INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

Departamento de Matemáticas y Física

Sustentabilidad y tecnología

PROYECTO DE APLICACIÓN PROFESIONAL (PAP)

Programa de Modelación Matemática para el Desarrollo de Planes y Proyectos de Negocio

ITESO, Universidad Jesuita de Guadalajara

4J07A Programa de Modelación Matemática para el Desarrollo de Planes y Proyectos de Negocio

Analítica avanzada de COVID-19 para el Estado de Jalisco

PRESENTAN

Programas educativos y Estudiantes

Lic. en Ingeniería Financiera. Pablo Andrés Duarte Robles

Lic. en Ingeniería Financiera. Pablo Alejandro Rivera Sánchez

Lic. en Ingeniería Financiera. Jesús Iván Lafarga Lizárraga

Profesor PAP: Diana Paola Montoya Escobar

Tlaquepaque, Jalisco, marzo del 2022

ÍNDICE

Contenido

REPORTE PAP 3
Presentación Institucional de los Proyectos de Aplicación Profesional 3
Resumen 0
1. Ciclo participativo del Proyecto de Aplicación Profesional 0
1.1 Entendimiento del ámbito y del contexto 1
1.2 Caracterización de la organización 2
1.3 Identificación de la(s) problemática(s) 3
1.4. Planeación de alternativa(s) 4
1.5. Desarrollo de la propuesta de mejora 5
1.6. Valoración de productos, resultados e impactos 46
1.7. Bibliografía y otros recursos 48
1.8. Anexos generales 49
2. Productos 49
3. Reflexión crítica y ética de la experiencia 50
3.1 Sensibilización ante las realidades 50
3.2 Aprendizajes logrados 51

REPORTE PAP

Presentación Institucional de los Proyectos de Aplicación Profesional

Los Proyectos de Aplicación Profesional (PAP) son experiencias socio-profesionales de los alumnos que desde el currículo de su formación universitaria- enfrentan retos, resuelven problemas o innovan una necesidad sociotécnica del entorno, en vinculación (colaboración) (co-participación) con grupos, instituciones, organizaciones o comunidades, en escenarios reales donde comparten saberes.

El PAP, como espacio curricular de formación vinculada, ha logrado integrar el Servicio Social (acorde con las Orientaciones Fundamentales del ITESO), los requisitos de dar cuenta de los saberes y del saber aplicar los mismos al culminar la formación profesional (Opción Terminal), mediante la realización de proyectos profesionales de cara a las necesidades y retos del entorno (Aplicación Profesional).

El PAP es un proceso acotado en el tiempo en que los estudiantes, los beneficiarios externos y los profesores se asocian colaborativamente y en red, en un proyecto, e incursionan en un mundo social, como actores que enfrentan verdaderos problemas y desafíos traducibles en demandas pertinentes y socialmente relevantes. Frente a éstas transfieren experiencia de sus saberes profesionales y demuestran que saben hacer, innovar, co-crear o transformar en distintos campos sociales.

El PAP trata de sembrar en los estudiantes una disposición permanente de encargarse de la realidad con una actitud comprometida y ética frente a las disimetrías sociales. En otras palabras, se trata del reto de "saber y aprender a transformar".

El Reporte PAP consta de tres componentes:

El primer componente refiere al ciclo participativo del PAP, en donde se documentan las diferentes fases del proyecto y las actividades que tuvieron lugar durante el desarrollo de este y la valoración de las incidencias en el entorno.

El segundo componente presenta los productos elaborados de acuerdo con su tipología.

El tercer componente es la reflexión crítica y ética de la experiencia, el reconocimiento de las competencias y los aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.

Resumen

El propósito general del PAP “Programa de Modelación Matemática para el Desarrollo de Planes y Proyectos de Negocio” es, a partir de datos del COVID-19 generados por el gobierno, analizar la información que estos nos pueden arrojar para poder identificar variables y/o relaciones de interés entre las mismas que nos ayuden a identificar y analizar el comportamiento de la enfermedad. En objetivo final de proyecto es, después de tener un análisis detallado de los datos, crear modelos de Inteligencia Artificial sobre una variable de salida de interés, en este caso, se creó un modelo de clasificación en donde se determina, con base a las características de un paciente, si este va a morir o no. Para alcanzar dicho objetivo, el proyecto se dividió en tres fases principales, entendimiento de los datos, análisis de los datos y creación del modelo, cada una de estas fases se complementan entre sí. Con lo anterior, se llegó a tener datos y visualizaciones relevantes del COVID-19 que pueden ayudar a identificar nuevas olas de contagios, así como el mencionado modelo que, con ciertas medidas de exactitud, determina si un paciente morirá o no. Toda esta información se obtuvo con el propósito de ayudar a tener un mejor control y entendimiento de la pandemia.

1. Ciclo participativo del Proyecto de Aplicación Profesional

El PAP es una experiencia de aprendizaje y de contribución social integrada por estudiantes, profesores, actores sociales y responsables de las organizaciones, que de manera colaborativa construir sus conocimientos para dar respuestas a problemáticas de un contexto específico y en un tiempo delimitado. Por tanto, la experiencia PAP supone un proceso en lógica de proyecto, así como de un estilo de trabajo participativo y recíproco entre los involucrados.

El proyecto siguió una metodología de ciencia de datos llamada, en inglés, Team Data Science Process (TDSP), dicha metodología funciona de manera ágil e iterativa con el objetivo de proporcionar soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente. Este proceso proporciona un ciclo de vida para estructurar el desarrollo de los proyectos de ciencia de datos, fomentando la colaboración y el aprendizaje en equipo para llegar a un resultado final de manera satisfactoria.

Durante todo el proyecto nos apegamos a este ciclo de vida de ciencia de datos, el cual se divide principalmente en 3 fases, la primera es el entendimiento del negocio, que en este caso sería el entendimiento o contexto de la pandemia y el COVID-19. La segunda fase es la adquisición y entendimiento de datos, así como la creación de modelos con los mismos (modelado, son dos tareas diferentes que se complementan entre sí. Finalmente, la última fase es el despliegue del resultado (modelo final).

De manera más específica, el proyecto siguió estas fases mediante 4 bloques, seguimos un cronograma de actividades dividido en dichos bloques. El primer bloque consiste del entendimiento del problema del COVID-19 y el entendimiento de los datos, este bloque corresponde a la primera fase anteriormente mencionada. El segundo bloque consiste en realizar un análisis exhaustivo de los datos para encontrar relaciones entre los datos, mientras que el tercer bloque consiste en, ya habiendo realizado el análisis de datos, generar modelos de regresión y/o clasificación del COVID-19, así como pronósticos de casos confirmados o número de personas hospitalizadas. Los dos bloques anteriores corresponden a la fase dos explicada anteriormente. Por último, el último bloque, que corresponde a la última fase anteriormente mencionada, consiste en la entrega del proyecto y de los resultados obtenidos durante todos los bloques.

1.1 Entendimiento del ámbito y del contexto

La enfermedad por COVID-19 es una enfermedad infecciosa causada por el virus SARS-CoV-2. Tanto fiebre como la dificultad para respirar son síntomas conocidos que pueden surgir de una infección por COVID-19, la mayoría de la gente que ha superado la enfermedad se recupera completamente en el plazo de unas semanas, pero parece que el virus en algunos casos puede dañar los pulmones, el corazón y el cerebro lo que aumentaría el riesgo de salud a largo plazo. La reunión mundial de la salud advirtió que no sólo son preocupantes los casos graves y las muertes por virus en el mundo, sino también los efectos secundarios de la enfermedad, los cuales aún se desconocen, así como todas las secuelas derivadas de un contagio de un virus ya que se ha observado que algunos pacientes desarrollan una forma

crónica de la enfermedad, con síntomas como agotamiento permanente y problemas cognitivos.

La pandemia de COVID-19 ha llevado a una pérdida dramática de vidas humanas en todo el mundo y presenta un desafío tanto para la salud pública, la economía y el mundo del trabajo. El impacto causado por la pandemia en la economía y en la vida social de las personas es devastador, millones de personas están en riesgo y han entrado en extrema pobreza, millones de empresas han quebrado mientras otras aún enfrentan el riesgo de dejar de existir y millones de trabajadores han perdido sus trabajos y estabilidad económica. Hoy en día, ha habido más de 500 millones de casos y más de 6 millones de muertes alrededor de 212 países, además, la enfermedad ha afectado la infraestructura sanitaria de los países, los mercados financieros y la economía de forma directa.

Este proyecto se origina debido a la aparición de la pandemia a nivel mundial a inicio del año 2020. El gobierno de Jalisco y el resto del país comenzó a recolectar todos los datos sobre la enfermedad conforme fueron llegando, con el objetivo de poder tener información que se pudiera analizar y actuar en contra de la problemática. Al tener acceso a los datos del SINAVE, podemos aplicar diferentes técnicas de análisis de series de tiempo con la tecnología que tenemos disponible hoy en día, aprovechando la creciente popularidad del análisis de datos y modelado para la toma de decisiones.

1.2 Caracterización de la organización

El escenario en el que se desarrolló el PAP fue en conjunto con el Gobierno de Jalisco, utilizando sus bases de datos y estando en comunicación a través del docente (Diana Paola Montoya Escobar).

El proyecto fue conformado por un equipo de trabajo de 10 personas que cumplieron con el rol de analista de datos, entre todos nos encargamos de analizar diferentes datos y variables sobre el COVID-19 en varios niveles (estatal, nacional, etc). Cada semana realizamos la misma entrega acerca del mismo análisis, sin embargo, este equipo de 10 personas estaba

subdividido en 3 equipos, esto con el propósito de que, en cada entrega, a pesar de analizar lo mismo, se tuvieran hallazgos y perspectivas diferentes que nos ayudaran a encontrar diferentes relaciones en las bases de datos analizadas.

Cada semana se realizaron análisis, visualizaciones, cálculos y se presentaron en distintos formatos (sobre todo en Jupyter Notebooks que contenían el código y las salidas de este). Cada equipo realizaba un reporte con sus hallazgos y conclusiones, con el fin de recopilar todo para este reporte final.

El propósito final de hacer estos análisis entre todos fue poder entender la base de datos y los diferentes comportamientos de la enfermedad con el fin de crear un modelo de clasificación que nos ayude a determinar, de acuerdo con las características del paciente, si este sobrevive o muere al padecer la enfermedad.

1.3 Identificación de la(s) problemática(s)

A inicios del año 2020, como ya sabemos, la enfermedad comenzó a expandirse a distintos países, ocasionado que varios impusieran distintas medidas para prevenir la propagación de la enfermedad. Esto fue un fenómeno nunca antes visto y el mundo entero tuvo que encontrar nuevas formas de seguir funcionando y respetar estas medidas (como el teletrabajo, que revolucionó la forma en que la sociedad es productiva).

Dichas medidas generaron varias problemáticas tanto económicas como sociales, las cuales se han ido solucionando con el paso del tiempo, como por ejemplo la adaptación de trabajar desde casa. Sin embargo, la problemática más importante ha sido en el ámbito de la salud, el COVID-19 es una enfermedad agresiva que ha ocasionado varias muertes y problemas de salud graves a diferentes personas alrededor del mundo. Cuando recién surgió la enfermedad, nadie sabía exactamente cómo se comportaba, cuáles eran los diferentes síntomas ni cómo combatirla, pero, con el paso del tiempo, hemos controlado hasta cierto punto la enfermedad gracias a diferentes medidas y a la creación de las vacunas.

A pesar de todos los avances que hemos tenido, aún existe una gran preocupación en el tema de la salud, han aparecido diferentes olas a lo largo de estos dos años que nos toman por sorpresa, principalmente debido al surgimiento de una nueva variante. Como cada variante es diferente, es muy complicado predecir cómo se va a comportar y cuáles son sus características, sin embargo, mediante la información y los datos almacenados de todos los casos que se han generado, podemos darnos una idea del comportamiento general de la enfermedad, crear modelos de predicción y clasificación, así como encontrar relaciones entre diferentes variables que nos ayuden a controlar y darnos una idea del comportamiento de cada variante u ola nueva que surja.

1.4. Planeación de alternativa(s)

La solución del problema se dará gracias al desarrollo de herramientas de visualización y análisis de datos, así como la implementación de técnicas de machine learning para realizar predicciones, clusters y análisis para la toma de decisiones con los datos generados durante la pandemia sobre el comportamiento del COVID-19 en Jalisco. Se utilizaron diversas herramientas tales como R y Python para la manipulación, limpieza, exploración y visualización de los datos, así como para poder modelar con el fin de tener un panorama más claro sobre qué acciones tomar.

La realización de este proyecto es importante debido a la situación de emergencia global que surgió desde el año 2020. El poder tener datos sobre el futuro sobre posibles escenarios de contagios, defunciones y otras variables de interés para el gobierno de Jalisco, puede salvar vidas y mejorar las condiciones de muchas personas. De igual manera, el tener datos reales del pasado puede ayudar a identificar patrones que puedan volver a repetirse en un futuro.

El plan de trabajo del proyecto se divide en 3 fases, el entendimiento del problema y de la base de datos, análisis de los datos y modelado de los mismos. Siguiendo esta pequeña estructura se cubrirá gran parte de la alternativa propuesta anteriormente, es decir, tendremos información de relevante del COVID obtenida mediante en análisis de los datos y con esta podremos obtener diferentes modelos que nos aporten más información de la enfermedad.

1.5. Desarrollo de la propuesta de mejora

A lo largo del PAP, se realizaron varias entregas que contenían diferentes análisis de los datos para poder entenderlos de una mejor manera, así como encontrar relaciones y nuevas variables (calculadas a partir de las existentes) que nos mostraran el comportamiento del COVID-19. Los productos presentados a continuación se dividen en cada una de estas entregas.

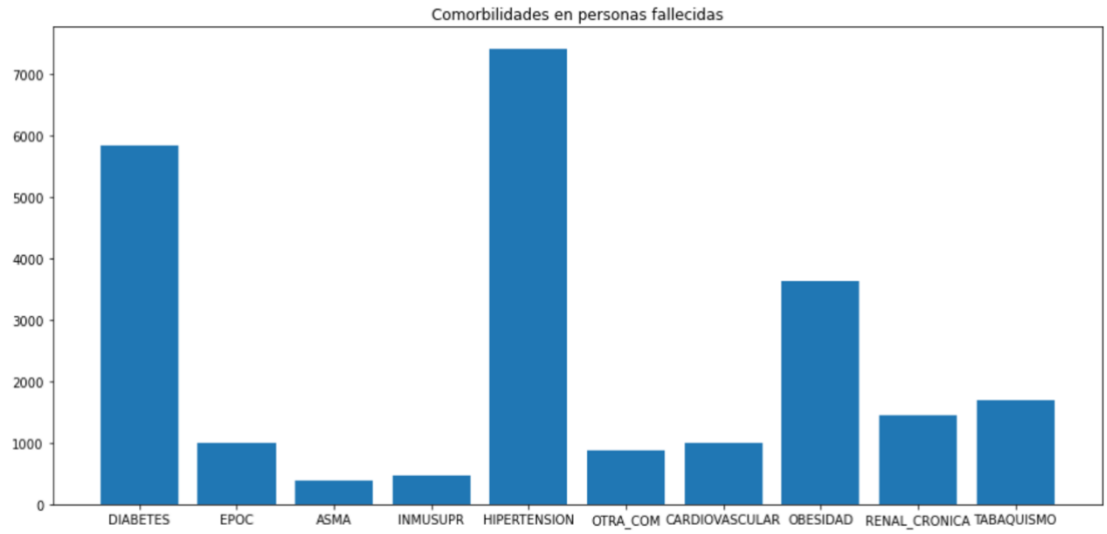
Entrega 1 – Entender los datos

Esta entrega fue la primera en la cual trabajamos con la base de datos de SINAVE, por lo que primero se realizó la limpieza y transformación de dicha base de datos para poder trabajar con ella. De igual manera, se filtró para que únicamente muestre los casos de Jalisco, todo lo anterior se realizó en un código que se utilizó para limpiar, transformar y filtrar las bases de datos más actualizadas con las que trabajamos las siguientes entregas.

Posteriormente, tratamos de responder algunas preguntas que nos habíamos planteado, estas fueron:

1) ¿Qué características más prominentes tienen las personas que han fallecido? (Comorbilidades)

La primera pregunta que podríamos responder con los datos es ¿qué comorbilidad era más común entre los fallecidos? Manipulamos los datos y llegamos a la siguiente gráfica.



Un dato interesante al que llegamos es que alrededor del 78% de las personas que fallecieron padecían de por lo menos una de estas comorbilidades, por lo que podríamos concluir que la probabilidad de fallecer por COVID aumenta drásticamente si se padece al menos una. Como podemos observar, la gráfica muestra que las 3 comorbilidades más comunes en las personas fallecidas son hipertensión, diabetes y obesidad.

2) ¿Hay manera de "predecir" o identificar la formación de un pico antes de que suceda? Esto con la idea de evitarlo o minimizarlo, es decir, cuando se empiece a ver el comportamiento de un pico o cuando empiecen los indicios de uno (si es que lo hay), tomar medidas para que este no sea tan grave.

Esta pregunta la tratamos de responder más no encontramos la manera de hacerlo utilizando únicamente los datos disponibles. Se podría realizar algún modelo predictivo para poder predecir picos, o también, por ejemplo, observar las diferencias entre los casos por día y si esta es mayor a x, entonces puede que sea el inicio de un pico. Sin embargo, sin más cálculos o datos es muy difícil hacerlo.

3) ¿Cuántos casos salieron positivos con prueba de laboratorio pero negativo con prueba de antígeno o viceversa? Esto para identificar la efectividad de la última y poder ver cuántos casos pueden pasar desapercibidos.

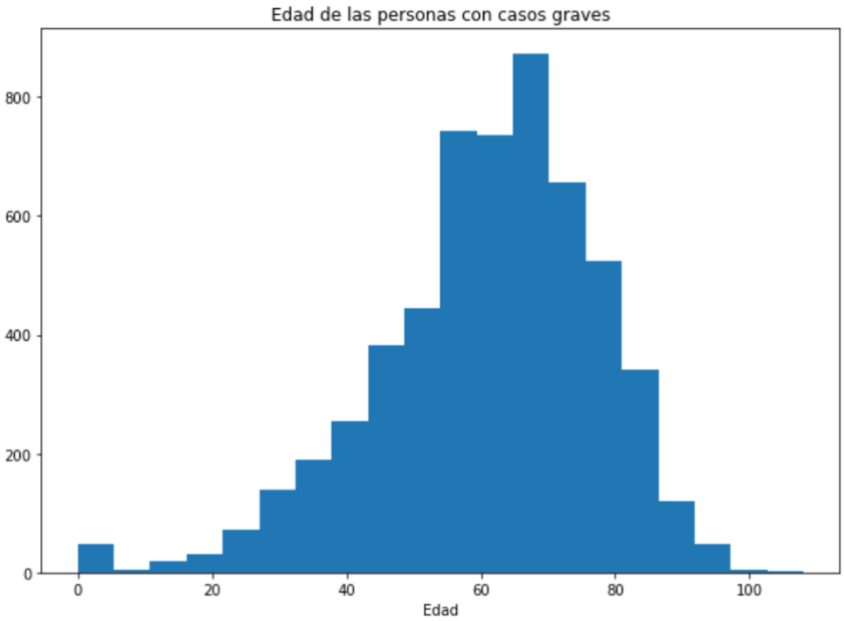
Para la pregunta anterior, se calcularon las veces en donde la prueba de antígeno y la prueba de laboratorio coincidían en su resultado (ya sea positivo o negativo). También se calculó el número de veces que estas no coincidían, es decir, que una saliera positiva y la otra negativa y viceversa. Se obtuvieron los siguientes resultados.

	Coincidencia	Lab. P y Ant. N	Lab. N y Ant. P
Número pruebas	15237	3765	452

Calcular un porcentaje de efectividad es un poco complicado, sin embargo, con base a los datos podemos observar que ambas pruebas coincidieron alrededor de 15,000, mientras que han sido diferentes la una con la otra en alrededor de 4,000 ocasiones. En 3765 ocasiones la prueba de laboratorio fue positiva y la de antígeno negativa, por otra parte, en 452 ocasiones la prueba de laboratorio fue negativa y la de antígeno fue positiva.

