Universidad de Valladolid

ESCUELA DE INGENIERÍA INFORMÁTICA DE SEGOVIA

Grado en Ingeniería Informática de Servicios y Aplicaciones

Análisis, mediante técnicas de Machine Learning, de la efectividad de las medidas aplicadas contra el COVID-19 en Castilla y León.

Alumno: Álvaro Fuentes Valverde

Tutores: José Vicente Álvarez Bravo y Anibal Bregón Bregón

"Sólo podemos ver poco del futuro,
pero lo suficiente para darnos cuenta
de que hay mucho que hacer."
-Alan Turing

Agradecimientos

En primer lugar, me gustaría dar las gracias a mis tutores José Vicente Álvarez y Aníbal Bregón, sin los cuales este proyecto no habría podido desarrollarse con éxito. También querría incluir en este agradecimiento a todo el personal docente de la Escuela de Ingeniería Informática de Segovia por el gran esfuerzo y dedicación mostrado a lo largo de toda mi trayectoria como estudiante.
En segundo lugar, querría agradecer a mi familia todo su soporte y apoyo en las decisiones que he tomado a lo largo de mi formación, sin ella no habría llegado a ser la persona que soy hoy en día. Mencionar también a mis amigos de toda la vida, los cuales siempre han estado a mi lado y nunca me han dejado de apoyar al igual que mis compañeros y amigos de la Universidad cuya ayuda y consejos han sido esenciales a lo largo de mi etapa universitaria.
Y finalmente me gustaría agradecer y dedicar este proyecto a todas aquellas personas que han estado en primera linea en la lucha contra la pandemia, desde el personal sanitario que se ha jugado la salud física y mental por salvar al mayor número de personas, hasta todos aquellos trabajadores o personas que de una forma u otra han estado y siguen presentes en la lucha contra el COVID-19.

Resumen

Debido a la pandemia COVID-19, España ha tenido que aplicar una serie de medidas y restricciones para la reducción del gran número de contagios y muertes provocadas por el virus. Estas restricciones han sido distintas en cada una de las Comunidades Autónomas en las que se divide el país, teniendo cada una de estas el poder de aplicar diferentes medidas en cada zona, causando un gran impacto en los habitantes y economía de cada territorio.

En este proyecto se han utilizado técnicas de Machine Learning como Clustering y Gradient Boosting Trees que, partiendo de una serie de datos de salud, medidas aplicadas y movilidad en Castilla y León, han permitido obtener las medidas más eficaces a aplicar en cada territorio. De esta forma, se facilita la toma de decisiones descartando aquellas restricciones ineficaces cuyo impacto en la población y en la economía es notable. Para este estudio y búsqueda de las mejores medidas se han creado distintas gráficas y análisis acerca de los datos obtenidos. Los métodos y modelos construidos en el proyecto se han utilizado posteriormente sobre zonas de salud de test para la evaluación final del trabajo realizado.

Palabras claves: COVID-19, Medidas y restricciones, Castilla y León, Zonas de salud,Machine Learning, Clustering, Gradient Boosting Trees, Evaluación.

Abstract

Due to the COVID-19 pandemic, Spain has had to apply a series of measures and restrictions to reduce the large number of infections and deaths caused by the virus. These restrictions have been different in each of the Autonomous Communities in which the country is divided, each of these having the power to apply different measures in each area, causing a great impact on the inhabitants and economy of each territory.

In this project, Machine Learning techniques such as Clustering and Gradient Boosting Trees have been used which, based on a series of health data, applied measures and mobility in Castilla y León, have allowed to obtain the most effective measures to be applied in each territory. In this way, decision-making is facilitated by discarding those ineffective restrictions whose impact on the population and the economy is significant. For this study and search for the best measures, different graphs and analyses have been created on the data obtained. The methods and models constructed in the project were subsequently used on test health zones for the final evaluation of the work carried out.

Keywords: COVID-19, Measures, Castilla y León, Health zones, Machine Learning, Clustering, Gradient Boosting Trees, Evaluation.

Índice general

Índice de figuras

Índice de tablas

Parte I

Memoria del Proyecto

Capítulo 1

Descripción del proyecto

1.1. Introducción

La pandemia provocada por el virus SARS-CoV-2 [3], denominada científicamente con el nombre COVID-19, ha cambiado no solamente nuestras vidas sino las del mundo entero. Este virus detectado el 31 de diciembre de 2019 (día en el que se notificó el primer brote), provocó en apenas unos meses la paralización completa del mundo entero al extenderse rápidamente por todo el globo tal y como podemos ver en las gráficas de contagios y muertes mostradas en las Figuras 1.1 y 1.2 o en el impacto económico mostrado en la Figura 1.3 . Una de las razones de su rápida propagación se debe a las múltiples formas de contagio, entre las que destaca la expulsión de partículas al hablar o estornudar (aerosoles) [9].

Los distintos síntomas del virus , similares a los de la gripe, tienen efectos graves en sectores de la población considerados de riesgo, como por ejemplo las personas mayores de 65 años, personas con enfermedades cardiovasculares o respiratorias, o personas afectadas por otras enfermedades graves. Esto produjo un colapso total en los sistemas sanitarios de todo el mundo, llegando a situaciones criticas como la alta presión en hospitales y UCIs debido al alto ratio de contagio del virus.

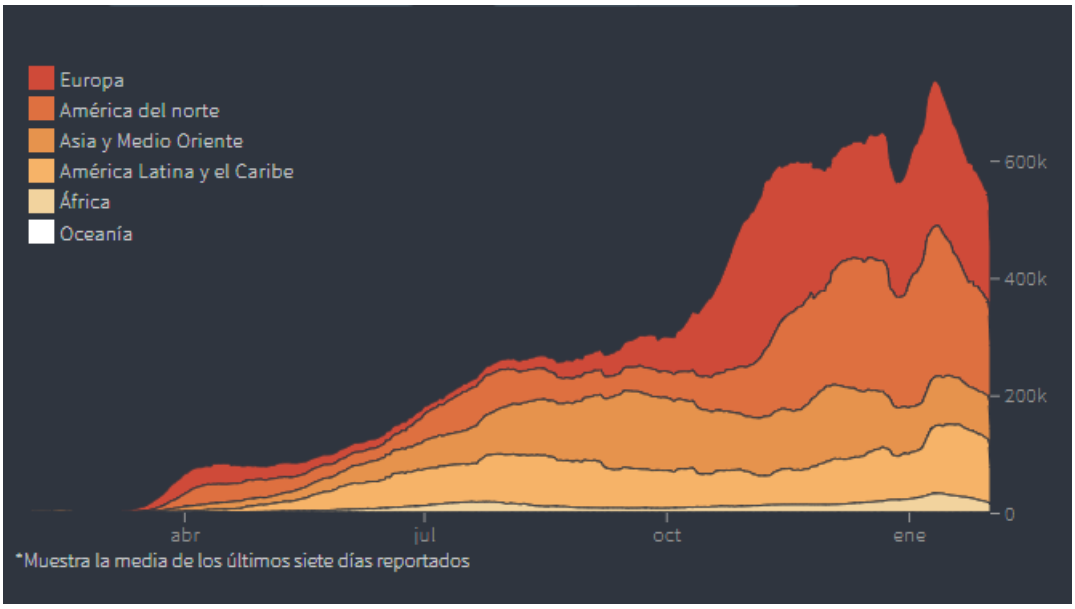


Figura 1.1: Gráfico de contagios globales desde el inicio de la pandemia.[32]

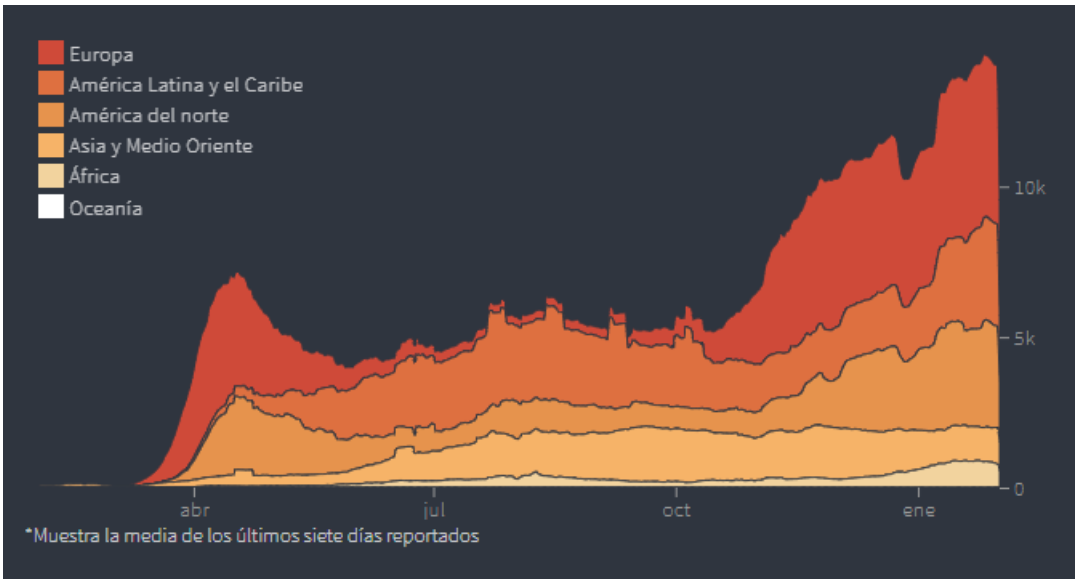


Figura 1.2: Gráfico de muertes globales desde el inicio de la pandemia.[32]

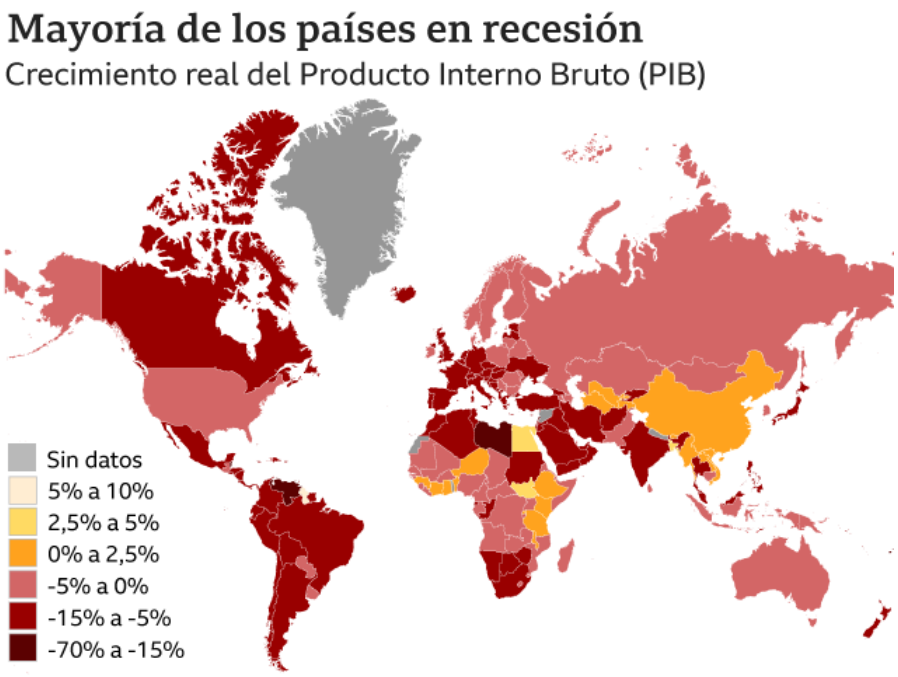


Figura 1.3: Mapa del impacto económico del COVID-19. [26]

Dadas estas premisas, todos los gobiernos aplicaron diversas medidas y restricciones intentando frenar la rápida propagación del virus, para así, liberar la presión ejercida sobre los sistemas sanitarios. El objetivo principal de estas medidas era el de conseguir la reducción de los contagios y las muertes provocadas por el COVID-19 de la forma más efectiva y rápida posible. Es por ello que muchos de los gobiernos optaron por el confinamiento domiciliario de toda su población paralizando así toda actividad no esencial. Dicha medida fue eficaz y consiguió doblegar la curva de contagios en la mayoría de los países que la aplicaron. Sin embargo, y debido a las consecuencias tanto económicas como sociales de dicha medida, esta fue considerada como alternativa final en la lucha contra las distintas nuevas olas que surgirían en el futuro. A raíz de ello, surgieron nuevas pero más leves restricciones que permitieran el descenso de la curva de contagios con un menor impacto en la población y la economía de cada país.

Dichas medidas, aunque amparadas bajo un objetivo común y una referencia global regida principalmente por la Organización Mundial de la Salud, fueron diferentes en cierto grado dependiendo del lugar, obteniendo como consecuencia mejores resultados en unas naciones que en otras. Al amparo de esta situación, se abrió una batalla no solo en el ámbito sanitario, sino en muchos otros campos de conocimiento como la estadística, la investigación farmacéutica o la Inteligencia Artificial.

1.2. Motivación

Este TFG se centra en España, uno de los países más afectados por el COVID-19 debido a diversos factores:

- **Demografía**: España es uno de los países con la media de edad más alta del mundo. Observando la gráfica de la media de edad en varios países (Figura 1.4) podemos apreciar como España se sitúa como uno de los países más envejecidos del mundo [2]. Debido a este hecho, España se ha visto más afectada en el aumento de pacientes en estado grave.

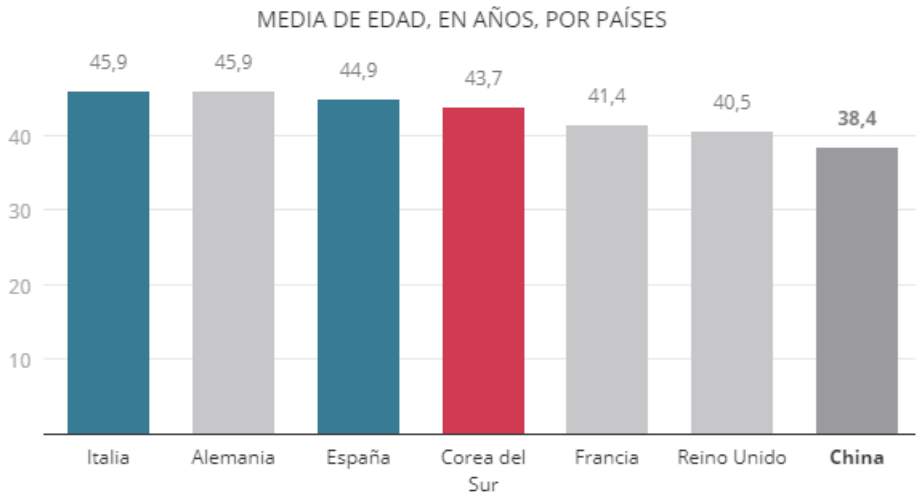


Figura 1.4: Gráfica de la media de edad en años entre países. [2]

- **Estructura social mediterránea**: Las costumbres y modo de vida en la zona mediterránea han sido el caldo de cultivo idóneo para la rápida propagación del virus, tanto la forma en la que nos relacionamos, caracterizada por el contacto físico y el carácter familiar, como la alta relación entre personas mayores y jóvenes han sido determinantes en la evolución de los contagios y muertes en comparación con otros países.

- **Economía**: El sector turístico y hostelero, el cual representa el 12 % del **PIB** de la economía española [18], ha sido el principal afectado por la pandemia debido al gran número de restricciones por movilidad entre países, la ausencia total del turismo en la zona peninsular o la poca adaptabilidad al teletrabajo. Se puede apreciar dicho impacto en la gráfica mostrada en la Figura 1.5, donde a partir del segundo trimestre de 2020 (2020T2) se observa una gran caída.

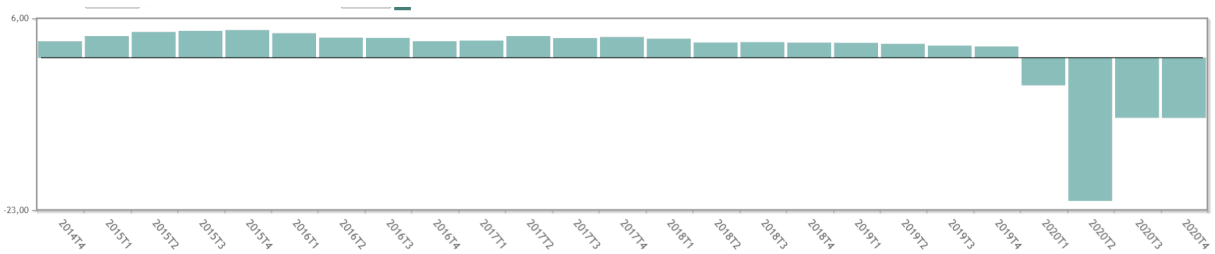


Figura 1.5: PIB de España por años y trimestres.[19]

- **Sistema sanitario:** Pese a poseer un sistema sanitario superior en calidad a la media global, este se ha sido expuesto a una presión nunca vista, lo que ha dejado ver el deterioro del mismo a lo largo de los años, especialmente en la atención primaria. Comunidades como Castilla y León han sido de las más afectadas por la gran dependencia de la población de este tipo de centros.

Es por estos factores por los que España es uno de los países que más ha tenido que maniobrar en cuanto a la aplicación de medidas, delegando la aplicación de estas en las distintas Comunidades Autónomas en las que se divide el país. Así, cada Comunidad Autónoma tiene la potestad de aplicar las medidas que considere oportunas dentro de su territorio para poder frenar el número de contagios. De esta manera, y siempre bajo una supervisión y marco definido por el Gobierno Central, se ha creado un sistema en el cual la aplicación de restricciones y medidas queda en manos de las distintas Comunidades y zonas dentro de las mismas.

Este modo de gestión y lucha contra la pandemia ha traído dificultades a la hora de aplicar restricciones efectivas de forma generalizada (en cada comunidad o territorio) debido a las diferencias entre cada una de las zonas de salud de dichas áreas; dando como resultado medidas ineficaces que han afectado tanto a la población como a la economía sin lograr resultados realmente efectivos.

A la vista de lo anteriormente expuesto, y con idea de aportar algo de luz a lo sucedido, se ha propuesto realizar un trabajo de investigación que permita determinar cuales de las medidas aplicadas han sido más eficaces en cada una de las zonas de salud afectadas de la Comunidad de Castilla y León. De este modo, en futuras olas se aplicarán en cada localización aquellas medidas más eficaces que permitan el descenso de los contagios y así evitar el colapso socio-económico que conlleva.

Para el desarrollo de esta investigación se ha elegido como caso de estudio la denominada *segunda ola* en España (que se circunscribe al periodo comprendido entre el 1 de julio y el 18 de diciembre de 2020). Se ha optado por ella en lugar de la primera debido a la disponibilidad de un marco estadístico e informativo mucho más preciso y amplio.

1.3. Objetivos y limitaciones

Dentro de este proyecto se han identificado una serie de objetivos principales, desglosados en sub-objetivos con el fin de facilitar el cumplimiento de estos:

1.3.1. Objetivos del proyecto

- **OBJ-1** Obtención y procesamiento de los datos relacionados con el COVID-19 en las zonas de salud bajo estudio.
  - **OBJ-1.1**-Estudio de los principales portales y fuentes de obtención de datos relacionados con el COVID-19 en Castilla y León.
  - **OBJ-1.2**-Recopilación y procesamiento de datos de salud relacionados con el COVID-19.
  - **OBJ-1.3**-Recopilación, búsqueda y contraste de los principales documentos oficiales e información relacionados con las medidas aplicadas.
  - **OBJ-1.4**-Obtención, transformación y procesamiento de los datos de movilidad de cada una de las zonas de salud dentro del periodo estudiado.
  - **OBJ-1.5**-Análisis de todos los datos de salud, medidas aplicadas y movilidad obtenidos.
- **OBJ-2** Búsqueda y elección de modelos y métodos de aprendizaje para la obtención de resultados
  - **OBJ-2.1**-Estudio de mejores métodos y modelos para la adquisición de los mejores resultados.
  - **OBJ-2.2**-Realización de comparativa entre los distintos modelos y métodos encontrados para determinar el más adecuado para el proyecto.
  - **OBJ-2.3-**Ajuste y mejora de los métodos y modelos, para alcanzar los resultados más precisos.
  - **OBJ-2.4-**Obtención de las mejores medidas haciendo uso de los métodos y modelos creados sobre los datos disponibles.
- **OBJ-3** Aplicación del método de análisis y modelo a una zona de test.
  - **OBJ-3.1**-Recopilación y creación de los datasets necesarios de la zona de test.
  - **OBJ-3.2**-Aplicación del método de aprendizaje a los datos de test
  - **OBJ-3.3-**Aplicación del modelo de aprendizaje a los datos de test para la obtención de las mejores medidas

1.3.2. Limitaciones

Durante el desarrollo del proyecto se han encontrado ciertas limitaciones que pasamos a enumerar:

- **Búsqueda de la información necesaria**: la búsqueda de fuentes fiables y oficiales sobre los datos de la pandemia puede ocasionar problemas de cara a la recopilación de información, debido a la dificultad de acceso a las fuentes oficiales de las zonas estudiadas o la necesidad de permisos de acceso a información protegida.

- **Veracidad de la información obtenida**: la información obtenida puede ser, en algunos casos, contradictoria o confusa al ser obtenida de diferentes fuentes gubernamentales o extraoficiales, necesitando ser contrastada para verificar su veracidad.

- **Constante actualización y cambio en los datos**: al tratarse de un proyecto que basa su creación en un tema de actualidad, los datos utilizados o modelos generados podrán necesitar de una constante actualización y revisión de cara a garantizar la mayor exactitud posible.

- **Limitaciones en cuanto a la cantidad de información disponible**: la información disponible puede verse limitada o inaccesible debido al breve periodo de tiempo en el que se han comenzado a obtener datos precisos sobre la pandemia y la generación de estos.

1.4. Estructura del proyecto

Capítulo 1. Introducción: En este primer capítulo se realizará una introducción al tema principal del proyecto y una explicación de las motivación para su desempeño junto con los objetivos principales del mismo.

Capítulo 2. Metodología: Se realizará una explicación del tipo de proceso de desarrollo del proyecto, así como las herramientas utilizadas

Capítulo 3. Gestión del proyecto: Capítulo en el cual se detallará la planificación seguida para el desarrollo del proyecto junto al presupuesto y balance del mismo.

Capítulo 4. Dominio del problema: Se realizará un estudio y análisis del entorno donde se aplicará nuestro proyecto destacando todos aquellos aspectos importantes a tener en cuenta a la hora de comenzar la construcción de nuestra herramienta. También se realizará una comparación de nuestro proyecto con herramientas y trabajos similares.

Capítulo 5. Obtención de datos: Dentro de este capítulo se abordarán todos los aspectos relacionados con la búsqueda, obtención y transformación de los datos necesarios en el proyecto. Adicionalmente se proporcionará una definición minuciosa de todas aquellas variables usadas para la correcta comprensión del trabajo realizado.

Capítulo 6. Análisis de datos: Se realizará un análisis de los datos obtenidos para la obtención de una visión global preliminar.

Capítulo 7. Métodos de aprendizaje: Clustering: Explicación teórica y práctica del método de aprendizaje clustering, junto al análisis de los resultados obtenidos.

Capítulo 8. Modelos de aprendizaje: Boosting : Explicación teórica y práctica de los modelos de aprendizaje boosting usados, junto al análisis de los resultados obtenidos.

Capítulo 9. Evaluación: Capítulo en el que se aplicará y evaluará todo el trabajo realizado, usando zonas independientes a las ya usadas.

Capítulo 10: Conclusiones y trabajo futuro: En el ultimo capítulo se expondrán aquellas conclusiones obtenidas con la realización del proyecto, junto al aprendizaje adquirido, nuevas ideas y ampliaciones futuras.

Capítulo 2

Metodología de Trabajo

En este capítulo se explicará la metodología usada para el desarrollo del proyecto. Se comenzará explicando el tipo y funcionamiento de la metodología escogida y el por qué de su elección.

2.1. Metodología SCORE

Debido al tipo de proyecto en el que nos encontramos (investigación), se ha optado por la utilización de la metodología ágil **SCORE**, la cual, esta orientada a proyectos de investigación dentro del ámbito educativo. Este tipo de metodología esta basado en **SCRUM**.

Como podemos observar, este tipo de metodología es ideal dentro de nuestro marco de trabajo, donde se hacen necesarias la flexibilidad y capacidad de adaptación para la toma de decisiones a lo largo del desarrollo del proyecto debido a la incógnita de los resultados obtenidos.

2.1.1. Origen

Debido al aumento del número de trabajos de investigación dentro del ámbito universitario y al alto número de estudiantes al cargo de cada tutor, se hacía imposible el correcto seguimiento y por tanto apoyo a cada uno de estos proyectos, por lo que el proceso de desarrollo se hacia lento y complejo. Es por ello que los profesores Michael Hicks y Jeffrey S. Foster, pertenecientes a la Universidad de Meryland, tomaron la decisión de adaptar uno de los tipos de marco de trabajo existente más eficiente (**SCRUM**) para la planificación

y seguimiento de proyectos dirigidos al sector de la investigación universitaria, naciendo así el concepto de *SCRUM for research* o ***SCORE****[14]*.

Esta nueva metodología busca el mantenimiento de la calidad y motivación en cada una de las investigaciones realizadas, llevando a los alumnos a la adquisición de una autonomía propia durante el proceso de desarrollo del proyecto con un seguimiento y apoyo adecuados.

2.1.2. Adaptación desde SCRUM

Como hemos podido ver la metodología **SCORE** nace como una adaptación de **SCRUM**, por lo que se hace necesario la correcta comprensión de esta para la explicación de la adaptabilidad de una metodología a otra.

Por otro lado, **SCRUM** consiste en un conjunto de prácticas, roles y procesos que permiten la entrega de productos de forma incremental aportando un valor añadido al producto final obtenido [11].

El principal flujo de trabajo en el que esta basado **SCRUM** para el desarrollo de software es el mostrado a continuación:

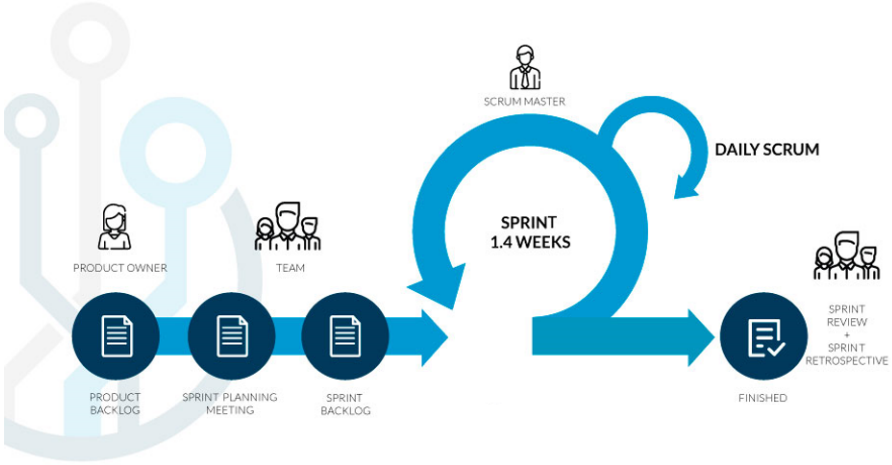


Figura 2.1: Flujo de trabajo SCRUM [36].

Como podemos ver en la Figura 2.1, el trabajo dentro de este marco de trabajo se divide entre diferentes equipos donde existe un rol denominado **scrum master** encargado del seguimiento y la correcta consecución de objetivos.

Los distintos periodos en los que se obtiene una versión funcional del producto poseen el nombre de **sprints**, siendo todos ellos de la misma duración dentro de un marco de

tiempo de 1 a 4 semanas.

Al inicio de cada sprint, estos equipos de desarrollo realizan una reunión con el nombre de **sprint planning**, donde se tiene como objetivo la identificación y comunicación del trabajo a realizar durante el sprint que está a punto de comenzar. Esta reunión se realiza en conjunto con el **product owner**, cuyo objetivo es la creación de un canal de comunicación entre los **stakeholders** y el equipo de desarrollo.

Durante el desarrollo de un sprint se realizarán reuniones diarias cortas, no superiores a 15 minutos, denominadas **daily**, en las cuales se abordará el trabajo realizado y los distintos bloqueos surgidos desde la última reunión diaria junto al trabajo a desempeñar ese día dentro del sprint. En el caso de encontrar bloqueos, será el **scrum master** el encargado de plantear soluciones para la resolución de dichos problemas o bloqueos.

Cada uno de estos **sprints** tendrá a su fin dos reuniones denominadas **sprint review** y **sprint retrospective**:

- **Sprint review**: Reunión realizada para la revisión del producto obtenido y su versión, junto a la determinación de futuras adaptaciones con la participación de los **stakeholders** involucrados en el proyecto.
- **Sprint retrospective**: Reunión realizada para la evaluación del trabajo desempeñado a lo largo del sprint. También se especificarán aquellos elementos a mejorar de cara al siguiente.

El principal atractivo detectado en **SCRUM** para la construcción de la metodología **SCORE** son esas reuniones diarias o **dailys** realizadas durante cada sprint. Estas reuniones serán adaptadas bajo el nombre de **status meetings** y **on-demand meetings** [14]:

- **Status meetings:** reuniones similares a las **dailys** pero realizadas solamente en dos o tres días a lo largo de una semana.
- **On-demand meetings:** reuniones realizadas entre alumnos y profesores para la resolución de bloqueos y problemas surgidos durante las reuniones periódicas realizadas.

2.1.3. Status meeting

Las **status meetings** o reuniones de estado, serán una de las piezas centrales de la metodología **SCORE**, análogas a las **SCRUM meetings** o **dailys**. Estas reuniones se realizarán a lo largo de la semana y solamente durante tres días fijos y no diariamente.

Esto se deberá a la limitación de horario existente entre alumnos y profesores. Dichas reuniones fijas a lo largo de la semana no excederán los 15 minutos de duración y en ellas no se deberá profundizar en detalles técnicos específicos de cada proyecto, ya que esto pertenecerá al tipo de reuniones **on demand** contempladas en esta metodología.

Durante estas reuniones todos los estudiantes tutelados por el mismo profesor describirán el trabajo realizado en cada uno de sus proyectos desde la última reunión y los obstáculos encontrados, así como el trabajo a desempeñar hasta el siguiente encuentro. Adicionalmente las **status meetings** aportan una forma de motivación hacia los alumnos a través de la participación de estos tutores revelando el trabajo desempeñado en su labor como profesores, sean parte o no de los proyectos tutelados, para fomentar la aparición de nuevas ideas o curiosidades enriquecedoras para los alumnos.

Gracias a esa interacción entre los distintos alumnos surgen aspectos positivos como la aportación de ideas de personas ajenas al proyecto y exposición del trabajo realizado al grupo, obteniendo así un *feedback* enriquecedor frente a la exposición final o el aprendizaje sobre tipos de herramientas útiles usadas en otros proyectos.

2.1.4. On-demand technical meetings

Como se ha mencionado anteriormente, en las **status meetings** no se abordarán los detalles técnicos y específicos acerca del proyecto al ser reuniones cortas y ágiles.

Es por ello por lo que dispondremos de las **on-demand technical meetings** donde se abordarán todas aquellas discusiones sobre la investigación relacionada con resultados obtenidos, modelos, técnicas.... Estas reuniones no dispondrán de un espacio fijo en el tiempo, si no que serán propuestas por el alumno según se requiera de su necesidad dentro del proyecto. Estas reuniones serán propuestas siempre y cuando el alumno haya realizado un trabajo en profundidad del tema a abordar, evitando así reuniones poco productivas en cuanto al avance del proyecto y resolución de obstáculos.

Las reuniones **on-demand** surgen como adaptación de las reuniones denominadas **spring planning** donde se establece el trabajo a realizar durante el **spring** a comenzar. Dentro de la metodología **SCORE** el objetivo de este tipo de reuniones vendrá cubierto por las reuniones bajo demanda en las cuales se detallarán las lineas de investigación a seguir y las tareas a realizar.

Este tipo de reuniones dota al proceso de desarrollo del proyecto de gran flexibilidad y adaptación a las necesidades del alumno y el tutor gracias a la realización de las mismas (únicamente cuando es necesario) favoreciendo un desarrollo fluido y una optimización del tiempo mucho más eficiente.

2.1.5. Otros elementos de SCORE

Como hemos podido observar las **status-meeting y las on-demand meetings** son la piedra angular en el funcionamiento de la metodología **SCORE**. Pero este tipo de metodología aporta otra serie de elementos aparte de estas reuniones para mejorar el proceso de desarrollo de los proyectos de investigación:

- **Lugar de desarrollo de trabajo:** Dentro de las posibilidades existentes en las instalaciones universitarias usadas, el tutor deberá proporcionar un lugar de desarrollo e investigación al alumno cercano a este, con el objetivo de crear un canal de comunicación accesible y dotarlo de las herramientas adecuadas para obtener un proyecto de calidad.
- **Interacción social: SCORE** plantea un marco de trabajo donde tanto tutor y alumno dispongan de encuentros sociales con el objetivo de favorecer y fomentar esa interacción entre ambos. Algunos ejemplos son comidas para la celebración de objetivos con el grupo de investigación en conjunto, realización de descansos para el café, etc.
- **Grupos de aprendizaje**: **SCORE** incita al tutor a la realización de grupos de aprendizaje entre los alumnos tutorizados para favorecer el intercambio de ideas u opiniones que enriquezcan los distintos proyectos en desarrollo.

2.1.6. Adaptación al proyecto

La metodología **SCORE** proporciona una forma de desarrollo de trabajo que se adapta al proyecto de investigación llevado a cabo. Sin embargo, dada la naturaleza de nuestro proyecto y la situación epidemiológica en la que nos encontramos a consecuencia del COVID-19, se hace necesaria la adaptabilidad de esta metodología para su correcta aplicación.

Debido a las recomendaciones de las autoridades sanitarias para el cuidado y protección contra la pandemia, toda aquella interacción social y presencial propuesta por la metodología **SCORE** queda eliminada en cada uno de los tipos de reuniones vistos, siendo sustituida por reuniones realizadas de forma virtual e intercambio de mensajes a través de diferentes herramientas como Microsoft Teams o Gmail.

Por esta razón, las denominadas **status meetings** quedan reducidas a este intercambio de mensajes con los tutores asignados al proyecto, pasándose a llamar **status messages** y variando esa asignación fija de tres días a la semana por una variabilidad semanal sin ninguna fijación. Todos aquellos mensajes intercambiados entre alumno y

tutores tratan dudas puntuales surgidas en el momento de desarrollo, sin ser demasiado técnicas ni extensas en explicación para la resolución de pequeñas dudas u obstáculos.

Las reuniones por demanda o **on-demand meetings** se mantienen dentro de nuestro proyecto de una manera similar a la propuesta por **SCORE** por su gran flexibilidad a la hora de ser solicitadas; manteniendo su objetivo principal en la resolución de problemas más profundos y técnicos y en el planteamiento de las tareas a realizar a futuro. Al igual que las **status meetings**, estas reuniones quedan exentas de la interacción social descrita por **SCORE** pasando a ser realizadas de manera telemática mediante reuniones virtuales con ambos tutores a través de la herramienta Microsoft Teams.

Ante la necesidad de organización y medición necesaria en cualquier proceso de desarrollo, se ha propuesto una división del proyecto por bloques de trabajo. Dichos bloques podrán interpretarse como una adaptación del concepto **sprint** del marco de trabajo **SCRUM** teniendo cada bloque un proceso de desarrollo entre 1 y 4 semanas.

Cada una de estos bloques contendrá una o más reuniones por demanda para la revisión, consecución y creación de las tareas que las componen. Por lo tanto, estas **on-demand meetings** acogerán aquellas funcionalidades vistas en SCRUM de **sprint planning**, **sprint review** y **sprint retrospective** para cada una de ellas.

Finalmente, podemos ver como esta adaptación realizada de la metodología **SCORE** nos proporciona una forma de desarrollo flexible, adaptable, con una gran comunicación dentro del entorno y situación en la que ha tenido que ser desarrollada la investigación. Todos los beneficios de **SCORE** vienen dados por las distintas formas de interacción con los tutores a través de las herramientas de **status messages** o reuniones por demanda establecidas.

SCORE también nos ha permitido resolver correctamente esa división del proyecto en bloques que nos ha aportado una mayor capacidad de gestión y visión de alcance del mismo para lograr todos los objetivos marcados.

2.2. Herramientas utilizadas

Para la construcción y elaboración del proyecto han sido utilizadas una serie de herramientas enumeradas y explicadas a continuación.

- **OpenRefine:** Herramienta gratuita para el manejo y exploración de datos desordenados que ofrece un sistema de limpieza y transformaciones de un formato a otro pudiendo ampliarse con servicios web y datos externos. Desarrollada inicialmente por Google con el nombre de Google Refine, actualmente se trata de una herramienta sostenida por la comunidad.

- **Microsoft Excel:** Hoja de cálculo desarrollada por Microsoft que cuenta con sistemas de cálculo, representación gráfica, tablas calculares y un lenguaje de programación con el nombre de Visual Basic.
- **Jupyter Notebooks:** Proyecto de código abierto para el desarrollo de software. Creado a partir de Python, proporciona un soporte de entornos de ejecución en varios tipos de lenguajes de programación, siendo los principales Julia, Python y R. Jupyter ofrece un entorno informático interactivo en la web para la creación de notebooks y documentos de tipo JSON que siguen un sistema de versionado con una lista ordenada de celdas de entrada y salida que pueden contener código, texto, gráficos matemáticos o textos enriquecidos con la extensión .ipynb.
- **Google Colab:** Servicio alojado en la nube (cloud) basado en los Notebooks de Jupyter que permite el uso gratuito de recursos como GPUs o TPUs de Google, así como las librerías Scikit-learn, Pytorch, TensorFlow, Keras y OpenCV.
- **Google Earth:** Herramienta que proporciona un conjunto completo de datos geoespaciales de manera pública, como mapas detallados de cada país, imágenes panorámicas de calles, accidentes rurales, maquetas de diferentes localizaciones, zonificaciones de distintos lugares, etc.
- **Overleaf:** Sitio web que permite la creación de documentos en LaTeX posibilitando la compilación del código LaTeX de manera automática, generando los resultados de manera casi simultanea. Proporciona un gran número de plantillas para la creación de documentos.
- **Trello:** Herramienta de administración que emplea el sistema **Kanban** para el registro de actividades con tarjetas virtuales, permitiendo agregar listas, adjuntar archivos, etiquetar eventos, agregar comentarios y compartir tableros.
- **Microsoft Teams:** Plataforma de comunicación que provee de un espacio de trabajo mediante el uso de chats, reuniones por vídeo y almacenamiento e integración de aplicaciones.

2.3. Tecnologías utilizadas

Las distintas tecnologías usadas como lenguajes de programación o bibliotecas se enumeran a continuación:

- **Bibliotecas python:** El lenguaje usado para la construcción de los modelos y métodos del proyecto ha sido Python, que nos ofrece un gran número de bibliotecas para el manejo y representación de datos, así como bibliotecas necesarias para la creación de los modelos implementados:

- **Numpy:** Biblioteca usada para dar soporte en la creación de vectores y matrices, proporcionando una gran colección de funciones matemáticas de alto nivel.
- **Pandas:** Biblioteca utilizada para el manejo y análisis de estructuras de datos, permitiendo la fácil lectura y escritura de ficheros en diferentes formatos, acceso a los datos mediante índices o filas y columnas, métodos para la reordenación, división y combinación de conjuntos de datos.
- **Matplotlib:** Biblioteca usada para la generación de gráficos a partir de los datos disponibles en diferentes formatos. Hace uso de la extensión matemática **NumPy**.
- **Seaborn:** Biblioteca que permite la generación de gráficos. Basada en **Matplotlib**, proporciona una interfaz de alto nivel sencilla y fácil de usar.
- **Sklearn:** Biblioteca de software libre usada para aprendizaje automático. Incluye algoritmos de clasificación, regresión y análisis de grupo entre los cuales podemos encontrar Kmeans, Gradient Boosting, DBS o máquinas de vectores entre otros. Está diseñada para poder trabajar con las bibliotecas NumPy y SciPy.
- **Pickle:** Biblioteca usada para la exportación e importación de modelos de aprendizaje.

Capítulo 3

Gestión del proyecto

En este capítulo se presenta la planificación y estimación del proyecto dentro del tiempo recomendado para la realización del Trabajo de Fin de Grado, el cual no debe exceder las 300 horas.

Tal y como se ha mostrado en el Capítulo 2, se ha hecho uso del marco de trabajo **SCORE** para el desarrollo del proyecto, dividiéndose este en distintos bloques de trabajo de diferente duración cada una.

Cabe destacar, que al ser este un proyecto de investigación en el cual el desarrollo y logro de los objetivos finales es variante y difuso, la labor de estimación de un tiempo preciso de acuerdo al tiempo real de desarrollo del proyecto se ha hecho complejo.

A continuación veremos en detalle la estimación del esfuerzo realizada para cada una de las tareas que componen el proyecto y su planificación en el tiempo. Además se mostrará la forma de estimación del presupuesto económico del proyecto junto a un balance final en el que se han incluido los distintos bloqueos u obstáculos encontrados a la hora de llevar a cabo el desarrollo del mismo y que han alterado la estimación realizada.

3.1. Estimación del esfuerzo

Uno de los requisitos previos a realizar ante una planificación temporal es la estimación del esfuerzo y tiempo de las tareas que componen nuestro proyecto. En este proyecto se ha utilizado el marco de trabajo **SCORE**, usando esta última las denominadas *historias de usuario* para la correcta estimación del esfuerzo a través de los puntos de historia.

Para una correcta comprensión de la adaptación realizada entre las tareas y las his-

torias de usuario, los puntos de historia pasarán a llamarse Punto de Tarea (PT), que indicarán mediante un valor numérico el tiempo necesario para la realización de cada tarea.

La asignación de los PT a cada una de las tareas se ha realizado en consenso entre el alumno y los tutores encargados de guiar el proyecto.

Las distintas tareas que han compuesto el proyecto, así como los puntos de tareas de cada una de ellas, han sido:

Nombre tarea	**Puntos de tarea (PT)**
TAREA 0.1-BÚSQUEDA Y RECOPILACIÓN PORTALES Y PLATAFORMAS DE DATOS	2 PT
TAREA 0.2-ANÁLISIS DE VARIABLES ESTADÍSTICAS MÁS RELEVANTES	1 PT
TAREA 0.3-BÚSQUEDA DE TRABAJOS SIMILARES	1 PT
TAREA 0.4-INSTALACIÓN Y APRENDIZAJE DE HERRAMIENTAS NECESARIAS	2 PT
TAREA 0.5-OBTENCIÓN DATOS DE SALUD	1 PT
TAREA 0.6-OBTENCIÓN DATOS MEDIDAS	1 PT
TAREA 0.7- CREACIÓN DE DATASETS Y VARIABLES	5 PT

Tabla 3.1: Tarea 0-Preparación y obtención de la información

Nombre tarea	**Puntos de tarea (PT)**
TAREA 1.1-BÚSQUEDA Y COMPARACIÓN MÉTODO DE ANÁLISIS DE DATOS	3 PT
TAREA 1.2-CONSTRUCCIÓN MÉTODO DE CLUSTERING	3 PT
TAREA 1.3-ANÁLISIS DE RESULTADOS OBTENIDOS CON MÉTODO CLUSTERING	4 PT

Tabla 3.2: Tarea 1-Búsqueda y aplicación de métodos de análisis

Nombre tarea	**Puntos de tarea (PT)**
TAREA 2.1-BÚSQUEDA DE INFORMACIÓN DE MOVILIDAD DISPONIBLE	1 PT
TAREA 2.2-OBTENCIÓN DE INFORMACIÓN SOBRE MOVILIDAD	5 PT
TAREA 2.3-ANÁLISIS Y PROCESAMIENTO DE datasets OBTENIDOS	3 PT

Tabla 3.3: Tarea 2-Obtención y procesamiento datos movilidad

Nombre tarea	**Puntos de tarea (PT)**
TAREA 3.1-CREACIÓN DATASET MEDIDAS POR FECHA	1 PT
TAREA 3.2-BÚSQUEDA Y COMPARACIÓN DE MEJORES MODELOS DE APRENDIZAJE	2 PT
TAREA 3.3-CREACIÓN DE MODELOS	4 PT
TAREA 3.4-COMPARACIÓN DE RESULTADOS OBTENIDOS CON MODELOS	1 PT
TAREA 3.5-CREACIÓN DE MODELO ELEGIDO	4 PT

Tabla 3.4: Tarea 3-Creación de modelos de aprendizaje

Nombre tarea	**Puntos de tarea (PT)**
TAREA 4.1-CREACIÓN DATASETS ZONA DE EVALUACIÓN	2 PT
TAREA 4.2-OBTENCIÓN DATOS MOVILIDAD ZONA DE EVALUACIÓN	2 PT
TAREA 4.3-APLICACIÓN DE MÉTODO CLUSTERING	1 PT
TAREA 4.4-APLICACIÓN MODELOS OBTENIDOS DE ZONAS ANTERIORES.	2 PT
TAREA 4.5-EVALUACIÓN DE RESULTADOS OBTENIDOS	2 PT

Tabla 3.5: Tarea 4-Evaluación

Nombre tarea	**Puntos de tarea (PT)**
TAREA 5.1-CREACIÓN APARTADO DESCRIPCIÓN DEL PROYECTO	2 PT
TAREA 5.2-DOCUMENTACIÓN METODOLOGÍA DE TRABAJO Y GESTIÓN DEL PROYECTO	2 PT
TAREA 5.3-DOCUMENTACIÓN DOMINIO DEL PROBLEMA	2 PT
TAREA 5.4-DOCUMENTACIÓN DE OBTENCIÓN Y ANÁLISIS DE DATOS	4 PT
TAREA 5.5-DOCUMENTACIÓN MÉTODOS Y MODELOS USADOS	2 PT
TAREA 5.6-DOCUMENTACIÓN DE FASE DE EVALUACIÓN	2 PT
TAREA 5.7-DOCUMENTACIÓN CONCLUSIÓN, REFERENCIAS Y BIBLIOGRAFÍA	1 PT
TAREA 5.8-REVISIONES DE MEMORIA	2 PT

Tabla 3.6: Tarea 5-Documentación

Se obtiene finalmente una suma total de 70 Puntos de Tarea (PT).

3.2. Planificación temporal

Nuestro proyecto tal y como hemos podido ver ha sido divido en bloques de trabajo (adaptación del concepto sprint visto en **SCRUM**) cuya duración en el tiempo se establece de 1 a 4 semanas, teniendo en nuestro caso cada bloque una duración fija de 2 semanas. El conjunto de estos bloques y su desarrollo compone una visión global de toda la planificación llevada a cabo para realizar el proyecto al completo en el tiempo de duración estipulado de 300 horas de acuerdo a la carga de créditos ETCS del mismo. El número de horas diarias dedicadas al proyecto será de 3 horas aproximadamente, exceptuando fines de semana y días festivos.

A continuación se muestra la estimación de la duración de cada uno de estos bloques, correspondiendo únicamente a aquellos días en los que se ha trabajado en el proyecto junto a las reuniones bajo demanda realizadas en cada una de ellas:

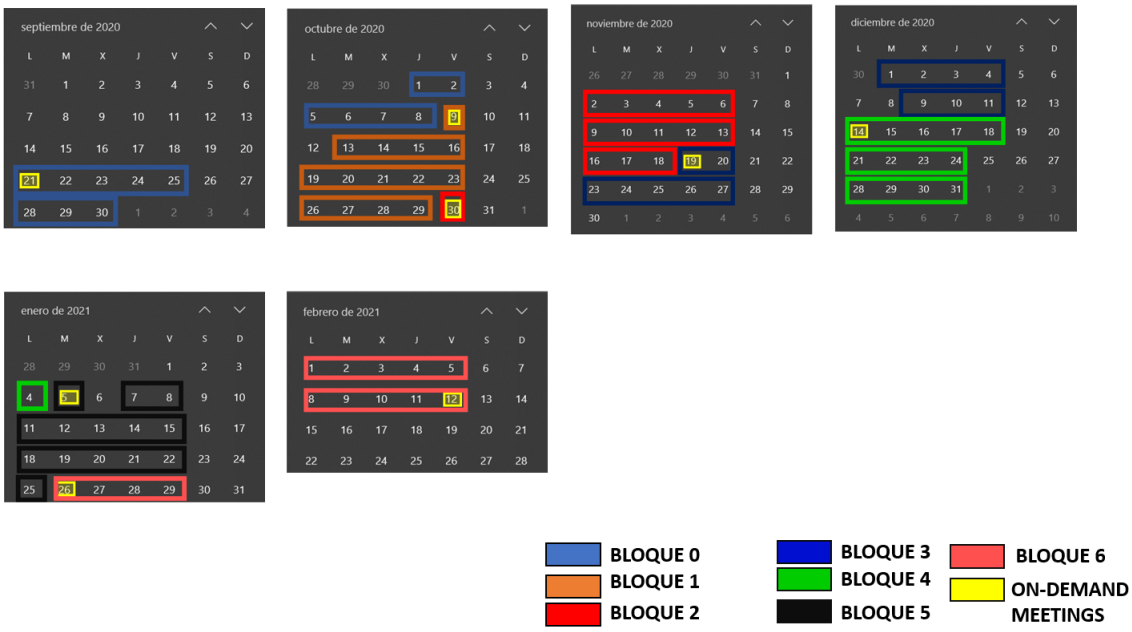


Figura 3.1: Calendario bloques del proyecto

Tal y como podemos ver en la Figura 3.1, los bloques en los que se ha dividido el proyecto poseen un tiempo de duración idéntico entre ellos. Esto se debe al reparto de las tareas del proyecto en cada una de los bloques, obteniendo así unos periodos de trabajo iguales junto a una estimación de las reuniones bajo demanda que serán realizadas para la revisión y propuesta de tareas al comienzo de esos periodos.

La tareas desarrolladas en cada uno de los bloques se muestran a continuación:

Bloque	**Periodo (DD-MM-YYYY)**	**Tareas**	**Total PT**
BLOQUE 0	21-09-2020 08-10-2020	0.1, 0.2, 0.3, 0.4, 5,1, 5.2	10 PT
BLOQUE 1	09-10-2020 29-10-2020	0.5, 0.6, 0.7 , 1.1	10 PT
BLOQUE 2	30-10-2020 18-11-2020	1.2, 1.3, 2.1, 5.3	10 PT
BLOQUE 3	19-11-2020 11-12-2020	2.2, 2.3, 3.1, 5.7	10 PT
BLOQUE 4	14-12-2020 04-01-2021	3.3, 3.4, 3.5, 4.3	10 PT
BLOQUE 5	05-01-2021 25-01-2021	4.1, 4.2, 4.4, 4.5, 5.6	10 PT
BLOQUE 6	26-01-2021 12-02-2021	3.2 ,5.4, 5.5, 5.8	10 PT
TOTAL	————	————	70 PT

Tabla 3.7: Tareas por Bloques

Como conclusión, a través del número de días de duración del proyecto y el valor total de puntos de tarea, podemos obtener el tiempo estimado en el proyecto y la equivalencia de ese tiempo en cada punto de tarea asignado:

Días totales empleados en proyecto	**Equivalencia en horas**
100 DÍAS	300 HORAS
Número total de Puntos de Tarea (PT)	**Horas aproximadas por punto de tareas (PT)**
70 PT	4.28 HORAS (4 HORAS Y 28 MINUTOS)

Tabla 3.8: Cálculo del tiempo total

BLOQUE	**Horas empleadas**
BLOQUE 0	42.8 HORAS
BLOQUE 1	42.8 HORAS
BLOQUE 2	42.8 HORAS
BLOQUE 3	42.8 HORAS
BLOQUE 4	42.8 HORAS
BLOQUE 5	42.8 HORAS
BLOQUE 6	42.8 HORAS
TOTAL	**300 HORAS**

Tabla 3.9: Tiempo estimado de cada bloque

3.2.1. Trello

Se ha usado la herramienta Trello como parte de la gestión y administración de las tareas del proyecto. Dicha herramienta basada en un tablero **Kanban** permite gestionar mediante el uso de columnas las diferentes tareas creadas, teniendo los tutores del proyecto acceso a dicho tablero en todo momento.

Las distintas columnas de las que partimos se basarán en el concepto de *Estado de tarea* dado por los tableros **Kanban**:

- **TO DO**: Columna correspondiente a aquellas tareas pendientes de hacer. Esta columna estará situada a la cabeza debido a su importancia y relevancia a la hora de alcanzar nuestros objetivos.
- **DOING**: Columna correspondiente a aquellas tareas en proceso.
- **DONE**: Columna correspondiente a aquellas tareas completadas.

En el caso del estado **DONE**, este se repartirá en las columnas correspondientes a los 8 bloques de trabajo en los que esta dividido nuestro proyecto y a la columna denominada **ON DEMAND MEETINGS**, donde se incluyen todas las reuniones realizadas, además de los informes y dudas resueltas en cada una

BLOCKED: Columna usada para establecer aquellas tareas que se encuentran bloqueadas, ya sea por requisitos previos o por algún imprevisto. Esta columna se presenta vacía, ya que las tareas situadas en ella permanecen de manera temporal hasta que se resuelva el bloqueo.

Cada una de las tareas creadas tendrá asignado un color dependiendo del estado en el que se encuentre, asociando el rojo a aquellas tareas aun pendientes (**TO DO**), amarillo a aquellas en proceso (**DOING**) y verde a aquellas ya completadas y distribuidas entre las columnas de las Fases y Reuniones (**DONE**). En el caso de las tareas **BLOCKED** el color será el Azul Oscuro.

Dentro de cada tarea encontraremos una *checklist* que nos ofrece una serie de subtareas en las cuales estará dividida la propia tarea. En algunas de las tareas estarán incluidos documentos, enlaces o descripciones de utilidad junto a la fecha estimada de finalización de acuerdo a cada bloque.

