrar a pronosticar como está organizada la Tabla 47.

	Rango_Edad	Tasa_%-CerebroVascular	Tasa_%-DiabMellit	Tasa_%-Renal	Tasa_%-Tiroides	Tasa_%-Obesidad	Tasa_%-Fumar	Tasa_%-Cardiaca	Tasa_%-Respiratoria	Tasa_%-Cancer	Tasa_%-Autoinmune	Tasa_%-Vih	Tasa_%-Otros
0	0-9	0.000000	0.020408	0.000000	0.020408	0.020408	0.000000	0.142857	0.020408	0.061224	0.020408	0.020408	0.244898
1	10-19	0.000000	0.041667	0.104167	0.062500	0.062500	0.000000	0.000000	0.083333	0.041667	0.000000	0.020833	0.166667
2	20-29	0.012755	0.063776	0.038265	0.015306	0.109694	0.005102	0.030612	0.035714	0.073980	0.035714	0.028061	0.091837
3	30-39	0.013485	0.069502	0.041494	0.018672	0.123444	0.001037	0.028008	0.036307	0.046680	0.013485	0.019710	0.039419
4	40-49	0.003956	0.101538	0.043516	0.018462	0.133626	0.002198	0.024176	0.025495	0.020659	0.011429	0.012308	0.036044
5	50-59	0.007720	0.108899	0.058919	0.021942	0.089191	0.005486	0.028037	0.027631	0.032101	0.010768	0.006095	0.032507
6	60-69	0.010434	0.123210	0.061494	0.027528	0.058830	0.006549	0.043734	0.044622	0.028305	0.005883	0.002553	0.032967
7	70-79	0.016841	0.103175	0.061460	0.027003	0.031165	0.005710	0.062911	0.066105	0.028649	0.005517	0.001065	0.036779
8	80-89	0.022469	0.082233	0.055942	0.028492	0.016331	0.004285	0.072504	0.096942	0.027681	0.003475	0.000579	0.051888
9	90+	0.022203	0.064048	0.055081	0.030316	0.009821	0.001281	0.081127	0.116140	0.021776	0.002989	0.000000	0.056789

Tabla 47. Dataset por edad con tasas por Comorbilidad. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS.

Dado que el dataset de mortandad total de COVID-19 cuenta con 3 columnas y 4487 filas, si procuramos unir los dos en uno solo de forma manual, tendríamos que realizar 125636 repeticiones, si lo intentáramos por medio de un ciclo repetitivo doble y condicionado, tendríamos el problema de que no son del mismo tamaño y no todas las semanas mueren las personas de la misma edad, por lo que no se cumplirían las condiciones y sacaría del ciclo sin terminar. Por eso nuestra mejor opción es aprovechar la forma en la que están diseñados, pues en el caso del *dataframe* de comorbilidades se tiene una columna con valores únicos, que es el conjunto de la edad y funciona perfectamente como una llave primaria PK, y es la misma columna que se encuentra en el de casos de fallecidos por COVID-19, con lo que podemos unir ambas tablas por medio de una consulta SQL.

Con lo que ahora ya tenemos nuestros dos *dataframe* unidos en uno solo, que nos dice la tasa, y con el cual podemos realizar los respectivos pronósticos y análisis, en este caso, una de las dos comorbilidades que más han encontrado las autoridades e investigadores de salud nacional en Colombia, como lo es la diabetes mellitus.

Debemos tener presente que las tasas las tenemos en términos de entre 0 y 1 por lo que es necesario normalizar la columna de fallecidos en la misma escala, para no tener errores grandes en la medición, por lo que utilizaremos la normalización Min-Max, para que queden en los mismos términos Tabla 48.

	Fecha de muerte	Rango_Edad	Fallecidos	CerebroVascular	Tasa_%-CerebroVascular	FallCerVas
0	1/1/2021 0:00:00	0-9	0.000000	0	0.000000	0.000000
1	1/2/2021 0:00:00	0-9	0.000000	0	0.000000	0.000000
2	1/7/2021 0:00:00	0-9	0.000000	0	0.000000	0.000000
3	1/9/2021 0:00:00	0-9	0.000000	0	0.000000	0.000000
4	10/12/2020 0:00:00	0-9	0.000000	0	0.000000	0.000000
...	...	...	...	...	...	...
4482	9/7/2021 0:00:00	90+	0.085714	52	0.022203	0.190314
4483	9/8/2020 0:00:00	90+	0.102857	52	0.022203	0.228376
4484	9/8/2021 0:00:00	90+	0.057143	52	0.022203	0.126876
4485	9/9/2020 0:00:00	90+	0.045714	52	0.022203	0.101501
4486	9/9/2021 0:00:00	90+	0.005714	52	0.022203	0.012688

4487 rows × 6 columns

Tabla 48. Normalización de Columna Fallecidos. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS.

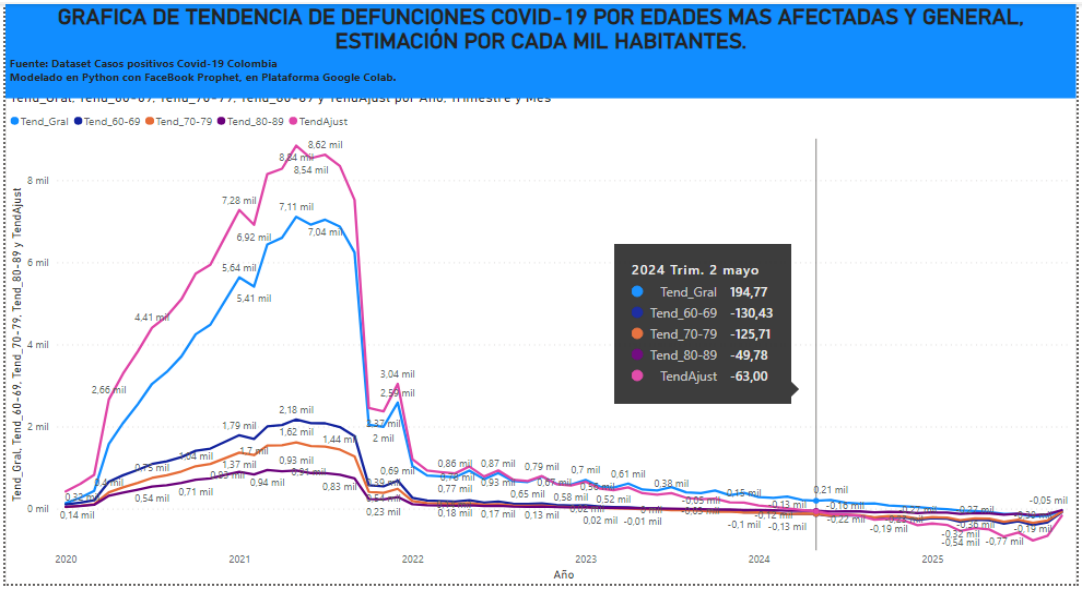
Una vez normalizada la columna de fallecidos procedimos a crear el *dataframe* del exceso de mortandad de fallecidos por COVID-19 con comorbilidad de diabetes mellitus y así sucesivamente los consolidamos, teniendo presente que este es el mismo procedimiento que realizamos para todas las comorbilidades. Procedemos a crear el modelo de pronóstico y así determinar si habrá aumento o descenso de muertes por COVID-19 en personas que presentan a la Diabetes Mellitus como su comorbilidad secundaria y también procedemos a crear el futuro que deseamos que pronostique en este caso, los próximos 100 periodos, con una frecuencia semanal, que como ya lo mencionamos en el punto 6.7, el día es muy disperso, el año es demasiado resumido, por lo que, las semanas y los meses son periodos temporales, en lo que mejor se puede consolidar la información Tabla 49.