4) Características de los casos positivos que fueron intubados o requirieron ir a la unidad de cuidados intensivos

De los casos positivos, analizamos brevemente la gravedad de ellos. Podemos observar que la siguiente gráfica cumple con la relación de riesgo y edad, las personas mayores tienden a tener mayores complicaciones que las personas jóvenes ocasionando que sean intubadas o admitidas a la UCI.

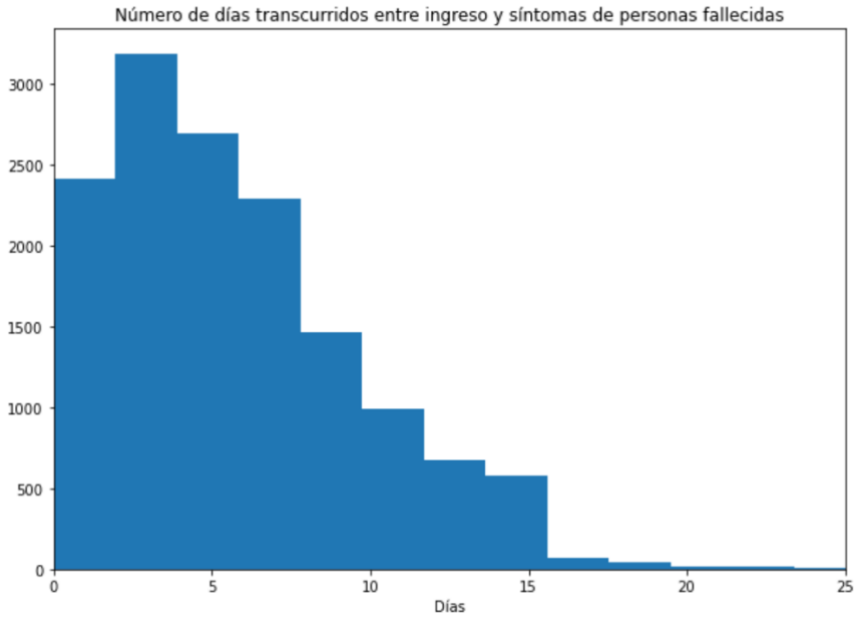


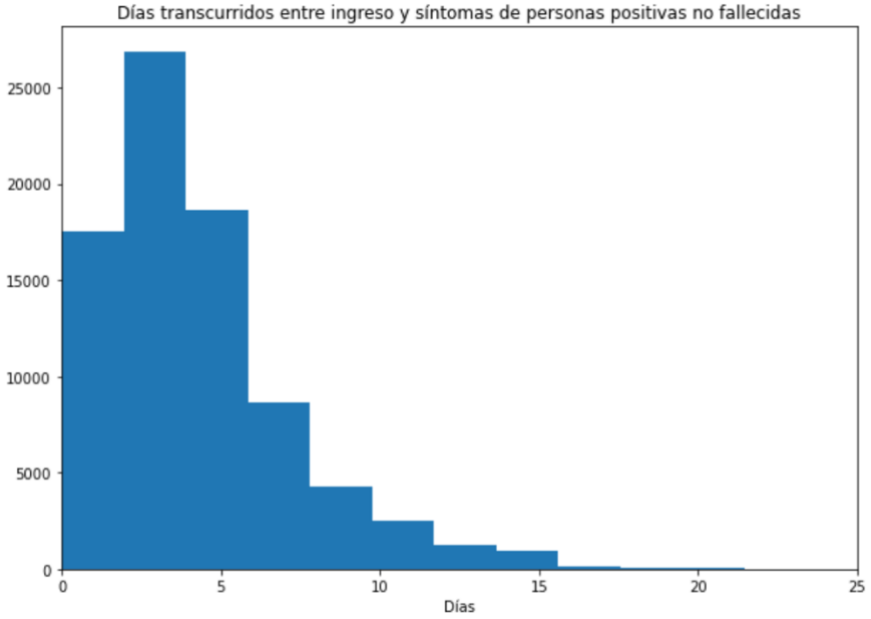
5) Porcentaje de casos positivos que requirieron UCI o fueron intubados.

Encontramos que las veces en las que una persona requiere intubación o ingresar a una UCI es muy pequeña, un 5.9% de los casos positivos requiere alguna de estas 2 medidas. Por otra parte, encontramos que casi un 27% de personas que fallecieron y estaban confirmadas como positivas requirieron alguna de estas 2 medidas, lo que nos podría indicar que la probabilidad de morir aumenta si el caso de COVID es grave (ingresa a una UCI o requiere intubación).

6) ¿Cuántos días, en promedio, espera la gente después de tener síntomas para ingresar a una unidad médica? Y también ¿este periodo de tiempo influye en el resultado del paciente (si se recupera o no)? Ver si entre más rápido se atiende una persona puede que sea mejor su pronóstico o no.

Para analizar y poder responder la pregunta anterior analizamos la diferencia en días entre la fecha de ingreso contra la fecha de la aparición de síntomas, tanto de personas que fallecieron como de personas con un resultado positivos que no fallecieron. Estos fueron los resultados.





Con base a las gráficas anteriores, podemos concluir que parece que no hay una relación fuerte entre el resultado del paciente (si fallece o no) y el número de días que esperan en ingresar a una unidad de salud. Ambas gráficas son similares, es decir, en promedio, las personas esperan la misma cantidad de días en atenderse y esto no influye mucho en su resultado. Sin embargo, parecería que en la gráfica de las personas fallecidas (superior), hay ligeramente más datos cargados hacia la derecha, esto indica que hay más personas que esperan un poco más de tiempo en atenderse.

Entrega 2 – Positividad y Rt

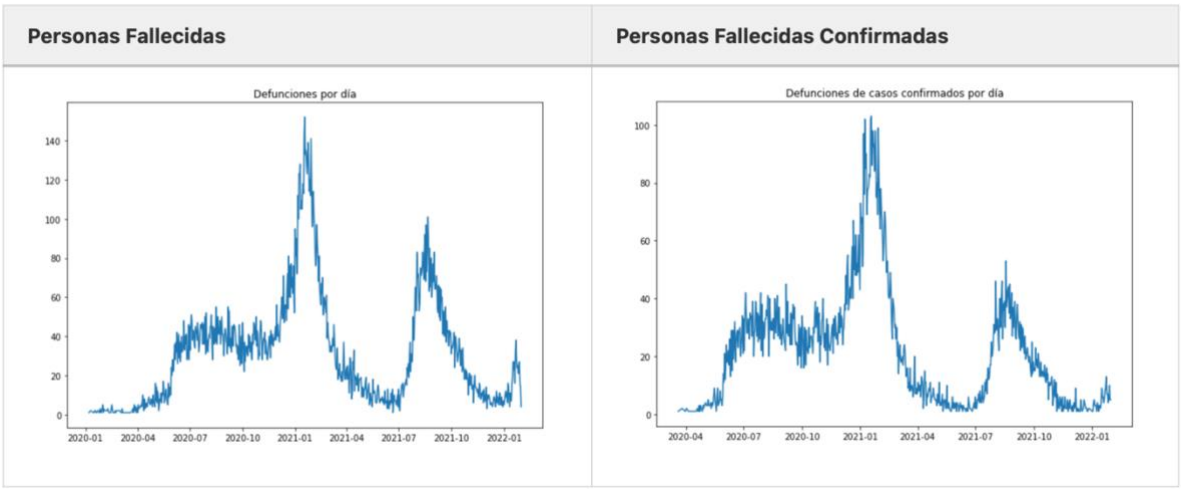
Para esta entrega, el trabajo realizado consistió en calcular ciertas variables (defunciones, mortalidad, positividad y la tasa de reproducción efectiva del virus) en diferentes visualizaciones o clasificaciones tales como fecha, sexo, edad, entre otras, todas únicamente del Estado de Jalisco. El objetivo principal de hacer este análisis es entender y poder darnos una idea de cómo el COVID afecta a cada una de las clasificaciones anteriores a través de diferentes métricas, las cuales nos pueden servir para analizar el comportamiento del COVID a través del tiempo.Como ya se mencionó anteriormente, dichas métricas fueron:

- Defunciones: En nuestro caso, analizamos tanto las defunciones totales, así como las defunciones de casos confirmados (resultado de laboratorio positivo).
- Mortalidad: Dicha métrica fue calculada como el número de defunciones confirmadas como positivas sobre los casos positivos, todo esto para cada clasificación.
- Positividad: La positividad, fue calculada como la división entre los casos positivos, entre la suma de los casos confirmados como positivos y negativos.
- **Tasa de reproducción efectiva del virus:** Finalmente, esta tasa que calcula el número de casos promedio que van a ser causados por una personas infectada, fue obtenida mediante la división de la suma de los casos confirmados de los 3 días más recientes, sobre la suma de casos confirmados de hace 3 días (brincándose 1 día).

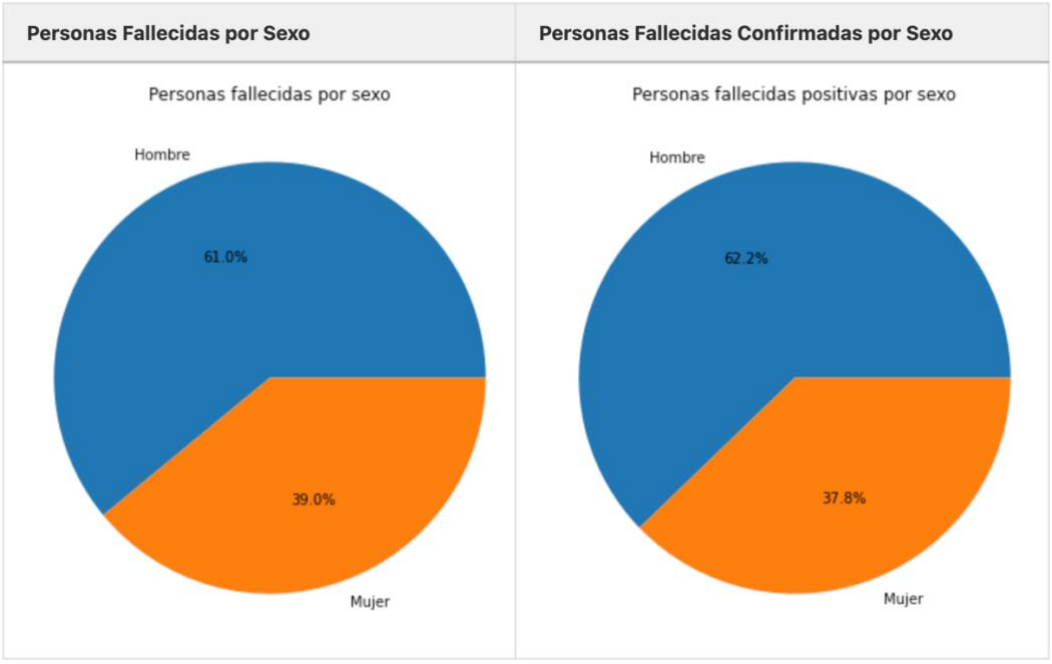
A continuación, se muestran los resultados de cada una de las métricas:

Defunciones

Primero, podemos observar la gráfica de las personas fallecidas por día en Jalisco, la cual coincide con las diferentes olas de contagio que han surgido en los últimos dos años. De igual manera, no hay ninguna diferencia relevante al comparar las defunciones totales y las confirmadas como positivos, este comportamiento se repite a lo largo de este pequeño análisis.

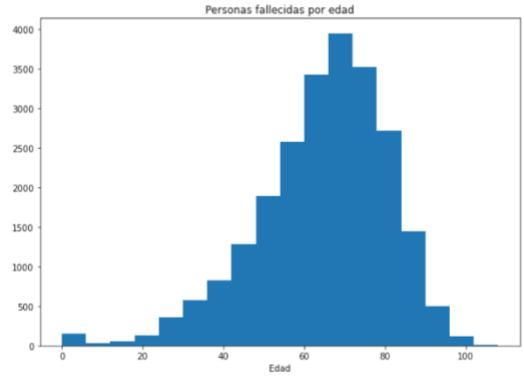


En la siguiente gráfica se muestra el porcentaje de fallecidos por sexo en Jalisco y, como podemos ver, la mayor parte de fallecidos (en ambos casos) son hombres, siendo alrededor de un 60%, mientras que las mujeres corresponden a alrededor de un 40%.

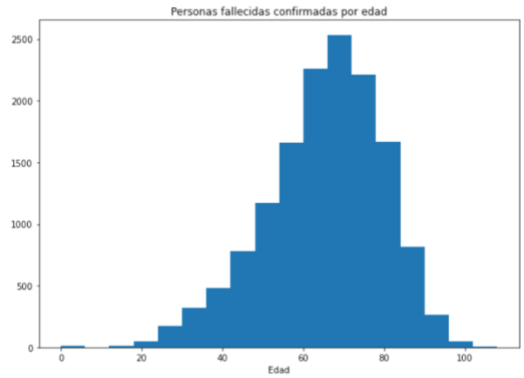


Como ya hemos escuchado y se nos ha informado en varias ocasiones, el COVID afecta gravemente a personas mayores, por lo que la probabilidad de fallecer de estas es alta. En la gráfica siguiente, podemos ver que esto se cumple en Jalisco, el histograma muestra que, a medida que la edad aumenta, el número de fallecidos también lo hace, encontrándo la mayor parte de fallecidos en un rango de 60 a 90 años de edad.

Personas Fallecidas por Edad

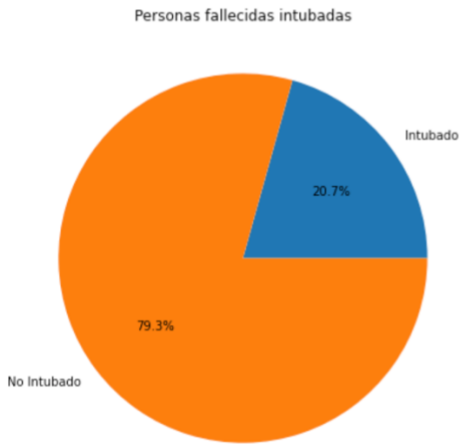


Personas Fallecidas Confirmadas por Edad

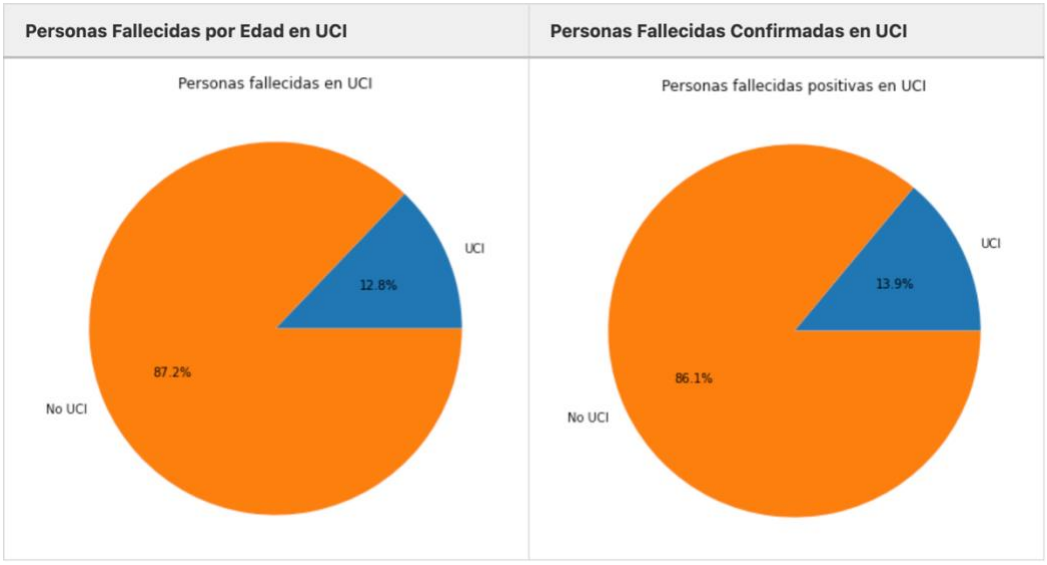


Las siguientes 2 gráficas muestran el porcentaje de personas fallecidas de acuerdo a su gravedad, es decir, si requirieron incubación o ingresar a la UCI. En ambos casos, vemos que la mayor parte de las personas fallecidas no requieren ninguna de estas dos medidas, es decir, mueren antes de llegar a ser casos considerados como "graves".

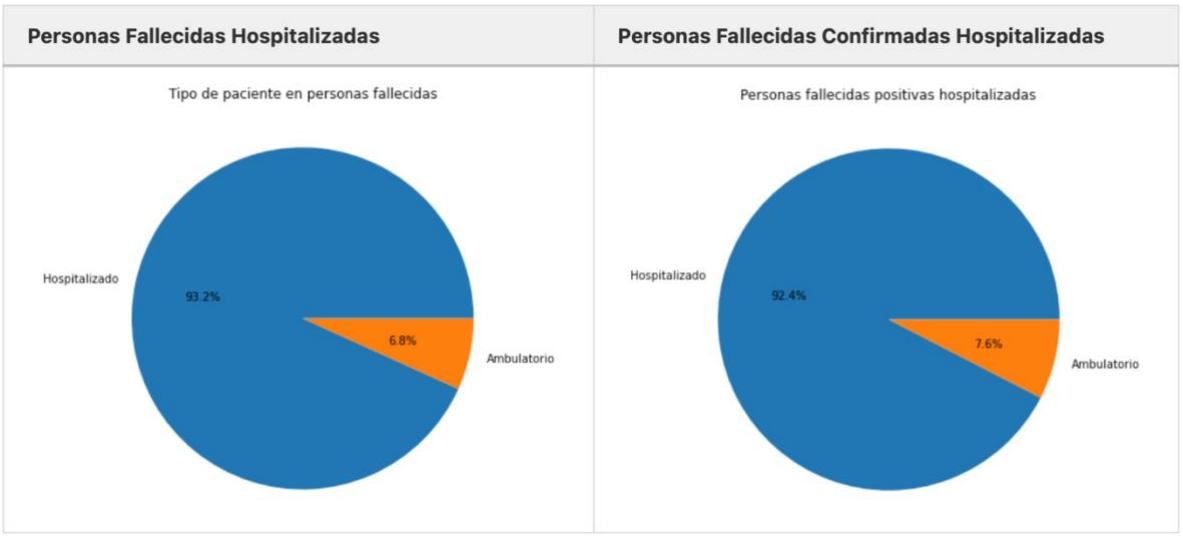
Personas Fallecidas por Edad Intubadas



Personas Fallecidas Confirmadas Intubadas



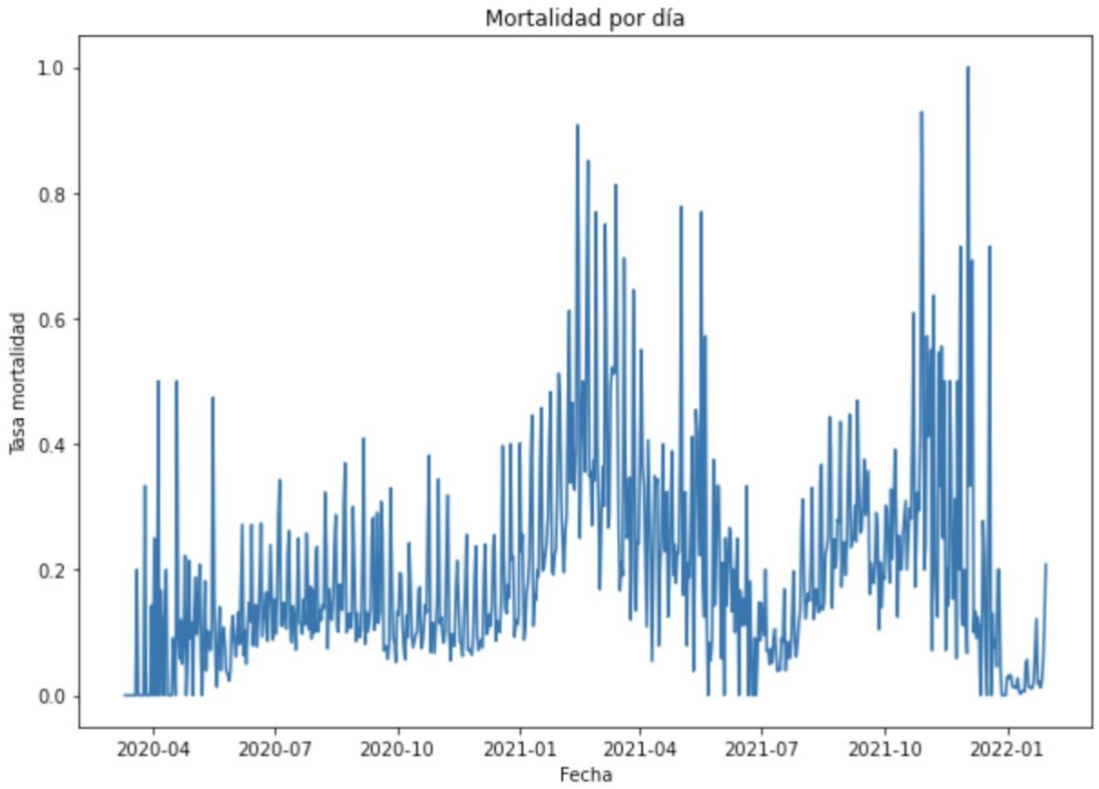
La siguiente visualización muestra el porcentaje de personas fallecidas de acuerdo a su tratamiento, en este caso, casi todas las personas fallecidas estaban hospitalizadas.



