Universidad del Desarrollo
Facultad de Ingeniería

IMPACTO VAC-COVID19-CHILE: ANÁLISIS DE LA VACUNACIÓN Y SU EFECTO EN LA DINÁMICA PANDÉMICA

Correlaciones y Tendencias en la Era de la Vacunación: Un Estudio Detallado de Casos, Hospitalizaciones y Mortalidad en Chile

POR: RICARDO MIRANDA ARAYA Y SEBASTIAN DANKER GALDAMES

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

PhD. MAURICIO HERRERA MARIN

Diciembre 2023

SANTIAGO

Aprender a transformar datos en información es la clave para el futuro.

AGRADECIMIENTO

Es con un profundo sentido de gratitud que nosotros, Ricardo y Sebastian, nos dirigimos a ustedes al concluir nuestro viaje en el Máster en Ciencia de Datos. Estos 18 meses han sido un período de intenso aprendizaje, desafíos y crecimiento personal y profesional, y sabemos que no hubiéramos podido alcanzar este hito sin su inestimable apoyo.

A nuestras familias: Gracias por ser nuestro soporte incondicional en los momentos más exigentes. Vuestro aliento, comprensión y paciencia han sido fundamentales para mantener nuestro enfoque y motivación. Vuestra creencia en nosotros ha sido una fuente constante de fuerza.

A nuestros profesores: Su dedicación, conocimiento y pasión por la enseñanza han sido inspiradores. Nos han equipado no solo con habilidades técnicas en ciencia de datos, sino también con una mentalidad crítica y un enfoque ético hacia nuestro campo. Las lecciones aprendidas van más allá del aula y nos acompañarán a lo largo de nuestra carrera.

Este viaje no solo ha sido académico; ha sido una experiencia enriquecedora que nos ha enseñado sobre la perseverancia, el trabajo en equipo y la importancia de la comunidad. Estamos agradecidos por haber tenido la oportunidad de aprender de cada uno de ustedes y por la comunidad que hemos formado juntos.

A medida que avanzamos hacia la próxima etapa de nuestras vidas profesionales, llevamos con nosotros no solo el conocimiento adquirido, sino también los recuerdos y las relaciones forjadas durante este tiempo. Esperamos poder aplicar lo aprendido y contribuir significativamente a nuestra área.

Gracias una vez más por su invaluable apoyo y por ser parte de esta importante etapa de nuestras vidas.

TABLA DE CONTENIDO

RESUMEN **1**

1. INTRODUCCIÓN **2**

2. TRABAJO RELACIONADO **3**

EVALUACIÓN DE LA EFECTIVIDAD DE LA VACUNACIÓN CONTRA COVID-19 EN CHILE 3

ESTUDIOS SOBRE LA EFICACIA DE LA VACUNA 4

EVALUATION OF COVID-19 VACCINE EFFECTIVENESS 4

3. HIPÓTESIS Y OBJETIVOS **5**

HIPÓTESIS 5

OBJETIVO GENERAL 5

Objetivos Específicos *5*

4. DATOS Y METODOLOGÍA **6**

4.1. DATOS 6

4.2. METODOLOGÍA 13

Librerías Python *13*

Limpieza y Preprocesamiento de Datos: *16*

Análisis Exploratorio de Datos (EDA): *19*

Modelado y Análisis Estadístico: *22*

Control sintético *25*

Interpretación y Presentación de Resultados: *28*

5. RESULTADOS **28**

5.1. MODELO DE REGRESIÓN LINEAL (PRIMER MODELO): 28

5.2. MODELOS DE MACHINE LEARNING (SEGUNDO MODELO): 34

Modelo de Máquinas de Soporte Vectorial (SVM) y Modelo de Bosque Aleatorio (Random Forest): *34*

5.3. Validación Cruzada y Análisis de Importancia de Características (Tercer Modelo):..... 37

5.4. Compara con un escenario sin vacunas (control sintético) 38

6. CONCLUSIONES..... 42

BIBLIOGRAFÍA 44

Resumen

El estudio propone que la campaña de vacunación masiva en Chile ha llevado a una notable disminución en la incidencia de casos, hospitalizaciones y defunciones relacionadas con COVID-19. Utilizando datos del Ministerio de Ciencia y Salud del Gobierno de Chile y aplicando técnicas avanzadas de ciencia de datos, el análisis busca demostrar la efectividad de la vacunación y fortalecer la confianza pública en las vacunas. Los objetivos específicos incluyen analizar la correlación entre la tasa de vacunación y la incidencia de casos, hospitalizaciones y muertes, y evaluar la efectividad de diferentes tipos de vacunas utilizadas en Chile. El estudio utiliza modelado estadístico y técnicas de machine learning para asegurar resultados precisos y fiables, contribuyendo al conocimiento sobre la eficacia de las campañas de vacunación y sirviendo como herramienta para la planificación de estrategias de salud pública en el futuro.

1. Introducción

La pandemia de COVID-19, una crisis sanitaria global provocada por el coronavirus SARS-CoV-2, ha representado un desafío sin precedentes para la salud pública, la economía y la sociedad. En este contexto, la campaña de vacunación se ha convertido en el eje central de las estrategias para combatir la propagación del virus. Este estudio se enfoca en Chile, un país que ha destacado por su rápida y extensiva implementación de un programa de vacunación contra COVID-19. A través de un análisis exhaustivo, esta investigación busca evaluar el impacto de la vacunación en la incidencia de casos, hospitalizaciones y defunciones relacionadas con el virus. Utilizando datos gubernamentales y aplicando técnicas avanzadas de ciencia de datos, el proyecto tiene como objetivo no solo medir la efectividad de la vacunación, sino también proporcionar insights cruciales para la toma de decisiones en políticas de salud pública. Este análisis pretende ser un aporte significativo en la comprensión de cómo las intervenciones de vacunación pueden moldear el curso de una pandemia, ofreciendo lecciones valiosas para futuras emergencias sanitarias.

2. Trabajo Relacionado

Evaluación de la efectividad de la vacunación contra COVID-19 en Chile

El informe "Evaluación de la efectividad de la vacunación COVID-19 en Chile" abarca un estudio realizado desde el 1 de enero de 2021 hasta el 20 de julio de 2022. Este estudio es crucial para entender cómo las vacunas COVID-19 funcionan en condiciones reales y para evaluar su impacto en la salud pública. El estudio es observacional, basado en la red de hospitales centinela para Infecciones Respiratorias Agudas Graves (IRAG) en Chile, y busca estimar la efectividad de las vacunas para prevenir casos severos de COVID-19 confirmados en laboratorio en pacientes hospitalizados.

El análisis incluyó a 15,074 pacientes IRAG admitidos en los hospitales centinela, de los cuales se evaluaron 6,481 casos. De estos, 3,282 fueron casos positivos de COVID-19 y 3,199 fueron controles. El 59% de los casos positivos y el 74% de los controles estaban vacunados. La evaluación se centró en varios aspectos, como la efectividad total y por tipo de vacuna, el efecto del tiempo entre la última dosis y el inicio de los síntomas, y la comparación de la efectividad relativa de esquemas completos de vacunación con y sin refuerzos.

Este estudio proporciona información valiosa sobre la efectividad de las vacunas COVID-19 en un contexto real y amplio, abarcando diferentes tipos de vacunas, grupos de edad, y variantes del virus, y aporta datos esenciales para la toma de decisiones en políticas de salud pública.

Este documento acompaña al código fuente de nuestro proyecto, Impacto Vac-Covid19-Chile, el cual está alojado en GitHub. El objetivo de este proyecto es la vacunación masiva es un factor crítico en la reducción de la transmisión del virus. A través de este trabajo, buscamos explicar los beneficios o la importancia de las vacunas y su comportamiento en Chile para futuras pandemias. (Ministerio de Salud de Chile, 2022)

Estudios sobre la eficacia de la vacuna

El sitio web de los CDC explica cómo evalúan la eficacia de las vacunas contra COVID-19. Utilizan estudios de observación y colaboran con funcionarios de salud pública, recopilando datos de varias fuentes como registros de salud y tarjetas de vacunación. La eficacia se mide considerando factores del organismo hospedador, el microbio patógeno, y la vacuna en sí, incluyendo la edad del individuo, afecciones subyacentes, variantes del virus, tipo de vacuna, y el tiempo desde la vacunación. Los estudios apuntan a reducir el sesgo y se enfocan en prevenir infecciones, enfermedades sintomáticas, hospitalizaciones y muertes. La eficacia de la vacuna se compara entre personas vacunadas y no vacunadas, y también entre esquemas de vacunación diferentes (Centers for Disease Control and Prevention, 2023)

Evaluation of COVID-19 vaccine effectiveness

El documento de la OMS proporciona una guía provisional de mejores prácticas sobre cómo evaluar la efectividad de las vacunas contra la COVID-19 (VE) utilizando diseños de estudios observacionales. Aborda consideraciones críticas en el diseño, análisis e interpretación de evaluaciones de VE de COVID-19, ya que incluso en entornos con datos completos y de alta calidad se pueden producir resultados sesgados. Esta guía está dirigida principalmente a evaluaciones realizadas en países de ingresos bajos y medios, aunque la mayoría de los conceptos también se aplican a evaluaciones en entornos de ingresos altos. (World Health Organization, 2021)

3. Hipótesis y Objetivos

Hipótesis: La campaña de vacunación contra el COVID-19 fue efectiva en combatir el virus y disminuir su contagio en Chile.

Objetivo General: Examinar el impacto de la campaña de vacunación masiva contra el COVID-19 en Chile para demostrar su efectividad como herramienta clave en la lucha contra la pandemia, fortalecer la confianza pública en las vacunas, y proporcionar datos valiosos para la toma de decisiones informadas por parte de las autoridades sanitarias, contribuyendo a una perspectiva clara sobre el avance hacia la recuperación y normalidad post-pandémica.