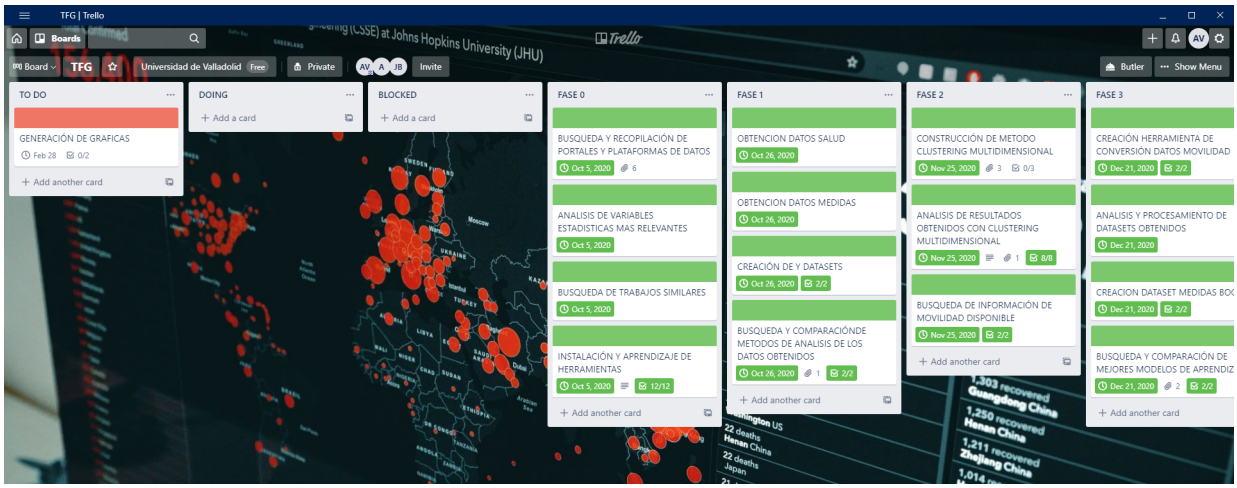


Figura 3.2: Tablero trello usado en el proyecto

3.3. Presupuesto

Dentro de un proyecto de investigación como este, se hace compleja la estimación de los costes que va a suponer debido al desconocimiento del número de tecnologías o herramientas necesarias para su desarrollo. Como ya hemos visto en apartados anteriores, la estimación tanto del tiempo como del esfuerzo se ha realizado de una manera muy relativa debido a la incógnita existente de los resultados que se obtendrán finalmente. Con estas premisas se intentará dar una estimación del coste de nuestro proyecto dividiéndose esta en los distintos importes existentes dentro del mismo.

3.3.1. Hardware y Software

Hardware:

Componente	**Uso (%)**	**Coste (€) por mes (6 meses)**	**Coste (€) total**
Ordenador de Trabajo	50 %	150€	900€
Internet	50 %	30€	180€

Tabla 3.10: Presupuesto Hardware

Software:

Herramienta	**Coste (€) por mes (6 meses)**	**Coste (€) total**
OpenRefine	0€	0€
Microsoft Excel (Student License)	0€	0€
Jupyter Notebooks	0€	0€
Google Colab	0€	0€
Trello	0€	0€
Microsoft Teams	0€	0€
TOTAL	**0€**	**0€**

Tabla 3.11: Presupuesto Software

3.3.2. Recursos humanos

Para el cálculo de costes de recursos humanos se usará como referencia el sueldo medio anual de un analista de datos Junior, rol con un trabajo similar al realizado en este proyecto. Este salario anual será dividido por el número de horas laborales aproximado existentes en un año para el cálculo de salario por horas.

Coste medio:

Rol de trabajo	**Sueldo Bruto anual (€)**	**Sueldo Bruto por horas (€)**
Analista de datos Junior	25000€	14.17€

Tabla 3.12: Coste rol de trabajo [12]

Coste medio en proyecto actual:

Rol de trabajo	**Coste por hora**	**Coste total (300 horas) (€)**
Analista de datos Junior	14.17€	4251.7€

Tabla 3.13: Presupuesto recursos humanos

3.3.3. Presupuesto total

Finalmente se obtiene el presupuesto total siendo este la suma total de los costes de hardware, software y recursos humanos:

Presupuesto	**Coste (€)**
Hardware	1080€
Software	0€
Recursos humanos	4251.7€
TOTAL	**5331.7€**

Tabla 3.14: Presupuesto Total

3.4. Balance

Finalizado el proyecto se ha realizado un balance del desarrollo realizado, destacando los cambios y distintos obstáculos que se han encontrado respecto a la planificación hecha. Dichos imprevistos han modificado la extensión de los bloques de trabajo en los que se divide el proyecto, quedando estos desiguales entre si (permitido dentro de la metodología SCORE).

Debido a la extensión de las tareas de documentación y revisión de la memoria (5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7 y 5.8) junto con un cambio en el caso de estudio del proyecto, el desarrollo del mismo se alargó mucho más de lo previsto adquiriendo estas tareas una carga de trabajo y tiempo muy superior al estimado causando un impacto en la duración de todos los bloques de trabajo del proyecto. Adicionalmente, en el periodo de tiempo comprendido entre el 22 de Marzo y 8 de Abril se paró el desarrollo del proyecto debido al proceso de revisión de la memoria por parte de los tutores y a las vacaciones de Semana Santa.

Durante el **Bloque 2** se tuvo que detener el desarrollo del proyecto durante dos días debido a causas personales, retomándose el flujo de trabajo pocos días después. Esta pequeña pausa provocó un alargamiento del **Bloque 3** del proyecto haciendo que los días festivos 7 y 8 de diciembre fueran tomados como días de trabajo.

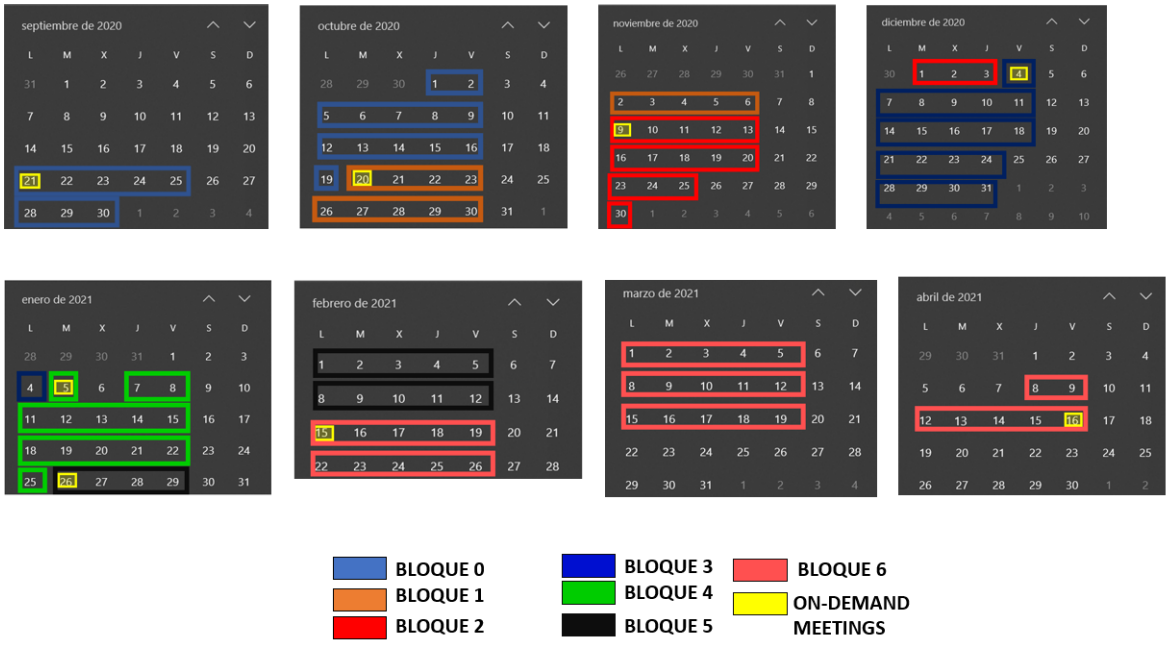


Figura 3.3: Planificación temporal balance

BLOQUE	**Periodo (DD-MM-YYYY)**	**Tareas**	**Total PT**
BLOQUE 0	21-09-2020 19-10-2020	0.1, 0.2, 0.3, 0.4, 5,1, 5.2	14 PT
BLOQUE 1	20-10-2020 06-11-2020	0.5, 0.6, 0.7 , 1.1	10 PT
BLOQUE 2	09-11-2020 03-12-2020	1.2, 1.3, 2.1, 5.3	12 PT
BLOQUE 3	04-12-2020 04-01-2021	2.2, 2.3, 3.1, 5.7	14 PT
BLOQUE 4	05-01-2021 25-01-2021	3.3, 3.4, 3.5, 4.3	10 PT
BLOQUE 5	26-01-2021 12-02-2021	4.1, 4.2, 4.4, 4.5, 5.6	12 PT
BLOQUE 6	15-02-2021 16-04-2021	3.2, 5.4, 5.5, 5.8	23 PT
TOTAL	————	————	95 PT

Tabla 3.15: Balance tareas por bloques

Por tanto, debido a estos imprevistos, el planteamiento temporal quedará extendido en 35 días más de lo estimado, contando el proyecto con una duración de 406,6 horas. Podemos observar cierta desviación respecto a la planificación inicial realizada, posiblemente debido a la complejidad de estimar en un principio la duración de un trabajo de investigación como este.

Días totales empleados en proyecto	**Equivalencia en horas**
135 DÍAS	406.6 HORAS
Número total de Puntos de Tarea (PT)	**Horas aproximadas por punto de tareas (PT)**
95 PT	4.28 HORAS (4 HORAS Y 28 MINUTOS)

Tabla 3.16: Cálculo del tiempo total

BLOQUES	**Horas empleadas**
BLOQUE 0	59.92 HORAS
BLOQUE 1	42.8 HORAS
BLOQUE 2	51.36 HORAS
BLOQUE 3	59.92 HORAS
BLOQUE 4	42.8 HORAS
BLOQUE 5	51.36 HORAS
BLOQUE 6	98.44 HORAS
TOTAL	**406.06 HORAS**

Tabla 3.17: Tiempo final de cada bloque

Capítulo 4

Dominio del problema

4.1. Datos sobre el COVID-19

Han sido muchas y numerosas las distintas formas de lucha contra la COVID-19. Una de las más significativas ha sido aquella relacionada con la obtención y análisis de datos para la predicción y aplicación de las distintas restricciones, que permiten reducir el número de contagios y muertes provocadas por el virus. Podemos dividir los datos usados en distintos tipos, salud (indicadores específicos) y movilidad.

4.1.1. Datos de salud sobre el COVID-19

Dentro de los datos de salud encontramos toda aquella información relacionada con el número de contagios detectados, pruebas realizadas, muertes provocadas por el virus y prevalencia del mismo en la población. Desde el inicio de la pandemia se han ido mejorando las técnicas de recopilación, ampliando el rango de estudio del virus, con el objetivo de conocer mejor su comportamiento y efectos.

Indicadores sobre el COVID-19

A partir de esos datos de salud ya mencionados, se han ido creando indicadores de seguimiento y formas de medición del riesgo e impacto que permitieran en cada uno de los territorios aplicar medidas en consecuencia para la reducción de la curva de contagios.

Entre los principales **indicadores y mediciones** [8] usados están:

- **Tasa de Incidencia:** representación de la aparición de nuevos casos en la población de riesgo. Calculada mediante la división de los casos nuevos por COVID-19 entre la población expuesta, siempre teniendo en cuenta el tiempo de seguimiento de dicha población
- **Incidencia Acumulada:** número de personas que contraen el virus en un periodo de tiempo concreto. Obtenida mediante la división del número de casos aparecidos en un periodo, entre el número de individuos libres de la enfermedad al inicio de dicho periodo. Los periodos comúnmente usados son los de 7 y 14 días.
- **Prevalencia:** proporción de la población que padece el virus. Proporciona un valor estático que refleja la magnitud de un problema en un momento concreto.
- **Mortalidad:** medida con la que se representa la muerte en una población en un momento concreto. Pueden ser usadas distintas medidas de mortalidad: mortalidad general (volumen de muertes ocurridas por todas las causas de enfermedad, en todos los grupos de edad y para ambos sexos), la mortalidad específica (mortalidad de un grupo específico de la población) o la letalidad (proporción de casos de COVID-19 que resultan mortales).

4.1.2. Datos de movilidad

La recopilación de datos de movilidad a lo largo de la pandemia ha adquirido una gran relevancia debido al efecto de esta en el aumento de contagios de cada territorio. Proyectos como el Estudio de Movilidad con Big Data [38], impulsado por el Ministerio de Transportes, Movilidad y Agenda Urbana del Gobierno de España, han permitido recopilar datos sobre las entradas y salidas de las diferentes provincias y zonas de salud que componen el país, para su posterior estudio y análisis.

Las distintas técnicas de obtención de datos unidas a herramientas tan potentes como Big Data, el aprendizaje automático o el aprendizaje profundo, han permitido crear un marco estadístico en el cual se ofrecen modelos y predicciones muy precisas y rápidas. Estas han hecho posible que el conocimiento acerca del virus haya crecido a lo largo del tiempo, permitiendo así anticiparse y aplicar medidas cada vez más efectivas para su eliminación.

4.2. Medidas contra el COVID-19

Una de las armas principales contra la propagación de la pandemia ha sido la aplicación de restricciones en la población de cada lugar, siempre aplicadas basándose en ese marco estadístico acerca de la pandemia proporcionado por los datos.

Entre las técnicas aplicadas en cada territorio encontramos el aumento de PCR realizadas de cara a una detección y aislamiento rápido de personas infectadas. Otras medidas a destacar son la reducción de la interacción social y restricción de la movilidad entre zonas afectadas, siendo factores claves frente al rápido aumento de contagios.

La ausencia o poco efectividad de muchas de las medidas aplicadas en cada zona, se ha podido deber a la particularidad de cada territorio en aspectos como la movilidad, la economía, la sociedad, el sistema sanitario...

Es por ello que en aquellos territorios en los que la aplicación de cada medida se realiza de una forma acorde a los datos (no sólo epidemiológicos de dicho territorio sino a datos como la movilidad o la actividad económica de cada zona) se aprecia una eficacia superior, observando periodos de vigencia de cada restricción menores en el tiempo y más efectivos.

A continuación se muestra en la Figura 4.1, un análisis realizado en Reino Unido donde se hizo un estudio mediante diferentes escenarios de los efectos de las medidas aplicadas en la curva de contagios de dicho país.

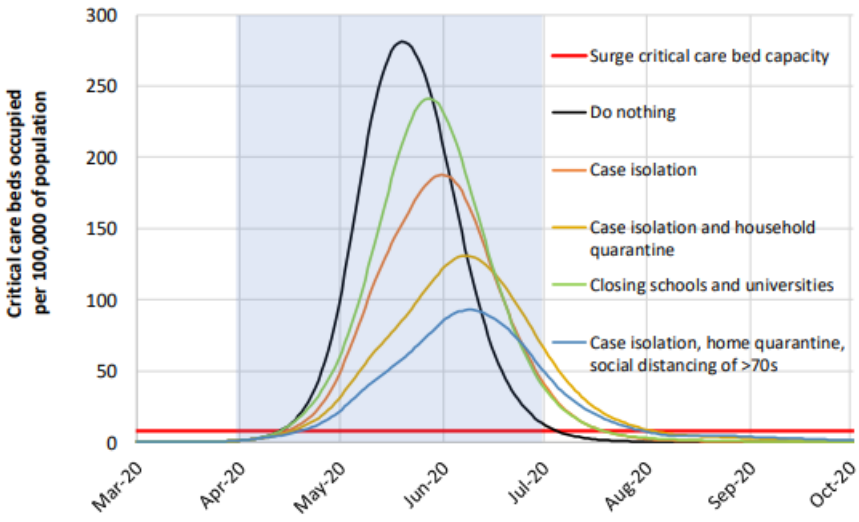


Figura 4.1: Escenarios de estrategia de mitigación para Reino Unido.[37]

4.3. Trabajos similares

4.3.1. Herramienta de predicción picos COVID-19

Un trabajo en la linea del desarrollado en este proyecto ha sido la creación de una herramienta que permite la predicción de picos futuros de contagios en función de los datos disponibles sobre la pandemia [21]. La herramienta desarrollada por investigadores de la Universidad del Egeo (Turquía) es capaz de describir con gran exactitud datos epidemiológicos como el número de contagiados y muertes por COVID-19, así como la capacidad de predicción de los posibles picos de contagios en las zonas aplicadas. Los parámetros utilizados para la creación del algoritmo usado por dicha herramienta fueron establecidos a partir de los datos públicos de la pandemia en China (datos relacionados con la biotecnología y el uso de distribuciones de Weibull). Tras su construcción fue probada en países como Francia, Brasil, Italia o Reino Unido, generando predicciones muy similares a los datos obtenidos *a posteriori*, tal y como se puede apreciar en la Figura 4.2, donde se muestran las distintas predicciones obtenidas para cada uno de estos países.

Analizando las diferencias entre la herramienta de predicción de picos y nuestro proyecto, vemos como en dicha herramienta el objetivo principal es la predicción de futuros picos de contagios, estando nuestro proyecto más centrado en el análisis y obtención de aquellas medidas más eficaces contra el COVID-19 en cada zona estudiada. Otra de las diferencias encontradas respecto a nuestro proyecto es el uso de un gran componente matemático para la obtención de predicciones y datos precisos. También podemos observar cómo los datos disponibles para la construcción de la herramienta son mucho más genéricos que los usados en nuestro proyecto, por lo que se obtiene un margen mucho mayor para el entrenamiento y por tanto, la generación de unos resultados concisos.

En cuanto a la semejanza de nuestro proyecto con la herramienta vista, podemos ver como los datos utilizados para su implementación son afines a los empleados en nuestro caso, exceptuando la cantidad y disposición en la fase de evaluación del análisis y predicciones realizadas.

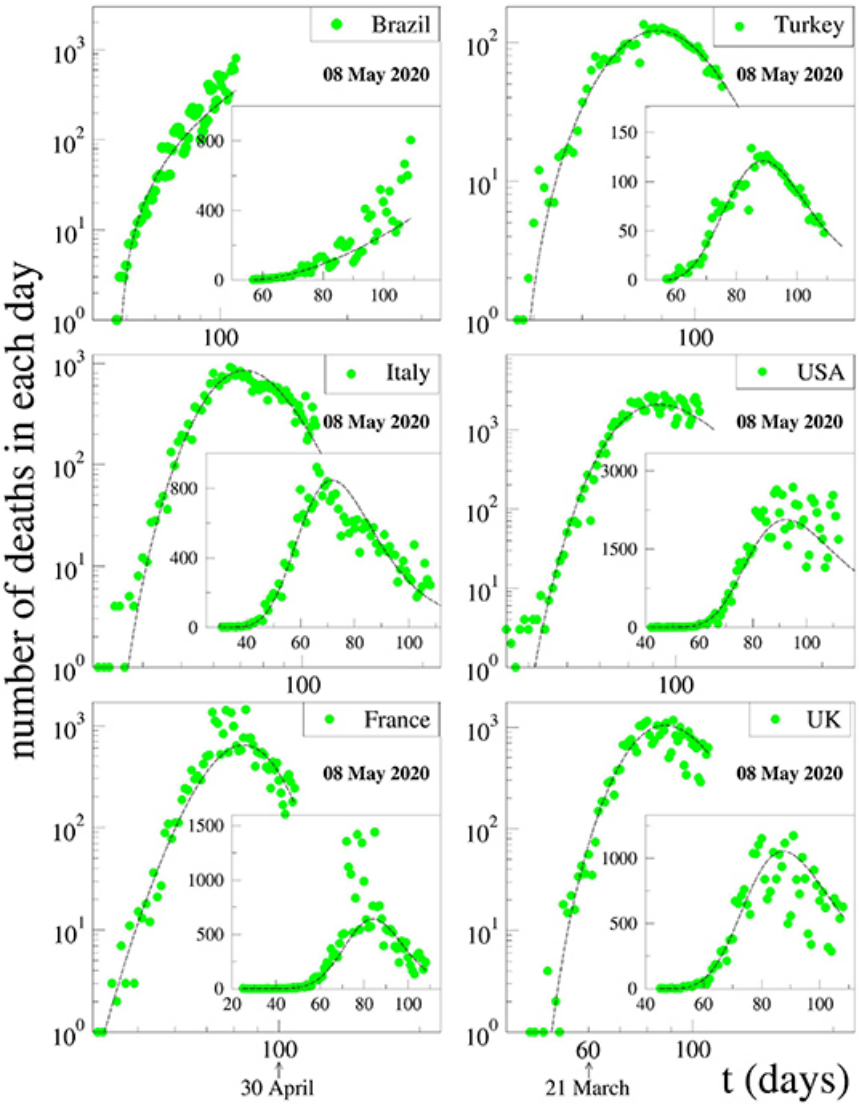


Figura 4.2: Predicciones de picos obtenidas por la herramienta en zonas de prueba. [21]

4.3.2. Algoritmo de predicción de propagación de enfermedades.

Otro de los proyectos similares al desarrollado es la herramienta BlueDot, que a través de un algoritmo impulsado mediante Inteligencia Artificial es capaz de predecir futuras enfermedades que contengan una alta probabilidad de propagación [29]. La empresa que da nombre a dicha herramienta fue creada por Kamran Khan en 2014, cuya motivación fueron los sucesos ocurridos en 2003 durante la epidemia de **SARS** (*Severe Acute Respiratory Syndrome*), lo que le llevó a la búsqueda de una forma de rastreo de enfermedades eficaz. La empresa cuenta en la actualidad con una plantilla compuesta por médicos y pro-

gramadores encargados de dar soporte a Bluedot a través de distintas tecnologías dentro del mundo de la Inteligencia Artificial.

BlueDot basa su funcionamiento en el rastreo de noticias en la red obteniendo información en plataformas de comunicación, blogs e incluso foros. Además, dispone de distintos mecanismos para la obtención de información, como procesadores del lenguaje natural usados en la traducción de toda aquella información obtenida sobre noticias en cada país, así como técnicas de aprendizaje automático para el análisis de dicha información.

La herramienta también emplea datos de movilidad como por ejemplo el acceso a información de billetes de avión y análisis de informes epidemiológicos de cada país. De esta manera, extrae una alta cantidad de datos de distintas fuentes, lo que le permite realizar predicciones incluso en aquellos países en los que los brotes de enfermedades son ocultados.

Los resultados obtenidos por BlueDot necesitan ser analizados por comités de expertos para comprobar su veracidad (debido al margen de error existente en la herramienta) al basar su funcionamiento principalmente en fuentes no fiables como pueden ser los de las noticias obtenidas en la red.

Contrastando la herramienta BlueDot con nuestro proyecto podemos apreciar el uso de esa información obtenida a partir de las noticias para su entrenamiento y posterior validación. En nuestro caso, la mayoría de la información relacionada con las medidas aplicadas es obtenida de distintos documentos, noticias y portales de prensa, siendo posteriormente contrastada y validada para su uso en los diferentes modelos. Sin embargo, podemos ver que la herramienta tiene un ámbito de actuación más allá de la COVID-19, a diferencia de nuestro proyecto, cuyo objeto de estudio es únicamente dicha pandemia.

4.3.3. Clustering para mitigar el impacto del COVID-19 en Malasia

Un trabajo muy semejante al desarrollado en este TFG es el proyecto de clustering basado en datos dinámicos para la mitigación del impacto económico provocado por el COVID-19 en Malasia [27]. La creación de este trabajo ha sido fruto de una colaboración entre las facultades de computación e Inteligencia Artificial de las Universidades de Malaysia Pahang, Birmingham City, Near East University y Fordham University.

En él se ha hecho uso del tipo de aprendizaje no supervisado clustering para la reducción del impacto económico/social provocado por el COVID-19 en Malasia. Así, se ha creado un algoritmo dinámico de clustering que, a través de una fusión inteligente de los datos diarios de salud y movilidad (simulada), realiza de forma dinámica agrupaciones (clusters) de zonas a confinar (alta incidencia detectada) y zonas fuera de peligro a las

que no se les debería aplicar confinamiento alguno.

Como se puede observar, el trabajo descrito guarda grandes similitudes con las herramientas y métodos usados en nuestro proyecto, siendo el objetivo principal de ambos la mitigación del impacto socioeconómico provocado por el COVID-19 haciendo uso del modelo de aprendizaje no supervisado clustering sobre datos de salud.

En cuanto a las diferencias entre ambos trabajos, podemos destacar como nuestro proyecto hace uso de datos más específicos, junto a unos datos de movilidad reales y no simulados. Respecto al clustering, el trabajo descrito anteriormente describe un flujo dinámico de obtención de resultados (clusters), mientras que en nuestro proyecto esa obtención es estática, ya que se realiza sobre datos pasados, obteniendo resultados que son usados posteriormente como entrada a otro modelo de aprendizaje (Gradient Boosting).

4.4. Comparativa trabajos similares con el proyecto en desarrollo

En la Tabla 4.1 se resume la comparación realizada entre los trabajos similares descritos y nuestro proyecto destacando similitudes y diferencias.

Herramienta/ Proyecto	**Similitudes**	**Diferencias**
Predicción de picos.	Uso de datos de salud.	Tiene como objetivo la predicción y no el análisis y obtención de medidas efectivas. Gran componente matemático. Gran volumen de datos usado.
Algoritmo de predicción de enfermedades BlueDot.	Uso de noticias para la obtención de información.	Marco de estudio mucho más amplio (otras pandemias y enfermedades).
Clustering para mitigar el impacto del COVID-19 sobre Malasia.	Uso de clustering para reducción de impacto del COVID-19. Uso de datos de salud y movilidad.	Marco de estudio más general (país). Uso de datos de movilidad simulados. Uso de un único método/modelo de aprendizaje.

Tabla 4.1: Tabla comparativa trabajos similares vs proyecto actual

Capítulo 5

Obtención y tratamiento de datos

5.1. Introducción

A lo largo de toda la pandemia han sido numerosos los cambios que se han realizado en la forma de recopilar los datos. Las distintas técnicas usadas de recopilación y procesamiento han ido cambiando a medida que se desarrollaba la pandemia. Es por ello por lo que nos hemos situado en el marco temporal de la *segunda ola* de España (cuyo inicio se sitúa pocas semanas después de la primera), debido a un mejor conocimiento del virus en esta segunda , y donde las técnicas de recopilación y registro han sido mejoradas para la obtención de datos más precisos y amplios, favoreciendo así la anticipación y la aplicación de medidas.

A lo largo de este capítulo nos centraremos en el origen, obtención, transformación y procesamiento de todos los datos correspondientes a la zonas de salud que han sido escogidas como caso de estudio.

5.2. Caso de estudio

Nos hemos enfocado en aquellas zonas de salud más afectadas dentro de la Comunidad de Castilla y León durante la *segunda ola*, cuya duración está se sitúa entre el 1 de julio y 18 de diciembre de 2020. Dichas zonas fueron escogidas por su alta incidencia y confinamientos, efectuados como consecuencia de esta.

Dichos confinamientos proporcionan un marco de estudio más apropiado al dado por otras zonas ya que ofrecen la posibilidad de obtener información mucho más detallada sobre la aplicación de medidas.

Las zonas de salud que se han utilizado como objeto de estudio han sido:

- **Cantalejo**
- **Aranda de Duero:**
  - **Aranda Sur**
  - **Aranda Rural**
  - **Aranda Norte**
- **Miranda de Ebro:**
  - **Miranda Oeste**
  - **Miranda Este**
- **Miranda del Castañar**
- **Mota del Marqués**
- **Peñafiel**
- **Medina del Campo:**
  - **Medina del Campo Urbano**
  - **Medina del Campo Rural**
- **Íscar**

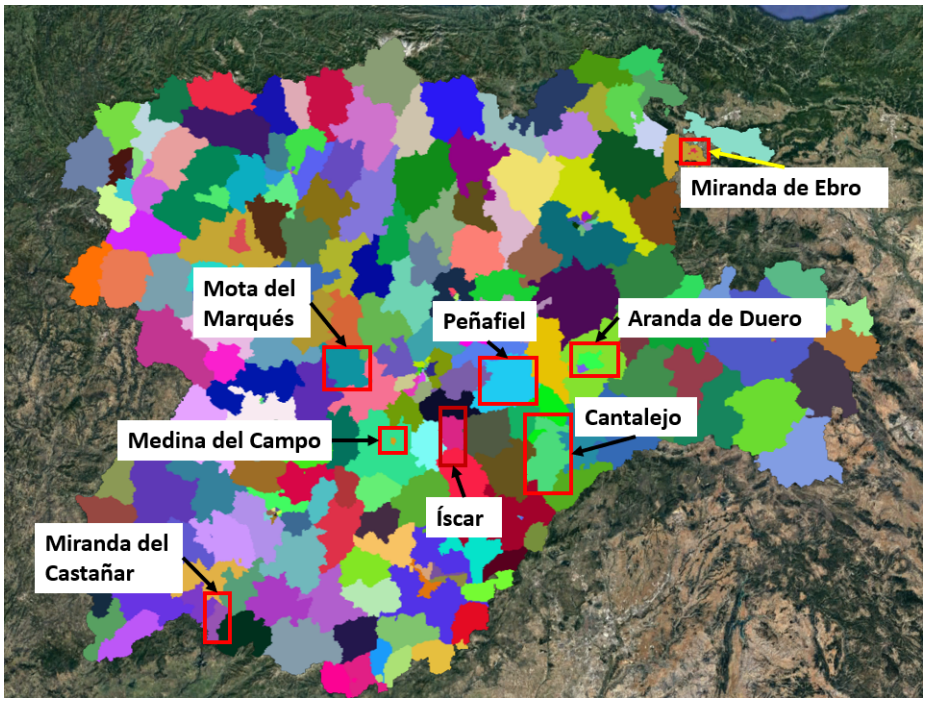


Figura 5.1: Localización de las zonas de salud escogidas

Las distintas zonas de salud presentadas corresponden a varias provincias dentro de la Comunidad Autónoma de Castilla y León tal y como se aprecia en la Figura 5.1. En ellas existen localidades como Aranda de Duero, Miranda de Ebro o Medina del Campo, donde todas sus zonas de salud se han visto afectadas por estos confinamientos.

5.3. Datos de salud y medidas aplicadas

5.3.1. Datos de salud

Estos datos nos han permitido determinar la situación epidemiológica de cada zona y han servido como base para todos los modelos y métodos construidos.

Portal de Datos Abiertos de la Junta de Castilla y León

En lo referente a los datos de salud, todos ellos han sido obtenidos usando como fuente de información el Portal de Datos Abiertos de la Junta de Castilla y León [23]. Este proyecto fue creado por dicha comunidad con el objetivo de aumentar la transparencia

en la información sobre la actividad en la comunidad, logrando así la participación y colaboración de los ciudadanos y empresas garantizando un intercambio de conocimiento.

Este proyecto se basa principalmente en la filosofía y práctica de Datos Abiertos (*Open Data*), que busca que ciertos datos estén disponibles de forma libre para todo el mundo, sin restricciones de *copyright*, patentes u otras formas de control. Los datos deben publicarse sin procesar (en bruto), bien estructurados y en formatos conocidos para así facilitar la reutilización.

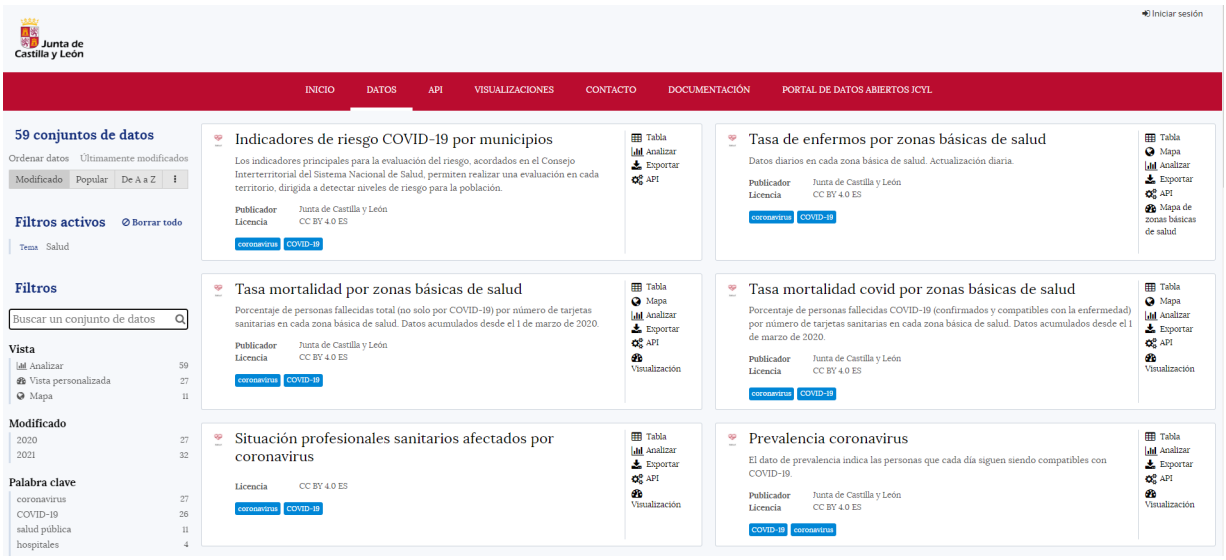


Figura 5.2: Portal de datos abiertos de la Junta de Castilla y León

Como podemos ver en la Figura 5.2, el portal ofrece una interfaz de acceso a os datos sencilla e intuitiva.

Filtrado de datos:

Tal y como se aprecia en la Figura 5.3, este portal nos ofrece herramientas como API's, métodos de visualización o múltiples herramientas de exportación en distintos formatos. Dentro del portal nos encontramos con datos de todo tipo organizados por categorías y etiquetas.

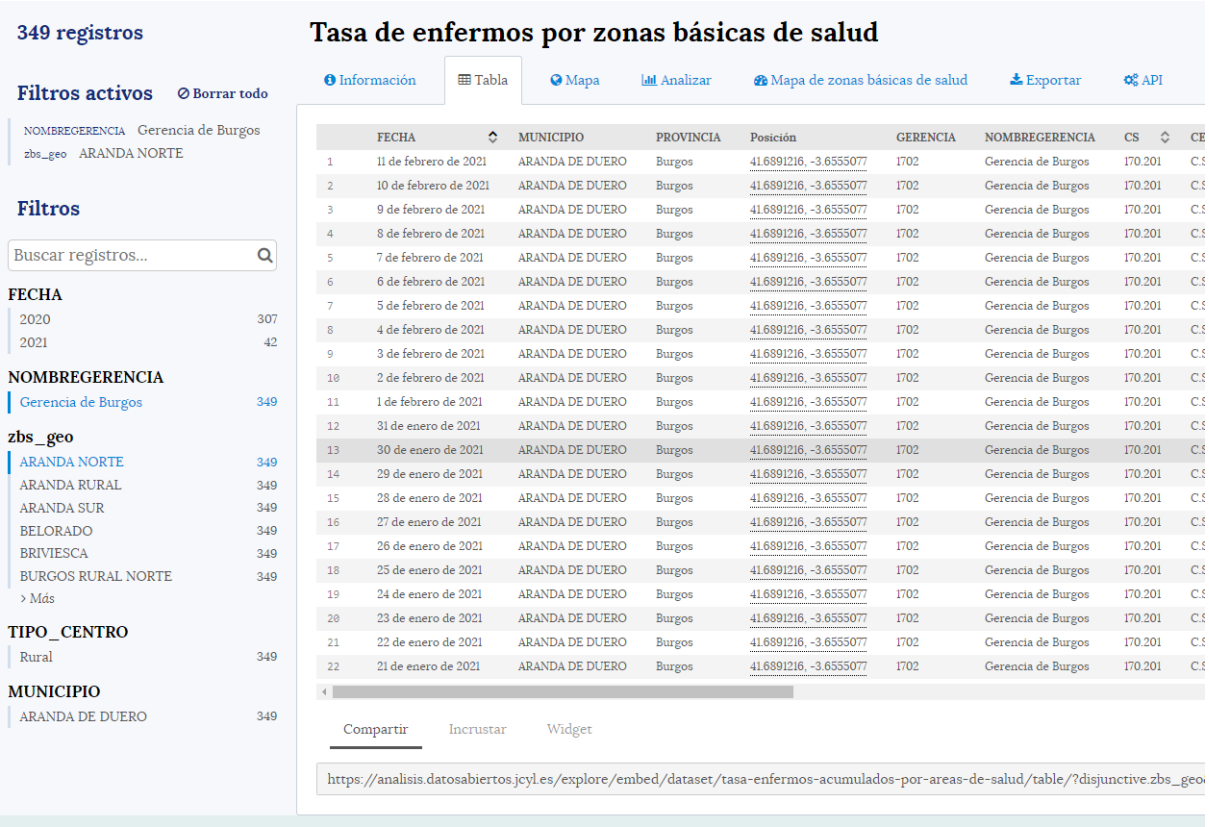


Figura 5.3: Panel de manejo de la información en Portal de datos abiertos

Para el desarrollo de este proyecto se han empleado aquellos datos situados en el Tema Salud y con la etiqueta COVID-19. Los datos consultados y extraídos han sido los siguientes:

Tasa de enfermos por zonas básicas de salud: Datos diarios sobre el número de enfermos en cada zona de salud de la Junta de Castilla y León. Las columnas extraídas de los datases han sido:

- **FECHA**: Fecha del día al que corresponden los datos de cada una de las filas en las que se organiza el dataset, el formato de esta será DD/MM/YYYY. Ejemplo: 01/08/2020.
- **NOMBRE GERENCIA**: Nombre de la gerencia a la que pertenece la zona de salud consultada. Ejemplo: Gerencia de Segovia.
- **CENTRO**: Nombre del Centro de Salud al que pertenece la zona consultada. Ejemplo: C.S. CANTALEJO.

- **PROVINCIA**: Nombre de la provincia a la que pertenece la zona de salud.
- **PCR REALIZADOS**: Número de pruebas PDIA (Pruebas diagnósticas de infección activa) realizadas en la zona de salud, incluyendo tanto pruebas PCR como de antígenos.
- **PCR POSITIVOS**: Número de pruebas PDIA (Pruebas diagnósticas de infección activa) que han obtenido un resultado positivo dentro de la zona de salud. Se incluyen positivos tanto de pruebas PCR como de antígenos.

Tasa de mortalidad COVID-19 por zonas básicas de salud: Porcentaje de personas fallecidas por COVID-19 (confirmados y compatibles con la enfermedad) por número de tarjetas sanitarias en cada zona básica de salud. Las columnas extraídas de los datases han sido:

- **FECHA**: Fecha del día al que corresponden los datos de cada una de las filas en las que se organiza el Dataset, el formato de esta será DD/MM/YYYY. Ejemplo: 01/08/2020.
- **NOMBRE GERENCIA**: Nombre de la gerencia a la que pertenece la zona de salud consultada. Ejemplo: Gerencia de Segovia.
- **CENTRO**: Nombre del Centro de Salud al que pertenece la zona de salud consultada. Ejemplo: C.S. CANTALEJO.
- **PROVINCIA**: Nombre de la provincia a la que pertenece la zona de salud.
- **FALLECIDOS**: Número de personas fallecidas por COVID-19 (confirmados y compatibles con la enfermedad).

Prevalencia coronavirus: Dato que indica el número de personas que cada día siguen siendo compatibles con COVID-19. Las columnas extraídas de los datasets obtenidos por cada una de las zonas de salud han sido:

- **FECHA**: Fecha del día al que corresponden los datos de cada una de las filas en las que se organiza el Dataset, el formato de esta será DD/MM/YYYY. Ejemplo: 01/08/2020.
- **NOMBRE GERENCIA**: Nombre de la gerencia a la que pertenece la zona de salud consultada. Ejemplo: Gerencia de Segovia.
- **CENTRO**: Nombre del Centro de Salud al que pertenece la zona consultada. Ejemplo: C.S. CANTALEJO.

- **PROVINCIA**: Nombre de la provincia a la que pertenece la zona de salud.
- **PREVALENCIA**: Dato que indica la tasa de prevalencia calculada teniendo en cuenta la población de tarjetas sanitarias de cada zona.

5.3.2. Datos sobre medidas aplicadas

La obtención de los datos de las medidas aplicadas en cada una de las zonas de salud estudiadas en el proyecto se ha hecho mediante la extracción de las mismas de los Boletines Oficiales publicados por la Junta de Castilla y León (*BOCYL*) [25].

Así se ha realizado un estudio y recopilación de todas las fechas en las que se aplicaron medidas para acceder rápidamente a aquellos documentos publicados dentro de la plataforma BOCYL que contuvieran información oficial sobre dichas medidas. Podemos distinguir entre varios tipos de medidas, aquellas aplicadas en conjunto a toda la Comunidad Autónoma y aquellas aplicadas a cada zona de salud individualmente del resto por los diferentes confinamientos.

Medidas utilizadas

A continuación se exponen, por fechas de aplicación, cada una de las medidas de cumplimiento obligatorio en la Comunidad Autónoma de Castilla y León, afectando a todas las zonas de salud analizadas. Quedarán resaltadas (usando un tipo de letra más oscura) aquellas medidas usadas en el proyecto, elegidas debido a su gran impacto y relevancia tanto en la población como en la economía:

17 AGOSTO:

- **Cierre discotecas, salas de baile, sin actuaciones musicales.**
- Distancia interpersonal de 1.5 metros en barra y entre mesas.
- **10 personas en mesa máximo**.
- **Cierre establecimientos 1:00 AM (limite ocio nocturno)**.
- Pruebas PCR a la entrada de centros socio-sanitarios.
- Limites visitas residentes hospitales 1 persona.
- Cribado PCR en caso de brote.
- Aumento de control sobre botellones.

- **Prohibición de fumar en la vía publica sin una distancia mínima de 2 metros**.

21 AGOSTO:

- **Limite en sectores de actividad al 50 %**.
  - Reducción de personas en entierros de 75 a 50.
  - Grupos al aire libre de 150 personas a 100 personas.
  - Grupos museos de 25 personas a 10 personas.
- **75 % aforo en hostelería (50 % barra)**.
- 80 % aforo en terrazas.
- **Cierre de peñas**.
- Limite horario nocturno salas de juego, atracciones de feria 1:00 AM.
- Autorización para eventos multitudinarios.

17 OCTUBRE:

- **Límite de reuniones a 6 personas** .
- **Límite de 6 personas en mesa**.
- **Prohibido consumo en barra y de pie en establecimientos**.
- **50 % aforo hostelería en caso de salas de 40 comensales y 75 % en caso de más**.

24 OCTUBRE:

- **Toque de queda nocturno a partir de las 22:00 PM**.

30 OCTUBRE:

- **Cierre perimetral de la Comunidad Autónoma de Castilla y León**.

6 NOVIEMBRE:

- Creación semáforo/niveles de alerta.
- **Cierre hostelería** .
- **Cierre de centros comerciales**.
- **Cierre de gimnasios e instalaciones deportivas**.
- **Suspendidas visitas a residencias de mayores**.

  En la Figura 5.4 se muestra de forma más clara las medidas usadas dentro del proyecto.

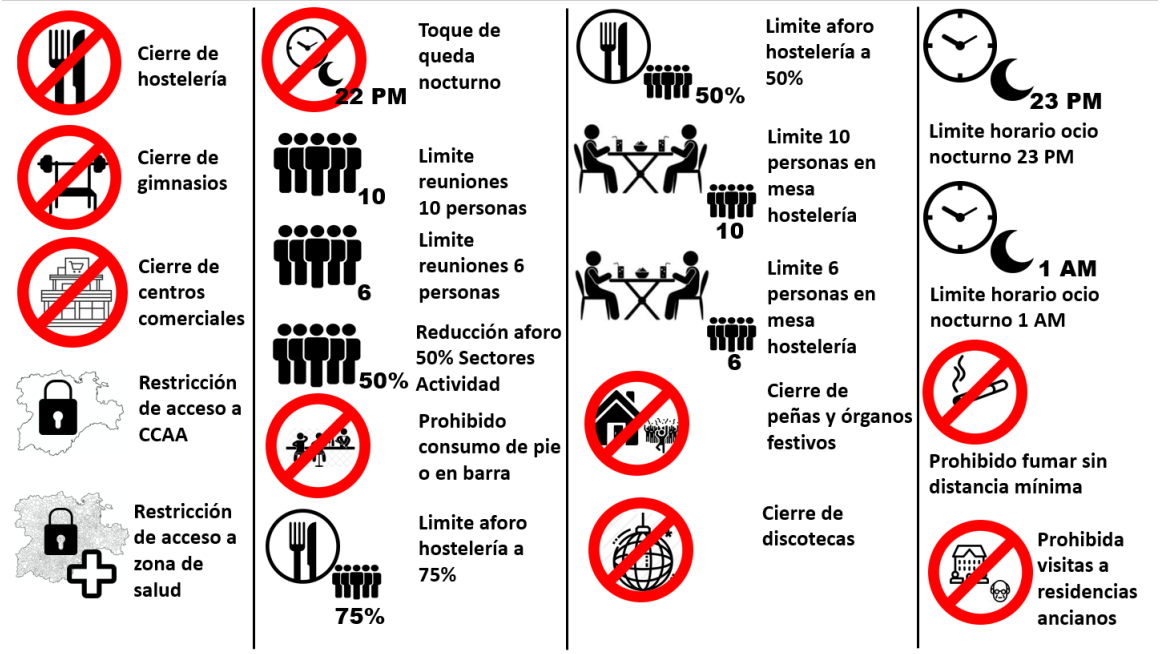


Figura 5.4: Medidas utilizadas en proyecto

Centrándonos en los confinamientos, los datos de medidas y restricciones aplicadas han sido obtenidos consultando los boletines oficiales de la Junta de Castilla y León disponibles para cada confinamiento, así como diferentes diarios y periódicos, siempre contrastando la información de todas las fuentes para la máxima precisión.

Las medidas aplicadas durante los confinamientos no difieren de aquellas aplicadas a toda la Comunidad, aunque si se diferencian en la fecha de aplicación, siendo muchas de las medidas de los confinamientos anteriores a su aplicación a nivel de Comunidad y no habiendo ninguna posterior al día **24 de octubre** (día en el que se estableció el Estado de Alarma [31] en todo el país y por lo tanto fin de todos los confinamientos). Dichos confinamientos han tenido una duración aproximada de 14 días en la mayoría de zonas, aplicándose medidas similares por zona.

Zona de salud	**Fecha inicio de confinamiento**	**Fecha fin de confinamiento**
CANTALEJO	22/08/2020	04/09/2020
ARANDA SUR	07/08/2020	21/08/2020
ARANDA RURAL	07/08/2020	21/08/2020
ARANDA NORTE	07/08/2020	21/08/2020
MIRANDA OESTE	26/09/2020	24/10/2020
MIRANDA ESTE	26/09/2020	24/10/2020
MIRANDA DEL CASTAÑAR	26/09/2020	13/10/2020
MOTA DEL MARQUÉS	13/10/2020	24/10/2020
PEÑAFIEL	22/09/2020	06/10/2020
MEDINA DEL CAMPO URBANO	29/09/2020	13/10/2020
MEDINA DEL CAMPO RURAL	29/09/2020	13/10/2020
ÍSCAR CONFINAMIENTO 1	02/08/2020	16/08/2020
ÍSCAR CONFINAMIENTO 2	18/09/2020	17/10/2020

Tabla 5.1: Periodos de aplicación de confinamientos en cada zona de salud

Para su mejor comprensión, a continuación se muestra un panel acerca de qué medidas han sido aplicadas en cada uno de los confinamientos de cada zona:

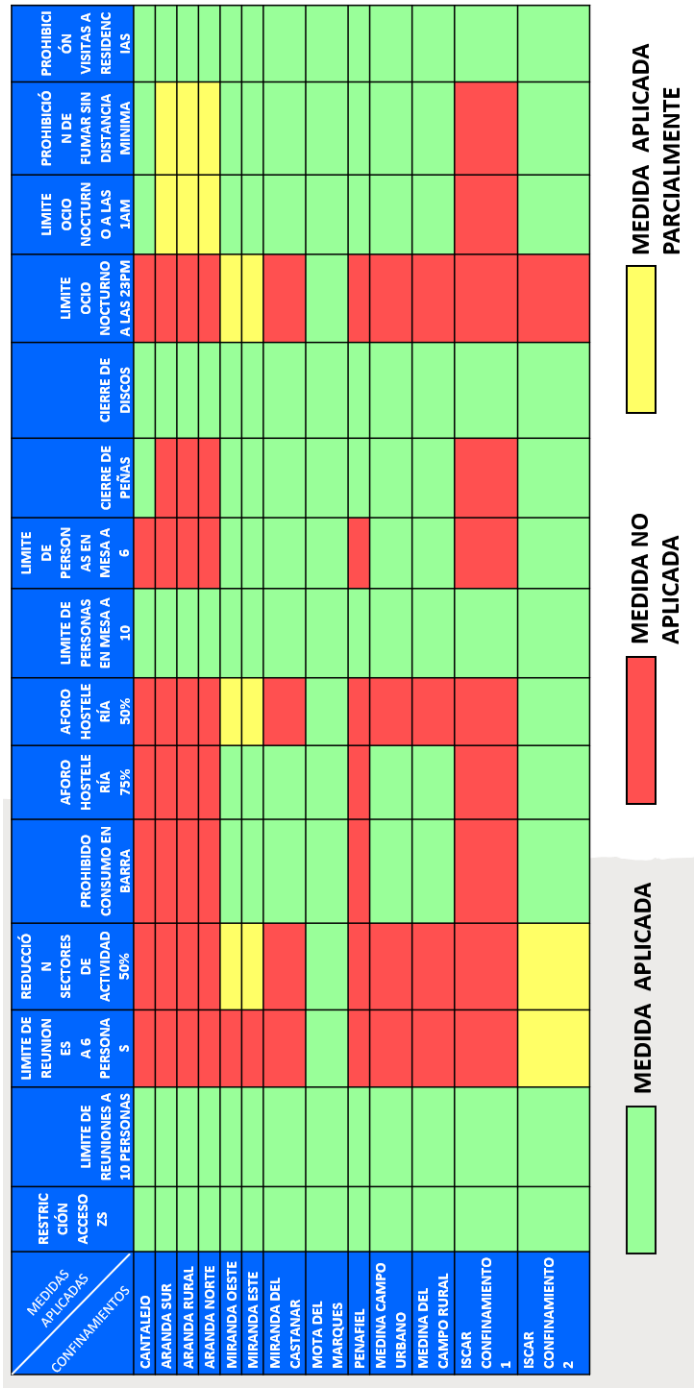


Figura 5.5: Medidas aplicadas durante los confinamientos

5.3.3. Creación de datasets y variables

Toda la información relacionada con la salud y las medidas aplicadas ha sido recogida dentro de un mismo dataset empleando la herramienta de Excel.

Los datos se dividen por zonas de salud y de manera cronológica durante la *segunda ola*, que va desde el 1 de julio hasta el 18 de diciembre de 2020 (fecha determinada como fin de la ola).

Variables

Nuestro dataset dentro de Excel se estructura como una tabla en la que cada columna corresponderá a una variable y cada fila a los datos de una determinada fecha y zona de salud.

Las columnas que forman la tabla serán aquellas relacionadas con las variables de salud y medidas. Dichas columnas contendrán un nombre único de variable. A continuación se muestran todas las variables de salud y medidas aplicadas a través de una serie de tablas donde se indica el tipo de dato de la variable junto a una breve descripción de la misma:

- **Tabla variables de salud (Tabla 5.2):** Muestra todas aquellas variables relacionadas con datos de salud sobre el COVID-19.
- **Tabla variables matemáticas de salud (Tabla 5.3):** Indica aquellas variables relacionadas con los diferentes indicadores de incidencia de contagios y muertes en periodos de tiempo definidos.
- **Tabla variables medidas aplicadas 1 (Tabla 5.4):** Muestra aquellas variables relacionadas con las medidas estudiadas en el proyecto.
- **Tabla variables medidas aplicadas 2 (Tabla 5.5):** Muestra aquellas variables relacionadas con las medidas estudiadas en el proyecto.

Nombre variable	**Tipo**	**Descripción**
FECHA	STRING	Fecha a la que corresponden los datos de la fila
NOMBRE_GERENCIA	STRING	Nombre de la provincia de gerencia a la que pertenecen los datos
CENTRO	STRING	Nombre de la zona de salud a la que pertenecen los datos
PROVINCIA	STRING	Provincia a la que pertenecen los datos
MOV_CLUSTER	INT	Valor dado a cada una de las subidas y bajadas obtenidas en cada zona mediante la técnica de clustering usada (esta variable será explicada en capítulos posteriores)
PREVALENCIA	INT	Número de personas compatibles con COVID-19
PCR_REALIZADOS	INT	Número de pruebas PCR realizadas
PCR_POSITIVOS	INT	Número de pruebas PCR con resultado positivo
FALLECIDOS	INT	Número de personas fallecidas a causa de la COVID-19
VIAS_IMPORTANTES	BOOLEAN	Indicador de la existencia de carreteras con alto tránsito muy cercanas a la zona de salud.
POBLACION	INT	Indicador del número de habitantes de cada zona en función de las tarjetas sanitarias registradas en dicha zona.

Tabla 5.2: Variables de salud

Nombre variable	**Tipo**	**Descripción**
PORC_PCR_XDIAS	FLOAT	Porcentaje de la situación epidemiológica de cada zona en relación al número de PCR realizadas y número de PCR positivas obtenidas
POSI_VENT_MOVXDIAS	FLOAT	Resultado de la aplicación de una ventana móvil sobre los datos de PCR positivas en el marco temporal X pudiendo ser este de 4, 7 y 14 días
POSI_IA_XDIAS	FLOAT	Resultado del cálculo de la Incidencia acumulada de PCR positivas usando el marco temporal X pudiendo ser este de 4, 7 y 14 días
REALI_VENT_XDIAS	FLOAT	Resultado de la aplicación de una ventana móvil sobre los datos de PCR realizadas en el marco temporal X pudiendo ser este de 4, 7 y 14 días
REALI_IA_XDIAS	FLOAT	Resultado del cálculo de la Incidencia acumulada de PCR realizadas usando el marco temporal X pudiendo ser este de 4, 7 y 14 días
FALL_VENT_MOVX	FLOAT	Resultado de la aplicación de una ventana móvil sobre los datos de fallecidos por COVID-19 en el marco temporal X pudiendo ser este de 4, 7 y 14 días
FALL_IA_XDIAS	FLOAT	Resultado del cálculo de la Incidencia acumulada de fallecidos por COVID-19 usando el marco temporal X pudiendo ser este de 4, 7 y 14 días

Tabla 5.3: Variables matemáticas de salud

Nombre variable	**Tipo**	**Descripción**
CIERRE_HOSTELERIA	BOOLEAN	Cierre de toda la hostelería y restauración
CIERRE_GIMNASIOS	BOOLEAN	Cierre de todos los gimnasios e instalaciones deportivas
CIERRE_C_COMERCIALES	BOOLEAN	Cierre de los centros comerciales y grandes establecimientos
REST_ACCESO_CCAA	BOOLEAN	Cierre perimetral de la CA de Castilla y León
REST_ACCESO_ZS	BOOLEAN	Restricción de acceso a la zona de salud
TOQUE_NOCTURNO	BOOLEAN	Toque de queda nocturno a las 22:00PM
LIMIT_REUN_10	BOOLEAN	Limitación de las reuniones sociales a 10 personas máximo
LIMIT_REUN_6	BOOLEAN	Limitación de las reuniones sociales a 6 personas máximo
RED_PERS_SACTV_50	BOOLEAN	Reducción del número de personas permitidos en los sectores de actividad al 50 %
PROHIB_CONSUM_BARRA	BOOLEAN	Prohibido el consumo en barra o de pie en los establecimientos de hostelería

Tabla 5.4: Variables medidas aplicadas 1

Nombre variable	**Tipo**	**Descripción**
AFORO_HOST_75	BOOLEAN	Limitación al 75 % del aforo permitido en el interior de los recintos de hostelería
AFORO_HOST_50	BOOLEAN	Limitación al 50 % del aforo permitido en el interior de los recintos de hostelería
LIMIT_MESA_10	BOOLEAN	Reducción del número de personas permitidas a 10 máximo en mesas de hostelería
LIMIT_MESA_6	BOOLEAN	Reducción del número de personas permitidas a 6 máximo en mesas de hostelería
CIERRE_PENAS	BOOLEAN	Cierre de peñas y organizaciones con motivos festivos
CIERRE_DISCO	BOOLEAN	Cierre de discotecas y establecimientos de ocio nocturno
LIMIT_OC_NOCT_23PM	BOOLEAN	Limite del horario de apertura del ocio nocturno a las 23:00 PM
LIMIT_OC_NOCT_1AM	BOOLEAN	Limite del horario de apertura del ocio nocturno a la 1:00 AM
PROHIB_FUMAR	BOOLEAN	Prohibición de fumar en la vía pública si no se respeta una distancia mínima de dos metros
PROHIB_VISIT_RESI	BOOLEAN	Prohibidas las visitas a residentes en residencias de ancianos

Tabla 5.5: Variables medidas aplicadas 2

Cálculo de variables estadísticas de salud

Ventanas Móviles:

Debido a la naturaleza cambiante de los datos obtenidos sobre el COVID-19 y su variabilidad diaria se propone la creación de marcos de tiempo usando ventanas móviles, para así, poder observar correctamente y de una forma más clara todas las subidas y bajadas en cada una de las zonas de salud. Los marcos temporales utilizados han sido de 4, 7 y 14 días.