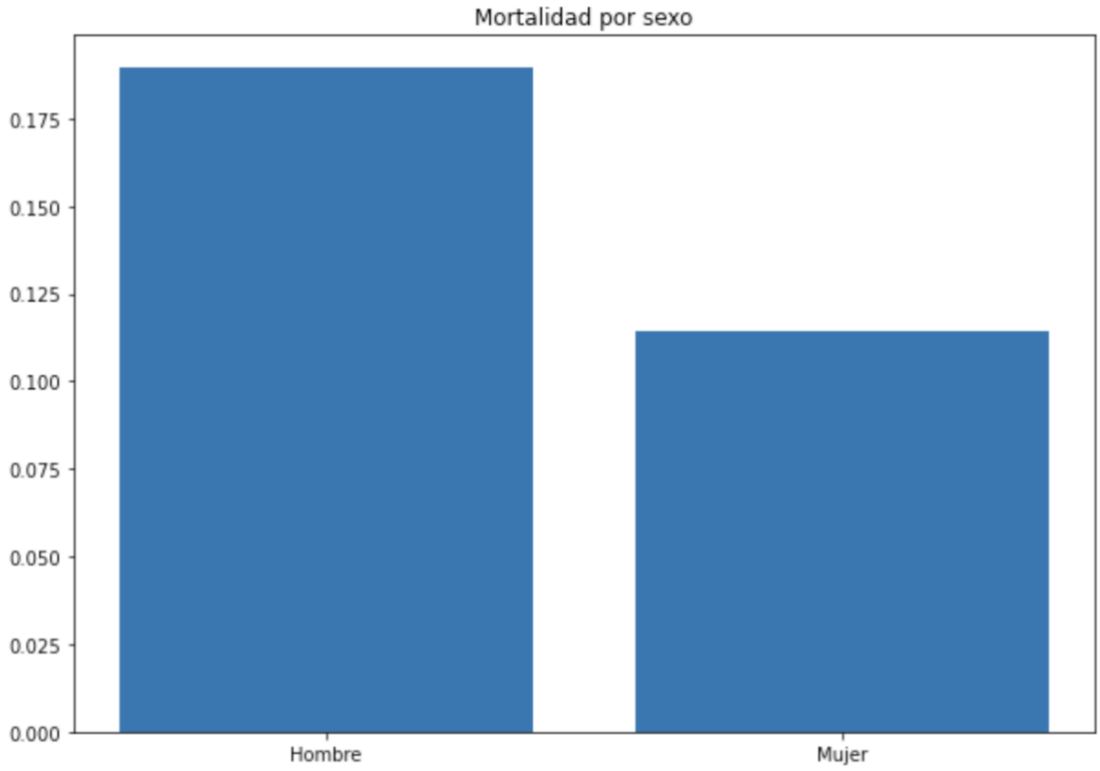
Mortalidad

Recordando, esta métrica fue calculada como las defunciones de casos confirmados (positivos), entre el total de casos positivos, y no muestra la mortalidad del virus.

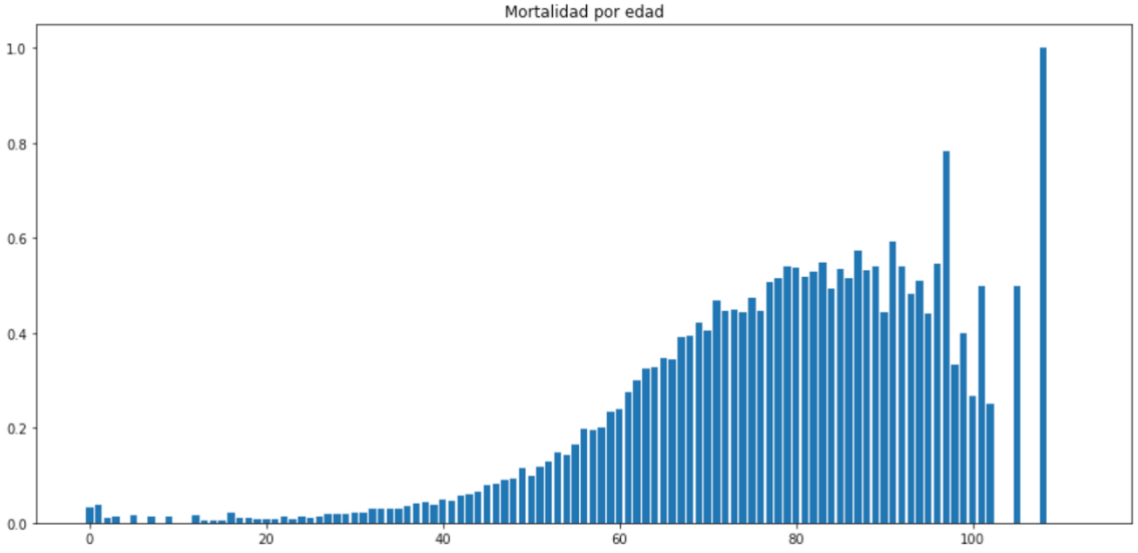
Primero, encontramos que la mortalidad total del Estado de Jalisco es de un 15.16%, sería interesante comparar con los demás Estados para ver en qué posición nos encontramos. La siguiente gráfica muestra la tasa de mortalidad por día en Jalisco. Al igual que las defunciones, se observan picos en la tasa cuando ha habido picos en contagios.



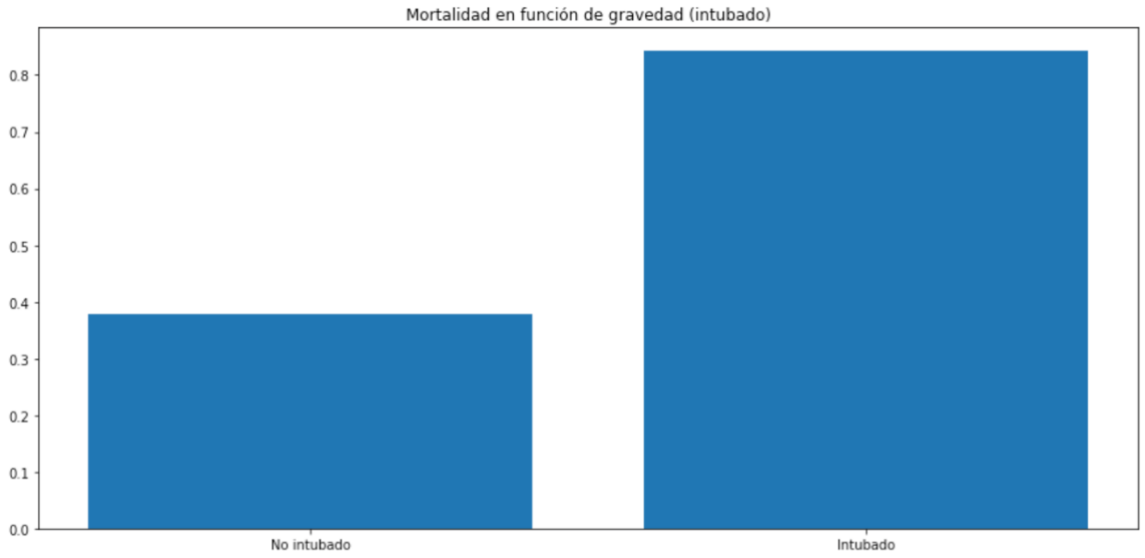
Posteriormente, podemos visualizar la misma tasa de mortalidad pero ahora clasificada por sexo. Podemos observar que la tasa de mortalidad de los hombres es más alta que la de las mujeres, lo cual coincide con lo encontrado anteriormente en la métrica de defunciones, donde había más hombres fallecidos.



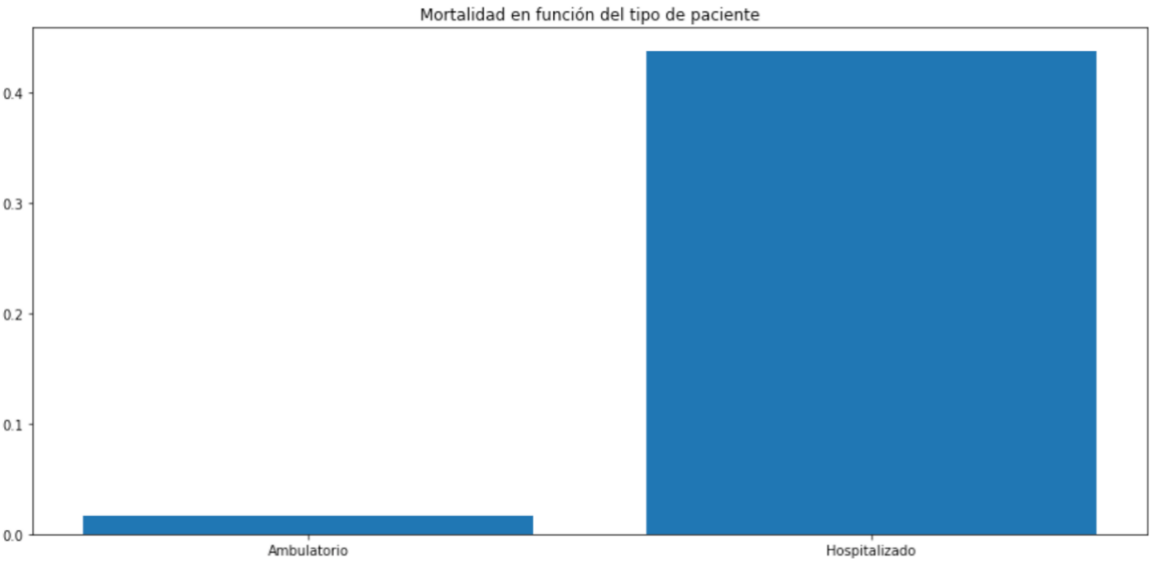
Al obtener la tasa por edad, vemos que, al igual que las defunciones, esta tiende a aumentar a medida que la edad también aumenta, indicando que los adultos mayores son más propensos a morir. Recalcando que la edad está representada por el eje de las x.



En la siguiente gráfica podemos observar la tasa de mortalidad en función de la gravedad del paciente, es decir, si fue intubado o no. En este caso, la tasa de mortalidad es casi el doble en personas intubadas que en personas no intubadas, por lo que las probabilidades de morir aumentan si el caso es grave.



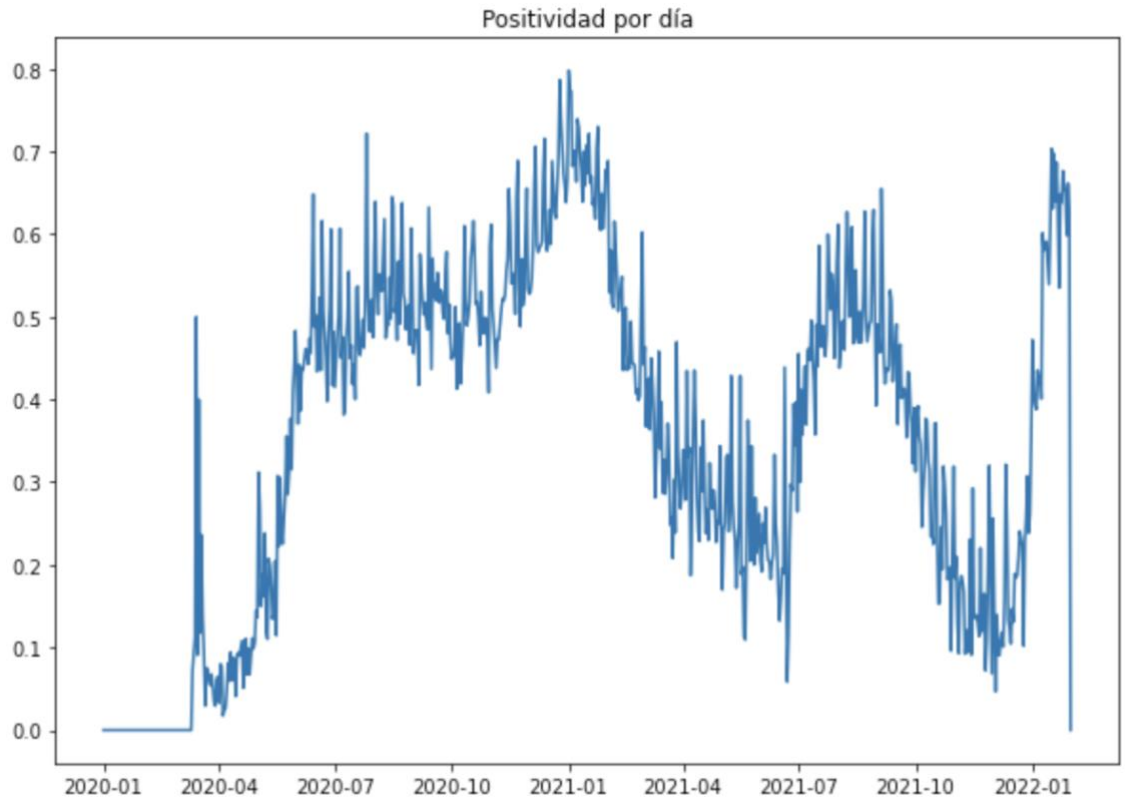
Al analizar la mortalidad por tipo de paciente, la tasa es mayor en pacientes hospitalizados. Esto tiene sentido ya que la intubación se hace por lo general en un centro de salud.



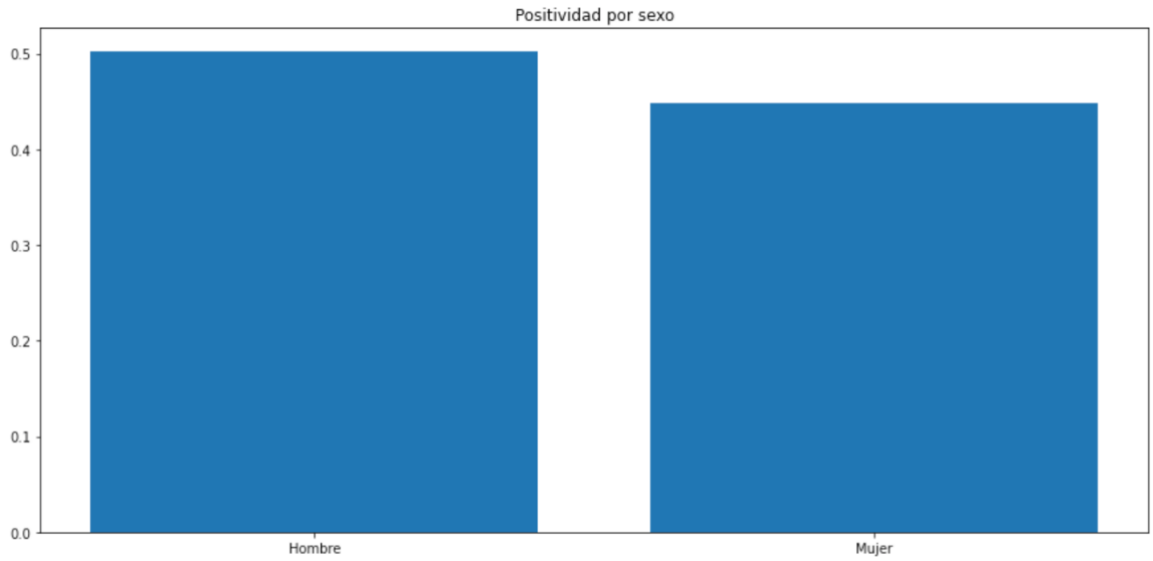
Positividad

Recordando, esta métrica fue calculada como el número de casos positivos (confirmados) dividido entre el número de casos positivos más casos negativos.

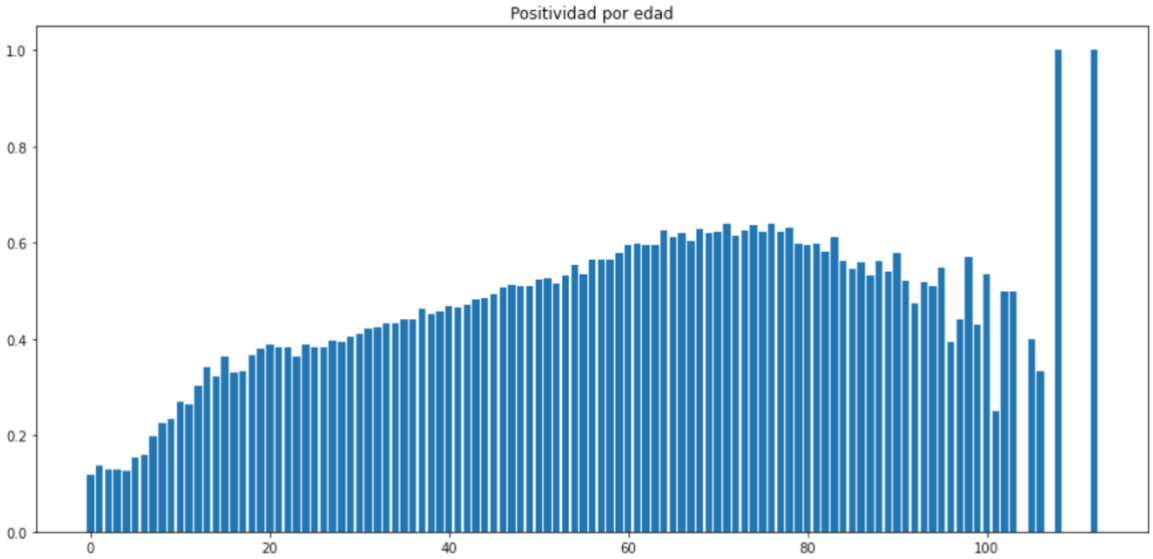
Con respecto a esta métrica, encontramos que la positividad en Jalisco, de acuerdo a la base de datos, ha sido de un 47.36%. Si analizamos la positividad por día, como se muestra en la siguiente gráfica, vemos que, al igual que las 2 métricas anteriores, esta aumenta con respecto a los casos. Las subidas y bajadas en la positividad coinciden con los picos que hemos tenido con respecto a casos positivos.