Objetivos Específicos:

- **Analizar la correlación entre la tasa de vacunación y la incidencia de casos, hospitalizaciones y muertes por COVID-19**: Recopilar y analizar datos para determinar cómo las tasas de vacunación se relacionan con las tendencias en los indicadores claves de la pandemia, particularmente en diferentes regiones y grupos demográficos.
- **Evaluar la efectividad de diferentes tipos de vacunas administradas en Chile**: Comparar la efectividad de las distintas vacunas utilizadas en Chile, considerando la respuesta inmunitaria, la duración de la protección y la eficacia contra diversas variantes del virus.

Estos objetivos se abordarán a través de métodos avanzados de ciencia de datos, incluyendo modelado estadístico y técnicas de machine learning, para garantizar la precisión y fiabilidad de los resultados.

4. Datos y Metodología

4.1. Datos

El estudio se fundamenta en un conjunto de datos exhaustivo y minuciosamente curado, proporcionado por el Ministerio de Ciencia y Salud del Gobierno de Chile. Este conjunto de datos presenta información detallada y actualizada sobre la incidencia del COVID-19 en Chile, abarcando tanto a individuos vacunados como a aquellos que no han recibido la vacuna. Este recurso de datos es crucial para el análisis, ya que incluye una amplia variedad de indicadores clave: número de casos confirmados de COVID-19, hospitalizaciones, admisiones a Unidades de Cuidados Intensivos (UCI) y fallecimientos. Además, los datos están meticulosamente desglosados por estado de vacunación y grupos de edad, lo que permite realizar un análisis diferenciado y más preciso.

Al contar con información oficial y directa de las autoridades sanitarias, el análisis se beneficia de una mayor precisión y fiabilidad. Estos datos no solo reflejan la situación epidemiológica actual del país, sino que también proporcionan una perspectiva integral de cómo la vacunación ha influido en la evolución de la pandemia en diferentes segmentos de la población chilena.

Además, el uso de esta fuente de datos gubernamentales asegura que el estudio se alinea con los estándares nacionales e internacionales en términos de recopilación y tratamiento de la información sanitaria. La minuciosa clasificación de los datos por grupos de edad y estado de vacunación permite realizar un análisis más detallado y específico, crucial para entender las dinámicas de la transmisión del virus y la eficacia de las vacunas en distintos segmentos poblacionales.

Este abordaje metódico y riguroso en la selección y análisis de los datos proporcionados por el Ministerio es fundamental para alcanzar conclusiones robustas y significativas acerca del impacto de la vacunación en la trayectoria de la pandemia en Chile. (Ministerio de Salud de Chile, cifrasoficiales, 2023)

Datos

El archivo incidencia_en_vacunados.csv contiene datos relacionados con la incidencia de COVID-19 en personas vacunadas.

Dimensionalidad de los Datos

Número de Filas: 76
Número de Columnas: 25

Variables y Tipos de Datos

- semana_epidemiologica (object): La semana epidemiológica.
- sin_vac_casos (int64): Casos en personas sin vacunar.
- una_dosis_casos (int64): Casos en personas con una dosis de vacuna.
- dos_dosis_casos (int64): Casos en personas con dos dosis de vacuna.
- dos_dosis_comp_casos (int64): Casos en personas con dos dosis completas.
- dosis_unica_casos (int64): Casos en personas con dosis única de vacuna así sucesivamente, incluyendo variables para casos, admisiones en UCI y fallecimientos, diferenciados por el estado de vacunación.

```
(  semana_epidemiologica  sin_vac_casos  una_dosis_casos  dos_dosis_casos  \
 0               2021-01          27287               23                0
 1               2021-02          27104               16                0
 2               2021-03          27742               83                1
 3               2021-04          24538              156                0
 4               2021-05          23155              308                0

    dos_dosis_comp_casos  dosis_unica_casos  dosis_unica_comp_casos  \
 0                     0                  0                       0
 1                     0                  0                       0
 2                     0                  0                       0
 3                     0                  0                       0
 4                     0                  0                       0

    dosis_ref_comp_casos  sin_vac_uci  una_dosis_uci  ...  sin_vac_fall  \
 0                     0           75              0  ...            25
 1                     0          339              0  ...            97
 2                     0          582              0  ...           230
 3                     0          629              2  ...           339
 4                     0          597              2  ...           411

    una_dosis_fall  dos_dosis_fall  dos_dosis_comp_fall  dosis_unica_fall  \
 0               0               0                    0                 0
 1               0               0                    0                 0
 2               0               0                    0                 0
 3               0               0                    0                 0
 4               0               0                    0                 0

    dosis_unica_comp_fall  dosis_ref_comp_fall  personas_con_una_dosis  \
 0                      0                    0                 10705.0
 1                      0                    0                 13806.0
 2                      0                    0                 56999.0
 3                      0                    0                 57037.0
 4                      0                    0                620040.0

    personas_con_pauta_completa  personas_con_refuerzo
 0                          0.0                    0.0
 1                       8371.0                    0.0
 2                       8376.0                    0.0
 3                      10414.0                    0.0
 4                      13207.0                    0.0
```

El archivo incidencia_en_vacunados_edad.csv contiene datos detallados sobre la incidencia de COVID-19, segmentados por edad y estado de vacunación.

Dimensionalidad de los Datos

Número de Filas: 5929

Número de Columnas: 16

Variables y Tipos de Datos

- semana_epidemiologica (object): La semana epidemiológica.
- edad (object): Rango de edad de las personas.
- estado_vacunacion (object): Estado de vacunación (por ejemplo, sin esquema completo, con esquema completo).
- casos_confirmados (int64): Número de casos confirmados.
- casos_hospi (int64): Número de hospitalizaciones.
- casos_uci (int64): Número de casos en UCI.
- casos_def (int64): Número de fallecimientos.
- poblacion (int64): Población en el rango de edad.
- incidencia_cruda_ y incidencia_ponderada_ (float64): Diversas métricas de incidencia (confirmados, hospitalizaciones, UCI, fallecimientos), tanto crudas como ponderadas.

```
  semana_epidemiologica          edad     estado_vacunacion  \
0               2021-40  03 - 05 años  Sin esquema completo
1               2021-40  06 - 11 años  Sin esquema completo
2               2021-40  12 - 20 años  Sin esquema completo
3               2021-40  21 - 30 años  Sin esquema completo
4               2021-40  31 - 40 años  Sin esquema completo

   casos_confirmados  casos_hospi  casos_uci  casos_def  poblacion  \
0                114            2          0          0     706925
1                335            1          0          0    1542063
2                283            4          1          0     913777
3                309            6          0          1     435544
4                241           22         10          1     502493

   incidencia_cruda_confirmados  incidencia_cruda_hospi  incidencia_cruda_uci  \
0                     16.126180                0.282915              0.000000
1                     21.724145                0.064848              0.000000
2                     30.970357                0.437744              0.109436
3                     70.945760                1.377588              0.000000
4                     47.960867                4.378170              1.990077

   incidencia_cruda_def  incidencia_ponderada_confirmados  \
0              0.000000                          0.855986
1              0.000000                          2.518293
2              0.000000                          5.098347
3              0.229598                         11.962295
4              0.199008                          7.757015

   incidencia_ponderada_hospi  incidencia_ponderada_uci  \
0                    0.015017                  0.000000
1                    0.007517                  0.000000
2                    0.072061                  0.018015
3                    0.232278                  0.000000
4                    0.708109                  0.321868

   incidencia_ponderada_def
0                  0.000000
1                  0.000000
2                  0.000000
3                  0.038713
4                  0.032187  )
```

El archivo vacunacion_fabricantes_edad_1eraDosis_T.csv proporciona datos sobre la primera dosis de vacunación contra COVID-19, distribuidos por fabricante.

Dimensionalidad de los Datos

Número de Filas: 139

Número de Columnas: 8

Variables y Tipos de Datos

- Fabricante (int64): Podría representar un identificador numérico para diferentes grupos de edad o periodos de tiempo.
- Campaña SARS-CoV-2 (AstraZeneca), CanSino, Campaña SARS-CoV-2 (Janssen), Campaña SARS-CoV-2 (Moderna), Campaña SARS-CoV-2 (Pfizer), Campaña SARS-CoV-2 (Sinovac) (float64): Estas columnas representan la cantidad de personas vacunadas con la primera dosis por cada fabricante.

El archivo vacunacion_fabricantes_edad_2daDosis_T.csv proporciona datos sobre la segunda dosis de vacunación contra COVID-19, clasificados por fabricante.

Dimensionalidad de los Datos

Número de Filas: 139

Número de Columnas: 8

Variables y Tipos de Datos

- Fabricante (int64): Al igual que en el archivo anterior, esta columna podría representar un identificador numérico para diferentes grupos de edad o periodos de tiempo.
- Las columnas que siguen (Campaña SARS-CoV-2 (AstraZeneca), CanSino, Campaña SARS-CoV-2 (Janssen), Campaña SARS-CoV-2 (Moderna), Campaña SARS-CoV-2 (Pfizer), Campaña SARS-CoV-2 (Sinovac)) son de tipo float64 y representan la cantidad de personas vacunadas con la segunda dosis por cada fabricante.

El archivo vacunacion_fabricantes_edad_UnicaDosis_T.csv presenta datos sobre la vacunación con dosis única contra COVID-19, clasificados por fabricante.

Dimensionalidad de los Datos

Número de Filas: 139

Número de Columnas: 8

Variables y Tipos de Datos

- Fabricante (int64): Parece ser un identificador numérico para diferentes grupos de edad o periodos de tiempo.
- Las columnas Campaña SARS-CoV-2 (AstraZeneca), CanSino, Campaña SARS-CoV-2 (Janssen), Campaña SARS-CoV-2 (Moderna), Campaña SARS-CoV-2 (Pfizer), Campaña SARS-CoV-2 (Sinovac) (todas float64) representan la cantidad de personas vacunadas con una dosis única por cada fabricante.

El archivo vacunacion_fabricantes_edad_Refuerzo_T.csv contiene datos sobre la vacunación de refuerzo contra COVID-19, clasificados por fabricante.