	Fecha de muerte	Rango_Edad	Fallecidos	DM	Tasa_%-DiabMellit	FallDiaMel
0	1/1/2021 0:00:00	0-9	0.000000	1	0.020408	0.000000
1	1/2/2021 0:00:00	0-9	0.000000	1	0.020408	0.000000
2	1/7/2021 0:00:00	0-9	0.000000	1	0.020408	0.000000
3	1/9/2021 0:00:00	0-9	0.000000	1	0.020408	0.000000
4	10/12/2020 0:00:00	0-9	0.000000	1	0.020408	0.000000
...	...	...	...	...	...	...
4482	9/7/2021 0:00:00	90+	0.085714	150	0.064048	0.548981
4483	9/8/2020 0:00:00	90+	0.102857	150	0.064048	0.658778
4484	9/8/2021 0:00:00	90+	0.057143	150	0.064048	0.365988
4485	9/9/2020 0:00:00	90+	0.045714	150	0.064048	0.292790
4486	9/9/2021 0:00:00	90+	0.005714	150	0.064048	0.036599

4487 rows × 6 columns

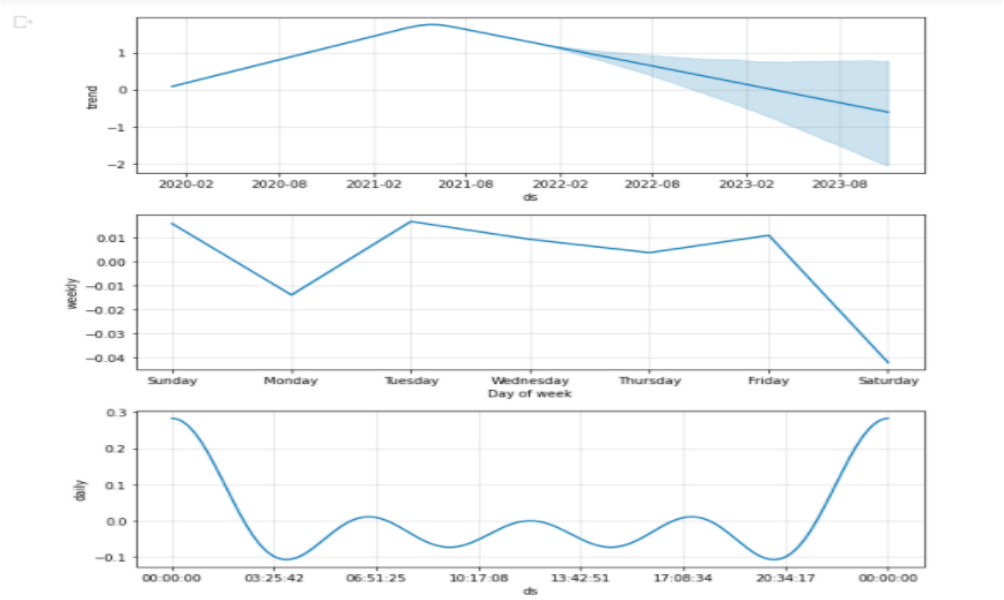
Tabla 49. Pronóstico de Diabetes Mellitus. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS.

Sin embargo, gracias la unificación de estas tablas, no solamente podemos filtrar por las comorbilidades, también lo podemos hacerlo por las edades y generar sus pronósticos, en este caso, las edades que más sufren la muerte por el COVID-19 y su respectivo pronostico se evidencia en la Grafica 25.



Grafica 25. Pronóstico de Muertes por COVID-19 por las edades más afectadas con Comorbilidades, estimación por cada mil habitantes. Fuente: Imagen generada de Power BI, modelado en Python con Facebook Prophet, en la plataforma Google Colab, tomando el set de datos

Al ajustar el pronóstico con la estacionalidad del COVID-19, la cual es estimada por la librería *Facebook Prophet (FP)* en aproximadamente semana y media, es cuando podemos identificar que para el segundo trimestre del 2024 se aproximan la mayoría de las estimaciones a 0, sin embargo, la tendencia general por cada mil habitantes es que llegue a 0 en el primer trimestre del 2025, esto porque esa estimación no está tomando en cuenta la estacionalidad del COVID-19. Adicionalmente podemos validar que a pesar de que las muertes por COVID-19 podrían a 0 en el año 2024, las muertes de enfermos de COVID-19 con comorbilidades tales como Diabetes Mellitus entre otras podrían llegar a 0 entre 4 trimestre de 2023 y el primero de 2024. Como bien lo muestra la gráfica de tendencia, el pronóstico lo inicio a partir de febrero de 2021 Grafica 26.



Grafica 26. Tendencia de Muertes Diabetes Mellitus. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS.

En la validación cruzada, podemos ver que hemos reducido la diferencia entre “y” y “yhat” Tabla 50.

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
0	2020-09-11	1.468695	0.137423	2.698077	0.996627	2020-09-10
1	2020-09-12	1.443698	0.170445	2.697410	1.059847	2020-09-10
2	2020-09-13	1.393368	0.109413	2.649665	1.123933	2020-09-10
3	2020-09-14	1.394402	0.150732	2.614659	1.102672	2020-09-10
4	2020-09-15	1.464623	0.251292	2.797051	1.237653	2020-09-10
...	...	...	...	...	...	...
641	2021-12-05	2.137777	0.934871	3.340145	3.185138	2020-12-09
642	2021-12-06	2.143253	1.045276	3.287184	3.492454	2020-12-09
643	2021-12-07	2.175011	1.029163	3.387225	2.757831	2020-12-09
644	2021-12-08	2.170441	1.001528	3.344821	0.688489	2020-12-09
645	2021-12-09	2.221296	1.077063	3.342164	0.234513	2020-12-09

646 rows × 6 columns

Tabla 50. Validación Cruzada del Modelo. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS.