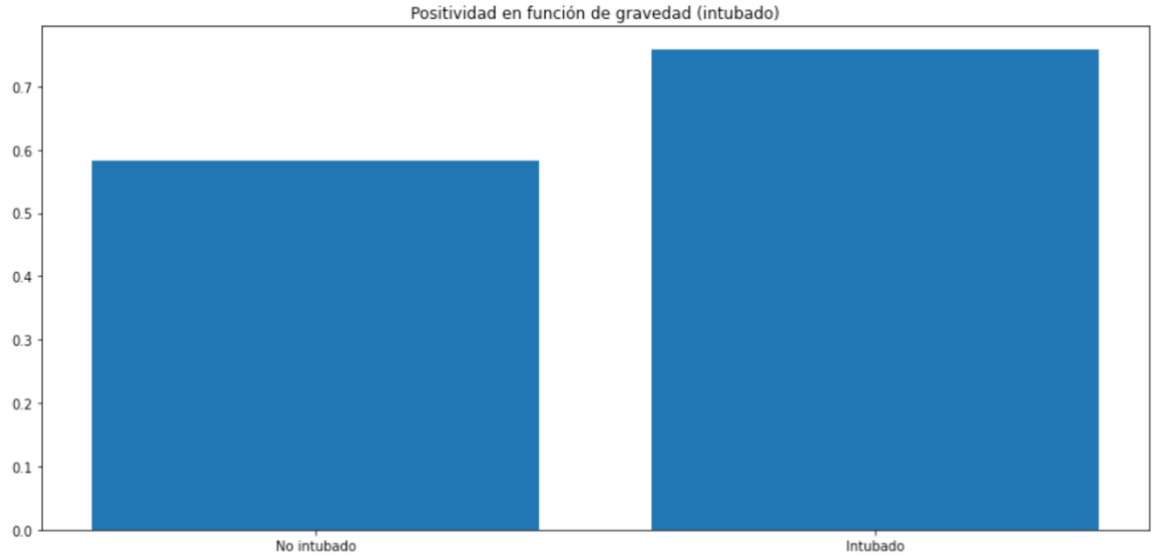
La siguiente gráfica muestra la positividad por sexo, la cual es ligeramente más alta en hombres. Nuevamente esta información coincide con las 2 variables anteriores, donde los hombres también son mayormente afectados por el virus.



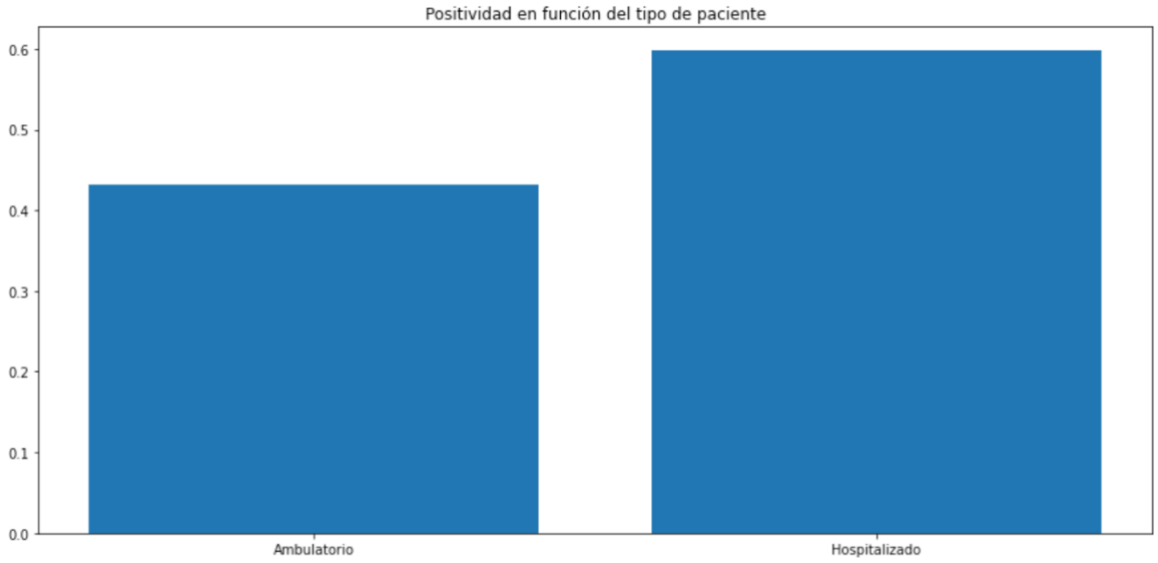
En cuanto a la positividad por edad, esta aumenta conforme la edad lo hace. Pensaríamos que esto debería ser a la inversa, que la positividad sea mayor en personas jóvenes, sin embargo, no lo es. Esto podría ser ocasionado por el hecho de que las personas jóvenes tienen más acceso y posibilidad de hacerse una prueba, mientras que personas mayores pueden estar un poco más limitadas. Esto ocasiona más casos negativos en jóvenes (denominador mayor) y por lo tanto una tasa de positividad menor.



En la siguiente gráfica, vemos la positividad en función de la gravedad del paciente y, como se puede observar, dicha tasa es más alta en personas que fueron intubadas.



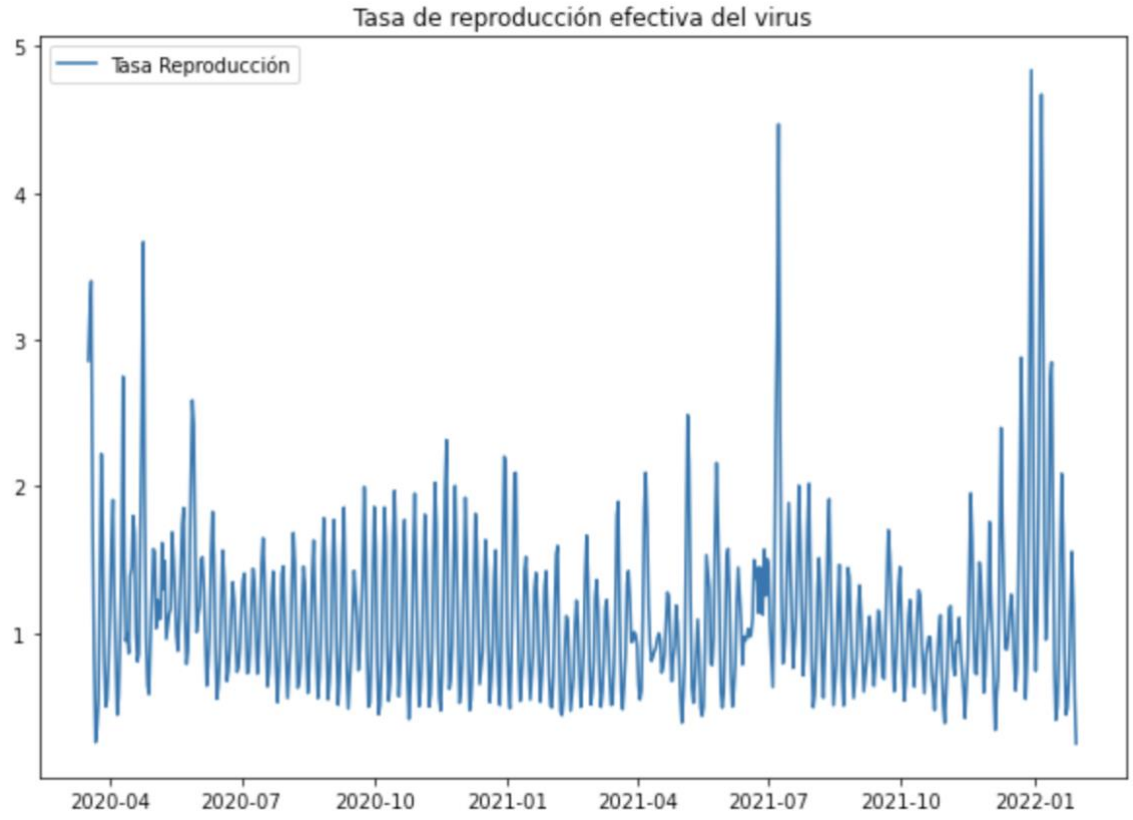
Finalmente, si analizamos la positividad por tipo de paciente, la tasa es más alta en pacientes hospitalizados que en pacientes no hospitalizados.



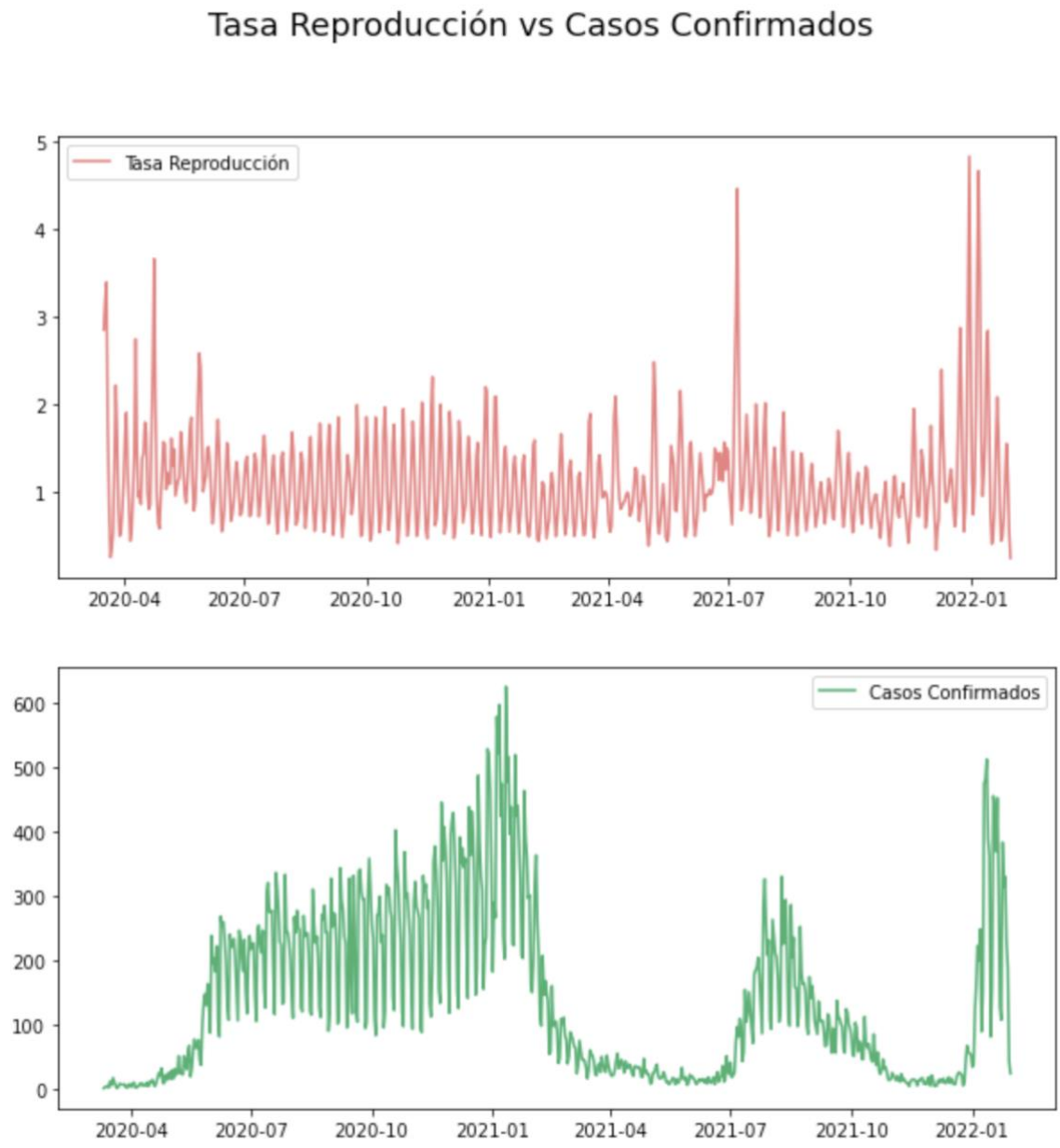
Tasa de reproducción efectiva del virus

Recordando, la tasa de reproducción efectiva calcula el número de casos promedio que van a ser causados por una persona infectada, es decir, cuántas personas van a ser contagiadas por cada persona ya contagiada.

La siguiente gráfica muestra la tasa de reproducción efectiva del virus diaria en Jalisco. A lo largo de este periodo de tiempo, la tasa ha sido en promedio 1.13, es decir, hay 1.13 contagiados por cada caso positivo en Jalisco, de igual manera, el máximo ha sido 4.83 y el mínimo ha sido 0.24.



En la siguiente gráfica se muestra el comportamiento de tasa de reproducción efectiva del virus contra los casos positivos. Podemos observar que también coincide más o menos con los picos de contagios vistos a lo largo del tiempo, sin embargo, recientemente la tasa ha sido muy alta lo cual tiene sentido debido a la variante más contagiosa OMICRON.

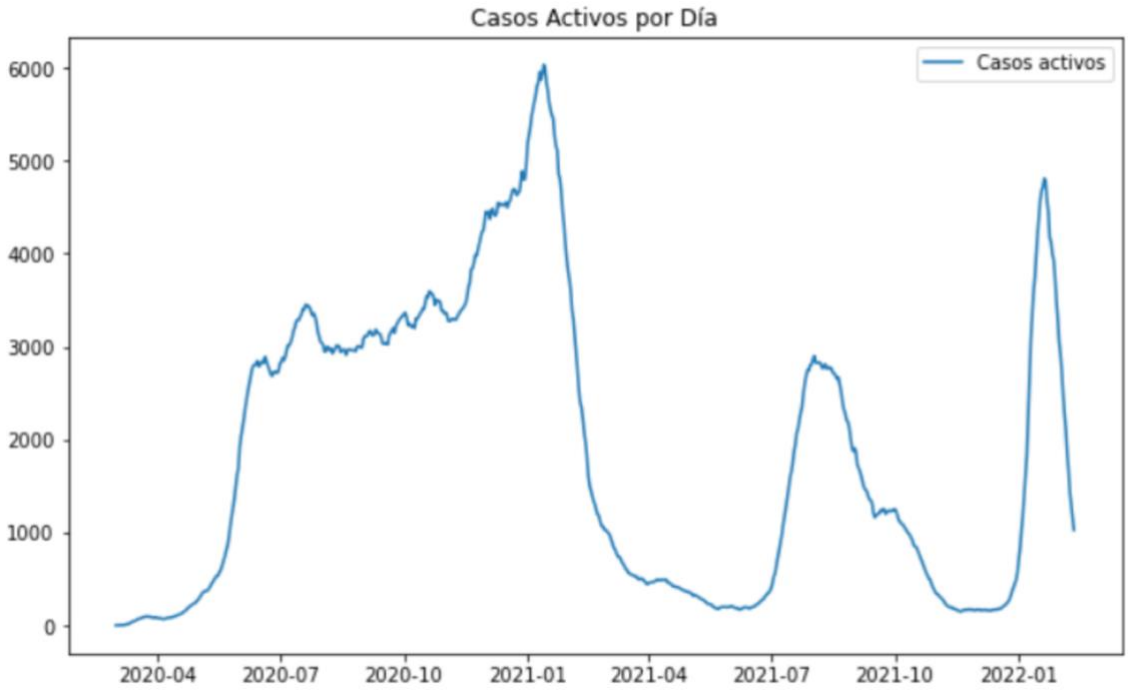


Más adelante analizaremos de manera un poco más detallada la tasa de reproducción efectiva del virus, veremos que nos puede ayudar a identificar ciertas tendencias en los contagios.

Entrega 3 – Análisis comparativo entre municipios y entre olas de contagio

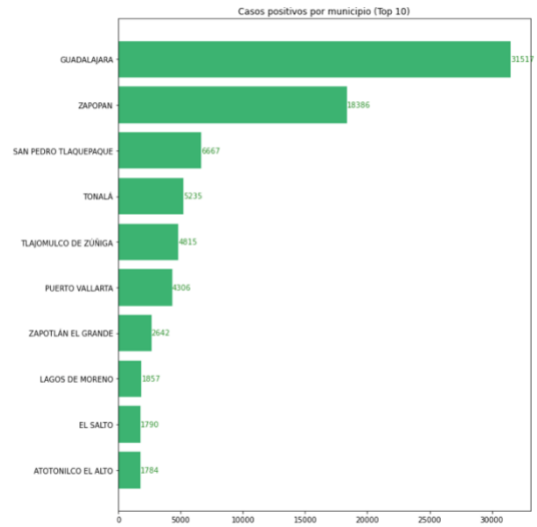
Esta entrega consistió en realizar el cálculo de los casos activos, que es una variable de interés a la hora de analizar el progreso de la pandemia. De igual manera, se realizó una comparación por municipio en Jalisco de las distintas variables de la entrega pasada, estas variables fueron casos positivos, defunciones, mortalidad y positividad. Por último, se realizó un análisis de las distintas olas de la pandemia, teniendo 3 en consideración.

Primero, iniciamos con los casos activos. Los casos activos al principio de la pandemia sobrepasaron a los que estuvieron presentes en las siguientes etapas, es decir, durante el primer año de la pandemia los casos activos, en promedio, fueron más que en las siguientes dos olas. Conforme pasó el tiempo y la población recibió vacunas se observa una disminución (a inicio del 2021), seguido un rebrote más adelante, el cual lo podríamos atribuir a la aparición de la variante OMICRON, la cual sabemos es más contagiosa. A continuación se muestra la gráfica de casos activos por día.

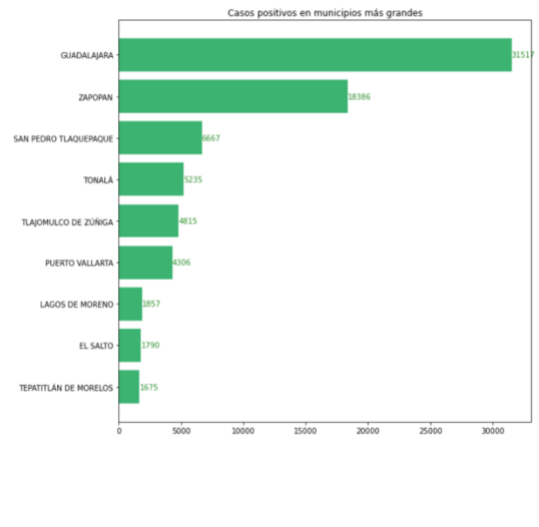


Pasando a la comparación por municipios del Estado de Jalisco, A continuación, se muestran los casos positivos por municipio (los 10 que más tuvieron), y los casos positivos en los municipios más poblados.

Casos positivos por municipio (top 10)



Casos positivos por municipio más poblado

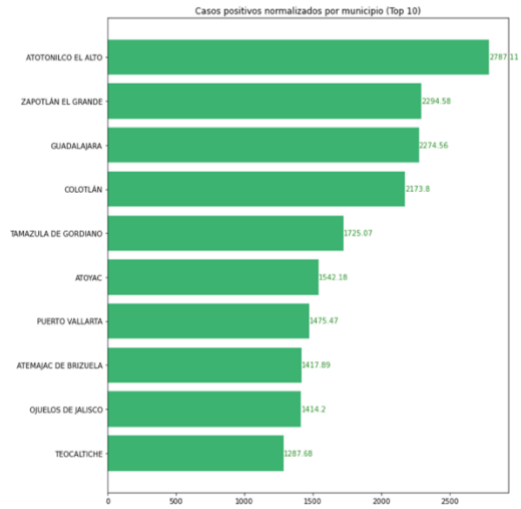


Lo primero que podemos observar con este comparativo, es que no necesariamente por tener una mayor población, se tienen más casos positivos, esto puede deberse a que ciertos territorios tienen mayor extensión o tienen mejor manejo de las medidas sanitarias, entonces mayor población no siempre causa más casos positivos.

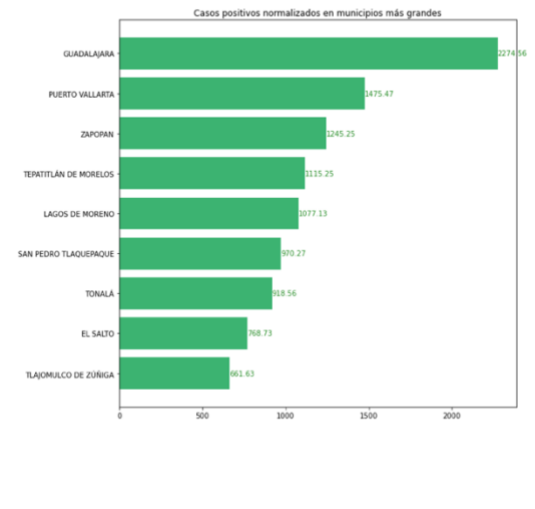
Sin embargo, las gráficas anteriores no son una comparación justa entre los municipios, para observar mejor estas comparaciones, es necesario normalizar, es decir, dividir entre la población y multiplicar por 100000.