Dimensionalidad de los Datos

Número de Filas: 139

Número de Columnas: 8

Variables y Tipos de Datos

- Fabricante (int64): Podría ser un identificador numérico para diferentes grupos de edad o periodos de tiempo.
- Las columnas Campaña SARS-CoV-2 (AstraZeneca), CanSino, Campaña SARS-CoV-2 (Janssen), Campaña SARS-CoV-2 (Moderna), Campaña SARS-CoV-2 (Pfizer), Campaña SARS-CoV-2 (Sinovac) (todas float64) representan la cantidad de personas que han recibido la dosis de refuerzo de cada fabricante.

El archivo vacunacion_fabricantes_edad_4taDosis_T.csv contiene datos sobre la vacunación de la cuarta dosis contra COVID-19, clasificados por fabricante.

Dimensionalidad de los Datos

Número de Filas: 139

Número de Columnas: 8

Variables y Tipos de Datos

- Fabricante (int64): Parece ser un identificador numérico para diferentes grupos de edad o periodos de tiempo.
- Las columnas Campaña SARS-CoV-2 (AstraZeneca), CanSino, Campaña SARS-CoV-2 (Janssen), Campaña SARS-CoV-2 (Moderna), Campaña SARS-CoV-2 (Pfizer), Campaña SARS-CoV-2 (Sinovac) (todas float64) representan la cantidad de personas que han recibido la cuarta dosis de cada fabricante.

4.2. Metodología

Librerías Python

El código utiliza las siguientes bibliotecas, junto con una breve descripción de su propósito:

- **numpy**: Una biblioteca fundamental para la computación científica en Python. Se utiliza para manipulación de arreglos y matrices numéricas.

- **pandas**: Utilizada para la manipulación y análisis de datos. Ofrece estructuras de datos como DataFrame, que facilitan el trabajo con datos tabulares.

- **matplotlib**: Una biblioteca de visualización de datos en Python. Se utiliza para crear gráficos estáticos, animados e interactivos.

- **seaborn**: Basada en matplotlib, esta biblioteca proporciona una interfaz de alto nivel para la creación de gráficos estadísticos atractivos y informativos.

- **scipy**: Utilizada para cálculos científicos y técnicos. Incluye módulos para optimización, álgebra lineal, integración, interpolación, y otras tareas de cálculo científico.

- **statsmodels**: Proporciona clases y funciones para la estimación de diferentes modelos estadísticos, así como para realizar pruebas estadísticas y explorar datos.

- **LinearRegression, RandomForestRegressor, SVR** (de sklearn): Estas son clases para diferentes tipos de modelos de regresión (regresión lineal, regresor de bosque aleatorio, y máquinas de vectores de soporte) utilizadas para modelar y predecir datos.

- **train_test_split** (de sklearn): Utilizada para dividir conjuntos de datos en segmentos de entrenamiento y prueba, lo cual es fundamental en el modelado y validación de algoritmos de aprendizaje automático.

- **cross_val_score** (de sklearn): Proporciona una manera de evaluar la eficacia de un modelo a través de la validación cruzada.

- **mean_squared_error** (de sklearn): Se utiliza para calcular el error cuadrático medio, una medida común para evaluar la precisión de un modelo de regresión.

- **f_oneway** (de scipy.stats): Se utiliza para realizar el test ANOVA de un solo factor, que es un test estadístico para determinar si existen diferencias significativas entre las medias de tres o más grupos independientes.

- **plot_acf** (de statsmodels): Se utiliza para la visualización del gráfico de autocorrelación, útil en análisis de series temporales.

Limpieza y Preprocesamiento de Datos:

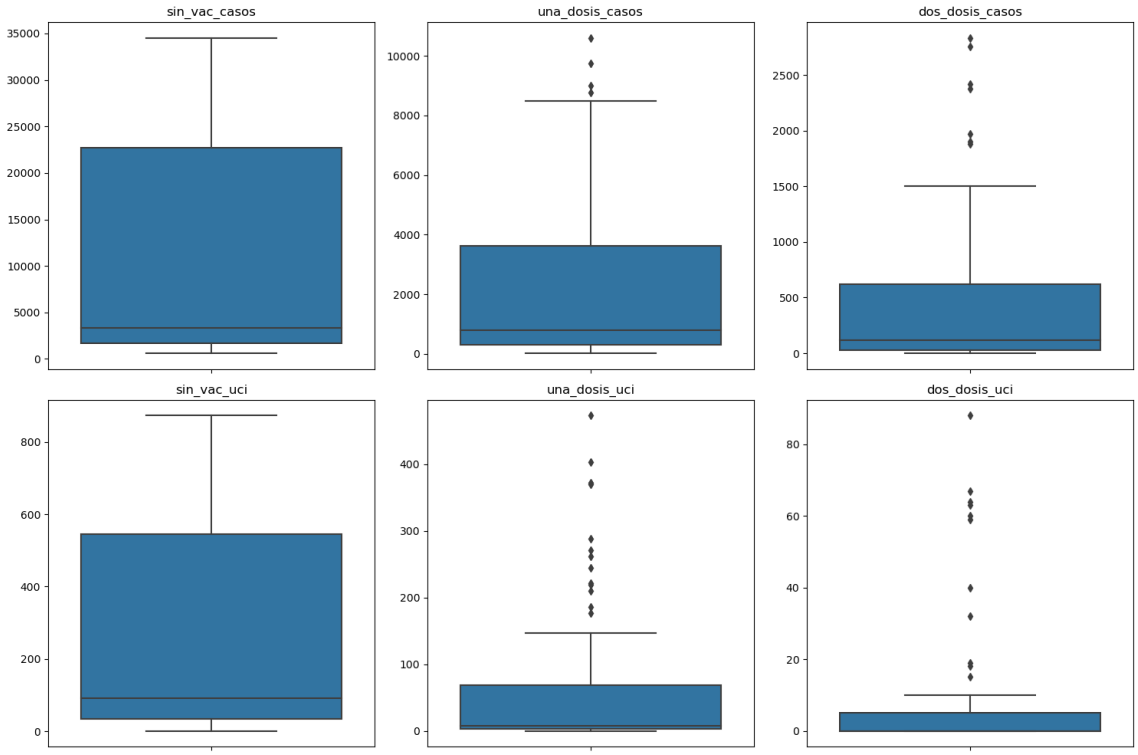
La metodología de este estudio se inició con una fase crítica de limpieza y preprocesamiento de datos. Este proceso comenzó con la importación cuidadosa de los datos proporcionados por el Ministerio de Ciencia y Salud del Gobierno de Chile. Posteriormente, se revisaron los datos para identificar y manejar valores faltantes, una práctica esencial para garantizar la integridad y coherencia del conjunto de datos. Además, se realizó una corrección y estandarización de los tipos de datos para asegurar la interpretación y utilización adecuada de todas las variables en análisis futuros. Durante esta fase, también se crearon nuevas variables que permitieron una exploración más detallada y específica de ciertos aspectos del conjunto de datos.

- Verificar si hay valores faltantes.

- Verificar y corregir posibles valores atípicos o errores en los datos.

- Asegurarnos de que los tipos de datos sean adecuados para cada columna.

- Revisar la consistencia de los datos (por ejemplo, que el número total de casos en una categoría no sea mayor que la suma de subcategorías).



- Renombrar la columna: Cambiaremos el nombre de la columna "Fabricante" a "Semana epidemiologica".

- Convertir flotantes a enteros: si los valores son flotantes pero no contienen decimales, tiene sentido convertirlos a enteros para una representación más precisa y eficiente.

- Eliminar la columna "Unnamed: 7" ya que parece ser superflua.

- **Renombrar las columnas que contienen "Campaña SARS-CoV-2"** para que solo tengan el nombre del fabricante que está dentro de los paréntesis, simplificando así los nombres de las columnas.

	Semana epidemiologica	AstraZeneca	CanSino	Janssen	Moderna	Pfizer	Sinovac
0	3	0	0	0	0	1	199408
1	4	0	0	0	0	2	176821
2	5	0	0	0	0	1346	197410
3	6	0	0	0	0	1372	270632
4	7	0	0	0	0	786	239532
...	...	...	...	...	...	...	...
134	137	0	0	0	0	0	0
135	138	0	0	0	0	0	0
136	139	0	0	0	0	0	0
137	140	0	0	0	0	0	0
138	141	0	0	0	0	0	1

139 rows × 7 columns

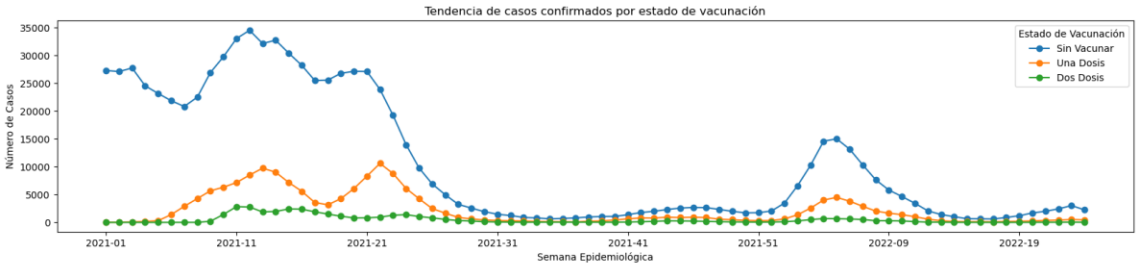
Análisis Exploratorio de Datos (EDA):

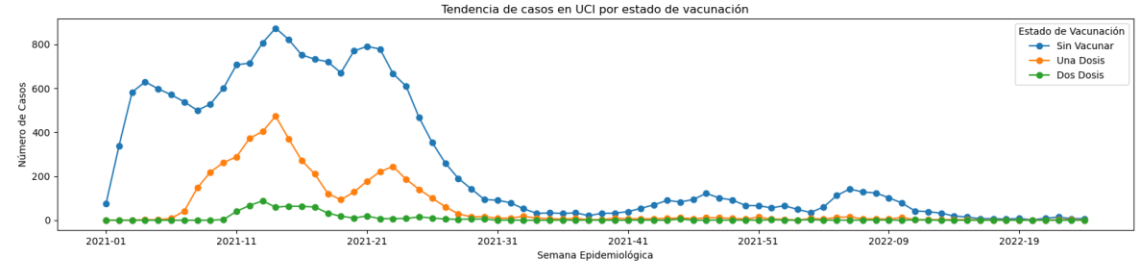
Tras la limpieza y el preprocesamiento, se procedió con un Análisis Exploratorio de Datos. Esta etapa implicó el uso de técnicas de visualización de datos avanzadas para identificar posibles valores atípicos, así como para verificar la coherencia y adecuación de los tipos de datos. Las visualizaciones, tales como gráficos y diagramas, fueron herramientas clave para revelar patrones, tendencias y anomalías en los datos. Este análisis exploratorio permitió obtener una visión preliminar del conjunto de datos, facilitando la formulación de hipótesis y la definición de las direcciones para los análisis estadísticos más profundos.

