****

**Procesamiento masivo de datos para estudio del COVID-19 en España**

|  |  |  |
| --- | --- | --- |
| Titulación:  Máster en Big Data y Data Science  Curso académico  2021-2022 | Alumno/a: López Buelga, Ana Isabel  D.N.I:32891324J  Director/a de TFM: Arroquia Cuadros, Benjamím | Convocatoria: |

00 Mes 2021

Índice

[Resumen 5](#_Toc99060815)

[1. Introducción 6](#_Toc99060816)

[2. Objetivos 7](#_Toc99060817)

[3. Estado del Arte y Marco teórico 8](#_Toc99060818)

[3.1. Fuente de datos 8](#_Toc99060819)

[3.2. Técnicas de Machine Learning aplicadas a la predicción en enfermedades 9](#_Toc99060820)

[4. Desarrollo del proyecto y resultados 10](#_Toc99060821)

[4.1. Metodología 10](#_Toc99060822)

[4.2. Planteamiento del problema 12](#_Toc99060823)

[4.3. Desarrollo del proyecto 12](#_Toc99060824)

[4.3.1. Fase de selección de datos 12](#_Toc99060832)

[4.3.2. Fase de preprocesamiento de datos 24](#_Toc99060833)

[4.3.3. Fase de transformación de datos 27](#_Toc99060834)

[4.3.4. Fase de minería de dato 31](#_Toc99060835)

[4.3.5. Fase de interpretación 31](#_Toc99060836)

[4.4. Resultados o conocimiento extraído 31](#_Toc99060837)

[5. Conclusión y trabajos futuros 32](#_Toc99060838)

[6. Referencias 33](#_Toc99060839)

[Anexos I 35](#_Toc99060840)

[Anexos II 36](#_Toc99060841)

Índice de ilustraciones

Ilustración 1. Modelo KDD Fuente: (articulo Cap.+2.+Conocimiento+en+bases+de+datos.pdf) 10

Ilustración 2 Boxblot del número de casos en España. Fuente: Elaboración propia 15

Ilustración 3 Histograma del número de casos en España. Fuente: Elaboración propia 16

Ilustración 4 Gráfico del número de casos en España por día. Fuente: Elaboración propia 16

Ilustración 5 : Gráfico de barras del número de casos en España por comunidad. Fuente: Elaboración propia 17

Ilustración 6 Gráfico lineal del número de casos en España por comunidad. Fuente: Elaboración propia 18

Ilustración 7 Tendencia de movilidad en las distintas ubicaciones por cada cc.aa. Fuente: Elaboración propia 21

Ilustración 8 Gráfico de la evolución en la vacunación por cc.aa. Fuente: Elaboración propia 24

Ilustración 9 : Situación de los datos de Melilla antes de interpolar. Fuente: Elaboración propia 25

Ilustración 10 : Situación de los datos de Ceuta antes de interpolar. Fuente: Elaboración propia 26

Ilustración 11 Situación de los datos de Melilla después de interpolar. Fuente: Elaboración propia 26

Ilustración 12 Situación de los datos de Ceuta después de interpolar. Fuente: Elaboración propia 26

Índice de tablas

Tabla 1Descripción del dataset casos\_técnica\_ccaa (Descripción de variables casos\_tecnica\_ccaa. s.f.). Fuente: Elaboración propia 13

Tabla 2. Relación entre código ISO 3166 y el código asignado a cada CCAA. Fuente: Elaboración propia 14

Tabla 3. Análisis exploratorio del dataset casos\_técnica\_ccaa. Fuente: Elaboración propia 14

Tabla 4 Análisis exploratorio del dataset casos\_técnica\_ccaa para cada cc.aa. Fuente: Elaboración propia 17

Tabla 5 : Descripción del dataset de movilidad (Google Mobility, s.f.) Fuente: Elaboración propia 19

Tabla 6 Análisis exploratorio del dataset de movilidad (Google Mobility, s.f.) Fuente: Elaboración propia 20

Tabla 7 Conteo de datos disponibles del dataset de movilidad (Google Mobility, s.f.) Fuente: Elaboración propia 20

Tabla 8 Descripción del dataset de vacunación (Vacunas, s.f.). Fuente: Elaboración propia 22

Tabla 9 : Análisis estadístico del dataset de vacunación (Vacunas, s.f.). Fuente: Elaboración propia 22

Tabla 10 : Conteo de datos del dataset de vacunación (Vacunas, s.f.). Fuente: Elaboración propia 23

Tabla 11 Descripción del dataset de ingresos (ingresos, s.f.). Fuente: Elaboración propia 24

Tabla 12 Análisis estadístico del dataset de ingresos. Fuente: Elaboración propia 25

Tabla 13 : Relación entre los viejos nombres y los nuevos nombres del dataset de movilidad. Fuente: Elaboración propia 30

Tabla 14 : Descripción de las variables del dataset unificado. Fuente: Elaboración propia 31

Tabla 15: Descripción de las columnas adicionales. Fuente: Elaboración propia 32

Tabla 16 Valores del error de los modelos con parametrización por defecto 35

Tabla 17 Valores del error de los modelos en la iteración I 36

Tabla 18: Mejor algoritmo y sus métricas de error por CCAA 37

Tabla 19: Valores del error de los modelos en la iteración I 38

Tabla 20 Valor de los parámetros y el error obtenido con cada combinación para cada algoritmo 39

# Resumen

El primer caso conocido de COVID-19, fue diagnosticado en Wuhan, provincia de Hubei, China y se extendió rápidamente por todo el mundo.

Poco más de dos años después, seguimos sufriendo sus consecuencias y el número de contagios sigue estando en las noticias de cada día a pesar de que la mayor parte de la población está vacunada y se han aplicado limitaciones de movilidad y contactos sociales.

Con este escenario, se va a desarrollar el presente Trabajo Fin de Máster que consiste en estudiar la relación de los contagios por Comunidad Autónoma en España, en relación con la movilidad de la gente y el índice de vacunación en cada comunidad.

Se parte de la hipótesis de que al reducir la movilidad y estar la gente vacunada, la incidencia de contagios baja y que, a partir de unos modelos de análisis y predicción se pueda corroborar dicha hipótesis, definiendo unos objetivos específicos, que abordarán el problema como un proyecto de Ciencia de Datos basado en la metodología KDD.

Para llegar a dichos objetivos, se construye el conjunto de datos a partir de los datos de casos por COVID-19, junto con los datos de vacunación y movilidad de la gente en cada cc.aa. española.

# Introducción

El 31 de diciembre de 2019, en Wuhan, China, se diagnosticó un brote epidémico de un tipo de neumonía desconocida, confirmándose unos días después que estábamos ante un nuevo coronavirus, el SAS-CoV-2. En febrero de 2020, la OMS dio nombre a esta enfermedad, Coronavirus Disease, 2019 (COVID-19), declarándose como pandemia el 11 de marzo de 2020 . (COVID-19: cronología de la actuación de la OMS, (s.f.))

En España, el primer caso de COVID-19 fue diagnosticado el 31 de enero de 2020 y en marzo de ese mismo año ya se contabilizaban alrededor de 6000 casos, dando lugar a la aprobación del estado de alarma y el comienzo de las restricciones y medidas tomadas para para la expansión del virus.

Las principales restricciones aplicadas fueron las que cortaban las relaciones sociales de la ciudadanía, así como su desplazamiento.

Por ello, y ayudándonos de la Ciencia de Datos y los datos publicados, se pueden crear distintas herramientas para monitorizar el estado de la pandemia, haciendo que la información sea legible para casi toda la población, además de realizar análisis que detecten patrones o tendencias en su comportamiento.

Algunos de esos análisis han servido para predecir la expansión del virus, así como su mortalidad, y ayudar en la toma de decisiones a la hora de aplicar restricciones sobre la población

Con los análisis mencionados anteriormente, se ve que las técnicas de <<>> han servido para poder llevar un seguimiento de la pandemia e implantar medidas que frenen su expansión. Pero sigue habiendo mucha mala información acerca de la efectividad de las mismas. Por ello, este trabajo se va a enfocar en estudiar la evolución de la incidencia del COVID-19 y la incidencia que ha tenido la movilidad y la vacunación en las cc.aa. españolas, mediante técnicas de Machine Learning y utilizando datos públicos sobre incidencia, movilidad y vacunación.

# Objetivos

Las medidas más implantadas a la hora de intentar poner freno a la pandemia han sido cerrar el ocio y limitar la movilidad de la gente para que se interactúe menos unos con otros.

El objetivo principal de este proyecto es estudiar la relación que hay entre el número de contagios, la movilidad de la gente en cada cc.aa., junto con la efectividad de la vacunación para poder predecir si está cerca el fin de la pandemia.

Para llegar a dicho objetivo, se establecen los siguientes objetivos específicos:

* Búsqueda de diferentes fuentes de datos de casos de COVID, datos de vacunación y datos de movilidad por cc.aa.
* Analizar los diferentes datasets encontrados y escoger los que sean más apropiados para nuestro estudio.
* Utilizando técnicas de Minería de Datos, analizar el comportamiento del virus con relación a la movilidad de la gente y los datos de vacunación para ver si el final de la pandemia está cerca.

# Estado del Arte y Marco teórico

En este apartado se analizarán las distintas fuentes de datos disponibles para nuestro estudio así como las técnicas disponibles que se pueden usar para el mismo y sus ventajas y desventajas a la hora de aplicarlas a nuestro estudio.

Empezaremos analizando los datos disponibles y luego se analizarán las técnicas.

## Fuente de datos

Internet nos da acceso a múltiples fuentes de datos sobre el COVID-19 de las cuales las más relevantes se describen a continuación.

Kaggle es uno de los repositorios más conocidos a la hora de extraer datos (Kaggle, s.f.). Esta empresa, subsidiaria de Google, permite a los científicos de datos encontrar y publicar datos para crear modelos de ciencia de datos. Este portal nos proporciona datos sobre la incidencia del covid, fallecidos o la situación de la ocupación en UCI. (Kaggle, COVID-19 in Spain, s.f.) Otra fuente de datos que podemos considerar es Datadista (Datadista, s.f), que también es un medio de comunicación especializado en el periodismo de investigación y datos que explican temas complejos de forma sencilla utilizando técnicas clásicas del periodismo de investigación junto con técnicas de análisis de datos y visualización. En su portal, encontramos la referencia a los datos del COVID-19 en España (Datadista, COVID-19, s.f.). Escovid19data (Escovid19data, s.f.) es un proyecto colaborativo que da información sobre la evolución del virus por provincias. Los datos oficiales de incidencia de coronavirus en España, los proporciona el Instituto de Salud Carlos III (ISCIII, s.f.), una institución cuyo objetivo es contribuir a mejorar la salud de todos los ciudadanos y luchar contra las enfermedades, a través del fomento de la investigación y la innovación en Ciencias de la Salud y Biomedicina, y de la prestación de servicios científicos-técnicos de referencia y programas docentes dirigidos al Sistema Nacional de Salud (ISCIII Quiénes somos Misión y Visión, s.f.). Los datos los podemos encontrar en el apartado de su web dedicado al COVID-19 (cnecovid, s.f.).

Con respecto a los datos de movilidad, el Ministerio de Transporte comenzó a recopilar dichos datos mientras duró el estado de alarma (Ministerio de transportes, movilidad y agenda urbana. Open data movilidad, s.f). Ese conjunto de datos se ha creado a partir del posicionamiento de los teléfonos móviles. Pero si queremos datos de más tiempo, tenemos que buscarlos en los dos gigantes de la comunicación móvil: Android (Android, s.f.), perteneciente a Google y Apple (Apple, s.f.). Ambos SO ofrecen datos de geolocalización de los dispositivos móviles de la gente. Google ofrece datos de anonimizados de movilidad local en varios países (Google Mobility, s.f.). Para España ofrece los datos por cc.aa. Muestran cómo ha cambiado el número de visitas a lugares públicos como supermercados y parques. Apple también ofrece datos de movilidad en diferentes países (Apple Mobility, s.f.) pero, a diferencia de Google, te muestra los dispositivos que se han desplazado dentro de una región y en que medio de transporte se han desplazado.

## Técnicas de Machine Learning aplicadas a la predicción en enfermedades

El uso de algoritmos de Machine Learning en sistemas de diagnóstico y prevención de enfermedades puede ayudar a mejorar los tratamientos de los pacientes y frenar la expansión de la enfermedad. Por ello su uso es cada vez más frecuente en el estudio de enfermedades. En el estudio (Qiang, X., Kou, Z., 2019), realizan una predicción sobre la mutación del virus de la gripe aviar utilizando técnicas de Random Forest, SVM, teorema de Bayes y KNN. En (Withanage, G.P., 2018) podemos ver como predicen la incidencia del virus del Dengue em el distrito de Gampha, Sri Lanka, usando técnicas de regresión a partir de las lluvias, la temperatura y humedad entre otros factores. (Iqbal, N., 2019) trata de predecir también la incidencia del dengue, comparando distintos algoritmos y consiguiendo una exactitud de un 92% con el algoritmo de LogitBoost.

El COVID-19 es una epidemia, por lo que los modelos matemáticos de predicción que se están desarrollando suelen estar basados en el modelo SIR (Susceptibles, Infectados y Recuperados). Como se explica en (Martín Barroso, V., 2020), el modelo SIR, está basado en (Kermack, W.O., 1972), y permite caracterizar la evolución de una epidemia que se propaga mediante contagio entre los individuos de una población de tamaño N que permanece constante. En cada momento del tiempo t los individuos pueden encontrarse en tres posibles estados o compartimentos:

* Susceptibles S(t): individuos sanos que pueden ser contagiados por los infectados.
* Infectados I(t): individuos portadores de la enfermedad.
* Recuperados R(t): individuos que se han recuperado de la enfermedad.

Pero en (Ardabili SF, 2020) podemos ver como los investigadores aseguran que el modelo SIR o sus variantes no son los más precisos para hacer previsiones de la epidemia de COVID-19 ya que ésta no tiene una tendencia estacionaria y sigue un patrón muy variable.

Sobre la relación del índice de vacunación con la incidencia del COVID-19 no se han encontrado artículos que describan un análisis.

# Desarrollo del proyecto y resultados

## Metodología

El desarrollo del presente trabajo se ha realizado en base a la metodología KDD (Knowledge Discovery in Databases) (Timarán-Pereira, 2016).