La elección de estos marcos temporales se basa en los periodos usados en los estudios estadísticos de los datos relacionados con el COVID-19 en los que los marcos temporales de 7 y 14 días ofrecen información de la tendencia de los datos en el tiempo e índice de probabilidad de rebrotes [17] .

El marco temporal de 4 días ha sido escogido para obtener una forma más de observación de la tendencia de los datos a través del tiempo en un periodo de tiempo menor.

Para la obtención de estas ventanas móviles se han usado las variables **PCR_REALIZADAS**, **PCR_POSITIVAS y FALLECIDOS**, debido a que son las que más información nos aportan acerca del impacto de la pandemia en cada zona. Se ha hecho uso de la función rolling().mean() sobre dichas columnas, lo que nos genera una media aritmética de los datos en función del marco de tiempo indicado (4, 7 o 14 días), obteniendo las variables **POSI_VENT_MOVXDIAS, REALI_VENT_XDIAS y FALL_VENT_XDIAS**.

Incidencia Acumulada: Alineándonos con la comunidad científica y con la variable de Incidencia Acumulada por cada 100000 habitantes en los periodos de 7 y 14 días usada para la medición y determinación de la situación epidemiológica de cada zona, se ha creado una variable basada en dicha incidencia con los datos de salud disponibles de cada zona.

$$IA = \frac{Numero_{casos-en-un-periodo}}{Numero_{individuos-libres-inicio-periodo}} \times 100000 \tag{5.1}$$

La obtención de esta variable de Incidencia Acumulada ha diferido un poco en la fórmula teórica vista, ya que se ha sustituido el numerador por los resultados obtenidos con las ventanas móviles explicadas anteriormente, y el denominador por la población total basada en el número de tarjetas sanitarias registradas de cada zona de salud, correspondiente a la variable **POBLACIÓN**. Esto se ha debido a las limitaciones de obtención de información encontradas:

$$POSI\_IA\_XDIAS = \frac{POSI\_VENT\_MOVXDIAS}{POBLACION} \times 100000 \tag{5.2}$$

$$REALI\_IA\_XDIAS = \frac{REALI\_VENT\_MOVXDIAS}{POBLACION} \times 100000 \tag{5.3}$$

$$FALL\_IA\_XDIAS = \frac{FALL\_VENT\_MOVXDIAS}{POBLACION} \times 100000 \tag{5.4}$$

Los resultados obtenidos al aplicar esta fórmula han sido almacenados en las variables **POSI_IA_XDIAS, REALI_IA_XDIAS y FALL_IA_XDIAS.** Pudiendo ser el valor de la **X** en **XDIAS**, 4, 7 y 14, correspondientes a los periodos estudiados.

Porcentaje PCR:

Una de las principales variables a tener en cuenta a la hora de determinar la evolución de los contagios en cada una de las zonas estudiadas es el número de PCR realizadas. La disminución del número de PCR realizadas repercute directamente en el número de PCR positivas, ya que al hacerse menos pruebas disminuye la detección de casos positivos y por tanto la información real sobre el número de contagios en cada zona.

Es por ello, que hemos creado una variable denominada porcentaje PCR (**PORC_PCR_XDIAS**) que nos permite obtener un valor aproximado de la situación epidemiológica de cada zona pese a la reducción de pruebas realizadas, obteniendo así unos datos más precisos. A mayor porcentaje, más probabilidades hay que en la zona haya una gran incidencia oculta debido a la falta de pruebas realizadas.

Para la obtención de esta variable se realiza una división entre la variable **POSI_IA_XDIAS** obtenida anteriormente y la variable **REALI_IA_XDIAS** en un marco de tiempo X determinado, multiplicando el resultado de esta división por 100 para obtener así un porcentaje que de tal manera nos de información en cada uno de los marcos de tiempo estudiados (4, 7 y 14 días):

$$PORC\_PCR\_XDIAS = \frac{POSI\_IA\_XDIAS}{REALI\_IA\_XDIAS} \times 100 \tag{5.5}$$

Todos los cálculos realizados se han hecho de acuerdo al formato numérico aceptado por Excel y haciendo uso de todas las operaciones matemáticas que la herramienta nos ofrece.

5.4. Datos movilidad

La obtención de los datos de movilidad en cada una de las zonas de salud estudiadas ha sido clave a la hora de analizar los resultados obtenidos por los métodos y modelos utilizados. Estos datos nos han permitido obtener una explicación inicial a las causas de las subidas de contagios, como las bajadas en cada zona.

Inicialmente nos hemos centrado en aquellos datos que nos han permitido saber el número de entradas a cada una de las zonas estudiadas y el motivo de la entrada, para así, determinar cuanta gente entraba a dicha zona desde fuera y cómo afectaba al aumento de contagios.

Los tres principales motivos por los que se han dividido las entradas a cada zona de salud han sido:

- **Viajes a lugar habitual de residencia:** Aquellas entradas justificadas que tenían como actividad de destino la vuelta a la residencia habitual en la que vive.
- **Viajes por trabajo:** Aquellas entradas justificadas a la zona de salud por motivos laborales.
- **Viajes por otros motivos:** Entradas a la zona estudiada por motivos diferentes a los ya mencionados. Entre este tipo de entradas pueden estar aquellas dadas por turismo, visita a familiares, ocio y entretenimiento, procesos administrativos... todas ellas consideradas **entradas no justificadas.**

También ha sido objeto de estudio la movilidad en el interior de la zona de salud entre aquellos municipios más afectados para poder determinar si esto ha sido causa principal del aumento y disminución de contagios. Los municipios estudiados han sido:

- Pedrajas (dentro de la zona de salud de Íscar) y la movilidad entre Íscar y Pedrajas.
- Pesquera de Duero (dentro de la zona de salud de Peñafiel) estudiando la movilidad entre Pesquera de Duero y Peñafiel.

5.4.1. Proyecto de estudio de movilidad con Big Data

La principal fuente de información usada para la obtención de entradas producidas a cada zona, ha sido el proyecto de estudio de la movilidad con Big Data, llevado a cabo por el Ministerio de Transportes, Movilidad y Agenda Urbana del Gobierno de España [38].

El propósito de este estudio es describir la movilidad a nivel nacional, autonómico, provincial y local, para apoyar los esfuerzos de monitorización de la progresión de la enfermedad y evaluar la efectividad de las medidas tomadas para restringir la movilidad. Todo ello, con el objetivo de toma de las mejores decisiones durante el tiempo de duración de la pandemia.

El estudio a su vez, usa como referencia y comparación una semana correspondiente a la denominada antigua normalidad, coincidiendo esta con el periodo del 14 al 20 de febrero de 2020. Gracias a esta referencia se tiene información de la movilidad en cada zona cuando no existían restricciones.

Dentro de este proyecto existe una actualización diaria de los datos y la información, siendo la información de cada día la correspondiente al estudio realizado de los 3 días anteriores al mismo (tiempo mínimo de tratamiento y disponibilidad de los datos de movilidad).

La principal herramienta usada por el estudio para la obtención de los datos de movilidad ha sido la ubicación de los dispositivos móviles de la ciudadanía (siempre cumpliendo las leyes y directivas de protección de datos vigentes) [10].

El funcionamiento de esta herramienta de obtención se basa principalmente en el análisis de una muestra de gran tamaño, procesando la información y generando indicadores correspondientes a un periodo de respuesta de tres días. El proyecto reutiliza el trabajo y datos obtenidos en el uso del Big Data en trabajos anteriores de movilidad del Gobierno de España.

El proyecto realiza un estudio diario y horario de cada una de las zonas en las que se divide el país, para evaluar y comparar la evolución y tendencia de la movilidad en comparación con la semana de referencia escogida. Se analiza tanto la movilidad en el interior de los territorios, como aquella proveniente del exterior, usando indicadores sobre el número de viajes diarios de la población.

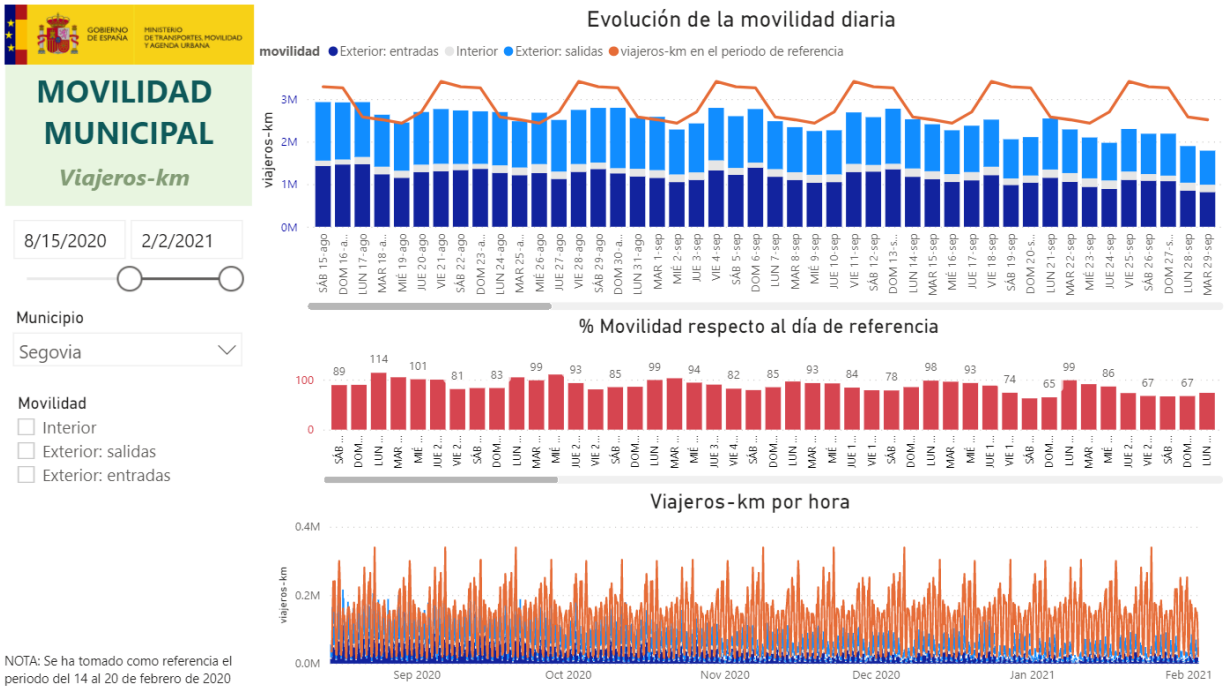


Figura 5.6: Aplicación online del proyecto de estudio de movilidad con Big Data

El estudio proporciona una aplicación online, tal y como podemos ver en la Figura 5.6, que a través de una interfaz muestra los gráficos de la evolución de toda la movilidad diaria registrada a nivel nacional, autonómico y local, tanto por horas como por distancia.

En el caso de querer acceder a los datos obtenidos por el estudio, el Gobierno de España provee de un portal de datos abiertos [39] ,donde el contenido está estructurado (en un primer nivel) en dos carpetas correspondientes a dos matrices maestras, la matriz de viajes (maestra 1) y la matriz de viajes por persona (maestra 2). Cada una de estas carpetas tiene un segundo nivel donde los ficheros e información son mostrados por días y meses completos (tanto del periodo estudiado como del periodo de referencia usado). El portal también incluye información de la zonificación empleada y la representación geográfica de la misma

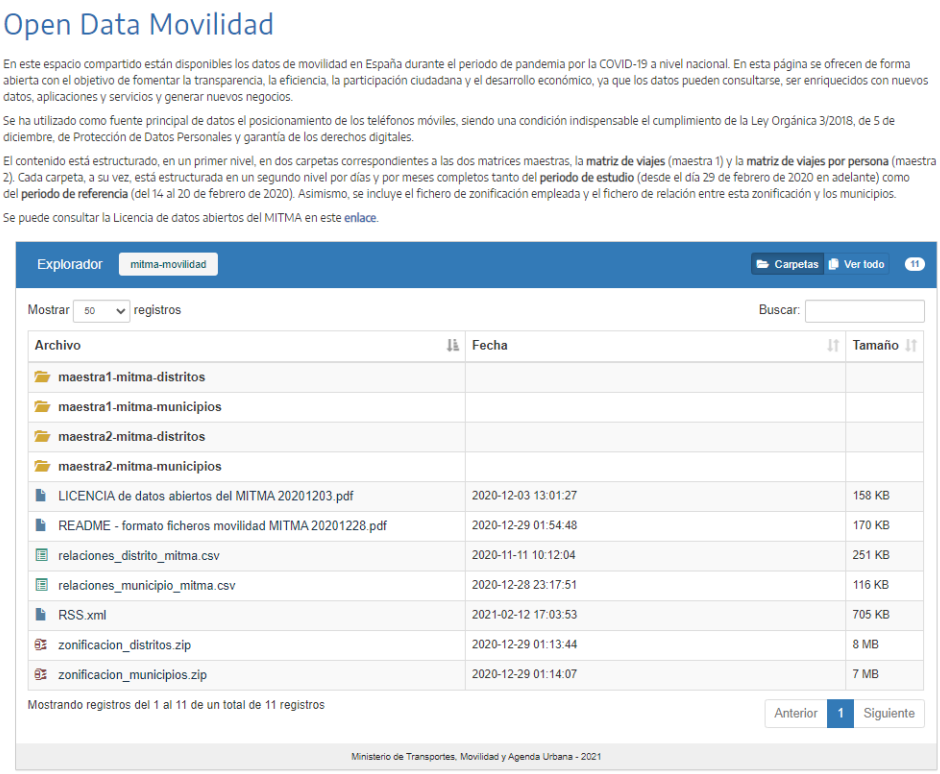


Figura 5.7: Portal de datos abierto del proyecto de estudio de movilidad con Big Data

5.4.2. Estructura de la información

Dentro de este portal de datos abiertos se encuentran una serie de archivos y documentos que proporcionan todos los datos de movilidad obtenidos con el estudio. En nuestro proyecto nos centraremos en aquellas carpetas correspondientes a la matriz de viajes (maestra 1) cuyo contenido será el siguiente:

"maestra1-mitma-distritos": Carpeta donde podemos encontrar las matrices de viajes por días y por meses completos actualizadas diariamente.

Esta matriz matriz de viajes (maestra 1) por distritos contiene el número de viajes y de viajeros por km, para cada día y cada combinación de origen, destino, actividad de origen, actividad de destino, residencia, periodo horario y distancia (por rangos).

Los ficheros de texto (.txt) donde esta contenida la información, están divididos en su interior con campos o variables, separadas por '|'(barra vertical) y valores numéricos basados en el uso del '.'(punto) como separador decimal. Dichas variables serán:

fecha|origen|destino|actividad_origen|actividad_destino|residencia
|edad|periodo|distancia|viajes|viajes_km

Teniendo los ficheros la estructura mostrada en la Figura:

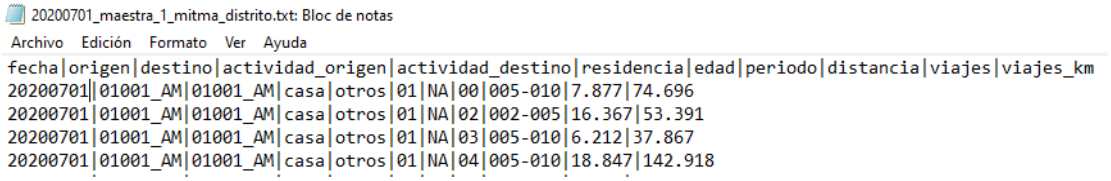


Figura 5.8: Fichero de texto datos movilidad

Las variables origen y destino hacen referencia al código de distrito o a una agrupación de estos en el caso de zonas con poca población. Dicho código, es obtenido gracias a la división por distritos realizada, la cuál se pude visualizar importando el archivo de tipo Shapefile de zonificación a través de la herramienta Google Earth. Dicha zonificación por distritos se muestra en la Figura 5.9.

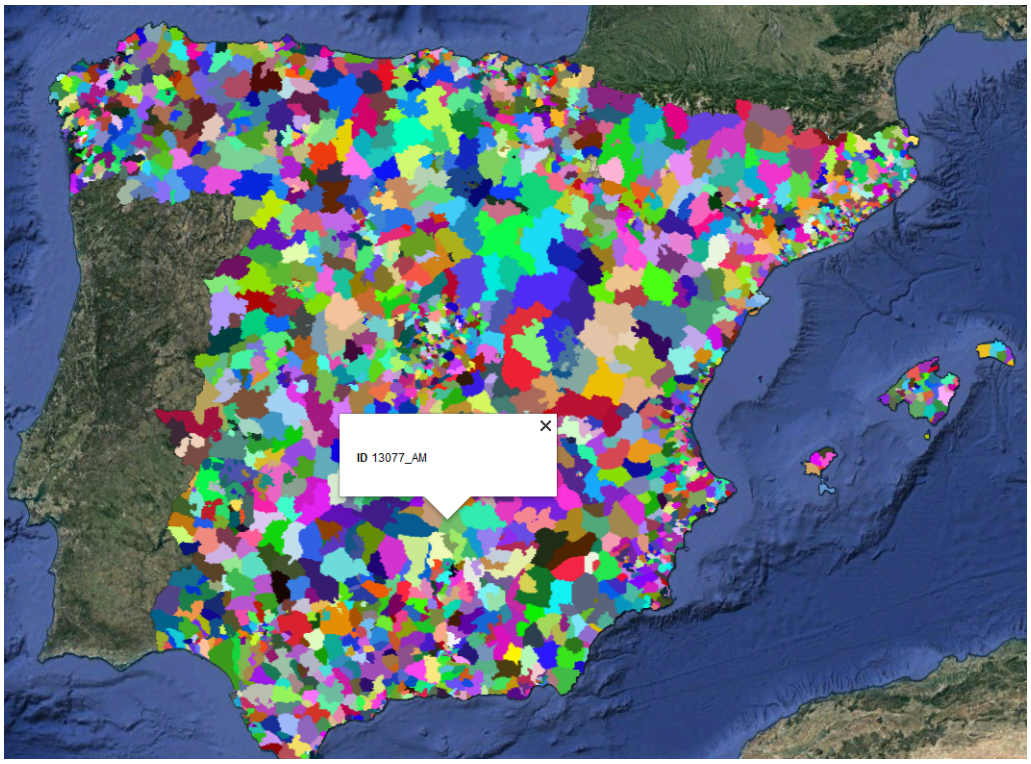


Figura 5.9: Muestra de la zonificación por distritos en la herramienta Google Earth

"maestra1-mitma-municipios": Carpeta donde podemos encontrar las matrices de viajes por días y por meses completos actualizadas diariamente.

Dicha matriz de viajes por municipios contiene el número de viajes y de viajes por km para cada día y cada combinación de origen, destino, periodo y distancia (por rangos):

fecha|origen|destino|periodo|distancia|viajes|viajes_km

Los ficheros de texto poseerán las misma estructura que los vistos ya explicados para la matriz de viajes de distritos (**maestra1-mitma-distritos**). En esta caso la zonificación realizará una división de los territorios en municipios y no en distritos tal y como podemos ver en la Figura 5.10.

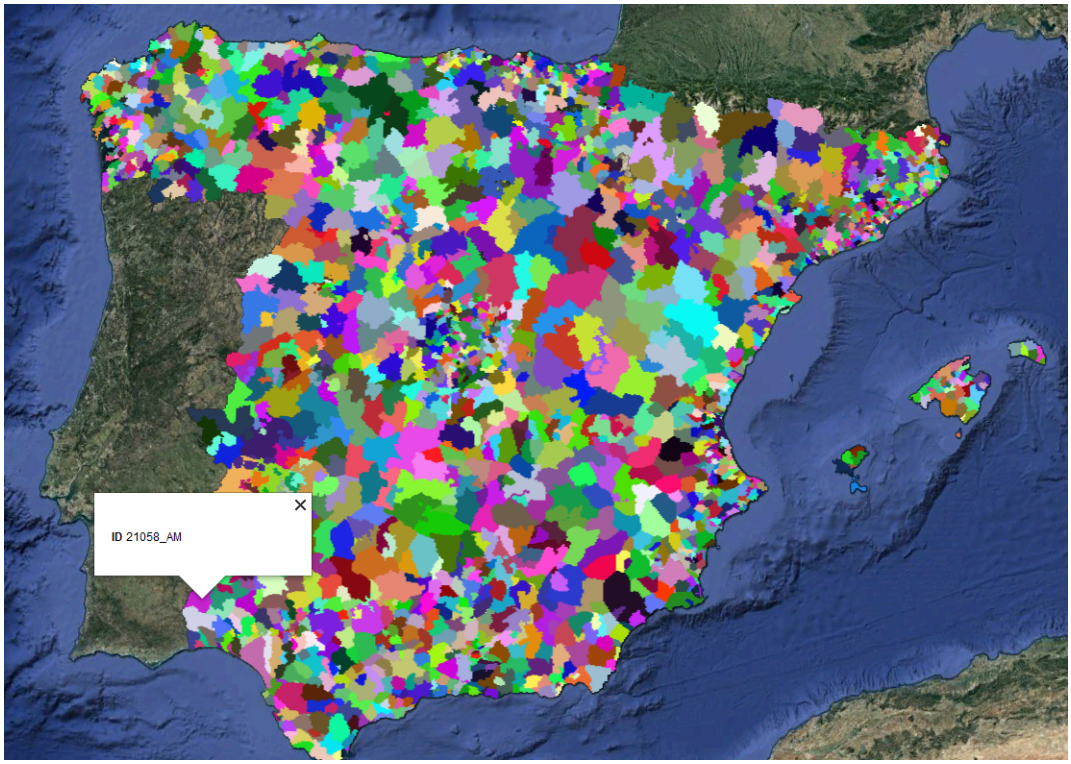


Figura 5.10: Muestra de la zonificación por municipios en la herramienta Google Earth

5.4.3. Creación de datasets y variables de movilidad

Debido al formato y tamaño de los ficheros de texto proporcionados en el portal de datos abiertos, se han creado una serie de herramientas de extracción, procesamiento y conversión de datos para poder crear así un dataset de movilidad acorde a nuestro proyecto y compatible con las herramientas que hemos utilizando dentro de este para un análisis

adecuado de la información.

Obtención y conversión de datos diarios

Mediante Jupyter Notebooks hemos creado una herramienta de recopilación y transformación a partir de los ficheros txt obtenidos del portal de datos abiertos . Esta herramienta realizará una transformación de los datos del formato txt al formato csv, creando así ficheros diarios del número de entradas a cada zona. Los pasos seguidos para su obtención han sido los siguientes:

1. Descarga y almacenamiento en local de todos los ficheros txt diarios del portal de datos abiertos correspondientes al marco de tiempo de estudio (*segunda ola*).
2. Obtención de los códigos de distrito de cada una de las zonas de salud estudiadas.
3. Introducción del código de distrito dentro del script, correspondiente a la zona de salud deseada para la generación de datasets diarios de dicha zona. Se deberá proporcionar también la ruta de guardado de los ficheros diarios generados.
4. Ejecución de todos los script para la creación de los ficheros por días en formato csv de cada zona de salud.

```
In [1]:
#----------------------------GENERADOR DATASETS JUNIO
import pandas as pd
first_num = [0, 1, 2, 3 ]
second_num = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9 ]

for x in first_num:
  for y in second_num:
    if y==0 and x==0:

        y=1

    if x==3 and y==1:
      break;
    else:
      data = pd.read_csv(r'd:\Users\Alvaro\Desktop\DATASETS MOVILIDAD\DATOS\JUNIO\202006%s%s maestra 1 mitma distrito.txt'
      % (x, y), sep="|", header=None)
      data.columns = ["fecha", "origen", "destino", "actividad_origen", "actividad_destino", "residencia",
                      "edad", "periodo", "distancia", "viajes", "viajes_km"]

      test=data['destino'].values=='4019403'



      test2=data[test]
      df1=test2
      print('GENERANDO FICHERO DESTINO DEL DIA %s%s DE JUNIO...' % (x, y))
      df1.to_csv(r'd:\Users\Alvaro\Desktop\DATASETS MOVILIDAD\DATOS\JUNIO\SEGOVIA I\%s%sJUNIO.csv' % (x, y), index = False)

C:\Users\Alvaro\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3146: DtypeWarning: Columns (0,1,5,6,7,9,10) have
mixed types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,

GENERANDO FICHERO DESTINO DEL DIA 01 DE JUNIO...
```

Figura 5.11: Parte de herramienta de recopilación y conversión usada para la obtención de datos diarios

El motivo por el que se ha usado Jupyter Notebooks en lugar de Google Colab, ha sido la necesidad de almacenamiento de los ficheros de texto descargados del portal de datos abiertos, ya que, debido al tamaño de estos ficheros, se hace complejo su procesamiento en Colab a causa de las limitaciones de la versión trial en cuanto a almacenamiento.

Obtención datos mensuales

Una vez generados y almacenados todos los ficheros en formato csv de las entradas por días, haremos uso de una herramienta de obtención y transformación implementada en Google Colab, que a partir de estos ficheros csv diarios, nos generará automáticamente ficheros de movilidad por meses de cada zona de salud. Los pasos a seguir son:

1. Introducción de ficheros csv por días de cada zona de salud.
2. Ejecución del script.
3. Almacenamiento de los ficheros csv de movilidad por meses de cada zona de salud.

Este script genera una serie de variables relacionadas con el números de viajes realizados (**viajes**) y los motivos de estos (**actividad_destino**). Dichas variables se generarán realizando una suma de todas las entradas efectuadas en un día, diferenciando entre motivos de entrada (entradas por vuelta a residencia habitual, por trabajo u otros motivos). En la Figura 5.12 podemos ver un fragmento de los datos obtenidos con este script.

	fecha	destino	sum_viajes_total	sum_viajes_casa	sum_viajes_trabajo	sum_viajes_otros	i
1	20200701	4019406	10630.172	7029.114	681.557	2919.501	
2	20200702	4019406	10999.675	7588.352	774.384	2636.939	
3	20200703	4019406	10865.745	7063.868	829.426	2972.451	
4	20200704	4019406	8620.791	6167.038	543.671	1910.082	
5	20200705	4019406	7339.282	5472.279	405.343	1461.660	
6	20200706	4019406	10116.711	7049.756	801.959	2264.996	
7	20200707	4019406	8944.980	6034.100	693.043	2217.837	
8	20200708	4019406	9905.278	6714.052	827.887	2363.339	
9	20200709	4019406	10170.970	6758.154	837.601	2575.215	
10	20200710	4019406	9369.172	6173.519	798.033	2397.620	
11	20200711	4019406	8255.832	5710.530	569.177	1976.125	
12	20200712	4019406	7573.273	5450.326	343.319	1779.628	
13	20200713	4019406	9992.904	6921.010	509.318	2562.576	
14	20200714	4019406	9500.928	6735.092	523.853	2241.983	
15	20200715	4019406	9546.502	6766.530	537.952	2242.020	
16	20200716	4019406	9669.618	6633.151	795.727	2240.740	
17	20200717	4019406	9587.567	6713.792	511.768	2362.007	
18	20200718	4019406	7915.363	5505.833	486.582	1922.948	
19	20200719	4019406	7183.309	5177.531	422.254	1583.524	
20	20200720	4019406	9492.365	6951.480	514.826	2026.059	
21	20200721	4019406	9032.315	6140.940	529.495	2361.880	
22	20200722	4019406	10090.814	6959.066	677.833	2453.915	
23	20200723	4019406	9654.622	6621.820	748.753	2284.049	
24	20200724	4019406	9227.066	6144.062	654.666	2428.338	

Figura 5.12: Salida generada por herramienta de obtención de datos de movilidad por meses

Generación dataset global

Una vez obtenidos los ficheros csv de movilidad por meses de cada zona de salud, mediante la herramienta **OpenRefine**, se unifican todos estos ficheros mensuales, obteniendo un dataset global de la movilidad en todas las zonas.

Openrefine permite agrupar y organizar todos los ficheros csv mensuales en un mismo dataset los datos quedan organizados por zonas de salud, correspondiendo cada fila a un día determinado en una zona de salud específica. Cada columna representará una variable, señaladas en la Tabla 5.6.

Una vez creado el dataset global, se siguen los pasos descritos anteriormente para la obtención de los datos de referencia usados en el estudio, los cuales corresponden al

periodo temporal del 14 al 20 de febrero (periodo perteneciente a la antigua normalidad). Estos datos son integrados en el dataset global de movilidad. Las variables de nuestro dataset global son las siguientes:

Nombre variable	**Tipo**	**Descripción**
FECHA	STRING	Fecha del día en el que se registró la movilidad
DESTINO	STRING	Nombre del lugar de estudio en el que se ha realizado toda la movilidad de entradas
SUM_VIAJES_TOTAL	FLOAT	Suma total del número de viajes realizados cada día a la zona de salud estudiada
SUM_VIAJES_CASA	FLOAT	Suma del número de viajes realizados cada día a la zona de salud con el motivo de volver al lugar de residencia
SUM_VIAJES_TRABAJO	FLOAT	Suma del número de viajes realizados cada día a la zona de salud por motivos laborales .
SUM_VIAJES_OTROS	FLOAT	Suma del número de viajes realizados cada día a la zona de salud por motivos no esenciales.

Tabla 5.6: Variables dataset movilidad

Trabajaremos con este dataset de movilidad global usando la herramienta de hojas de cálculo Excel.

Preparación de los datos

Una vez integrado nuestro dataset de movilidad dentro de Excel, se hará uso de los datos de los periodos de referencia obtenidos y los datos disponibles de cada zona de salud, para identificar aquellos valores más altos en las variables **SUM_VIAJES_TOTALES, SUM_VIAJES_CASA, SUM_VIAJES_TRABAJO y SUM_VIAJES_OTROS**, señalándolos con el color rojo y obteniendo una visión más clara de aquellos días donde se detectó una movilidad relativamente alta. Esta identificación de valores de movilidad altos se realiza usando las herramientas de filtrado y búsqueda disponibles en el propio Excel.

A través de la variable **FECHA** se han indicado aquellos días o periodos de tiempo significativos en los que se ha podido dar una mayor o menor movilidad, entre estos eventos destacan las fiestas locales y nacionales, puentes (rosa) y periodos de confinamiento (negrita) de cada una de las zonas de salud.

FECHA	DESTINO	SUM_VIAJES_TOTAL	SUM_VIAJES_CASA	SUM_VIAJES_TRABAJO	SUM_VIAJES_OTROS
14/12	ARANDA DUERO	52857.304	22936.877	3481.331	26439.096
15/12	ARANDA DUERO	56830.503	25516.626	2990.157	28323.719999999998
16/12	ARANDA DUERO	53073.72	23876.869000000006	3087.816	26109.034999999996
17/12	ARANDA DUERO	53781.864	23966.882	3114.1829999999995	26700.799
18/12	ARANDA DUERO	56171.738	23787.571	3418.084	28966.083
19/12	ARANDA DUERO	43242.916000000005	18528.918	3372.72	21341.278000000002
20/12	ARANDA DUERO	33914.369999999995	16382.318	2339.1749999999997	15192.877
21/12	ARANDA DUERO	53999.432	23210.229000000003	3470.2030000000004	27319.0
22/12	ARANDA DUERO	54591.669	23831.867000000002	3319.6609999999996	27440.141000000003
23/12	ARANDA DUERO	54691.99600000001	23423.502	3217.5190000000002	28050.975000000006
24/12	ARANDA DUERO	45722.452	18438.472	3492.413000000001	23791.567000000003
25/12	ARANDA DUERO	33016.134	15489.059000000001	2013.7949999999998	15513.279999999999
26/12	ARANDA DUERO	41216.068	17640.842	3237.4550000000004	20337.771
27/12	ARANDA DUERO	31973.407	15163.451	2328.7380000000003	14481.217999999999
28/12	ARANDA DUERO	46806.215	20054.806	2921.101	23830.307999999997
29/12	ARANDA DUERO	45921.373	18592.18	2980.9329999999995	24348.260000000002
30/12	ARANDA DUERO	49593.049000000006	21170.158000000003	2922.577	25500.314000000002
31/12	ARANDA DUERO	42839.515	17398.982	3173.6670000000004	22266.866
01/07	MIRANDA EBRO	95895.589	42407.626	2668.86	50819.10299999999
02/07	MIRANDA EBRO	96610.44499999999	41898.352	2732.626	51979.467
03/07	MIRANDA EBRO	93983.59	40528.047	2524.1459999999997	50931.397
04/07	MIRANDA EBRO	83225.30700000002	35479.138	2412.4040000000005	45333.765
05/07	MIRANDA EBRO	74416.668	35263.625	1979.77	37173.273
06/07	MIRANDA EBRO	88128.937	39646.672000000006	2370.929	46111.335999999996

Figura 5.13: Fragmento de dataset final de movilidad creado en Excel

Capítulo 6

Análisis de datos

En este capítulo se aborda el análisis realizado a los datos obtenidos para el proyecto. Dicho análisis está dividido entre los tres tipos de datos principales utilizados: salud, medidas aplicadas y movilidad. Se muestran las gráficas generadas y aquellas conclusiones obtenidas de cada análisis.

6.1. Análisis datos salud

Se realiza un análisis de aquellos datos de salud obtenidos para la adquisición de una imagen global y preliminar que nos permita apreciar la evolución de la pandemia en cada una de las zonas estudiadas.

Gracias a la generación de distintas gráficas, se han analizado las variables obtenidas de las zonas de salud estudiadas. En las gráficas obtenidas se ha representado a través del eje X la fecha correspondientes a cada dato y en el eje Y los valores de cada una de las variables estudiadas. Las gráficas mostradas reflejarán los datos de cada zona a través de colores, como se muestra en la Figura 6.1.

Figura 6.1: Leyenda de colores asignados a cada zona

6.1.1. Carpeta imágenes gráficas

Debido al gran número de días estudiados en cada gráfica y al limite de tamaño disponible para las mismas en la memoria de este proyecto, se facilita una carpeta ligada a este documento, donde se ubican todas las gráficas usadas en alta resolución para una correcta lectura y comprensión por parte del lector.

Las organización de las imágenes dentro de esta carpeta seguirá la misma estructura que la vista en este capítulo, teniendo cada imagen el nombre de la Figura a la que corresponden.

6.1.2. Datos diarios

A continuación se muestran las gráficas obtenidas para cada variable diaria:

PCR positivas: Se ha generado una gráfica del número de pruebas PCR diarias realizadas con un resultado positivo, es decir, personas contagiadas

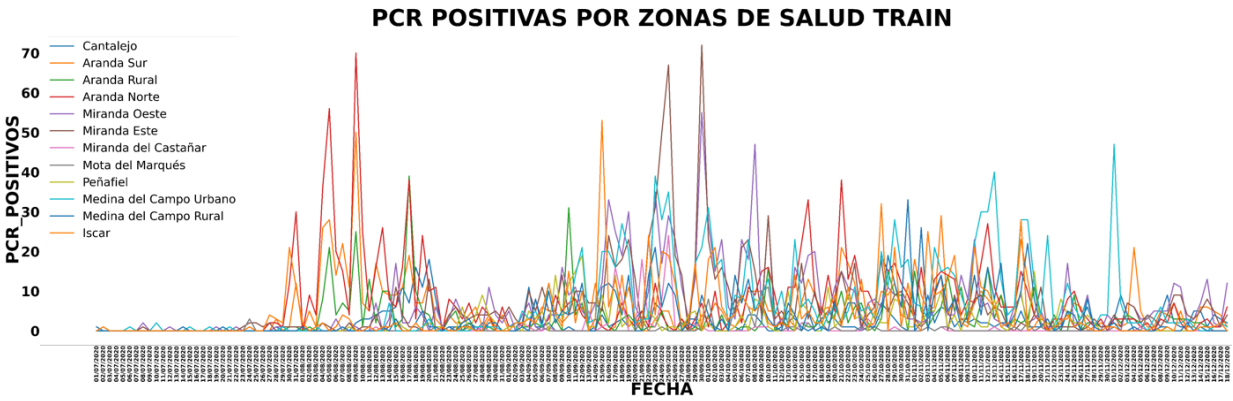


Figura 6.2: Gráfica de PCR positivas realizadas en las zonas de salud estudiadas

Observando la gráfica de la Figura 6.2 acerca del número de positivos detectados en cada zona de salud, se aprecia como a partir de agosto en algunas zonas como Aranda de Duero o Cantalejo, se comienza a detectar un alto número de contagios con un posterior descenso, posiblemente debido a las distintas medidas de confinamiento aplicadas.

En el caso de los meses posteriores, se distingue el crecimiento de una gran curva de contagios en todas las zonas de salud, llegando a su máximo en septiembre y octubre, meses en lo cuales se aplicaron la mayoría de confinamientos y medidas dentro de la Comunidad. El efecto de dichas medidas es apreciable en los meses posteriores, donde se observa un claro descenso de las curvas en la mayoría de zonas de salud. En zonas como Medina del Campo Rural o Aranda Sur esta curva descendió más lentamente.

PCR realizadas: Se ha obtenido una gráfica del número de pruebas PCR diarias realizadas en total para la detección del virus en aquellas personas infectadas.

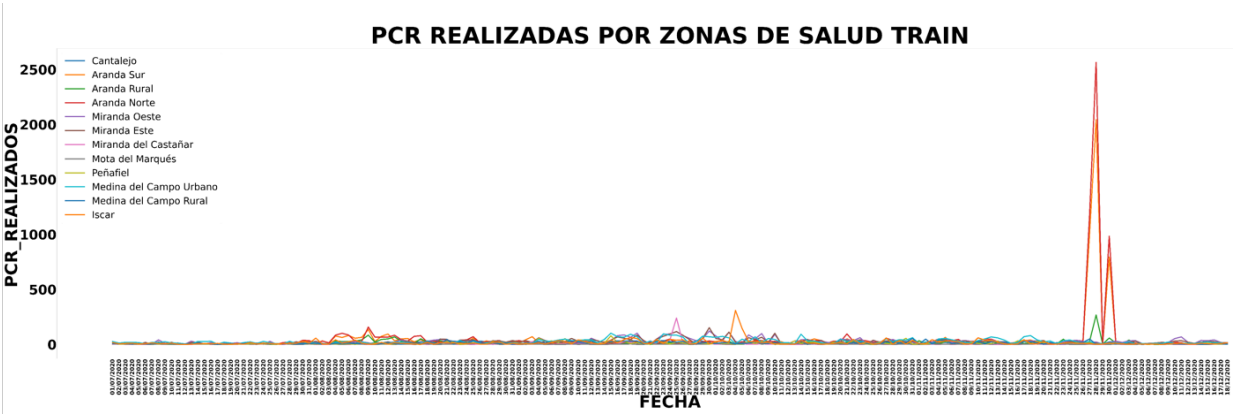


Figura 6.3: Gráfica de PCR diarias realizadas en las zonas de salud estudiadas

Tal y como se puede apreciar a primera vista en la Figura 6.3, se observan dos grandes picos correspondientes a los cribados masivos realizados en las zonas de salud de Aranda de Duero por su alta incidencia.

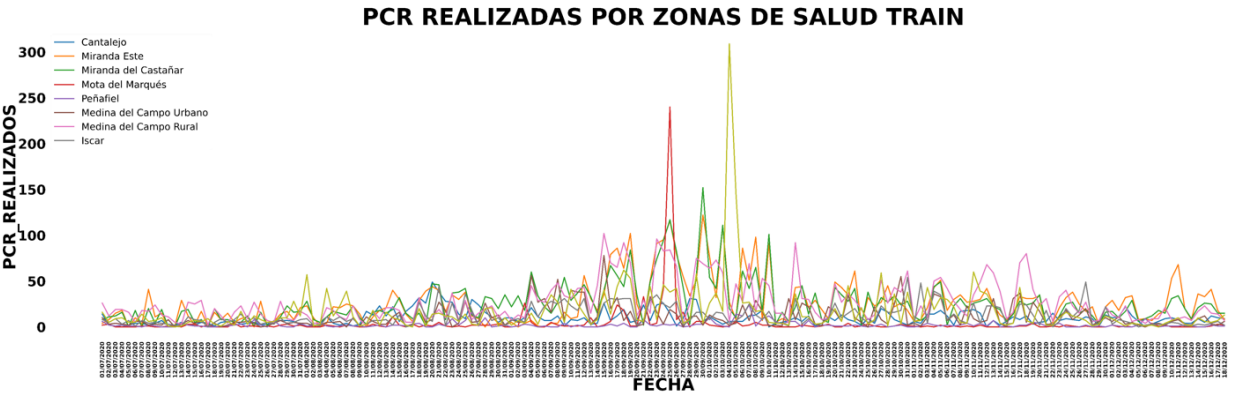


Figura 6.4: Gráfica de PCR diarias realizadas en zonas de salud sin cribados masivos

Si se descartan aquellas zonas en las cuales se realizaron los cribados masivos, se obtiene una gráfica (Figura 6.4) en la que se aprecia con más detalle el número de PCR realizadas. Así, se puede apreciar que en la mayoría de zonas el número de pruebas realizadas aumentó en los meses de agosto, septiembre y octubre (meses donde se sitúan la mayoría de confinamientos aplicados).

Fallecidos: Gráfica del número de personas fallecidas diariamente a causa del COVID-19 o con síntomas semejantes.

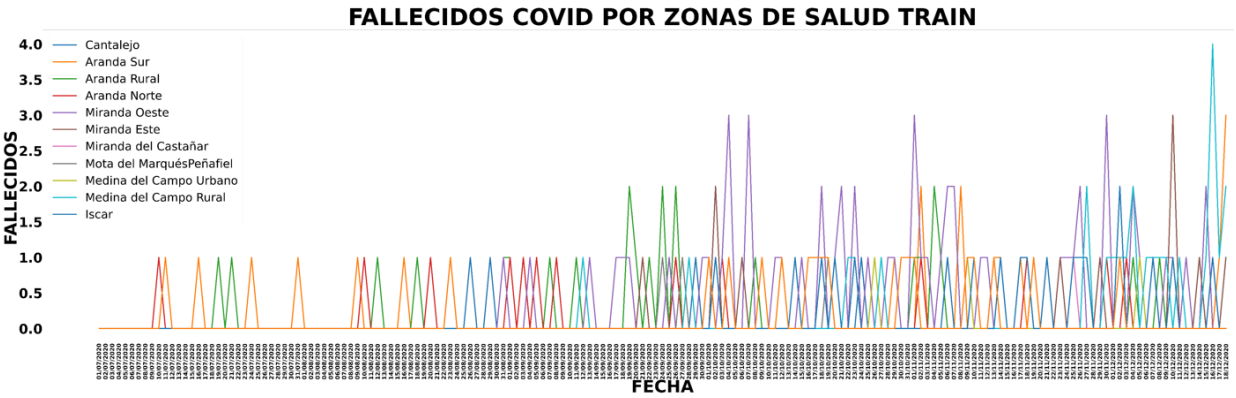


Figura 6.5: Gráfica de fallecidos diarios por COVID-19 en las zonas de salud estudiadas

El número de fallecidos aumentó durante los meses posteriores a agosto, localizándose este incremento en zonas como Aranda, Miranda de Ebro o Medina, donde el número de habitantes es superior al resto de zonas de salud estudiadas.

Prevalencia: Obtención del número de personas que siguen siendo compatibles con el virus cada día.

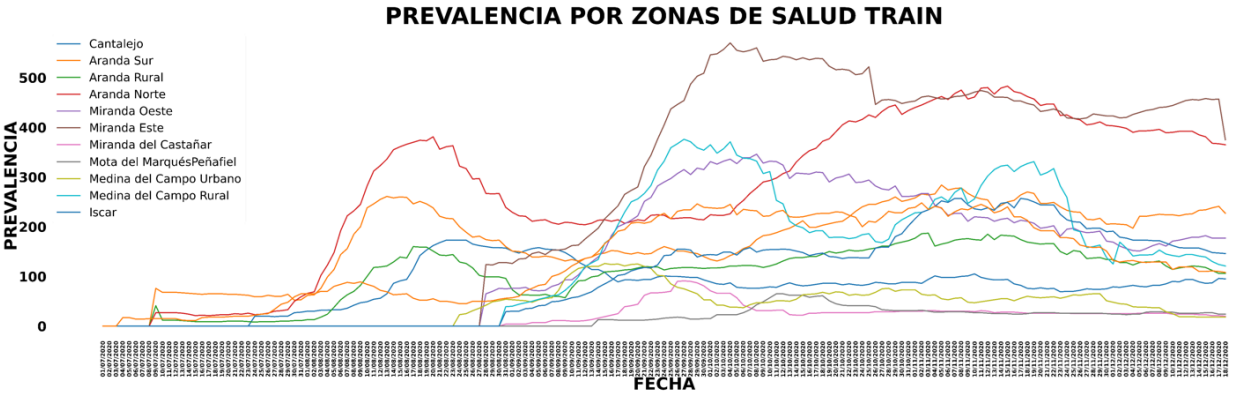


Figura 6.6: Gráfica de prevalencia en las zonas de salud estudiadas

Tal y como se puede apreciar en la Figura 6.6, la prevalencia era baja al principio de la *segunda ola* en todas las zonas de salud por la cercanía al fin de la *primera ola*. A medida que se progresaba en el tiempo esta prevalencia crecía, teniendo unos valores muy altos desde mediados de septiembre.

Problemáticas:

Uno de los factores que no se ha tenido cuenta a la hora de analizar estos datos diarios, es la población en cada una de las zonas estudiadas. Debido a la alta población de zonas como Aranda de Duero, Medina del Campo o Miranda de Ebro, Figura 6.7, se podría deducir que dichas zonas han sido las más afectadas, algo que puede llegar a distorsionar la realidad. Es por ello, por lo que se debe tener en cuenta el factor población en todas las zonas para determinar el impacto en función al número de personas que residen en cada lugar.

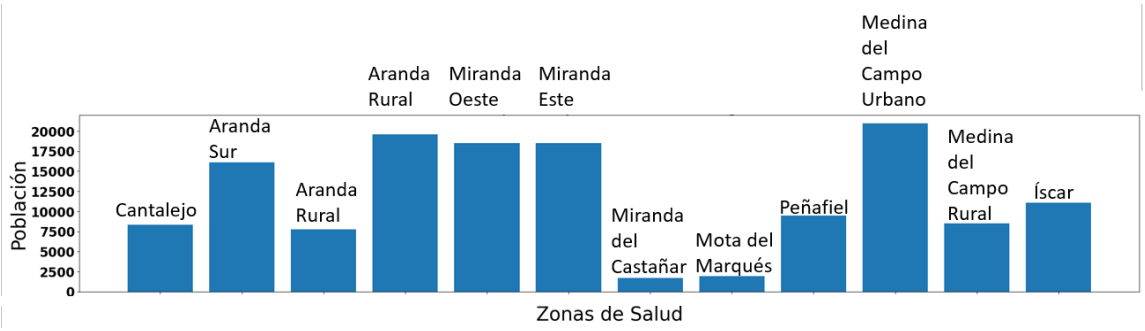


Figura 6.7: Gráfica de población por tarjeta sanitaria en cada zona de salud

Otro de los inconvenientes que se encuentra al analizar los datos diarios, es la forma de las gráficas obtenidas, las cuales dificultan la labor de análisis y obtención de una tendencia que permita averiguar de manera sencilla e intuitiva las subidas y bajadas generales en cada una de las zonas (gran número de picos).

En el siguiente apartado se realizará un análisis gráfico de aquellas variables, como la incidencia acumulada o el número de PCR positivas por cada realizada, que ofrecen una forma de medición de los datos a través de distintos periodos de tiempo coincidentes con el ratio de actuación del virus y la población de cada zona.

6.1.3. Datos periódicos

Las variables analizadas en cada uno de los periodos (ya descritas en el apartado 5.3.3), son las siguientes:

- **IA PCR positivas:**Incidencia acumulada de pruebas PCR realizadas con resultado positivo en un periodo de tiempo determinado, siendo la variable **POSI_IA_XDIAS**
- **IA PCR Realizadas:**Incidencia acumulada de pruebas PCR realizadas en un periodo de tiempo determinado, siendo la variable **REALI_IA_XDIAS**

- **IA fallecidos:**Incidencia acumulada de fallecidos a causa del COVID-19 en un periodo de tiempo determinado. siendo la variable **FALL_IA_XDIAS**
- **Porcentaje PCR:** Porcentaje de positivos obtenido respecto a las PCR realizadas en un periodo determinado de tiempo, siendo la variable **PORC_PCR_XDIAS**.

IA PCR positivas

IA 4 días PCR positivas

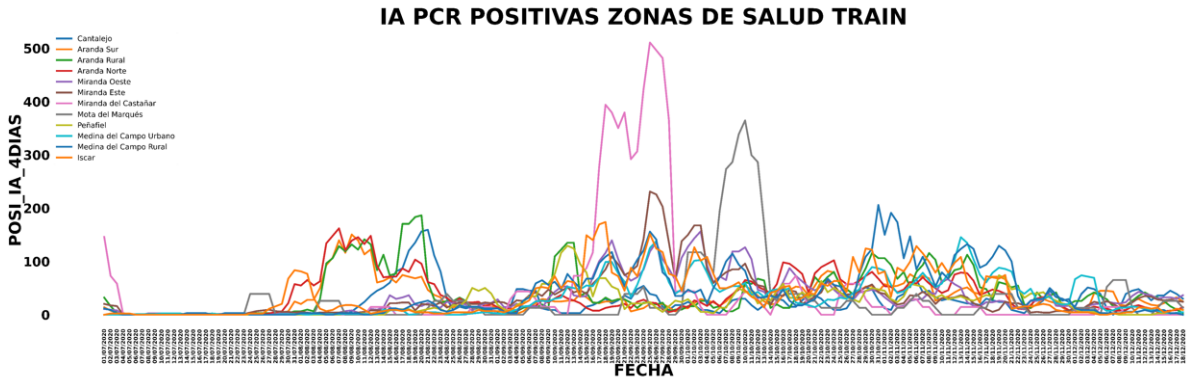


Figura 6.8: Gráfica IA 4 dias PCR positivas

IA 7 días PCR positivas

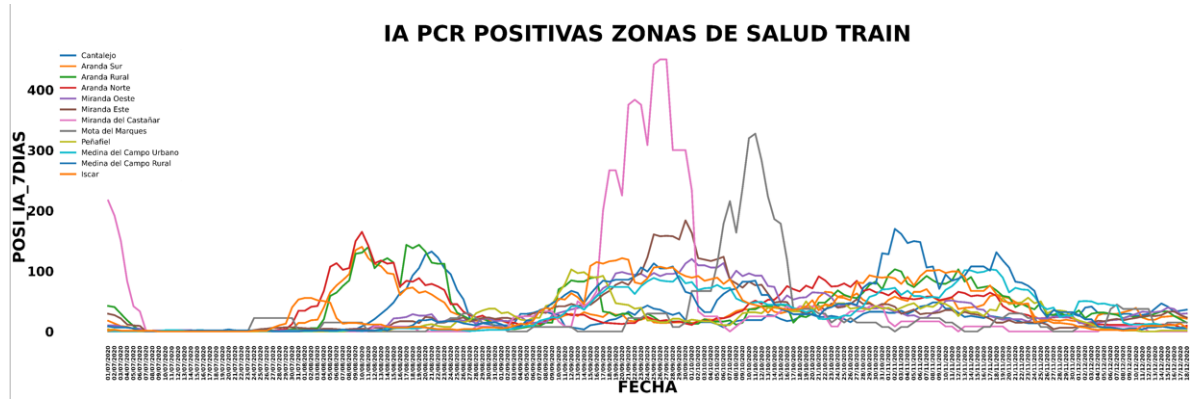


Figura 6.9: Gráfica IA 7 dias PCR positivas

IA 14 días PCR positivas

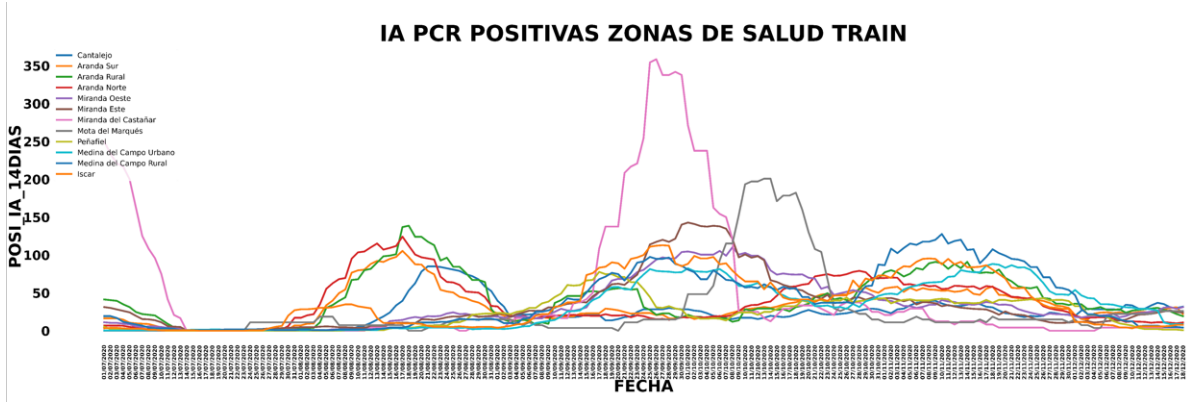


Figura 6.10: Gráfica IA 14 dias PCR positivas

Una vez obtenidas las gráficas correspondientes a la Incidencia Acumulada en distintos periodos del número de casos de COVID-19 detectados, se observa con una mayor claridad la tendencia y forma de las distintas curvas de contagios por cada zona. En las Figuras 6.8, 6.9, 6.10, se puede apreciar de forma clara las distintas subidas y bajadas existentes dentro de la *segunda ola*.