A continuación vemos la importancia de escalar para poder comparar magnitudes diferentes, los municipios que parecían tener más casos positivos, están más abajo una vez que normalizamos.

Casos positivos normalizados por municipio (top 10)

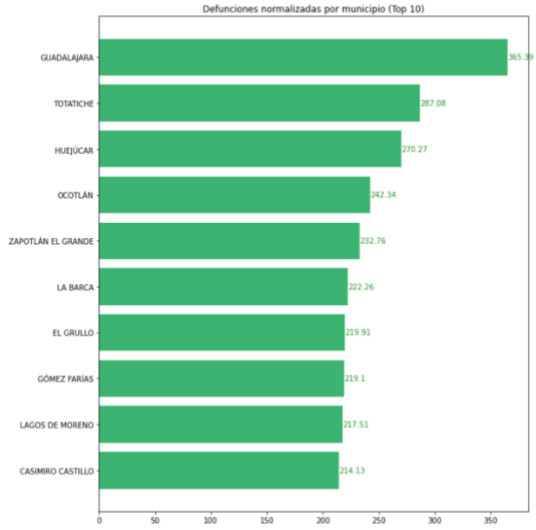


Casos positivos normalizados por municipio más poblado

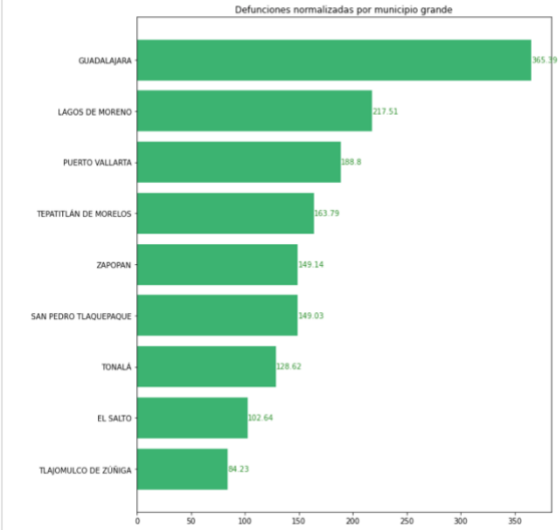


La siguiente gráfica muestra las defunciones por municipio, esta gráfica ya muestra los datos normalizados por cada 100,000 habitantes. Se puede apreciar que la cantidad de muertes no depende mucho del tamaño de la población, hay más varias defunciones por cada 100,000 habitantes en municipios más pequeños que los más poblados, lo cual podría deberse a una falta de acceso a servicios de salud.

Defunciones normalizadas por municipio (top 10)

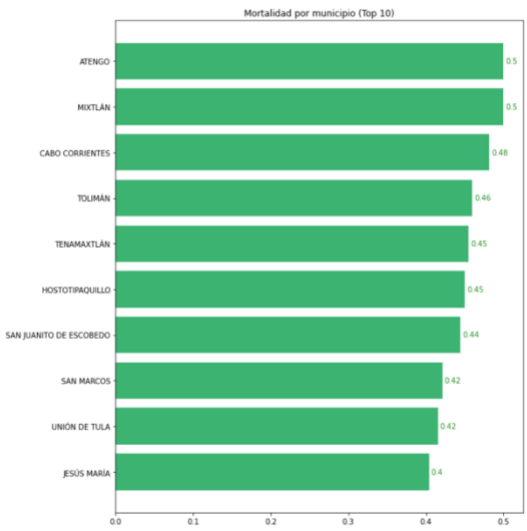


Defunciones normalizadas por municipio más poblado

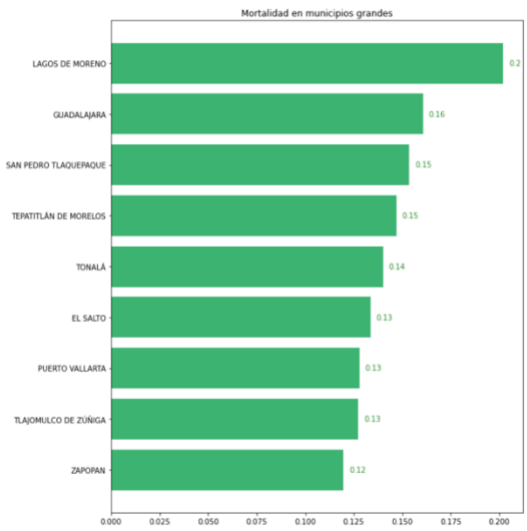


En cuanto a la mortalidad por municipio, la siguiente gráfica muestra que esta fue o ha sido menor en los municipios más poblados mientras que ha sido alta en municipios pequeños. Al igual que con las defunciones, es probable que esto se deba a que existe un mejor sistema de atención en municipios grandes ya que son zonas más pobladas.

Mortalidad por municipio (top 10)

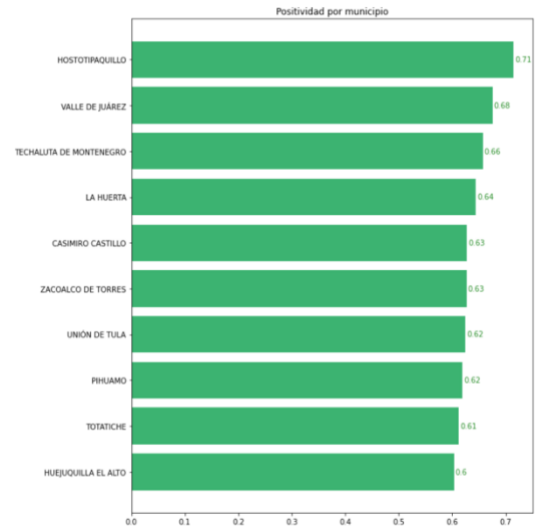


Mortalidad por municipio más poblado

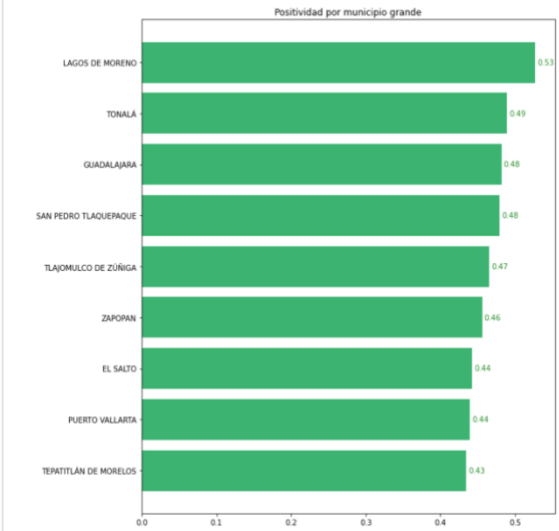


La positividad es más alta en municipios pequeños, sin embargo, es muy similar a la de los municipios más poblados. De igual manera, hay que considerar que el impacto en este indicador será mucho mayor si salen 10 personas positivas en un municipio muy pequeño que en uno muy grande, esto debido a que hay una menor población total.

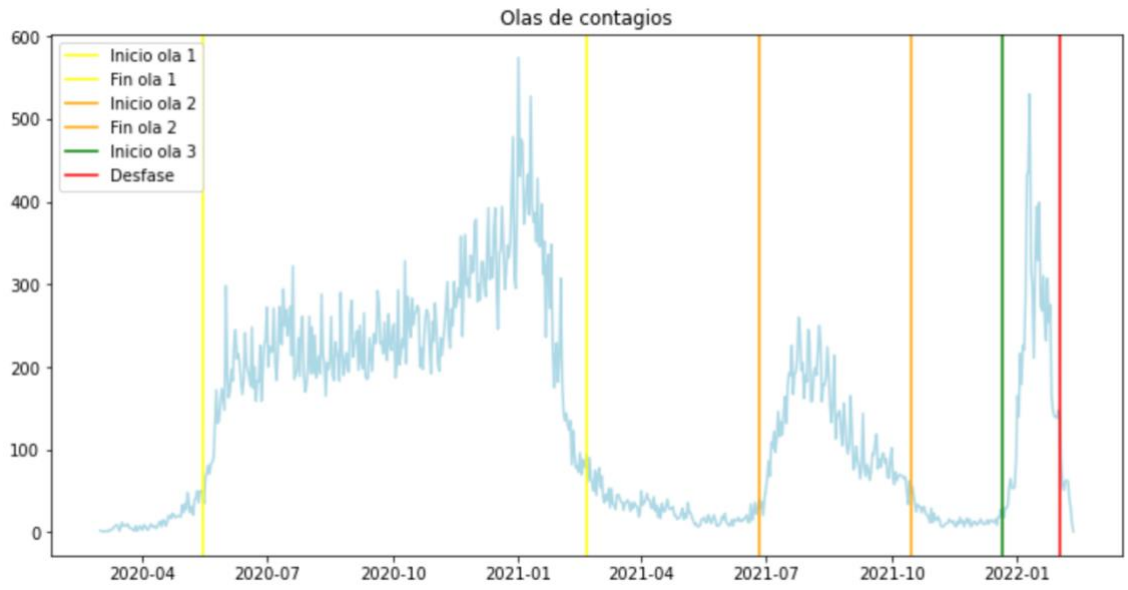
Positividad por municipio (top 10)



Positividad por municipio más poblado



A continuación, se presenta cómo se dividió la serie de tiempo de casos positivos en 3 olas.

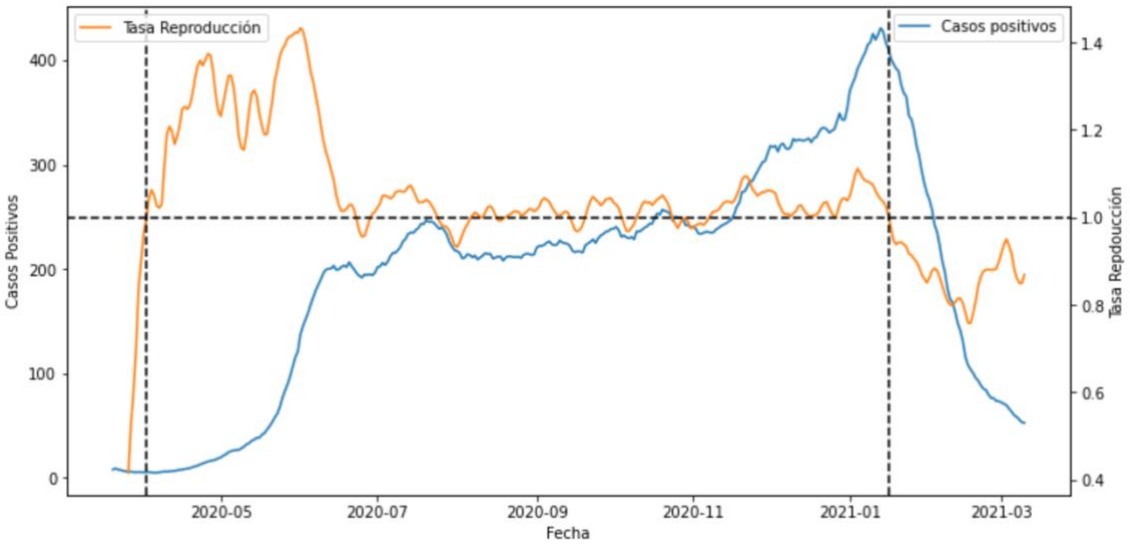


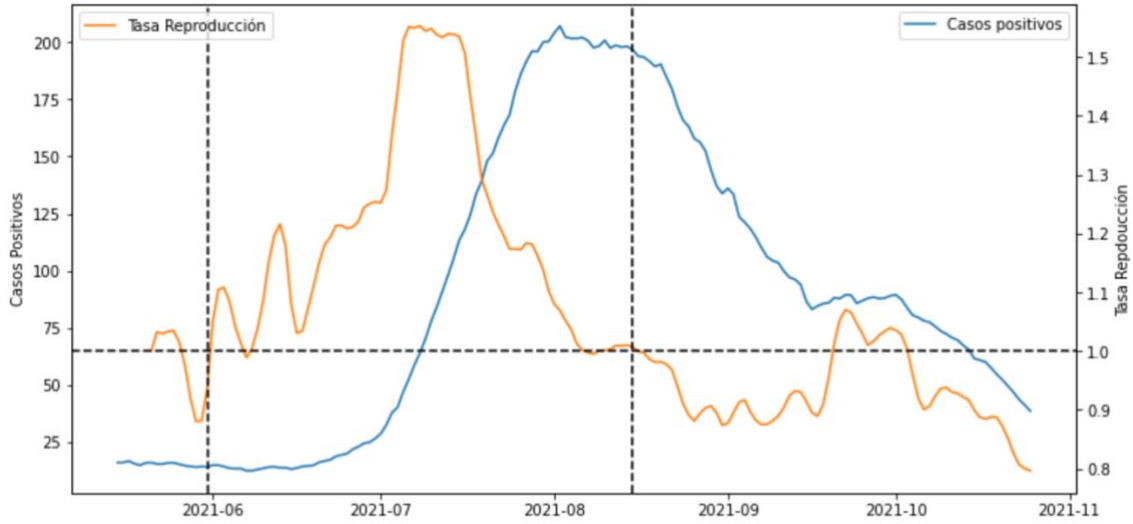
De cada una de las olas anteriores, se obtuvo la serie de tiempo de los casos positivos, las defunciones, la mortalidad, la positividad y la tasa de reproducción efectiva del virus, sin embargo, para poder comparar entre las 3, se obtuvieron también los promedios de estas medidas a lo largo de cada ola. A continuación, se muestran los resultados.

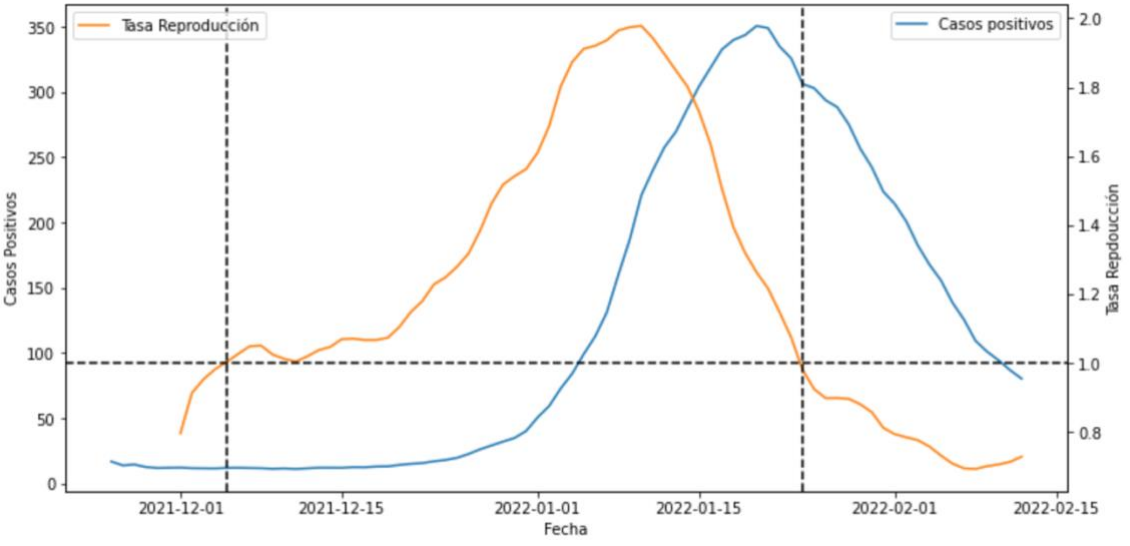
	Casos positivos totales	Casos negativos totales	Defunciones totales	Tasa reproducción promedio	Mortalidad total	Positividad total
Ola 1	68036	59837	10309.0	1.017452	0.151523	0.532059
Ola 2 (Delta)	13800	16664	2461.0	1.050438	0.178333	0.452994
Ola 3 (Omicron)	8796	6943	226.0	1.336796	0.025693	0.558867

Las olas de contagio siguen los patrones y las características que hemos escuchado sobre el virus. La primera ola se caracterizó por ser la más larga, además, al ser al inicio de la aparición del virus, no teníamos mucha idea de cómo se comportaba y las vacunas no existían, es por esto por lo que fue la que más casos positivos y defunciones tuvo. Sin embargo, al analizar la ola de la variante Delta, podemos observar que ha sido la más letal, teniendo la mortalidad más alta de las 3, concordando con la información que se nos hizo llegar la cual afirmaba que era más letal. De igual manera, la ola de la variante Omicron se caracterizó por ser altamente contagiosa pero no tan letal, esto lo vemos en la variable de mortalidad, la cual fue la más baja de las 3 olas, y en la variable de tasa de reproducción promedio, la cual fue la más alta de todas indicando que por cada persona contagiada, en promedio, se contagiaban 1.33 personas. Además de lo anterior, hay que considerar que la vacunación ha ayudado a que el virus ya no sea tan mortal, ayudando también a la disminución en la mortalidad.

El último punto a analizar en esta entrega tiene que ver con una relación importante entre los casos positivos y la tasa de reproducción efectiva del virus, a continuación, se muestran unas gráficas de cada ola (siendo la primera gráfica la primera ola y la última la tercera ola) para analizar dicha relación.


Tasa Reproducción
Casos positivos
Casos Positivos
Tasa Repdoucción
Fecha
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03


Tasa Reproducción
Casos positivos
Casos Positivos
Tasa Repdoucción
2021-06
2021-07
2021-08
2021-09
2021-10
2021-11


Tasa Reproducción
Casos positivos
Casos Positivos
Tasa Repdoucción
Fecha
2021-12-01
2021-12-15
2022-01-01
2022-01-15
2022-02-01
2022-02-15

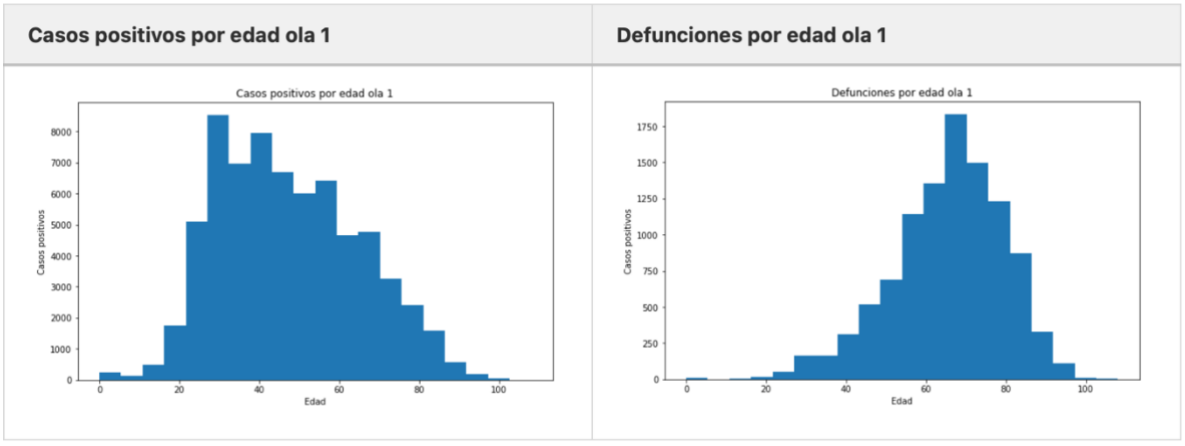
Analizando detenidamente las gráficas anteriores, especialmente las líneas punteadas, podemos observar que, en las 3 olas, cuando la tasa de reproducción efectiva del virus cruza el 1, es decir, se infecta más de una persona por infectado, hay un alza en los casos positivos a lo largo de los días posteriores. De igual manera, si la tasa de reproducción del virus disminuye después de haber estado por encima del 1, se puede observar una baja en los casos al pasar los días. Lo anterior, puede ser un buen indicador para controlar el surgimiento de nuevas olas, pudiendo activar las medidas necesarias unos días antes de que empiecen los contagio masivos y así tener una mejor probabilidad de evitar una nueva ola.