En términos generales, gracias a la unión de las 2 tablas, la normalización de la columna de fallecidos y un pequeño cambio de estacionalidad, hemos logrado reducir sustancialmente el error de las estimaciones Tabla 51.

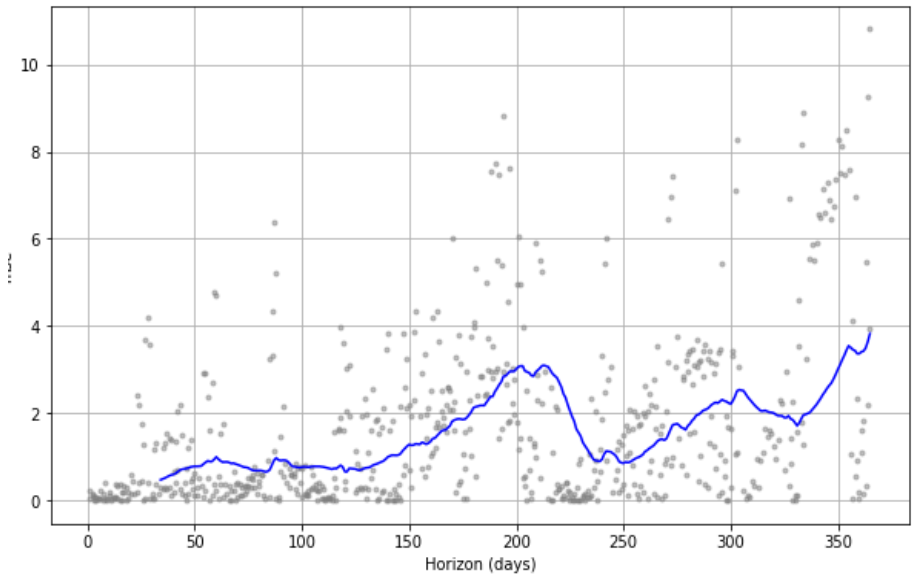
INFO:fbprophet:Skipping MAPE because y close to 0

	horizon	mse	rmse	mae	mdape	coverage
0	34 days	0.465027	0.681929	0.494952	0.291986	0.921875
1	35 days	0.483766	0.695533	0.509534	0.303986	0.921875
2	36 days	0.505221	0.710789	0.525163	0.307966	0.921875
3	37 days	0.531422	0.728987	0.546987	0.318277	0.906250
4	38 days	0.556721	0.746137	0.569254	0.328075	0.890625
...	...	...	...	...	...	...
321	361 days	3.403597	1.844884	1.617829	0.853807	0.453125
322	362 days	3.422542	1.850011	1.625875	0.853807	0.437500
323	363 days	3.493542	1.869102	1.647635	0.954196	0.421875
324	364 days	3.641078	1.908161	1.687067	1.250054	0.390625
325	365 days	3.823316	1.955330	1.730609	1.901248	0.359375

326 rows × 6 columns

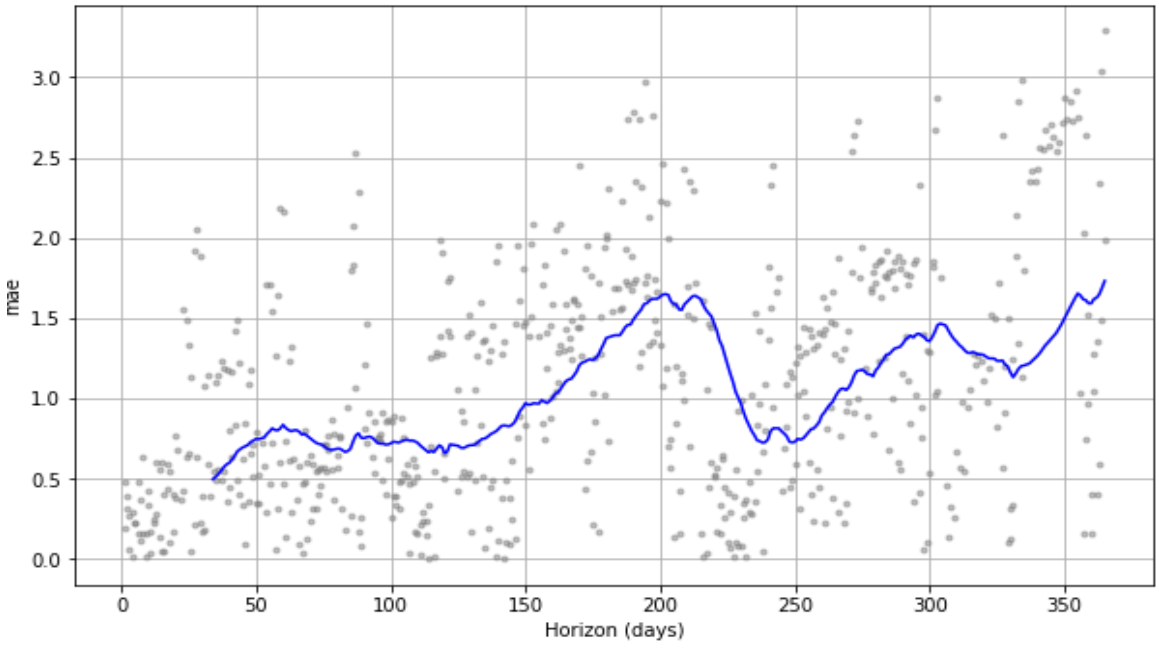
Tabla 51. Validación de error. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS.

Inclusive el mismo programa indica que logramos reducir el error porcentual absoluto a 0, por lo que debemos centrarnos en las otras mediciones, como en este caso el Error Cuadrático Medio MSE, en el cual podemos ver la diferencia entre el estimado y el pronóstico en la Grafica 27.



Grafica 27. Error Cuadrático Medio MSE. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS.

Dado que en 365 días el error cuadrático medio MSE máximo encontrado es de 3.82 precisamente en el último punto de la medición. Por lo que la diferencia entre la estimación y la real es relativamente baja. En el error absoluto MAE, también podemos validar que es mínima la diferencia entre el "y" y "yhat", Grafica 28.