Se aprecia como durante el mes de agosto, el número de contagios comenzó a aumentar en zonas como Aranda de Duero y Cantalejo, donde se aplicaron por ello los primeros confinamientos y se logró una bajada de la curva como consecuencia de dichas medidas.

A partir de septiembre se observa un gran pico de casos detectados en la mayoría de las zonas estudiadas, destacando aquellas con menor población como Miranda del Castañar o Mota del Marqués. Dicha subida se normaliza en octubre, donde en la mayoría de zonas se aprecia una bajada con una ligera subida en el mes posterior, para descender finalmente en diciembre de manera general.

Cabe destacar que dichos aumentos en la curva de contagios se corresponden con la aplicación *a posteriori* de los diferentes confinamientos y medidas estudiados.

PCR Realizadas

IA 4 días PCR realizadas

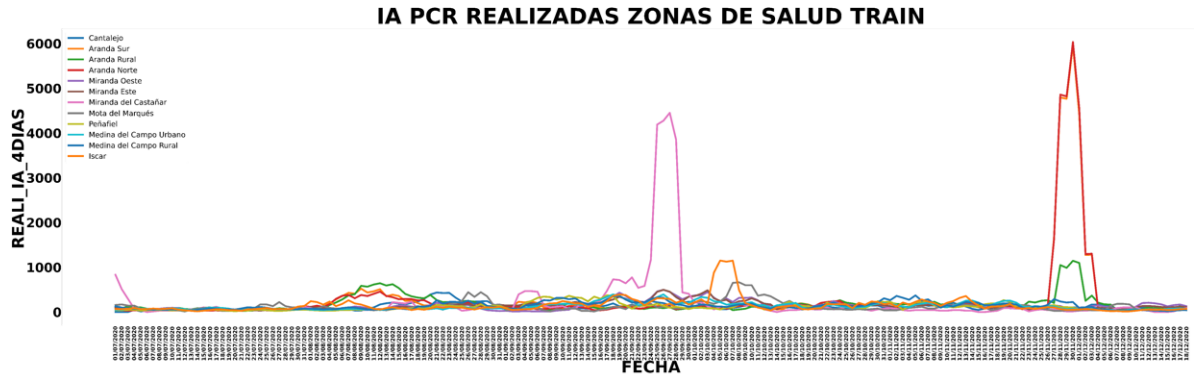


Figura 6.11: Gráfica IA 4 días PCR realizadas

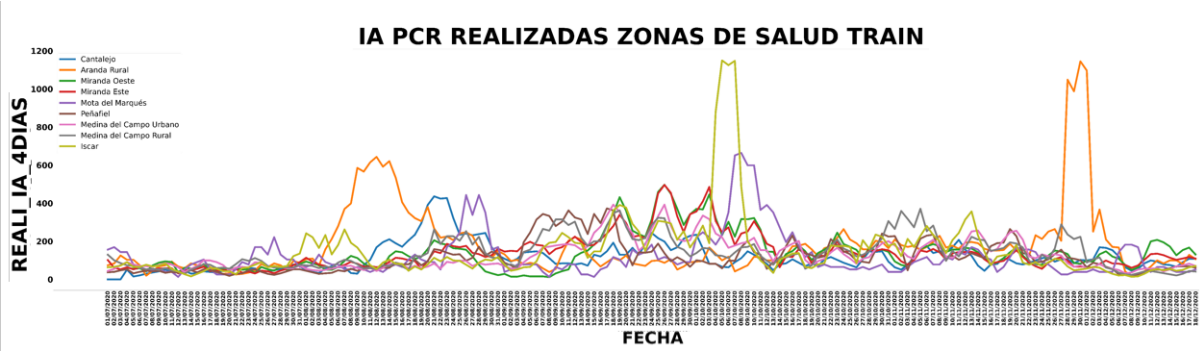


Figura 6.12: Gráfica IA 4 días PCR realizadas sin cribados

IA 7 días PCR realizadas

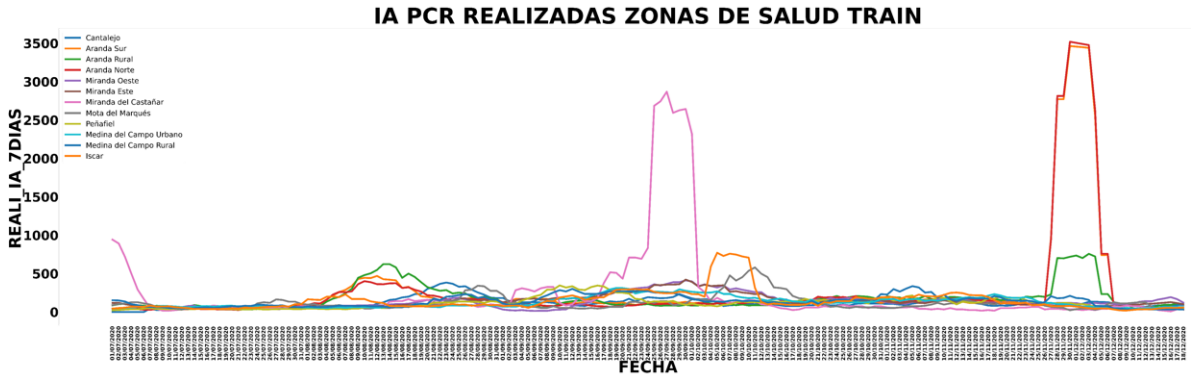


Figura 6.13: Gráfica IA 7 días PCR realizadas

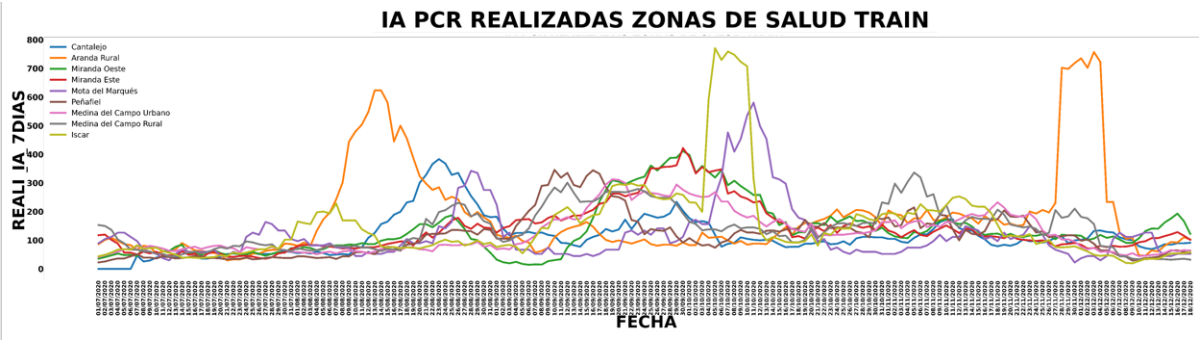


Figura 6.14: Gráfica IA 7 días PCR realizadas sin cribados

IA 14 días PCR realizadas

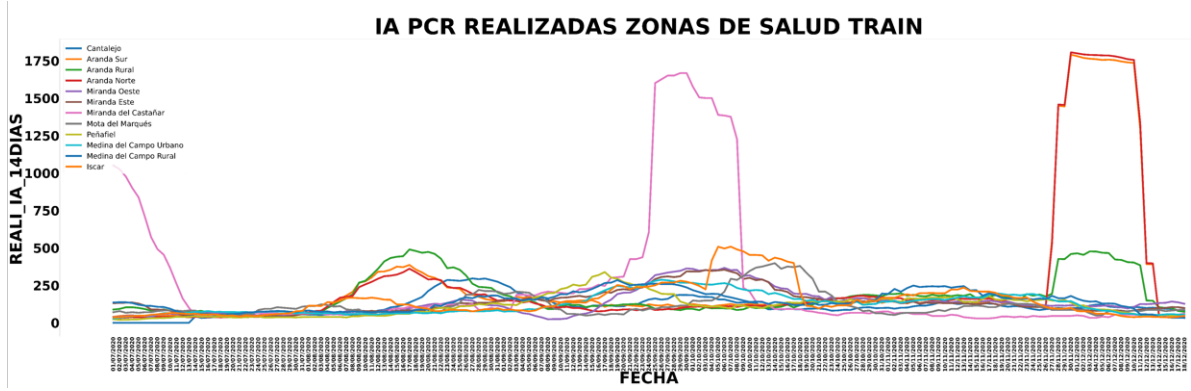


Figura 6.15: Gráfica IA 14 días PCR realizadas

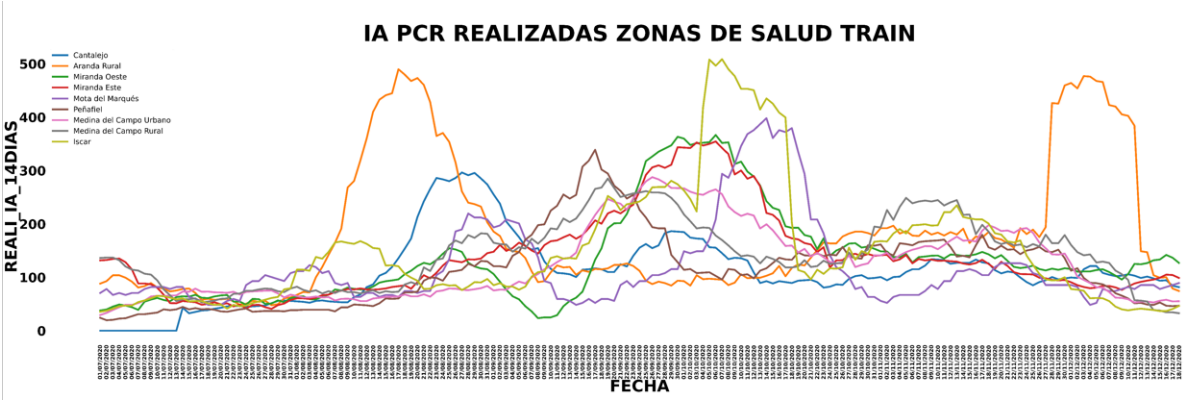


Figura 6.16: Gráfica IA 14 días PCR realizadas sin cribados

Observando las gráficas obtenidas sobre la IA del número de pruebas realizadas, se detecta como en las zonas de Miranda del Castañar o Aranda de Duero Norte y Sur destacan grandes picos correspondientes a los cribados masivos realizados en dichas zonas. Estos cribados fueron realizados durante los meses de septiembre y diciembre, más críticos, para doblegar la curva de contagios. En el caso de Miranda del Castañar, su cribado detectó un gran número de infectados, algo que podemos validar con las gráficas obtenidas. En el caso de Aranda de Duero el cribado masivo realizado en diciembre no detectó tantos contagios.

Para el análisis de la IA de número de pruebas PCR realizadas, se ha generado un tipo de gráfica en cada periodo donde quedan eliminadas aquellas zonas que realizaron cribados masivos (Figuras 6.12, 6.14, 6.16). En estas gráficas se aprecia como a partir de agosto se realizan más pruebas en zonas afectadas como Cantalejo o Aranda. En los meses posteriores de septiembre y octubre aumentó el número de pruebas en el resto de zonas, coincidiendo con la gran curva de contagios observada en el caso de la IA de PCR positivas ya analizada. Esto es algo normal, ya que sin la realización de pruebas no se podrían haber detectado la mayoría de casos activos.

IA fallecidos

IA 4 días fallecidos

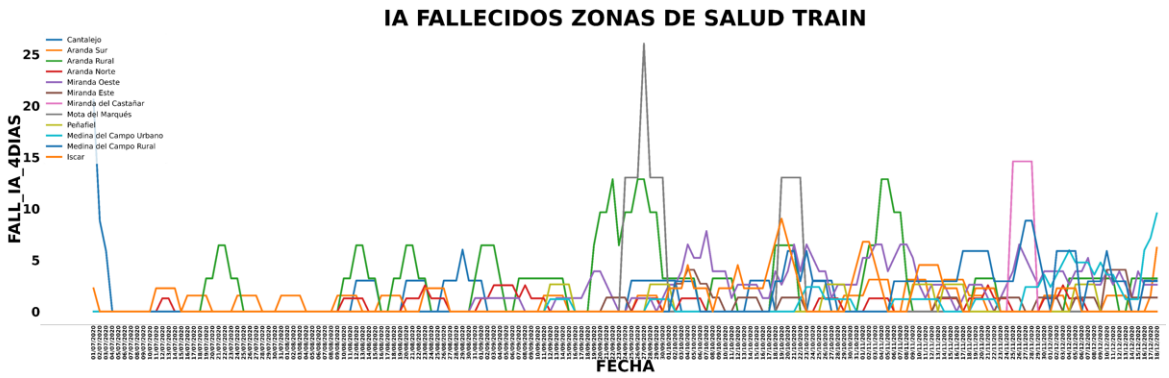


Figura 6.17: Gráfica IA 4 dias fallecidos por COVID-19

IA 7 días fallecidos

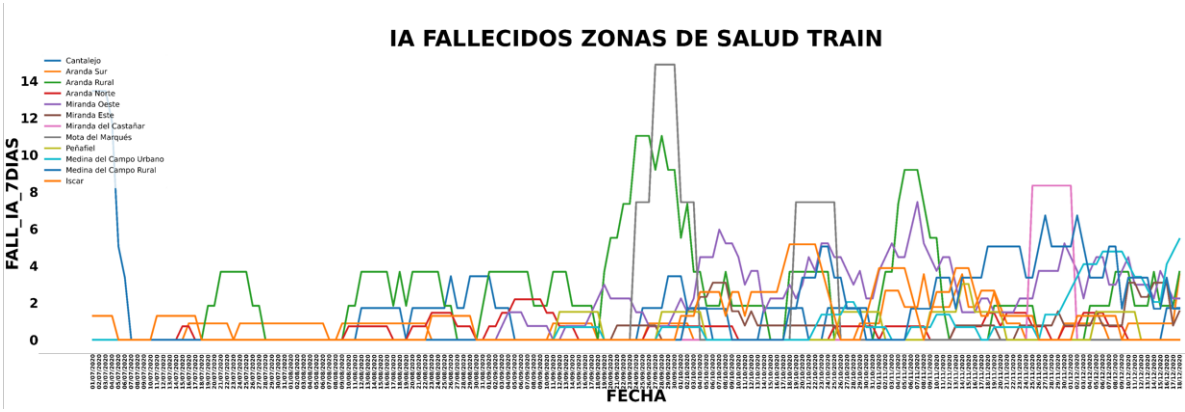


Figura 6.18: Gráfica IA 7 días fallecidos por COVID-19

IA 14 días fallecidos

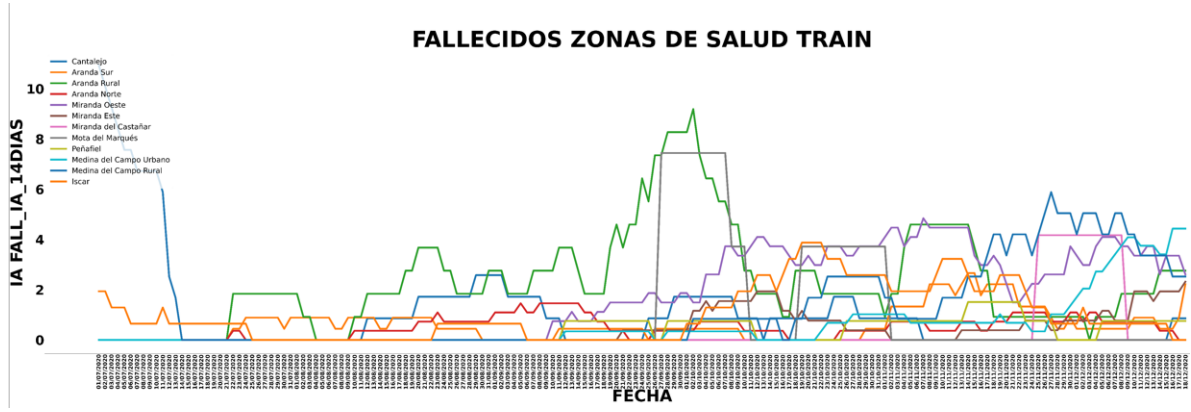


Figura 6.19: Gráfica IA 14 días fallecidos por COVID-19

Analizando las gráficas obtenidas respecto a la IA de número de fallecidos por COVID en los distintos periodos, destaca un gran aumento a mediados de septiembre en zonas como Aranda Rural o Mota del Marqués (ambas zonas rurales con una población envejecida). En el resto de zonas esta tendencia se incrementa en menor medida en los meses posteriores, manteniéndose a lo largo de toda la *segunda ola*. Algunas zonas como Peñafiel, Aranda Norte o Miranda Este muestran una incidencia muy baja.

Porcentaje PCR

Porcentaje PCR 4 días

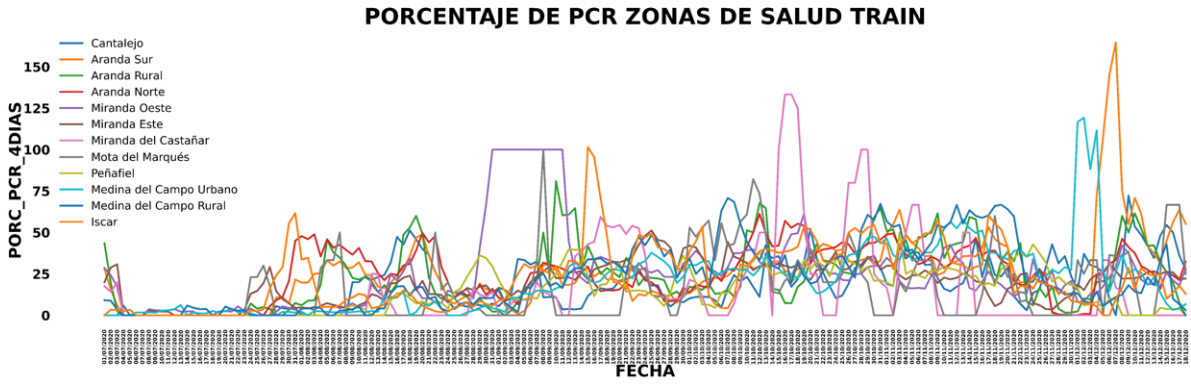


Figura 6.20: Gráfica porcentaje PCR 4 días

Porcentaje PCR 7 días

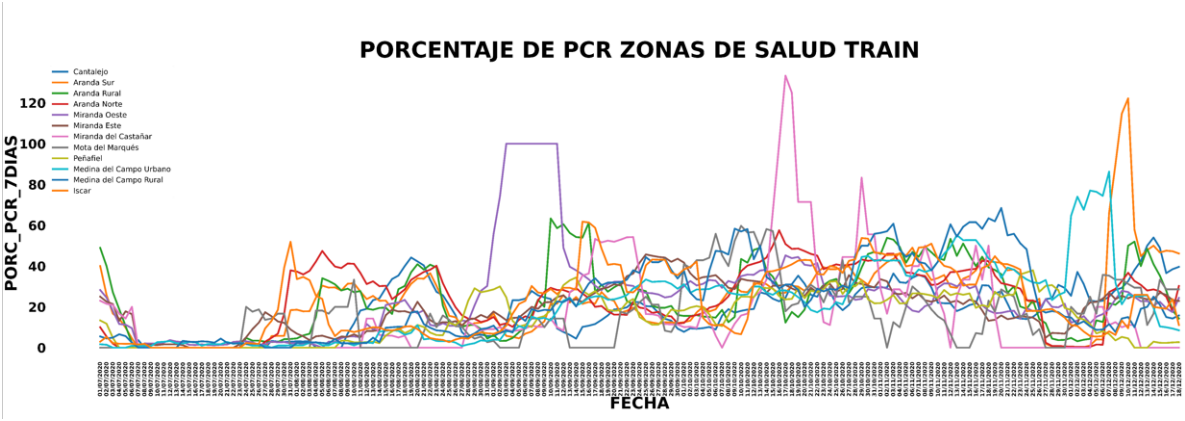


Figura 6.21: Gráfica porcentaje PCR 7 días

Porcentaje PCR 14 días

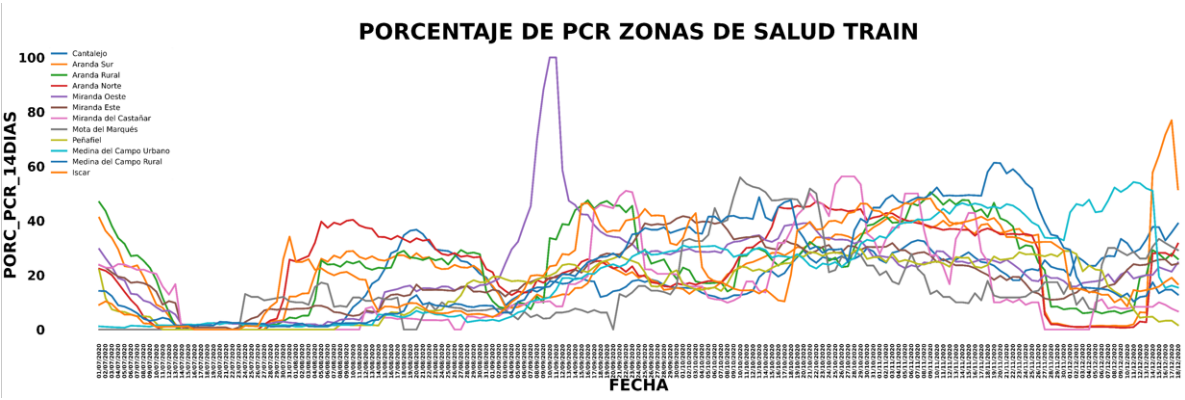


Figura 6.22: Gráfica porcentaje PCR 14 días

Tal y como se ha visto en los análisis de la IA del número de pruebas PCR realizadas (Figuras 6.12, 6.14, 6.16) y la IA de PCR positivas (Figuras 6.8, 6.9, 6.10), ambas gráficas están estrechamente relacionadas, pues la bajada o subida del número de pruebas realizadas repercute directamente en el número de casos positivos, y por tanto, en la correcta detección de la verdadera incidencia de cada zona. Es por ello por lo que se realiza un análisis de la variable **PORC_PCR_XDIAS** para así obtener de forma más precisa la evolución de la pandemia en cada una de las zonas y relacionarla con las conclusiones obtenidas anteriormente.

Observando las gráficas generadas en los distintos periodos de tiempo (Figuras 6.20,

6.21, 6.22),se puede apreciar como la curva de contagios correspondiente a esta *segunda ola* comenzó a finales de julio y principios de agosto en las zonas de Aranda de Duero , Cantalejo e Íscar (en menor medida), descendiendo ligeramente a final de mes. Durante los meses de septiembre, octubre y noviembre en la mayoría de las zonas dicha curva de contagios fue creciendo progresivamente hasta alcanzar grandes picos en zonas como Miranda Oeste. Finalmente, en diciembre dicha curva comenzó a descender en mayor o menor medida en la mayoría de zonas (a excepción de Aranda Norte y Cantalejo).

De este modo todas las gráficas estudiadas guardan una semejanza entre si en todas las representaciones anteriores. Dicha relación se basa en la detección de una gran incidencia y un posterior confinamiento en la mayoría de las zonas durante septiembre, octubre y noviembre. Podemos observar cómo aparentemente todas estas medidas aplicadas lograron reducir los contagios a partir del comienzo del invierno.

Como conclusión, destacar cómo la gráfica correspondiente al periodo de 14 días (Figuras 6.10, 6.16, 6.19 y 6.22)permiten apreciar con mayor claridad la efectividad de las medidas aplicadas para doblegar la curva de contagios, pues es necesario un tiempo mínimo de 14 días para observar los efectos de las medidas aplicadas en la tendencia de número de contagios.

6.2. Análisis datos movilidad

Al igual que en los datos de salud, se ha realizado un análisis de los datos de movilidad, para observar la tendencia de esta en los distintos periodos correspondientes a la *segunda ola*.

Gracias a los datos de movilidad de cada zona hemos podido determinar el tipo de actividad de cada una de estas. Dicho estudio se ha realizado analizando los tres tipos de movilidad obtenidos anteriormente, siendo estos las entradas por vuelta a la residencia habitual, por trabajo y por otros motivos.

Para este estudio las zonas de Aranda de Duero (Sur, Rural y Norte), Miranda de Ebro (Este y Oeste) o Medina del Campo (Urbano y Rural) se han encapsulado dentro de cada distrito para su análisis.

A continuación se muestra un resumen con los resultados y clasificación dada a cada zona en función de los datos de movilidad, mostrándose también los datos función del porcentaje de población:

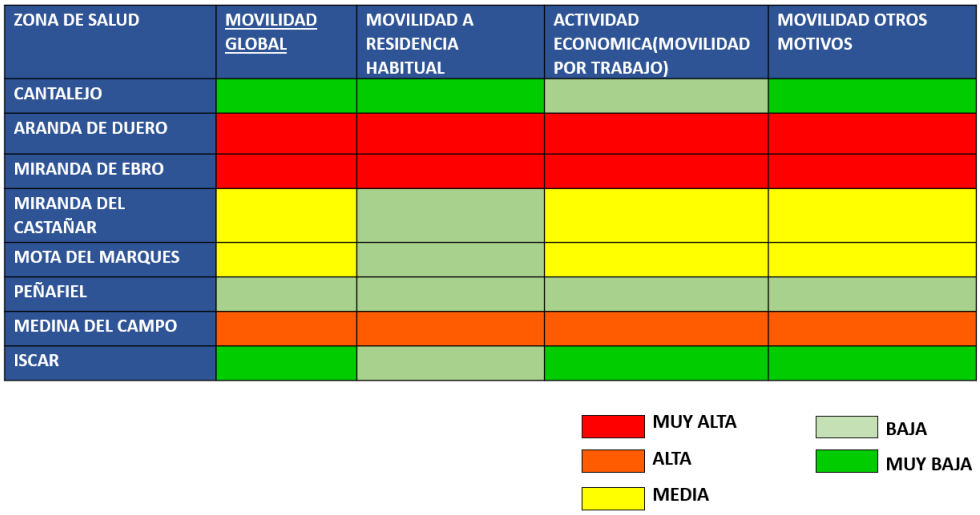


Figura 6.23: Panel de clasificación de la movilidad de cada zona de salud

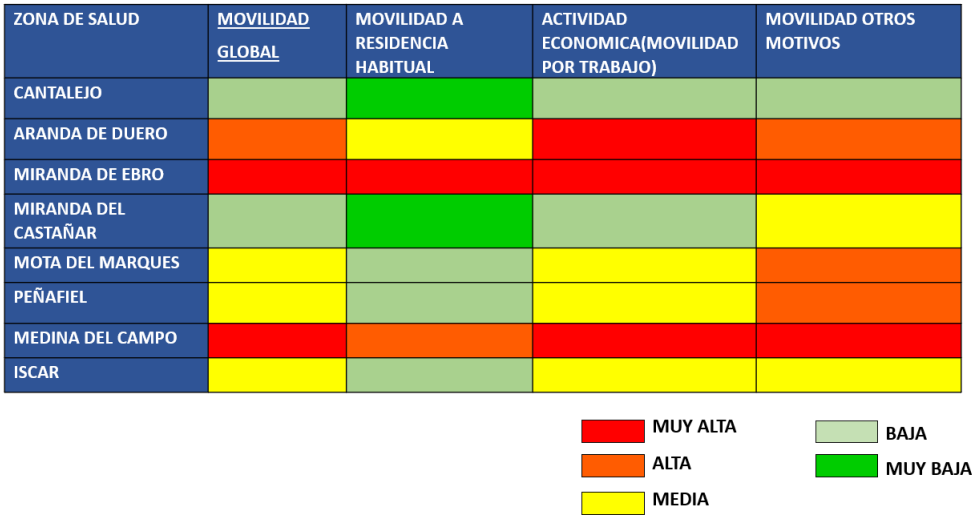


Figura 6.24: Panel de clasificación en función de la población de la movilidad de cada zona de salud

En las Figuras 6.23 y 6.24, se puede apreciar como las zonas de Aranda de Duero, Medina del Campo y Miranda de Ebro poseen una alta movilidad respecto al resto de zonas estudiadas. Esto puede traducirse en una mayor incidencia en el número de contagios o en la aplicación de medidas más restrictivas. Es por ello por lo que estudiaremos mediante diferentes gráficas la movilidad en cada una de las zonas usadas como caso de estudio, clasificándolas en dos grupos para un mejor análisis de las mismas:

- **Zonas de alta movilidad:** zonas donde se ha detectado una alta movilidad en todos los tipos estudiados y que corresponden a Aranda de Duero, Medina del Campo y Miranda de Ebro.

- **Zonas de movilidad intermedia o baja:** zonas donde se ha detectado una movilidad intermedia o baja y que son analizadas conjuntamente para una mayor claridad. Estas zonas corresponden a Cantalejo, Miranda de Castañar, Mota del Marqués, Peñafiel e Íscar.

Al igual que en apartado anterior, las gráficas obtenidas de cada zona de salud se han representado a través del eje X, donde se establecen las fechas correspondientes a cada dato, y el eje Y, donde se especifican el número de viajes de cada uno de los tipos de movilidad estudiados. Las gráficas mostradas reflejarán los datos de cada zona de salud a través de colores, siendo cada color el de una zona tal y como se muestra en las Figuras 6.25 y 6.26.

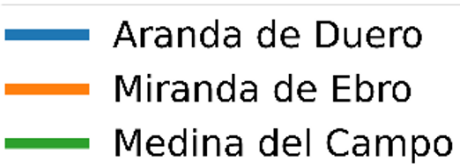


Figura 6.25: Leyenda de las gráficas de zonas de alta movilidad

Figura 6.26: Leyenda de las gráficas de zonas de media y baja movilidad

6.2.1. Movilidad de entradas

Zonas de alta movilidad

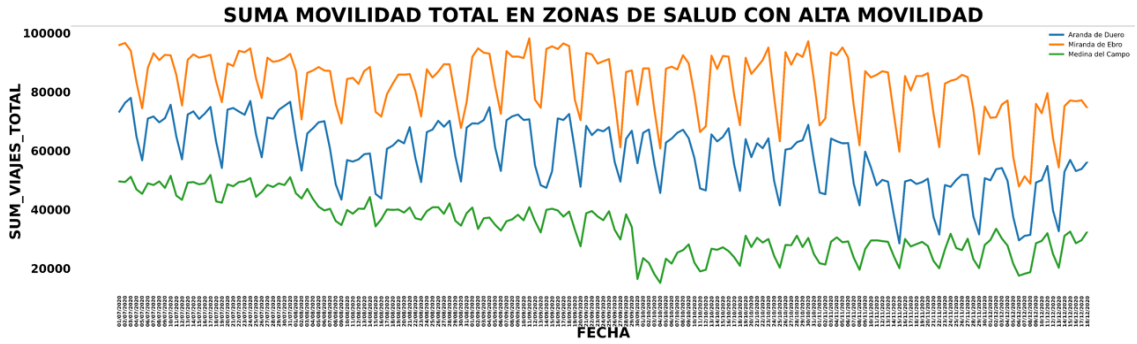


Figura 6.27: Gráfica de entradas totales a zonas de alta movilidad

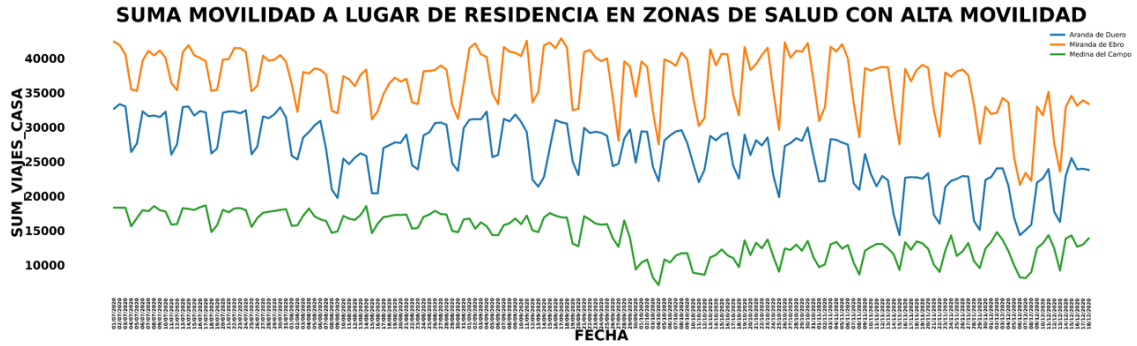


Figura 6.28: Gráfica de entradas a residencia habitual en zonas de alta movilidad

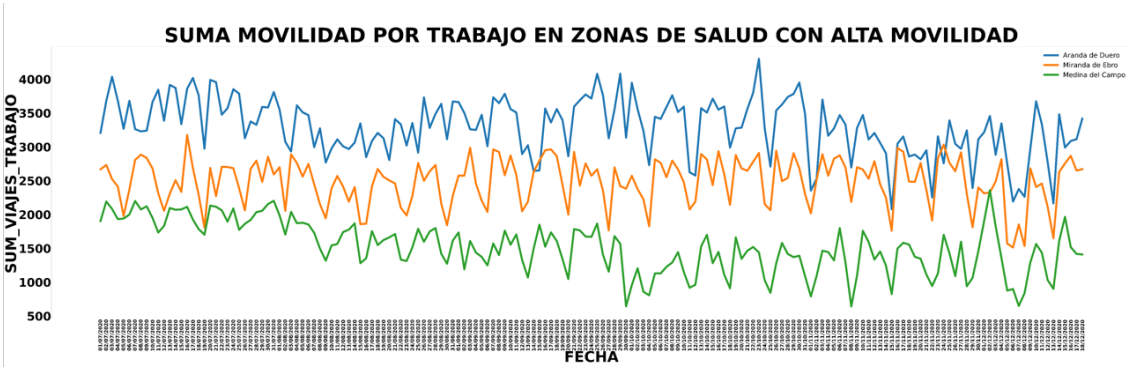


Figura 6.29: Gráfica de entradas por trabajo a zonas de alta movilidad

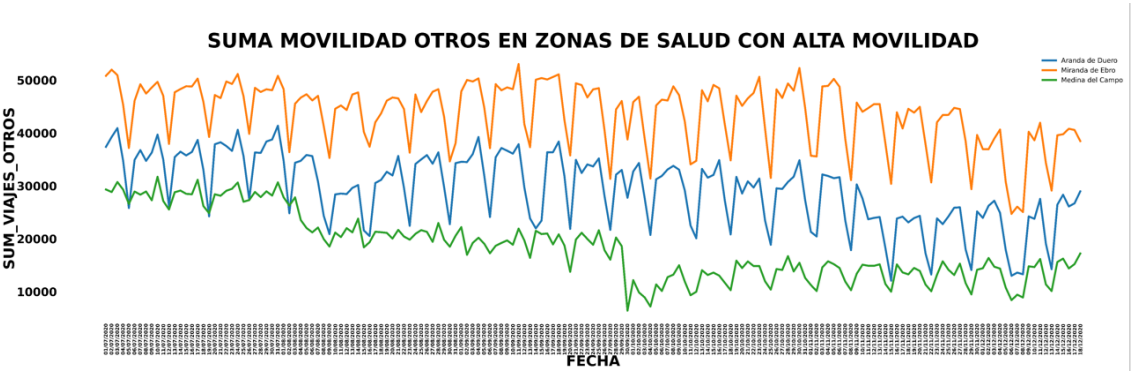


Figura 6.30: Gráfica de entradas por otros motivos a zonas de alta movilidad

Destacar que las zonas de Miranda de Ebro y Aranda de Duero poseen una mayor movilidad respecto a Medina del Campo. En estos dos primeros distritos se puede ver como la tendencia de la movilidad total (Figura 6.27) se mantiene a lo largo de la *segunda ola*, descendiendo ligeramente en agosto, septiembre y noviembre, con aquellos confinamientos o restricciones de movilidad aplicados. Esta tendencia constante en la movilidad se debe principalmente a que muchas de las vías de comunicación más relevantes del país atraviesan dichas zonas, por lo que la movilidad durante todo el año es intensa. El número de entradas por vuelta al lugar de residencia (Figura 6.28) aumentó en el mes de septiembre como consecuencia del fin de las vacaciones de verano e inicio del curso escolar. En el caso de la movilidad por trabajo (Figura 6.29) o por otros motivos (Figura 6.30), ambas se mantuvieron constantes a lo largo de la *segunda ola* obteniendo ciertas bajadas posiblemente como consecuencia de los confinamientos mencionados.

En la zona de Medina del Campo se observa como el confinamiento realizado a finales del mes de septiembre tuvo un gran impacto en la movilidad. Dicho confinamiento provocó un descenso en todos los tipos de entradas manteniéndose dicha tendencia baja

hasta diciembre, donde se detectó una ligera subida sobre todo en aquella movilidad por trabajo. Es curioso observar como los picos detectados al analizar los contagios en diciembre (Figura 6.22), coinciden con este aumento de la movilidad por trabajo en la zona de Medina.

Finalmente se puede concluir que aquellas zonas clasificadas como zonas de movilidad alta sufrieron descensos en el número de entradas como consecuencia de los confinamientos y medidas realizados. Es posible apreciar, realizando una comparación con las gráficas de Incidencia Acumulada (apartado 6.1.3), como estas zonas se corresponden con aquellas curvas detectadas como altas. Por tanto se puede deducir que la movilidad es un factor importante en cuanto al número de contagios y medidas aplicadas.

Zonas de movilidad intermedia o baja

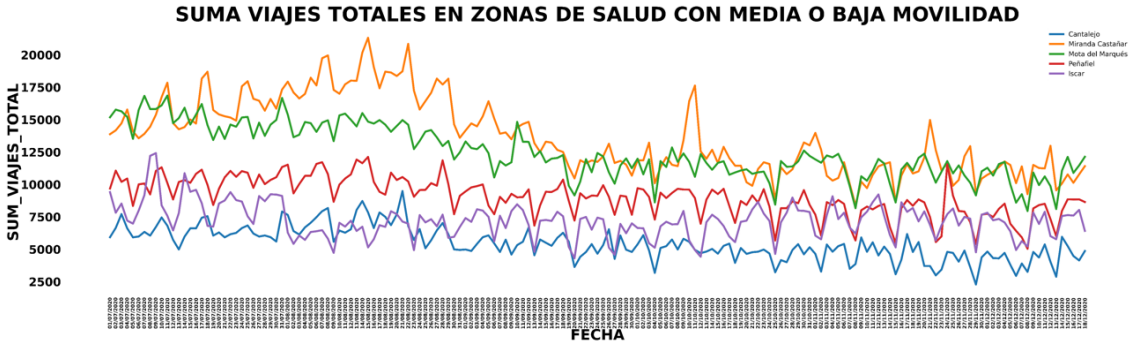


Figura 6.31: Gráfica de entradas totales a zonas de alta movilidad

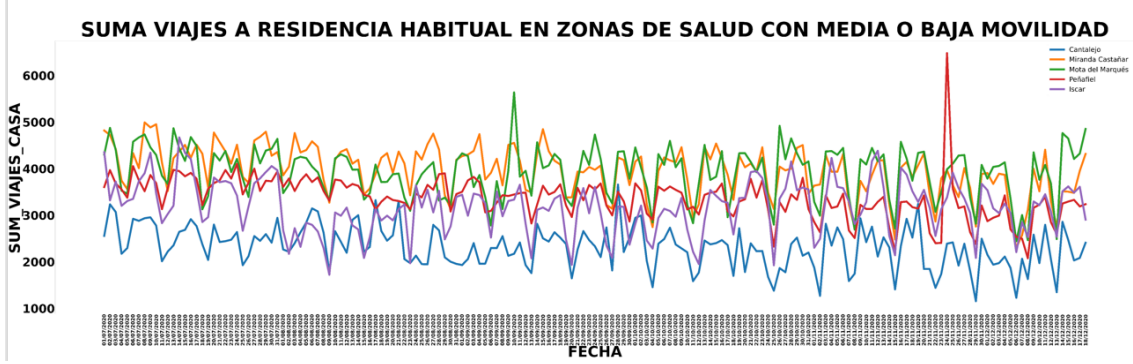


Figura 6.32: Gráfica de entradas a residencia habitual en zonas de alta movilidad

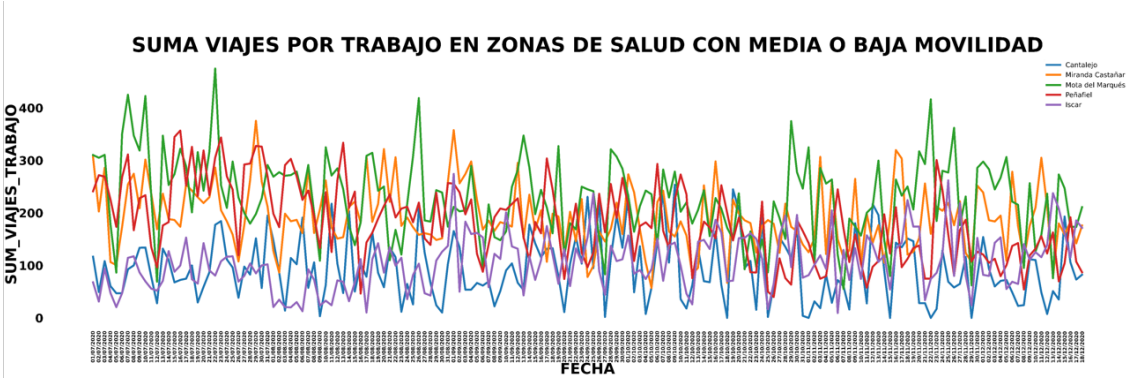


Figura 6.33: Gráfica de entradas por trabajo a zonas de alta movilidad

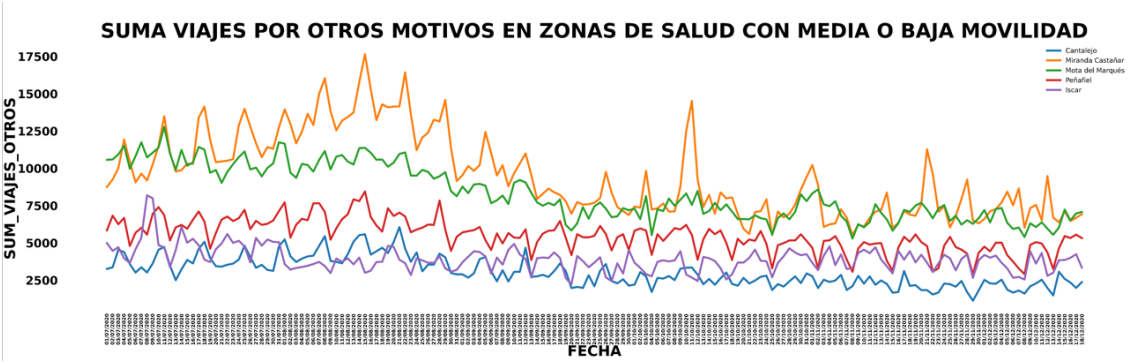


Figura 6.34: Gráfica de entradas por otros motivos a zonas de alta movilidad

Observando la gráfica de aquellas zonas con un número de entradas inferior al visto anteriormente en Aranda, Miranda de Ebro y Medina, se puede apreciar como en este caso la movilidad total (Figura 6.31) en dichas zonas se ve reducida a partir de finales de septiembre y principios de octubre. Este periodo coincide con muchos de los confinamientos aplicados a dichas zonas con el objetivo de bajar la incidencias de contagios vistas en el apartado 6.1.3. La tendencia baja en la movilidad se mantiene en la mayoría de las zonas estudiadas hasta el fin de la *segunda ola*.

Analizando el número de entradas por vuelta a la residencia habitual (Figura 6.32), en la mayoría de las zonas este tipo de movilidad muestra una tendencia constante, exceptuando Íscar donde en la primera mitad de agosto experimentó una bajada en este tipo de movilidad.

En el caso de la movilidad por trabajo (Figura 6.33), se han detectado grandes variaciones y picos que impiden obtener una tendencia clara de las entradas realizadas. Pese a este obstáculo, si que se ha podido visualizar en zonas como Cantalejo, Miranda del

Castañar o Mota del Marqués, cierta periodicidad en los datos obtenidos, lo que nos indica que la movilidad por trabajo es constante y se realiza siempre en los mismo días de la semana. A su vez, en las zonas de Íscar y Peñafiel, la movilidad también ofrece una tendencia más clara de las subidas y bajadas, posiblemente relacionada con temporadas de trabajo existentes en ambas zonas, lo que podría ser un factor importante en el número de contagios producidos.

Tanto en las entradas por vuelta a la residencia habitual como en las entradas por trabajo los confinamientos y medidas aplicadas de movilidad han tenido efectos muy leves debido a la clasificación de este tipo de movilidad como **entradas justificadas**.

Profundizando en las entradas realizadas por otros motivos (Figura 6.34), se puede observar como claramente en los meses de verano (donde apenas existían medidas de restricción) la movilidad por causas **no justificadas** era alta. Esta tendencia alta posteriormente se vio afectada por los confinamientos y medidas aplicadas, bajando el número de entradas por otros motivos en todas las zonas de salud. Por lo tanto, una vez que se cerró perimetralmente la zona de salud o la Comunidad Autónoma, el número de entradas no justificadas bajó, manteniéndose hasta el final de la *segunda ola* con algunos picos altos puntuales.

6.2.2. Movilidad entre zonas afectadas

Se ha realizado un estudio de la movilidad dentro las zonas de salud de Peñafiel e Íscar para determinar si la relación de entradas de un municipio afectado a otro ha sido relevante en la evolución de los casos de contagios.

La elección de ambas zonas ha sido motivada por la continua y alta incidencia de los municipios de Pedrajas y Pesquera de Duero, correspondientes a las zonas de salud de Íscar y Peñafiel respectivamente.

Se ha establecido cada uno de los municipios principales de Peñafiel e Íscar como origen siendo los municipios de Pesquera de Duero y Pedrajas el destino de las entradas, esto se ha debido a la alta incidencia y menor población de estos dos últimos.

Al igual que en los análisis anteriores, se ha obtenido un panel donde se especifica la movilidad de cada municipio en función al tipo:

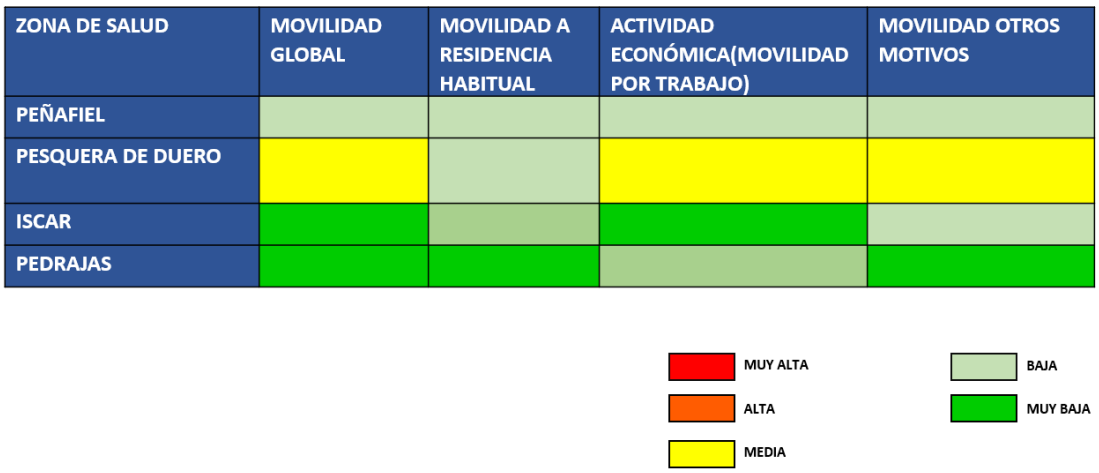


Figura 6.35: Panel de clasificación de la movilidad en los municipios de Peñafiel y Pesquera de Duero

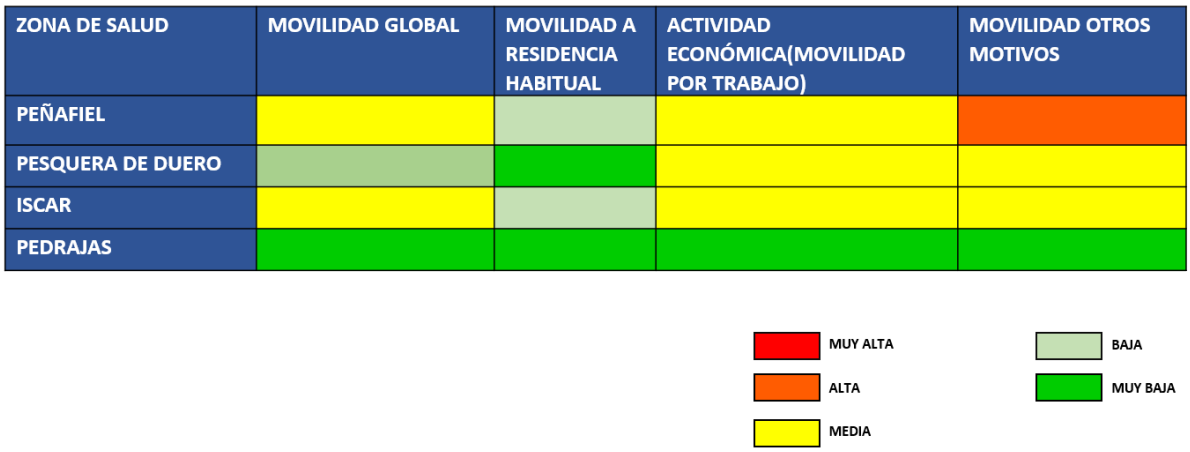


Figura 6.36: Panel de clasificación de la movilidad en los municipios de Íscar y Pedrajas

Los colores asignados en las gráficas a cada uno de los municipios y zonas estudiadas han sido:

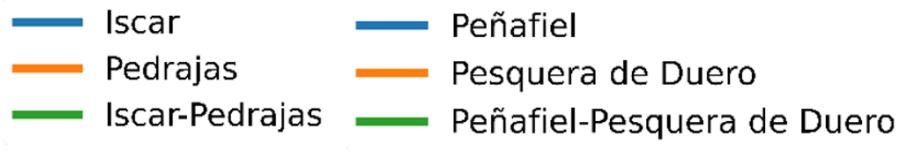


Figura 6.37: Leyendas de gráficas movilidad entre zonas afectadas

Como se puede ver en la Figura 6.37, el color verde corresponderá a aquella movilidad de entrada de un lugar a otro.

Movilidad Peñafiel-Pesquera de Duero:

Se ha estudiado la movilidad entre Peñafiel (origen) y Pesquera de Duero(destino) con el objetivo de determinar si la movilidad entre ambas puede ser útil para la justificación de las subidas y bajadas de la curva de contagios.

Movilidad total

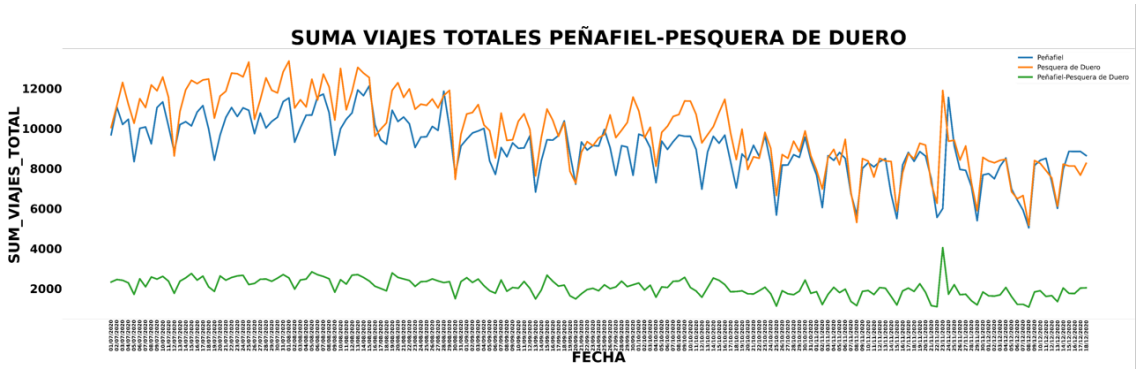


Figura 6.38: Gráfica de entradas totales desde Peñafiel a Pesquera de Duero

Movilidad vuelta a residencia habitual

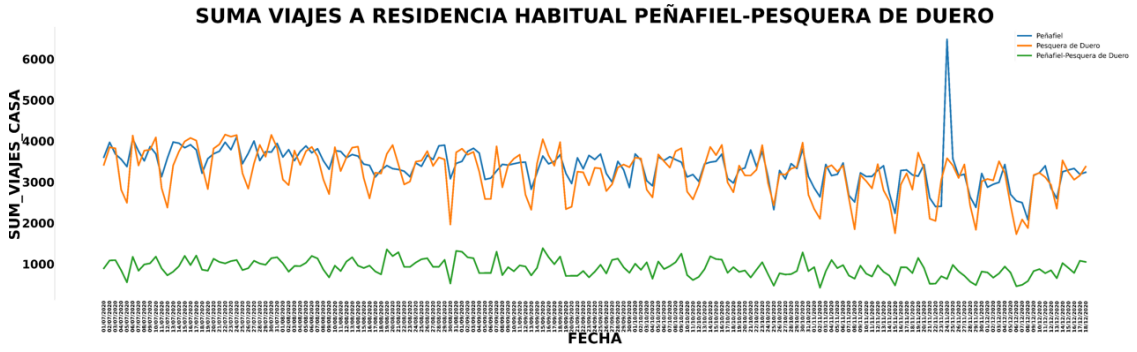


Figura 6.39: Gráfica de entradas a residencia habitual desde Peñafiel a Pesquera de Duero

Movilidad por trabajo

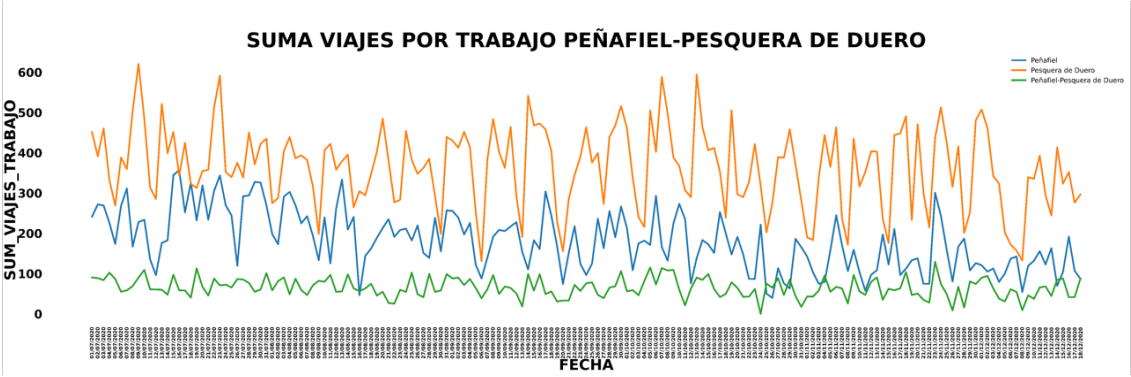


Figura 6.40: Gráfica de entradas por trabajo desde Peñafiel a Pesquera de Duero

Movilidad por otros motivos

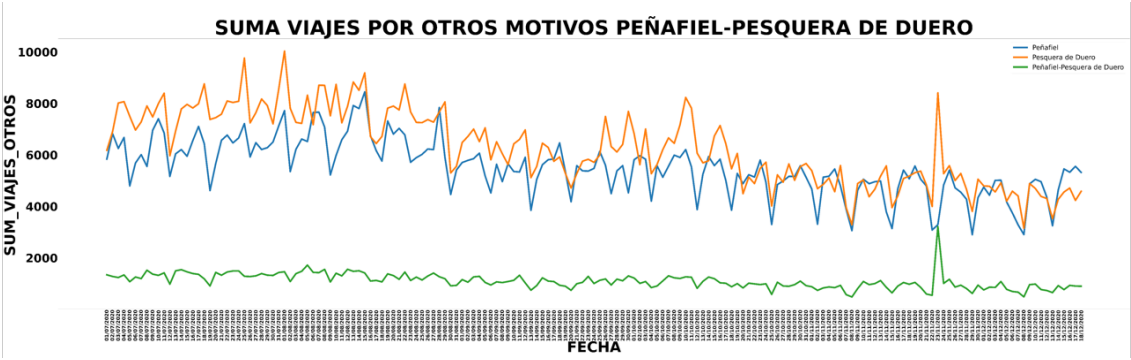


Figura 6.41: Gráfica de entradas por otros motivos desde Peñafiel a Pesquera de Duero

Observando la suma de viajes totales realizados desde el municipio de Peñafiel a Pesquera de Duero (Figura 6.38), se puede apreciar como en los meses de octubre y noviembre existen picos en las tres tendencias estudiadas, lo que nos indica que hubo días donde la actividad entre ambos municipios fue grande.