Entrega 4 – Distribución de edad por ola

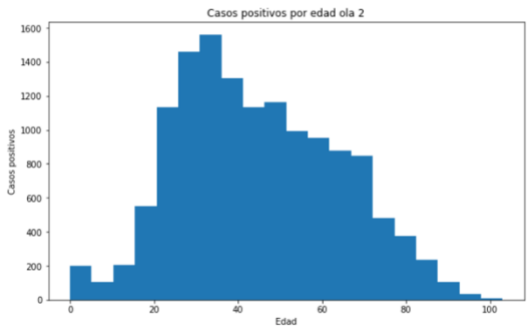
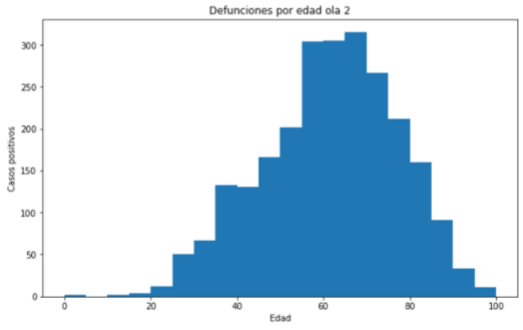
En esta entrega se analizaron los casos positivos y defunciones por edad, para cada ola, con el fin de observar quiénes fueron los más afectados en cada rubro. Se definieron las siguientes fechas como los inicios y fines de las olas:

- Ola 1: Inicio el 15 de mayo de 2020 y fin el 20 de febrero de 2021
- Ola 2: Inicio el 26 de junio de 2021 y fin el 15 de octubre de 2021
- Ola 3: Inicio el 21 de diciembre de 2021 hasta el día de hoy (menos 10 días)

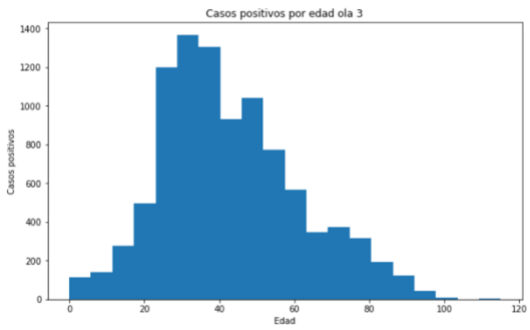
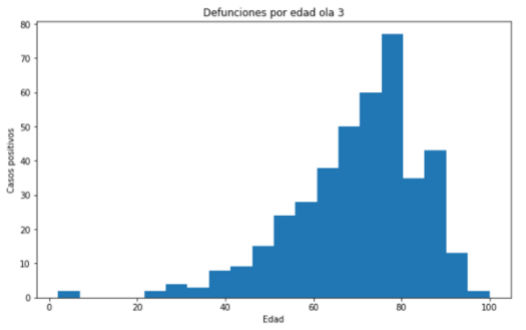
Tomando en cuenta lo anterior, la siguiente gráfica muestra los resultados de la ola número 1. Podemos observar que la mayor parte de los casos positivos se concentran entre 20 y 60 años, y el rango con mayores defunciones oscila entre los 60 y 80 años, confirmado que en efecto la enfermedad afecta de manera más grave a personas mayores.



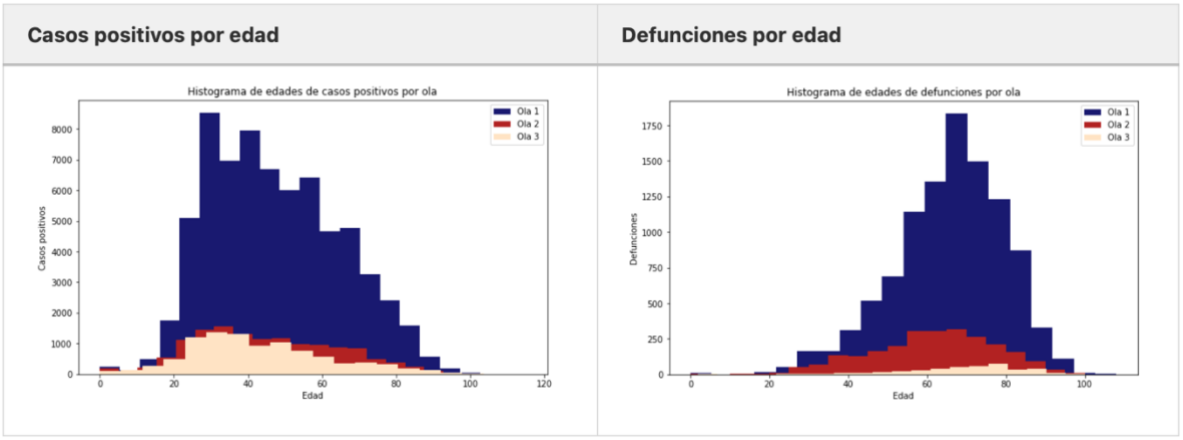
Al graficar los resultados de la ola número 2, la de la variante Delta, vemos que los resultados son relativamente similares, teniendo la mayor concentración de casos positivos las edades de entre los 20 y 50 años para esta ola. En cuanto a defunciones, la mayoría estaban también entre los 60 y 80 años de edad. Hay que considerar que para este momento ya habían vacunas, por lo que más adelante compararemos el número de defunciones entre olas.

Casos positivos por edad ola 2	Defunciones por edad ola 2
	

Para la última ola, la de la variante Omicron, la siguiente gráfica muestra que los casos positivos se concentran entre los 20 y 40 años, mientras que las defunciones siguen estando entre 60 y 80 años. Ambos totales disminuyeron considerablemente, sobre todo las defunciones.

Casos positivos por edad ola 3	Defunciones por edad ola 3
	

Podemos analizar las anteriores variables, casos positivos y defunciones, de todas las olas y hacer una comparación. Esto se muestra en la siguiente gráfica.



Aquí se observa qué ola fue más devastadora, siendo esta la primera ola, pero hay que tomar en cuenta que la primera ola duró mucho más que las siguientes, además de que no habían vacunas, medidas ni conocimiento por parte de los gobiernos para poder tomar acción en contra del virus. Parecería ser que cada ola es menos grave que la anterior, sin embargo, todas han tenido un impacto fuerte en la sociedad.

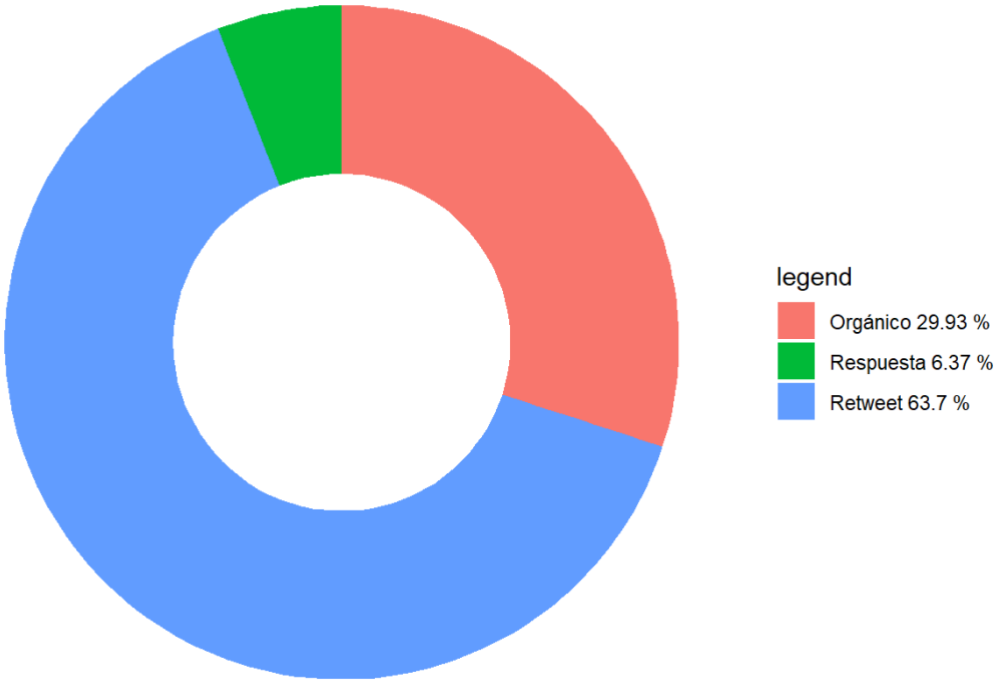
A partir de este análisis de las edades por ola, pudimos observar que las edades de casos positivos se repitieron en las 3 olas, siendo los jóvenes los que más se infectaron. Suponemos que esto se debe a que son las personas que más salen, debido a responsabilidades como la escuela y trabajo, además de posible negligencia de su parte (salir de fiesta, etc.). En cuanto a las defunciones, los más grandes (60-80 años), fueron los que más fueron perjudicados cuando salieron positivos. También, podemos ver la diferencia entre cuántas personas fueron infectadas o fallecieron en cada ola, observando que gracias a la vacunación y medidas preventivas, la frecuencia de ambas variables disminuyó conforme el paso del tiempo. Sería necesario estandarizar en función del tiempo los positivos y defunciones por ola para poder tener una comparación más objetiva.

Entrega 5 – Análisis de texto

Para el análisis de texto, se realizó una búsqueda de 6000 tweets con las palabras clave "COVID México", para posteriormente realizar un análisis de sentimiento, lugar de procedencia del usuario, followers, retweets, origen del tweet, fecha y hora de publicación, entre otras cosas.

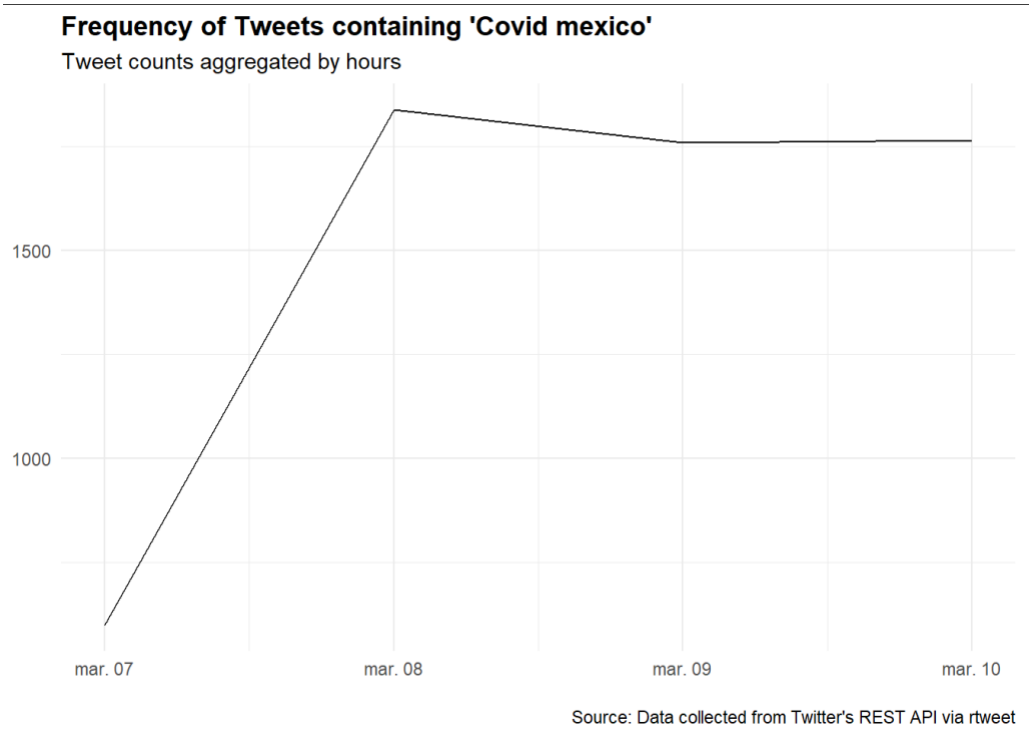
Las visualizaciones obtenidas se presentan a continuación:

- Proporción de tweets orgánicos, retweets y respuestas



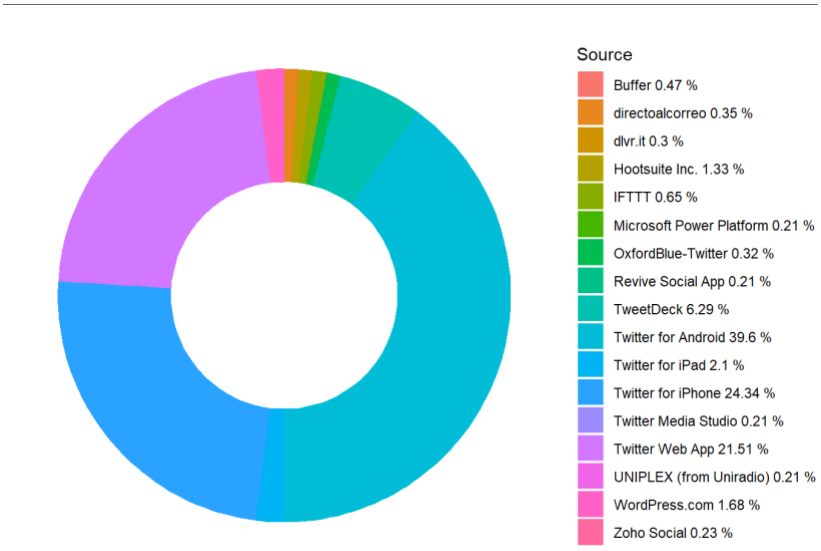
Un tweet orgánico se refiere a un tweet publicado directamente de alguna cuenta. Podemos observar que la mayor parte de los tweets obtenidos fueron retweets, casi 2/3 del total.

- Frecuencia de publicación por día



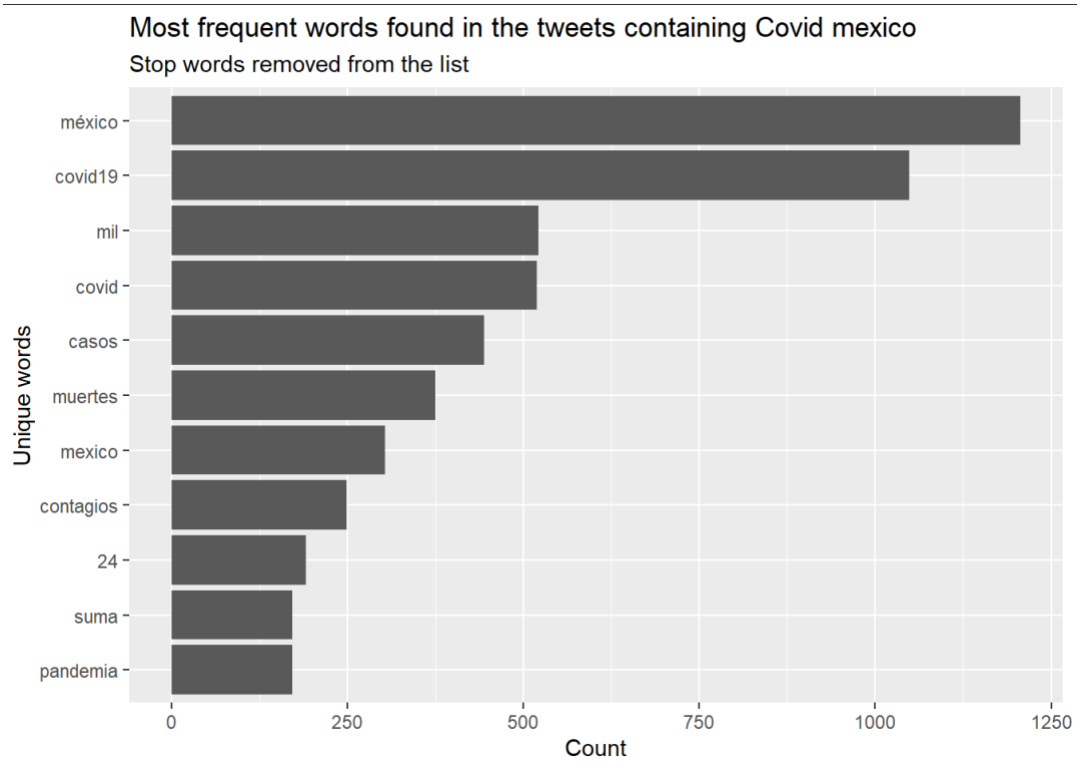
El 8 de marzo se observó una subida en el número de tweets, con una leve disminución en relevancia en los días posteriores.

- Origen de los tweets



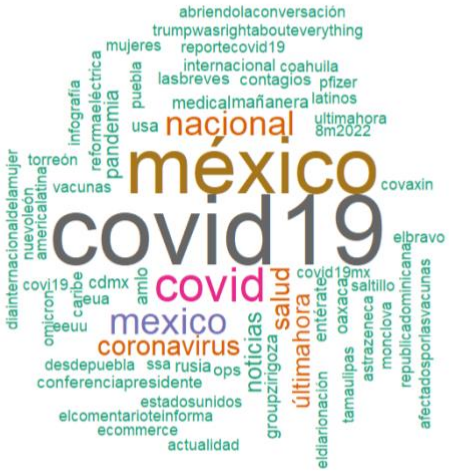
La gran mayoría de los tweets fueron publicados desde aparatos personales, lo que indica que son personas las que están publicando.

- Palabras más utilizadas en los tweets



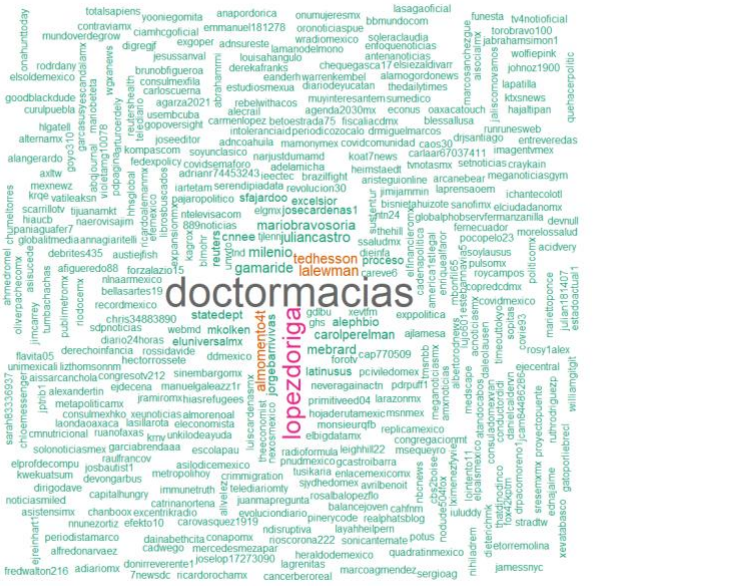
Las palabras más frecuentes nos hablan de casos y muertes, algo que podemos esperar escuchar hablando de "COVID México".