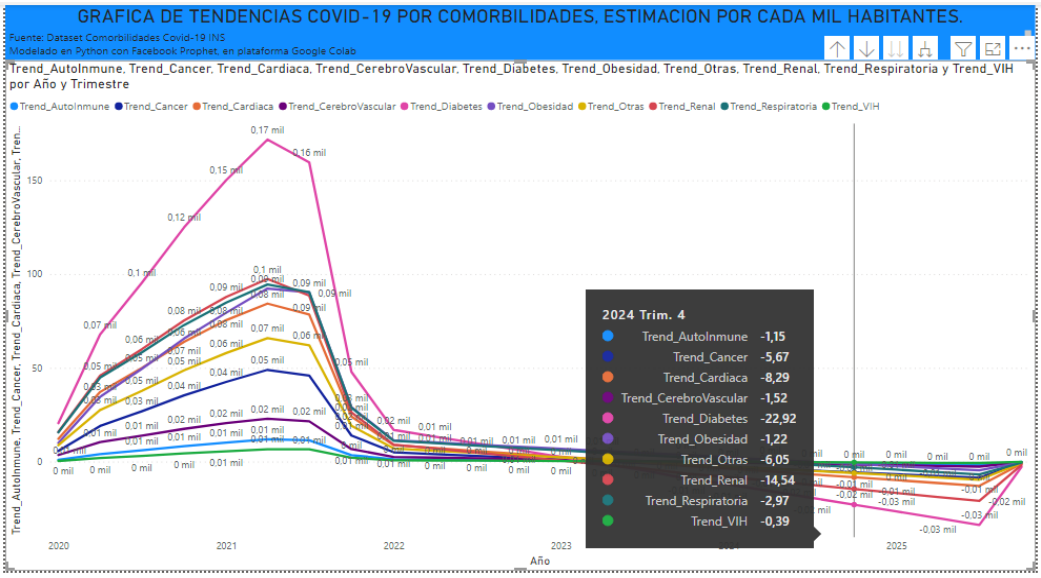
Grafica 28. Error Absoluto MAE. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de INS.

En este punto de la investigación hemos cumplido con los objetivos principales de la investigación, sin embargo, queremos ver si la maquina puede aprender a clasificar las comorbilidades, por lo que en el **Apéndice B**, con el *dataset* ofrecido por los Estados Unidos, en donde están clasificadas 23 comorbilidades, entrenaremos a la máquina,

para que aprenda a encontrar la enfermedad de base, de acuerdo a las comorbilidades de los pacientes, que han diagnosticado miles de médicos que han aportado, para la creación del *dataset* en el mencionado país. Posterior a esto, utilizando el mismo *dataset* de los Estados Unidos donde están clasificadas 23 comorbilidades y en el **Apéndice B** podemos visualizar el resto de las comorbilidades para saber cuándo llegaran a 0 muertes por comorbilidades asociadas al virus.

8. PRONOSTICO DE LAS MUERTES POR COVID-19 ASOCIADAS A LAS COMORBILIDADES.

Uno de los objetivos era generar un modelo de pronóstico para las muertes COVID-19 con comorbilidades, por lo que en la Grafica 29 de tendencias podemos evidenciar la estimación por cada mil habitantes, muestra que para para el último trimestre del 2024, todas las comorbilidades llegarían a 0 muertes.



Grafica 29. Tendencia de muertes Covid-19 por comorbilidades, estimación por cada 1000 habitantes, creada En Power Bi, Generada en Python, con la plataforma Google Colab, Fuente: Set De Datos Descargado de la INS.

Sin embargo, en el resumen, de las muertes por comorbilidades, tenemos que las llegadas 0 muertes van en un rango de entre el 1 trimestre del 2023 y 3 trimestre del 2024, Así que, de continuar así la tendencia, en un máximo 2 años el virus dejará su fuerte tasa de mortandad Tabla 52.

AÑO	TRIMESTRE	MES	DIA	TOTAL MUERTES	COMORBILIDAD
2023	I	MARZO	12	-0.005789555	Renal
2023	II	ABRIL	23	-0.016102458	Diabetes
2023	II	JUNIO	4	-0.001842709	Cancer
2023	III	DICIEMBRE	6	-0.0009553	AutoInmune
2023	III	AGOSTO	13	-0.00378877	Otras
2023	III	JULIO	23	-0.007851884	Cardiaca
2023	IV	DICIEMBRE	24	-0.000492944	VIH
2023	IV	NOVIEMBRE	5	-0.000308596	CerebroVascular
2024	II	ABRIL	21	-0.002406658	Respiratoria
2024	III	AGOSTO	18	-0.006633149	Obesidad

Tabla 52. Pronóstico de defunciones Covid-19 por comorbilidades, diseñado en Python con Facebook Prophet, modelado en Power Bi, Fuente: Dataset de caso positivos Covid-19 Colombia y Comorbilidades INS.

9. DESPLIEGUE

Para el despliegue del modelo final, que consta de los modelos de regresión para la mortalidad de Pacientes con comorbilidades CONTAGIADOS DE COVID-19 en COLOMBIA, el diseño de la visualización y las herramientas empleadas se muestra en la en la Figura 1 y se puede encontrar disponible en el siguiente link: https://github.com/milenabb88/ModeloDatosCovid

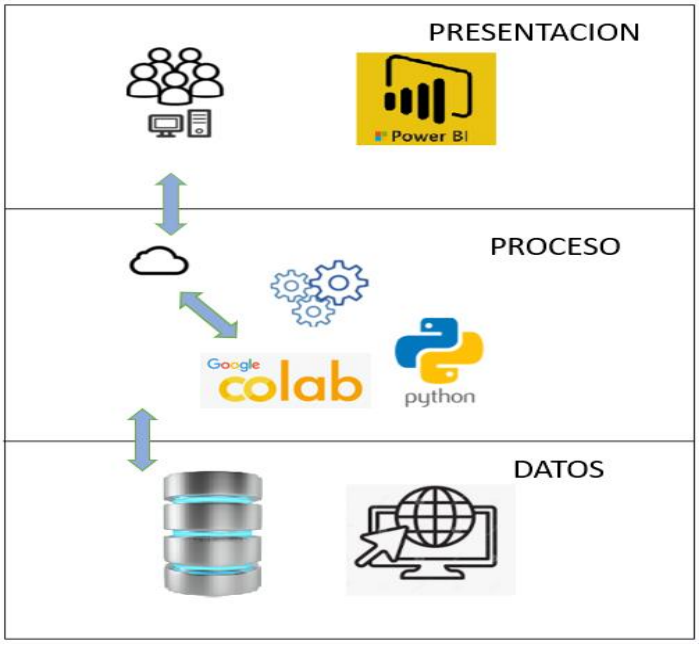


Figura 1. Despliegue del proceso. Fuente: Propia

10. CONCLUSIONES

Dado que desde el inicio se generaron unos objetivos claros como, generar modelos de datos de diferentes fuentes de información, en esta investigación se tuvieron en cuenta diferentes *datasets*, se realizó el respectivo trabajo de minería, y reclasificación para la homologación de estas y poder hacer que los datos interactuaran entre sí, pese a que llegaban de distintas fuentes.