Analizando la gráfica correspondiente al número de entradas por vuelta a residencia habitual (Figura 6.39), se puede ver como la semejanza entre las tendencias estudiadas es alta, por lo que se deduce que las entradas a Pesquera por vuelta a la residencia habitual provienen en su mayoría de la zona de Peñafiel. Esta observación nos lleva a pensar en la gran cantidad de entradas producidas a Peñafiel desde Pesquera por trabajo u otros motivos.

La gráfica de movilidad por trabajo (Figura 6.40) muestra como la mayoría de entradas por trabajo realizadas a Pesquera de Duero provenían de Peñafiel.

En el caso de la movilidad por otros motivos (Figura 6.41) no se han obtenido resultados que indiquen que el número de entradas realizado al municipio de Pesquera de Duero tuvieran como origen Peñafiel. Destacar únicamente un pico detectado a finales del mes de noviembre.

Movilidad Íscar-Pedrajas

Movilidad total

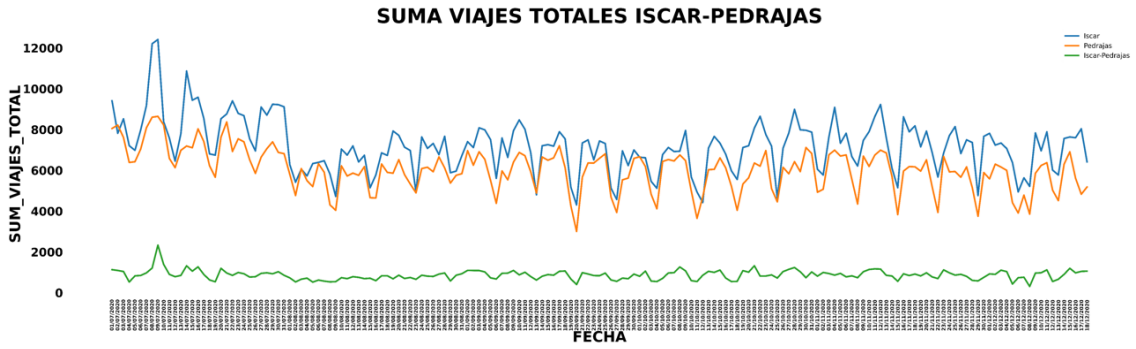


Figura 6.42: Gráfica de entradas totales desde Íscar a Pedrajas

Movilidad vuelta a residencia habitual

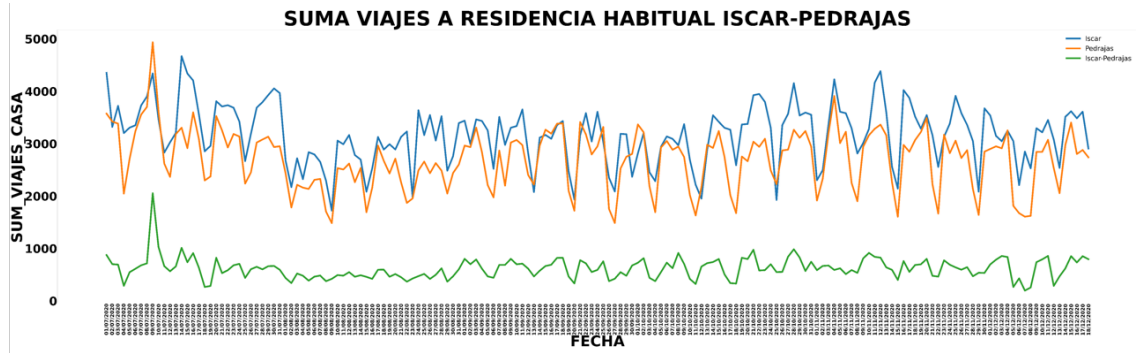


Figura 6.43: Gráfica de entradas a residencia habitual desde Íscar a Pedrajas

Movilidad por trabajo

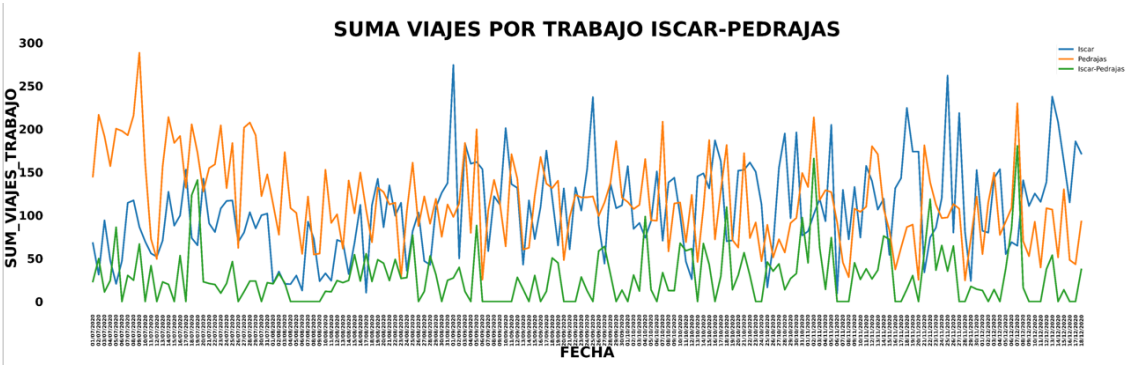


Figura 6.44: Gráfica de entradas por trabajo desde Íscar a Pedrajas

Movilidad por otros motivos

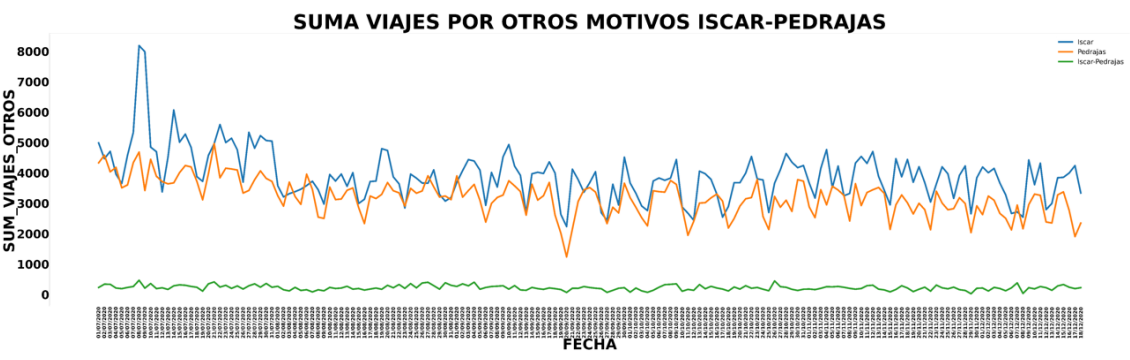


Figura 6.45: Gráfica de entradas por otros motivos desde Íscar a Pedrajas

Analizando la suma de viajes totales realizados desde Íscar a Pedrajas (Figura 6.42), al igual que en el análisis anterior, se puede apreciar la existencia de picos coincidentes en las tres tendencias estudiadas, lo que nos indica que hubo días donde la actividad entre ambos municipios era grande.

Al observar la gráfica correspondiente al número de entradas por vuelta a residencia habitual (Figura 6.43), podemos ver como la tendencia de entradas desde Íscar y otras zonas a Pedrajas es muy semejante, por lo que podemos deducir que las entradas que tienen como destino este municipio tenían como origen en su mayoría la zona de Íscar. Estos resultados nos llevan a pensar también en la gran cantidad de entradas producidas a Íscar desde Pedrajas por trabajo u otros motivos.

Analizando la gráfica de entradas por trabajo (Figura 6.44), podemos observar como el número de entradas por trabajo en Pedrajas provenía en su totalidad de Íscar.

En el caso de la movilidad por otros motivos (Figura 6.45) y al igual que en el caso anterior, no se han obtenido resultados que indiquen que el número de entradas realizado por otros motivos al municipio de Pedrajas tuvieran como origen Íscar.

Finalmente, y una vez realizado los análisis entre los distintos municipios, podemos concluir que la movilidad por trabajo es un factor muy importante a considerar ya que, es el motivo principal de relación entre los municipios estudiados lo que puede estar altamente relacionado con las distintas subidas y bajadas de contagios detectadas. La importancia en la movilidad por trabajo hace que nos fijemos también en la movilidad provocada por la vuelta a la residencia habitual, ya que el número de entradas por trabajo está directamente relacionado con el número de entradas de vuelta a la zona de residencia.

Capítulo 7

Métodos de aprendizaje: Clustering

7.1. Introducción

Una de las fases más importantes del proyecto realizado ha sido la aplicación de los diferentes métodos y modelos a los datos. A través de su uso, hemos ido obteniendo resultados que han guiado la elección de los mejores tipos y técnicas para la construcción final de una herramienta que permita alcanzar los objetivos marcados.

Al encontrarnos dentro de un trabajo de investigación, las técnicas y métodos usados no estaban especificados desde un principio, por lo que ha sido necesaria la búsqueda y análisis de técnicas dentro del campo de la Inteligencia Artificial que nos permitieran trabajar con los datos disponibles, obteniendo así resultados acordes a los objetivos previstos.

7.2. Clustering

Situándonos en este contexto de incertidumbre (donde los resultados a obtener no eran claros), se ha optado por la elección de la técnica de aprendizaje no supervisado clustering [6].

El clustering o algoritmo de agrupamiento, consiste en la división de un conjunto de datos de entrada (datos de salud) en grupos o subconjuntos (clusters), de manera que los elementos que componen cada grupo compartan características o patrones indetectables a primera vista. Todos los resultados obtenidos dependerán del tipo de algoritmo clustering elegido y la medida de similitud usada para la comparación entre elementos (distancia).

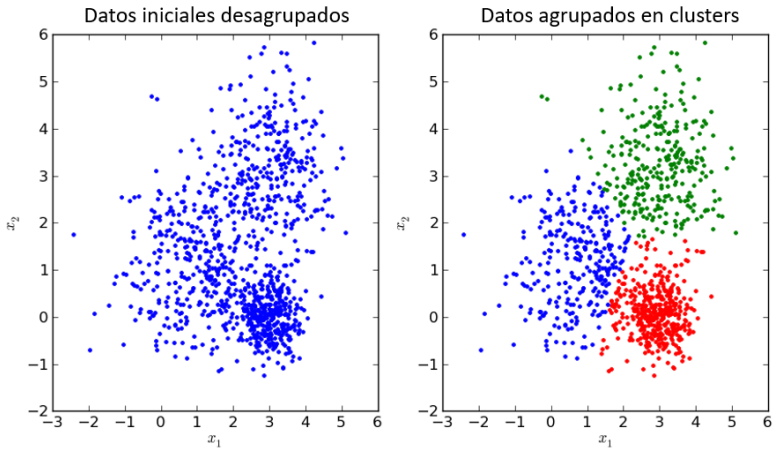


Figura 7.1: Ejemplo de aplicación de clustering [20].

Tal y como podemos ver en la Figura 7.1, este método se basa principalmente en la agrupación de elementos según su proximidad. Los elementos próximos entre sí pertenecerán al mismo cluster y aquellos elementos que sean lejanos unos de otros pertenecerán a clusters diferentes.

En ciertos datos la aplicación de clustering puede hacer aparecer outliers (elementos que no pertenecen a ningún subgrupo o cluster), tal y como se puede apreciar en la Figura 7.2. En el caso de este proyecto siempre se intentará evitar (sí es posible) la aparición de esos outliers.

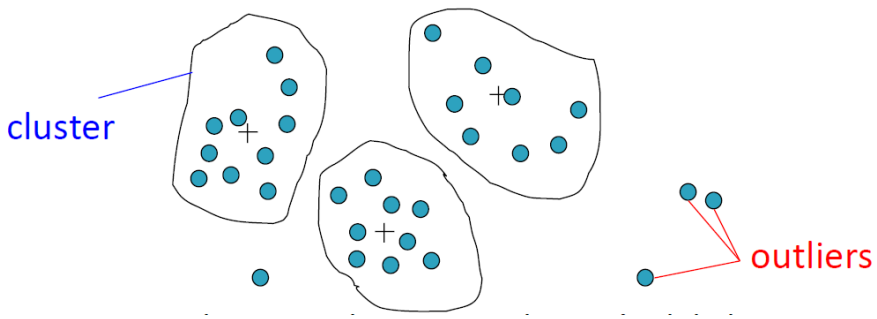


Figura 7.2: Outliers en clustering[30]

La definición de la distancia entre elementos es la clave del clustering y lo que determina la especificidad de cada problema. Su cálculo se basa principalmente en la obtención de la distancia entre dos vectores, x e y, dimensionales. Este cálculo se puede realizar de múltiples formas siendo las más comunes:

- **Distancia Euclídea:** distancia entre dos puntos a partir del teorema de Pitágoras denominado espacio euclídeo.

$$\sqrt{|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2} \tag{7.1}$$

- **Distancia de Manhattan:** obtención de la distancia entre dos puntos a través de la suma de las diferencias absolutas de sus coordenadas

$$|x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}| \tag{7.2}$$

- **Distancia de Minkowski:** generalización de las distancias euclídeas (q=1) y de Manhattan (q=2) vistas anteriormente.

$$\sqrt[q]{|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q} \tag{7.3}$$

- **Distancia de Chebyshev:** conocida como la distancia del tablero de ajedrez, ya que se basa en el número de movimientos necesarios para que la pieza del rey llegue de una casilla a otra dentro del tablero de dicho juego.

$$d_{max}(x_i, y_j) = max_z|x_{i_z} - x_{j_z}| \tag{7.4}$$

La Figura 7.3 representa visualmente cada una de las distancias explicadas anteriormente:

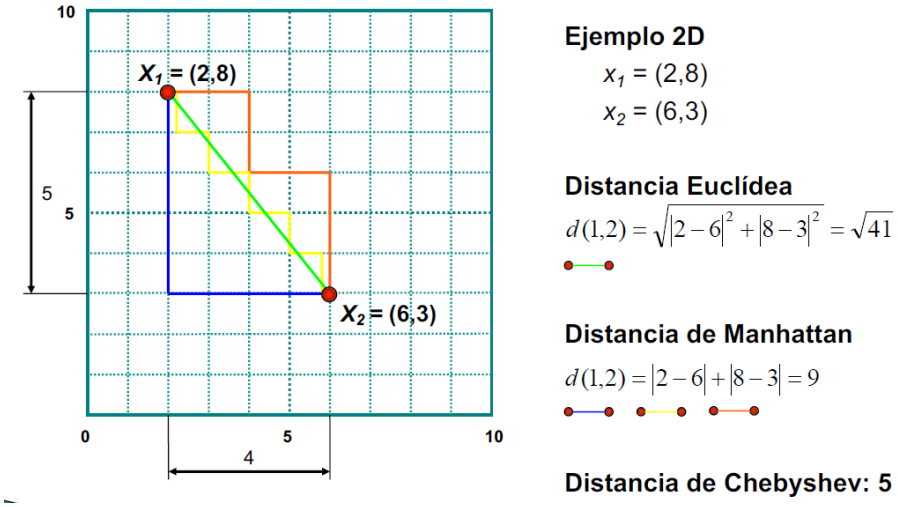


Figura 7.3: Ejemplo de formas de cálculo de distancia en clustering

7.2.1. Tipos de clustering

Dentro de la técnica de clustering, se encuentran diversos tipos:

- **Clustering por particiones:** uso de algoritmos de agrupamiento por particiones. En este tipo de clustering se tienen una serie de clusters denominados k con un centroide o medoid asociado, el cual es usado de manera iterativa para la asignación de los elementos a cada cluster. Ejemplos de este tipo de clustering son K-means y K-medoids.

- **Clustering jerárquico:** método de minería de datos usado para la agrupación de datos en cluster, quedando estos anidados en jerarquías siguiendo la forma de un árbol. Ejemplos: Diana, Agnes, BIRCH o CHAMELEON.

- **Métodos basados en densidad:** método de clustering que usa como criterio la densidad de puntos en lugar de la distancia para la creación de grupos o clusters. Ejemplos: DBSCAN.

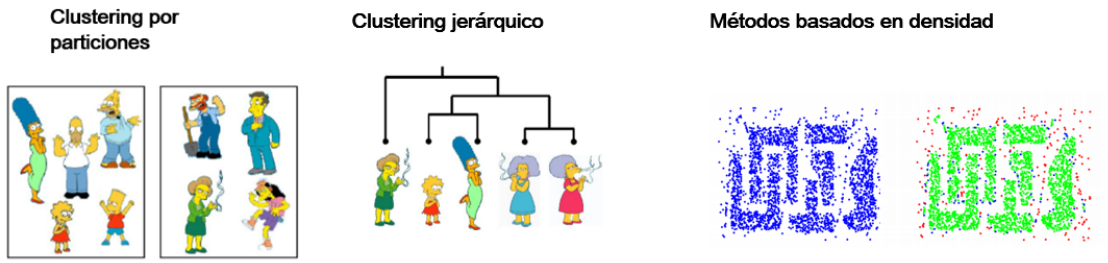


Figura 7.4: Tipos de clustering

7.2.2. Clustering por particiones

Dentro de este proyecto se ha escogido el tipo de clustering basado en particiones, debido a la experiencia previa en el uso de este tipo de algoritmos y en la sencillez de construcción que ofrece respecto a otros. Este tipo de agrupamiento se basa principalmente en la fijación de un número k de clusters conocido, los cuales poseen un punto geométrico denominado centroide usado para la asignación de datos a cada cluster de manera iterativa. En la Figura 7.5 podemos observar un ejemplo del proceso de creación de clusters por particiones.

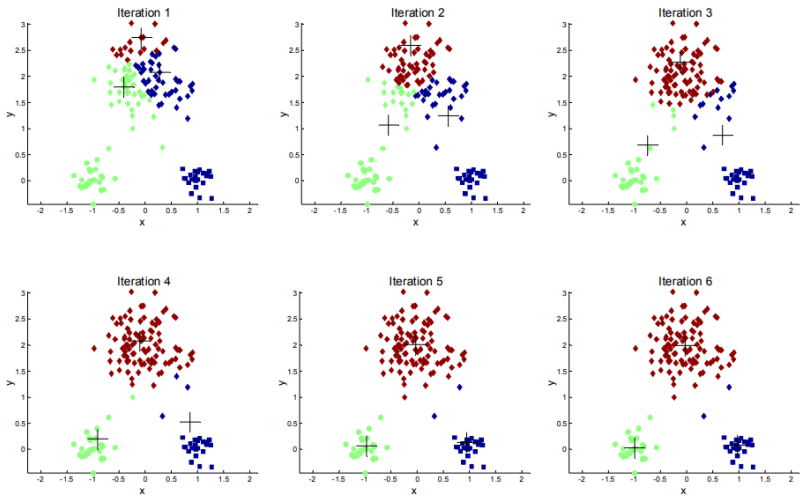


Figura 7.5: Funcionamiento clustering por particiones [4].

Los algoritmos de particionado estudiados y usados con los datos obtenidos han sido K-means y K-medoids.

K-means

K-means es un algoritmo de agrupamiento por particiones donde se fija un número de k clusters específico. Cada uno de estos clusters tendrá un centro geométrico denominado centroide que será usado para la creación de los subgrupos. A continuación se explican todo los procesos y herramientas necesarias para el funcionamiento de K-means.

Determinación del número k de clusters [28]:

La determinación del número k de clusters a utilizar es un factor muy importante dentro de nuestro algoritmo de partición, ya que un número de clusters erróneo puede dar lugar a una agrupación de datos incorrecta o poco precisa que afecte a los resultados obtenidos. No existe ningún criterio que permita obtener de manera exacta el número de clusters óptimo para realizar el agrupamiento, sin embargo, existen técnicas como el método de Calinsky, Gap o método del codo (elbow), que nos ofrecen una orientación a la hora de elegir dicho número. En este proyecto se hará uso del método del codo o método elbow debido a la experiencia previa con él. El método de elbow hace uso de los valores de inercia obtenidos al aplicar el algoritmo K-means a un número de clusters entre 1 y N. Dicha inercia esta formada por la suma de las distancias al cuadrado de cada elemento

del cluster a su centroide:

$$^{N}\sum_{i_{=0}} ||x_{i-\mu}||^2 \tag{7.5}$$

A través de una gráfica se representa la inercia obtenida respecto del número de clusters con el objetivo de apreciar un cambio brusco que nos indique el número óptimo de clusters a usar.

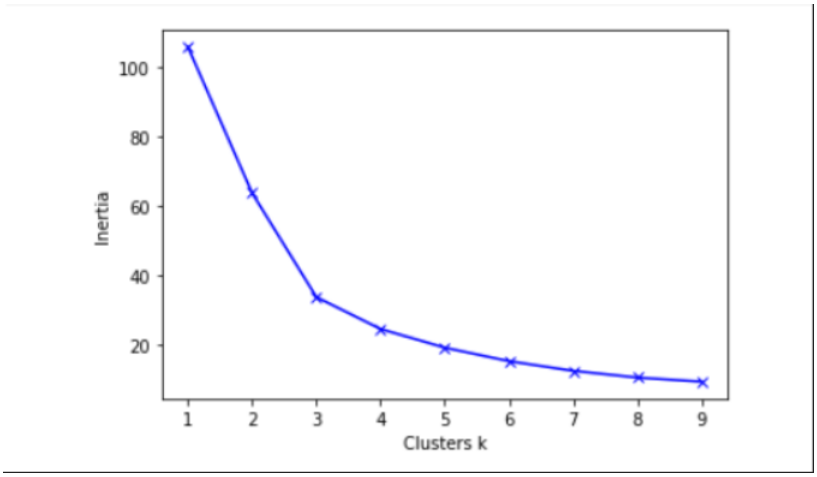


Figura 7.6: Ejemplo gráfica elbow

Como podemos ver en la Figura 7.6 , la gráfica obtenida posee la forma de un brazo, siendo su codo (elbow) el que nos indica el valor de k clusters recomendado a usar.

Determinación de centroides:

Partiendo de un número k de clusters, cada uno de estos grupos determinará un centroide C_i, escogiendo aquellos valores que minimicen la función objetivo:

$$Coste(C) = \sum_{i=1}^{k} \sum_{x \in C_i} d^2(m_i, x) \tag{7.6}$$

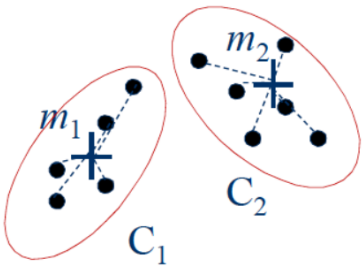


Figura 7.7: Ejemplo determinación de centroides

Proceso de aprendizaje:

Los diferentes puntos a agrupar son asignados al centroide más cercano haciendo uso de las métricas vistas anteriormente. De esta manera, se van construyendo los k clusters especificados inicialmente en el algoritmo. Todo este proceso es realizado de manera iterativa, actualizando en cada iteración los centroides en función de las asignaciones de puntos a cada cluster. Este proceso se detiene una vez que los centroides adquieren una estabilidad (la función objetivo no podrá minimizarse por debajo de un umbral dado o cuando los centroides no se mueven).

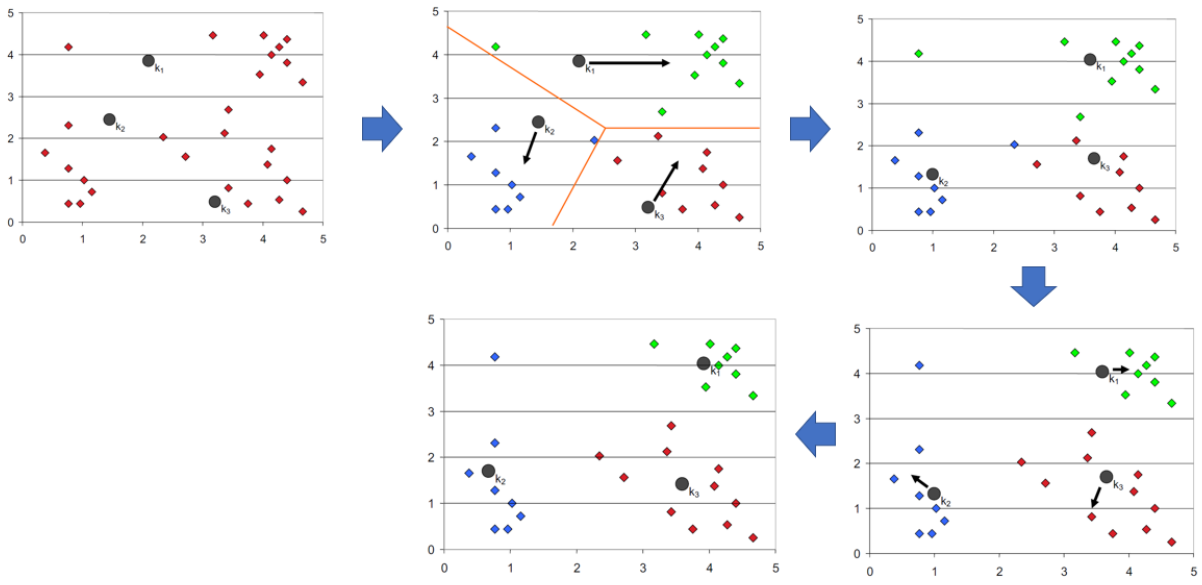


Figura 7.8: Ejemplo del proceso de k-means

Desventajas:

El algoritmo K-means posee una serie de desventajas que debemos tener en cuenta a la hora de su utilización:

- **Inicialización:** El algoritmo necesita de heurísticas adecuadas para la estimación de los centroides de cara a la obtención de unos resultados apropiados.
- **Forma del cluster:** K-means tiene problemas a la hora de trabajar con clusters de distinto tamaño, densidad o convexidad.
- **Sensibilidad a outliers:** Una de las principales desventajas que posee el algoritmo K-Means es la sensibilidad a outliers, lo que puede ser un gran problema a la hora de obtener unos resultados precisos y claros.

K-medoids:

El algoritmo K-medoids, similar en funcionamiento al algoritmo K-means, posee diferencias respecto a este que le permiten ser más robusto ante la presencia de outliers.

Determinación del número k de clusters:

Al igual que en K-means, para la determinación del número de clusters se usará el método de elbow descrito anteriormente (Sección 7.2.2).

Proceso:

En el caso de K-medoids, el uso de centroides es sustituido por los denominados medoids. Los medoids hacen referencia a la elección de un objeto existente dentro de un cluster siendo este el más central. Para la elección de estos medoids, se puede usar el método PAM (Partitioning Around Medoids) [4], que realiza una serie de pasos para la elección correcta de los objetos más centrales de cada cluster, obteniendo finalmente una agrupación de los datos en clusters:

1. Selección arbitraria de elementos dentro de los k cluster como medoids iniciales.
2. Asignación de cada objeto restante al medoid más cercano.
3. Aleatoriamente, seleccionar un objeto considerado no-medoid (O-random).
4. Cálculo total del coste de cambio.

5. Intercambio de objeto considerado medoid con aquel objeto aleatorio (O-random) que suponga el mínimo coste de reemplazo. Si ese mínimo es negativo volveremos al paso 1.

6. En el caso de que ese mínimo sea positivo para cada objeto no seleccionado se buscara el objeto medoid más similar y se detendrá el proceso.

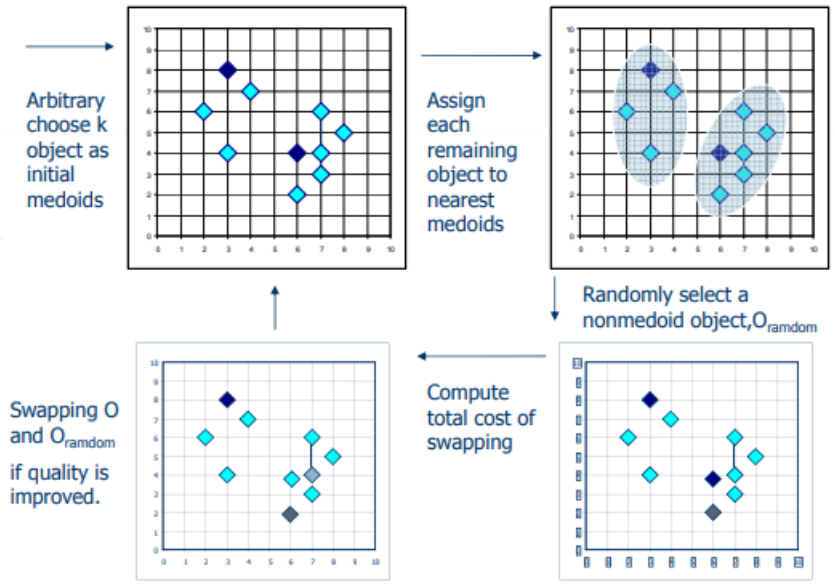


Figura 7.9: Diagrama método PAM. [4]

Desventajas K-medoids:

- **Variabilidad de resultados obtenidos:** Una de las desventajas del algoritmo K-Medoids descrito, es que puede obtener resultados diferentes para distintas ejecuciones en el mismo conjunto de datos.
- **Forma del cluster:** K-medoids también genera problemas a la hora de trabajar con clusters de distinto tamaño, densidad o convexidad.

7.3. Construcción de los algoritmos

Una vez analizadas las distintas técnicas de clustering a usar en el proyecto, se realizará su construcción e implementación en el contexto dado [1]. Los distintos algoritmos de clustering han sido aplicados con el objetivo de obtener una visión preliminar del estado epidemiológico en el tiempo de cada una de las zonas de salud estudiadas. El objetivo es clasificar y agrupar los datos de cada zona en clusters que nos permitan obtener distintos niveles de gravedad epidemiología. La construcción de los métodos se realizará en el lenguaje Python a través de la plataforma online Colab de Google. Se usarán bibliotecas como sklearn, pandas o matplotlib, relacionadas con el aprendizaje automático y el tratamiento y análisis de datos.

7.3.1. Preparación de los datos

Para la implementación de los algoritmos de clustering se hace uso únicamente de aquellos datos relacionados con la salud de las zonas estudiadas (apartado 5.3). Se emplea el dataset de salud generado con anterioridad, compuesto únicamente por las variables de salud ya descritas en las Tablas 5.2 y 5.3. Aquellas variables relacionadas con las medidas aplicadas o la movilidad, no serán usadas para la aplicación de clustering, siendo eliminada mediante los métodos proporcionados por la biblioteca pandas.

Dentro de los datos de salud solo se usarán aquellas variables relacionadas con la IA de contagios (**POSI_IA_XDIAS**), pruebas realizadas (**REALI_IA_XDIAS**) y fallecimientos (**FALL_IA_XDIAS**) por periodos de tiempo, junto a las variables de porcentaje PCR (**PORC_PCR_X_DIAS**). Se almacenarán las variables de nombre de la zona de salud (**CENTRO**) y fecha (**FECHA**) de cada dato usado, para la posterior representación gráfica de los resultados obtenidos por el clustering.

	PORC_PCR_4DIAS	PORC_PCR_7DIAS	PORC_PCR_14DIAS	POSI_IA_4DIAS	POSI_IA_7DIAS	POSI_IA_14DIAS	REALI_IA_4DIAS	REALI_IA_7DIAS	REALI_IA_14DIAS	FALL_IA_4DIAS	FALL_IA_7DIAS	FALL_IA_14DIAS
0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
1	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
2	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
3	5.882353	0.0	0.0	3.006976	0.0	0.0	51.118595	0.0	0.0	0.0	0.0	0.0
4	0.000000	0.0	0.0	0.000000	0.0	0.0	15.034881	0.0	0.0	0.0	0.0	0.0

Figura 7.10: Dataset usado para clustering

La razón por la que se han elegido únicamente los datos de salud calculados en los periodos de tiempo de 4, 7 y 14 días se ha debido al comportamiento del virus, que debe ser medido en los distintos periodos de tiempo. Otro de los motivos lo encontramos en la gran capacidad de análisis e interpretación de dichos datos en comparación con aquellos

obtenidos diariamente. Las variables pertenecientes a las **ventanas móviles** realizadas sobre los datos también han sido eliminadas debido a la existencia de la variable de Incidencia Acumulada, la cual representa la misma información de manera más precisa y teniendo en cuenta factores como la población de cada zona.

Debido a esto, las técnicas de clustering aplicadas a los datos se han divido en tres tipos según el periodo correspondiente de las variables utilizadas (4, 7 y 14 días). Como podemos apreciar, el clustering se hará sobre varias variables siendo este de tipo **multi-dimensional**.

7.3.2. Creación del método

Una vez realizada la transformación y procesamiento del dataset a usar, comenzará la construcción de los algoritmos de clustering.

El primer paso a realizar es la normalización de las variables usadas para que el modelo pueda trabajar con estas de una manera correcta. Para ello, se hará uso de la función de preprocesado **MinMaxScaler()**, la cual normaliza nuestros datos para situarlos en un rango entre 0 y 1.

```
[5]  scaler = preprocessing.MinMaxScaler()
     df_normal = scaler.fit_transform(df)

     df_normal

     array([[0.        , 0.        , 0.        ],
            [0.        , 0.        , 0.        ],
            [0.        , 0.        , 0.        ],
            ...,
            [0.1754386 , 0.01799982, 0.02036763],
            [0.19047619, 0.02159979, 0.02251159],
            [0.16666667, 0.02159979, 0.02572753]])
```

Figura 7.11: Código de normalización datos clustering

Como podemos ver en la Figura 7.11, se almacenará en la variable scaler la aplicación de la función MinMaxScaler(), para posteriormente crear un nuevo dataset (df_normal) con todos los datos ya normalizados en un rango de 0 a 1.

A continuación y tal como se ha explicado en el apartado 7.2.2 , se hace el cálculo y aplicación de la inercia para la determinación mediante el método de elbow del número de k clusters recomendado. Se realiza la búsqueda de dicho número en un rango de 1 a 9 clusters, guardando en las variables correspondientes a los tipos de clustering usados (**kmeansModel** y **kmedoidsModel**), los resultados obtenidos con cada iteración del bucle for, almacenándose posteriormente en la variable **inercia** (usada para la representación gráfica).

K-means:

```
[ ]
    inertia = []
    K = range(1,10)
    for k in K:
        kmeansModel = KMeans(n_clusters=k).fit(df_normal)
        kmeansModel.fit(df_normal)
        inertia.append(kmeansModel.inertia_)
```

```
[ ] plt.plot(K, inertia, 'bx-')
    plt.xlabel('Clusters k')
    plt.ylabel('Inertia')
    plt.show()
```

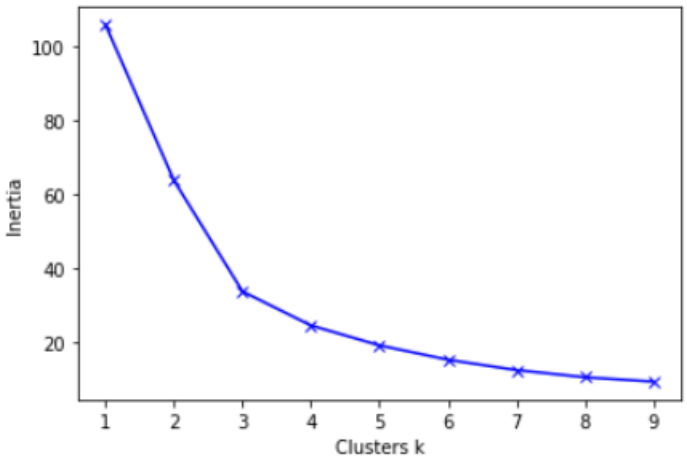


Figura 7.12: Código determinación k clusters en K-means

K-medoids:

```
inertia = []
K = range(1,10)
for k in K:
    kmedoidsModel = KMedoids(n_clusters=k).fit(df_normal)
    kmedoidsModel.fit(df_normal)
    inertia.append(kmedoidsModel.inertia_)
```

```
plt.plot(K, inertia, 'bx-')
plt.xlabel('Clusters k')
plt.ylabel('Inertia')
plt.show()
```

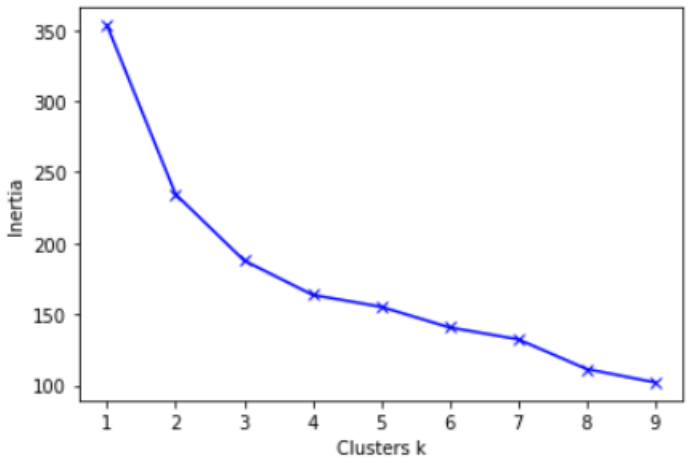


Figura 7.13: Código determinación k clusters en K-medoids

Finalmente obtenemos **3** clusters como número ideal en ambos tipos de clustering. En K-means se aprecia visualmente la forma del codo (elbow) indicando el valor 3 del eje X (Clusters k). En el caso de K-medoids este codo se hace más complejo de vislumbrar, pero al igual que en K-means podemos comprobar que el número de clusters indicado como correcto es 3. El siguiente paso a realizar será la aplicación de los métodos K-means y K-medoids al conjunto de datos mostrando las coordenadas dadas a cada centroide.

```
[ ]
KMeans = KMeans(n_clusters=3, random_state= 123).fit(df_normal)
centroids = KMeans.cluster_centers_
print(centroids)
```

```
[[0.09362709 0.02909212 0.0531528 ]
 [0.33932927 0.16639436 0.09977993]
 [0.0823143  0.31756374 0.86594091]]
```

```
KMeans.labels_ = map(change_labels, KMeans.labels_)


labels = pd.DataFrame(KMeans.labels_) #This is where the label outpu
labeledConfin = pd.concat((df,labels),axis=1)
labeledConfin = labeledConfin.rename({0:'labels'},axis=1)
labeledConfin.head()
```

	PORC_PCR_14DIAS	POSI_IA_14DIAS	REALI_IA_14DIAS	labels
0	0.0	0.0	0.0	0
1	0.0	0.0	0.0	0
2	0.0	0.0	0.0	0
3	0.0	0.0	0.0	0
4	0.0	0.0	0.0	0

Figura 7.14: Código aplicación algoritmo K-means

En el caso de K-medoids se mostrarán los medoids obtenidos.

```
[26]
KMedoids = KMedoids(n_clusters=3, random_state= 123).fit(df_normal)
centroids = KMedoids.cluster_centers_
print(centroids)
```

```
[[0.21611722 0.07295286 0.06701165]
 [0.38190955 0.19475286 0.10123267]
 [0.02857143 0.00538515 0.03741653]]
```

```
KMedoids.labels_ = map(change_labels, KMedoids.labels_)


labels = pd.DataFrame(KMedoids.labels_) #This is where the label output of
labeledConfin = pd.concat((df,labels),axis=1)
labeledConfin = labeledConfin.rename({0:'labels'},axis=1)
labeledConfin.head()
```

	PORC_PCR_14DIAS	POSI_IA_14DIAS	REALI_IA_14DIAS	labels
0	0.0	0.0	0.0	0
1	0.0	0.0	0.0	0
2	0.0	0.0	0.0	0
3	0.0	0.0	0.0	0
4	0.0	0.0	0.0	0

Figura 7.15: Código aplicación algoritmo K-medoids

A través de la variable **labels** se especifica el cluster al que ha sido asignado cada dato, siendo los valores de label los correspondientes al número de k clusters comenzando por 0. Esta variable **labels** se insertará dentro del dataset al que se le ha aplicado el algoritmo de clustering, teniendo así registrado en una variable el grupo asignado a cada dato del dataset. El paso siguiente será el de representar los clusters obtenidos para el conjunto de datos estudiado.

La Figura 7.16, muestra los resultados de los clusters u agrupaciones obtenidas con las variables utilizadas en el proceso de aprendizaje (uso de la función *pairplot* sobre las labels obtenidas anteriormente).

```
sns.pairplot(labeledConfin,hue='labels')
```

```
<seaborn.axisgrid.PairGrid at 0x7f42c3564cd0>
```

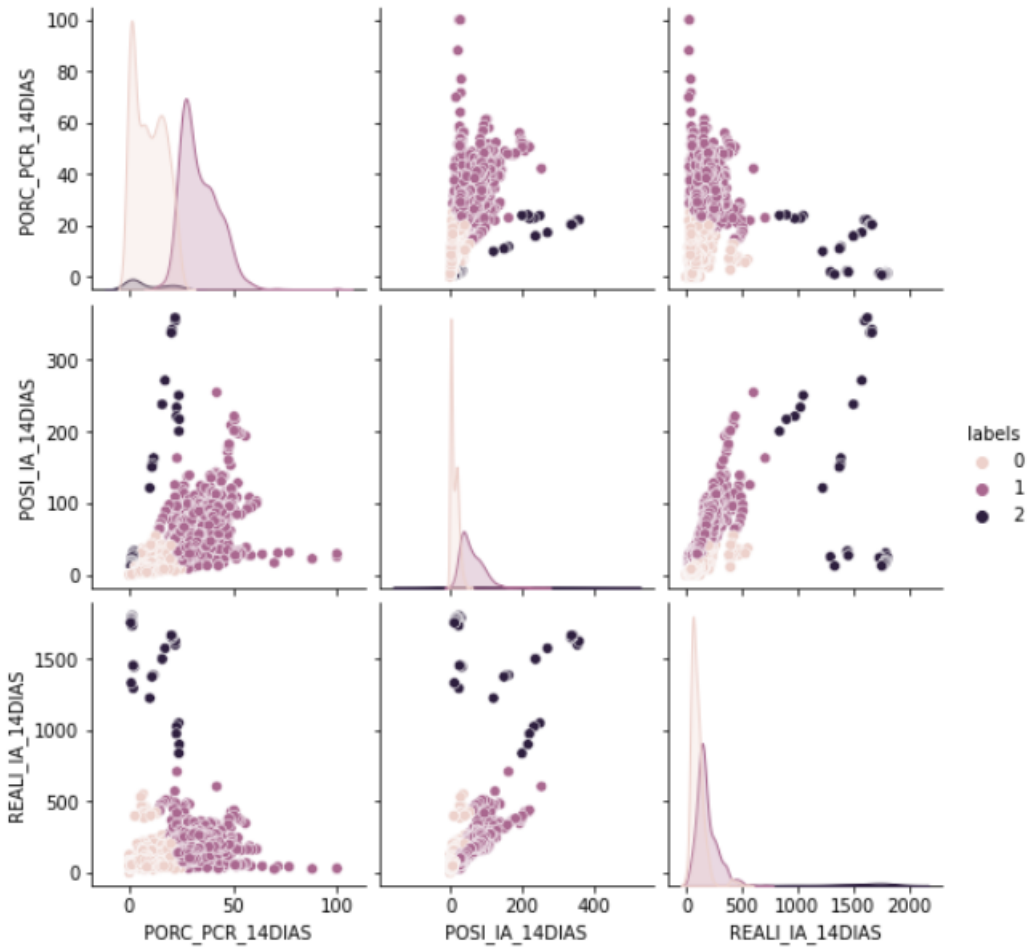


Figura 7.16: Visualización resultados clustering

Mediante el uso del método *sairplot*, proporcionado por la biblioteca *seaborn*, obtenemos un resultado visual de las agrupaciones o clusters creados para cada variable usada, tal y como podemos ver en la Figura 7.17. En este caso se hace uso de una variable constante para la muestra de los clusters obtenidos en función de cada una de las variables usadas en la clasificación. De esta manera, se podrá obtener mediante esta función un análisis visual claro del clustering multidimensional realizado por cada variable.

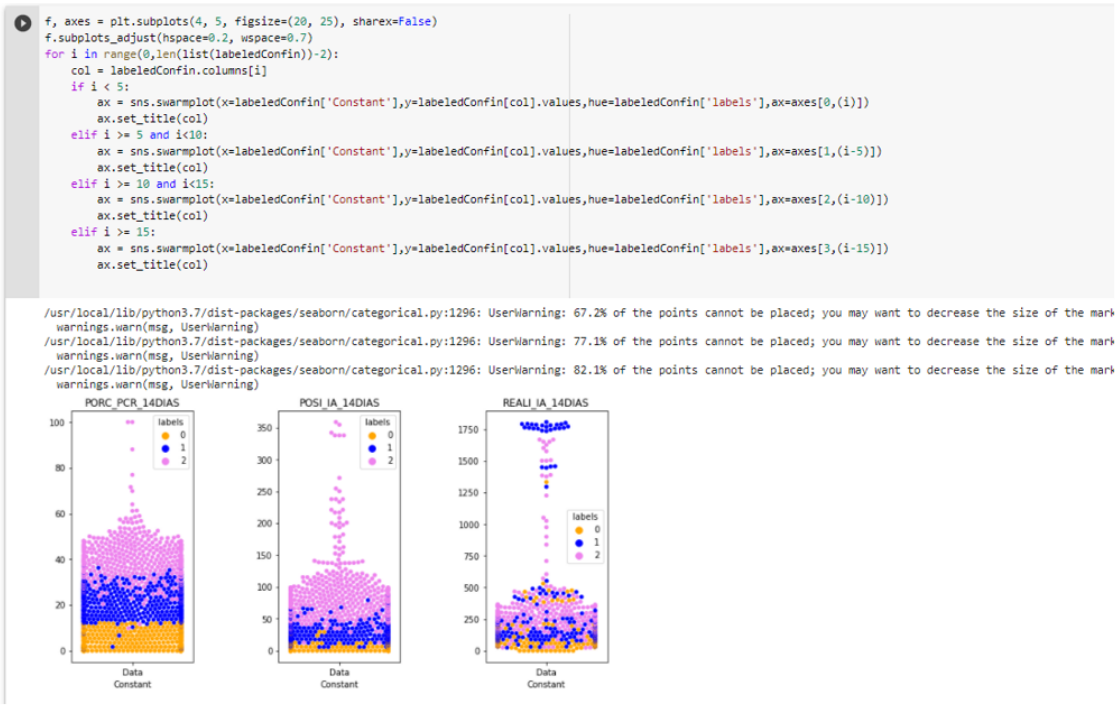


Figura 7.17: Método *sairplot* visualización resultados clustering en cada variable

7.3.3. K-means vs K-medoids

Como hemos podido observar anteriormente, ambos algoritmos de agrupamiento por particiones son similares en funcionamiento, sin embargo, el algoritmo K-medoids nos ofrece una mayor robustez de clasificación frente outliers, algo que estará muy presente en nuestro proyecto debido al tipo de datos usados y a la propia naturaleza de este. Es por ello por lo que se ha realizado una comparación entre los resultados obtenidos con ambos algoritmos para determinar cuales la mejor opción a elegir.

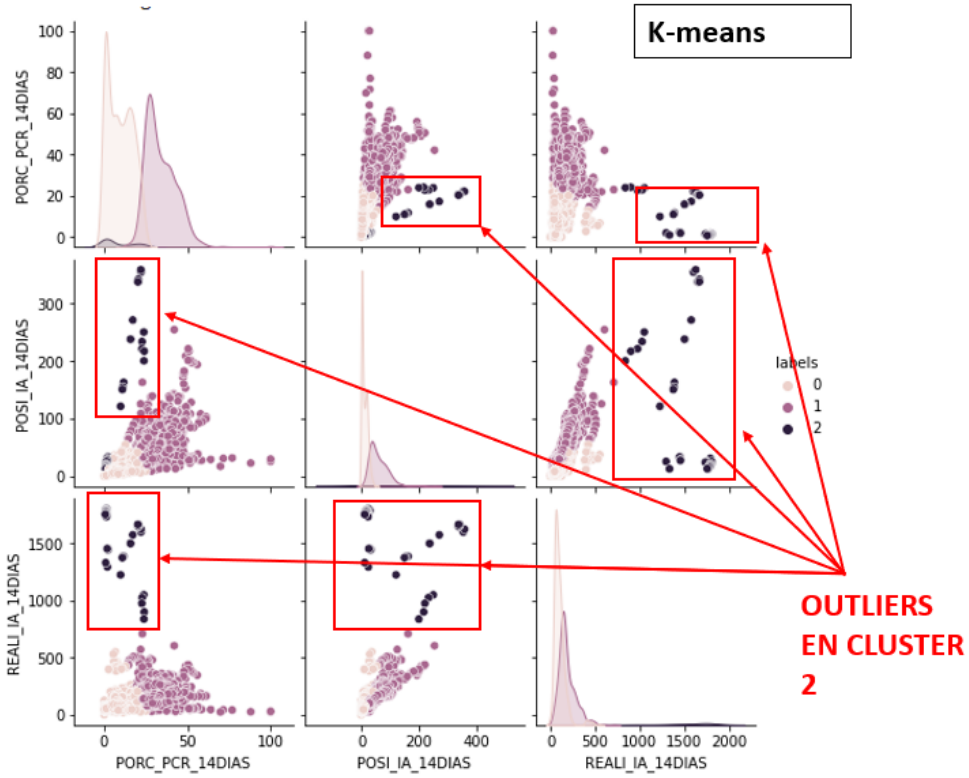


Figura 7.18: Resultados algoritmo K-means

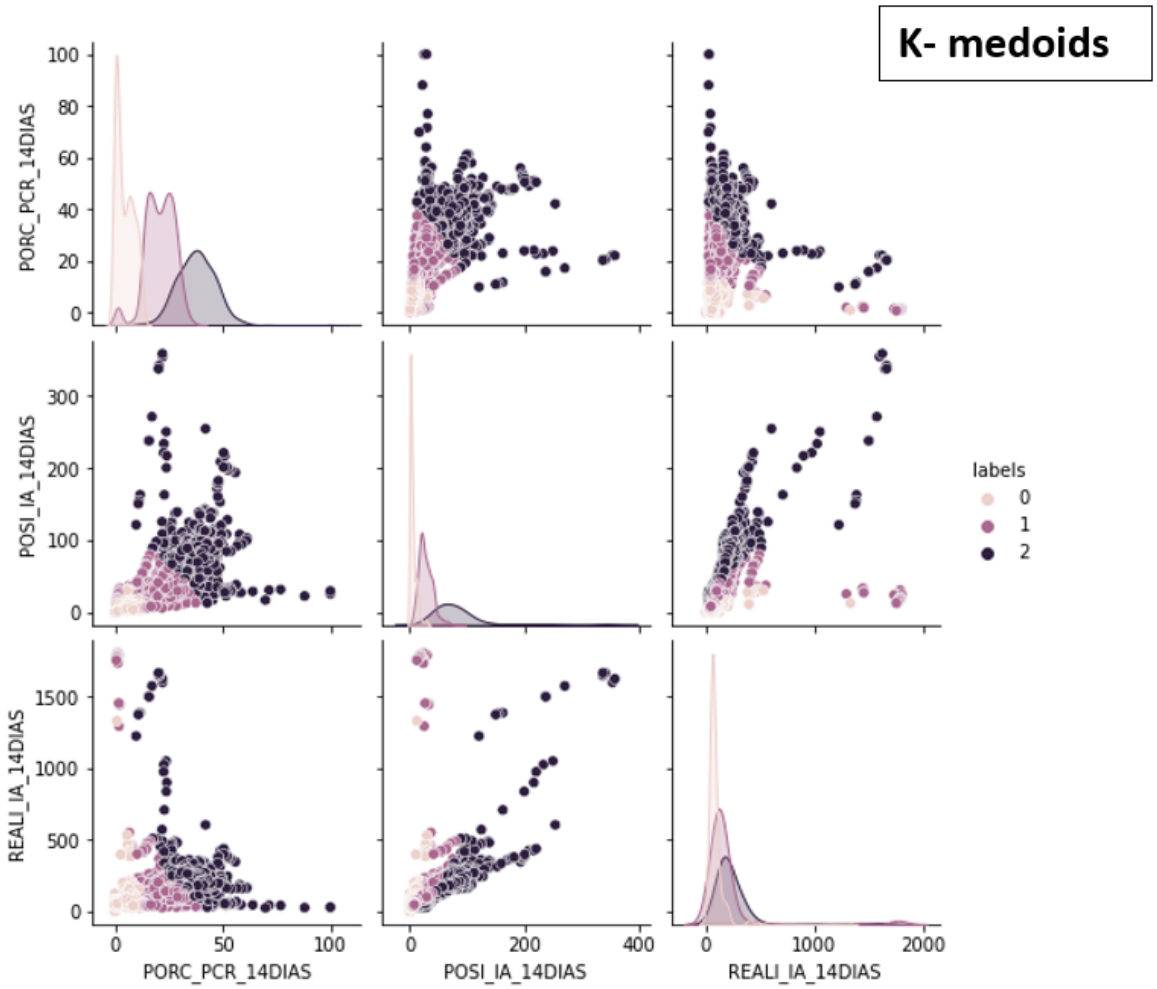


Figura 7.19: Resultados algoritmo K-medoids

Como podemos apreciar en las Figuras 7.18 y 7.19, el algoritmo K-medoids para los mismos datos y número de k clusters que el K-means, ofrece unos mejores resultados en cuanto a la clasificación y tratamiento de outliers (indicados en los resultados del algoritmo K-means mediante **cuadros rojos**). Debido a esta razón, en el proyecto se hará uso de este algoritmo en lugar de K-means, debido a su mayor robustez y clasificación de la información.