Dicha metodología se puede definir como un proceso automático en el que se combinan etapas de descubrimiento con etapas de análisis.

El proceso consiste en extraer patrones del conjunto de datos para su posterior análisis.

El proceso KDD se divide en las siguientes etapas:

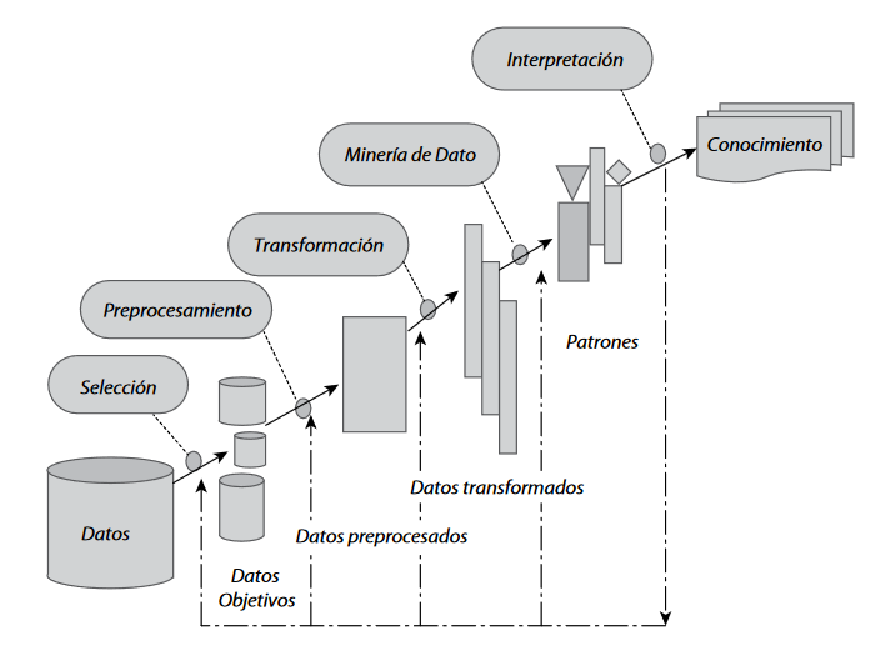


Ilustración . Modelo KDD Fuente: (articulo Cap.+2.+Conocimiento+en+bases+de+datos.pdf)

**Selección**

En esta etapa se va a seleccionar los datos con los que se va a trabajar, de entre las fuentes citadas. Los pasos son:

* Investigar sobre las distintas fuentes de datos que hay disponibles para nuestro análisis.
* Escoger el dataset que mejor se adapte a nuestros objetivos
* Analizar que significan todas las columnas de datos que tenemos
* Tenemos los datos del covid del Instituto de Salud Carlos III, ya que hemos visto que la mayoría de las fuentes de datos investigadas hacen referencia a este dataset
* Los datos de movilidad de Google, porque son los más completos

**Procesamiento**

O limpieza de datos: e analiza la calidad de los datos, se aplican operaciones básicas como el cribado de datos ruidosos y se seleccionan estrategias para el manejo de datos desconocidos datos nulos, datos duplicados y técnicas estadísticas para su reemplazo

**Transformación**

Esta etapa consiste en comprobar que transformación es necesaria aplicar a los datos para poder cruzarlos y estandarizarlos:

* Eliminar las columnas que no sean relevantes para nuestros análisis
* Investigar si es necesario separar datos de una columna en varias para un mejor análisis

**Minería de dato**

Esta es la etapa de búsqueda y descubrimiento de patrones. Para ello se aplicarán técnicas de minería de datos con modelos descriptivos para analizar los datos y una vez analizados los datos, se procederá a usar modelos predictivos para predecir la futura incidencia.

Los modelos descriptivos son aquellos que identifican patrones que explican los datos y como se relacionan. Algunos ejemplos pueden ser: reglas de asociación, los patrones secuenciales, los clustering y las correlaciones

Los modelos predictivos son aquellos que pretenden estimar valores futuros o desconocidos de variables que se consideren de interés (variables objetivo o dependientes), usando otras variables categorizadas como independientes.

En esta etapa entra la aplicación de las distintas técnicas de minería de datos, como pueden ser:

* Clasificación
* Clustering
* Asociación
* Patrones secuenciales

**Interpretación**

En esta etapa se van a interpretar los patrones descubiertos en la etapa anterior y consolidar todo el conocimiento extraído.

Si se ve que hay necesidad de repetir las etapas anteriores para unos mejores resultados, se decide aquí.

Los datos utilizados para realizar este TFM, así como el código de implementación de los distintos algoritmos puede ser encontrado en (López, A., s.f.)

## Planteamiento del problema

El objetivo de este proyecto es estudiar la relación que hay entre el número de contagios, la movilidad de la gente en cada cc.aa., junto con la efectividad de la vacunación para poder predecir si está cerca el fin de la pandemia. Es decir, que factores de movilidad han sido los que más han bajado la misma, así como comprobar que el índice de vacunación ha sido uno de los factores que ha ayudado a bajar la incidencia.

## Desarrollo del proyecto

El proyecto se desarrollará siguiendo las fases del modelo KDD



### Fase de selección de datos

Se han evaluado las distintas fuentes de datos citadas en el apartado 3.1 de este trabajo y se han escogido los siguientes conjuntos de datos:

* Dataset casos de COVID: se ha escogido el proporcionado por el Instituto de Salud Carlos III (ISCIII., s.f.) ya que nos proporciona los datos separados por cc.aa. y parece ser el origen de datos de los demás conjuntos citados.
* Dataset de movilidad: se ha escogido el dataset de datos de Google (Google Mobility, s.f.) ya que contiene datos de sitios donde se reúne la gente separados por cc.aa. y con fecha reciente. Se descartan los datos de (Apple Mobility, s.f.) porque hacen referencia al medio de transporte dentro de las ciudades. También se descartan los datos de (Ministerio de transportes, movilidad y agenda urbana, s.f.) ya que sus datos es un conjunto acotado de datos que engloba únicamente los datos de movilidad del estado de alarma.
* Dataset de vacunación: se ha escogido el dataset proporcionado por (Vacunas, s.f).
  + - 1. Dataset casos COVID por CC.AA.

Como se ha comentado anteriormente, para obtener todos los datos de casos detectados en España por CC.AA. se ha escogido el dataset proporcionado por el Instituto Carlos III (ISCIII., s.f.). El dataset en formato CSV se puede obtener en (dataset COVID, s.f.). Dicho CSV es actualizado semanalmente. Para este estudio, el rango de datos que vamos a utilizar es el comprendido entre las fechas 31/01/2020 y 15/03/2022, (desde el primer caso conocido en España) que nos da un total de 14725 filas de registros y 8 columnas que definen los atributos. La descripción de la tabla se puede ver a continuación:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Descripción** | **Tipo** |
| ccaa\_iso | Código ISO de la CCAA de residencia | Texto |
| fecha | Fecha de inicio de síntomas o, en su defecto, la fecha de diagnóstico menos 6 días. | Texto |
| num\_casos | Número de casos por técnica diagnóstica y CCAA de residencia | Numérico |
| num\_casos\_prueba\_pcr | Número de casos con prueba de laboratorio PCR o técnicas moleculares | Numérico |
| num\_casos\_prueba\_test\_ac | Número de casos con prueba de laboratorio de test rápido de anticuerpos | Numérico |
| num\_casos\_prueba\_ag | Número de casos con prueba de laboratorio de test de detección de antígeno | Numérico |
| num\_casos\_prueba\_elisa | Número de casos con prueba de laboratorio de serología de alta resolución (ELISA/ECLIA/CLIA) | Numérico |
| num\_casos\_prueba\_desconocida | Número de casos sin información sobre la prueba de laboratorio | Numérico |

Tabla Descripción del dataset casos\_técnica\_ccaa (Descripción de variables casos\_tecnica\_ccaa. s.f.). Fuente: Elaboración propia

Los datos en este dataset se dividen en fila por día y cc.aa. Es decir, para un día D, tendremos 19 registros. El número total de registros es 19 cc.aa. \* 775 días.

Ccaa\_iso se puede considerar una variable categórica y el código que sigue para referenciar las cc.aa es el ISO 3166 (ISO 3166, s.f). En nuestro dataset, se sigue dicha norma pero se quita la referencia al país. Por ejemplo, para referirnos al Principado de Asturias, el código en ISO 3166 sería ES-AS, pero en nuestro dataset, el código es AS. Teniendo esto en cuenta, en nuestro dataset, cada código se correspondería a:

|  |  |  |
| --- | --- | --- |
| **Código ISO-3166** | **ccaa\_iso** | **Nombre de la CCAA** |
| **ES-AN** | **AN** | Andalucía |
| **ES-AR** | **AR** | Aragón |
| **ES-AS** | **AS** | Principado de Asturias |
| **ES-CN** | **CN** | Canarias |
| **ES-CB** | **CB** | Cantabria |
| **ES-CM** | **CM** | Castilla-La Mancha |
| **ES-CL** | **CL** | Castilla-León |
| **ES-CT** | **CT** | Cataluña |
| **ES-EX** | **EX** | Extremadura |
| **ES-GA** | **GA** | Galicia |
| **ES-IB** | **IB** | Islas Baleares |
| **ES-RI** | **RI** | La Rioja |
| **ES-MD** | **MD** | Comunidad de Madrid |
| **ES-MC** | **MC** | Región de Murcia |
| **ES-NC** | **NC** | Comunidad Foral de Navarra |
| **ES-PV** | **PV** | País Vasco |
| **ES-VC** | **VC** | Comunidad Valenciana |
| **ES-CE** | **CE** | Ceuta |
| **ES-ML** | **ML** | Melilla |

Tabla . Relación entre código ISO 3166 y el código asignado a cada CCAA. Fuente: Elaboración propia

Realizamos un análisis exploratorio de los datos numéricos y obtenemos los siguientes resultados

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **variable** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **num\_casos** | 14725 | 765,9804414 | 2175,465624 | 0 | 33 | 156 | 597 | 45922 |
| **num\_casos\_prueba\_pcr** | 14725 | 370,346893 | 742,3585761 | 0 | 21 | 109 | 374 | 10161 |
| **num\_casos\_prueba\_test\_ac** | 14725 | 0,32991511 | 3,205126891 | 0 | 0 | 0 | 0 | 96 |
| **num\_casos\_prueba\_ag** | 14725 | 351,8610526 | 1600,676446 | 0 | 0 | 16 | 126 | 41399 |
| **num\_casos\_prueba\_elisa** | 14725 | 0,36196944 | 2,483943314 | 0 | 0 | 0 | 0 | 64 |
| **num\_casos\_prueba\_desconocida** | 14725 | 43,08061121 | 270,5287006 | 0 | 0 | 0 | 0 | 6046 |

Tabla . Análisis exploratorio del dataset casos\_técnica\_ccaa. Fuente: Elaboración propia

Para ver una mejor distribución de la variable num\_casos, la representamos en gráficas. Primero en un gráfico boxplot (figura 1) en el que vemos que el número total de casos diagnosticados suele ser menor de 10000, dándose casos aislados de más de 30000 casos diarios. Estos casos aislados, son los denominados outliers que para nuestro estudio los vamos a tener en cuenta ya que pueden corresponderse con los picos de las distintas olas.

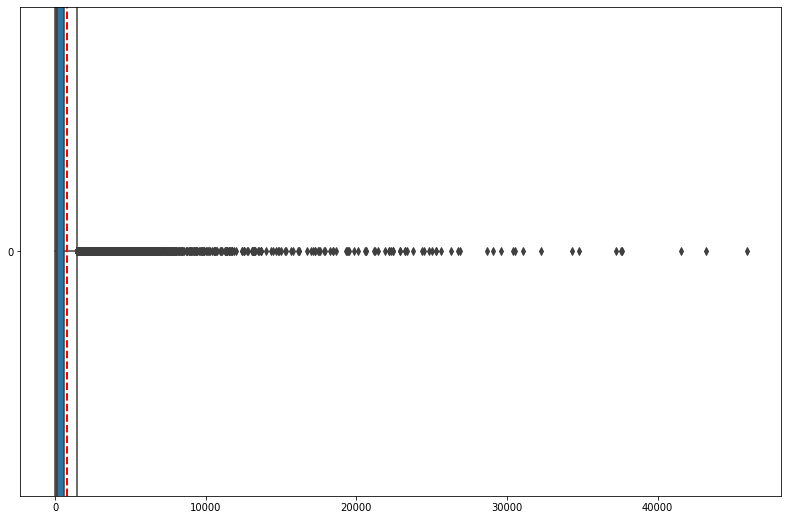


Ilustración Boxblot del número de casos en España. Fuente: Elaboración propia

Por otro lado, si representamos los datos como un histograma, observamos una asimetría positiva que nos confirma que la mayoría de casos diarios está por debajo de los 10000 casos

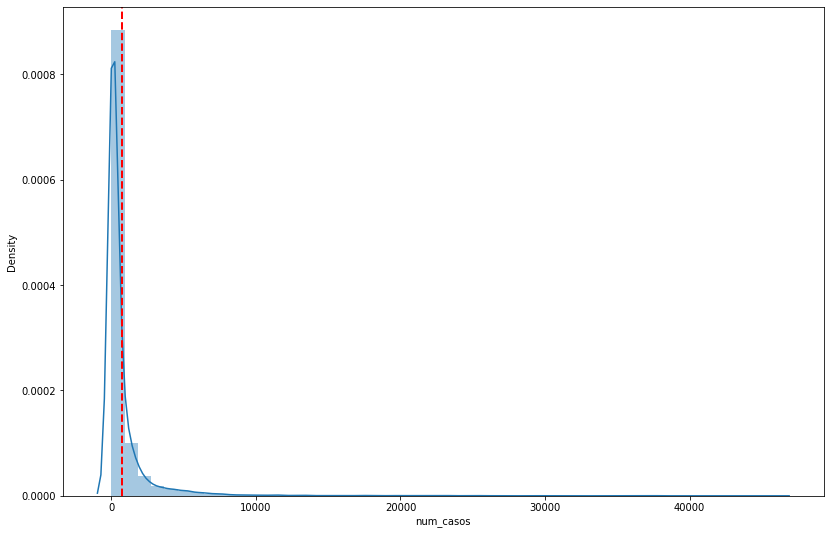


Ilustración Histograma del número de casos en España. Fuente: Elaboración propia

Con un gráfico lineal, mostramos el número de casos totales en España y en él, podemos apreciar claramente las 6 olas que hemos sufrido y se aprecia claramente como la 6ª ola batió récords de contagios superando ampliamente las anteriores.