Por otro lado, se entró a generar todo un preproceso cumpliendo con la metodología CRIPS-DM, pasando por análisis iniciales, generando una política seria para evaluar las regresiones, con las cuales pudiéramos afirmar que las variables guardaban una correlación, bien fuera media como en el caso de la edad, alta como en el caso de las comorbilidades o ninguna como sucedió con el sexo, en relación con la muerte por covid-19.

En este estudio se realizó un modelo de regresión en el cual sometimos las variables independientes del estudio inicial, para determinar su grado de correlación con respecto a las muerte por covid, es decir se intentó determinar si las variables sexo y edad, presentaban algún tipo de relación con respecto a la variable dependiente que es la muerte por covid, pese a que se esperaba un gran resultado con la edad, ya que es la más mencionada, en los medios, pudimos determinar que la variable edad, no era un factor determinante, para el fallecimiento por covid.

También pudimos realizar el estudio de tiempos de los nuevos contagios por departamento, el cual nos enseñó cómo se fue dispersando el virus por todo el país y cada cuanto aparecían nuevos casos, que fue fundamental para el estudio de la difusión viral, que fue complementado con la serie temporal de la sintomatología expresada por los pacientes, de esta forma se pudo determinar las personas de que edades fueron más expuestas al virus y su periodicidad por departamento.

Gracias a los buenos resultados obtenidos con los diferentes sets de datos, se logró que realizar los respectivos entrenamientos de la máquina, para que aprendiera a predecir en los diferentes procesos de Machine Learning, como lo fueron el aprendizaje supervisado en las regresiones, el aprendizaje reforzado en los pronósticos y el aprendizaje no supervisado, en las clasificaciones de los grupos de enfermedades de base a las cuales pertenece cada comorbilidad del COVID-19, con los cuales se generaron las diferentes estimaciones que ya se presentaron, convirtiéndose en una gran fuente de investigación no solamente para este trabajo sino para los múltiples estudios que se pueden llegar a realizar.

APENDICE A

A.1. CORRELACION MULTIPLE ENTRE LAS VARIABLES INDEPENDIENTES Y LA VARIABLE DEPENDIENTE.

Cabe mencionar que cuando tenemos un sistema de interacciones simplificado, como el de las comorbilidades, que pueden generar interacción directa entre las mismas, también podemos realizar el estudio de correlaciones entre las mismas variables puesto que hay pacientes que pueden tener múltiples comorbilidades, que como ya lo vimos al principio también son las enfermedades de base, y una morbilidad se puede convertir en otra, por lo que pueden interactuar entre ellas mismas, es decir entre comorbilidades, y entre las mismas pueden deteriorar la calidad de vida del paciente, por eso generamos el mismo procedimiento, que en la correlación simple, con la diferencia de que aquí mostramos todas las correlaciones en una matriz, por lo que creamos un cuadro, de correlaciones para poder simplificar el proceso.

Mediante una matriz *tidy que genera correlaciones por medio del método de Pearson, con lo que* podemos implementar el proceso, de consolidar todas las relaciones presentes, simplemente debemos tener presente que hay que resetear el índice como lo hacemos con la función *groupby*, para que se genere la tabla plana y así lograr que la matriz inicie a correlacionar automáticamente una a una todas las variables y realice el proceso de forma repetitiva hasta que culmine el proceso Tabla A1.

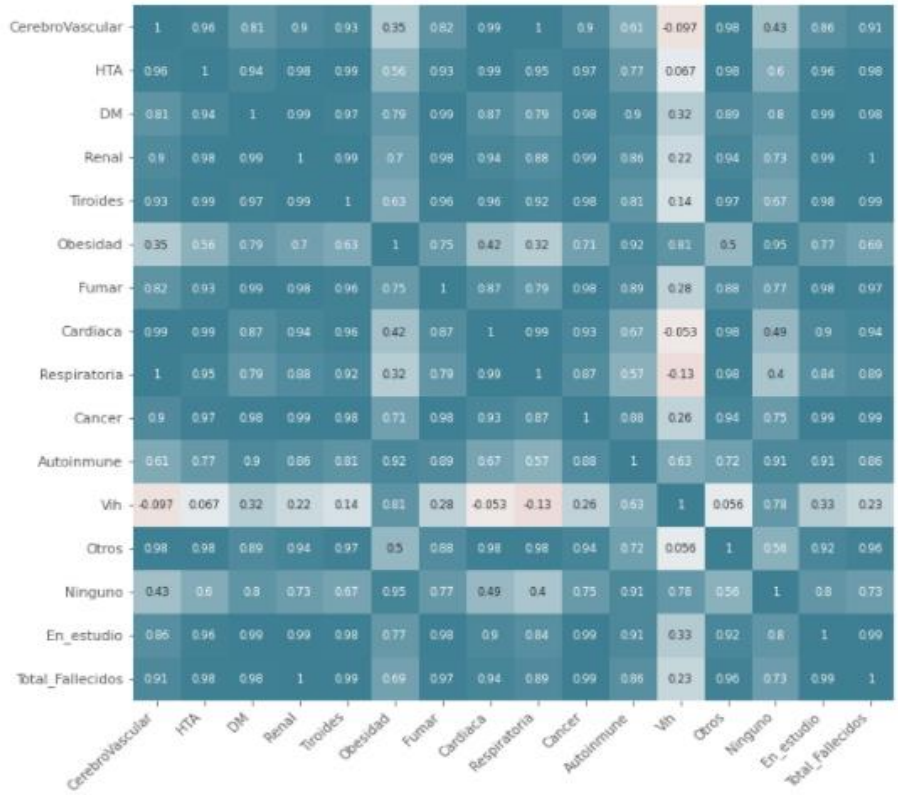
	variable_1	variable_2	r	abs_r
243	Total_Fallecidos	Renal	0.997576	0.997576
63	Renal	Total_Fallecidos	0.997576	0.997576
128	Respiratoria	CerebroBascular	0.995766	0.995766
8	CerebroBascular	Respiratoria	0.995766	0.995766
244	Total_Fallecidos	Tiroides	0.994931	0.994931
...	...	...	...	...
27	HTA	Vih	0.067065	0.067065
203	Otros	Vih	0.055752	0.055752
188	Vih	Otros	0.055752	0.055752
183	Vih	Cardiaca	-0.053286	0.053286
123	Cardiaca	Vih	-0.053286	0.053286

240 rows × 4 columns

Tabla A1. Matriz de Correlación de Variables. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

En este caso obtuvimos que se repitió el proceso 240 veces frente a las 10 variables en el proceso de comparar la variable 1 contra la dos obtener su relación y convertir esa relación en un valor absoluto.