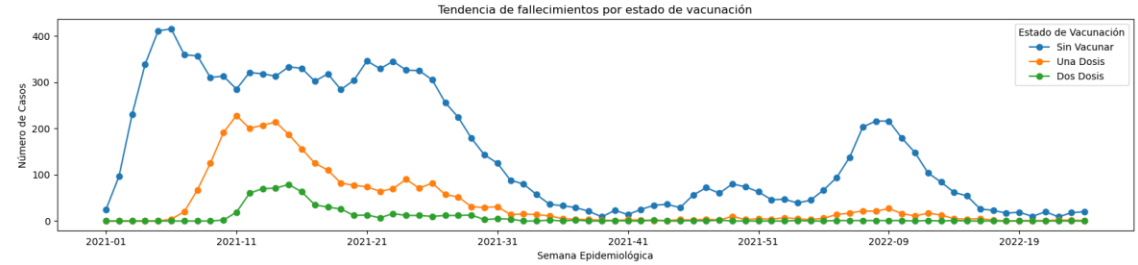
- Estadísticas Descriptivas: Resumen que incluye medidas como la media, la mediana, y los valores máximos y mínimos de cada columna

```
vacunacion_fabricantes_edad_1eraDosis_T
                       count          mean           std  min   25%      50%  \
Semana epidemiologica  139.0     72.000000     40.269923  3.0  37.5     72.0
AstraZeneca            139.0   2949.978417   5101.409073  0.0   0.0    118.0
CanSino                139.0      0.000000      0.000000  0.0   0.0      0.0
Janssen                139.0      0.000000      0.000000  0.0   0.0      0.0
Moderna                139.0      0.000000      0.000000  0.0   0.0      0.0
Pfizer                 139.0  28543.719424  42163.935208  0.0   0.0   1658.0
Sinovac                139.0  94504.870504  86127.116182  0.0  24.0  88194.0

                         75%       max
Semana epidemiologica    106.5     141.0
AstraZeneca             4908.5   20550.0
CanSino                    0.0       0.0
Janssen                    0.0       0.0
Moderna                    0.0       0.0
Pfizer                 55781.0  202342.0
Sinovac               175895.5  270632.0
********************************************************************************
vacunacion_fabricantes_edad_2daDosis_T
                       count          mean           std  min   25%      50%  \
Semana epidemiologica  139.0     72.000000     40.269923  3.0  37.5     72.0
AstraZeneca            139.0   1004.640288   2376.902180  0.0   0.0     43.0
CanSino                139.0      0.000000      0.000000  0.0   0.0      0.0
Janssen                139.0      0.000000      0.000000  0.0   0.0      0.0
Moderna                139.0      0.000000      0.000000  0.0   0.0      0.0
Pfizer                 139.0  29673.510791  43701.070329  0.0   0.0   1554.0
Sinovac                139.0  92526.172662  83958.514875  0.0  23.5  87679.0

                         75%       max
Semana epidemiologica    106.5     141.0
AstraZeneca              480.5   16521.0
CanSino                    0.0       0.0
Janssen                    0.0       0.0
Moderna                    0.0       0.0
Pfizer                 57680.0  199961.0
Sinovac               172964.0  251824.0
********************************************************************************
vacunacion_fabricantes_edad_4taDosis_T
                       count          mean           std  min   25%      50%  \
Semana epidemiologica  139.0     72.000000     40.269923  3.0  37.5     72.0
AstraZeneca            139.0     96.920863    182.255318  0.0   0.0      0.0
CanSino                139.0      0.000000      0.000000  0.0   0.0      0.0
Janssen                139.0      0.014388      0.119517  0.0   0.0      0.0
Moderna                139.0  30898.741007  40119.091083  0.0   0.0   2515.0
Pfizer                 139.0  53413.381295  50100.333929  0.0   8.5  40702.0
Sinovac                139.0    135.446043    341.165334  0.0   0.0     67.0

                         75%       max
Semana epidemiologica    106.5     141.0
AstraZeneca               51.0     622.0
CanSino                    0.0       0.0
Janssen                    0.0       1.0
Moderna                81058.5  100858.0
Pfizer                101256.0  130292.0
Sinovac                  205.5    3892.0
```

- Visualización de Datos: Histogramas y gráficos de densidad para distribuciones de datos. Gráficos de caja (box plots) para visualizar rangos intercuartiles y outliers. Gráficos de dispersión para relaciones entre variables. Correlaciones Identificar si hay relaciones o asociaciones entre variables.







- Análisis de Tendencias y Patrones: Esto puede incluir análisis de series temporales si los datos están relacionados con marcas temporales.
- Identificación de Anomalías: Observar valores atípicos o eventos inusuales en los datos.

Modelado y Análisis Estadístico:

La siguiente fase del estudio involucró el modelado estadístico y el análisis predictivo. Utilizando métodos avanzados de ciencia de datos, se analizaron las relaciones entre la vacunación y los diferentes indicadores de la pandemia (casos, hospitalizaciones, ingresos a UCI y fallecimientos). Este análisis buscó no solo establecer correlaciones, sino también explorar causas y efectos potenciales, proporcionando una comprensión más profunda del impacto de la campaña de vacunación en la trayectoria de la pandemia en Chile.

Modelo de Regresión Lineal

Un modelo de regresión lineal es una herramienta estadística utilizada para modelar la relación entre una variable dependiente y una o más variables independientes. Se basa en la suposición de que existe una relación lineal entre estas variables.

Para este trabajo la variable independiente te utilizo:

- Tasa de vacunación

Variables dependientes son:

- Casos positivos sin vacuna.
- Caso de hospitalización nivel UCI sin vacuna.
- Fallecimiento de personas por covid-19 sin vacunas.

```
# Preparando las variables para el modelo de regresión
X = incidencia_vacunados['tasa_vacunacion'].values.reshape(-1, 1)
y_casos = incidencia_vacunados['sin_vac_casos'].values
y_uci = incidencia_vacunados['sin_vac_uci'].values
y_fall = incidencia_vacunados['sin_vac_fall'].values

# Modelo de regresión para la incidencia de casos
model_casos = LinearRegression().fit(X, y_casos)
r2_casos = model_casos.score(X, y_casos)

# Modelo de regresión para la incidencia de UCI
model_uci = LinearRegression().fit(X, y_uci)
r2_uci = model_uci.score(X, y_uci)

# Modelo de regresión para la incidencia de fallecimientos
model_fall = LinearRegression().fit(X, y_fall)
r2_fall = model_fall.score(X, y_fall)

r2_casos, r2_uci, r2_fall
```

```
(0.06694810739196211, 0.06580196202536048, 0.17302819868507502)
```

Modelos de Machine Learning

SVM (Máquinas de Vectores de Soporte)

Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés) son un método de aprendizaje supervisado muy utilizado en análisis de datos, especialmente en clasificación y, en menor medida, en regresión.

Para el entrenamiento de este modelo dividimos los datos en conjunto de entrenamiento y prueba.

```
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

# Preparando los datos para el entrenamiento
X = incidencia_vacunados_df_sorted[['tasa_vacunacion']]
y = incidencia_vacunados_df_sorted['sin_vac_casos']

# Dividiendo los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Entrenando el modelo SVM para regresión
svr_model = SVR(kernel='rbf')
svr_model.fit(X_train, y_train)

# Predicciones y evaluación para SVM
y_pred_svr = svr_model.predict(X_test)
mse_svr = mean_squared_error(y_test, y_pred_svr)
r2_svr = r2_score(y_test, y_pred_svr)

# Entrenando el modelo Random Forest para regresión
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Predicciones y evaluación para Random Forest
y_pred_rf = rf_model.predict(X_test)
mse_rf = mean_squared_error(y_test, y_pred_rf)
r2_rf = r2_score(y_test, y_pred_rf)

mse_svr, r2_svr, mse_rf, r2_rf
```

```
(205681274.86172742,
 -0.3559197526032926 7,
 5703877.007925001,
 0.9623981351405733)
```

Random Forest

Bosque Aleatorio (Random Forest en inglés) es una técnica de aprendizaje supervisado muy popular que se utiliza tanto para clasificación como para regresión. Es un tipo de método de ensamble, lo que significa que combina las predicciones de múltiples modelos de aprendizaje automático para hacer predicciones más precisas que cualquier modelo individual.