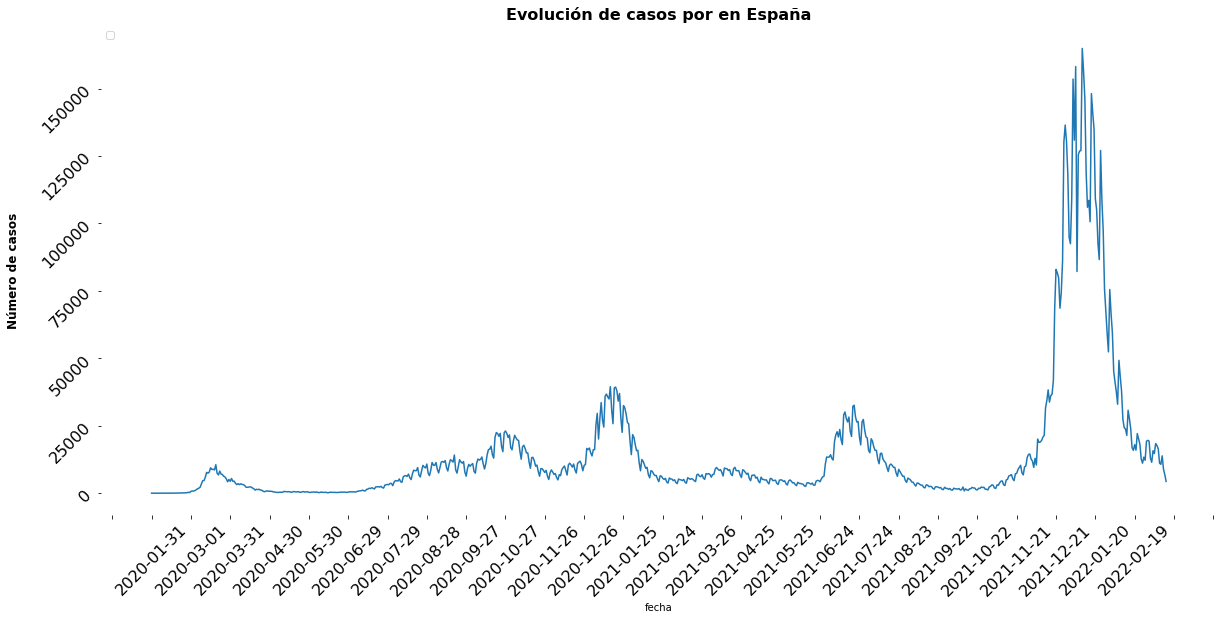


Ilustración Gráfico del número de casos en España por día. Fuente: Elaboración propia

Este trabajo se centra en encontrar una tendencia de la incidencia del virus pro CCAA, es decir, queremos saber el número de casos de cada CCAA. Por ello vamos a analizar dicho dato.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ccaa\_iso** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **CT** | 775 | 2996,283 | 5982,40108 | 0 | 586,5 | 1242 | 2451 | 45922 |
| **MD** | 775 | 2074,39 | 3199,03045 | 2 | 401 | 1213 | 2268 | 23768 |
| **AN** | 775 | 1771,73 | 2436,2972 | 0 | 244 | 1028 | 2040 | 17261 |
| **VC** | 775 | 1705,77 | 3628,71532 | 0 | 134 | 413 | 1498,5 | 25590 |
| **CL** | 775 | 860,84 | 1589,99199 | 0 | 105 | 331 | 896,5 | 11535 |
| **PV** | 775 | 853,4103 | 1724,72642 | 0 | 149,5 | 416 | 769 | 13499 |
| **GA** | 775 | 717,5974 | 1409,74473 | 0 | 67 | 219 | 613,5 | 8389 |
| **CM** | 775 | 615,0374 | 1062,82215 | 0 | 88,5 | 266 | 634 | 6957 |
| **AR** | 775 | 520,609 | 1026,64753 | 0 | 89,5 | 233 | 474,5 | 7178 |
| **MC** | 775 | 495,2955 | 1084,32995 | 0 | 51 | 112 | 451,5 | 7984 |
| **CN** | 775 | 405,7535 | 883,237399 | 0 | 67,5 | 152 | 267 | 6892 |
| **IB** | 775 | 336,5097 | 658,292429 | 0 | 36 | 104 | 334,5 | 4434 |
| **EX** | 775 | 316,351 | 601,396736 | 0 | 32 | 90 | 324 | 3699 |
| **NC** | 775 | 292,7381 | 617,647406 | 0 | 47 | 113 | 261 | 5355 |
| **AS** | 775 | 257,1561 | 527,265152 | 0 | 21 | 80 | 255,5 | 3816 |
| **CB** | 775 | 165,5703 | 342,645144 | 0 | 26 | 66 | 144 | 2352 |
| **RI** | 775 | 117,3974 | 222,565734 | 0 | 14 | 52 | 113 | 1705 |
| **ML** | 775 | 26,40387 | 45,3930095 | 0 | 2 | 13 | 30 | 372 |
| **CE** | 775 | 24,78581 | 57,6592329 | 0 | 1 | 9 | 24 | 470 |

Tabla Análisis exploratorio del dataset casos\_técnica\_ccaa para cada cc.aa. Fuente: Elaboración propia

Por CCAA los datos cambian y vemos que el número de contagios varía mucho dependiendo de cada CCAA. Por ejemplo, las comunidades menos afectadas son las ciudades autónomas de Ceuta y Melilla y las más afectadas Cataluña y Comunidad de Madrid, aquellas que agrupan el mayor porcentaje de población y puede que tengan una mayor movilidad de la gente. El número de casos por CCAA se puede ver representado en la figura. La media y la mediana son valores muy alejados por los picos de las olas.

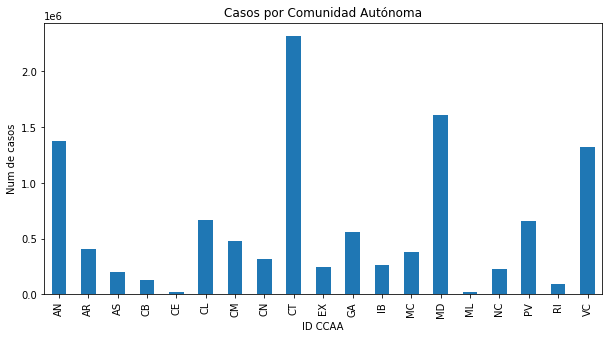


Ilustración : Gráfico de barras del número de casos en España por comunidad. Fuente: Elaboración propia

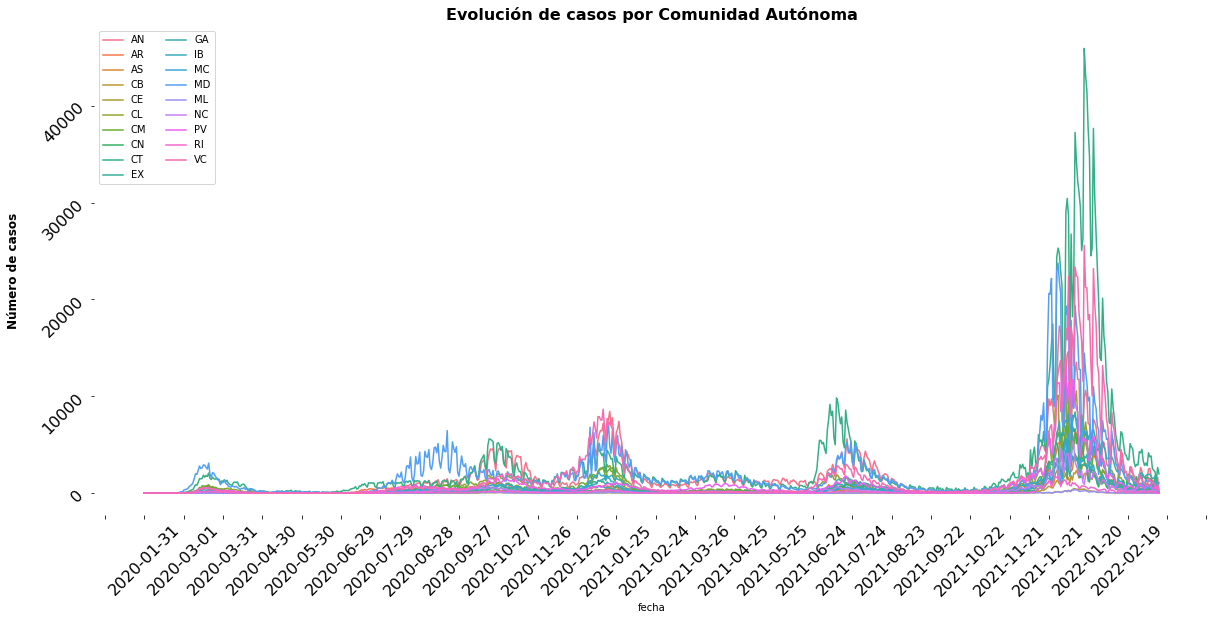


Ilustración Gráfico lineal del número de casos en España por comunidad. Fuente: Elaboración propia

En este gráfico podemos ver el comportamiento del virus en estos dos años, donde los picos son las 6 olas que ha habido. Comparándolo con la figura 3, vemos las olas algo más desplazadas ya que no incidieron en todas las comunidades al mismo tiempo y de la misma forma. Con este gráfico volvemos a constatar que las comunidades donde más casos hemos tenido han sido Cataluña (CN), Andalucía (AN) y la Comunidad de Madrid (MD) ya que sus gráficas son las que sobresalen por encima del resto.

* + - 1. Dataset de movilidad

Para analizar la movilidad de la gente en las CCAA españolas, hemos escogido el dataset de datos Google (Google Mobility, s.f.). El dataset en formato CSV se actualiza cada X tiempo y para este estudio vamos a escoger el mismo rango de fechas que el dataset de datos Covid, es decir, del 15/02/2020 y 15/03/2022, (desde el primer caso conocido en España). El dataset está formado por un total de 47868 registros y 13 columnas o variables. La descripción de las variables se puede ver a continuación en la Tabla 5.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Descripción** | **Tipo** |
| country\_region\_code | Código de país o región. En nuestro dataset (ES) | Texto |
| country\_region | Nombre del país o región. En nuestro dataset (Spain) | Texto |
| sub\_region\_1 | Nombre de la CCAA en inglés. Ver Tabla 1 | Texto |
| sub\_region\_2 | Nombre de la provincia en inglés. | Texto |
| metro\_area | área metropolitana (no aplica en los datos de España) | Numérico |
| iso\_3166\_2\_code | Código de región o estado de todos los países codificados en ISO-3166-2. Ver Tabla 1 | Texto |
| census\_fips\_code | Código de identificación de condados de EEUU | Numérico |
| place\_id | ID que usa Google para identificar el lugar donde se tomaron los datos | Texto |
| date | Fecha del dato | Texto |
| retail\_and\_recreation\_percent\_change\_from\_baseline | Tendencias de movilidad de lugares como restaurantes, cafeterías, centros comerciales, parques de atracciones, museos, bibliotecas y cines | Numérico |
| grocery\_and\_pharmacy\_percent\_change\_from\_baseline | Tendencias de movilidad de lugares como supermercados, almacenes de comida, mercados de productos agrícolas, tiendas de comida especializada y farmacias | Numérico |
| parks\_percent\_change\_from\_baseline | Tendencias de movilidad de lugares como parques locales, parques nacionales, playas públicas, puertos deportivos, parques caninos, plazas y jardines públicos | Numérico |
| transit\_stations\_percent\_change\_from\_baseline | Tendencias de movilidad de lugares como estaciones de transporte público (por ejemplo, de metro, tren y autobuses) | Numérico |
| workplaces\_percent\_change\_from\_baseline | Las tendencias en movilidad de lugares de trabajo. | Numérico |
| residential\_percent\_change\_from\_baseline | Las tendencias en movilidad de zonas residenciales | Numérico |

Tabla : Descripción del dataset de movilidad (Google Mobility, s.f.) Fuente: Elaboración propia

El análisis estadístico nos hace darnos cuenta de que los campos numéricos metro\_area y census\_fips\_code están vacíos. Son dos variables que podemos eliminar de nuestro estudio ya que no tienen datos que nos den información. Las demás variables se corresponden con los valores de movilidad y vemos que van desde el -100, valor que indica 0 movilidad y se corresponde con los meses de confinamiento a valores de más de 500 en los lugares de ocio al aire libre como parques, playas o jardines públicos. Aunque vemos que no tenemos datos para todos los días en todas las CCAA. Los datos totales disponibles pueden verse en la tabla 7. Podemos apreciar que en las ciudades autónomas de Ceuta y Melilla tenemos muchos valores nulos e incluso un rango de fechas menor.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **variable** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| metro\_area | 0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| census\_fips\_code | 0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| retail\_and\_recreation\_percent\_change\_from\_baseline | 14390 | -26,0349 | 24,4858 | -97 | -35 | -21 | -10 | 63 |
| grocery\_and\_pharmacy\_percent\_change\_from\_baseline | 13741 | -1,03202 | 24,02671 | -94 | -9 | 2 | 12 | 205 |
| parks\_percent\_change\_from\_baseline | 14350 | 12,60307 | 54,70334 | -93 | -16 | 6 | 33 | 543 |
| transit\_stations\_percent\_change\_from\_baseline | 13420 | -20,3475 | 26,21614 | -93 | -33 | -18 | -4 | 120 |
| workplaces\_percent\_change\_from\_baseline | 14391 | -22,7045 | 19,25588 | -92 | -31 | -18 | -11 | 70 |
| residential\_percent\_change\_from\_baseline | 14061 | 6,03371 | 7,616586 | -10 | 1 | 4 | 8 | 46 |

Tabla Análisis exploratorio del dataset de movilidad (Google Mobility, s.f.) Fuente: Elaboración propia