Por medio del siguiente *heatmap* obtenemos nuestra matriz de correlaciones, o la matriz de confusión Accuracy Grafica A1.



Grafica A1. Matriz de Correlación. Fuente: Imagen generada desde Python tomando el set de datos copiado de INS.

Logramos visualizar Grafica A1, en qué casos es más fuerte la correlación de Pearson, y en el caso de la obesidad y el VIH/SIDA, podríamos determinar, dado que sus relaciones son relativamente bajas tanto en la estimación OLS, como en la correlación de Pearson, las defunciones de los enfermos de COVID-19 que padecen de VIH/SIDA y Obesidad, son causadas por la misma comorbilidad que por adquisición del virus.

APENDICE B

CLASIFICACION AUTOMATICA DE LOS GRUPOS DE ENFERMEDADES DATASET USA.

Dado que el *dataset* de los Estados unidos tiene la clasificación internacional de enfermedades, procederemos a entrenar la máquina para que realice la clasificación y genere un diagnóstico del grupo al que pertenece la enfermedad que posee el paciente. Luego de hacer la reducción del *dataframe* y la respectiva limpieza de datos, logramos, reducirlo con las columnas que necesitamos en la Tabla B1.

	Data As Of	Start Date	End Date	Group	Year	Month	State	Condition Group	Condition	ICD10_codes	Age Group	COVID-19 Deaths	Number of Mentions	Flag
0	08/29/2021	01/01/2020	08/28/2021	By Total	NaN	NaN	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	520.0	540.0	NaN
1	08/29/2021	01/01/2020	08/28/2021	By Total	NaN	NaN	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	25-34	2348.0	2405.0	NaN
2	08/29/2021	01/01/2020	08/28/2021	By Total	NaN	NaN	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	35-44	6191.0	6367.0	NaN
3	08/29/2021	01/01/2020	08/28/2021	By Total	NaN	NaN	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	45-54	17515.0	18048.0	NaN
4	08/29/2021	01/01/2020	08/28/2021	By Total	NaN	NaN	United States	Respiratory diseases	Influenza and pneumonia	J09-J18	55-64	42471.0	43677.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
285655	08/29/2021	04/01/2021	04/30/2021	By Month	2021.0	4.0	Puerto Rico	COVID-19	COVID-19	U071	All Ages	197.0	197.0	NaN
285656	08/29/2021	05/01/2021	05/31/2021	By Month	2021.0	5.0	Puerto Rico	COVID-19	COVID-19	U071	All Ages	184.0	184.0	NaN
285657	08/29/2021	06/01/2021	06/30/2021	By Month	2021.0	6.0	Puerto Rico	COVID-19	COVID-19	U071	All Ages	35.0	35.0	NaN
285658	08/29/2021	07/01/2021	07/31/2021	By Month	2021.0	7.0	Puerto Rico	COVID-19	COVID-19	U071	All Ages	29.0	29.0	NaN
285659	08/29/2021	08/01/2021	08/28/2021	By Month	2021.0	8.0	Puerto Rico	COVID-19	COVID-19	U071	All Ages	190.0	190.0	NaN

Tabla B1. Dataset inicial de Comorbilidades. Fuentes: Imagen generada desde Google Colab tomando el set de datos descargado de CDC.

Sin embargo, para visualizar las comorbilidades, procedemos a reducir las columnas, de modo que podamos ver toda la composición, por lo que procedemos a generar el diccionario de datos, para ver como lo está asumiendo la maquina Imagen B1.

```
{'Influenza and pneumonia': 'J09-J18', 'Chronic lower respiratory diseases': 'J40-J47', 'Adult respiratory distress syndrome': 'J80', 'Respiratory failure': 'J96'
```

Imagen B1. Reducción de Columnas. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de CDC.

Una vez obtenida, procedemos a convertir en números estas comorbilidades, y todas las columnas que podamos convertir en números.

	Group	Year	Month	ConditionGroup	Condition	ICD10_codes	AgeGroup	COVID-19Deaths	NumberOfMentions	NumCond
37260	By Month	2020.0	1.0	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	0.0	0.0	1
37261	By Month	2020.0	2.0	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	0.0	0.0	1
37262	By Month	2020.0	3.0	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	9.0	9.0	1
37263	By Month	2020.0	4.0	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	27.0	30.0	1
37264	By Month	2020.0	5.0	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	19.0	19.0	1
...	...	...	...	...	...	...	...	...	...	...
285655	By Month	2021.0	4.0	COVID-19	COVID-19	U071	All Ages	197.0	197.0	23
285656	By Month	2021.0	5.0	COVID-19	COVID-19	U071	All Ages	184.0	184.0	23
285657	By Month	2021.0	6.0	COVID-19	COVID-19	U071	All Ages	35.0	35.0	23
285658	By Month	2021.0	7.0	COVID-19	COVID-19	U071	All Ages	29.0	29.0	23
285659	By Month	2021.0	8.0	COVID-19	COVID-19	U071	All Ages	190.0	190.0	23

177842 rows × 10 columns

Tabla B2. Conversión Campos en variables Numéricas. Fuentes: Imagen generada desde Google Colab tomando el set de datos descargado de CDC.

Adicional a esto, para complejizar el proceso de aprendizaje, utilizaremos otro tipo de clasificación más puntual, como lo es la de los códigos de las enfermedades ICD-10.

ICD10_codes
J09-J18
J09-J18
J09-J18
J09-J18
J09-J18
...
U071
U071
U071
U071
U071

Tabla B3. Clasificación de Comorbilidades según código ICD10. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de CDC.

Luego de todo el proceso, ya tenemos un *dataframe* reducido, sin todos los errores, pero con la suficiente información para que empiece a entrenar Tabla B4.