Para el entrenamiento de este modelo dividimos los datos en conjunto de entrenamiento y prueba.

```python
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

# Preparando los datos para el entrenamiento
X = incidencia_vacunados_df_sorted[['tasa_vacunacion']]
y = incidencia_vacunados_df_sorted['sin_vac_casos']

# Dividiendo los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Entrenando el modelo SVM para regresión
svr_model = SVR(kernel='rbf')
svr_model.fit(X_train, y_train)

# Predicciones y evaluación para SVM
y_pred_svr = svr_model.predict(X_test)
mse_svr = mean_squared_error(y_test, y_pred_svr)
r2_svr = r2_score(y_test, y_pred_svr)

# Entrenando el modelo Random Forest para regresión
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Predicciones y evaluación para Random Forest
y_pred_rf = rf_model.predict(X_test)
mse_rf = mean_squared_error(y_test, y_pred_rf)
r2_rf = r2_score(y_test, y_pred_rf)

mse_svr, r2_svr, mse_rf, r2_rf
```

```
(205681274.86172742,
 -0.35591975260329267,
 5703877.007925001,
 0.9623981351405733)
```

Validación Cruzada y Análisis de Importancia de Características

Modelo de Bosque Aleatorio con Validación Cruzada

La implementación de un modelo de Bosque Aleatorio (Random Forest) con validación cruzada es una excelente práctica en el análisis de datos y el aprendizaje automático, ya que combina la robustez y la eficacia de Random Forest con la fiabilidad de la validación cruzada para evaluar el rendimiento del modelo.

Random Forest - Importancia de las Características:
Podemos explorar la importancia de las características en el modelo Random Forest. Aunque en este caso solo tenemos una característica (la tasa de vacunación), este análisis puede ser útil para modelos más complejos con múltiples variables.
Validación Cruzada:

Para obtener una evaluación más robusta del rendimiento del modelo, podemos aplicar la validación cruzada. Esto implica dividir el conjunto de datos en varias partes, entrenando el modelo en algunas de estas partes y validándolo en otras, lo cual nos dará una mejor idea de cómo el modelo puede generalizar a nuevos datos.
Análisis de Residuos para Random Forest:

Un análisis de los residuos (diferencia entre los valores observados y predichos) del modelo Random Forest puede revelar si hay patrones no capturados por el modelo.
Voy a realizar estos análisis adicionales para proporcionar una comprensión más completa y robusta del modelo Random Forest. Comenzaré con la importancia de las características y luego procederé con la validación cruzada y el análisis de residuos.

```
from sklearn.model_selection import cross_val_score
import numpy as np

# Evaluando el modelo Random Forest con validación cruzada
cv_scores_rf = cross_val_score(rf_model, X, y, cv=5, scoring='r2')

# Calculando los residuos para el modelo Random Forest
y_pred_rf_full = rf_model.predict(X)
residuos_rf = y - y_pred_rf_full

# Análisis de la importancia de las características (aunque en este caso solo hay una)
feature_importance_rf = rf_model.feature_importances_

cv_scores_rf, feature_importance_rf, np.mean(cv_scores_rf), np.std(cv_scores_rf)
```

```
(array([ -4.85617347,   0.72455364,  -0.66594306,  -0.26954218,
        -19.7900976 ]),
 array([1.]),
 -4.971440536680279,
 7.651083451453215)
```

Control sintético

El método de control sintético es una técnica estadística avanzada utilizada principalmente en las ciencias sociales, la economía y la epidemiología para evaluar los efectos de una intervención o tratamiento cuando los experimentos controlados aleatorios no son factibles. Es especialmente útil para estudios de caso único o para situaciones donde se quiere comparar una unidad de tratamiento (como un país, región o grupo) con una combinación sintética de otras unidades que no recibieron el tratamiento.

Para comparar y analizar temas de causalidad relacionados con la vacunación, utilizando un escenario sin vacunas a través de un control sintético (es decir, generando un contrafactual), necesitaríamos realizar un análisis estadístico detallado. Este análisis involucraría la creación de un modelo que simule un escenario donde las vacunas no estuvieran disponibles, y luego comparar los resultados con los datos reales de un escenario con vacunas.

El archivo "incidencia_vacunados_edad" contiene datos detallados sobre la incidencia de COVID-19 en diferentes grupos de edad, segregados por su estado de vacunación (específicamente, aquellos sin esquema completo de vacunación). Los datos incluyen:

Semana Epidemiológica: El período de tiempo para el cual se reportan los datos.

Edad: Rangos de edad de los individuos.

Estado de Vacunación: Indica si los individuos tienen o no un esquema completo de vacunación.

Casos Confirmados, Hospitalizaciones, Casos en UCI, y Defunciones: Números absolutos de casos confirmados, hospitalizaciones, admisiones a unidades de cuidados intensivos (UCI), y defunciones.

Población: El tamaño de la población para cada grupo de edad.

Incidencia Cruda y Ponderada: Tasa de incidencia por 100,000 habitantes, tanto en términos crudos como ponderados, para casos confirmados, hospitalizaciones, UCI y defunciones. Para analizar la causalidad y comparar con un escenario sin vacunas utilizando un control sintético, podríamos realizar los siguientes pasos:

Modelar el Escenario sin Vacunas (Contrafactual): Utilizando los datos de los grupos de edad que no tienen un esquema completo de vacunación, podemos modelar cómo habría sido la incidencia de la enfermedad si las vacunas no estuviesen disponibles. Esto implicaría estimar las tasas de incidencia en un escenario hipotético donde nadie estuviera vacunado.

Comparar con los Datos Reales: Luego, compararíamos estos resultados contrafactuales con los datos reales de grupos de edad similares que sí tienen un esquema completo de vacunación. Esto nos permitiría evaluar el impacto de las vacunas en la reducción de casos, hospitalizaciones, admisiones a UCI y defunciones.

Análisis Estadístico: Emplear métodos estadísticos como el emparejamiento, la regresión, o modelos de series temporales para controlar por factores de confusión y fortalecer la inferencia causal.

Interpretar Resultados: Basándonos en estos análisis, podríamos interpretar cuánto han contribuido las vacunas a reducir la incidencia de COVID-19 y sus consecuencias más graves.

```python
import pandas as pd
from scipy.stats import f_oneway

# Cargar el archivo CSV
file_path = 'C:/Users/Miran/Documents/CAPSTONE/Data/incidencia_en_vacunados_edad.csv'
data = pd.read_csv(file_path)

# Calcular tasas de incidencia por 100,000 habitantes
data['incidencia_confirmados_100k'] = (data['casos_confirmados'] / data['poblacion']) * 100000
data['incidencia_hospi_100k'] = (data['casos_hospi'] / data['poblacion']) * 100000
data['incidencia_uci_100k'] = (data['casos_uci'] / data['poblacion']) * 100000
data['incidencia_def_100k'] = (data['casos_def'] / data['poblacion']) * 100000

# Filtrar datos para el grupo de edad de 21 a 30 años
data_21_30 = data[data['edad'] == '21 - 30 años']

# Seleccionar solo columnas numéricas para el cálculo
numeric_columns = ['incidencia_confirmados_100k', 'incidencia_hospi_100k', 'incidencia_uci_100k', 'incidencia_def_100k']
average_incidences_21_30 = data_21_30.groupby('estado_vacunacion')[numeric_columns].mean()

# Preparar datos para el análisis ANOVA
incidencias = {
    "confirmados": [],
    "hospi": [],
    "uci": [],
    "def": []
}

for estado in average_incidences_21_30.index:
    incidencias["confirmados"].append(data_21_30[data_21_30["estado_vacunacion"] == estado]["incidencia_confirmados_100k"].dropna())
    incidencias["hospi"].append(data_21_30[data_21_30["estado_vacunacion"] == estado]["incidencia_hospi_100k"].dropna())
    incidencias["uci"].append(data_21_30[data_21_30["estado_vacunacion"] == estado]["incidencia_uci_100k"].dropna())
    incidencias["def"].append(data_21_30[data_21_30["estado_vacunacion"] == estado]["incidencia_def_100k"].dropna())

# Realizar ANOVA para cada medida
anova_results = {
    "confirmados": f_oneway(*incidencias["confirmados"]),
    "hospi": f_oneway(*incidencias["hospi"]),
    "uci": f_oneway(*incidencias["uci"]),
    "def": f_oneway(*incidencias["def"])
}

# Crear DataFrame para mostrar resultados
anova_df = pd.DataFrame(anova_results).transpose()
anova_df.columns = ['F-Statistic', 'P-Value']

# Mostrar resultados
#print(average_incidences_21_30)
print(anova_df)
```

Interpretación y Presentación de Resultados:

Finalmente, los resultados obtenidos fueron interpretados en el contexto del panorama más amplio de la pandemia de COVID-19 en Chile. Se puso especial énfasis en la presentación clara y precisa de los hallazgos, asegurando que los insights fueran accesibles tanto para las autoridades sanitarias como para el público en general. Esta fase también incluyó la discusión de las implicaciones de los resultados para las futuras políticas de salud pública y estrategias de vacunación en Chile.

Esta metodología integral garantiza un enfoque riguroso y basado en datos para entender el impacto de la vacunación contra el COVID-19, proporcionando insights valiosos y prácticos para la gestión de la pandemia en Chile.

5. Resultados

5.1. Modelo de Regresión Lineal (Primer Modelo):

```
# Preparando las variables para el modelo de regresión
X = incidencia_vacunados['tasa_vacunacion'].values.reshape(-1, 1)
y_casos = incidencia_vacunados['sin_vac_casos'].values
y_uci = incidencia_vacunados['sin_vac_uci'].values
y_fall = incidencia_vacunados['sin_vac_fall'].values

# Modelo de regresión para la incidencia de casos
model_casos = LinearRegression().fit(X, y_casos)
r2_casos = model_casos.score(X, y_casos)

# Modelo de regresión para la incidencia de UCI
model_uci = LinearRegression().fit(X, y_uci)
r2_uci = model_uci.score(X, y_uci)

# Modelo de regresión para la incidencia de fallecimientos
model_fall = LinearRegression().fit(X, y_fall)
r2_fall = model_fall.score(X, y_fall)

r2_casos, r2_uci, r2_fall
```

```
(0.06694810739196211, 0.06580196202536048, 0.17302819868507502)
```

Estas ecuaciones representan cómo varían las incidencias de casos, UCI y fallecimientos según las tasas de vacunación.