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CC.AA** | **date** | **retail\_and\_recreation\_percent\_change\_from\_baseline** | **grocery\_and\_pharmacy\_percent\_change\_from\_baseline** | **parks\_percent\_change\_from\_baseline** | **transit\_stations\_percent\_change\_from\_baseline** | **workplaces\_percent\_change\_from\_baseline** | **residential\_percent\_change\_from\_baseline** |
| Andalusia | 760 | 760 | 677 | 760 | 760 | 760 | 760 |
| Aragon | 760 | 760 | 760 | 760 | 760 | 760 | 760 |
| Asturias | 760 | 760 | 664 | 759 | 760 | 760 | 760 |
| Balearic Islands | 760 | 760 | 760 | 760 | 760 | 760 | 760 |
| Basque Country | 760 | 760 | 760 | 760 | 760 | 760 | 760 |
| Canary Islands | 760 | 760 | 760 | 760 | 760 | 760 | 760 |
| Cantabria | 760 | 760 | 750 | 742 | 760 | 760 | 760 |
| Castile and León | 760 | 760 | 760 | 760 | 760 | 760 | 760 |
| Castile-La Mancha | 760 | 760 | 674 | 760 | 760 | 760 | 760 |
| Catalonia | 760 | 760 | 760 | 760 | 760 | 760 | 760 |
| Ceuta | 753 | 735 | 733 | 735 | 11 | 735 | 566 |
| Community of Madrid | 760 | 760 | 759 | 760 | 760 | 760 | 760 |
| Extremadura | 760 | 760 | 759 | 760 | 760 | 760 | 760 |
| Galicia | 760 | 760 | 671 | 760 | 760 | 760 | 760 |
| La Rioja | 760 | 760 | 757 | 743 | 735 | 760 | 760 |
| Melilla | 754 | 735 | 639 | 735 | 514 | 736 | 575 |
| Navarre | 760 | 760 | 757 | 756 | 760 | 760 | 760 |
| Region of Murcia | 760 | 760 | 670 | 760 | 760 | 760 | 760 |
| Valencian Community | 760 | 760 | 671 | 760 | 760 | 760 | 760 |

Tabla Conteo de datos disponibles del dataset de movilidad (Google Mobility, s.f.) Fuente: Elaboración propia

En el gráfico, se puede apreciar la tendencia a la movilidad de la gente estos dos años por CCAA. Lo primero que notamos es la bajada de movilidad en marzo de 2020, cuando comenzó el confinamiento. A partir de mayo 2020, comenzó la desescalada, y la tendencia a la movilidad comienza a subir. Vemos que en los lugares como supermercados o farmacias (grocery\_and\_pharmacy\_percent\_change\_from\_baseline) se aprecia una tendencia más o menos lineal, salvo en Cantabria en el verano de 2021 que aumentó mucho. Por otro lado, en parques y playas (parks\_percent\_change\_from\_baseline), se ve que la mayor movilidad es en los meses de verano y parece que en las comunidades costeras

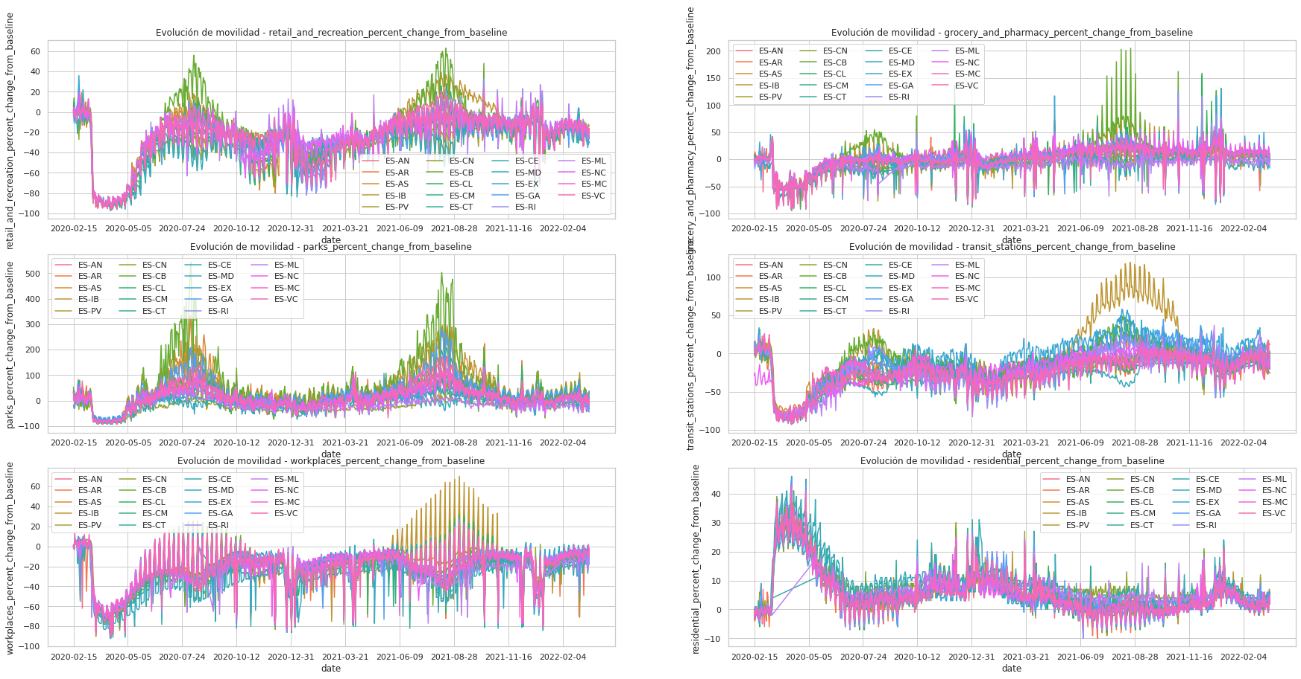


Ilustración Tendencia de movilidad en las distintas ubicaciones por cada cc.aa. Fuente: Elaboración propia

* + - 1. Dataset de vacunación en España

Para analizar los datos de vacunación en España, se ha escogido el dataset (Vacunas, s.f). En este dataset podemos ver los datos separados por fecha y CCAA y nos muestran el número de vacunas entregadas a cada comunidad total y de cada farmacéutica, así como el número de personas vacunadas con una dosis o la pauta completa. El dataset está formado por un total de 5605 registros y 15 columnas. Las fechas entre las que se comprenden los datos es desde 2021-01-04 hasta 2022-03-04. La descripción de las variables se puede ver en la tabla 8:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Descripción** | **Tipo** |
| fecha | Fecha de publicación del dato | Texto |
| cod\_ine | Código ine de la CCAA | Numérico |
| CCAA | Nombre de la CCAA | Texto |
| dosis\_Pfizer | Nº de dosis entregadas ese día de Pfzier | Numérico |
| dosis\_Moderna | Nº de dosis entregadas ese día de Moderna | Numérico |
| dosis\_AstraZeneca | Nº de dosis entregadas ese día de AstraZeneca | Numérico |
| dosis\_Janssen | Nº de dosis entregadas ese día de Janssen | Numérico |
| dosis\_totales | Nº total de dosis entregadas ese día | Numérico |
| dosis\_administradas | Nº de dosis administradas ese día | Numérico |
| porcentaje\_dosis\_administradas | Porcentaje de dosis administradas por 100 habitantes | Numérico |
| porcentaje\_entregadas | Porcentaje sobre entregadas | Numérico |
| personas\_una\_dosis | Personas con al menos una dosis | Numérico |
| personas\_pauta\_completa | Personas con pauta completa | Numérico |
| porcentaje\_pauta\_completa | Porcentaje con pauta completa | Numérico |
| fecha\_ultima\_vacuna | Fecha de la última vacuna registrada | Texto |

Tabla Descripción del dataset de vacunación (Vacunas, s.f.). Fuente: Elaboración propia

El análisis estadístico del dataset completo nos dice que más del 87% de la población de las CCAA está vacunada y que la CCAA que más dosis ha puesto, ha contabilizado más de 16M de vacunas.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **index** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| dosis\_Pfizer | 5605 | 1826320,505 | 2450395,266 | 1005 | 171915 | 873147 | 2358690 | 11421630 |
| dosis\_Moderna | 5472 | 390333,1488 | 609771,9834 | 200 | 23500 | 151500 | 477490 | 3508880 |
| dosis\_AstraZeneca | 5130 | 380285,4561 | 439023,0527 | 400 | 68775 | 228000 | 466170 | 1805800 |
| dosis\_Janssen | 4161 | 100949,9183 | 115891,4133 | 250 | 22145 | 58600 | 122515 | 525450 |
| dosis\_totales | 5472 | 2726683,94 | 3626599,303 | 1205 | 282015 | 1351943,5 | 3522840 | 17673830 |
| dosis\_administradas | 5605 | 2494885,23 | 3426052,344 | 153 | 189103 | 1201701 | 3124143 | 16880745 |
| porcentaje\_dosis\_administradas | 5605 | 100,0724746 | 68,43836376 | 0,01 | 26,07 | 116,54 | 157,44 | 222,95 |
| porcentaje\_entregadas | 5605 | 89,10785727 | 11,38958323 | 1,3 | 86,2 | 91,7 | 95,3 | 184,2 |
| personas\_una\_dosis | 4389 | 1644869,115 | 1811672,508 | 9092 | 439746 | 944063 | 1939135 | 7451485 |
| personas\_pauta\_completa | 5434 | 1178415,208 | 1638536,547 | 0 | 83003 | 524112 | 1565251,5 | 7061199 |
| porcentaje\_pauta\_completa | 5605 | 45,79008742 | 32,98592509 | 0 | 7,78 | 56,62 | 78,18 | 87,62 |

Tabla : Análisis estadístico del dataset de vacunación (Vacunas, s.f.). Fuente: Elaboración propia

En la tabla 10 se puede observar el conteo de datos de este dataset por cc.aa.. Vemos que hay 295 datos y que encontramos nulos en las variables “personas\_una\_dosis” y “personas\_pauta\_completa”, mientras que “dosis\_administradas” y el “porcentaje\_pauta\_completa” no tienen valores nulos que podrían usarse para estimar los nulos.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CCAA** | **fecha** | **cod\_ine** | **dosis\_administradas** | **porcentaje\_dosis\_administradas** | **personas\_una\_dosis** | **personas\_pauta\_completa** | **porcentaje\_pauta\_completa** |
| Andalucía | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Aragón | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Asturias | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Baleares | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| C Valenciana | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Canarias | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Cantabria | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Castilla La Mancha | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Castilla y Leon | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Cataluña | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Ceuta | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Extremadura | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Galicia | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| La Rioja | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Madrid | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Melilla | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Murcia | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| Navarra | 295 | 295 | 295 | 295 | 231 | 286 | 295 |
| País Vasco | 295 | 295 | 295 | 295 | 231 | 286 | 295 |

Tabla : Conteo de datos del dataset de vacunación (Vacunas, s.f.). Fuente: Elaboración propia

Analizamos el avance de la vacunación por cada CCAA en la figura 7 y vemos que el avance ha sido positivo y prácticamente al mismo ritmo en todas las CCAA, ya que las curvas de las gráficas que miden porcentajes siguen la misma tendencia para todas las comunidades. También observamos que las fechas en las que más dosis se han administrado va de abril a agosto de 2021 siendo donde la curva tiene una pendiente positiva más pronunciada.

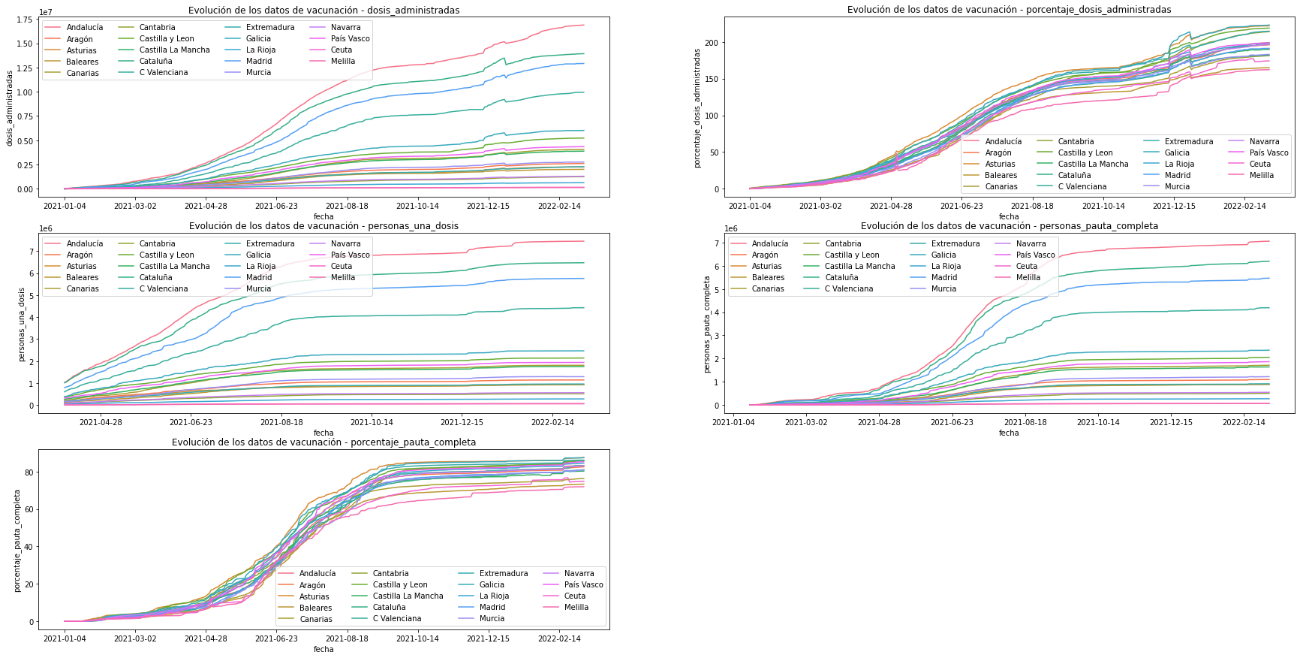


Ilustración Gráfico de la evolución en la vacunación por cc.aa. Fuente: Elaboración propia

* + - 1. Dataset de ingresos

Para conocer el estado de los ingresos en hospital debidos a la enfermedad COVID-19, se ha escogido el dataset (ingresos, s.f.). En este dataset vemos los datos ordenados por fecha y comunidad autónoma, con datos desde el 01/01/2020 hasta el hasta el. Contiene un total de 15295 filas y 7 columnas, detalladas en la tabla 11.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Descripción** | **Tipo** |
| Fecha | Fecha de los datos | Texto |
| cod\_ine | Código ine de la CCAA | Numérico |
| CCAA | Nombre de la CCAA | Texto |
| Casos | Número total de casos por covid | Numérico |
| Fallecidos | Número de fallecidos por covid | Numérico |
| Hospitalizados | Número total de pacientes hospitalizados por Covid | Numérico |
| UCI | Número total de pacientes ingresados en las UCIS de los hospitales con Covid | Numérico |

Tabla Descripción del dataset de ingresos (ingresos, s.f.). Fuente: Elaboración propia

En el análisis estadístico de los datos se puede ver en la tabla 12. Si comparamos los máximos de las variables con la media de las mismas, vemos que está última parece ser baja. Esto puede ser debido al rango de datos, que va desde el 01 de enero de 2020, antes de que se declarara el primer caso de covid en España.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **index** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| cod\_ine | 15409 | 10 | 5,4774033 | 1 | 5 | 10 | 15 | 19 |
| Casos | 15409 | 734,47615 | 2177,5028 | 0 | 24 | 137 | 538 | 46169 |
| Fallecidos | 15409 | 6,5611007 | 16,421801 | 0 | 0 | 1 | 6 | 334 |
| Hospitalizados | 15409 | 32,722305 | 76,895333 | 0 | 2 | 10 | 32 | 1933 |
| UCI | 15409 | 3,1239535 | 6,6821143 | 0 | 0 | 1 | 3 | 135 |

Tabla Análisis estadístico del dataset de ingresos. Fuente: Elaboración propia

Para ver la distribución de los datos, podemos fijarnos en la ilustración 9, un gráfico boxplot de las variables numéricas donde podemos ver los datos atípicos como pueden ser los valores de más de 300 casos en Fallecidos , valores de más de 1500 hospitalizados o de más de 100 camas ocupadas en UCI. Estos valores pueden deberse a los picos de las olas. La ilustración 10 es un histograma donde se ve la tendencia de los datos, más concentrados en valores cercanos al mínimo que al máximo obtenido en los valores estadísticos (véase tabla 12).