7.3.4. Clustering en periodos de tiempo

Una vez construido y elegido el algoritmo de clustering a usar en nuestro proyecto, será aplicado a nuestro dataset diferenciando entre cada uno de los periodos estudiados (4, 7 y 14 días). El objetivo será el de detectar aquel periodo que ofrezca los mejores resultados.

El número de k clusters usado en los tres periodos ha sido de 3, ya que tanto la gráfica de elbow como las pruebas realizadas previamente nos han permitido determinar este número de k clusters como el mejor.

Eliminación variable de fallecidos:

Las variables utilizadas finalmente en cada uno de los periodos han sido aquellas relacionadas con la incidencia acumulada de PCR realizadas (**REALI_IA_XDIAS**) y PCR positivas (**POSI_IA_XDIAS**), así como la variable porcentaje de PCR (**PORC_PCR_XDIAS**). La variable de IA de fallecidos ha sido eliminada debido a los malos resultados obtenidos al aplicar el algoritmo K-medoids y a la ausencia de resultados precisos usando esta.

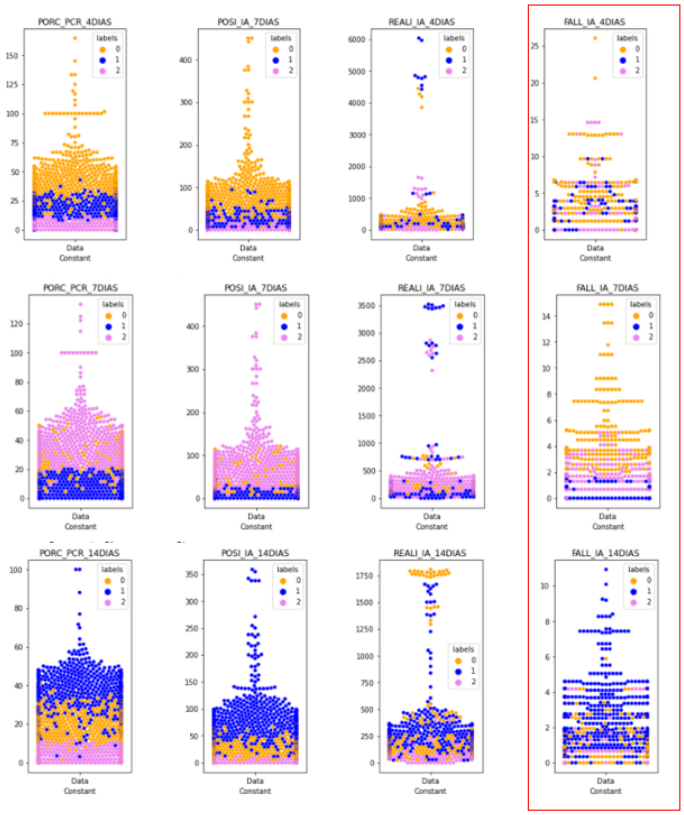


Figura 7.20: Resultados clustering con variable IA fallecidos

Como podemos ver en la Figura 7.20, donde se muestran las agrupaciones o clusters obtenidos para cada una de las variables usadas, aquellos resultados correspondientes a la variable **FALL_IA_XDIAS** son los menos precisos en cuanto a la determinación de agrupaciones o clusters.

A continuación se muestran los resultados obtenidos con cada uno de los periodos.

Período de 4 días:

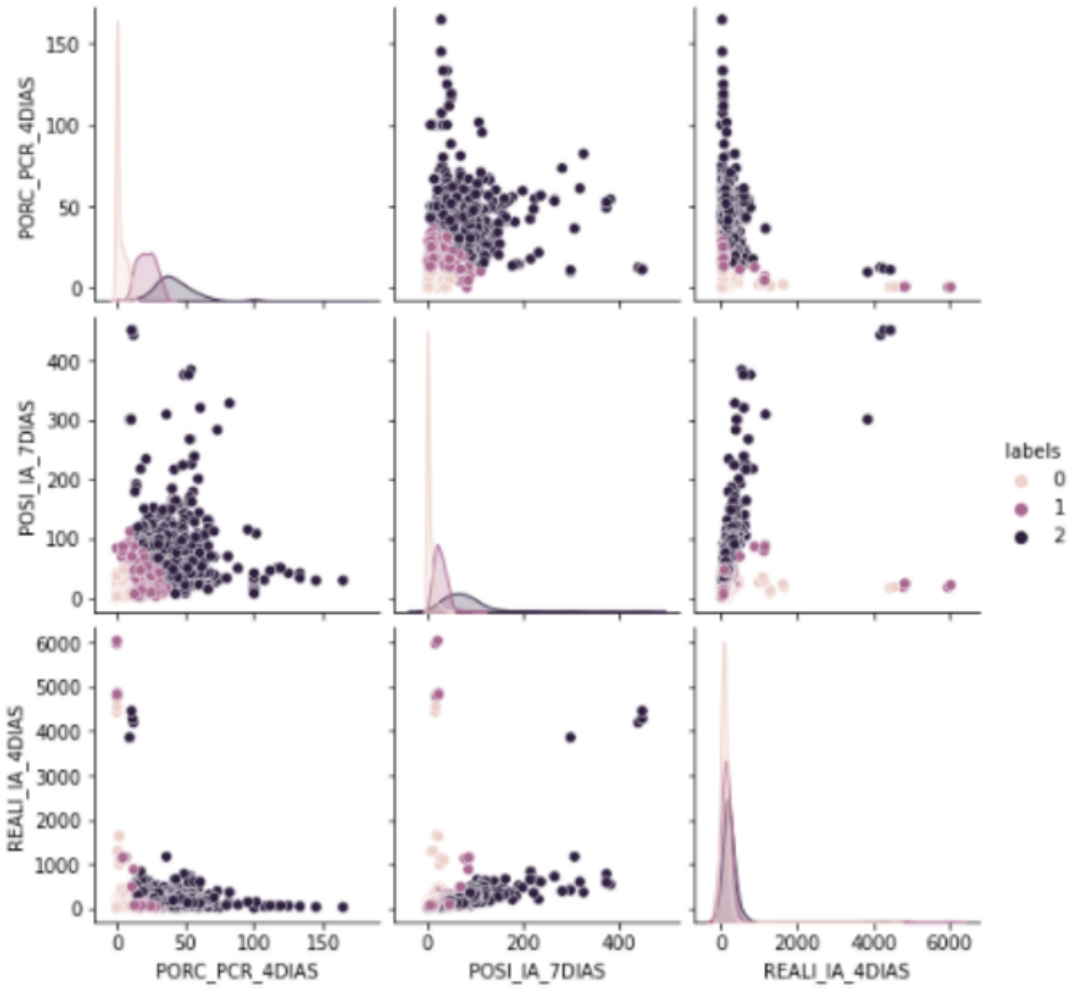


Figura 7.21: Resultados clustering periodo de 4 días

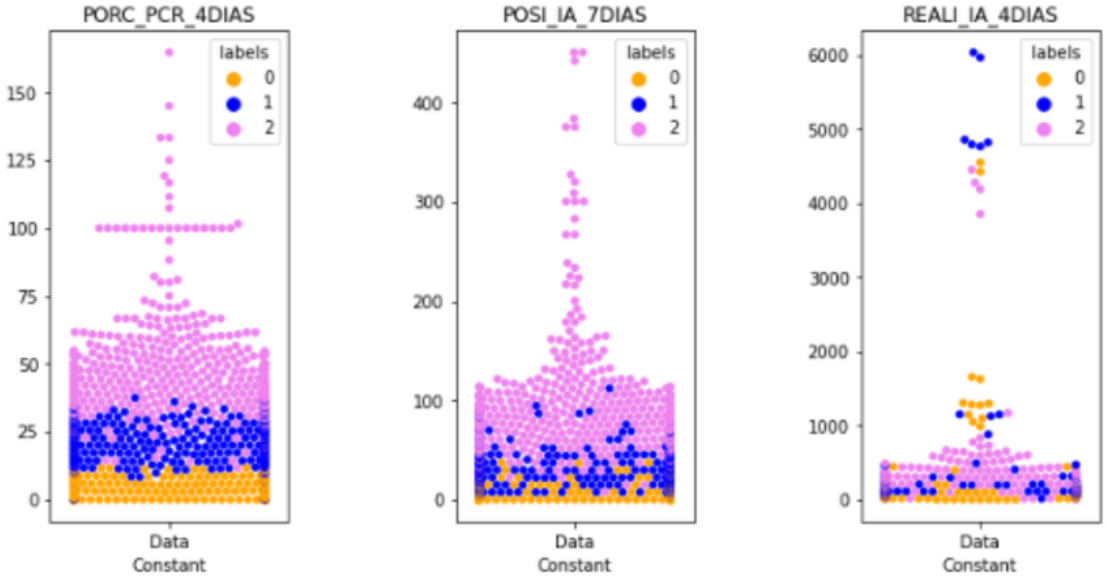


Figura 7.22: Resultados de librería gráfica *seaborn* periodo de 4 días

Período de 7 días:

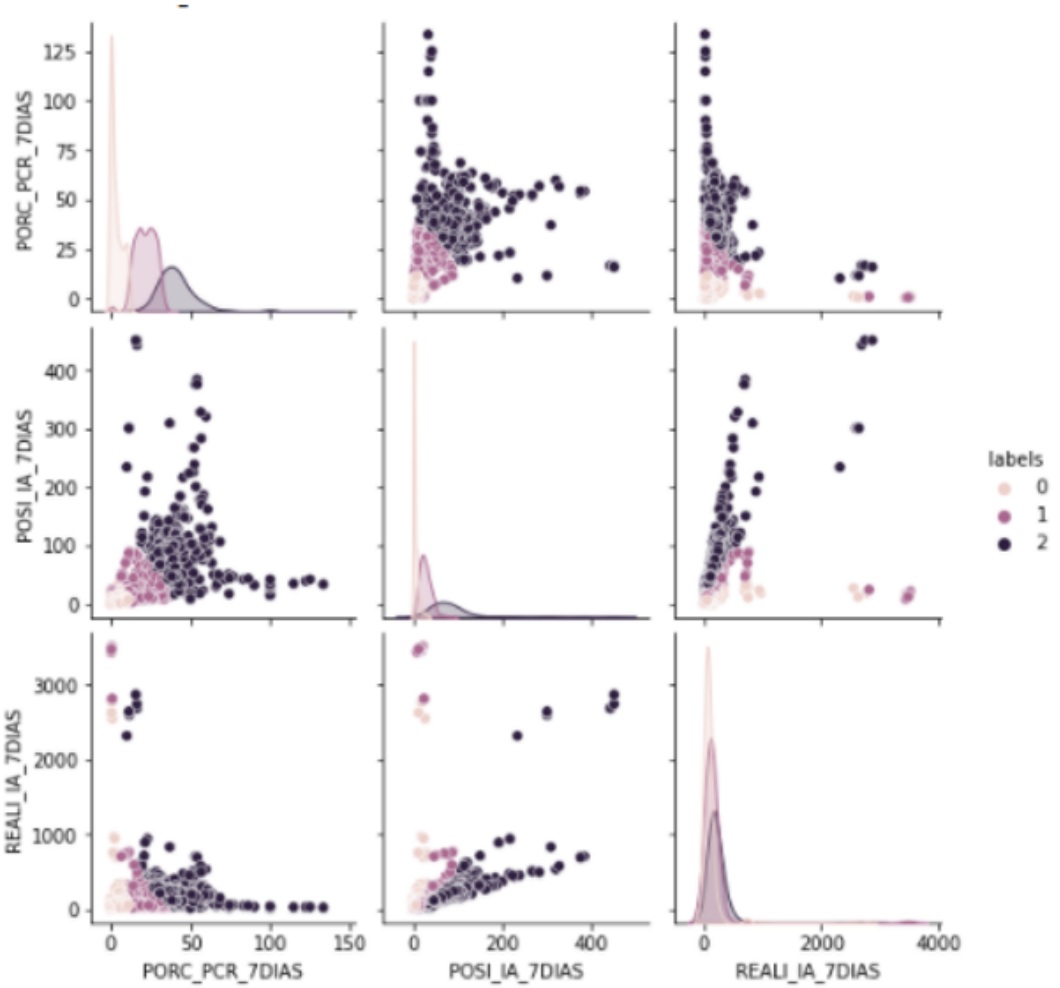


Figura 7.23: Resultados clustering periodo de 7 días

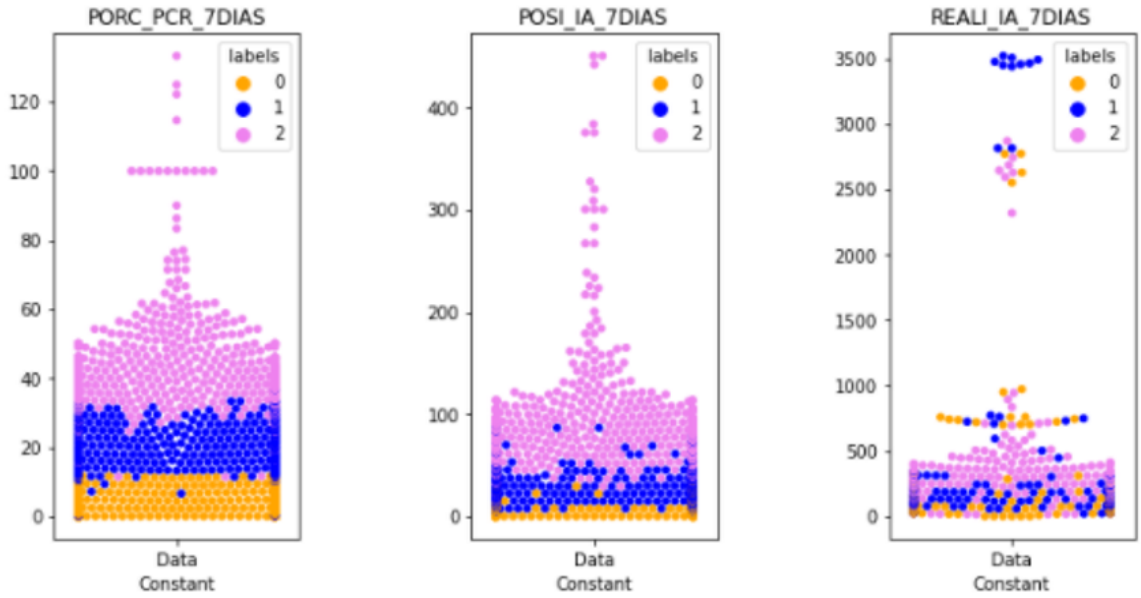


Figura 7.24: Resultados de librería gráfica *seaborn* periodo de 7 días

Período de 14 días:

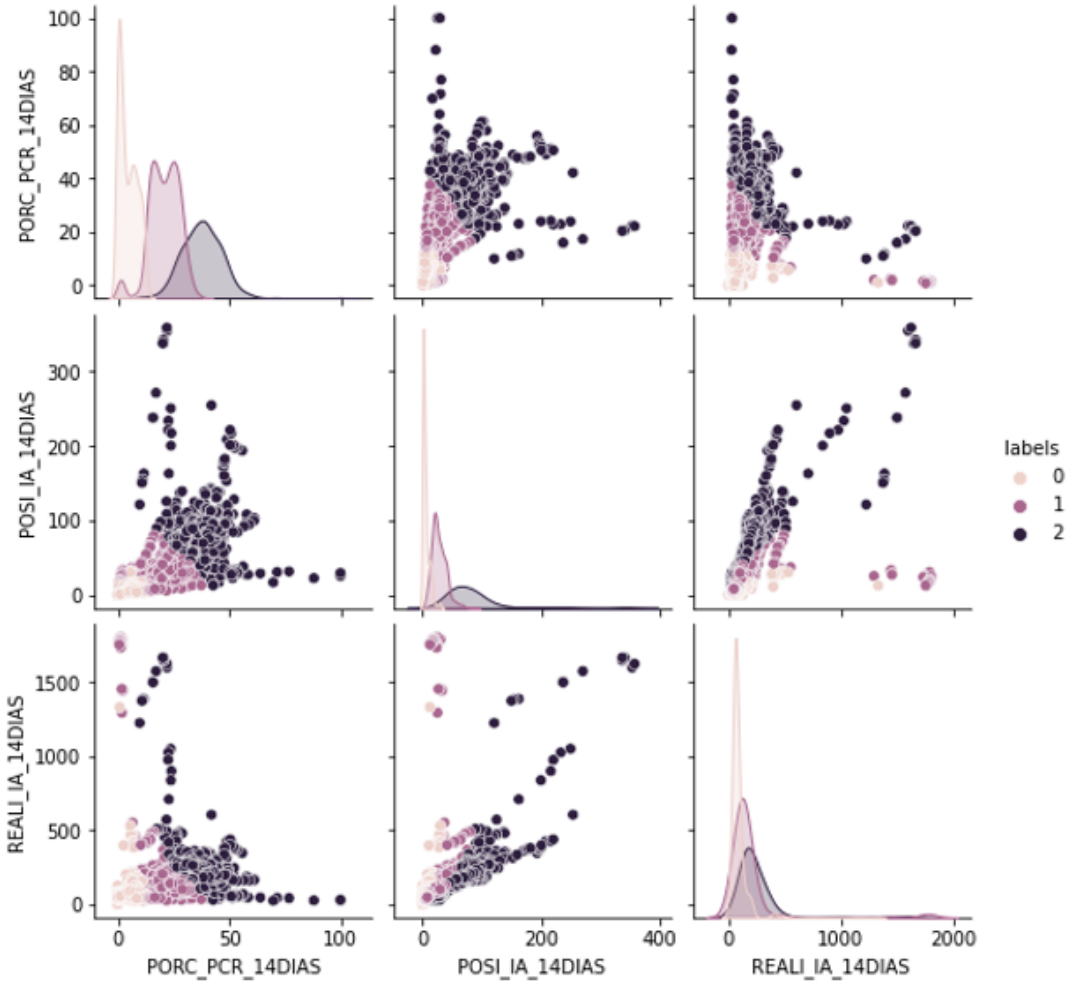


Figura 7.25: Resultados clustering periodo de 14 días

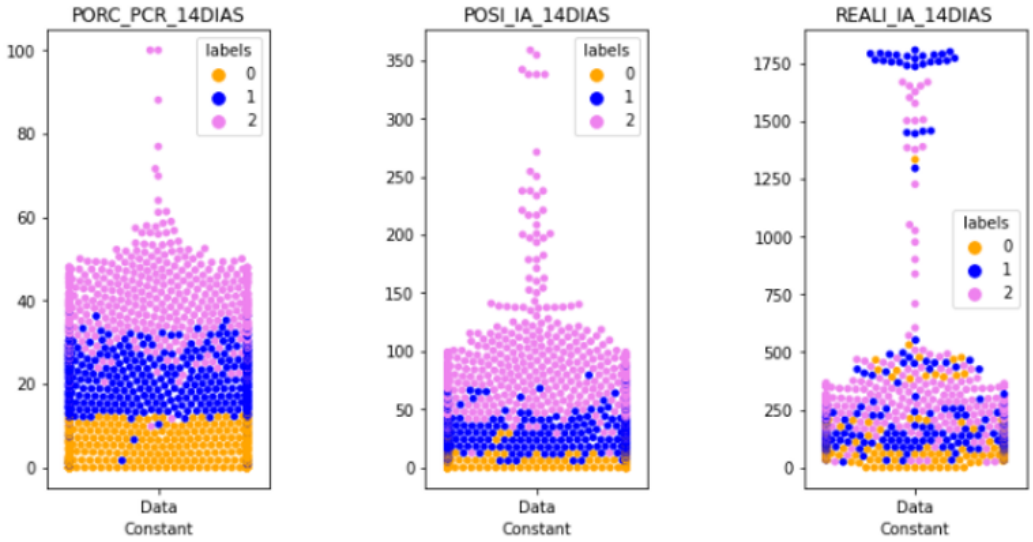


Figura 7.26: Resultados de librería gráfica *seaborn* periodo de1 14 días

Podemos apreciar como los resultados obtenidos en los tres periodos son bastante buenos, ya que el algoritmo consigue determinar de manera muy clara los tres tipos de subgrupos existentes para cada una de las variables. Los periodos de 7 y 14 días (Figuras 7.23, 7.24, 7.25 y 7.26) muestran una división más clara de los clusters que el periodo correspondiente a los 4 días Figuras (7.21 y 7.22).

Finalmente optamos por la elección del periodo de 14 días en lugar de el de 7 usando como principal motivo los análisis realizados en el apartado 6.1.3, donde se ha podido ver como las gráficas correspondientes a dicho periodo (Figuras 6.10, 6.16, 6.19 y 6.22) proporcionaban una visión clara de la tendencia del virus. Otra de las razones para la elección de este periodo se debe a que el tiempo de obtención de resultados respecto a la aplicación de medidas para las bajadas del virus es aproximadamente de 14-15 días, por lo que los resultados obtenidos con este cluster permitirán determinar qué medidas han tenido efecto en cada una de las zonas estudiadas.

7.3.5. Resultados de aplicación clustering a los datos de las zonas de salud estudiadas

Creación código de colores:

Una vez determinado el tipo de algoritmo de clustering a usar y el periodo de tiempo de los datos usados, se analizan los resultados obtenidos en cada una de las zonas estudiadas. Previo al análisis, se ha configurado a través de una función el etiquetado de los clusters generados para asignar colores adecuados a lo que dichos subgrupos representan:

- **Mal (2):** Estado correspondiente a una situación epidemiológica de riesgo extremo. El color dado a este estado ha sido el **rojo**.
- **Regular (1):** Estado correspondiente a una situación epidemiológica de riesgo moderado.El color dado a este estado ha sido el **amarillo**.
- **Bien (0):** Estado correspondiente a una situación epidemiológica por debajo de los niveles de alerta. El color dado a este estado ha sido el **verde**.

Figura 7.27: Leyenda de colores usada en clustering

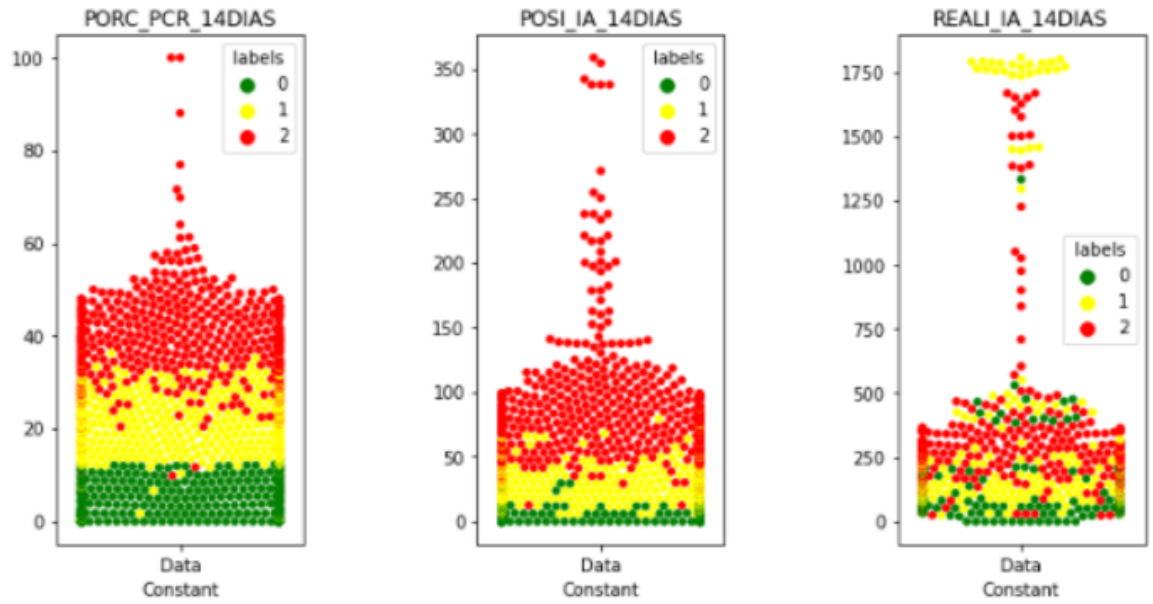


Figura 7.28: Clustering con código de colores establecido

Gráficas Clustering:

El análisis de los resultados se realizará mediante la generación de gráficas de cada zona de salud, donde se podrá observar el cluster correspondiente a los datos de cada zona en el periodo de tiempo de duración de la *segunda ola*.

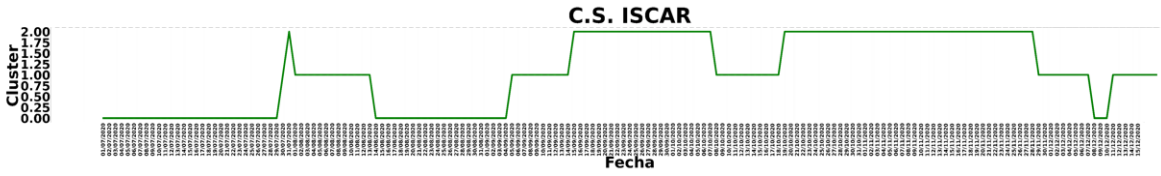


Figura 7.29: Resultados clustering en zona de Íscar

En la gráfica de ejemplo mostrada, se puede apreciar los distintos clusters a los que se han asignado los datos diarios de la zona de salud de Íscar (Figura 7.29). En el eje Y se encuentran las tres etiquetas dadas a cada cluster, 0 (bien), 1 (regular) y 2 (mal), teniendo en el eje X la fecha de cada dato. Se han creado gráficas como la mostrada de ejemplo para cada zona de salud, ya que gracias a ellas se ve de forma clara los distintos estados epidemiológicos obtenidos por el algoritmo de clustering en los que ha estado cada zona a lo largo de la *segunda ola*.

Validación de las gráficas:

La forma de la mayoría de las gráficas obtenidas se asemeja a la tendencia observada en el análisis preliminar realizado en el Apartado 6.1.3. Gráficas como la correspondiente a la Figura 6.22 muestran una tendencia similar a las vistas en el clustering.

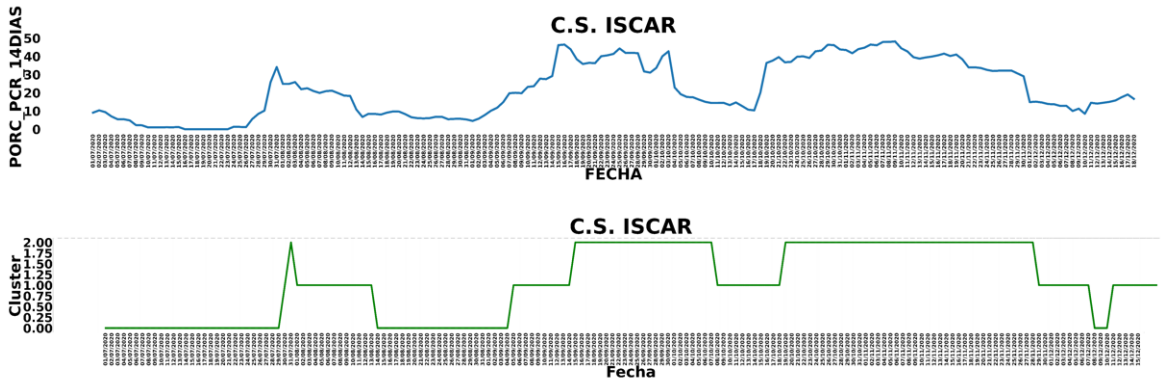


Figura 7.30: Comparación de gráficas clustering y porcentaje PCR en la zona de Íscar

Como se ve en el ejemplo mostrado (Figura 7.30), la gráfica de Porcentaje PCR en el periodo de 14 días en Íscar, posee una forma muy similar a la gráfica de los resultados del clustering en esa misma zona.

Análisis de resultados

Consideraciones previas:

Las bajadas y subidas obtenidas en cada gráfica son analizadas para determinar las diferentes causas, y así, poder obtener información sobre las medidas más efectivas y los factores que han provocado el aumento de los contagios.

Análisis de bajadas:

En el caso de las bajadas detectadas, habrá que situarse aproximadamente 14 días antes para obtener las medidas aplicadas que provocaron dicho descenso de contagios 14 días después. Esto se basa en el tiempo aproximado de actuación de las medidas empleadas frente al COVID-19.

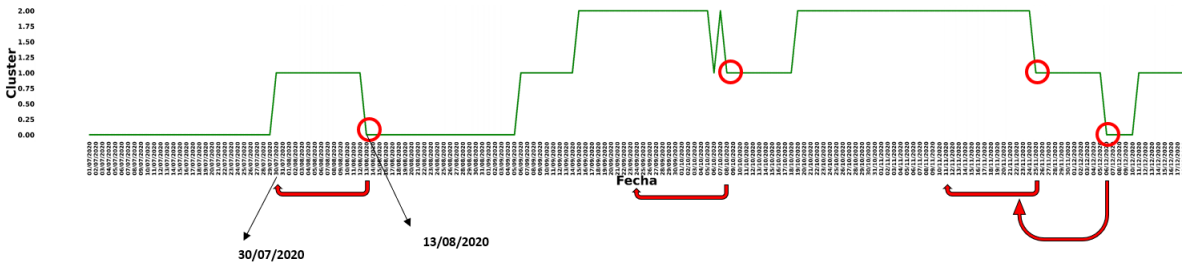


Figura 7.31: Obtención de las causas de las bajadas detectadas en gráfica clustering

Para las subidas no será necesario situarnos 14 días antes, ya que las causas de estas pueden tener un efecto más inmediato.

Outliers en gráficas:

Los picos obtenidos cuya duración no sea superior a la de dos días (inclusive), no serán tenidos en cuenta para el análisis de subidas y bajadas debido a que serán considerados como outliers.

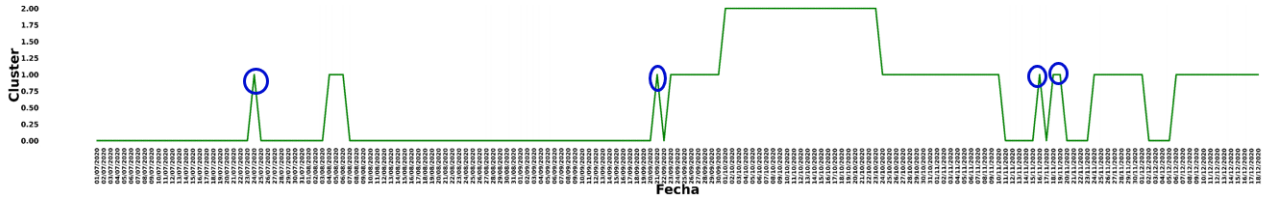


Figura 7.32: Tratamiento de picos obtenidos en gráficas clustering

Tipos de subidas y bajadas:

Para un mejor análisis se ha creado una calificación de los tipos de subidas y bajadas detectados:

- **Subidas:**
  - **Subida grande:** Cambio del cluster 0 al cluster 2.
  - **Subida intermedia:** Cambio del cluster 1 al cluster 2.
  - **Subida baja:** Cambio del cluster 0 al cluster 1.
- **Bajadas:**
  - **Bajada baja:** Cambio del cluster 1 al cluster 0.

- **Bajada intermedia:** Cambio del cluster 2 al cluster 1.
- **Bajada grande:** Cambio del cluster 2 al cluster 0.

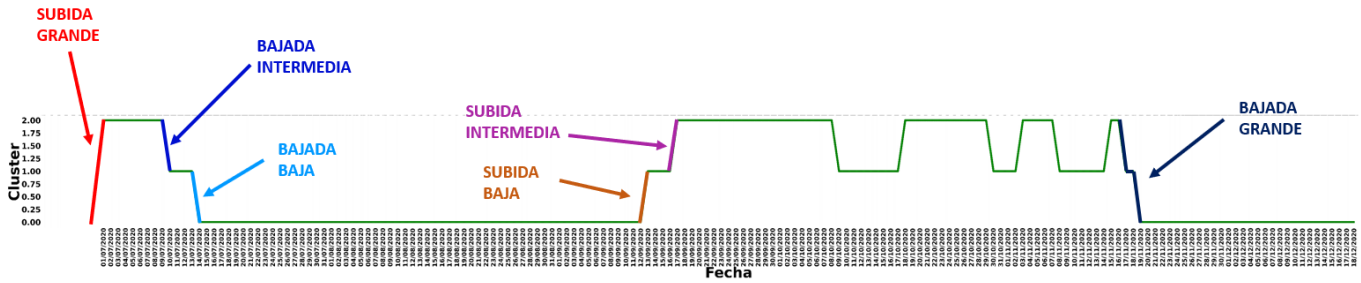


Figura 7.33: Gráfica tipos de bajadas y de subidas

Tablas de análisis por zona de salud:

Se ha recogido en una serie de tablas cada uno de los análisis hechos a los tipos de subidas y bajadas de cada zona, siendo la estructura de cada tabla la mostrada en la Figura 7.34.

NOMBRE DE LA ZONA DE SALUD	CONFINAMIENTO	FECHA INICIO DE CONFINAMIENTO EN ZS	FECHA FIN DE CONFINAMIENTO EN ZS
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
TIPO DE SUBIDA	FECHA EN LA QUE SE PRODUJO LA SUBIDA	N/A	CAUSAS QUE HAN PROVOCADO LA SUBIDA.
TIPO DE BAJADA	FECHA EN LA QUE SE PRODUJO LA BAJADA	FECHA 14 DÍAS ANTES EN LA QUE SE APLICARON LAS MEDIDAS O CAUSAS QUE PROVOCARÓN LA BAJADA	MEDIDAS O CAUSAS QUE HAN PROVOCADO LA BAJADA DE CONTAGIOS

Figura 7.34: Estructura de las tablas usadas

Tablas zonas de salud estudiadas:

CANTALEJO	CONFINAMIENTO:	22/08	4/09
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
SUBIDA BAJA	11/08		GRAN MOVILIDAD TOTAL Y OTROS
SUBIDA INTERMEDIA	17/08		GRAN MOVILIDAD TOTAL
BAJADA INTERMEDIA	30/08	16/08	MEDIDAS 17 AGOSTO
SUBIDA INTERMEDIA	05/11		MOVILIDAD POR PUENTE 1/11. GRAN MOVILIDAD POR TRABAJO.
BAJADA INTERMEDIA	10/11	26/10	MEDIDAS 17 Y 24 DE OCTUBRE
SUBIDA INTERMEDIA	08/12		AUMENTO DE MOVILIDAD OTROS POR FESTIVOS 7/12 Y 8/12

Figura 7.35: Tabla análisis zona de salud Cantalejo

ARANDA SUR	CONFINAMIENTO	07/08	21/08
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
SUBIDA BAJA	30/07		GRAN MOVILIDAD TOTAL
SUBIDA INTERMEDIA	06/08		GRAN MOVILIDAD TOTAL
BAJADA INTERMEDIA	23/08	09/08	MEDIDAS CONFINAMIENTO. BAJADA DRASTICA DE MOVILIDAD TOTAL EN CONFINAMIENTO.
BAJADA BAJA	03/09	20/08	MEDIDAS 17 Y 21 AGOSTO. EFECTOS CONFINAMIENTO. MOVILIDAD
SUBIDA BAJA	06/09		FIN CONFINAMIENTO. AUMENTO DE MOVILIDAD TOTAL TRAS CONFINAMIENTO
SUBIDA INTERMEDIA	18/10		MOVILIDAD POR PUENTE 12/10. GRAN MOVILIDAD TRABAJO
BAJADA INTERMEDIA	28/11	14/11	MEDIDAS NOVIEMBRE. CAIDA DE MOVILIDAD TOTAL.
SUBIDA INTERMEDIA	13/12		MOVILIDAD PUENTE 7/12, 8/12. GRAN MOVILIDAD POR TRABAJO DIA 10/12.

Figura 7.36: Tabla análisis zona de salud Aranda Sur

ARANDA RURAL	CONFINAMIENTO	07/08	21/08
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
SUBIDA BAJA	3/08		GRAN MOVILIDAD TOTAL
SUBIDA INTERMEDIA	07/08		GRAN MOVILIDAD TOTAL
BAJADA GRANDE	02/09	19/08	MEDIDAS 17 AGOSTO. EFECTOS CONFINAMIENTO. BAJADA DE MOVILIDAD GRANDE POR CONFINAMIENTO.
SUBIDA GRANDE	07/09		FIN MEDIDAS CONFINAMIENTO. AUMENTO DE LA MOVILIDAD TOTAL TRAS EL CONFINAMIENTO
BAJADA INTERMEDIA	24/09	10/09	DIAS 11/09, 12/09, 14/09 Y 15/09 CAIDA DE MOVILIDAD TOTAL. FIESTAS LOCALES NO TIENEN REPERCUSION EN CONTAGIOS
SUBIDA INTERMEDIA	20/10		GRAN MOVILIDAD POR TRABAJO- MOVILIDAD PUENTE 12/10
BAJADA INTERMEDIA	24/10	10/10	BAJADA DE MOVILIDAD TOTAL Y MOVILIDAD POR TRABAJO
SUBIDA INTERMEDIA	27/10		MOVILIDAD POR TRABAJO ALTA
BAJADA INTERMEDIA	27/11	13/11	MEDIDAS NOVIEMBRE. CAIDA DE MOVILIDAD TOTAL DESDE OCTUBRE

Figura 7.37: Tabla análisis zona de salud Aranda Rural

ARANDA NORTE	CONFINAMIENTO	07/08	21/08
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
BAJADA BAJA	06/07	22/06	BAJADA DE MOVILIDAD TOTAL DIAS PREVIOS (20/06, 21/06). MEDIDAS PRIMERA OLA
SUBIDA BAJA	30/07		MOVILIDAD TOTAL ALTA(CASA-TRABAJO-OTROS)
SUBIDA INTERMEDIA	04/08		MOVILIDAD TOTAL ALTA(CASA-TRABAJO-OTROS)
BAJADA INTERMEDIA	29/08	15/08	MEDIDAS CONFINAMIENTO. BAJADA DE MOVILIDAD TOTAL DURANTE CONFINAMIENTO.
SUBIDA INTERMEDIA	12/10		MOVILIDAD PUENTE 12/10. MOVILIDAD TRABAJO
BAJADA INTERMEDIA	26/11	14/11	MEDIDAS NOVIEMBRE, BAJADA DE MOVILIDAD TOTAL DESDE OCTUBRE

Figura 7.38: Tabla análisis zona de salud Aranda Norte

MIRANDA OESTE	CONFINAMIENTO	26/09	24/10
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
BAJADA BAJA	8/07	24/06	EFECTO MEDIDAS PRIMERA OLA
SUBIDA BAJA	14/08		GRAN MOVILIDAD TOTAL Y OTROS. MOVILIDAD POR FIESTAS 15/08.
SUBIDA INTERMEDIA	06/09		GRAN MOVILIDAD TOTAL EN AUMENTO A PARTIR DE SEPTIEMBRE. AUMENTO MOVILIDAD POR FIESTAS LOCALES 31/08.
BAJADA INTERMEDIA	01/11	18/10	MEDIDAS 17 OCTUBRE. MEDIDAS CONFINAMIENTO. BAJADA DE MOVILIDAD POR CONFINAMIENTO. MOVILIDAD TOTAL BAJA A PARTIR DE NOVIEMBRE

Figura 7.39: Tabla análisis zona de salud Miranda Oeste

MIRANDA ESTE	CONFINAMIENTO	26/09	24/10
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
BAJADA BAJA	11/07	27/06	EFECTOS MEDIDAS PRIMERA OLA.
SUBIDA BAJA	15/08		MOVILIDAD POR FIESTAS LOCALES. MOVILIDAD TOTAL Y OTROS ALTA
SUBIDA INTERMEDIA	18/09		MOVILIDAD TOTAL, CASA Y OTROS MUY ALTA MOVILIDAD GRANDE POR FIESTAS 12/09.
BAJADA INTERMEDIA	07/11	24/10	MEDIDAS CONFINAMIENTO. MOVILIDAD TOTAL Y OTROS EN CONFINAMIENTO REDUCIDA

Figura 7.40: Tabla análisis zona de salud Miranda Este

MIRANDA DEL CASTANAR	CONFINAMIENTO	26/09	13/10
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
SUBIDA BAJA	06/09		GRAN MOVILIDAD TOTAL.
SUBIDA INTERMEDIA	16/09		GRAN MOVLIDAD TOTAL
BAJADA INTERMEDIA	09/10	25/09	MEDIDAS CONFINAMIENTO. GRAN CAIDA DE MOVILIDAD DURANTE CONFINAMIENTO.
SUBIDA INTERMEDIA	18/10		AUMENTO DE LA MOVILIDAD DIAS FINALES DE CONFINAMIENTO (NO CUMPLIDO AL FINAL)
BAJADA INTERMEDIA	09/11	26/10	MEDIDAS 24 OCTUBRE.
SUBIDA INTERMEDIA	14/11		EFECTOS DEL PUENTE 1/11
BAJADA GRANDE	19/11	05/11	MEDIDAS 6 DE NOVIEMBRE. MOVILIDAD TOTAL BAJA

Figura 7.41: Tabla análisis zona de salud Miranda del Castañar

MOTA DEL MARQUES	CONFINAMIENTO	13/10	24/10
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
SUBIDA BAJA	23/07		MOVILIDAD OTROS ALTA
BAJADA BAJA	31/07	17/07	MOVILIDAD TOTAL EN DESCENSO RESPECTO A DIAS ANTERIORES
SUBIDA BAJA	03/08		MOVILIDAD OTROS ALTA
BAJADA BAJA	07/08	23/07	LIGERO DESCENSO DE MOVILIDAD TOTAL
SUBIDA BAJA	09/08	25/07	MOVILIDAD TOTAL ALTA
BAJADA BAJA	18/08	04/08	MEDIDAS 17 AGOSTO
SUBIDA BAJA	20/09		MOVILIDAD POR FIESTAS LOCALES. MOVILIDAD TOTAL Y CASA DIA 10/09 MUY ALTA
SUBIDA INTERMEDIA	30/09		MOVILIDAD CASA ALTA
BAJADA INTERMEDIA	24/10	10/10	MEDIDAS CONFINAMIENTO. MOVILIDAD TOTAL Y CASA BAJA
BAJADA BAJA	13/11	30/10	MEDIDAS 24 Y 30 OCTUBRE. MOVILIDAD TOTAL BAJA
SUBIDA BAJA	15/11		MOVILIDAD POR PUENTE 1/11

Figura 7.42: Tabla análisis zona de salud Mota del Marqués

PEÑAFIEL	EºCONFINAMIENTO	22/09	06/10
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
BAJADA BAJA	2/07	18/06	EFECTOS MEDIDAS PRIMERA OLA
SUBIDA BAJA	26/08		AUMENTO DE MOVILIDAD POR FIESTAS LOCALES. MOVILIDAD OTROS ALTA. GRAN MOVILIDAD ENTRE MUNICIPIOS PEÑAFIEL Y PESQUERA DE DUERO
SUBIDA INTERMEDIA	16/09		MOVILIDAD TOTAL ALTA DIAS PUNTUALES 4 Y 12 GRAN MOVILIDAD POR TRABAJO ENTRE MUNICIPIOS PEÑAFIEL Y PESQUERA DE DUERO
BAJADA INTERMEDIA	25/09	11/09	BAJADA DE MOVILIDAD TOTAL RESPECTO A DIAS ANTERIORES BAJADA DE MOVILIDAD ENTRE MUNICIPIOS PEÑAFIEL Y PESQUERA DE DUERO
BAJADA BAJA	10/12	26/11	BAJADA DE MOVILIDAD TOTAL BAJADA MOVILIDAD POR TRABAJO ENTRE MUNICIPIOS PEÑAFIEL Y PESQUERA DE DUERO

Figura 7.43: Tabla análisis zona de salud Peñafiel

MEDINA CAMPO URBANO	CONFINAMIENTO	29/09	13/10
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
SUBIDA BAJA	07/09		MOVILIDAD OTROS MUY ALTA
SUBIDA INTERMEDIA	22/09		MOVILIDAD TOTAL EN AUMENTO. MOVILIDAD OTROS ALTA
BAJADA INTERMEDIA	15/10	01/10	MEDIDAS CONFINAMIENTO. BAJADA DE MOVILIDAD TOTAL Y OTROS
SUBIDA INTERMEDIA	29/10		FIN MEDIDAS CONFINAMIENTO. MOVILIDAD OTROS AUMENTA POR PUENTE 1/11
BAJADA INTERMEDIA	15/12	01/12	MOVILIDAD TOTAL BAJA

Figura 7.44: Tabla análisis zona de salud Medina del Campo urbano

MEDINA CAMPO RURAL	CONFINAMIENTO	29/09	13/10
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
BAJADA BAJA	4/07	20/06	EFECTOS MEDIDAS PRIMERA OLA
SUBIDA BAJA	04/09		MOVILIDAD TOTAL ALTA. GRAN MOVILIDAD POR OTROS Y TRABAJO
SUBIDA INTERMEDIA	15/09		MOVILIDAD TOTAL ALTA. GRAN MOVILIDAD OTROS
BAJADA INTERMEDIA	24/10	10/10	MEDIDAS CONFINAMIENTO. BAJADA DE MOVILIDAD TOTAL
SUBIDA INTERMEDIA	28/10		FIN MEDIDAS CONFINAMIENTO
BAJADA INTERMEDIA	03/12	19/11	MOVILIDAD TOTAL BAJA

Figura 7.45: Tabla análisis zona de salud Medina del Campo rural.

ISCAR	CONFINAMIENTOS:	02/08/2020-16/08/2020	18/09/2020-17/10/2020
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
SUBIDA BAJA	29/07		GRAN MOVILIDAD TOTAL Y OTROS AUMENTO MOVILIDAD POR TRABAJO Y OTROS ENTRE PEDRAJAS E ISCAR
BAJADA BAJA	14/08	31/07	MEDIDAS PRIMER CONFINAMIENTO. BAJADA DE MOVILIDAD TOTAL POR CONFINAMIENTO BAJADA MOVILIDAD POR TRABAJO ENTRE PEDRAJAS E ISCAR
SUBIDA BAJA	04/09		FIN MEDIDAS PRIMER CONFINAMIENTO AUMENTO DE MOVILIDAD(TRABAJO-OTROS) AUMENTO MOVILIDAD POR TRABAJO ENTRE PEDRAJAS E ISCAR
SUBIDA INTERMEDIA	14/09		MOVILIDAD TOTAL, TRABAJO Y OTROS ALTA.
BAJADA INTERMEDIA	09/10	25/09	MEDIDAS SEGUNDO CONFINAMIENTO BAJADA DE MOVILIDAD TOTAL (OTROS NO) BAJADA MOVILIDAD POR TRABAJO ENTRE PEDRAJAS E ISCAR
SUBIDA INTERMEDIA	19/10		CONFINAMIENTO NO RESPETADO (MOVILIDAD OTROS ALTA) BAJADA MOVILIDAD POR OTROS ENTRE PEDRAJAS E ISCAR (CONFINAMIENTO A PEDRAJAS)
BAJADA INTERMEDIA	30/11	16/11	MEDIDAS NOVIEMBRE BAJADA MOVILIDAD POR TRABAJO Y OTROS ENTRE PEDRAJAS E ISCAR
BAJADA BAJA	08/12	24/11	MOVILIDAD TOTAL BAJA RESPECTO A DIAS ANTERIORES BAJADA MOVILIDAD POR TRABAJO Y OTROS ENTRE PEDRAJAS E ISCAR
SUBIDA BAJA	10/12		MOVILIDAD TOTAL SUBE. MOVILIDAD POR PUENTE 7/12, 8/12 AUMENTO MOVILIDAD POR TRABAJO ENTRE PEDRAJAS E ISCAR

Figura 7.46: Tabla análisis zona de salud Íscar.

Conclusiones obtenidas:

Como se puede observar en las tablas de análisis creadas, una de las principales causas de subidas y bajadas ha sido la movilidad. En muchas zonas, los aumentos de movilidad han sido provocados por periodos estivales como el verano o la proximidad a festivos locales o nacionales, junto con el fin o ausencia de medidas restrictivas (confinamientos). Este aumento de la movilidad ha provocado la mayoría de las subidas detectadas por el algoritmo de clustering y por tanto el aumento de los contagios en cada zona.

Tal y como se había analizado en el apartado 6.2.2. En las zonas de Íscar y Peñafiel (Figuras 7.46 y 7.43) aquella movilidad realizada entre municipios ha sido determinante en las subidas y bajadas detectadas.

Medidas de limitación de la movilidad como aquellas implantadas con los confinamientos, cierres perimetrales de la CA, cierre de hostelería o el toque de queda han sido causa de muchas de las bajadas detectadas, debido a su impacto en el número de entradas o movilidad interna de cada zona.

Otras restricciones como las relacionadas con la limitación de la actividad social o el número de reuniones también han sido relevantes en algunas de las zonas estudiadas.

Por tanto, podemos concluir que el factor más relevante a la hora de reducir el número de contagios (como ya se presuponía) es la movilidad. Aquellas medidas que permitan reducir el número de entradas a cada zona serán consideradas de gran relevancia o efectividad en muchas de las zonas estudiadas.

Capítulo 8

Modelos de aprendizaje: Boosting

8.1. Introducción

Una vez aplicado el método de aprendizaje clustering sobre el conjunto de datos usado, se han obtenido una serie de clusters que permiten clasificar los datos epidemiológicos de cada zona en función de su gravedad (bien, regular o mal). Esta clasificación en subgrupos nos ofrece la visualización de las subidas y bajadas de contagios en cada una de las zonas, pudiendo analizar las causas en detalle.

Toda esta información es muy útil, pero no es suficiente para poder lograr el verdadero objetivo del proyecto: la obtención de aquellas medidas más efectivas en cada zona. Es por ello, por lo que a partir de la información dada por los clusters sobre las distintas subidas y bajadas, se ha decidido crear un modelo de aprendizaje supervisado que permita obtener qué medidas han sido realmente las más efectivas en cada zona y cuáles no.

La elección de un modelo de aprendizaje supervisado viene condicionada por la información y pruebas preliminares obtenidas (movilidad de entrada a las zonas de salud,medidas aplicadas, resultados de clustering...). Ahora se posee información específica de cada zona de salud estudiada, lo que permite validar los resultados obtenidos por el modelo de aprendizaje utilizado.

El modelo escogido para la determinación de las medidas más efectivas de cada zona ha sido el modelo Gradient Boosting Tree [34]. La elección de este modelo se debe a su funcionamiento basado en ensembles. A continuación profundizaremos en el concepto de ensemble y Boosting (tipo de ensemble), necesarios para la comprensión del funcionamiento del modelo usado.

8.2. Ensembles

Los ensembles son técnicas de agrupación donde múltiples modelos de Machine Learning combinan sus salidas para la resolución de problemas complejos. Cada modelo se encarga de obtener una predicción que posteriormente es combinada con los resultados de los otros modelos para obtener una única predicción [16]. Por ejemplo, en un ensemble compuesto por 15 árboles donde se obtienen 10 predicciones correspondientes a un cliente que va a darse de baja y 5 que no, la predicción final obtenida de nuestro ensemble será que dicho cliente se dará de baja en nuestro sistema [13].

En la Figura 8.1 se muestra un esquema básico sobre la idea de ensemble, podemos observar como los distintos modelos que forman los ensembles hacen uso de un conjunto de datos que puede ser igual para todos los modelos o estar repartido entre cada uno.

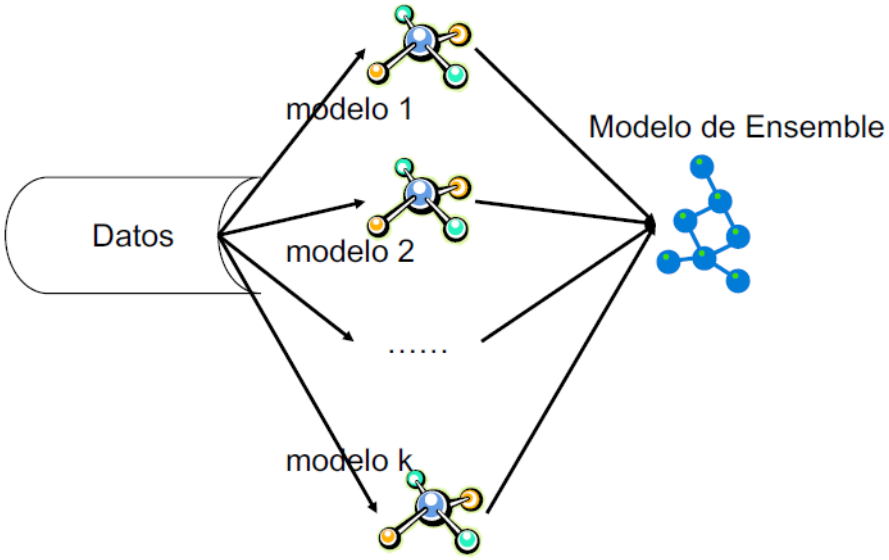


Figura 8.1: Formación de ensembles [7].

La idea principal de los ensembles se encuentra en la selección de modelos bases a agrupar [33], siendo muchas veces escogido un tipo único de algoritmo de aprendizaje en todos los modelos que componen el ensemble (ensemble homogéneo). Existen casos en los que se emplean distintos algoritmos de aprendizaje en los modelos agrupados (ensemble heterogéneo), combinando así algoritmos que individualmente no obtienen buenos resultados, pero que combinados con otros tipos mejoran mucho. Un aspecto muy importante en la elección de modelos es la combinación que se obtendrá en un futuro. Si elegimos modelos con un sesgo (bias) y varianza distinto se deberá usar un método de agregación (Bagging, Boosting, Stacking...) que tienda a reducir ambos parámetros para mantenerlos los más bajo posibles y así evitar el denominado problema sesgo-varianza.

La clave de los métodos de ensemble radica en que los modelos que forman a cada uno de estos sean lo más diversos posibles, haciendo que los errores no se correlacionen y obteniendo un resultado final construido a partir de pequeños resultados dados por dichos modelos (mayor generalización).

8.2.1. Ventajas y desventajas de los ensembles

En este apartado se indican todas aquellas ventajas y desventaja de la técnica de agrupación ensemble usada en nuestro proyecto.