- Hashtags más frecuentes



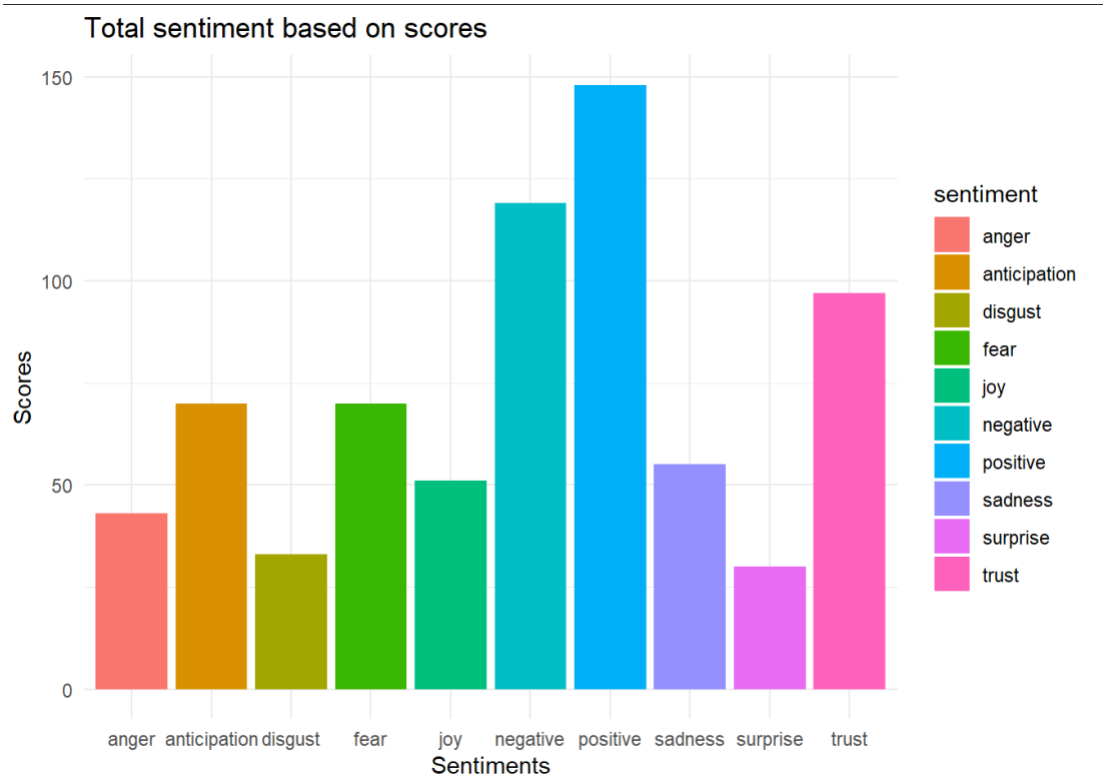
En cuanto a hashtags observamos que covid19, México, nacional, últimahora son los más relevantes, seguía habiendo bastante alerta por la pandemia en ese momento.

- Cuentas de donde se originan más retweets



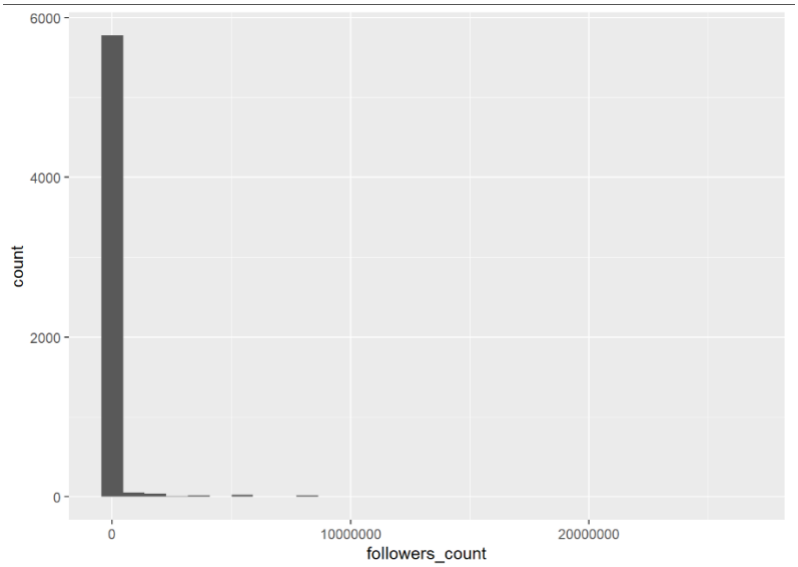
doctormacias, lopezdoriga, entre otros, generan más retweets.

- Análisis de sentimientos



Hay más sentimientos positivos que negativos, pero los negativos superan al sentimiento de confianza, que está en tercer lugar.

- Popularidad de los usuarios que tweetean



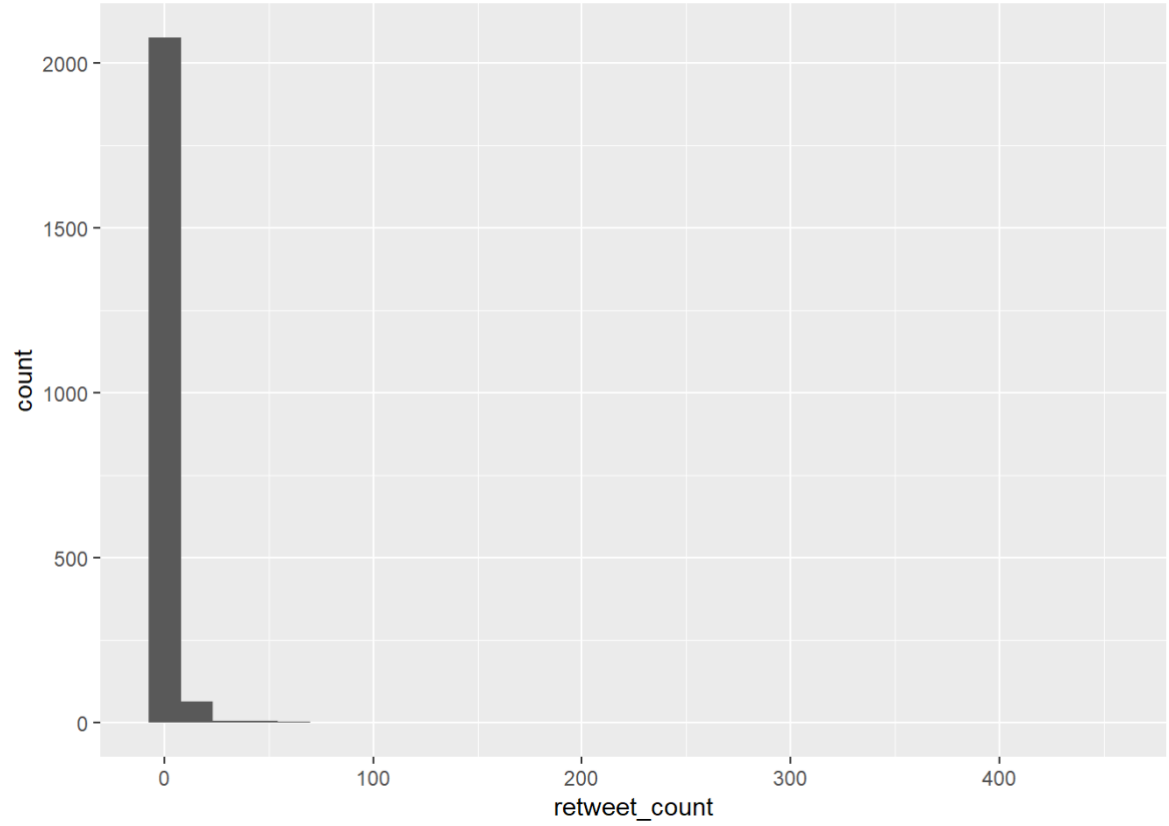
La mayor parte de los usuarios tienen pocos followers.

- Cuentas más populares

```
## # A tibble: 6 x 4
##   Twitter_Account followers_count location         text
##   <chr>                     <int> <chr>            <chr>
## 1 TheEconomist           26387141 London           "Today on "The Intelligence"~
## 2 Reuters                24709628 Around the world "U.S. leaning toward ending ~
## 3 Reuters                24709628 Around the world "U.S. leaning toward ending ~
## 4 CNNEE                  21103265 En todas partes  "Más de seis millones de per~
## 5 AristeguiOnline         9116895 Ciudad de México "#Entérate | Covid-19: Méxic~
## 6 AristeguiOnline         9116895 Ciudad de México "La diputada panista Margari~
```

The Economist, Reuters, CNN en español y AristeguiOnline son las cuentas más populares, algunas con más de 20 millones de followers.

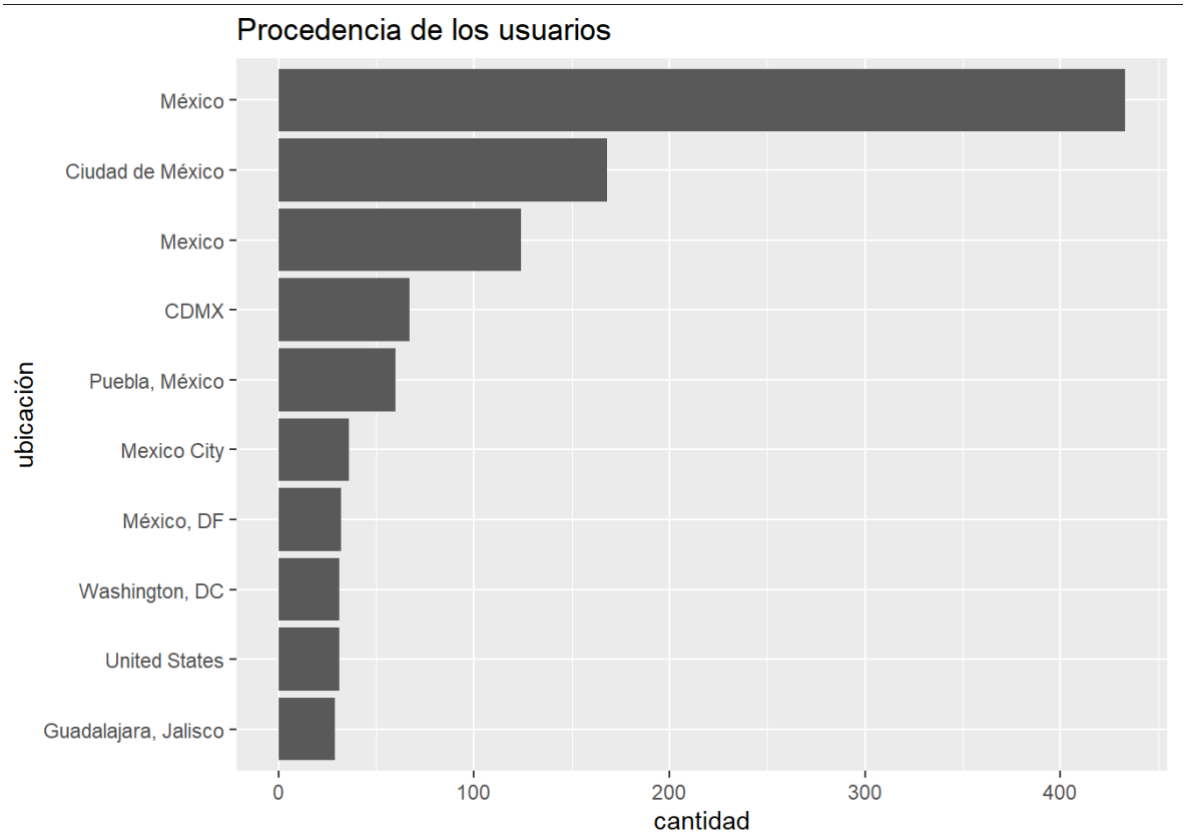
- Tweets más populares




```
## # A tibble: 1 x 5
##   Twitter_Account retweet_count followers_count location     text
##   <chr>                   <int>           <int> <chr>        <chr>
## 1 doctormacias              449          379394 Leon, Mexico COVID-19 por la va~
```

El tweet más popular es de doctormacias, mientras que los otros tweets casi no generan retweets.

- Procedencia de los usuarios



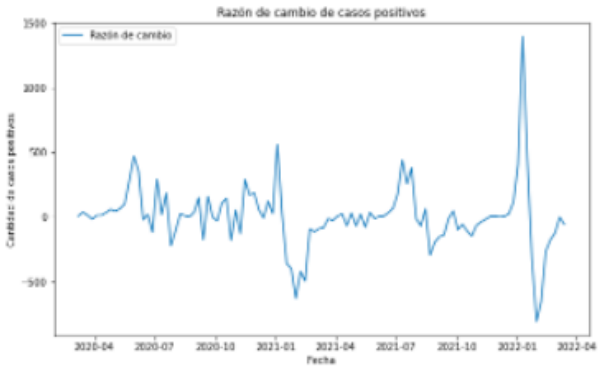
Casi todos los usuarios provienen de México, lo que es algo esperado dado el tema de la búsqueda.

Entrega 7 – Razones de Cambio

En esta entrega, se analizaron las razones de cambio de los casos confirmados y las defunciones (por diferencia relativa y porcentual), para luego observar las correlaciones entre la tasa de reproducción y la positividad. Obtuvimos los siguientes resultados.

La siguiente gráfica muestra la razón de cambio tanto porcentual como relativa de los casos positivos a través del tiempo.

Razón de cambio de casos confirmados

Razón de cambio casos confirmados (diferencia relativa)	Razón de cambio casos confirmados (diferencia porcentual)
	

La razón de cambio es en su mayoría positiva, por lo que se podría decir que los confirmados han estado creciendo más que disminuyendo.

Razón de cambio de defunciones

Razón de cambio defunciones (diferencia relativa)	Razón de cambio defunciones (diferencia porcentual)
	

En cuanto a defunciones, podemos ver que en un punto sí fue negativo con caídas muy fuertes, por lo que esta variable sí ha visto más disminuciones comparado con la de los casos positivos confirmados.

Correlaciones

Con Rt

	Correlación con casos positivos	Correlación con casos positivos (%)	Correlación con defunciones	Correlación con defunciones (%)
R_t	0.620445	0.953277	0.194194	0.509292

Con Positividad

	Correlación con casos positivos	Correlación con casos positivos (%)	Correlación con defunciones	Correlación con defunciones (%)
Positividad	-0.028097	-0.164427	0.283225	0.131893

Las defunciones han visto una disminución en cuanto a su razón de cambio, mientras que los casos positivos confirmados lo contrario. Hay una fuerte correlación positiva entre la tasa de reproducción efectiva y los casos confirmados positivos (como es de esperarse), mientras que las defunciones es mucho menor correlacionada, pero sigue habiendo un poco de correlación positiva.

En cuanto a la correlación de casos confirmados con la positividad, dio una correlación baja y negativa. Positividad y defunciones están correlacionadas de forma positiva pero mínima.

Entrega 8 – Análisis de clasificación de defunciones

Para esta entrega, el trabajo realizado consistió en obtener un modelo de clasificación que determine si una persona, **ya identificada como caso positivo de COVID-19 por una prueba de laboratorio**, va a morir o no. Es decir, al introducir datos nuevos, el modelo nos devuelve uno de dos posibles resultados, la persona muere o la persona no muere. Dicho modelo se basará en ciertas variables que se explicarán más adelante y en la base de datos del SINAVE, la cual cuenta con varios casos positivos.

Variables de entrada y variable objetivo

Primero que nada, definimos que la variable objetivo sea la fecha de defunción del paciente, recordando que sí existe dicha fecha significa que la persona falleció y, en caso de que no existe, la persona no falleció. Lo anterior hará que el modelo pueda determinar si la persona fallecerá o no. Para trabajar de una mejor manera, dicha variable se convirtió en variable dummy, donde el número 1 indica que la persona había fallecido (existía una fecha de defunción), mientras que un 0 indica que la persona no falleció.

Las variables de entrada del modelo se eligieron pensando en la importancia que estas podrían tener en el resultado del paciente, dichas variables fueron las siguientes 18:

- Sexo
- Tipo de paciente
- Edad
- Intubado
- UCI
- Neumonía
- Embarazo
- Diabetes
- EPOC
- Asma
- Inmunosupresión
- Hipertensión
- Otra comorbilidad
- Cardiovascular
- Obesidad
- Renal crónica
- Tabaquismo
- Otro casoç

Todas las variables anteriores, salvo la edad, también se convirtieron en variables dummies. Por ejemplo, si la persona padecía diabetes, se marcó con un 1, mientras que si la persona no padecía diabetes se marcó con el número 0. Con esto, todas las variables se volvieron 1 y 0 menos la edad. A continuación, se muestra una pequeña referencia de cómo quedaron los datos.

Personas Fallecidas

SEXO	TIPO_PACIENTE	EDAD	INTUBADO	UCI	NEUMONIA	EMBARAZO
0	0	53	0	0	0	0
1	0	23	0	0	0	0
1	0	28	0	0	0	0
1	0	23	0	0	0	0
1	0	31	0	0	0	0
...	...	...	...	...	...	...
1	1	78	0	0	1	0
1	1	50	0	0	0	0
1	1	79	0	1	1	0
1	1	35	0	0	1	0
1	1	62	1	1	1	0

Personas Fallecidas Confirmadas

Defuncion
0
0
0
0
0
...
1
1
1
1
1

Desbalance de las clases

La base de datos presenta alrededor de 82,000 casos positivos confirmados por prueba en donde el paciente no falleció, mientras que hay alrededor de 15,000 casos positivos confirmados por prueba donde el paciente falleció, es decir, existe un desbalance de clases habiendo más no fallecidos que fallecidos.

Por lo tanto, tomamos una muestra aleatoria del 18% de casos positivos confirmados por prueba en donde el paciente no falleció, teniendo así 15,000 muestras de cada una de las clases, es decir, 15,000 personas que fallecieron y 15,000 personas que no, dando un total de 30,000 datos para obtener el modelo.

Split y pre-procesado de variables

Teniendo los datos para el modelo, realizamos un split de los mismos, en donde tomamos el 80% de ellos como datos de entrenamiento y el 20% restante como datos de prueba. A continuación, se muestran las cantidades exactas de datos.

```
Train set: (23872, 18) (23872,)
Test set: (5968, 18) (5968,)
```

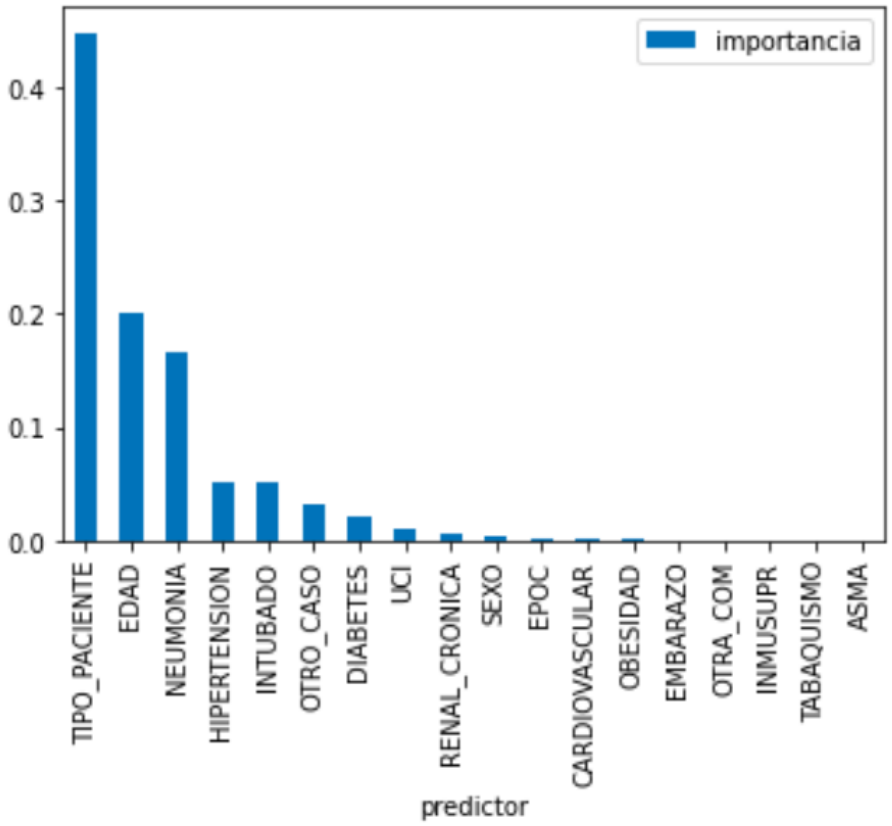
Como podemos observar, los datos entrenamiento de entrada son 23,872 y consta de 18 variables, mientras que los datos de salida del entrenamiento también son 23,872, pero, en este caso únicamente la variable de si la personas falleció o no.
Posteriormente, y como pudimos ver en la pequeña muestra de las variables de entrada, la edad es un valor muy diferente a las demás variables, siendo esta la única que no muestra solamente 1 y 0. Por lo tanto, decidimos normalizar las variables de entrada con el máximo de cada una, esto para que la edad quede como número positivo. Debido a que las demás variables son 1 y 0, el máximo será 1, por lo que estas no cambian, sin embargo, la variable de edad sí cambia y quedan valores más cercanos a 1 y 0.