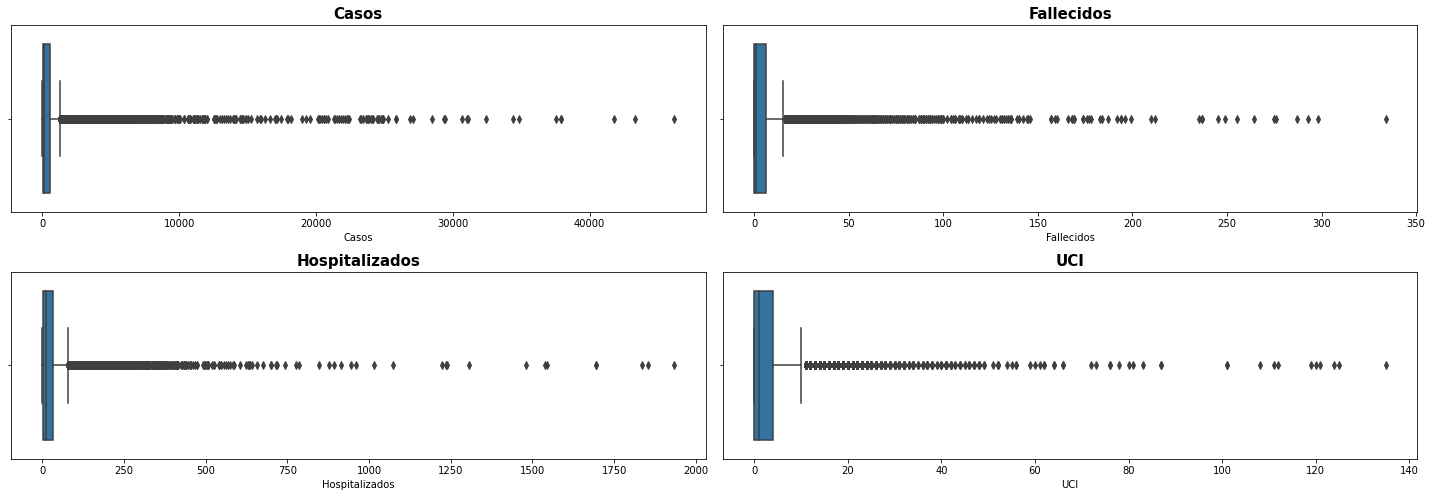


Ilustración 9 Bloxplot casos, fallecidos e ingresos

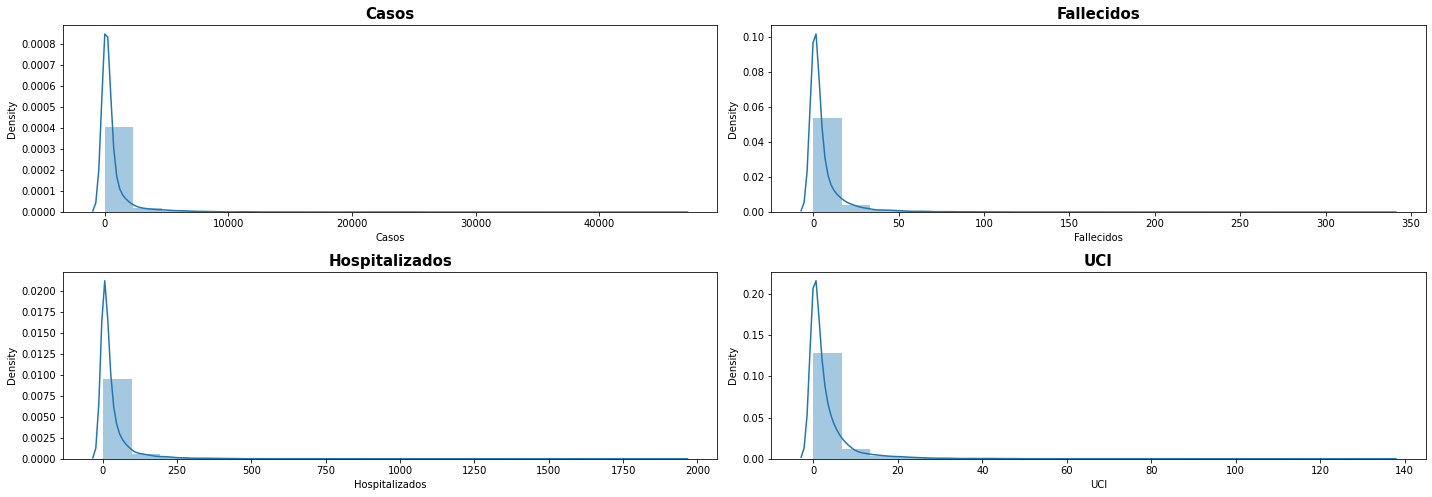


Ilustración 10 Histograma casos, fallecidos e ingresos

En la ilustración 11, vemos la evolución de las distintas variables en el tiempo, mientras en las primeras olas, la incidencia de las variables parece crecer y decrecer al mismo ritmo, vemos como en las dos últimas olas, el número de casos es elevado pero las hospitalizaciones y la mortalidad parecen ir descendiendo

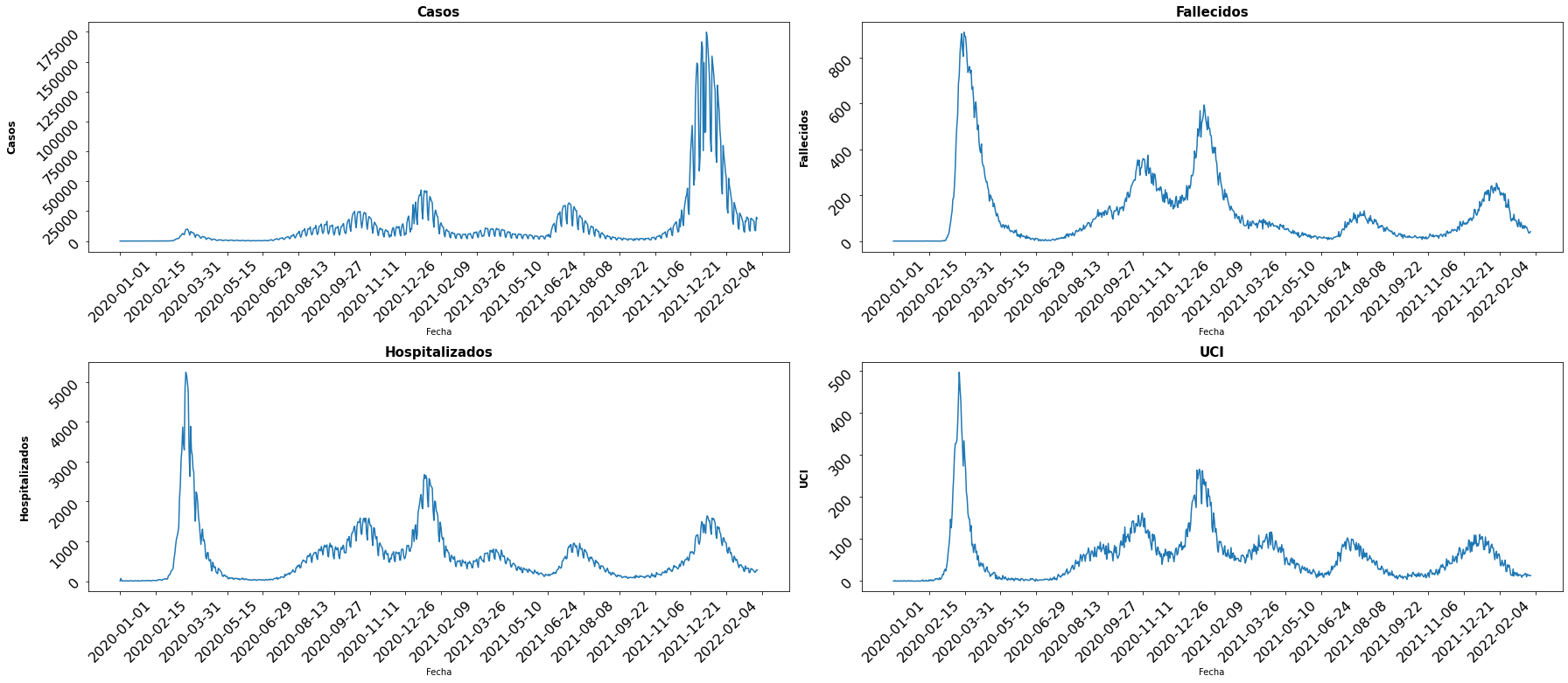


Ilustración 11 Evolución por día de los casos, fallecimientos y hospitalizaciones

### Fase de preprocesamiento de datos

Durante el análisis de los datos se ha observado que no todos son relevantes para este estudio, por lo que se va a eliminar aquellas variables que no nos aporten conocimiento.

* Dataset Covid:
  + La variable numérica num\_casos es el total de casos de cada CCAA para cada día, por lo que es la variable que se utilizará para nuestro estudio. Las demás contienen datos irrelevantes por lo que se descartarán.
* Dataset de movilidad:
  + Las variables country\_region y country\_region code indican el país al que pertenecen los datos. Este trabajo es sobre el comportamiento del covid19 en España, por lo que el valor es el mismo para todos los registros y por eso se descartarán dichas columnas.
  + La variable sub\_region\_2 indica la provincia en la que se tomaron los datos. Nuestro estudio es por comunidad autónoma, por lo que filtramos los datos con la condición de que esta variable tenga valor nulo (así indica los casos totales por CCAA) y luego la descartamos ya que no nos interesa los datos por provincias.
  + En el análisis estadístico, hemos visto que census\_fips\_code y metro\_area son variables con valores nulos en todo el dataset. Indican datos relacionados con condados y áreas metropolitanas en los países que existan. En España no tenemos esa división de territorio por lo que también se pueden descartar dichas variables.
* Dataset de vacunación:
  + Se descartan las columnas dosis\_Pzifer, dosis\_Moderna, dosis\_AstraZeneca y dosis\_Janssen, dosis\_totales y porcentaje\_entregadas ya este estudio no se basa en la farmacéutica que elaboró cada vacuna y dichas variables se corresponden con las dosis entregadas. Que una dosis haya sido entregada a la CCAA no significa que se haya puesto a su población y por eso este dato no es útil en el estudio.
  + Se elimina la variable fecha\_ultima\_vacuna ya que tampoco es relevante para este estudio.
* Dataset ingresos:
  + Se descarta la columna de “Casos”, ya que este dato lo hemos obtenido con el dataset covid (dataset COVID, s.f)

También se han de buscar aquellos valores nulos. En los dataset COVID e ingresos no se han encontrado valores nulos. Por otro lado, en el dataset de movilidad, tal y como se muestra en la tabla Z, se han encontrado varios nulos, sobre todo en las ciudades de Ceuta y Melilla. Al ser una cantidad de datos significativa, se ha optado por interpolar los datos faltantes ya que eliminarlos sería reducir demasiado el set de datos para esas dos comunidades autónomas. En las ilustraciones 12 y 13, vemos el estado de los datos sin interpolar y las ilustraciones 14 y 15 representa los datos interpolados.