Ventajas:

- Mejora de rendimiento predictivo. Las predicciones basadas en distintos modelos tienen mayor estabilidad debido a la reducción de la varianza y el ruido (influencia de outliers).
- Aplicación a problemas de clasificación y regresión.
- Menor esfuerzo en la limpieza y preprocesamiento de datos usados.
- Gran escalabilidad.
- Efectivos en la reducción del *overfitting* y bias mediante los distintos tipos de ensemble (Bagging y Boosting), pudiendo así generalizar correctamente nuevas observaciones.

Desventajas:

- Dificultad en la interpretación de los resultados obtenidos (no se permite su interpretación gráfica).
- Gran exigencia computacional.
- Incapacidad de extrapolación fuera del rango de predictores obtenidos en los datos de entrenamiento.
- Perdida de información al usar predictores basados en datos continuos.

8.2.2. Tipos

Los tipos de ensemble más utilizados son [7]:

- **Votación:** Tipo de ensemble compuesto por distintos algoritmos de Machine Learning que entrenan con los mismos datos dando cada uno de los algoritmos una salida. Se elegirá como resultado final la salida que sea votada por la mayoría.
- **Bagging:** Ajuste de múltiples modelos teniendo cada uno un subconjunto específico de datos de entrenamiento. El resultado de todos los modelos usados (media de predicciones) se combina para dar un resultado final. Un ejemplo claro de este tipo de ensemble son los modelos *Random Forest*.
- **Boosting:** Ajuste secuencial de múltiples modelos sencillos denominados weak learners, de tal forma que cada uno de estos modelos aprenda de los errores del anterior. Un ejemplo de este tipo de ensembles son los Gradient Boosting Trees, donde esos weak learners están compuestos por árboles con pocas ramificaciones.
- **Stacking:** Ajuste de varios modelos distintos que entrenan con el mismo conjunto de datos y cuya salida es usada como entrada de un modelo encargado de tomar la decisión final.

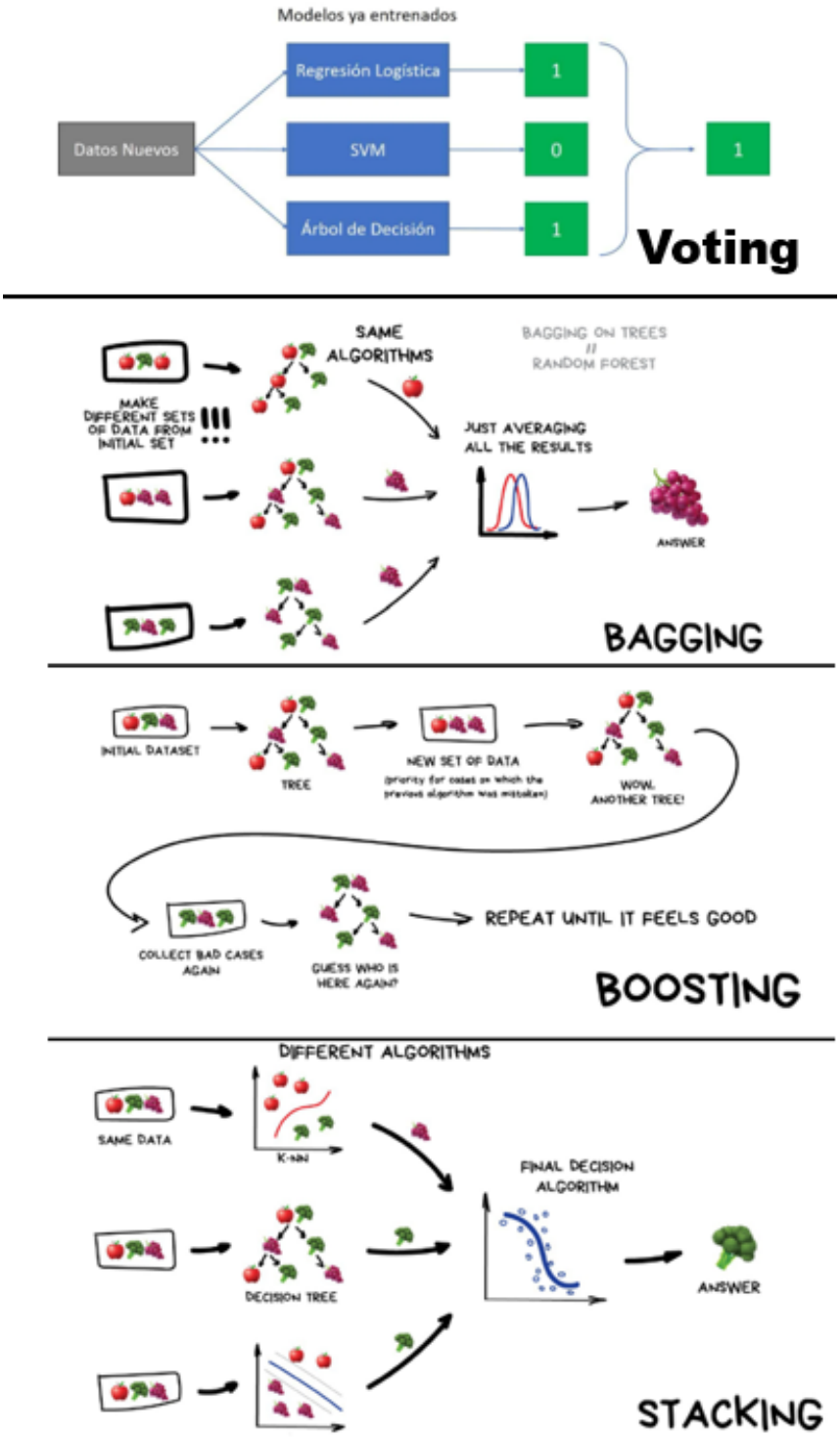


Figura 8.2: Tipos de ensembles

8.2.3. Problema sesgo-varianza

Al igual que la mayoría de métodos de Machine Learning, los modelos que componen los ensembles sufren el problema de equilibrio entre los conceptos de bias (sesgo) y varianza:

Bias:

El término *bias* o sesgo es usado en Machine Learning para referirnos a la diferencia existente entre las predicciones obtenidas y los valores reales. Refleja la capacidad del modelo para conocer la verdadera relación existente entre la variable predictora y la variable de respuesta. Generalmente, los algoritmos lineales poseen un sesgo alto ya que su aprendizaje y entendimiento es sencillo, pero su flexibilidad es muy reducida (rendimiento bajo en problemas complejos). Entre estos algoritmos encontramos: regresión lineal, análisis de discriminante lineal y regresión logística [5].

Varianza:

El término *varianza* se refiere al grado en el que el modelo varía según los datos utilizados en el entrenamiento. Idealmente, el modelo no debe modificarse demasiado debido a pequeños cambios en los datos de entrenamiento. Si esto sucede es porque el modelo almacena los datos en lugar de aprender la verdadera relación entre la variable predictora y la variable de respuesta. Aquellos algoritmos de aprendizaje automático con una gran varianza se sienten fuertemente influenciados por las especificaciones de los datos de entrenamiento, entre estos encontramos: árboles de decisión, support vector machines y K-NN.

Ensembles en problema sesgo-varianza:

El objetivo de cualquier algoritmo o modelo de aprendizaje automático será el de obtener un equilibrio entre sesgo y varianza. Para poder lograr dicho objetivo se hace uso de los ensembles, que mediante la obtención de un único modelo a partir de la combinación de otros consigue lograr un equilibrio entre el sesgo y la varianza. Los tipos de ensemble más utilizados para la obtención de dicho equilibrio son el Bagging y el Boosting.

- **Bagging:** Uso de modelos con muy poco bias y gran varianza, mediante su agregación se lográ reducir la varianza sin apenas aumentar el bias.
- **Boosting:** Uso de modelos con muy poca varianza y gran bias, ajustando de forma secuencial los modelos se logra reducir el bias sin influir apenas en la varianza.

8.3. Boosting

En nuestro proyecto se hace uso del tipo de ensemble Boosting, que tal y como podemos ver en el ejemplo mostrado en la Figura 8.3, basa su funcionamiento en la mejora de los errores obtenidos por modelos anteriores, denominados weak learners (clasificador débil), repitiendo el proceso hasta obtener los mejores resultados posibles. Dentro de nuestro proyecto haremos uso de estos clasificadores (árboles) para el entrenamiento y obtención aquellas medidas más efectivas, corrigiendo cada modelo los errores generados por anteriores (fallos en determinación de medidas eficaces), para finalmente obtener aquellas medidas más eficaces en cada una de las zonas estudiadas.

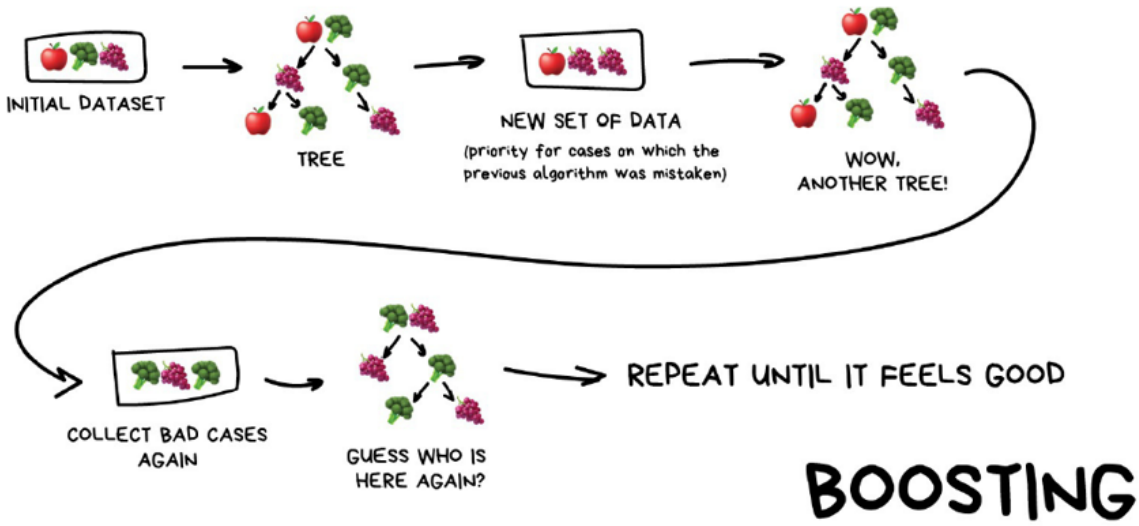


Figura 8.3: Funcionamiento modelos GBT

Profundizando en los tipo de ensemble Boosting, observamos como su funcionamiento esta basado en la agregación de clasificadores débiles (weak learners), teniendo cada uno de estos diferentes pesos en función de los resultados (predicciones) de clasificadores anteriores [40]. Por tanto, la agregación de un clasificador débil al ensemble provoca un cambio en la estructura de los pesos, haciendo que los datos mal clasificados ganen peso y aquellos correctos lo pierdan; los weak learners se centrarán en aquellos casos que fueron mal clasificados (medidas aplicadas) por los clasificadores anteriores.

8.3.1. Ventajas y desventajas del Boosting

Ventajas:

- Rapidez del entrenamiento.
- Implementación fácil y sencilla.
- Adaptación a datos usados en el proyecto (resultados clustering y medidas aplicadas).

Desventajas:

- Sensibilidad a datos incompletos (fallos con hipótesis complejas o débiles).
- Ralentización del algoritmo por aparición de puntos de corte.
- Tendencia al sobreajuste (*overfitting*), se necesita de una forma de obtención del número óptimo de modelos a incluir.

8.3.2. Estrategias Boosting

En este apartado se describen las distintas técnicas y estrategias de Boosting en las que se fundamentan los modelos Gradient Boosting Trees:

AdaBoost

El principal algoritmo usado en Boosting es AdaBoost, algoritmo capaz de aprender a partir de clasificadores débiles (weak learners), sin embargo existen muchos otros como LPBoost, BrownBoost, XGBoost, MadaBoost o LogitBoost, estando estos dentro del marco de AnyBoost, el cual basa su funcionamiento en el descenso del gradiente en un espacio funcional (Gradient Boosting). A continuación veremos el funcionamiento del algoritmo AdaBoost (problema de clasificación binaria), el cual proporciona la base para entender el funcionamiento del modelo Gradient Boosting usado en el proyecto. El funcionamiento de AdaBoost esta basado en tres puntos principales:

- Uso del tipo de modelo **weak learner**, el cual es capaz de predecir la variable respuesta con un porcentaje de acierto superior al dado de manera aleatoria. Este **weak learner** se traduce en un árbol de pocos nodos (en el caso de los árboles de regresión).

- Mismo peso de inicio para todas las observaciones que forman el set de entrenamiento.
- Codificación de la variable respuesta en dos clases como +1 y -1.

Una vez establecidos los puntos principales, se inicia un proceso de iteraciones. La primera iteración ajusta un **weak learner**, empleando los datos de entrenamiento y pesos iniciales, usando posteriormente este (ya almacenado y ajustado) para la predicción e identificación de las observaciones de entrenamiento.

Este proceso hará que los pesos de las observaciones se actualicen disminuyendo el peso de aquellas bien clasificas y aumentando el de las mal clasificadas. Por lo tanto, el **weak learner** obtendrá un peso total proporcional al total de aciertos, estableciendo así una mayor influencia en el ensemble en función a dichos aciertos.

Así se podrá realizar otra iteración en la cual se llamará de nuevo al **weak learner** para su ajuste, usando esta vez los pesos ya actualizados de la iteración anterior. El nuevo **weak learner** obtenido se almacenará obteniendo un nuevo modelo del ensemble.

Este proceso se ira repitiendo M veces hasta generar un total de **M weak learner**. La clasificación de nuevas observaciones se realizará mediante la obtención de la predicción de cada uno de los **weak learners** que forman el ensemble. De esta manera, se agregarán los resultados de dichos **weak learners** , ponderando el peso de cada uno acorde al peso que se le asigna en el ajuste. El objetivo de esta estrategia es la predicción correcta por parte de cada nuevo **weak learner** en base a las observaciones que los anteriores no han sido capaces de predecir.

Binning

Uno de los principales problemas que encontramos en el Boosting son los *thresholds* o puntos de corte. Dichos puntos producen un cuello de botella o ralentización de nuestro algoritmo, llegando a obtener procesos de compilación muy grandes.

La estrategia usada para su búsqueda y detección es la denominada discrete binning. Dicha estrategia se basa en una técnica de preprocesamiento de datos utilizada para reducir los efectos de pequeños errores de observación. Los valores originales de los datos que caen en un pequeño intervalo dado (*bin*), se sustituyen por un valor representativo de ese intervalo (a menudo el valor central). El binning de datos es una forma de agrupar números de valores más o menos continuos en un número menor de *bins*.

Esta técnica es característica de modelos **GBT** como HistGradientboosting o XGBT, que al implementarla adquieren mayor velocidad en el proceso de aprendizaje.

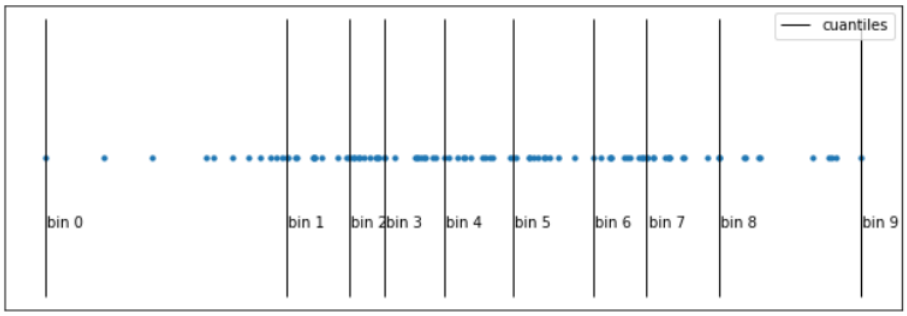


Figura 8.4: Ejemplo binning. [34]

Early stopping

Como se ha mencionado anteriormente, uno de los principales problemas de los ensembles es el sobreajuste o *overfitting* provocado durante el entrenamiento. Este *overfitting* está fuertemente relacionado con el número de **weak learners** o árboles usados. Es por ello por lo que al usar Boosting se debe poseer una técnica que nos permita determinar el número de árboles o **weak learners** ideales a usar de cara a obtener los mejores resultados y evitar el sobreajuste.

Como solución a este problema, en la mayoría de implementaciones y modelos se utiliza la denominada *early stopping* o parada temprana, que nos permite parar el proceso de ajuste de nuestro modelo justo en el momento anterior a dejar de obtener los mejores resultados.

Dentro de nuestro proyecto esta parada temprana será controlada mediante los argumentos **validation_fraction , n_iter_no_change y tol** .

8.4. Modelo Gradient Boosting Trees

El modelo Gradient Boosting Trees está compuesto por un conjunto (ensemble) de árboles de decisión individuales que son entrenados de manera secuencial, cambiando la importancia (peso) de las observaciones, de tal forma que cada nuevo árbol intenta mejorar los errores del anterior (Boosting). En cada uno de estos árboles individuales las observaciones se van repartiendo a través de bifurcaciones denominadas nodos, construyendo así la estructura del árbol hasta alcanzar un nodo terminal. La predicción de cada nueva observación se realiza añadiendo las predicciones de todos los árboles que conforman el modelo.

Entre las principales ventajas del modelo Gradient Boosting Trees encontramos :

- Gran capacidad en la exploración de datos, permitiendo la identificación rápida y eficiente de variables (predictores) más importantes (en este caso cada una de las medidas aplicadas).
- Selección automática de predictores.
- Manejo de predictores numéricos y categóricos sin necesidad de crear variables *dummy* o *one-hot-encoding*.
- En caso de no disponer de ningún predictor en alguna observación, se puede lograr realizar el proceso de predicción haciendo uso de las observaciones que pertenecen al último nodo alcanzado.
- Gran capacidad de exploración de datos. Rápida identificación de variables (predictores) más importantes.
- No es necesario el cumplimiento de distribuciones específicas en los datos.
- Menor necesidad de limpieza y pre-procesado de datos lo que facilita el trabajo con la información.
- Poco impacto de outliers.

Como vemos este tipo de modelo es el más adecuado teniendo en cuenta la tarea a realizar, ya que a partir de los datos obtenidos en el clustering se construyen árboles de decisión usando cada medida como predictor, generando así los predictores más importantes y por tanto las medidas más eficaces en cada zona.

Algoritmo Gradient Boosting:

Gradient Boosting basa su funcionamiento en el algoritmo Adaboost 8.3.2, permitiendo emplear cualquier función de coste diferenciable. El proceso empleado en el entrenamiento es el siguiente:

1. Ajuste de un primer **weak learner** f1 para la predicción de la variable de respuesta y cálculo de los residuos con y-f1(x). Seguidamente, se realiza el ajuste de un nuevo modelo f2 cuyo objetivo es la corrección de los errores hechos por el modelo f1.

$$f_1(x) \approx y \tag{8.1}$$

$$f_2(x) \approx y - f_1(x) \tag{8.2}$$

2. A continuación, en una nueva iteración se calcularán los residuos de los dos modelos f1 y f2 de manera conjunta para que un tercer modelo f3 trate de corregirlos. Este proceso se realizará M veces, de tal forma que cada nuevo modelo f(x) minimice los residuos del anterior.

$$f_3(x) \approx y - f_1(x) - f_2(x) \tag{8.3}$$

3. Esta minimización iterativa de los residuos puede provocar el denominado *overfitting*, por lo que se hará uso de un valor de regularización como el *learning rate* para limitar la influencia de cada modelo en el ensemble.

$$f_1(x) \approx y \tag{8.4}$$

$$f_2(x) \approx y - \lambda f_1(x) \tag{8.5}$$

$$f_3(x) \approx y - \lambda f_1(x) - \lambda f_2(x) \tag{8.6}$$

$$y \approx \lambda f_1(x) + \lambda f_2(x) + f_3(x) + ... + \lambda f_m(x) \tag{8.7}$$

8.4.1. Importancia de los predictores

Uno de los principales problemas de los modelos Gradient Boosting Trees es su interpretación, pues están compuestos por una combinación de árboles de decisión y es muy complicado poder obtener una representación gráfica que nos permita saber de forma visual los predictores más importantes.

Como solución a esta problemática se han creado estrategias para la cuantificación de la importancia de los predictores, permitiendo un análisis de los resultados obtenidos menos abstracto.

Las medidas usadas para la obtención de la influencia son la pureza de nodos y la importancia por permutación.

Pureza de los nodos:

La forma de obtención del incremento de pureza de los nodos se realiza a través de cada uno de los predictores usados, donde se calcula el descenso medio del Root Mean Squared Error (RMSE) conseguido por cada división (provocada por el predictor para el

conjunto de árboles que forman el ensemble). Este es la desviación cuadrática (dispersión) sobre la regresión que hace el modelo.

Estos predictores se corresponden a las medidas aplicadas tanto en confinamientos como a nivel de CCAA. Cuanto mayor sea el valor de dicho descenso o RMSE, mayor será la aportación del predictor al modelo, y por tanto, mayor efectividad tendrá la medida a la que corresponde dicho predictor dentro de la zona de salud estudiada.

Importancia de los predictores en el modelo
####ZONA DE SALUD: C.S. ARANDA NORTE ####

	predictor	importancia
6	LIMIT_REUN_10	0.325178
0	CIERRE_HOSTELERIA	0.212905
19	PROHIB_VISIT_RESI	0.135904
3	REST_ACCESO_CCAA	0.108955
10	AFORO_HOST_75	0.099216
8	RED_PERS_SACTV_50	0.066633
4	REST_ACCESO_ZS	0.021805
12	LIMIT_MESA_10	0.020812
5	TOQUE_NOCTURNO	0.008593
2	CIERRE_C_COMERCIALES	0.000000

Figura 8.5: Ejemplo de obtención pureza por nodos

Importancia por permutación:

Identificación de la influencia de cada predictor a través de una métrica de evaluación del modelo. En nuestro caso usaremos el **Root Mean Squared Error.**

El valor asignado o asociado a cada predictor se obtendrá mediante el proceso descrito a continuación:

1. Creación del conjunto de árboles que componen el modelo. Hiperparámetros como la profundidad, learning rate usado o número de árboles, serán obtenidos a través del uso de la validación cruzada.

2. Cálculo de RMSE (rmse0)
3. Para cada uno de los predictores p permutaremos en todos los árboles que componen el modelo los valores de dicho predictor, manteniendo el resto de valores constante.
4. Recalcular el **RMSE** tras permutación (rmsep)
5. Cálculo del incremento del **RMSE** debido a la permutación del predictor p.

$$\%Incremento_p = \frac{(rmse - j - rmse_0)}{rmse_0} \times 100 \tag{8.8}$$

En el caso de que el predictor permutado contribuya al aumento del error del modelo, este **se considerará como predictor influyente**. La obtención de un incremento del error inferior o igual a 0 será considerada como una influencia nula por parte del predictor en el modelo.

En nuestro proyecto se hace uso de este cálculo para validar los resultados obtenidos con la pureza de los nodos y obtener finalmente las medidas más efectivas. Cada predictor será denominado como una *feature* o característica.

####ZONA DE SALUD: C.S. ARANDA NORTE ####

	importances_mean	importances_std	feature
6	0.517252	0.048665	LIMIT_REUN_10
0	0.103974	0.013706	CIERRE_HOSTELERIA
8	0.014944	0.011689	RED_PERS_SACTV_50
10	0.009900	0.010252	AFORO_HOST_75
3	0.006694	0.004261	REST_ACCESO_CCAA
19	0.002486	0.001509	PROHIB_VISIT_RESI
12	0.001756	0.001357	LIMIT_MESA_10
4	0.000000	0.000000	REST_ACCESO_ZS
5	0.000000	0.000000	TOQUE_NOCTURNO

Figura 8.6: Ejemplo de obtención importancia por permutación.

8.5. Construcción de modelos de aprendizaje

A continuación se muestra todo el proceso de construcción de los modelos de aprendizaje usados, así como la preparación de los datos, tipos de modelos usados, comparación de resultados y análisis de conclusiones obtenidas.

8.5.1. Preparación de los datos

División por zonas de salud:

Partiendo del dataset global donde están incluidas todas las zonas de salud usadas como caso de estudio, se ha realizado una división por zonas creando así datasets individuales de cada una de ellas.

El motivo de esta decisión, es la aplicación de los modelos GBT a los datos de cada zona para obtener aquellas medidas más relevantes en cada una de estas, obteniendo posteriormente un modelo de cada zona.

Predictores:

Para la construcción del ensemble se han usado todas aquellas variables y datos relacionados con las medidas aplicadas en cada una de las zonas de salud estudiadas (tablas 5.4 y 5.5), es por ello por lo que en nuestro dataset hemos eliminado todas las variables relacionadas con la salud. Dichas medidas funcionarán como los predictores de los diferentes árboles de decisión que se usen en la creación de los modelos GBT.

Variable de respuesta:

La variable de respuesta usada como referencia para la construcción del árbol se creará a partir de los resultados obtenidos del clustering aplicado previamente. La variable denominada **MOV_CLUSTER** reflejará todos aquellos tipos de subidas y bajadas según los tipos dados en cada una de las zonas de salud.

Para la creación de esta variable se ha asignado a cada uno de los tipos de subidas y bajadas fijado en el proceso de clustering anterior (7.3.5), una serie de números o calificaciones del más alto (subidas) al más bajo (bajadas):

Subidas:

- **Subida grande: 6**
- **Subida intermedia: 5**

- **Subida baja: 4**

Bajadas:

- **Bajada baja: 3**
- **Bajada intermedia: 2**
- **Bajada grande: 1**

Dicho valor será asignado a cada linea de la gráfica en el tiempo una vez se produzca una subida o una bajada. Para el valor inicial dado a las lineas de comienzo de las gráficas se han observado las subidas y bajadas previas al periodo estudiado (*segunda ola*).

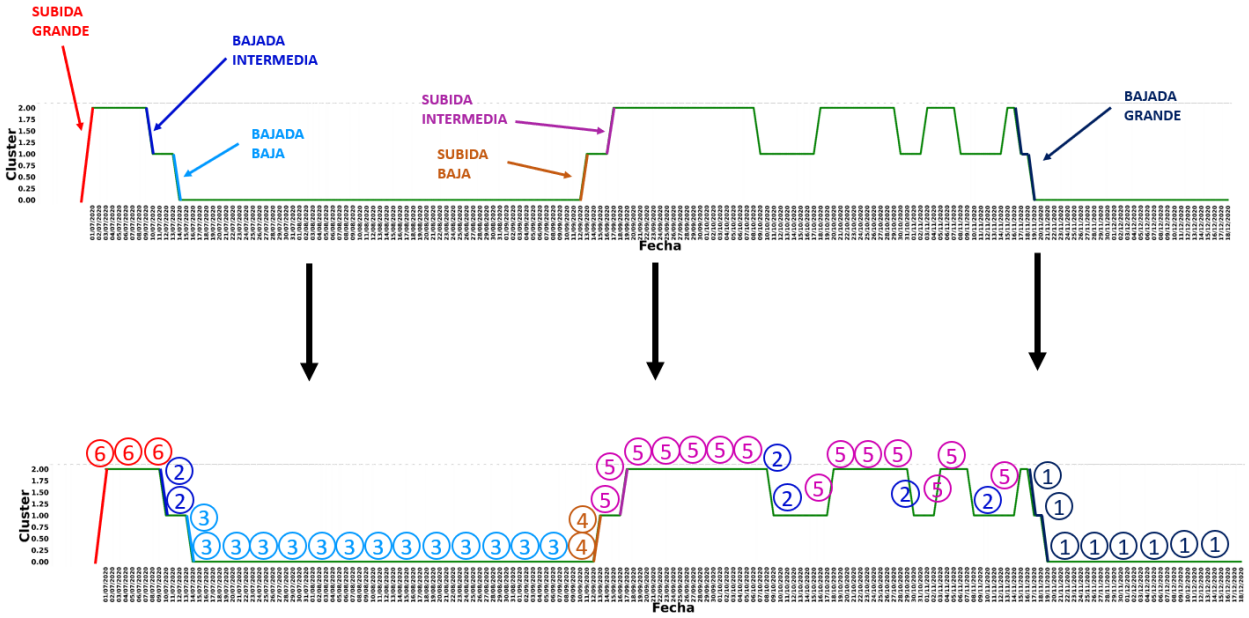


Figura 8.7: Creación variable MOV_CLUSTER

	MOV_CLUSTER	CIERRE_HOSTELERIA	CIERRE_GIMNASIOS	CIERRE_C_COMERCIALES	REST_ACCESO_CCAA	REST_ACCESO_ZS	TOQUE_NOCTURNO	LIMIT_REUN_10	LIMIT
0	3	0	0	0	0	0	0	0	
1	3	0	0	0	0	0	0	0	
2	3	0	0	0	0	0	0	0	
3	3	0	0	0	0	0	0	0	
4	3	0	0	0	0	0	0	0	

Figura 8.8: Ejemplo dataset usado en modelos GBT

8.5.2. Creación y comparación Modelos GBT

En este proyecto se han creado tres tipos de implementaciones GBT para la comparación y obtención de los mejores resultados. Dichas implementaciones poseen un funcionamiento basado en los modelos GBT, diferenciándose en la forma de estimar los mejores hiperparámetros para el entrenamiento y composición de los ensembles. Los tres tipos usados han sido:

- **Gradient Boosting Trees:** Hace uso del estimador GradientBoostingRegressor para la búsqueda de los mejores hiperparámetros de construcción del ensemble.
- **HistGradientBoosting Trees [22]:** Ensemble más rápido que GradientBoostingRegressor para grandes conjuntos de datos. Durante el entrenamiento el árbol de crecimiento aprende en cada punto de división si las muestras con valores perdidos deben ir al hijo izquierdo o al derecho (basándose en la ganancia potencial). Si no se encuentran valores perdidos para una característica determinada durante el entrenamiento, las muestras con valores perdidos se asignan al hijo que tenga más muestras.
- **XGBoost Trees:** Tipo de ensemble más rápido que HistGradientBoosting. A diferencia de las implementaciones anteriores la parada temprana necesita de una división previa de los datos de entrenamiento para la obtención de los mejores hiperparámetros.

Implementación:

Los pasos seguidos para la implementación de los diferentes modelos posee una estructura semejante en los tres tipos empleados:

1.Carga de bibliotecas y modelos a usar:

Se importan todos las bibliotecas y modelos necesarios para la creación de los modelos.

2.Preparación de los datos:

Se eliminan aquellas variables no relevantes para usar en nuestros modelos, quedando únicamente aquellas usadas como predictores (medidas) y respuesta (subidas y bajadas detectadas por clustering).

3.Creación del modelo:

División de nuestro dataset en grupos de test y train para el entrenamiento y posterior cuantificación de la capacidad predictiva de los modelos GBT.

Configuración inicial de hiperparámetros de nuestro modelo. Estos parámetros serán configurados con unos valores iniciales (sin criterio) para la observación de la eficacia del modelo. Posteriormente y mediante validación cruzada, se buscarán los mejores hiperparámetros haciendo uso de la parada temprana (8.3.2. Dichos parámetros serán:

- **n_estimators:** Número de árboles que compondrán nuestro ensemble
- **loss:** Función de pérdida
- **max_feautures:** Número de predictores considerados en cada división. En este caso quedará configurada en auto para el uso de todos los predictores.
- **random_state:** Semilla para hacer que los resultados sean reproducibles.

```
# División conjunto de datos en train y test
# ----------------------------------------------
X_train, X_test, y_train, y_test = train_test_split(
                                        df.drop(columns = "MOV_CLUSTER"),
                                        df['MOV_CLUSTER'],
                                        random_state = 123
                                    )
# Creación modelo GBT
# ----------------------------------------------
modelo = GradientBoostingRegressor(
            n_estimators = 10,
            loss         = 'ls',
            max_features = 'auto',
            random_state = 123
         )

# Train modelo
# ----------------------------------------------
modelo.fit(X_train, y_train)
```

Figura 8.9: Hiperparámetros iniciales de modelos GBT para train

```
El error (rmse) del train es: 0.8851041608505791
```

Figura 8.10: Valor rmse train

El parámetro de medición RMSE estará basado en la desviación cuadrática media (dispersión) sobre la regresión que hace el modelo. Dicha medida dependerá del rango de valores con el que se trabaja, que en nuestro caso será como máximo de 6 (correspondiente a la calificación de cada tipo de subida y bajada realizado). Por tanto, el RMSE se considerará adecuado si no supera el valor de 1 dentro de ese rango a la hora del entrenamiento (train) y evaluación (test).

4.Predicción y evaluación del modelo:

Una vez entrenado el modelo se evaluará con el conjunto de test creado previamente la capacidad predictiva. Al igual que en el entrenamiento (Figura 8.10), se mostrará el RMSE.

5.Búsqueda de mejores hiperparámetros (Grid search):

El entrenamiento y evaluación previas se realizan con hiperparámetros definidos sin criterio alguno. A través de grid search se realizan y analizan diferentes combinaciones para la obtención de los mejores valores óptimos a usar con cada uno de los modelos.

Esta búsqueda de los mejores hiperparámetros se realiza mediante la validación cruzada, donde a través de una serie de hiperparámetros iniciales dados, se inicia un proceso de búsqueda que se detiene una vez se encuentren los mejores valores.Dicha detención se realizará usando la parada temprana (véase en el apartado 8.3.2).

Los hiperparámetros iniciales dados para la búsqueda mediante validación cruzada poseerán distintos valores de cara realizar esa búsqueda de los mejores resultados. En el caso del número de árboles a incluir (**n_estimators**), se establecerá un número muy alto de árboles para la posterior detención mediante la parada temprana en el número óptimo a incluir (haciendo uso del *learning rate*).

```
# Hiperparámetros iniciales evaluados
# ------------------------------------------------
param_grid = {'max_features'  : ['auto', 'sqrt', 'log2'],
              'max_depth'     : [None, 1, 3, 5, 10, 20],
              'subsample'     : [0.5, 1],
              'learning_rate' : [0.001, 0.01, 0.1]
             }
```

Figura 8.11: Ejemplo hiperparámetros iniciales grid search

En cada uno de los tipos de modelo GBT usados se especificará el tipo de estimador correspondiente, en el que están incluidos estos hiperparámetros iniciales de evaluación junto a la configuración de la parada temprana.

En las implementaciones GBT e HistGradientBoosting el conjunto test usado para la parada temprana se extraerá automáticamente de los datos train usados para cada ajuste, integrándose automáticamente en la función grid (**GridSearchCV()**) por validación cruzada. En el caso de la implementación XGBT, este conjunto test usado en la parada temprana deberá ser separado manualmente del conjunto de datos train.

```
# Grid search con validación cruzada
# ------------------------------------------------
grid = GridSearchCV(
       #Uso del estimador GBT para la búsqueda de mejores parámetros
         estimator  = GradientBoostingRegressor(
                        #Establecimiento de número alto de arboles de decisión
                         n_estimators        = 1000,
                         random_state        = 123,
                         # Activación parada temprana
                         validation_fraction = 0.1,
                         n_iter_no_change    = 5,
                         tol                 = 0.0001
                      ),
        param_grid = param_grid,
        scoring    = 'neg_root_mean_squared_error',
        n_jobs     = multiprocessing.cpu_count() - 1,
        cv         = RepeatedKFold(n_splits=3, n_repeats=1, random_state=123),
        #Reentrenamiento con mejores valores obtenidos
        refit      = True,
        verbose    = 1,
        return_train_score = True
       )

grid.fit(X = X_train, y = y_train)
```

Figura 8.12: Busqueda por grid search validación cruzada GBT

Como podemos ver en la Figura 8.12, se establece las variables necesarias para la activación de la parada temprana, usada para la búsqueda de los mejores hiperparámetros. Dichas variables son:

- **validation_fraction:** Valor establecido para la división del conjunto de datos en train y test usado durante el proceso de parada temprana.
- **n_inter_no_change:** Número de etapas finales para la activación del proceso de parada temprana.
- **tol:** Valor de mejora usado para la detención del proceso de parada temprana.

Por tanto el proceso de parada temprana consiste en la especificación de una **validation_fraction** que establece la división de los conjuntos de datos usados para el entrenamiento (train) y para la validación (test). El modelo GBT se entrena en cada etapa utilizando el conjunto de entrenamiento y se evalúa utilizando el conjunto de test. Se continúa así hasta que las puntuaciones del modelo en las últimas **n_iter_no_change** etapas mejoren y alcancen el valor de la variable **tol**. Después de esto, se considera que el modelo ha convergido y se detiene(early stop) la adición de nuevas etapas.

En la variable grid se almacenan los mejores hiperparámetros encontrados por la validación cruzada, mostrando los valores usados en cada una de las iteraciones de búsqueda de los mejores hiperparámetros (Figura 8.13). Estos valores son:

- **param_learning_rate:** Valor dado al learning rate en cada iteración.
- **param_max_depth:** Profundidad del árbol usado en cada iteración.
- **param_max_feautures:** Operación usada para la elección del número de predictores en cada división.
- **param_subsample:** Relación entre 0 y 1 elegida para el entrenamiento en cada iteración.
- **mean_test_score:** Puntuación media obtenida para datos test en cada iteración.
- **std_test_score:** Puntuación estándar obtenida con datos test en cada iteración.
- **mean_train_score:** Puntuación media obtenida para datos train (entrenamiento) en cada iteración.
- **std_train_score:** Puntuación estándar obtenida con datos train (entrenamiento) en cada iteración.

	param_learning_rate	param_max_depth	param_max_features	param_subsample	mean_test_score	std_test_score	mean_train_score	std_train_score
83	0.1	1	log2	1	-0.943309	0.132827	-0.873648	0.060924
81	0.1	1	sqrt	1	-0.943309	0.132827	-0.873648	0.060924
86	0.1	3	sqrt	0.5	-0.944546	0.133177	-0.869125	0.048615
88	0.1	3	log2	0.5	-0.944546	0.133177	-0.869125	0.048615
82	0.1	1	log2	0.5	-0.951949	0.141048	-0.882145	0.063896
...	...	...	...	...	...	...	...	...
10	0.001	1	log2	0.5	-1.021828	0.088139	-0.985018	0.052882
8	0.001	1	sqrt	0.5	-1.021828	0.088139	-0.985018	0.052882
12	0.001	3	auto	0.5	-1.022324	0.081723	-0.954447	0.100369
26	0.001	10	sqrt	0.5	-1.025449	0.090665	-0.948228	0.094813
34	0.001	20	log2	0.5	-1.025449	0.090665	-0.948228	0.094813

100 rows × 8 columns

Figura 8.13: Resultados iteraciones de búsqueda mejores hiperparámetros.

Una vez realizado el proceso de validación cruzada se muestran los valores óptimos encontrados para nuestro modelo, realizando un re-entrenamiento con esos valores. En nuestro caso, dicho re-entrenamiento se hará automáticamente almacenando el modelo resultante en la variable **.best_estimator_** , que pasará a denominarse **modelo_final** para posteriormente mostrar el RMSE.

6.Importancia de los predictores:

Una vez obtenido nuestro modelo final se mostrarán aquellos resultados relacionados con la importancia de los predictores a través de la pureza de los nodos (Apartado 8.4.1) y la importancia por permutación (Apartado 8.4.1). Esto nos ayudará a determinar qué predictores han sido más relevantes en el entrenamiento y predicción de nuestro modelo, y por tanto, las medidas más efectivas en cada una de las zonas de salud.

Importancia por pureza de nodos:

Se muestra mediante una tabla la importancia de cada uno de los predictores (medidas) en el modelo. Gracias a esta tabla veremos claramente aquellas medidas más efectivas.

Importancia de los predictores en el modelo
####ZONA DE SALUD: C.S. MIRANDA OESTE ####

	predictor	importancia
3	REST_ACCESO_CCAA	0.492049
8	RED_PERS_SACTV_50	0.238930
12	LIMIT_MESA_10	0.206396
0	CIERRE_HOSTELERIA	0.049969
16	LIMIT_OC_NOCT_23PM	0.012415
6	LIMIT_REUN_10	0.000227
4	REST_ACCESO_ZS	0.000014
18	PROHIB_FUMAR	0.000000
17	LIMIT_OC_NOCT_1AM	0.000000
15	CIERRE_DISCO	0.000000
14	CIERRE_PENAS	0.000000
13	LIMIT_MESA_6	0.000000
10	AFORO_HOST_75	0.000000
11	AFORO_HOST_50	0.000000
1	CIERRE_GIMNASIOS	0.000000
9	PROHIB_CONSUM_BARRA	0.000000
7	LIMIT_REUN_6	0.000000
5	TOQUE_NOCTURNO	0.000000
2	CIERRE_C_COMERCIALES	0.000000
19	PROHIB_VISIT_RESI	0.000000

Figura 8.14: Ejemplo resultados importancia de predictores

Como podemos ver en la Figura 8.14 , la medida de restricción de acceso a la CA ha sido una de las más efectivas.

En el caso de la implementación HistGradientBoosting, estos valores no podrán ser visualizados debido a la ausencia de la variable para cada predictor **feauture_importances_** .

Importancia por permutación:

A través de la importancia media (**importances_mean**) de cada uno de los predictores en las permutaciones realizadas del modelo, validaremos los resultados de importancia obtenidos anteriormente (Figura 8.14).

####ZONA DE SALUD: C.S. MIRANDA OESTE ####

	importances_mean	importances_std	feature
3	0.857644	0.063211	REST_ACCESO_CCAA
12	0.343093	0.028289	LIMIT_MESA_10
0	0.123720	0.024555	CIERRE_HOSTELERIA
8	0.112309	0.012084	RED_PERS_SACTV_50
16	0.060635	0.019160	LIMIT_OC_NOCT_23PM
6	0.052298	0.007857	LIMIT_REUN_10
4	0.005560	0.001938	REST_ACCESO_ZS
18	0.000000	0.000000	PROHIB_FUMAR
17	0.000000	0.000000	LIMIT_OC_NOCT_1AM
15	0.000000	0.000000	CIERRE_DISCO
14	0.000000	0.000000	CIERRE_PENAS
13	0.000000	0.000000	LIMIT_MESA_6
10	0.000000	0.000000	AFORO_HOST_75
11	0.000000	0.000000	AFORO_HOST_50
1	0.000000	0.000000	CIERRE_GIMNASIOS
9	0.000000	0.000000	PROHIB_CONSUM_BARRA
7	0.000000	0.000000	LIMIT_REUN_6
5	0.000000	0.000000	TOQUE_NOCTURNO
2	0.000000	0.000000	CIERRE_C_COMERCIALES
19	0.000000	0.000000	PROHIB_VISIT_RESI

Figura 8.15: Ejemplo resultados importancia por pureza de nodos.

Se puede observar cómo se obtiene como medida más influyente en las iteraciones aquella ya detectada con una gran importancia (restricción de acceso a CA).

Exportar modelos finales:

Una vez obtenidos los modelos finales mediante el uso de los mejores hiperparámetros, estos serán exportados haciendo uso de la biblioteca pickle. Estos modelos finales podrán ser aplicados en otras zonas de salud diferentes a las usadas en el proyecto.

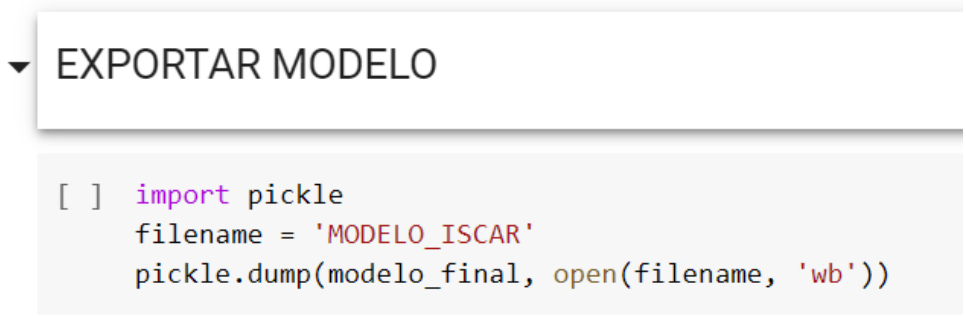


Figura 8.16: Ejemplo de exportación de los modelos finales obtenidos.

Comparación modelos obtenidos:

Se ha realizado una comparación de los resultados obtenidos al aplicar los tres tipos de modelos construidos a las zonas de salud estudiadas. Dicha comparación nos ha permitido determinar el mejor modelo a usar en nuestro proyecto basándonos en factores como la velocidad de búsqueda y compilación de los mejores parámetros y la obtención del menor **RMSE** en cada una de las zonas estudiadas.

A continuación se muestra una tabla donde se han recogido los resultados obtenidos con cada implementación para cada una de las zonas:

TIEMPO DE COMPILACION												TOTAL-MEDIA
	ARANDA SUR	ARANDA RURAL	ARANDA NORTE	MIRANDA OESTE	MIRANDA ESTE	MIRANDA DEL CASTAÑAR	MOTA DEL MARQUÉS	PEÑAFIEL	MEDINA CAMPO URBANO	MEDINA CAMPO RURAL	ISCAR	
GBT	47 sec	30.5 sec	43.9 sec	56.6 sec	50 sec	29.5 sec	39.5 sec	36 sec	52.4 sec	27.6 sec	19.7 sec	39.33 sec
HGBT	78 sec	46.7 sec	78 sec	96 sec	72 sec	78 sec	27.5 sec	55.5 sec	96 sec	52.9 sec	54.6 sec	66.8 sec
XGBT	30 sec	30.2 sec	32.5 sec	38.3 sec	43.3 sec	42.7 sec	41.9 sec	32.5 sec	35.2 sec	35.2 sec	32.5 sec	35.21 sec
rmse MODELO FINAL												TOTAL-MEDIA
	ARANDA SUR	ARANDA RURAL	ARANDA NORTE	MIRANDA OESTE	MIRANDA ESTE	MIRANDA DEL CASTAÑAR	MOTA DEL MARQUÉS	PEÑAFIEL	MEDINA CAMPO URBANO	MEDINA CAMPO RURAL	ISCAR	
GBT	1.133	1.29	1.05	0.17	0.62	1.18	0.63	0.42	0.51	1.046	1.02	0.989
HGBT	1.15	1.31	1.059	0.30	0.62	1.2	0.63	0.40	0.43	1.006	1.02	0.847
XGBT	1.138	1.32	1.07	0.16	0.61	1.2	0.63	0.42	0.46	1.29	1.02	0.829

Figura 8.17: Comparación implementaciones GBT.

Como podemos ver en la Figura 8.17, la implementación XGBT ofrece un menor tiempo medio global de compilación frente a las otras implementaciones, además de un **RMSE** global menor que el resto de modelos (estando para la mayoría de zonas por debajo del valor 1 estipulado como máximo permitido). Es por ello por lo que se ha optado por el uso de este modelo para la obtención de aquellas medidas más efectivas en cada una de las zonas estudiadas. La implementación HistGradientBoosting no ha sido seleccionada debido a su problemática con la muestra de la importancia por pureza de nodos.

8.5.3. Análisis y obtención de resultados

Se ha aplicado la implementación XGBT a cada una de las zonas para obtener las medidas más efectivas, usándose también en el dataset global donde se encuentran todas las zonas estudiadas, con el objetivo de obtener aquellas medidas eficaces en todas las zonas.

A continuación se muestran los resultados obtenidos al analizar tanto la importancia de los predictores como la importancia por iteración. Dichos resultados serán mostrados a través de una tabla con la estructura y leyenda indicados en la Figura8.18. Las medidas mostradas en la tabla tendrán asignado el nombre de variable dado en cada dataset, siguiendo la nomenclatura ya vista en el Apartado 5.3.3.

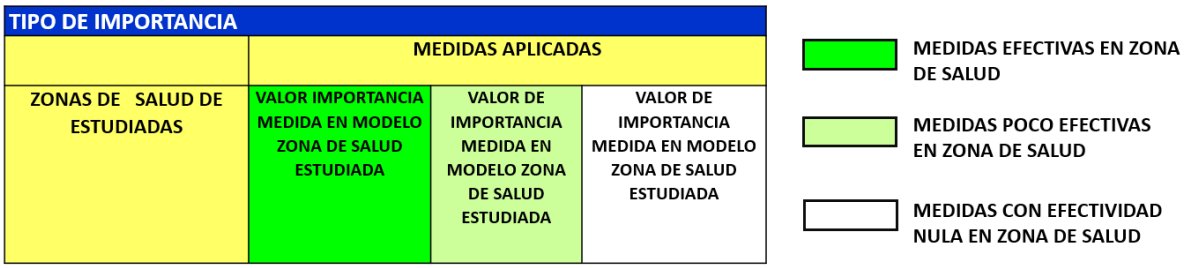
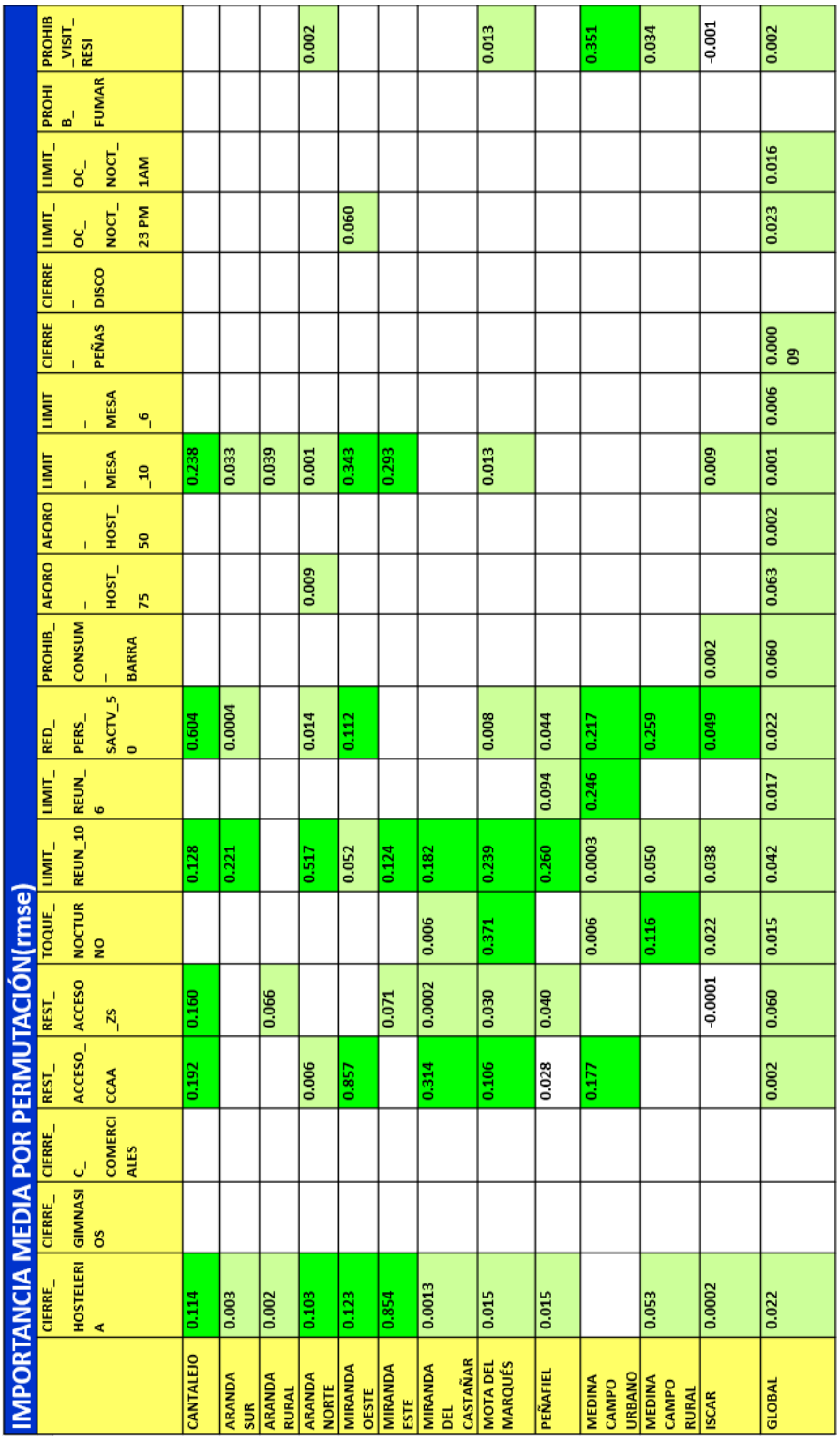


Figura 8.18: Estructura tabla resultados XGBT en zonas de salud.

En cada una de las filas correspondientes a cada zona de salud serán consideradas como **medidas efectivas** aquellos predictores que tengan una influencia (incremento RMSE en modelo XGBT) **superior a 0.1**. Los predictores por debajo de este valor y que no sean iguales o menores que 0, serán considerados como **medidas poco efectivas**.En el caso de obtener una importancia 0 o negativa, será considerada como **medida de nula eficacia** en la zona de salud.

IMPORTANCIA PREDICTORES(rmse)	CIERRE_ HOSTELERIA	CIERRE_ GIMNASIOS	CIERRE_ C_ COMERCIALES	REST_ ACCESO _ CCAA	REST_ ACCESO _ZS	TOQUE_ NOCTURNO	LIMIT_ REUN_10	LIMIT_ REUN_6	RED_ PERS_ SACTV_50	PROHIB_ CONSUM_ BARRA	AFORO _ HOST_ 75	AFORO _ HOST_ 50	LIMIT _ MESA _10	LIMIT _ MESA _6	CIERRE _ PEÑAS	CIERRE _ DISCO	LIMIT_ OC_ NOCT_ 23 PM	LIMIT_ OC_ NOCT_ 1AM	PROHIB_ FUMAR	PROHIB_VISIT_RESI
CANTALEJO	0.220			0.029	0.251		0.266		0.147				0.084							
ARANDA SUR	0.138						0.435		0.047					0.196						0.181
ARANDA RURAL	0.217				0.387								0.394							
ARANDA NORTE	0.212			0.108	0.021	0.008	0.325		0.066		0.099		0.020							0.135
MIRANDA OESTE	0.049			0.492			0.0002		0.238				0.206				0.012			
MIRANDA ESTE	0.424			0.051	0.077	0.002	0.117		0.046								0.018			0.030
MIRANDA DEL CASTAÑAR	0.198			0.213	0.144	0.076	0.133	0.088		0.037										0.106
MOTA DEL MARQUÉS	0.106			0.190	0.126	0.199	0.178	0.012	0.018				0.035							0.132
PEÑAFIEL	0.016			0.010	0.022		0.467	0.412	0.070											
MEDINA CAMPO URBANO	0.001			0.064	0.005	0.034	0.016	0.155	0.072		0.016									0.633
MEDINA CAMPO RURAL	0.078			0.064		0.084	0.099		0.669											0.002
ISCAR	0.249				0.048	0.171	0.107	0.035	0.100	0.155			0.055							0.076
GLOBAL	0.074			0.058	0.281	0.056	0.048	0.060	0.047	0.044	0.097	0.023		0.010	0.022	0.001 7	0.040	0.078		0.051

Figura 8.19: Resultados medidas más efectivas en cada zona de salud según importancia de predictor.