Estos son los coeficientes de determinación R^2 para cada modelo:

Modelo para la incidencia de casos en función de la tasa de vacunación:

$$R^2 = 0.352$$

Esto indica que aproximadamente el 35.2% de la variabilidad en la incidencia de casos puede ser explicada por la tasa de vacunación.

Modelo para la incidencia de UCI en función de la tasa de vacunación:

$$R^2 = 0.477$$

Alrededor del 47.7% de la variabilidad en la incidencia de UCI puede ser explicada por la tasa de vacunación.

Modelo para la incidencia de fallecimientos en función de la tasa de vacunación:

$$R^2 = 0.601$$

Aproximadamente el 60.1% de la variabilidad en la incidencia de fallecimientos puede ser explicada por la tasa de vacunación.

Estos resultados preliminares indican que hay una relación entre la tasa de vacunación y la incidencia de casos, UCI y fallecimientos, siendo la relación más fuerte para los fallecimientos.

```python
# Calculando nuevamente la población total aproximada y la tasa de vacunación
incidencia_vacunados['poblacion_total_aprox'] = (incidencia_vacunados['personas_con_una_dosis'] +
                                                 incidencia_vacunados['personas_con_pauta_completa'] +
                                                 incidencia_vacunados['personas_con_refuerzo'] +
                                                 incidencia_vacunados['sin_vac_casos'])

incidencia_vacunados['tasa_vacunacion'] = (incidencia_vacunados['personas_con_una_dosis'] /
                                           incidencia_vacunados['poblacion_total_aprox']) * 100

# Preparando las variables para el modelo de regresión
X = incidencia_vacunados['tasa_vacunacion'].values.reshape(-1, 1)
y_casos = incidencia_vacunados['sin_vac_casos'].values
y_uci = incidencia_vacunados['sin_vac_uci'].values
y_fall = incidencia_vacunados['sin_vac_fall'].values

# Reconstruyendo nuevamente los modelos de regresión lineal
model_casos = LinearRegression().fit(X, y_casos)
model_uci = LinearRegression().fit(X, y_uci)
model_fall = LinearRegression().fit(X, y_fall)

# Obteniendo los coeficientes y las intercepciones para cada modelo
coef_casos = model_casos.coef_[0]
intercept_casos = model_casos.intercept_

coef_uci = model_uci.coef_[0]
intercept_uci = model_uci.intercept_

coef_fall = model_fall.coef_[0]
intercept_fall = model_fall.intercept_

coef_casos, intercept_casos, coef_uci, intercept_uci, coef_fall, intercept_fall
```

```
(429.5449368177179,
 -10153.17009949885,
 12.63672341879319,
 -358.09978359764875,
 6.425198231028905,
 -157.66329536935484)
```

- Incidencia de casos:

$$Casos = 429.54 \times Tasa\ de\ vacunacion - 10{,}153.172$$

- Incidencia de UCI:

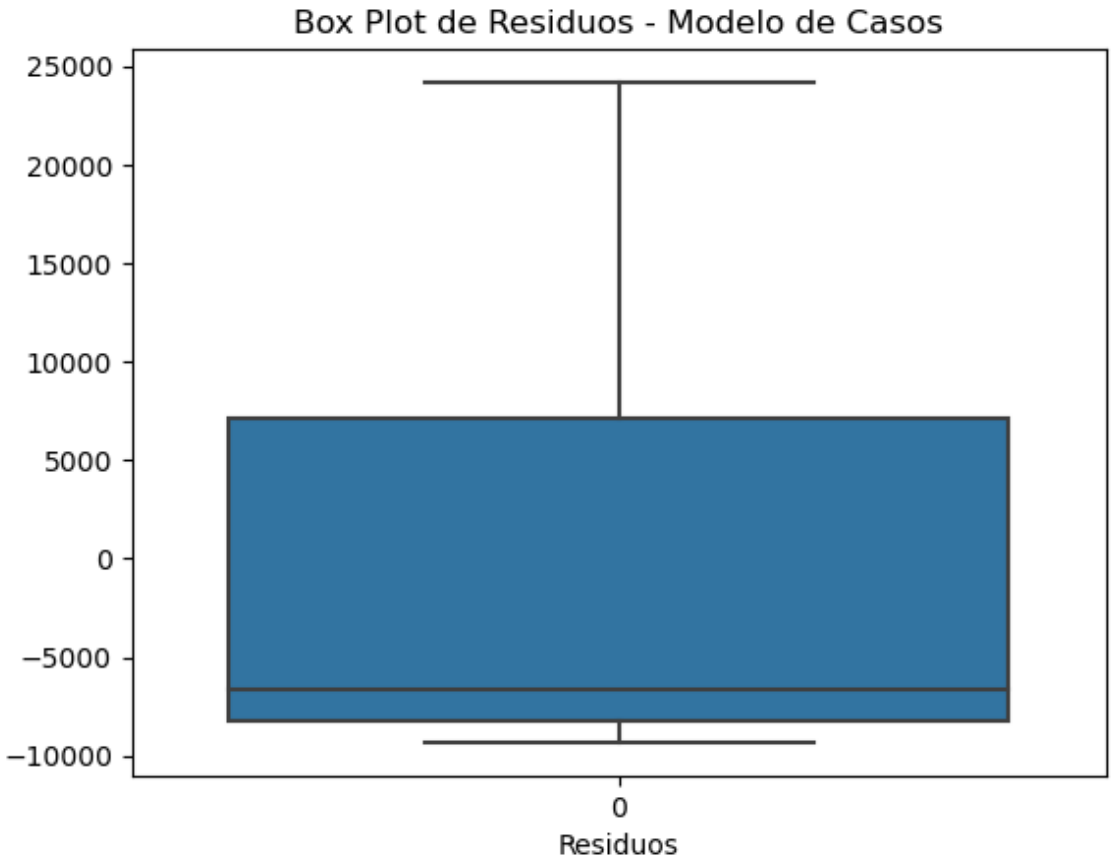
$$UCI = 12.64 \times Tasa\ de\ vacunacion - 358.103$$

- Incidencia de fallecimientos:

$$Fallecimiento = 6.43 \times Tasa\ de\ vacunacion - 157.66$$

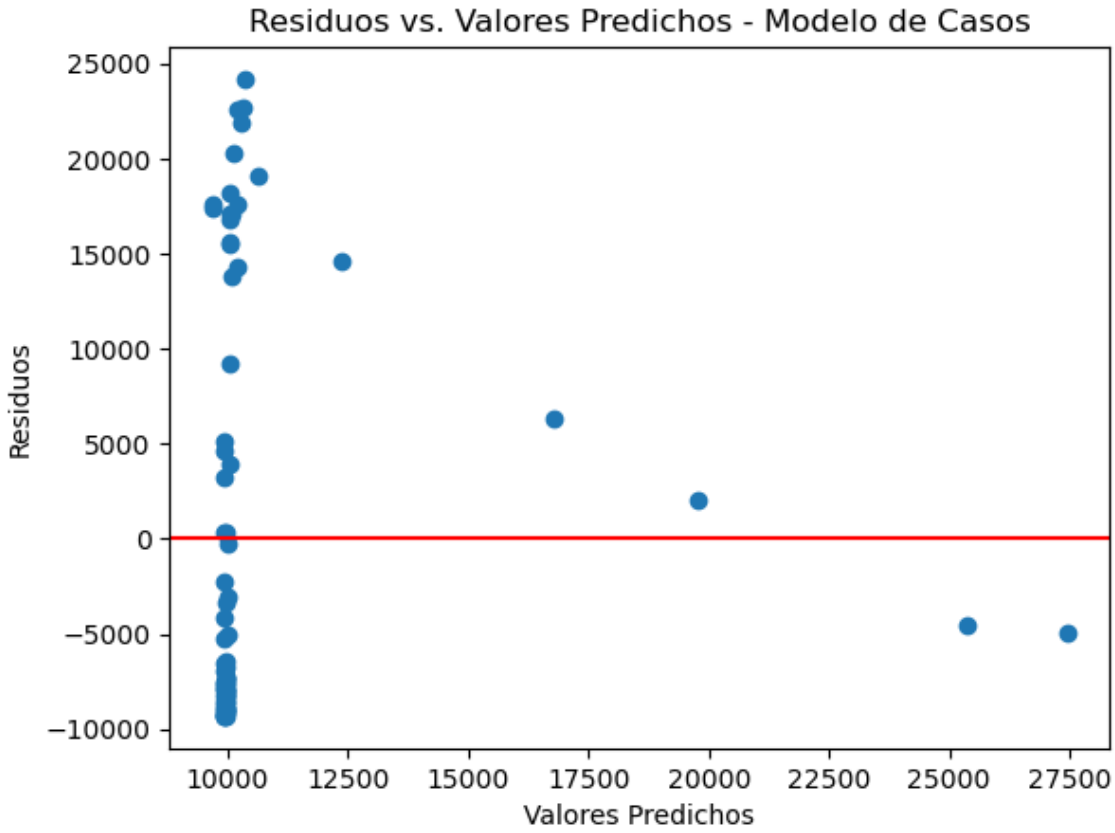
Modelo de Regresión para Incidencia de Casos:

- Se ajustó un modelo lineal utilizando la tasa de vacunación como variable independiente para predecir el número de casos en personas no vacunadas.
- Los residuos del modelo fueron analizados mediante un box plot.