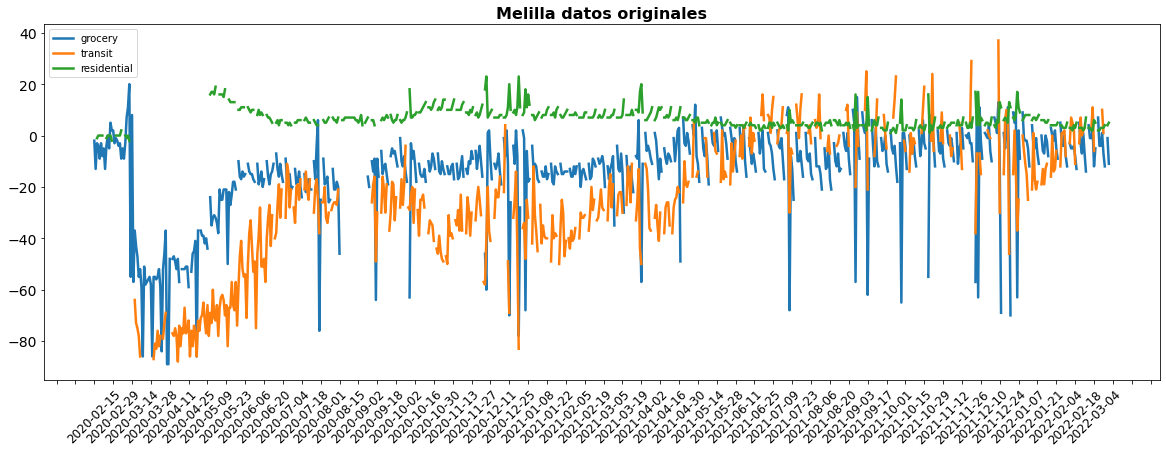


Ilustración : Situación de los datos de Melilla antes de interpolar. Fuente: Elaboración propia

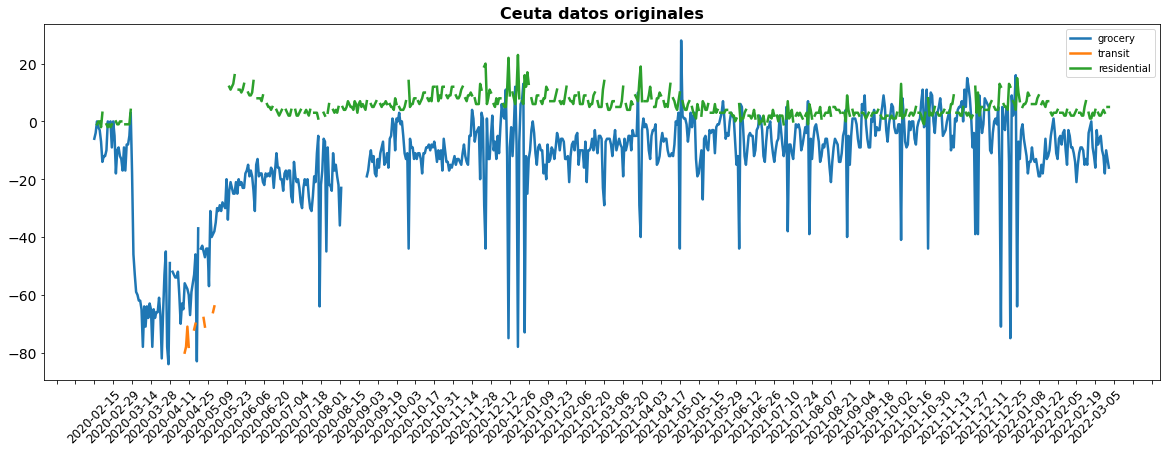


Ilustración : Situación de los datos de Ceuta antes de interpolar. Fuente: Elaboración propia

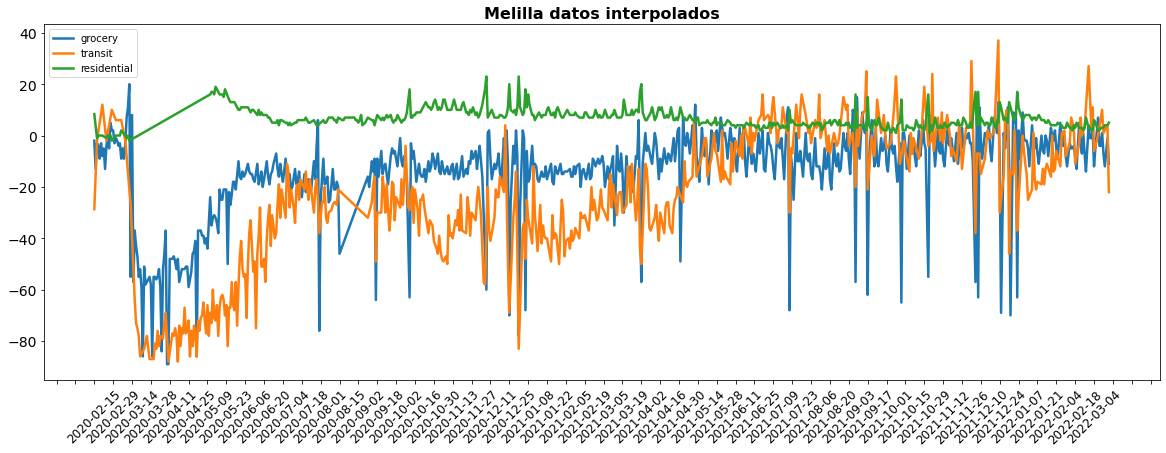


Ilustración Situación de los datos de Melilla después de interpolar. Fuente: Elaboración propia

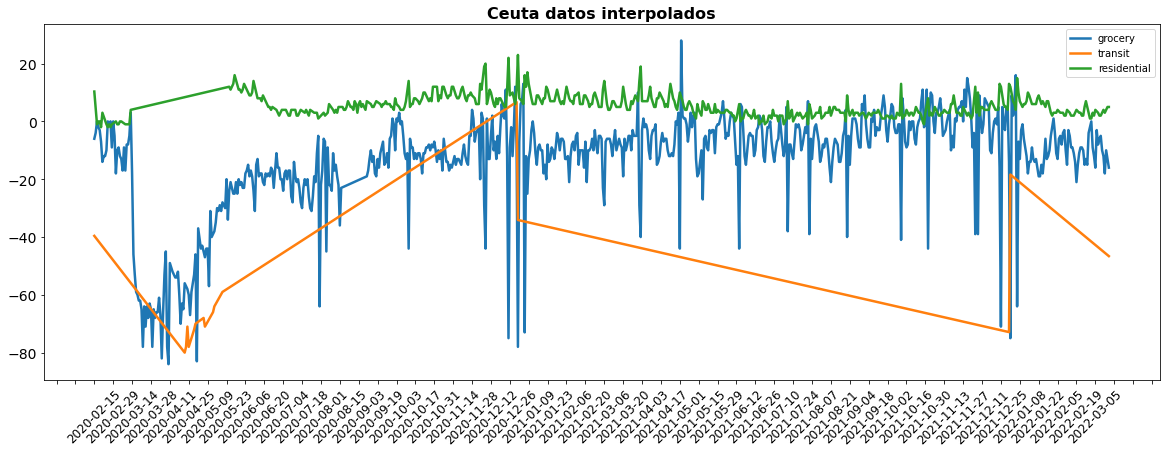


Ilustración Situación de los datos de Ceuta después de interpolar. Fuente: Elaboración propia

En el dataset de vacunación, los nulos los encontramos en los datos de “personas\_una\_dosis” y “personas\_pauta\_completa”, sobre todo en los primeros días. Para la columna “personas\_una\_dosis”, se ha optado por rellenar los nulos con la suma de los valores de la columna “dosis\_administradas” + valor de “personas\_una\_dosis” del día anterior para esa misma CCAA, ya que esa columna contiene un valor acumulativo y dado que la segunda vacuna se puso al menos un mes más tarde que la primera, pueden coincidir dichos valores. Para la columna “personas\_paura\_completa” se ha observado que los valores nulos van en el rango de fechas desde 04/01/2021 hasta 15/01/2021 y se corresponde con un valor de la columna “porcentaje\_pauta\_completa” = 0. Por lo que se van a rellenar los nulos con el valor 0.

### Fase de transformación de datos

En las fases anteriores, hemos seleccionado los datos a usar para este análisis. Dichos datos vienen de distintas fuentes y tienen distintos formatos. En esta fase se van a realizar las transformaciones necesarias para estandarizar las distintas fuentes de datos y poder unirlas en un único dataset que será nuestros datos de entrada en la fase siguiente; minería de dato. Las distintas transformaciones que se realizarán son:

* Unificar los códigos ISO y los nombres de las CCAA en los tres dataset de entrada. Para ello se ha creado manualmente la tabla códigos\_ccaa.csv (códigos\_ccaa, s.f.)
* Estandarizar el nombre y formato de la variable que indica la fecha de la medida. El nombre común para todos los dataset será fecha y el formato DD-MM-YYYY
* Renombrar las columnas del dataset de movilidad como se indica en la tabla

|  |  |
| --- | --- |
| **Viejo nombre** | **Nuevo nombre** |
| sub\_region\_1 | CCAA |
| iso\_3166\_2\_code | ccaa\_iso |
| date | fecha |
| retail\_and\_recreation\_percent\_change\_from\_baseline | lug\_ocio |
| grocery\_and\_pharmacy\_percent\_change\_from\_baseline | lug\_mercado |
| parks\_percent\_change\_from\_baseline | lug\_parque |
| transit\_stations\_percent\_change\_from\_baseline | lug\_estacion |
| workplaces\_percent\_change\_from\_baseline | lug\_trabajo |
| residential\_percent\_change\_from\_baseline | lug\_residencial |

Tabla : Relación entre los viejos nombres y los nuevos nombres del dataset de movilidad. Fuente: Elaboración propia

* En el dataset de datos\_covid, modificamos el valor ccaa\_iso para que case con los estándares ISO\_3166, añadiendo la cadena “ES-“
* El dataset de vacunación le añadimos la columna ccaa\_iso, cruzando los datos de la columna cod\_ine, en la tabla
* El dataset de ingresos le añadimos la columna ccaa\_iso, cruzando los datos de la columna cod\_ine, en la tabla

Una vez transformados cada uno de los dataset individualmente, procedemos a su unión, con lo que nos queda un dataset de 14725 registros y 15 columnas como se describe en la tabla 12.

|  |  |  |
| --- | --- | --- |
| **variable** | **Descripción** | **Tipo** |
| ccaa\_iso | Código ISO de la CCAA de residencia | Texto |
| fecha | fecha de los datos | datetime |
| num\_casos | Número de casos por técnica diagnóstica y CCAA de residencia | Numérico |
| lug\_ocio | Tendencias de movilidad de lugares como restaurantes, cafeterías, centros comerciales, parques de atracciones, museos, bibliotecas y cines | Numérico |
| lug\_mercado | Tendencias de movilidad de lugares como supermercados, almacenes de comida, mercados de productos agrícolas, tiendas de comida especializada y farmacias | Numérico |
| lug\_parque | Tendencias de movilidad de lugares como parques locales, parques nacionales, playas públicas, puertos deportivos, parques caninos, plazas y jardines públicos | Numérico |
| lug\_estacion | Tendencias de movilidad de lugares como estaciones de transporte público (por ejemplo, de metro, tren y autobuses) | Numérico |
| lug\_trabajo | Las tendencias en movilidad de lugares de trabajo. | Numérico |
| lug\_residencial | Las tendencias en movilidad de zonas residenciales | Numérico |
| Fallecidos | Número de fallecidos por covid | Numérico |
| Hospitalizados | Número de personas hospitalizadas por covid | Numérico |
| UCI | Número de personas hospitalizadas en UCI por covid | Numérico |
| dosis\_administradas | Nº de dosis administradas ese día | Numérico |
| porcentaje\_dosis\_administradas | Porcentaje de dosis administradas por 100 habitantes | Numérico |
| personas\_una\_dosis | Personas con al menos una dosis | Numérico |
| personas\_pauta\_completa | Personas con pauta completa | Numérico |
| porcentaje\_pauta\_completa | Porcentaje con pauta completa | Numérico |
| CCAA | Nombre de la CCAA | Texto |
| cod\_ine | Código numérico asignado a cada CCAA |  |

Tabla : Descripción de las variables del dataset unificado. Fuente: Elaboración propia

Al unir varios dataframes con distinto rango de fechas, nos han quedado valores nulos. El rango de fechas del dataset de movilidad es desde 15/02/2020 hasta 15/03/2022, el rango de fechas del dataset Covid es desde 31/01/2020 hasta 15/03/2022 y el rango de fechas del dataset de vacunación va desde 04/01/2021 hasta 04/03/2022. Al hacer el merge de los datos, se ha hecho left join con el dataset de covid al ser el que tiene mayor rango de fechas. Por lo que las variables de los demás dataset son las que contienen los nulos y se tratarán de la siguiente manera:

* Las variables cuyo origen es el dataset de movilidad contienen valores nulos desde el 31/01/2020 hasta el 15/02/2020. En esos días no había restricciones de movilidad y son fechas que no coinciden con ningún puente o fiesta que haga incrementar la movilidad por lo que dichos valores nulos se sustituyen con los datos de movilidad de los días desde 15/02/2020 hasta 15 días después.
* Las variables cuyo origen es el dataset de vacunas contienen nulos en dos periodos: desde el 31/01/2020 hasta el 04/01/2021, que es el periodo donde no había vacuna y desde 04/03/2022 hasta el 15/03/2022 que no hay datos disponibles. En el primer periodo, se rellenará con el valor 0 los nulos y en el segundo periodo se copiarán los datos de la fecha 04/03/2022, al ser los datos más recientes que tenemos.
* También hemos observado que el dataset de vacunación no tiene datos de los fines de semana, por lo que dichos datos se interpolarán con los del viernes.

Para poder elaborar un modelo óptimo en los siguientes pasos, se ha añadido información al dataframe. Se añaden las columnas indicadas en la tabla 13.

|  |  |  |
| --- | --- | --- |
| **variable** | **Descripción** | **Tipo** |
| dia\_sem | Número de día de la semana cuya correspondencia es:   * 0 : Lunes * 1 : Martes * 2 : Miércoles * 3 : Jueves * 4 : Viernes * 5 : Sábado * 6 : Domingo | Numérico |
| dia\_year | Número de día del año. Rango [1 : 365-366] | Numérico |
| dia\_mes | Día del mes [1 : 31] | Numérico |
| is\_weekend | Indica si el día es fin de semana (sábado o domingo ) | Numérico |

Tabla : Descripción de las columnas adicionales. Fuente: Elaboración propia

### Fase de minería de dato

* + - 1. Análisis de correlaciones

En la ilustración 16, se puede ver la correlación entre los distintos atributos que componen el dataset. Se puede apreciar que hay una correlación con tendencia positiva entre el número de casos y los datos de fallecimientos e ingresos, así como con la vacunación la cual puede ser debida a la incidencia de la variante ómicron, cuyos máximos están en diciembre 2021 (ver ilustración 4) y coinciden con los máximos de vacunación (ver ilustración 8). También vemos que existe correlación entre el porcentaje de vacunación de pauta completa con el movimiento entre lugares de ocio, mercados, parque o estaciones y por el contrario del movimiento entre lugares residenciales, que la correlación es negativa. Con esto podemos decir que a medida que la gente se iba vacunando, dejo de pasar tanto tiempo en casas y empezó a visitar lugares públicos. Por otro lado, también observamos una correlación negativa entre el porcentaje de vacunados con pauta completa y los ingresos y fallecimientos. Esto puede ser debido a la eficacia de las vacunas.