	Group	Year	Month	ConditionGroup	Condition	ICD10_codes	AgeGroup	COVID-19Deaths	NumberOfMentions	NumCond	NumCod	RangoEdad
37260	By Month	2020.0	1.0	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	0.0	0.0	1	1	1.0
37261	By Month	2020.0	2.0	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	0.0	0.0	1	1	1.0
37262	By Month	2020.0	3.0	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	9.0	9.0	1	1	1.0
37263	By Month	2020.0	4.0	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	27.0	30.0	1	1	1.0
37264	By Month	2020.0	5.0	Respiratory diseases	Influenza and pneumonia	J09-J18	0-24	19.0	19.0	1	1	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...
285655	By Month	2021.0	4.0	COVID-19	COVID-19	U071	All Ages	197.0	197.0	23	6	9.0
285656	By Month	2021.0	5.0	COVID-19	COVID-19	U071	All Ages	184.0	184.0	23	6	9.0
285657	By Month	2021.0	6.0	COVID-19	COVID-19	U071	All Ages	35.0	35.0	23	6	9.0
285658	By Month	2021.0	7.0	COVID-19	COVID-19	U071	All Ages	29.0	29.0	23	6	9.0
285659	By Month	2021.0	8.0	COVID-19	COVID-19	U071	All Ages	190.0	190.0	23	6	9.0

177842 rows × 12 columns

Tabla B4. Dataframe con clasificación de enfermedades. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de CDC.

En este caso le declaramos que la columna que deseamos entrenar para la predicción es la numérica de las diferentes comorbilidades, dado que aún en la clasificación categórica, es necesario que estas se conviertan en números, por eso la otra parte importante del proceso, es estandarizar las columnas numéricas y las categóricas, como parte del preproceso, borrándole el nombre de las columnas para que inicie el entrenamiento.

Al igual que hicimos en el *FbProphet*, realizamos la validación cruzada del proceso, para ver qué tan grande es la diferencia entre la verdadera y la predicha. Que para este caso serían estas 122. 441 filas como se refleja en Tabla B5.

	Group	Year	Month	ConditionGroup	Condition	ICD10_codes	AgeGroup	COVID-19Deaths	NumberOfMentions	NumCod	RangoEdad
153495	By Month	2021.0	4.0	Circulatory diseases	Hypertensive diseases	I10-I15	25-34	0.0	0.0	6	2.0
110587	By Month	2020.0	8.0	All other conditions and causes (residual)	All other conditions and causes (residual)	A00-A39, A42-B99, D00-E07, E15-E64, E70-E90, F...	75-84	53.0	91.0	6	7.0
106482	By Month	2020.0	3.0	Respiratory diseases	Chronic lower respiratory diseases	J40-J47	25-34	0.0	0.0	2	2.0
130527	By Month	2020.0	8.0	Circulatory diseases	Hypertensive diseases	I10-I15	45-54	0.0	0.0	6	4.0
189205	By Month	2020.0	6.0	Respiratory diseases	Influenza and pneumonia	J09-J18	85+	83.0	85.0	1	8.0
...	...	...	...	...	...	...	...	...	...	...	...
101323	By Month	2020.0	4.0	All other conditions and causes (residual)	All other conditions and causes (residual)	A00-A39, A42-B99, D00-E07, E15-E64, E70-E90, F...	45-54	0.0	0.0	6	4.0
126088	By Month	2020.0	9.0	Circulatory diseases	Ischemic heart disease	I20-I25	25-34	0.0	0.0	6	2.0
87373	By Month	2021.0	2.0	Intentional and unintentional injury, poisonin...	Intentional and unintentional injury, poisonin...	S00-T98, V01-X59, X60-X84, X85-Y09, Y10-Y36, Y...	65-74	14.0	20.0	6	6.0
170198	By Month	2021.0	7.0	Intentional and unintentional injury, poisonin...	Intentional and unintentional injury, poisonin...	S00-T98, V01-X59, X60-X84, X85-Y09, Y10-Y36, Y...	75-84	0.0	0.0	6	7.0
235465	By Month	2020.0	6.0	Respiratory diseases	Adult respiratory distress syndrome	J80	0-24	0.0	0.0	3	1.0

122441 rows × 11 columns

Tabla B5. Validación Cruzada. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de CDC.

Como podemos observar, en este caso, sin haber visto las etiquetas de las columnas, y teniendo los datos dentro de una matriz numérica, la maquina aprendió fácilmente, pues las diferencias entre el grupo de comorbilidades y la predicción son relativamente pequeñas Tabla B6.

	NumCond	predicción
232863	13	12.988481
204092	7	6.996012
92326	23	22.994193
252856	20	20.003174
257283	20	20.000652
...	...	...
229093	17	16.999203
257376	20	19.996612
93480	6	6.000467
210857	17	17.001882
208283	5	5.000362

30611 rows × 2 columns

Tabla B6. Clasificador de comorbilidades. Fuente: Imagen generada desde Google Colab tomando el set de datos descargado de CDC.

Convertimos la matriz en un *dataframe*, al cual le aislamos la columna a predecir y la predicción que realizó, No fue muy exacta, en ciertos puntos, pero si en la mayoría, dado que, si fue muy cercana, por lo que podemos augurar que tenemos un muy buen clasificador de comorbilidades, dependiendo de la edad, y el código ICD10, superando por mucho los resultados que inicialmente esperábamos.

9. BIBLIOGRAFIA

[1] Banco Mundial, “Evaluación Externa de la Calidad de la Atención en el Sector de la Salud en Colombia,” 2019. https://www.bancomundial.org/es/topic/health/publication/external-assessment-of-quality-of-care-in-the-health-sector-in-colombia (accessed Apr. 04, 2021).

[2] Ministerio de salud y protección Social, “Los retos del sistema de salud que dejó la pandemia por covid-19,” *Pagina Web*, 2020. https://www.minsalud.gov.co/Paginas/Los-retos-del-sistema-de-salud-que-dejo-la-pandemia-por-covid-19.aspx (accessed Apr. 04, 2021).

[3] M. Soto, “Modelado de Datos: Definición, Usos y Tipos.” https://www.tecnologias-informacion.com/modeladodatos.html (accessed May 05, 2021).

[4] grapheverywhere, “Machine Learning Qué es, tipos, ejemplos y cómo implementarlo,” 2020. https://www.grapheverywhere.com/machine-learning-que-es-tipos-ejemplos-y-como-implementarlo/ (accessed May 04, 2021).

[5] P. S. J. Barrios, “Comparación de técnicas de aprendizaje por refuerzo jugando a un videojuego de tenis.,” 2019.

[6] “Metodología investigación: Relación entre variables cuantitativas,” *Relación entre variables cuantitativas*, 2001. https://www.fisterra.com/formacion/metodologia-investigacion/relacion-entre-variables-cuantitativas/#:~:text=las dos variables.-,Correlación,la relación entre las variables. (accessed Nov. 15, 2021).

[7] “Correlación lineal y Regresión lineal simple,” *Correlación lineal y Regresión lineal simple*, 2016. https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal (accessed Nov. 15, 2021).