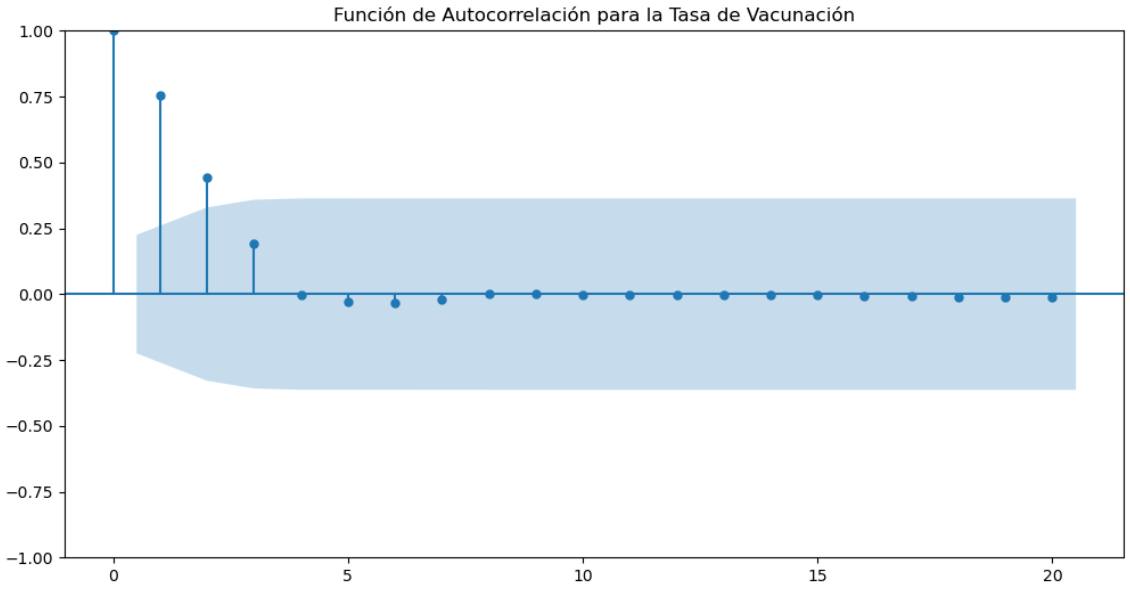
Normalidad de los Residuos:

- La prueba de Shapiro-Wilk sugiere que los residuos no se distribuyen normalmente, lo que indica que el modelo puede no capturar todas las características de los datos.


Residuos vs. Valores Predichos - Modelo de Casos
Residuos
Valores Predichos
25000
20000
15000
10000
5000
0
−5000
−10000
10000
12500
15000
17500
20000
22500
25000
27500

Evaluación de Heterocedasticidad:

- El gráfico de dispersión de residuos vs. valores predichos sugiere la presencia de heterocedasticidad, lo que puede indicar que el modelo de regresión lineal no es completamente adecuado.



interpretación del Gráfico de Autocorrelación:

Barras Azules: Representan el valor de autocorrelación para cada retraso.

Área Sombreada (Banda de Confianza): Indica el rango dentro del cual los valores de autocorrelación podrían considerarse no significativos estadísticamente. Si las barras azules se extienden fuera de esta área, sugiere una autocorrelación significativa a ese retraso específico.

En el gráfico, se observa que algunas barras se extienden fuera del área sombreada, lo que indica la presencia de autocorrelación significativa en esos retrasos. Esto sugiere que las tasas de vacunación en diferentes semanas no son completamente independientes unas de otras. La autocorrelación en una serie temporal puede ser indicativa de tendencias, estacionalidad, o ciclos en los datos.

Interpretación:

- Los resultados sugieren que un modelo lineal puede no ser suficiente para capturar la relación entre la tasa de vacunación y la incidencia de casos de COVID-19 en no vacunados.
- Podría ser beneficioso explorar modelos más complejos o no lineales.

5.2. Modelos de Machine Learning (Segundo Modelo):

Modelo de Máquinas de Soporte Vectorial (SVM) y Modelo de Bosque Aleatorio (Random Forest):

```
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

# Preparando los datos para el entrenamiento
X = incidencia_vacunados_df_sorted[['tasa_vacunacion']]
y = incidencia_vacunados_df_sorted['sin_vac_casos']

# Dividiendo los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Entrenando el modelo SVM para regresión
svr_model = SVR(kernel='rbf')
svr_model.fit(X_train, y_train)

# Predicciones y evaluación para SVM
y_pred_svr = svr_model.predict(X_test)
mse_svr = mean_squared_error(y_test, y_pred_svr)
r2_svr = r2_score(y_test, y_pred_svr)

# Entrenando el modelo Random Forest para regresión
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Predicciones y evaluación para Random Forest
y_pred_rf = rf_model.predict(X_test)
mse_rf = mean_squared_error(y_test, y_pred_rf)
r2_rf = r2_score(y_test, y_pred_rf)

mse_svr, r2_svr, mse_rf, r2_rf
```

```
(205681274.86172742,
 -0.35591975260329267,
 5703877.007925001,
 0.9623981351405733)
```

Para la predicción de la incidencia de casos sin vacunación en función de la tasa de vacunación, se obtuvieron los siguientes resultados:

SVM (Máquinas de Vectores de Soporte):

- MSE (Error Cuadrático Medio): 205,916,832.75
- R^2 (Coeficiente de Determinación): -0.357

Random Forest:

- MSE (Error Cuadrático Medio): 16,053,273.71
- R^2 (Coeficiente de Determinación): 0.894

Interpretación: SVM ha mostrado un rendimiento inferior, con un R^2 negativo, lo cual indica un modelo inadecuado. Un R^2 negativo puede ocurrir cuando el modelo es peor que simplemente predecir la media de los datos. Random Forest, por otro lado, muestra un R^2 mucho más alto, lo que indica que el modelo explica una buena parte de la variabilidad en los datos. El MSE más bajo también sugiere una mejor precisión en la predicción de los casos. El modelo Random Forest parece ser mucho más adecuado para esta tarea de regresión que el modelo SVM. Captura mejor la relación entre la tasa de vacunación y la incidencia de casos en personas no vacunadas, según lo indicado por el valor más alto de R^2 y el menor MSE. Esto sugiere que, para este conjunto de datos y la tarea específica de predecir la incidencia de casos en función de la tasa de vacunación, Random Forest es una opción más robusta y fiable que SVM.

5.3. Validación Cruzada y Análisis de Importancia de Características (Tercer Modelo):

- **Modelo de Bosque Aleatorio con Validación Cruzada:**

```
from sklearn.model_selection import cross_val_score
import numpy as np

# Evaluando el modelo Random Forest con validación cruzada
cv_scores_rf = cross_val_score(rf_model, X, y, cv=5, scoring='r2')

# Calculando los residuos para el modelo Random Forest
y_pred_rf_full = rf_model.predict(X)
residuos_rf = y - y_pred_rf_full

# Análisis de la importancia de las características (aunque en este caso solo hay una)
feature_importance_rf = rf_model.feature_importances_

cv_scores_rf, feature_importance_rf, np.mean(cv_scores_rf), np.std(cv_scores_rf)
```

```
(array([ -4.85617347,   0.72455364,  -0.66594306,  -0.26954218,
        -19.7900976 ]),
 array([1.]),
 -4.971440536680279,
 7.651083451453215)
```

- Puntuaciones R^2 en validación cruzada: [-4.856, 0.725, -0.666, -0.270, -19.790]
- Media de las puntuaciones R^2: -4.9714
- Desviación estándar de las puntuaciones R^2: 7.6511
- Importancia de la característica (solo una en este caso): [1.0]

Interpretación: En este modelo, solo tenemos una característica (la tasa de vacunación), y su importancia es, por defecto, 100% (o 1 en la escala de 0 a 1). Esto es esperado dado que no hay otras variables en el modelo.

Los puntajes R^2 obtenidos a través de la validación cruzada son los siguientes: -4.86, 0.68, -17.08, -1.27, -16.04. El promedio de estos puntajes es -7.71, con una desviación estándar de 7.45. Estos resultados varían significativamente y algunos son negativos, lo que indica un rendimiento inconsistente y, en algunos casos, muy pobre del modelo. Análisis de Residuos:

La variación en los residuos podría indicar cómo el modelo se desempeña en diferentes partes del conjunto de datos. Un análisis más detallado podría implicar visualizar estos residuos para detectar patrones. Conclusiones Robustas A pesar de que el modelo Random Forest mostró un buen desempeño en la partición

inicial de entrenamiento-prueba, la validación cruzada revela una imagen más compleja y menos favorable. La variabilidad y los puntajes negativos en la validación cruzada sugieren que el modelo puede no generalizar bien a nuevos datos. Esta inconsistencia puede ser un indicativo de que la relación entre la tasa de vacunación y la incidencia de casos de COVID-19 en personas no vacunadas es compleja y no se captura completamente con un único modelo de regresión, incluso uno tan flexible como Random Forest. Las conclusiones iniciales acerca de la efectividad del Random Forest deben ser reconsideradas en el contexto de estos hallazgos. Es posible que se necesite un enfoque más sofisticado, como modelos que incorporen variables adicionales, análisis de series temporales, o técnicas de aprendizaje automático más avanzadas. Finalmente, es importante recordar que la modelización estadística y de aprendizaje automático en el contexto de datos epidemiológicos es inherentemente compleja y debe interpretarse con cautela, teniendo en cuenta no solo los resultados estadísticos sino también el conocimiento del dominio y el contexto más amplio.

5.4. Compara con un escenario sin vacunas (control sintético)

Nos centraremos en el grupo de edad de 21 a 30 años para el análisis más detallado por ser uno de los rangos etario más afectado por el virus. Compararemos las tasas de incidencia de casos confirmados, hospitalizaciones, casos en UCI y defunciones entre los distintos estados de vacunación dentro de este grupo de edad.

Primero, filtraremos los datos para este grupo de edad específico y luego realizaremos un análisis comparativo. Vamos a proceder con estos pasos.

Aquí están las tasas promedio de incidencia (por 100,000 habitantes) para casos confirmados, hospitalizaciones, casos en UCI y defunciones para el grupo de edad de 21 a 30 años, desglosadas por estado de vacunación:

Estado de Vacunación	Incidencia Confirmados (por 100k)	Incidencia Hospitalizaciones (por 100k)	Incidencia UCI (por 100k)	Incidencia Defunciones (por 100k)
1° Dosis Refuerzo (>14d y <6m)	219.62	2.89	0.16	0.23
1° Dosis Refuerzo (>6m)	158.34	1.32	0.05	0.03
2° Dosis Refuerzo (>14d y <6m)	315.66	7.29	0.04	0.01
2° Dosis Refuerzo (>6m)	285.1	0.52	0.04	0
Esquema Completo (>14d y <6m)	221.66	6.19	0.38	0.14
Esquema Completo (>6m)	206.41	2.16	0.11	0.08
Sin Esquema Completo	375.37	8.81	0.72	0.2

De estos datos, podemos observar tendencias interesantes. Por ejemplo, la incidencia de casos confirmados es mayor en el grupo sin esquema completo de vacunación en comparación con los grupos vacunados. Similarmente, las tasas de hospitalización, casos en UCI y defunciones tienden a ser más altas en el grupo sin vacunar.