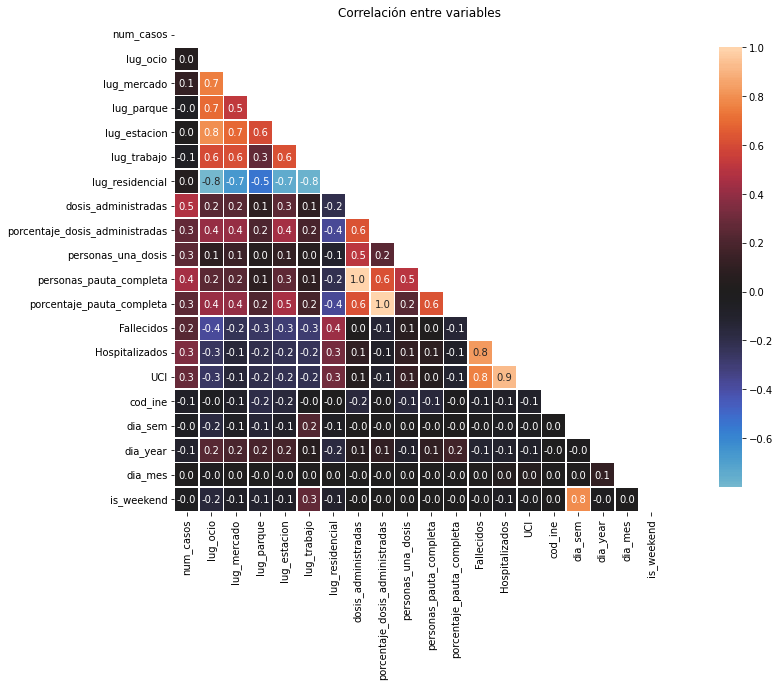


Ilustración Correlación entre variables. Fuente: Elaboración propia

El objetivo de este proyecto es ver como la movilidad y la vacunación afectan la incidencia del covid. En la ilustración 17 podemos ver esa relación. Vemos que los valores más altos se corresponden con los datos de vacunación y los más bajos con la asistencia a lugares públicos como parques o puestos de trabajo. No es el resultado esperado, ya que, habiendo vacuna, el número de casos debería haber tenido una correlación negativa con la vacunación, pero como se ha comentado anteriormente, la alta incidencia de la última ola en diciembre de 2021 puede ser la causante de estos resultados. Por ello vamos a analizar la correlación en un subconjunto de datos que abarque el rango de fechas desde 31/01/2020 hasta el 21/11/2021 que es cuando empieza la sexta ola en España según la ilustración 4.

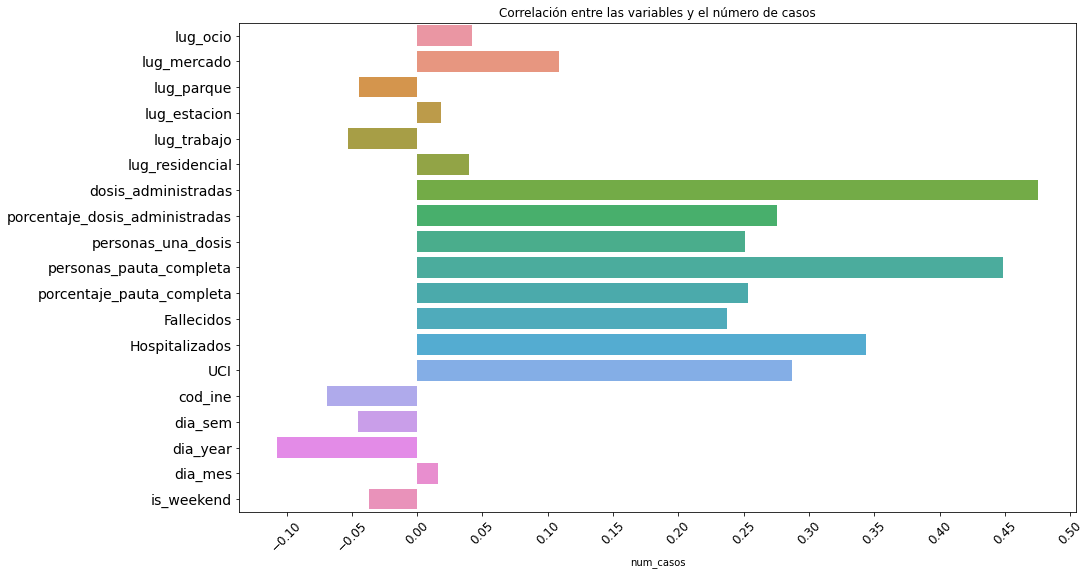


Ilustración Correlación entre variables y el número de casos

* + - 1. Análisis de árbol de decisión con valores por defecto

Para poder explicar mejor nuestro objetivo, se ha optado por usar modelos basados en árboles de decisión que nos permitan identificar las variables con más peso en las predicciones (Keiko A Herrick, 2013). Por otro lado, se ha demostrado que es de las mejores técnicas en predicción de datos en enfermedades (Michael J. Kane, 2014).

Así, los modelos inicialmente seleccionados son:

METER ESTAS REFERENCIAS EN EL ANEXO

* **RandomForest**. (Breiman, Random Forests. Machine Learning 45, 5–32 , 2001) que usa la técnica de Bagging o Bootstrap Aggregating (Breiman, Bagging predictors. Machine learning, 1996): “Bagging es un método para generar múltiples versiones de un predictor y usarlas para obtener un predictor agregado. La agregación promedia sobre las versiones al predecir un resultado numérico y hace un voto de pluralidad al predecir una clase”
* **AdaBoost**. (Sammut C., 2017) ,GradientBoosting. (Friedman, February 1999) y **XGBoost** que es una implementación open-source de Gradient Boosting diseñada para ser eficiente, flexible y portable. (XGBoost Documentation, s.f.). Estos modelos usan la técnica de Boosting (Freund, 1999), otra de las técnicas de conjuntos más utilizadas y que consiste en generar secuencialmente modelos débiles (weak learners) que “aprenden” de los errores del anterior e intentan minimizarlos.

Gradient --- Mirar los modelos que se usan y describirlos

METER AQUÍ LA EXPLICACION DE LOS ERRORES

Para aplicar el algoritmo, dividimos el dataset de entrada en dos en proporción 80% entrenamiento o train y 20% test

* Dataset entrenamiento: desde 31/01/2022 hasta 31/08/2021
* Dataset test: desde 01/09/2021 hasta 15/03/2022

Para evaluar la calidad de los modelos se ha optado por contabilizar los valores de los errores MAE, RMSE Y R2 ya que el uso de varias métricas ayudará a comprender las fortalezas y debilidades de un modelo en particular.

* **RMSE** (Error cuadrático medio) es la distancia media entre los valores reales y las predicciones realizadas por el modelo. Este error es más sensible a los valores atípicos al poner más peso en los errores más grandes. Se corresponde con la siguiente fórmula:
* **MAE** (Error absoluto medio) es el promedio de la diferencia entre los valores predichos y los valores reales. Su cálculo se corresponde con:
* **R2** (Coeficiente de determinación). Indica la proporción de datos que son explicados por el modelo. Cuanto más cerca de 1, mejor explicado está el modelo.

Ajustamos los modelos sobre los datos del dataset entrenamiento y hacemos la evaluación sobre los datos del dataset de test. La función utilizada para cada uno de los modelos se encuentra en la librería de scikit-learn y utilizará su definición por defecto. El resultado obtenido se puede ver en la tabla 16 e indica unos errores muy altos, es decir, los modelos no tienen suficientes datos para hacer una buena predicción por lo que se volverá a la fase de transformación de datos para ver que características se pueden modificar con el objetivo de bajar el error.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MAE** | **RMSE** | **SCORE** |
| RF | 1478,966 | 3842,662 | 0,155 |
| XGB | 1498,224 | 3902,494 | 0,128 |
| GB | 1528,746 | 3988,446 | 0,09 |
| AB | 1548,614 | 3962,463 | 0,101 |

Tabla Valores del error de los modelos con parametrización por defecto

* + - 1. Iteración I: Variación del número de casos

Para esta iteración vamos a centrarnos en el número de casos. El COVID-19 es un virus que se contagia de persona a persona, es decir, una persona que da positivo el día D-n, puede haber contagiado a una o varias personas que den positivo el día D. por lo que podemos asumir que el número de contagios del día D depende del número de contagios de los días anteriores D-n. Como las cuarentenas inicialmente eran de 10 días, añadiremos nuevos atributos que para un día D, añadan los datos de los 10 días anteriores (n є [1,10]).

Analizamos la correlación de los nuevos atributos con el número de casos del día actual (ilustración 18) y podemos ver que la correlación es positiva y en los primeros 7 días su valor es 0.9, cerca de ser perfecta. Por lo que podemos decir que los contagios del día D están influenciados por los contagios de los días anteriores.

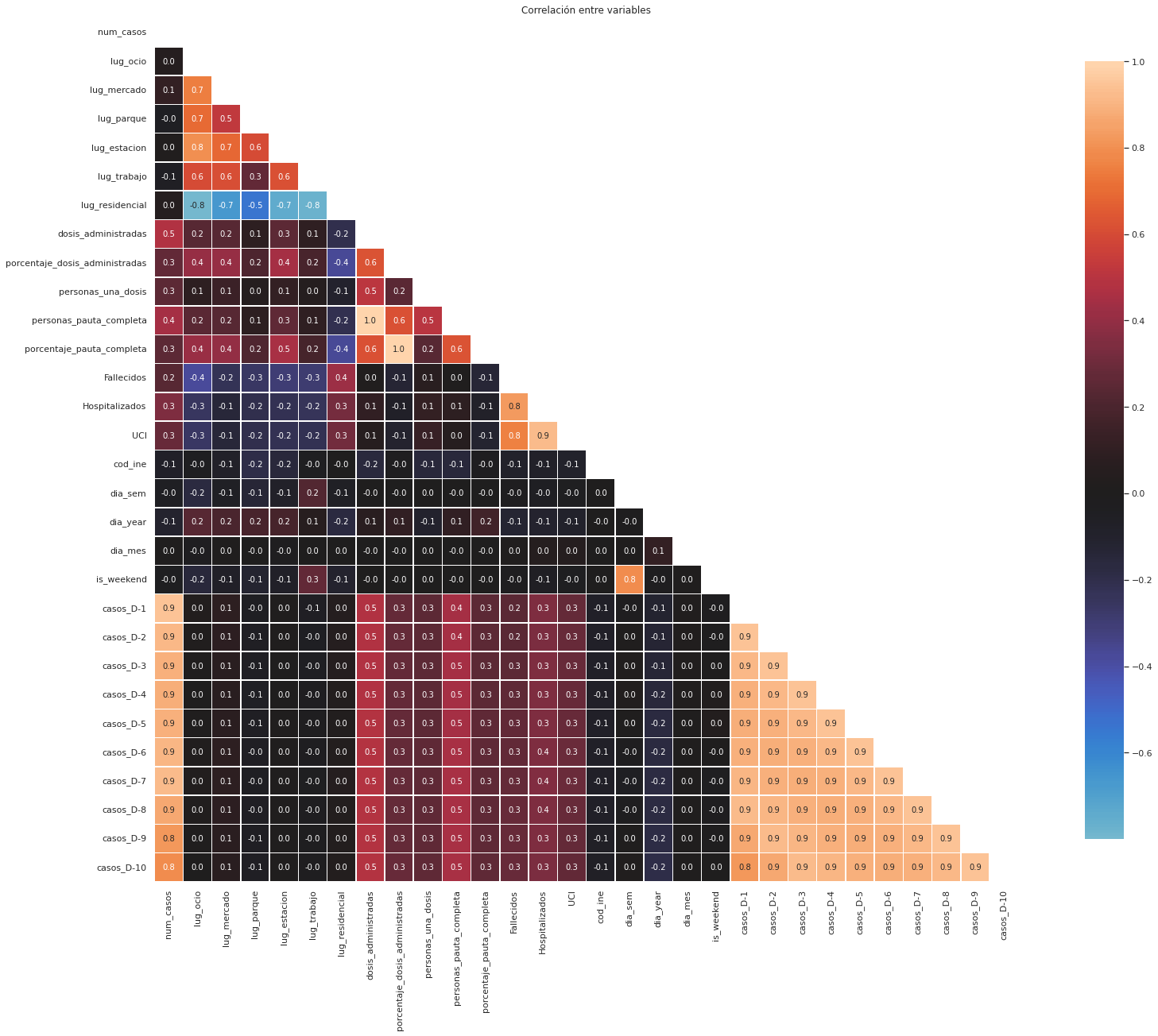


Ilustración Matriz de correlación de la iteración I

Calculamos los errores para los modelos de regresión con los nuevos atributos y vemos que hay una disminución notable de los errores, aunque los valores de error cuadrático siguen siendo altos (tabla 17). Por ello se opta por dividir los datos por comunidades y comprobar cuál sería el mejor modelo para cada una de ellas. Se pueden ver los datos en la tabla 18. Al separar los datos por CCAA se aprecia que en muchas de ellas el error es mucho más bajo que calculado con todos los datos juntos y además no sería óptimo aplicar el mismo modelo para todas.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Datos base** | | | **Iteración 1** | | |
| **MAE** | **RMSE** | **SCORE** | **MAE** | **RMSE** | **SCORE** |
| RF | 1478,966 | 3842,662 | 0,155 | 621,5 | 2574,71 | 0,565 |
| XGB | 1498,224 | 3902,494 | 0,128 | 614,954 | 2553,356 | 0,573 |
| GB | 1528,746 | 3988,446 | 0,09 | 571,288 | 2392,479 | 0,625 |
| AB | 1548,614 | 3962,463 | 0,101 | 597,035 | 2498,658 | 0,591 |

Tabla Valores del error de los modelos en la iteración I

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **nombre\_ccaa** | **cod\_ine** | **Modelo** | **MAE** | **RMSE** |
| Andalucía | 1 | GradientBoosting | 649,126 | 1443,634 |
| Aragón | 2 | XGBoost | 174,083 | 343,038 |
| Principado de Asturias | 3 | GradientBoosting | 131,988 | 263,71 |
| Cantabria | 6 | GradientBoosting | 67,583 | 144,587 |
| Ceuta | 18 | XGBoost | 15,302 | 23,962 |
| Castilla-León | 7 | GradientBoosting | 535,891 | 1117,723 |
| Castilla-La Mancha | 8 | XGBoost | 190,438 | 419,553 |
| Canarias | 5 | XGBoost | 335,1 | 653,404 |
| Cataluña | 9 | GradientBoosting | 3961,515 | 8780,937 |
| Extremadura | 11 | AdaBoost | 201,929 | 421,071 |
| Galicia | 12 | XGBoost | 284,142 | 518,263 |
| Islas Baleares | 4 | GradientBoosting | 159,389 | 333,163 |
| Región de Murcia | 14 | AdaBoost | 357,48 | 754,233 |
| Comunidad de Madrid | 13 | GradientBoosting | 1401,675 | 3331,812 |
| Melilla | 19 | RandomForest | 16,14 | 27,743 |
| Comunidad Foral de Navarra | 15 | AdaBoost | 180,653 | 441,255 |
| País Vasco | 16 | GradientBoosting | 457,436 | 1017,643 |
| La Rioja | 17 | AdaBoost | 63,541 | 130,293 |
| Comunidad Valenciana | 10 | GradientBoosting | 1619,482 | 3720,617 |

Tabla : Mejor algoritmo y sus métricas de error por CCAA

* + - 1. Iteración II: Variación de la movilidad

En la iteración I, hemos partido del supuesto de que los casos en días anteriores afectan a los casos del día en curso. De manera análoga podemos suponer que el movimiento de gente en el día D, está afectado por los días anteriores. Por ello se añaden los atributos D-n\*7 con n є [1, 4], que indican los datos de movilidad del mismo día de las 4 semanas anteriores. Se ha optado por añadir los datos de semana en semana porque no se puede comparar la movilidad entre fines de semana y días laborales ya que es muy diferente.