Modelado y tuneado de hiperparámetros

El modelado y tuneado de los hiperparámetros se realizó con una función ya integrada en Python, se realizó un grid como en la sesión de aprendizaje y se corrieron varias combinaciones de hiperparámetros para encontrar los mejores. Después, de igual manera con

una función de Python, se creó el modelo y se evaluó el desempeño de este. Dicho desempeño se muestra en la sección de resultados.

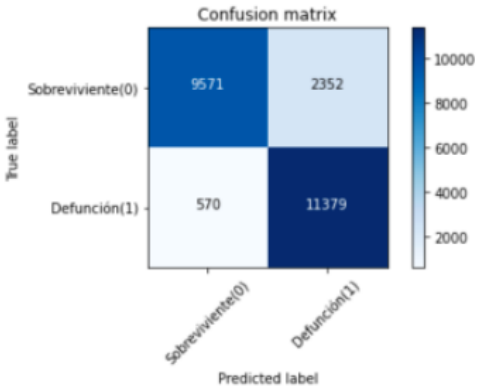
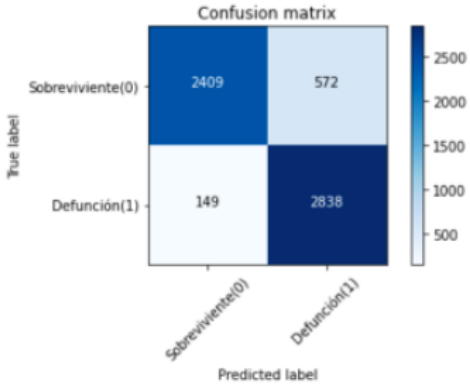
Un detalle importante para considerar es que, una vez que obtuvimos un modelo con "random forest", observamos la importancia de cada uno de los predictores (variables de entrada). Notamos que había varios que no aportan mucho al modelo, estos se muestran a continuación:

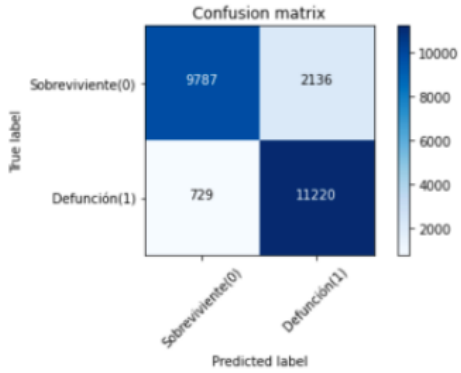
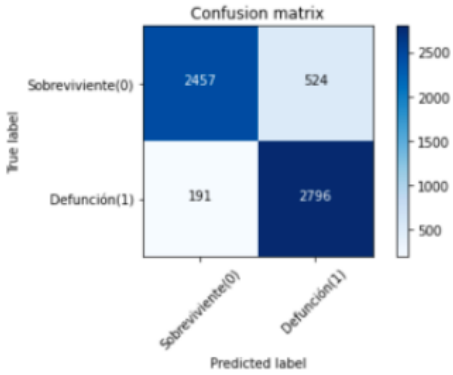


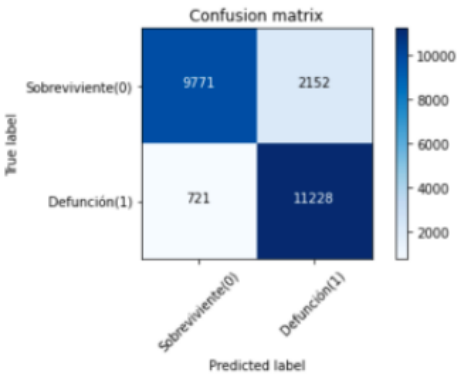
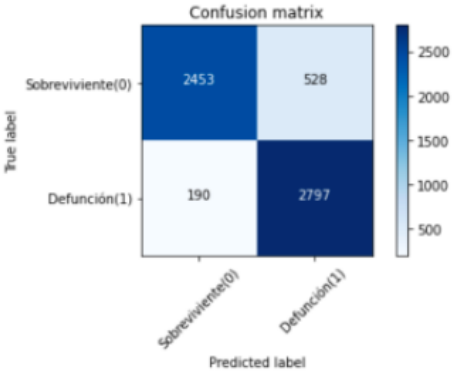
Considerando lo anterior, eliminamos todas las variables en las que la importancia de los predictores fuera menor a 0.0008 y volvimos a crear un nuevo modelo con estas nuevas variables. Los resultados de este también se muestran en la sección de resultados.

Resultados

A continuación, se muestran los resultados de las medidas de desempeño de los modelos calculados, así como sus matrices de confusión.

SVM Train	SVM Test
	

Random Forest Train	Random Forest Test
	

Random Forest 2 (LV) Train	Random Forest 2 (LV) Test
	

Con base a las matrices de confusión anteriores, obtuvimos las siguientes medidas de desempeño para cada modelo:

Medidas	Score_SVM test	Score_SVM train	Score_Random Forest test	Score_Random Forest train	Score_Random Forest less var test	Score_Random Forest less var train
Accuracy	0.879189	0.877597	0.880194	0.879985	0.879692	0.879650
Precision	0.832258	0.828709	0.842169	0.840072	0.841203	0.839163
Recall	0.950117	0.952297	0.936056	0.938991	0.936391	0.939660
F1 score	0.878570	0.876901	0.879813	0.879559	0.879298	0.879208
Jaccard defunciones	0.797415	0.795679	0.796354	0.796592	0.795733	0.796256
Jaccard sobrevivientes	0.769649	0.766109	0.774590	0.773554	0.773573	0.772778

Con base a la última tabla presentada, encontramos que el mejor modelo es el de Random Forest debido a que las medidas de desempeño son ligeramente superiores en comparación con los otros modelos, sin embargo, al ser todas similares, podríamos tomar cualquiera como un buen modelo. De igual manera, creemos que el modelo podría mejorarse y tener medidas mucho mejores, sin embargo, considerando que estamos trabajando con 30,000 datos, el desempeño es aceptable.

1.6. Valoración de productos, resultados e impactos

Con toda la información analizada a lo largo de las semanas de trabajo del proyecto, logramos encontrar ciertos patrones que ayudan a identificar y caracterizar la enfermedad del COVID-19. Primero, por ejemplo, como ya se nos ha informado, los datos arrojan que las personas que se ven más afectadas son gente de 3ra edad, de igual manera, la obtención de las variables como positividad, mortalidad, tasa de reproducción efectiva del virus, entre otras, nos ayudó a ver el comportamiento de la pandemia a través del tiempo y a través de las diferentes olas.

Una de las cosas más relevantes que encontramos y se presentó en la sección anterior, es el hecho de que cuando la tasa de reproducción efectiva del virus supera 1, es decir, por cada persona contagiada se contagia más de una, tiempo después viene un alza considerable en el número de casos positivos. De igual manera, si esta tasa, después de estar en por encima de 1, baja, habrá una tendencia también a la baja en los casos positivos días después. En general,

todos los cálculos y visualizaciones obtenidas ayudaron a tener un mejor entendimiento de la enfermedad desde varios niveles.

Cabe destacar también que, en nuestros análisis se confirmó que el tener una comorbilidad (Diabetes, Hipertensión, etc.) aumenta en gran medida el riesgo de morir por haber contraído la enfermedad. Esto lo pudimos notar tanto en la parte de análisis exploratorio y visualizaciones, como en la parte de modelado. La importancia que estás variables tenía en cuanto al resultado, era mayor que las de otro tipo de variables

En cuanto a la parte del modelado del proyecto, logramos obtener un modelo de clasificación de Random Forest con medidas de exactitud que se presentan en la sección de productos (2) de este proyecto. La idea del modelo era que al utilizar información del pasado, se pudiera determinar si una persona moriría o no de acuerdo con sus síntomas y otras variables relevantes como la edad. Dicho modelo, junto con el análisis de datos realizado, puede ser información relevante a futuro para que diferentes organizaciones de salud puedan tomar las medidas adecuadas contra el virus.

Asimismo, este PAP ayudó a adquirir experiencia analizando datos reales y relevantes en comparación de datos teóricos que usualmente revisamos en clase. Esto nos sirve para ver y experimentar problemas reales a los cuales nos podemos enfrentar al momento de querer analizar y obtener información relevante de cualquier base de datos. En pocas palabras, el proyecto nos da experiencia real sobre cómo tratar una base de datos y desde qué enfoques analizarla para que después de entenderla, se pueda crear e implementar algún modelo de inteligencia artificial para explicar algún comportamiento.

1.7. Bibliografía y otros recursos

- Gobierno de México. (2022). Datos Abiertos Dirección General de Epidemiología. Mayo 05, 2022, de Gobierno de México Sitio web: https://www.gob.mx/salud/documentos/datos-abiertos-152127
- Eisenberg, J.. (2020). Qué es el R0, el número que siguen los científicos para ver la intensidad del coronavirus. Mayo 05, 2022, de The Conversation Sitio web: https://theconversation.com/que-es-el-r0-el-numero-que-siguen-los-cientificos-para-ver-la-intensidad-del-coronavirus-137744
- Adam Rose. (2022). How has COVID-19 Impacted our Economy?. 27 de enero de 2022, de News Medical Sitio web: https://www.news-medical.net/news/20220114/How-has-COVID-19-Impacted-our-Economy.aspx
- Kimberly Chriscaden. (2020). Impact of COVID-19 on people's livelihoods, their health and our food systems. 27 de enero de 2022, de WHO Sitio web: https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people%27s-livelihoods-their-health-and-our-food-systems
- WHO. (-). Coronavirus. 27 de enero de 2022, de WHO Sitio web: https://www.who.int/es/health-topics/coronavirus#tab=tab_1
- The Economist. (-). Omicron latest: Omicron and the global economy. 27 de enero de 2022, de The Economist Sitio web: https://www.economist.com/omicron
- Banco Mundial. (2021). La economía mundial: en camino hacia un crecimiento firme, aunque desigual debido a los efectos perdurables de la COVID-19. 28 de enero de 2022, de Banco Mundial Sitio web: https://www.bancomundial.org/es/news/feature/2021/06/08/the-global-econom y-on-track-for-strong-but-uneven-growth-as-covid-19-still-weighs
- Brownlee, J.. (2020). Tune Hyperparameters for Classification Machine Learning Algorithms. Mayo 05, 2022, de Machine Learning Mastery Sitio web: https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/
- Amat, J.. (2020). Random Forest con Python. Mayo 05, 2022, de cienciadedatos.net Sitio web: Random Forest con Pytho

1.8. Anexos generales

Anexo 1. Repositorio con todos los scripts (Revisar solo los del Equipo 1 para este reporte)

- https://gitlab.com/dpmontoya/pap-mmd-p2022

Anexo 2. Cronograma del proyecto:

ISSUES DEL PROYECTO	Semana 1 22-ene	Semana 2 29-ene	Semana 3 05-feb	Semana 4 12-feb	Semana 5 19-feb	Semana 6 26-feb	Semana 7 05-mar	Semana 8 12-mar	Semana 9 19-mar	Semana 10 26-mar	Semana 11 02-abr	Semana 12 09-abr	Semana 13 16-abr	Semana 14 23-abr	Semana 15 30-abr	Semana 16 07-may	Semana 17 14-may
BLOQUE 1: Entender el problema y los datos																	
Entender el problema de COVID, principales preguntas que están surgiendo.																	
Entender los datos de SINAVE																	
Limpieza y transformación de los datos																	
BLOQUE 2:Análisis de datos																	
Análisis general, compartativa entre estados																	
Análisis entre olas de contagio																	
Análisis de casos hospitalizados																	
Razones de cambio y velocidades																	
Análisis por grupos de edad																	
Análisis de texto																	
Análisis extras																	
BLOQUE 3: Modelados																	
Pronóstico de casos confirmados, hospitalizados																	
Modelado de covid (regresión y/o clasificación)																	
BLOQUE 4: Entrega del proyecto																	
Reporte PAP																	
Presentación PAP																	

2. Productos

Los productos obtenidos en este proyecto son virtuales, estando contenidos en el repositorio del **Anexo 1**. Se trabajó brevemente en R al inicio del proyecto, después se realizaron todos los scripts en Python a excepción del análisis de texto de Twitter. El producto principal es el modelo de machine learning para clasificar defunciones y no defunciones, el cual se puede volver a entrenar con un set de datos nuevos para utilizarse en un futuro. La utilidad dependerá de cuánto tiempo más estemos en estado de alerta por la pandemia, pero podría reajustarse a otro Dataset de alguna otra enfermedad que surja a futuro, con sus debidas modificaciones a limpieza de datos y selección de variables importantes.

3. Reflexión crítica y ética de la experiencia

El RPAP tiene también como propósito documentar la reflexión sobre los aprendizajes en sus múltiples dimensiones, las implicaciones éticas y los aportes sociales del proyecto para compartir una comprensión crítica y amplia de las problemáticas en las que se intervino.

3.1 Sensibilización ante las realidades

Andrés Duarte: Desde el principio me fui sintiendo involucrado en lo que les sucedió a todas las personas que fueron contagiadas, al analizar cuántas personas fallecieron debido a complicaciones derivadas de haberse enfermado de Covid-19. Creo que la labor que nosotros realizamos en este proyecto es importante porque tiene un impacto social hasta cierto punto. La idea de analizar a fondo información de este estilo, tiene un trasfondo de intentar ayudar a los gobiernos a salvar vidas. Eso me hizo sentir feliz de estar realizando el trabajo, ya que para mí es importante que la labor que yo realice tenga un propósito. Si podemos aplicar conocimientos de Ciencia de Datos para algo tan importante como evitar fallecimientos, me parece que se está aplicando esfuerzo a una buena causa.

Pablo Rivera: Estos últimos dos años de pandemia han sido difíciles para todos, desafortunadamente la enfermedad ha afectado a varias personas tanto económicamente, pero sobre todo en temas de salud. En varias ocasiones me he sentido de cierta manera un tanto impotente, sintiendo que no hay nada que hacer ante esta situación más que cuidarme y seguir las medidas que el gobierno indica. Pensaba que era un problema que no podría ayudar a resolver, sin embargo, este proyecto me ayudó un poco a liberar este sentimiento de impotencia y vi una manera en la que, desde mi carrera, puedo ayudar a una situación real y difícil que, si se encuentra una solución, podría inclusive salvar vidas. Me emociona poder, en un futuro, poder resolver problemas para el beneficio de la sociedad aplicando herramientas que apliqué y aprendí en este proyecto, así como muchas otras que he ido adquiriendo a lo largo de la carrera.

Iván Lafárga: Durante el desarrollo de este pap me di cuenta una de las cosas más importantes que tienen que tener en cuenta todos los analistas al revisar los casos, esto es la sensibilidad del tema, desarrollando el proyecto tenemos muchos resultados, tratamiento de datos, desarrollo de conclusiones, lo más importante fue tener en cuenta que los números que estamos trabajando, así como las lecturas de estos mismos son personas con realidades, vida, familia y seres queridos. Tener en cuenta que la base de datos mayormente se trata de personas hay que tener en cuenta el lado humano al tener conclusiones sobre el trabajo realizado, así como las predicciones o decisiones sobre estas de estas misas, tomando en cuenta esto podemos utilizar nuestros conocimientos en función del beneficio de la sociedad aportando soluciones funcionales que logren salvar vidas, mantener a familias unidas y utilizar el poder del conocimiento, así como la práctica de este mismo para cambiar el mundo o nuestras circunstancias a un espacio mejor para los seres humanos y en general la tierra, considero tener un enfoque positivo el análisis de datos puede hacer grandes cambios de manera positiva haciendo uso de este para soluciones en beneficio del entorno que nos rodea.

3.2 Aprendizajes logrados

Andrés Duarte: Logré desarrollar capacidades de reportar hallazgos de manera escrita y verbal, esa habilidad de traducir código a una explicación concisa de lo que se hizo y encontró. Sentí el reto a la hora de realizar el modelado, y a la hora de apegarme al cronograma de actividades. Fue mejor de esa forma, ya que construimos todo el trabajo a lo largo del semestre para solamente recopilarlo, pero era estar constantemente trabajando y teniendo que cuidar el tiempo para no fallar con entregas. Creo que crecí como científico de datos.

Pablo Rivera: Durante el proyecto, aprendí primero que pasos a seguir al momento de analizar una base de datos y posteriormente crear algún modelo con herramientas de inteligencia artificial. Además, logré ver y aprender cómo llevar a cabo dicho análisis de datos de manera detallada, desde qué tipo de cosas buscar en los datos hasta como realizar ciertas visualizaciones de los mismos para poder entenderlos mejor. Finalmente, a pesar de haber llevado una clase sobre el tema, aprendí cómo aplicar un modelo de clasificación a una base de datos (la cual ya habíamos trabajado y analizado durante el proyecto), lo cual se me hizo enriquecedor ya que eran datos reales acerca de un problema real. Uno de los mayores retos creo que fue organizar los tiempos de las entregas, así como desarrollar el código de cada una de ellas en tiempo y forma. En algunas ocasiones surgían ciertos obstáculos con respecto a cosas que no estaba seguro de cómo programar, sin embargo, gracias al proyecto, me he dado cuenta que soy capaz y me gusta trabajar hasta encontrar una solución óptima a cualquier problema que me encuentre, sobre todo si tiene que ver con programación. Este proyecto, junto con otras materias que he llevado, me han ayudado a darme cuenta que me gustaría, en un futuro, dedicarme a temas relacionados al análisis de datos y a solucionar problemas con estos y con diferentes modelos que he aprendido a utilizar.

Ivan Lafarga: Logré concientizar mucho con los datos, saber que los números que veía en la pantalla eran personas me hizo involucrarme de manera distinta a otros proyectos por ser estos resultados seres humanos que contaban con familiares, amigos y seres queridos dejo de darle un valor numérico al trabajo convirtiéndolo en un trabajo más humano asumiendo valores y resultados como personas, cosa que nunca antes me había tocado experimentar , esto logro mover fibras sensibles en mi personas de las cuales también se cuenta como un aprendizaje dentro de este pap. Una de las enseñanzas que adquiero con esta experiencia es que los datos no se deben de trabajar directamente, tenemos que aprender, entender como los datos que obtenemos, un análisis exploratorio de los datos entendiendo sus circunstancias e influencias al momento de ser registrados para saber su correcta operación, uno de los retos durante el desarrollo de este proyecto fue entender que existen muchas soluciones para un problema pero solo uno es la más viable por lo tanto conocer, experimentar , trabajar y experimentar con distintas soluciones provoco un reto en la percepción de las soluciones del

mismo para encontrar la mejor solución con base a nuestro pensamiento crítico donde fuéramos muy puntuales al llegar a una solución dentro del método o del resultado contundente al final ya que muchas maneras de resolver un problema te expande la manera de ver el proyecto a muchas diferentes perspectivas de una misma problemática acercándonos a pensar en una gran cantidad formas distintas de ver algo que se considera que es lo mismo pero las diferentes maneras de ver un mismo problema te convierte a ver esto como muchos problemas distintos, esto fue una de las cosas que más lograron retarme en esta experiencia profesional. Coincidí lograr una sinergia entre la parte práctica y la teórica ya que no sólo se desarrolló código sino una interpretación de los resultados que surgían del mismo para exponerlos de manera sencilla como resultados puntuales, concisos y fácil de entender a pesar del difícil logro que fue llegar a ellos, así como saber interpretarlos de manera correcta la explicación debe ser sencilla y fácil de entender, como experiencia personal lograr esta sinergia fue muy importante para los aprendizajes de este curso además de darle una exposición visual muy corporativa, amigable y entendible para personas que no estuvieran muy relacionadas a la parte práctica del tema.