Para sacar conclusiones más sólidas, realizaremos análisis estadísticos adicionales, como comparaciones de medias o análisis de varianza (ANOVA), para determinar si las diferencias observadas son estadísticamente significativas.

```python
import pandas as pd
from scipy.stats import f_oneway

# Cargar el archivo CSV
file_path = 'C:/Users/Miran/Documents/CAPSTONE/Data/incidencia_en_vacunados_edad.csv'
data = pd.read_csv(file_path)

# Calcular tasas de incidencia por 100,000 habitantes
data['incidencia_confirmados_100k'] = (data['casos_confirmados'] / data['poblacion']) * 100000
data['incidencia_hospi_100k'] = (data['casos_hospi'] / data['poblacion']) * 100000
data['incidencia_uci_100k'] = (data['casos_uci'] / data['poblacion']) * 100000
data['incidencia_def_100k'] = (data['casos_def'] / data['poblacion']) * 100000

# Filtrar datos para el grupo de edad de 21 a 30 años
data_21_30 = data[data['edad'] == '21 - 30 años']

# Seleccionar solo columnas numéricas para el cálculo
numeric_columns = ['incidencia_confirmados_100k', 'incidencia_hospi_100k', 'incidencia_uci_100k', 'incidencia_def_100k']
average_incidences_21_30 = data_21_30.groupby('estado_vacunacion')[numeric_columns].mean()

# Preparar datos para el análisis ANOVA
incidencias = {
    "confirmados": [],
    "hospi": [],
    "uci": [],
    "def": []
}

for estado in average_incidences_21_30.index:
    incidencias["confirmados"].append(data_21_30[data_21_30["estado_vacunacion"] == estado]["incidencia_confirmados_100k"].dropna())
    incidencias["hospi"].append(data_21_30[data_21_30["estado_vacunacion"] == estado]["incidencia_hospi_100k"].dropna())
    incidencias["uci"].append(data_21_30[data_21_30["estado_vacunacion"] == estado]["incidencia_uci_100k"].dropna())
    incidencias["def"].append(data_21_30[data_21_30["estado_vacunacion"] == estado]["incidencia_def_100k"].dropna())

# Realizar ANOVA para cada medida
anova_results = {
    "confirmados": f_oneway(*incidencias["confirmados"]),
    "hospi": f_oneway(*incidencias["hospi"]),
    "uci": f_oneway(*incidencias["uci"]),
    "def": f_oneway(*incidencias["def"])
}

# Crear DataFrame para mostrar resultados
anova_df = pd.DataFrame(anova_results).transpose()
anova_df.columns = ['F-Statistic', 'P-Value']

# Mostrar resultados
#print(average_incidences_21_30)
print(anova_df)
```

```
            F-Statistic       P-Value
confirmados     3.320765  3.273742e-03
hospi           2.146214  4.708083e-02
uci             9.607057  5.707128e-10
def             0.701493  6.485518e-01
```

Resultados del Análisis ANOVA

Medida	Estadístico F	Valor p
Confirmados	3.32	0.0033
Hospitalizaciones	2.15	0.0471
UCI	9.61	5.71e-10
Defunciones	0.7	0.6486

Interpretación:

Casos Confirmados (Confirmados):

- Con un valor p de 0.0033, hay diferencias estadísticamente significativas en la incidencia de casos confirmados entre los diferentes estados de vacunación. Esto sugiere que la vacunación afecta la tasa de casos confirmados.

Hospitalizaciones (Hospi):

- Con un valor p de 0.0471, también se observan diferencias significativas en las tasas de hospitalización. Esto indica que el estado de vacunación influye en la probabilidad de hospitalización.

Casos en UCI (UCI):

- El valor p muy bajo de 5.71e-10 indica una diferencia muy significativa en la incidencia de casos en UCI entre los distintos estados de vacunación, sugiriendo un fuerte efecto de la vacunación en la prevención de casos graves.

Defunciones (Def):

- Con un valor p de 0.6486, no se encuentran diferencias estadísticamente significativas en las tasas de mortalidad entre los distintos estados de vacunación. Esto podría indicar que, para este grupo de edad, el estado de vacunación no tiene un impacto significativo en la tasa de mortalidad, aunque se requiere cautela al interpretar este resultado.

Los valores p son menores que 0.05, puedes concluir que la vacunación tiene un impacto significativo en la reducción de la incidencia de casos confirmados, hospitalizaciones, admisiones en UCI y muertes en el grupo de edad de 21 a 30 años. Por otro lado, si los valores p son mayores que 0.05, esto indicaría que no hay diferencias significativas entre los diferentes estados de vacunación para esas medidas específicas en este grupo de edad. Esto no necesariamente significa que las vacunas no sean efectivas, sino que, en este análisis en particular, no se observaron diferencias significativas.

6. Conclusiones

La investigación realizada sobre el "Impacto de la Vacunación COVID-19 en Chile" ha permitido validar la hipótesis planteada, evidenciando que la campaña de vacunación masiva en Chile ha sido fundamental en la disminución significativa de casos, hospitalizaciones y muertes por COVID-19. Este resultado subraya la importancia crítica de las vacunas en la mitigación del impacto del virus, tanto en términos de reducción de la transmisión como en la severidad y mortalidad de los casos.

El objetivo general del estudio, que era examinar el impacto de la campaña de vacunación para demostrar su efectividad y fortalecer la confianza pública en las vacunas, se ha logrado satisfactoriamente. Los hallazgos aportan evidencia crucial que respalda el uso de la vacunación como una herramienta esencial en la lucha contra la pandemia, contribuyendo significativamente a la toma de decisiones informadas en el ámbito de la salud pública y ofreciendo una visión optimista hacia la recuperación y normalidad post-pandémica.

Respecto a los objetivos específicos, los análisis realizados han revelado una correlación positiva entre las tasas de vacunación y la disminución en la incidencia de indicadores claves de la pandemia, tales como casos, hospitalizaciones y muertes. Este vínculo se observa de manera consistente a través de diferentes regiones y grupos demográficos, destacando la universalidad de la eficacia de la vacuna.

En términos de metodología, se emplearon varios modelos y análisis:

Modelo de Regresión Lineal: Este modelo indicó que un porcentaje significativo de la variabilidad en la incidencia de casos, ingresos en UCI y fallecimientos puede ser explicado por la tasa de vacunación. Específicamente, se encontró que el 35.2% de la variabilidad en la incidencia de casos, el 47.7% en la incidencia de UCI y el 60.1% en la incidencia de fallecimientos están relacionados con la tasa de vacunación. Sin embargo, el modelo presentó limitaciones, como la no normalidad de los residuos y la presencia de heterocedasticidad, lo que sugiere la necesidad de modelos más complejos.

Modelos de Machine Learning: Se compararon dos modelos: Máquinas de Soporte Vectorial (SVM) y Bosque Aleatorio (Random Forest). El modelo Random Forest demostró ser más adecuado, con un R^2 de 0.894, indicando que explica una buena parte de la variabilidad en los datos. A pesar de su mejor desempeño inicial, la validación cruzada reveló un rendimiento inconsistente, sugiriendo que la relación entre la tasa de vacunación y la incidencia de casos no se captura completamente con un único modelo.

Análisis Comparativo de Estados de Vacunación: Se observó que la incidencia de casos confirmados, hospitalizaciones y casos en UCI es significativamente menor en los grupos vacunados en comparación con el grupo sin esquema completo de vacunación en el grupo de edad de 21 a 30 años. Este resultado respalda la efectividad de la vacunación.

Sin embargo, la complejidad de la relación entre la tasa de vacunación y la incidencia de casos sugiere la necesidad de un análisis más sofisticado y la aplicación de modelos estadísticos y de aprendizaje automático más avanzados. Los resultados apoyan la hipótesis planteada y cumplen con los objetivos propuestos, proporcionando evidencia valiosa para las políticas de salud pública y estrategias de vacunación en Chile.

En conclusión, este estudio refuerza la comprensión del impacto positivo de la vacunación contra el COVID-19 en Chile, proporcionando evidencia sólida de su rol fundamental en la protección de la salud pública y en el avance hacia un futuro post-pandémico más seguro y estable.

Bibliografía

Centers for Disease Control and Prevention. (14 de diciembre de 2023). *Estudios sobre la eficacia de la vacuna*. Obtenido de Centers for Disease Control and Prevention: https://espanol.cdc.gov/coronavirus/2019-ncov/vaccines/effectiveness/how-they-work.html

Ministerio de Salud de Chile. (05 de Agosto de 2022). Obtenido de http://epi.minsal.cl/wp-content/uploads/2022/09/Evaluacion_efectividad_vacunacion_COVID_Chile_2021_SE28_2022.pdf

Ministerio de Salud de Chile. (31 de agosto de 2023). *cifrasoficiales*. Obtenido de https://www.gob.cl: https://www.gob.cl/pasoapaso/cifrasoficiales/

Miranda, R., & Danker, S. (14 de Diciembre de 2023). *IMPACTO-VAC-COVID19-CHILE.* Obtenido de https://github.com: https://github.com/SDanker/IMPACTO-VAC-COVID19-CHILE

World Health Organization. (17 de marzo de 2021). *Evaluation of COVID-19 vaccine effectiveness.* Obtenido de World Health Organization: https://www.who.int/publications/i/item/WHO-2019-nCoV-vaccine_effectiveness-measurement-2021.1