[8] “La OMS publicó la nueva clasificación internacional de enfermedades | Así Vamos en Salud - indicadores en salud normatividad derechos,” *La OMS publicó la nueva clasificación internacional de enfermedades*, 2018. https://www.asivamosensalud.org/publicaciones/noticias-especializadas/la-oms-publico-la-nueva-clasificacion-internacional-de (accessed Oct. 23, 2021).

[9] OPS/OMS, “Enfermedades no transmisibles - OPS/OMS | Organización Panamericana de la Salud,” *Ops*, 2018. https://www.paho.org/es/temas/enfermedades-no-transmisibles (accessed Apr. 25, 2021).

[10] “Vigilancia Enfermedades Transmisibles,” *ENFERMEDADES TRANSMISIBLES*. https://www.ins.gov.co/Direcciones/Vigilancia/paginas/transmisibles.aspx (accessed Oct. 23, 2021).

[11] “Nuevo coronavirus 2019,” 20202. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019?gclid=Cj0KCQiA0fr_BRDaARIsAABw4EuFOV8nG27mhLr-MQ3FrYrxs9NKs3QqpMKScY_eMx6JBhDORdv7mdoaAle2EALw_wcB%0Ahttps://www.who.int/es/emergencies/diseases/novel-coronavirus-2019%0Ahttps://www.who.in (accessed Apr. 24, 2021).

[12] Ministerio de Salud Pública, CORAPE, OPS, and OMS, “Enfermedad crónicas y COVID-19,” *Centro de Coordinación de Alertas y Emergencias Sanitarias*, 2020. https://www.paho.org/sites/default/files/enfermedades-cronicas-covid-19.pdf/ (accessed Apr. 25, 2021).

[13] “Base de datos: qué es, tipos y ejemplos - Significados,” *Base de datos*. https://www.significados.com/base-de-datos/ (accessed Oct. 23, 2021).

[14] “Tasa de mortalidad - Qué es, definición y concepto | 2021 | Economipedia.” https://economipedia.com/definiciones/tasa-de-mortalidad.html (accessed Apr. 25, 2021).

[15] DANE, “Que es el DANE,” p. 1, 2008.

[16] MINSALUD, “Sistema integral de la protección social.” http://www.sispro.gov.co/Pages/Observatorios/cancer.aspx (accessed May 05, 2021).

[17] “Datos abiertos.” https://gobiernodigital.mintic.gov.co/portal/Iniciativas/Datos-abiertos/ (accessed Oct. 23, 2021).

[18] “Instituto Nacional de Salud | Colombia Plataforma Estrategica,” *PLATAFORMA ESTRATEGICA*. https://www.ins.gov.co/Paginas/Plataforma-estrategica.aspx (accessed Oct. 23, 2021).

[19] A. M. Joshi, U. P. Shukla, and S. P. Mohanty, “Smart healthcare for diabetes: A COVID-19 perspective,” *arXiv*, 2020.

[20] S. Lam *et al.*, “Social determinates of health and COVID-19 mortality rates at the county level,” *2020 4th Int. Conf. Multimed. Comput. Netw. Appl. MCNA 2020*, pp. 159–165, 2020, doi: 10.1109/MCNA50957.2020.9264276.

[21] S. S. Prasad and S. N. Korra, “Medicine Allotment for COVID-19 Patients by Statistical Data Analysis,” pp. 665–669, 2021.

[22] J. A. Vega Rivero, J. C. Ruvalcaba Ledezma, I. Hernández Pacheco, M. del R. Acuña Gurrola, and L. López Pontigo, “La salud de las personas adultas mayores durante la pandemia de COVID-19,” *J. negat. no posit. results*, pp. 726–739, 2020, doi: 10.19230/jonnpr.3772.

[23] A. C. H. Yu, L. Demi, M. Muller, and Q. Zhou, “Ultrasound Imaging: A Silent Hero in COVID-19 and Lung Diagnostics,” *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 67, no. 11, pp. 2194–2196, 2020, doi: 10.1109/TUFFC.2020.3031444.

[24] W. L. Y. Wanng Ryan Yixiang, Qinsong Guo Tim, Guanhua Li Leo, Jiao Julia Yutian, “Predictions of COVID-19 Infection Severity Basedon Co-associations between the SNPs of Co-morbidDiseases and COVID-19 through Machine Learningof Genetic Data,” *2020 IEEE 8th Int. Conf. Comput. Sci. Netw. Technol.*, vol. 1, no. 0, pp. 1–5, 2020, doi: 10.1088/1751-8113/44/8/085201.

[25] N. Darapaneni *et al.*, “Comorbidity Impact on COVID-19,” *Proc. 2020 IEEE Int. Conf. Mach. Learn. Appl. Netw. Technol. ICMLANT 2020*, pp. 3–8, 2020, doi: 10.1109/ICMLANT50963.2020.9355994.

[26] C. R. Aquino-Canchari, R. del C. Quispe-Arrieta, and K. M. Huaman Castillon, “COVID-19 y su relación con poblaciones vulnerables,” *Rev. habanera cienc. méd*, vol. 19, pp. 1–18, 2020.

[27] Tianze Qiu, “’dwd $vvrfldwhg zlwk (sljhqhwlf &kdqjhv %urxjkw e\ 6$56 &ry,” *Data Assoc. with Epigenetic Chang. Brought by SARS-Cov-2*, pp. 1–5, 2020.

[28] A. Orús, “• Países con más casos de coronavirus | Statista,” *5 De Enero*, 2021. https://es.statista.com/estadisticas/1091192/paises-afectados-por-el-coronavirus-de-wuhan-segun-los-casos-confirmados/ (accessed May 04, 2021).

[29] OPS, “Informe de la evaluación rápida de la prestación de servicios para enfermedades no transmisibles durante la pandemia de COVID-19 en las Américas,” *Ops*, 2020.

[30] L. Linn, S. Oliel, and A. Baldwin, “La COVID-19 afectó el funcionamiento de los servicios de salud para enfermedades no transmisibles en las Américas - OPS/OMS | Organización Panamericana de la Salud,” *Organización Panamericana de la Salud/Organización Mundial de la Salud*, 2020. https://www.paho.org/es/noticias/17-6-2020-covid-19-afecto-funcionamiento-servicios-salud-para-enfermedades-no (accessed Apr. 30, 2021).

[31] J. Gallardo, “Modelos de proceso para proyectos de Data Mining (DM) CRISP-DM (Cross Industry Standard Process for Data Mining),” 2010.

[32] J. Amat, “Correlación lineal y regresión lineal simple en R,” *RPubs*, 2016. https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal%0Ahttps://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal%0Ahttps://rpubs.com/Joaquin_AR/223351 (accessed Nov. 15, 2021).