En la matriz de correlaciones para los datos de esta iteración, podemos ver que el número de casos no está muy relacionado con la movilidad de semanas anteriores, pero sí que la movilidad del día D, está bastante correlacionada con la de semanas anteriores. La matriz de correlaciones se puede ver en la ilustración19. Para una visualización mejor de la correlación de los nuevos atributos, se ha optado por no mostrar los correspondientes a la iteración 1, ya que su correlación era prácticamente 0.

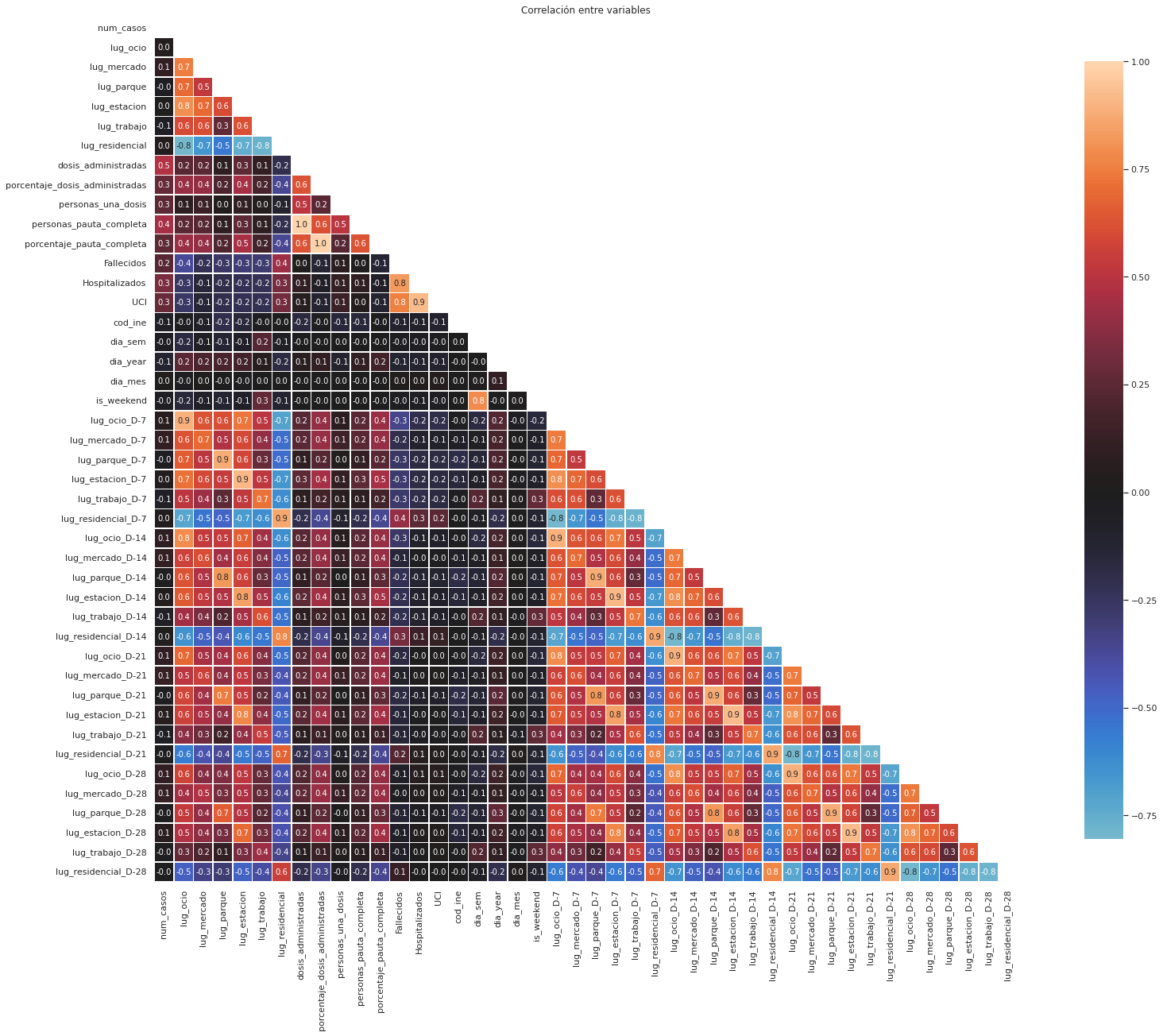


Ilustración 19 Matriz de correlación de la iteración II

En la iteración II vemos un ligero aumento de los errores al añadir los nuevos atributos y dado que la correlación del número de casos con los datos de movilidad de semanas anteriores es la misma que la correlación del número de casos con los datos de movilidad del mismo día, se descartarán los atributos de esta iteración.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Datos base** | | | **Iteración 1** | | | **Iteración 2** | | |
| **MAE** | **RMSE** | **SCORE** | **MAE** | **RMSE** | **SCORE** | **MAE** | **RMSE** | **SCORE** |
| RF | 1287,139 | 3498,258 | 0,198 | 621,5 | 2574,71 | 0,565 | 627,64 | 2596,997 | 0,558 |
| XGB | 1303,078 | 3579,2 | 0,16 | 614,954 | 2553,356 | 0,573 | 624,234 | 2577,272 | 0,565 |
| GB | 1262,985 | 3583,667 | 0,158 | 571,288 | 2392,479 | 0,625 | 572,016 | 2381,325 | 0,628 |
| AB | 1331,588 | 3698,025 | 0,103 | 597,035 | 2498,658 | 0,591 | 620,754 | 2580,323 | 0,563 |

Tabla : Valores del error de los modelos en la iteración I

* + - 1. Iteración III: Optimización de hiperparámetros

En los pasos anteriores hemos comprobado que añadir los atributos de casos en días anteriores ha bajado el error a más de la mitad mientras que en la iteración II ha subido un poco y por ello descartamos añadir los datos de dicha iteración. Para seguir ajustando el algoritmo, vamos a buscar los parámetros más óptimos del algoritmos (habitualmente llamado hiperparámetros) para cada CCAA. El planteamiento será del de buscar aquella parametrización que ofrezca los resultados de mayor calidad (con respecto a las métricas establecidas) y de mayor robustez. Scikit-learn proporciona diferentes herramientas para que la optimización de estos hiperparámetros pueda ser lo más sencilla posible. En concreto ofrece dos alternativas, la búsqueda en cuadrícula (grid search) y la búsqueda aleatoria (RandomSearch). Para asegurar la línea temporal de los datos, se utilizará la función TimeSeriesSplit.

En la tabla 20 se pueden ver los mejores hiperparámetros para cada algoritmo buscados con el método de RandomSearch. Podemos apreciar un descenso de los errores para los algoritmos de RandomForest y XGBoost respecto a los parámetros por defecto y un aumento de los errores en los otros dos algoritmos (Iteración I, véase tabla 18).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelo** | **Parametros** | **MAE** | **RMSE** | **Score** |
| RandomForest | n\_estimators : 16 | 620,709742 | 2574,910463 | 0,5653 |
| min\_samples\_split : 6 |
| min\_samples\_leaf : 2 |
| max\_depth : 12 |
| bootstrap : True |
| XGBoost | subsample : 0.7 | 582,297798 | 2462,14928 | 0,602539 |
| n\_estimators : 200 |
| min\_child\_weight : 4 |
| max\_depth : 2 |
| gamma : 0.1 |
| eta : 0.01 |
| colsample\_bytree : 0.8 |
| GradientBoosting | warm\_start : True | 606,703594 | 2504,429426 | 0,588772 |
| n\_estimators : 200 |
| min\_samples\_split : 4 |
| max\_features : auto |
| max\_depth : 3 |
| loss : ls |
| learning\_rate : 0.25 |
| AdaBoost | n\_estimators : 128 | 617,987379 | 2.581 | 0,563172 |
| loss : exponential |
| learning\_rate : 0.005 |

Tabla Valor de los parámetros y el error obtenido con cada combinación para cada algoritmo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **nombre\_ccaa** | **cod\_ine** | **Modelo** | **MAE** | **RMSE** |
| Andalucía | 1 | GradientBoosting | 649,126 | 1443,634 |
| Aragón | 2 | XGBoost | 161,579 | 340,37 |
| Principado de Asturias | 3 | XGBoost | 116,698 | 224,77 |
| Cantabria | 6 | GradientBoosting | 67,583 | 144,587 |
| Ceuta | 18 | AdaBoost | 13,719 | 26,423 |
| Castilla-León | 7 | GradientBoosting | 535,891 | 1117,723 |
| Castilla-La Mancha | 8 | XGBoost | 186,148 | 412,487 |
| Canarias | 5 | XGBoost | 319,136 | 623,904 |
| Cataluña | 9 | GradientBoosting | 3961,515 | 8780,937 |
| Extremadura | 11 | XGBoost | 187,013 | 376,254 |
| Galicia | 12 | XGBoost | 267,64 | 495,174 |
| Islas Baleares | 4 | XGBoost | 151,193 | 319,701 |
| Región de Murcia | 14 | XGBoost | 342,431 | 742,669 |
| Comunidad de Madrid | 13 | GradientBoosting | 1401,675 | 3331,812 |
| Melilla | 19 | GradientBoosting | 13,695 | 27,83 |
| Comunidad Foral de Navarra | 15 | AdaBoost | 180,653 | 441,255 |
| País Vasco | 16 | GradientBoosting | 457,436 | 1017,643 |
| La Rioja | 17 | AdaBoost | 63,541 | 130,293 |
| Comunidad Valenciana | 10 | GradientBoosting | 1619,482 | 3720,617 |

### Fase de interpretación

## Resultados o conocimiento extraído

# Conclusión y trabajos futuros

# Referencias

Android. (s.f.) Recuperado de https://www.android.com/

Apple. (s.f.). Recuperado de apple.com

Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. COVID-19 Outbreak Prediction with Machine Learning. Algorithms. 2020; 13(10):249. https://doi.org/10.3390/a13100249

COVID-19: cronología de la actuación de la OMS. (s.f.). Recuperado de https://www.who.int/es/news/item/27-04-2020-who-timeline---covid-19

cnecovid. (s.f.). Recuperado de https://cnecovid.isciii.es/covid19/

Datadista. (s.f). Recuperado de https://www.datadista.com/

Datadista. COVID 19. (s.f.) Recuperado de https://github.com/datadista/datasets/tree/master/COVID%2019

Escovid19data. (s.f.). Recuperado de https://lab.montera34.com/covid19/provincias.html

Google Mobility. (s.f.).Recuperado de https://www.google.com/covid19/mobility/?hl=es

Iqbal, N., Islam, M., Machine Learning for Dengue Outbreak Prediction: A Performance Evaluation of Different Prominent Classifiers. Informatica 43 (2019). https://doi.org/10.31449/inf.v43i1.1548

ISCIII. (s.f.). Recuperado de https://www.isciii.es

ISCIII Quiénes somos Misión y Visión. (s.f.) Recuperado de https://www.isciii.es/QuienesSomos/Paginas/QuienesSomos.aspx

Kaggle. (s.f.) Recuperado de https://www.kaggle.com/

Kaggle, COVID-19 in Spain. (s.f.) Recuperado de https://www.kaggle.com/danigarci1/covid19-in-spain

Kermack, W.O. y McKendrick (1972). A Contribution to the Mathematical Theory of Epidemics. Proceedings of the Royal

López, A. (s.f.) TFM Análisis COVID19. Recuperado de https://github.com/ailb1990/TFM\_analisis\_COVID19

Society of London. Series A, Containing Papers of a Mathematical and Physical Character. Vol. 115, N0. 772, pp. 700-721.

Martín Barroso, V. (2020) Una breve introducción al modelo SIR aplicado al caso del Covid-19. [ICEI Papers COVID-19; nº 12,]. https://www.ucm.es/icei/file/iceipapercovid12

Ministerio de transportes, movilidad y agenda urbana. Open data movilidad (s.f) Recuperado de mitma.es/ministerio/covid-19/evolucion-movilidad-big-data/opendata-movilidad

Qiang, X., Kou, Z. Scoring amino acid mutation to predict pandemic risk of avian influenza virus. BMC Bioinformatics 20, 288 (2019). https://doi.org/10.1186/s12859-019-2770-0

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y AlvaradoPérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas

de la formación profesional (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia.

doi: http://dx.doi.org/10.16925/9789587600490

Vacunas, s.f. Recuperado de https://raw.githubusercontent.com/datadista/datasets/master/COVID%2019/ccaa\_vacunas.csv

Withanage, G.P., Viswakula, S.D., Nilmini Silva Gunawardena, Y.I. et al. A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka. Parasites Vectors 11, 262 (2018). https://doi.org/10.1186/s13071-018-2828-2

# Anexos I

# Anexos II