IMPORTANCIA MEDIA POR PERMUTACIÓN(rmse)																						
	CIERRE_ HOSTELERIA	CIERRE_ GIMNASIOS	CIERRE_ C_ COMERCIALES	REST_ ACCESO_ CCAA	REST_ ACCESO _ZS	TOQUE_ NOCTURNO	LIMIT_ REUN_10	LIMIT_ REUN_6	RED_ PERS_ SACTV_50	PROHIB_ CONSUM _ BARRA	AFORO _ HOST_ 75	AFORO _ HOST_ 50	LIMIT _ MESA _10	LIMIT _ MESA _6	CIERRE _ PEÑAS	CIERRE _ DISCO	LIMIT_ OC_ NOCT_ 23 PM	LIMIT_ OC_ NOCT_ 1AM	PROHIB_ FUMAR	PROHIB_VISIT_RESI
CANTALEJO	0.114			0.192	0.160		0.128		0.604				0.238							
ARANDA SUR	0.003						0.221		0.0004				0.033							
ARANDA RURAL	0.002				0.066								0.039							
ARANDA NORTE	0.103			0.006			0.517		0.014		0.009		0.001							0.002
MIRANDA OESTE	0.123			0.857			0.052		0.112				0.343				0.060			
MIRANDA ESTE	0.854				0.071		0.124						0.293							
MIRANDA DEL CASTAÑAR	0.0013			0.314	0.0002	0.006	0.182													
MOTA DEL MARQUÉS	0.015			0.106	0.030	0.371	0.239		0.008				0.013							0.013
PEÑAFIEL	0.015			0.028	0.040		0.260	0.094	0.044											
MEDINA CAMPO URBANO				0.177		0.006	0.0003	0.246	0.217											0.351
MEDINA CAMPO RURAL	0.053					0.116	0.050		0.259											0.034
ISCAR	0.0002				-0.0001	0.022	0.038		0.049	0.002			0.009							-0.001
GLOBAL	0.022			0.002	0.060	0.015	0.042	0.017	0.022	0.060	0.063	0.002	0.001	0.006	0.000 09		0.023	0.016		0.002

Figura 8.20: Resultados medidas más efectivas en cada zona de salud según importancia por iteración.

Como podemos ver en las tablas generadas, en la mayoría de zonas la importancia de los predictores (Figura 8.19) coincide con la obtención de un valor distinto de 0 en la importancia en cada iteración (Figura 8.20). Esto nos permite validar que aquellos predictores con gran importancia son usados en el proceso de aprendizaje del modelo, y por tanto pueden ser considerados como medidas eficaces en la zona a la que corresponde el modelo.

Resultados finales:

Una vez realizada la obtención de las medidas más efectivas en cada zona de salud (Figura 8.19), se analizarán los resultados y conclusiones obtenidos, exponiendo por cada zona aquellas medidas más eficaces y las distintas razones que confirmen dichos resultados. Estos datos se validarán con la información obtenida y explicada en los capítulos de análisis de datos (Capítulo 6) y aplicación del método de aprendizaje clustering (Capítulo 7).

En las tablas de cada zona de salud se especificarán las medidas de más eficaces a menos en base a la importancia de predictor, incluyendo solo aquellas detectadas como **efectivas** según el patrón y estructura vistos en la Figura 8.18. También se especificarán las medidas determinadas como eficaces con la aplicación del modelo XGBT al dataset global.

Cantalejo:

MEDIDA	**IMPORTANCIA PREDICTOR(rmse)**
Limite de reuniones 10 personas	0.266
Restricción de acceso a CCAA	0.251
Cierre hostelería	0.220
Reducción de personas en sectores de actividad al 50 %	0.147

Tabla 8.1: Medidas más efectivas Cantalejo

Una de las medidas más eficaces en Cantalejo ha sido aquella relacionada con el limite de reuniones a 10 personas y la reducción de personas en los sectores de actividad al 50 %, aplicadas durante el confinamiento y el día 17 de octubre. Como se puede observar en los análisis de resultados de clustering de esta zona (Figura 7.35), las medidas aplicadas este día resultaron eficaces para la bajada de los contagios.

Otras medidas muy eficaces en Cantalejo que causaron un gran descenso en el número de entradas a la zona fueron la restricción de acceso a la Comunidad Autónoma y el cierre de la hostelería.

Aranda Sur:

MEDIDA	IMPORTANCIA PREDICTOR(rmse)
Limite de reuniones 10 personas	0.435
Limite de 10 personas en mesa hostelería	0.196
Prohibida visita a residencias	0.181
Cierre de hostelería	0.138

Tabla 8.2: Medidas más efectivas Aranda Sur.

En Aranda Sur las medidas más eficaces fueron las establecidas durante agosto, específicamente en los días 7 (cuando se aplicó el confinamiento y la medida de limite de reuniones a 10 personas) y 17 (cuando se aplicó la medida de limite de 10 personas en mesa de hostelería). La prohibición de visitas a residencias también resultó eficaz debido al gran número de residencias existentes en la zona por su alta población.

En cuanto a las medidas que afectan al número de entradas a la zona, encontramos el cierre de hostelería, que, como hemos podido observar, afecta en gran sobre todo al número de entradas en la mayoría de zonas estudiadas.

Se puede comprobar la veracidad de estos resultados con los ya obtenidos en la aplicación del clustering sobre esta zona (Figura 7.36)

Aranda Rural:

MEDIDA	IMPORTANCIA PREDICTOR(rmse)
Limite de 10 personas en mesa hostelería	0.394
Restricción de acceso a CCAA	0.387
Cierre de hostelería	0.217

Tabla 8.3: Medidas más efectivas Aranda Rural.

En la zona de salud de Aranda Rural las medidas más eficaces fueron las ya detectadas en el análisis preliminar de la aplicación del método de clustering (Figura 7.37). Cabe destacar que en esta zona predominan las limitaciones en la hostelería y aquellas con un gran impacto sobre la movilidad. Esto puede deberse al gran número de entradas detectado en la zona de Aranda de Duero (donde la actividad en el sector rural puede ser mucho mayor que en otros lugares) y el alto flujo de entradas en municipios con una baja población causa un gran impacto.

Aranda Norte:

MEDIDA	IMPORTANCIA PREDICTOR(rmse)
Limite de reuniones 10 personas	0.325
Cierre de hostelería	0.212
Prohibida visita a residencias	0.135
Restricción de acceso a CCAA	0.108

Tabla 8.4: Medidas más efectivas Aranda Norte.

En la zona de Aranda de Norte las medidas aplicadas durante el confinamiento como el limite de reuniones a 10 personas o la prohibición de visita a residencias, han sido detectadas como las más eficaces. Dicha detección puede interpretarse adecuadamente con los resultados ya obtenidos y reflejados en la Figura 7.38, donde el análisis realizado a las subidas y bajadas dadas por el método clustering nos muestran que el confinamiento en esta zona había sido muy efectivo.

Debido a la alta movilidad detectada en Aranda de Duero por ser zona colindante con una de las vías de comunicación más grandes del país, vemos como aquellas medidas causantes de una bajada del número de entradas a dicha zona son las más eficaces.

Miranda Oeste:

MEDIDA	IMPORTANCIA PREDICTOR(rmse)
Restricción de acceso a CCAA	0.492
Reducción de personas en sectores de actividad al 50 %	0.238
Limite de 10 personas en mesa hostelería	0.206

Tabla 8.5: Medidas más efectivas Miranda Oeste

Miranda Este:

MEDIDA	IMPORTANCIA PREDICTOR(rmse)
Cierre de hostelería	0.424
Limite de 10 personas en mesa hostelería	0.229
Limite de reuniones 10 personas	0.117

Tabla 8.6: Medidas más efectivas Miranda Este

Como se puede observar en el análisis de los datos de movilidad, Miranda de Ebro es una de las zonas con mayor número de entradas entre todas las estudiadas (Apartado 6.2.1), por tanto, no es extraño que nuestro modelo obtenga como medidas más efectivas las causantes de un gran impacto y bajada en estas, dicha bajada puede validarse a través de los resultados de clustering (Figuras 7.39 y 7.40). La movilidad en las zonas de dicho territorio es la gran causante de los contagios, por tanto, medidas como la restricción de acceso a la CA o el cierre de la hostelería demuestran una gran efectividad.

La limitación de reuniones a 10 personas, tanto en la calle como en las mesas de hostelería, también ha sido detectada como medida eficaz (medida aplicada en confinamientos).

Miranda del Castañar:

MEDIDA	**IMPORTANCIA PREDICTOR(rmse)**
Restricción de acceso a CCAA	0.213
Cierre de hostelería	0.198
Restricción de acceso a zona de salud	0.144
Prohibida visita a residencias	0.106

Tabla 8.7: Medidas más efectivas Miranda del Castañar.

En la zona de Miranda del Castañar (zona rural), se han obtenido como medidas más efectivas aquellas que de una forma u otra han logrado reducir el número de entradas de manera drástica. En este caso podemos observar como toda restricción de entrada a la zona (tanto provenientes de otros territorios como de otras Comunidades Autónomas) es eficaz en la bajada de contagios, destacando la restricción a la zona de salud y la prohibición de visitas a residencias, medidas características de los confinamientos.

Mota del Marqués:

MEDIDA	**IMPORTANCIA PREDICTOR(rmse)**
Toque de queda nocturno 22:00 PM	0.199
Restricción de acceso a CCAA	0.190
Limite de reuniones a 10 personas	0.178
Prohibida visitas a residencia	0.132
Restricción de acceso a zona de salud	0.126
Cierre de hostelería	0.106

Tabla 8.8: Medidas más efectivas Mota del Marqués

En la zona de Mota del Marqués se han obtenido como medidas más eficaces las relacionadas tanto con la restricción en la movilidad dentro de la zona (toque de queda nocturno), como la restricción de entradas a la CA. Al igual que las zonas de Miranda del Castañar o Aranda Rural, Mota del Marqués es considerada como una zona rural de baja población, donde un gran número de entradas tiene un gran impacto en la población y el aumento de contagios.

Medidas aplicadas durante los confinamientos como el limite de reuniones a 10 personas, la prohibición de visitas a residencias o la restricción de acceso a la zona de salud, también han resultado eficaces pese a la menor duración del confinamiento aplicado en esta zona (11 días).

Peñafiel:

MEDIDA	**IMPORTANCIA PREDICTOR(rmse)**
Limite de reuniones a 10 personas	0.467
Limite de reuniones a 6 personas	0.412

Tabla 8.9: Medidas más efectivas Peñafiel

En la zona de salud de Peñafiel se han obtenido como medidas más efectivas las causantes de una menor interacción social en la población como es el caso del limite de 10 o 6 personas en reuniones. Observando la Figura 7.43, vemos como una una vez aplicadas estas medidas (días 22/09 y 17/10 respectivamente), la zona de salud de Peñafiel no registró ninguna subida, por lo que confirmamos que mediante su aplicación y cumplimiento se mantuvo la curva de contagios a raya.

Uno de los motivos principales por lo que en esta zona no se han obtenido como efectivas las medidas con un gran impacto en el número de entradas, ha sido la movilidad interna existente entre municipios afectados (estudiadas en el apartado 6.2.2). Tal y como hemos podido analizar, este tipo de movilidad entre los municipios de Peñafiel y Pesquera de Duero ha sido la gran causante del aumento de contagios y subidas registradas, por lo que aquellas entradas provenientes de otras CCAA o zonas no ha afectado a Peñafiel. Medidas como la restricción de acceso a la zona de salud tampoco han sido efectivas debido a la ubicación de ambos municipios en la misma zona de salud.

Medina del Campo Urbano:

MEDIDA	IMPORTANCIA PREDICTOR(rmse)
Prohibida visita a residencias	0.633
Limite de reuniones a 6 personas	0.155

Tabla 8.10: Medidas más efectivas Medina del Campo Urbano

Medina del Campo Rural:

MEDIDA	IMPORTANCIA PREDICTOR(rmse)
Reducción de personas en sectores de actividad al 50 %	0.669

Tabla 8.11: Medidas más efectivas Medina del Campo Rural

Dentro de las zonas de salud de Medina del Campo (alto número de entradas), observamos como aquellas medidas relacionadas con la bajada de la movilidad no han tenido relevancia alguna a la hora de ser eficaces. Podemos apreciar como únicamente las medidas con un gran impacto en la interacción social o la relación con sectores de riesgo de la población, han sido consideradas como eficaces dentro de estas zonas.

Íscar:

MEDIDA	IMPORTANCIA PREDICTOR(rmse)
Cierre de Hostelería	0.249
Toque de queda nocturno	0.171
Prohibido el consumo en barra	0.155
Limite de reuniones a 10 personas	0.107

Tabla 8.12: Medidas más efectivas Íscar

En la zona de salud de Íscar se han detectado como medidas más eficaces el cierre de la hostelería y toque de queda nocturno, medidas que afectan tanto a las entradas realizadas a la zona como a la movilidad interna dentro de esta. Como se puede observar, muchas de las medidas detectadas están relacionadas con la hostelería, por lo que podemos deducir que las restricciones aplicadas sobre este sector resultan eficaces en esta zona.

Otra medida obtenida como efectiva ha sido el limite de reuniones a 10 personas (medida más eficaz en los dos confinamientos realizados en Íscar).

Global:

MEDIDA	**IMPORTANCIA PREDICTOR(rmse)**
Restricción de acceso a zona de salud	0.281

Tabla 8.13: Medidas más efectivas Global

Aplicando el modelo XGBT al dataset global (donde están incluidas todas las zonas de salud estudiadas), obtenemos como medida más efectiva la restricción de acceso a la zona de salud, medida aplicada en todos los confinamientos y cuyo efecto en la movilidad ha provocado una bajada drástica del número de entradas en todas las zonas estudiadas.

Medidas ineficaces:

Una vez obtenidas las medidas más eficaces en cada una de las zonas estudiadas, podemos observar como algunas de las restricciones estudiadas en el proyecto han tenido una efectividad muy baja o nula en la mayoría de las zonas. A continuación se indican que medidas han sido obtenidas como **no eficaces** por nuestro modelo.

MEDIDA	**SUMA IMPORTANCIA TOTAL**
Prohibición de fumar sin distancia mínima	0
Cierre de gimnasios e instalaciones deportivas	0
Cierre de centros comerciales y grandes establecimientos	0
Cierre de discotecas y locales de ocio nocturno	0.0017
Cierre de peñas y organizaciones con motivos festivos	0.022
Limitación al 50 % del aforo permitido en el interior delos recintos de hostelería	0.023
Limite de apertura de locales de ocio nocturno a las 23PM	0.07
Limite de apertura de locales de ocio nocturno a las 1AM	0.078
Limitación al 75 % del aforo permitido en el interior delos recintos de hostelería	0.212

Tabla 8.14: Medidas no efectivas en zona de salud estudiadas

Finalmente podremos afirmar que aquellas medidas con poca o nula eficacia en las zonas estudiadas podrían ser descartadas de cara a una menor restricción de la población e impacto económico.

Capítulo 9

Zonas de test

Una vez alcanzados los objetivos principales del proyecto, en cuanto a la obtención de las medidas más eficaces en cada zona de salud, se ha propuesto la aplicación de todos los conocimientos y aprendizajes adquiridos en una o más zonas totalmente diferentes a las ya usadas. El objetivo de esta fase será la comprobación de la efectividad y utilidad de la investigación realizada.

Las zonas elegidas para esta parte de evaluación han sido Segovia I, Segovia II y Segovia III, territorios pertenecientes a la capital de la provincia de Segovia. La elección de esta provincia se debe a la localización de la organización educativa a la que pertenecen todos los involucrados en el proyecto, la Escuela de Ingeniería Informática de Segovia.

Los pasos que se han seguido para el desarrollo de esta fase de test han sido idénticos a los ya vistos en las zonas de salud usadas como estudio, las cuales denominaremos en este capitulo como **zonas de entrenamiento**. El periodo de evaluación se centrará también en la *segunda ola*.

9.1. Datos zonas test

Al igual que en las zonas de entrenamiento, para cada una de las zonas de salud usadas de test se han obtenido los datos relacionados con la salud, las medidas aplicadas y la movilidad haciendo uso de los mismos portales y herramientas de obtención.

9.1.1. Datos de salud test

Los datos de salud referentes a las zonas de Segovia Capital han sido obtenidos del Portal de Datos Abiertos de la Junta de Castilla y León, siendo estos datos los mismos ya obtenidos para las zonas de entrenamiento (apartado 5.3). Las variables y datasets generados a partir de los datos de salud han seguido la misma nomenclatura y obtención ya vista en el apartado 5.3.3.

Toda la información de las zonas de test estudiadas ha sido integrada con los datos referentes a las zonas de salud usadas como entrenamiento, teniendo así un dataset global de todas las zonas de salud estudiadas en el proyecto.

Análisis datos de salud test:

A continuación se muestran las gráficas pertenecientes a la incidencia acumulada y el porcentaje de PCR en el periodo de 14 días (periodo con tendencia de contagios más clara). Cada una de las gráficas muestra los datos de las zonas de test comparándolos con aquellas zonas usadas para el entrenamiento, siguiendo la estructura ya descrita en el apartado 6.1.3.

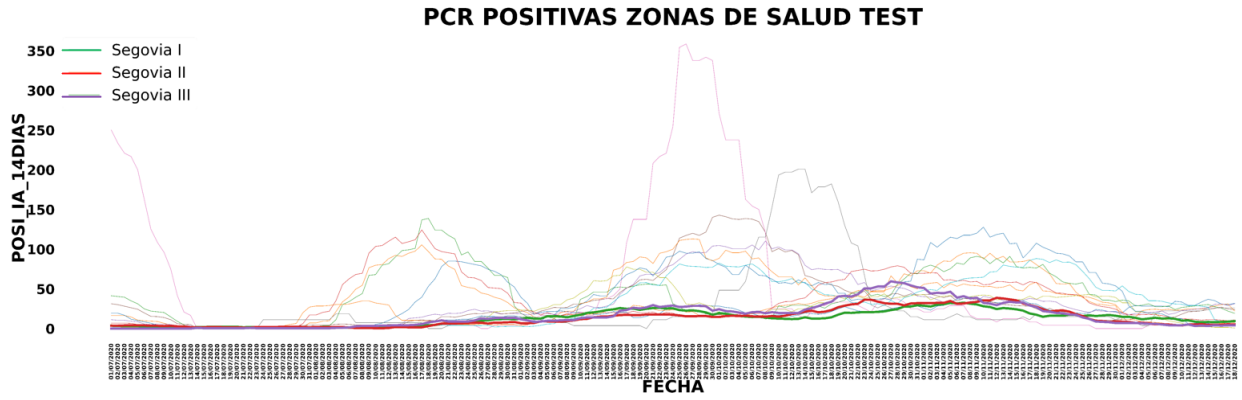


Figura 9.1: Gráfica de PCR positivas en zonas de salud test en el periodo de 14 días.

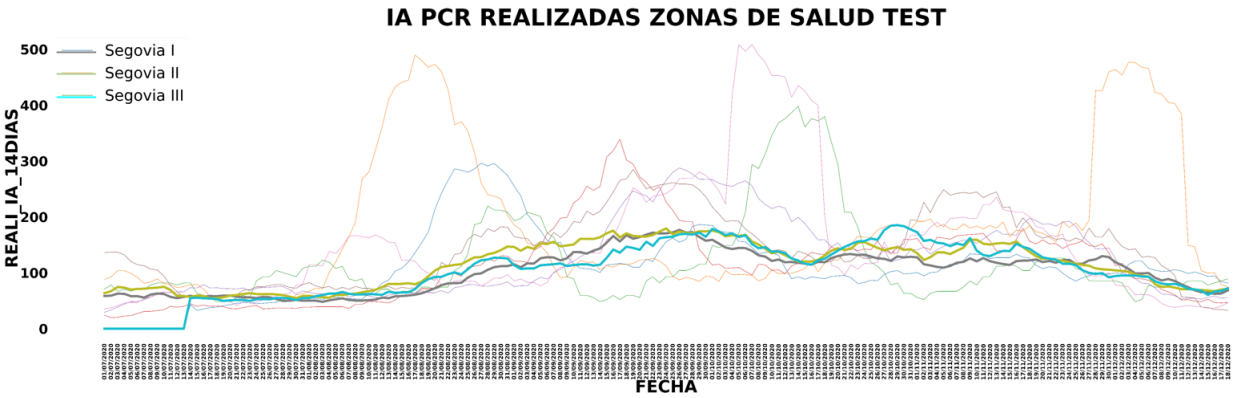


Figura 9.2: Gráfica de PCR realizadas en zonas de salud test en el periodo de 14 días.

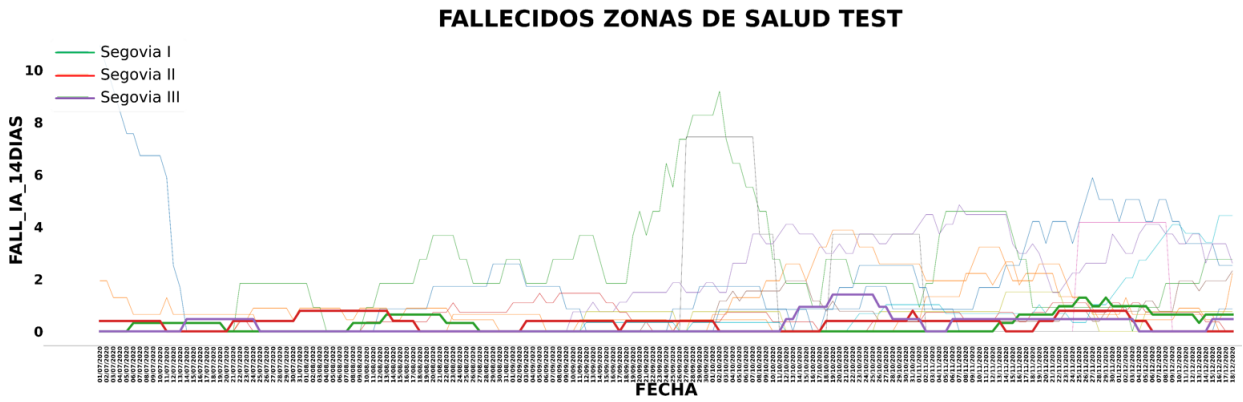


Figura 9.3: Gráfica de muertes por COVID-19 en zonas de salud test en el periodo de 14 días.

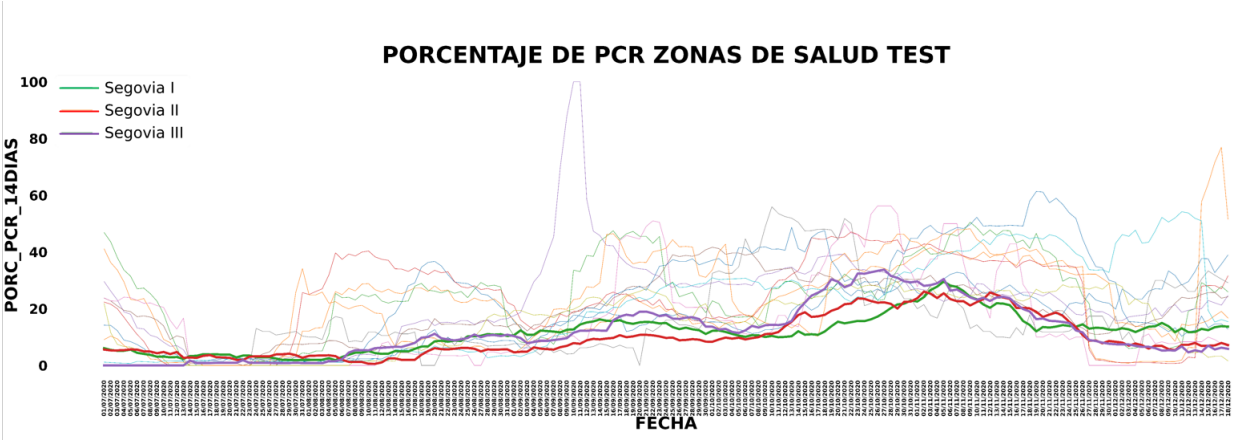


Figura 9.4: Gráfica de porcentaje PCR en zonas de salud test en el periodo de 14 días.

Podemos apreciar en las gráficas generadas como a partir de octubre los contagios detectados comenzaron a aumentar en las tres zonas de salud estudiadas descendiendo a finales de noviembre. Este descenso pudo ser provocado por todas las medidas aplicadas por la Junta de Castilla y León en toda la comunidad destacando aquellas aplicadas el 6 de noviembre. Se observa también una subida en el número de pruebas realizadas en agosto y septiembre y como la zona de salud Segovia III tuvo una mayor curva de contagios que las otras zonas evaluadas.

El número de fallecidos se mantiene bajo respecto a las zonas de entrenamiento aumentando ligeramente en octubre y noviembre, donde hemos detectado un gran número de contagios.

9.1.2. Datos de medidas test

Para la obtención de las medidas aplicadas en la provincia de Segovia se han usado los mismos métodos ya vistos en el apartado 5.3.2, consultando todos aquellos documentos oficiales publicados por la Junta de Castilla y León.

En el caso de estas zonas de test, pese a no aplicarse un confinamiento total como los vistos en las zonas de entrenamiento, en el periodo comprendido entre el 27 de noviembre y 8 de diciembre se aisló la provincia de Segovia del resto con el objetivo de mantener el bajo número de contagios registrado, por tanto, en nuestro proyecto consideraremos ese periodo de aislamiento como un confinamiento.

9.1.3. Datos de movilidad test

Para la obtención del número de entradas a las zonas de salud test, se han usado los mismos portales y herramientas descritos en el apartado 5.4, generando datasets y variables idénticas a las obtenidas en las zonas de entrenamiento.

Toda la información referente a la movilidad en las zonas de test ha sido guarda y analizada en un documento excel individual, siendo posteriormente añadida al dataset de movilidad de las zonas de entrenamiento, obteniendo así un dataset global de la movilidad en todas las zonas estudiadas en el proyecto.

Al igual que en las zonas de entrenamiento, se han creado dos paneles informativos donde se establece una calificación del tipo de número de entradas más frecuente en cada zona de salud de test:

CLASIFICACIÓN MOVILIDAD(NUM VIAJES)				
ZONA DE SALUD	MOVILIDAD GLOBAL	MOVILIDAD RESIDENCIA	ACTIVIDAD €/MOV TRABAJO	MOVILIDAD OTROS
SEGOVIA I	163556.48	70610.68	6024.17	86921.61
SEGOVIA II	110178.52	59894.26	4261.90	46022.35
SEGOVIA III	169958.86	55030.78	4991.41	109936.66

MUY ALTA
ALTA
MEDIA
BAJA
MUY BAJA

Figura 9.5: Panel calificación por tipo de movilidad zonas test

CLASIFICACIÓN MOVILIDAD (NUM VIAJES EN FUNCIÓN DE HABITANTES)				
ZONA DE SALUD	MOVILIDAD GLOBAL	MOVILIDAD RESIDENCIA	ACTIVIDAD €/MOV TRABAJO	MOVILIDAD OTROS
SEGOVIA I	11.15	3.61	0.32	7.21
SEGOVIA II	6.09	3.31	0.23	2.54
SEGOVIA III	7.4	3.19	0.27	3.93

MUY ALTA
ALTA
MEDIA
BAJA
MUY BAJA

Figura 9.6: Panel calificación por tipo de movilidad zonas test en función de la población

Entradas a zonas de salud

De manera idéntica a las zonas de entrenamiento, a continuación se muestran las gráficas del número de entradas de cada una de las zonas de salud de la capital de Segovia para un análisis preliminar de como ha variado la movilidad en el periodo de tiempo estudiado.

Basándonos en la estructura vista en el apartado 6.2, cada una de las gráficas generadas analiza la movilidad según el tipo de entrada, a saber: movilidad total, por vuelta a residencia habitual, por trabajo y por otros motivos. Las gráficas generadas muestran la comparación entre la movilidad en las zonas usadas como entrenamiento y las zonas usadas como test, conservando la misma estructura vista en las zonas de entrenamiento.

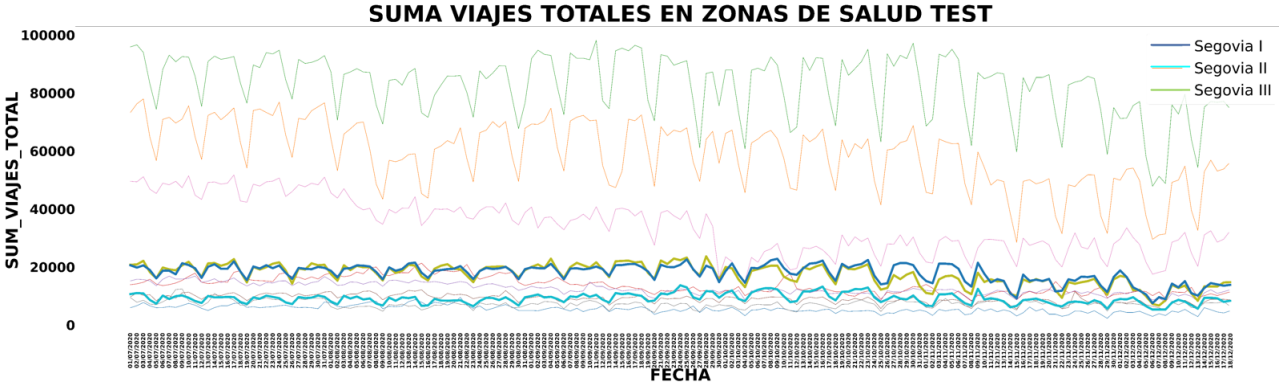


Figura 9.7: Gráfica de entradas a zonas salud test totales

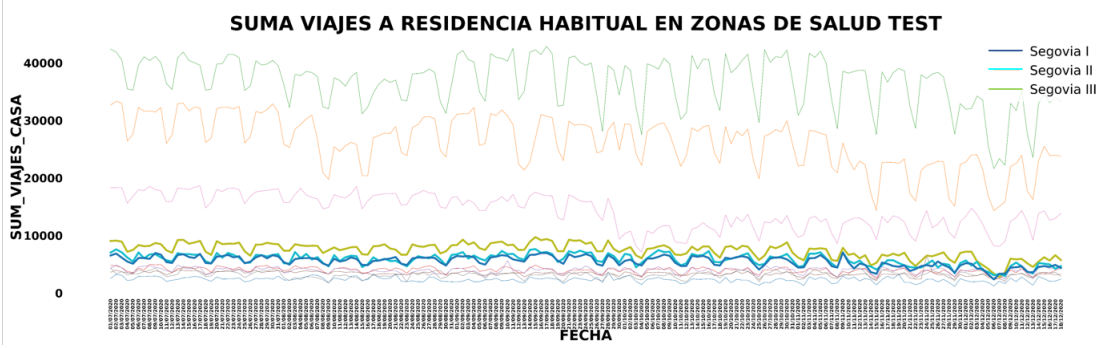


Figura 9.8: Gráfica de entradas a zonas salud test por vuelta a residencia habitual

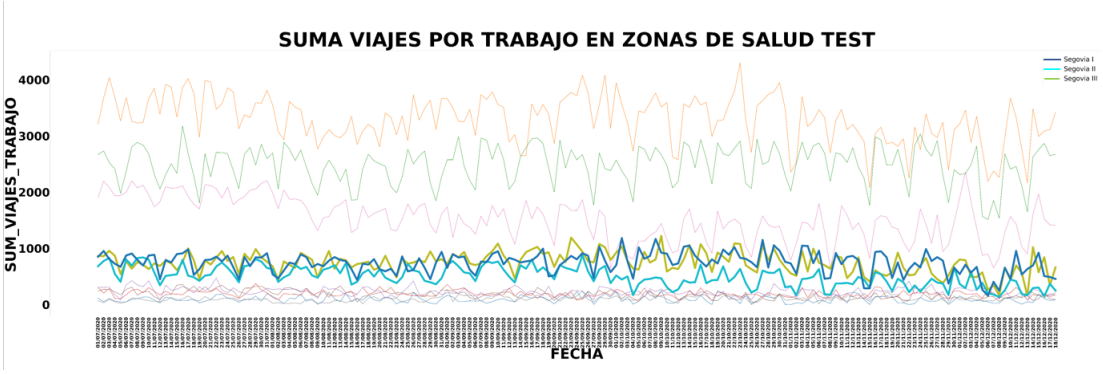


Figura 9.9: Gráfica de entradas a zonas salud test por trabajo

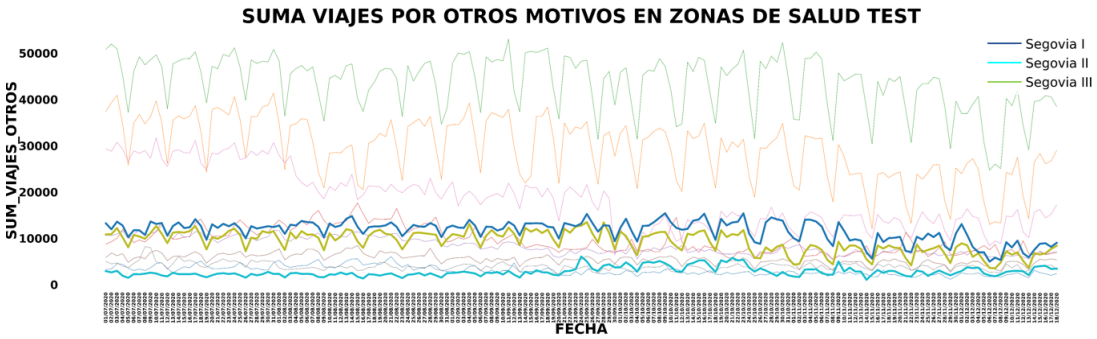


Figura 9.10: Gráfica de entradas a zonas salud test por otros motivos

A través de las gráficas generadas podemos observar como la zona de salud de Segovia II posee una actividad menor que las zonas I y III, y como a partir de noviembre el número de entradas a cada una de las zonas se redujo respecto a meses anteriores. Esta bajada pudo estar provocada por las medidas de cierre de la CA, toque de queda nocturno o cierre de hostelería, gimnasios y centros comerciales, medidas que como ya sabemos causaron un gran impacto en la movilidad.

Se observa como las zonas de Segovia I y III poseen mayor número de entradas por trabajo u otros motivos, lo que se refleja en un mayor número de contagios en el caso de la zona Segovia III.

Entradas a municipios y provincia

En este capitulo, se ha querido profundizar más acerca de la movilidad en las zonas de salud estudiadas. Es por ello, por lo que se ha realizado un estudio de las entradas

al municipio de Segovia Capital, así como a la provincia de Segovia. La motivación de este estudio se debe a que al ser estas zonas pertenecientes a una ciudad capital, en ellas se concentran toda la actividad de la provincia, y por tanto, todas las entradas provenientes de otras zonas y territorios. En este caso, el número de entradas no ha podido ser clasificado por tipos por lo que solamente se mostrarán las entradas totales al municipio y a la provincia de Segovia. Se ha creado un dataset único para el análisis de este tipo de entradas.

Municipios: Para el estudio del número de entradas al municipio de Segovia se ha hecho uso de la zonificación por municipios realizada por el proyecto Datos abiertos descrito en el apartado 5.4.2, descargando todos aquellos ficheros de texto almacenados en la carpeta "**maestra1-mitma-municipios**" y obteniendo los datos finales mediante la herramienta de conversión y obtención (apartado 5.4.3).

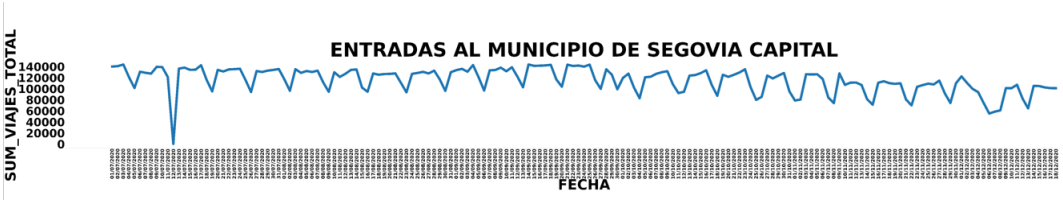


Figura 9.11: Gráfica de entradas a municipio Segovia Capital

La gráfica muestra como el número de entradas al municipio tuvo una tendencia constante a lo largo de la segunda ola, a excepción de noviembre y diciembre donde se detectó una ligera bajada, posiblemente provocada por las restricciones aplicadas.

El día 12/07/2020, posee un valor de 0 por la ausencia de registro de datos de movilidad en esa fecha.

Provincia:

La Figura 9.12 muestra la gráfica del número de entradas a la provincia de Segovia, siendo el eje Y el porcentaje de viajeros que entraron a la provincia usando como referencia un periodo perteneciente a la antigua normalidad (periodo antes del COVID-19).

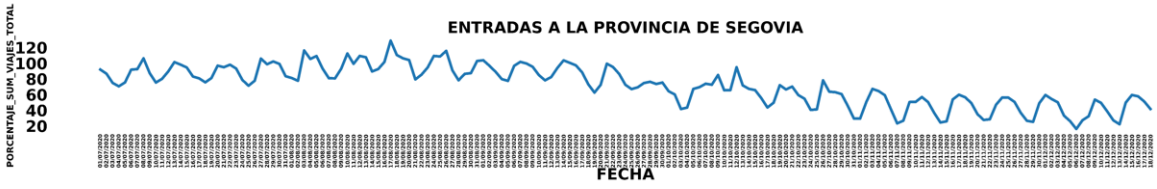


Figura 9.12: Gráfica de entradas a provincia de Segovia

La gráfica generada nos deja ver claramente como las entradas a la provincia fueron altas en julio y agosto, descendiendo a partir de octubre. Podemos apreciar como a partir del 30 de octubre el número de entradas a la provincia descendió bruscamente respecto a meses anteriores, probablemente a causa de la aplicación del cierre perimetral de la Comunidad de Castilla y León. Este cierre provocó un gran descenso en el número de entradas a Segovia por la limitación de entradas desde comunidades como Madrid.

9.1.4. Imágenes gráficas zonas test

Al igual que en el apartado 6.1.1, se proporciona una carpeta ligada a la memoria, donde se ubicarán todas las gráficas usadas en alta resolución, para una correcta lectura y comprensión por parte del lector.

9.2. Aplicación método de aprendizaje: Clustering

Se ha aplicado el método de aprendizaje clustering usado en las zonas de entrenamiento para la obtención de los distintos tipos de subidas y bajadas en las zonas de test evaluadas.

Datos:

Para la aplicación del clustering se ha hecho uso del dataset global de salud, donde se encuentran integradas tanto las zonas de entrenamiento como las de test. Se hace uso de este dataset para la clasificación de los datos epidemiológicos de las zonas de salud de Segovia dentro de un marco de estudio global. Dicho dataset tendrá la misma estructura descrita en el apartado 7.3.1.

Aplicación del método:

Se hará uso de la configuración final usada en las zonas de entrenamiento (variables y tipo de clustering) debido a los buenos resultados obtenidos.

El número de k clusters quedará fijado a 3 al igual que en implementaciones anteriores, obteniendo mediante la función de etiquetado, los tres estados (bien (0), regular (1) y mal (2)) descritos anteriormente (Apartado 7.3.5).

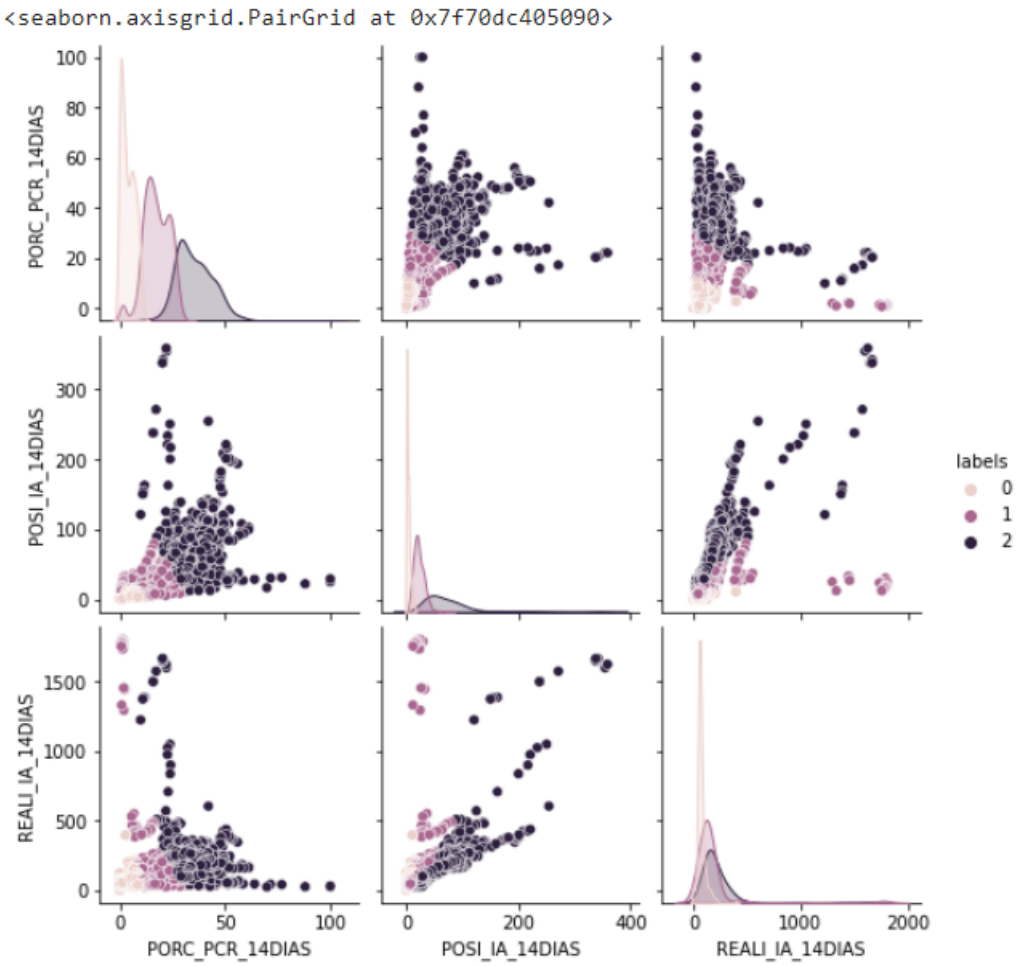


Figura 9.13: Resultados clustering zonas test

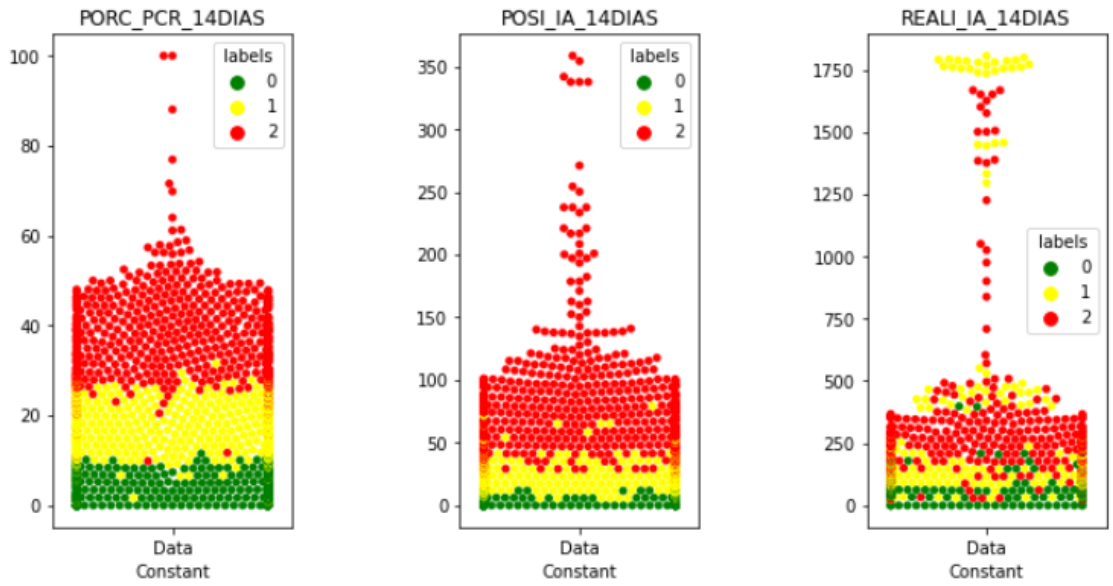


Figura 9.14: Resultados aplicación librería gráfica seaborn en zonas test.

Como podemos ver, los clusters obtenidos son claramente distinguibles para cada una de las variables usadas, por lo que podemos asegurar que la configuración usada en las zonas de entrenamiento también es efectiva en las zonas de test, siendo ambas muy semejantes debido a la integración de datos.

Análisis de resultados.

Al igual que en las zonas de entrenamiento, se han obtenidos distintas gráficas donde podemos observar todas las subidas y bajadas detectada en las zonas de test de Segovia Capital. Dichas subidas y bajadas serán clasificadas y analizadas en tablas siguiendo el mismo criterio (7.3.5) y estructura (7.34) ya vistos. Destacar como en las zonas de Segovia I y Segovia II se observa una menor alternancia entre subidas y bajadas.

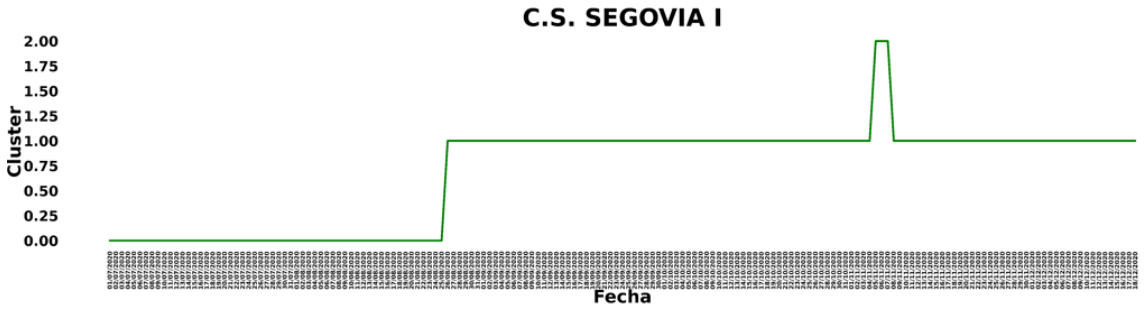


Figura 9.15: Gráfica resultados clustering Segovia I

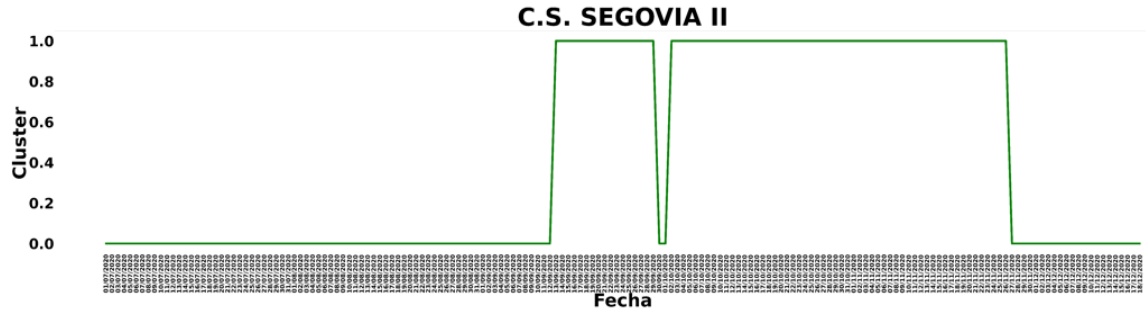


Figura 9.16: Gráfica resultados clustering Segovia II

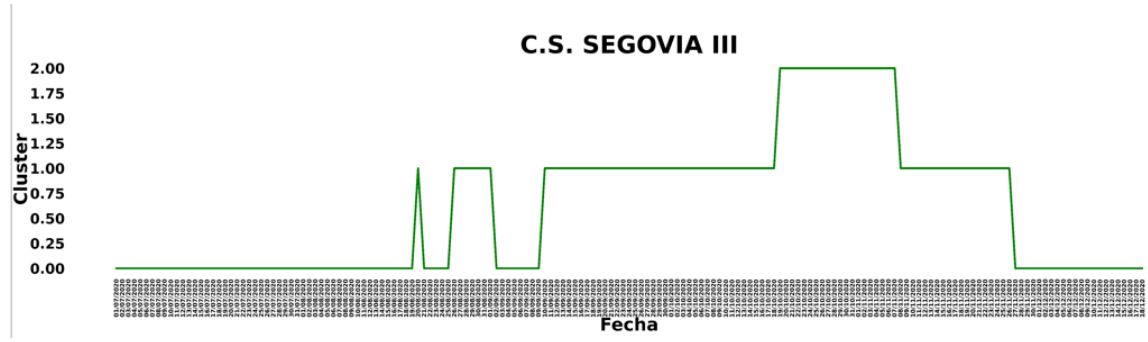


Figura 9.17: Gráfica resultados clustering Segovia III

Estás gráficas en alta definición serán adjuntadas a la carpeta mencionada en el apartado 9.1.4

Tablas de análisis:

SEGOVIA I			
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
SUBIDA BAJA	25/08		MOVLIDAD GENERAL, OTROS Y TRABAJO 21/08 ALTA(FESTIVO 15/08) EN ZS. NUMERO DE ENTRADAS A MUNICIPIO ALTO(FESTIVO 15/08) NUMERO DE ENTRADAS A PROVINCIA ALTO.
SUBIDA INTERMEDIA	04/11		AUMENTO DE MOVILIDAD A ZS POR FESTIVOS(26/10, 1/11). GRAN MOVILIDAD POR TRABAJO A ZS. GRAN NUMERO DE ENTRADAS A MUNICIPIO POR FIESTA 26/10.
BAJADA INTERMEDIA	08/11	25/10	BAJADA DE MOVILIDAD TOTAL TRAS FESTIVO(26/10) EN ZS. BAJADA DE ENTRADAS A MUNICIPIO TRAS FESTIVO 26/10 NUMERO DE ENTRADAS A PROVINCIA BAJO. MEDIDAS DE OCTUBRE

Figura 9.18: Tabla análisis clustering Segovia I

SEGOVIA II			
TIPO DE CAMBIO	FECHA	FECHA ORIGEN (14 DÍAS PREV)	MEDIDAS APLICADAS/CAUSAS
SUBIDA BAJA	12/09		GRAN MOVILIDAD TOTAL(CASA-TRABAJO) EN ZS GRAN NUMERO DE ENTRADAS AL MUNICIPIO GRAN NUMERO DE ENTRADAS A PROVINCIA
BAJADA BAJA	27/11	13/11	MEDIDAS 6 DE NOVIEMBRE BAJADA GRAN DE LA MOVILIDAD TOTAL A LA ZS ENTRADAS A MUNICIPIO BAJAS RESPECTO A MESES ANTERIORES ENTRADAS A PROVINCIA MUY BAJAS

Figura 9.19: Tabla análisis clustering Segovia II

SEGOVIA III			
TIPO DE CAMBIO	**FECHA**	**FECHA ORIGEN (14 DÍAS PREV)**	**MEDIDAS APLICADAS/CAUSAS**
SUBIDA BAJA	25/08		MOVILIDAD TOTAL ALTA. MOVILIDAD POR FESTIVO 15/08 (MOVILIDAD OTRO MUY ALTA) EN ZS NUMERO DE ENTRADAS A MUNICIPIO ALTAS NUMERO DE ENTRADAS A PROVINCIA ALTO
BAJADA BAJA	02/09	19/08	MEDIDAS 17 Y 21 AGOSTO LIGERA BAJADA DE NUMERO DE ENTRADAS A MUNICIPIO
SUBIDA BAJA	09/09		MOVILIDAD GENERAL A ZS ALTA. MOVILIDAD OTROS MUY ALTA(4/09) ALTO NUMERO DE ENTRADAS A MUNICIPIO ALTO NUMERO DE ENTRADAS A PROVINCIA
SUBIDA INTERMEDIA	18/10		MOVILIDAD GENERAL ALTA. MUY ALTO NUMERO DE ENTRADAS A ZS POR OTROS Y TRABAJO. NUMERO DE ENTRADAS A MUNICIPIO ALTO(FESTIVO 12/10)
BAJADA INTERMEDIA	08/11	25/10	MEDIDAS 17 Y 24 OCTUBRE BAJADA DE MOVILIDAD GENERAL EN ZS BAJADA DE NUMERO DE ENTRADAS A MUNICIPIO RESPECTO A DIAS ANTERIORES NUMERO DE ENTRADAS A PROVINCIA MUY BAJO
BAJADA BAJA	27/11	13/11	MEDIDAS 6 NOVIEMBRE MOVILIDAD GENERAL EN ZS BAJA NUMERO DE ENTRADAS A MUNICIPIO BAJO NUMERO DE ENTRADAS A PROVINCIA MUY BAJO

Figura 9.20: Tabla análisis clustering Segovia III

A través de las diferentes tablas creadas para el análisis de los resultados obtenidos mediante la aplicación de clustering, vemos como el número de entradas en las zonas de test ha sido un factor muy relevante en las subidas detectadas. Tanto las entradas a las zonas de salud, como las entradas al municipio y a la provincia, han sido las principales causas en las subidas registradas por el método clustering empleado. Las entradas previas a días festivos también han sido calificadas como causantes de los aumentos de contagios en cada una de las zonas.

Medidas como las aplicadas en octubre o noviembre, han sido clave para las bajadas detectadas por el clustering. En el caso de Segovia III las medidas aplicadas en agosto (relacionadas con las interacciones sociales) también tuvieron efectos en la bajada de contagios.

En resumen, podemos confirmar que al igual que observábamos en la fase de entrenamiento, en esta fase de evaluación la movilidad también es un factor muy relevante que determina las distintas subidas y bajadas detectadas por el método clustering.

9.3. Aplicación modelo de aprendizaje:GBT

Al igual que en la fase de entrenamiento, se hará uso de un modelo de aprendizaje GBT para la obtención de las medidas más efectivas en las zonas usadas como test (validando posteriormente los resultados con los obtenidos en la aplicación del clustering).

Se hará uso de la implementación XGBT, que tal y como hemos podido ver en la comparación realizada en el apartado 8.17 es la implementación más adecuada en cuanto a tiempo y efectividad.

Al igual que en el caso anterior, usaremos la importancia de cada predictor para la obtención de aquellas medidas más eficientes, siendo estos resultados validados mediante la importancia por permutación.

9.3.1. Datos

De la misma manera que en las zonas de entrenamiento, se crearán datasets individuales de cada una de las zonas de salud de Segovia capital, de los cuales, solo se hará uso de las variables de medidas aplicadas (predictores) y de la variable MOV_CLUSTER (variable de respuesta).

9.3.2. Aplicación de modelo

El modelo XGBT usado poseerá la misma configuración y elementos ya especificados para las zonas de entrenamiento (apartado 8.5.2). De esta manera, se obtendrán las medidas más relevantes, clasificadas a continuación por zona de salud usando como referencia la tabla ya vista en la Figura 8.18:

IMPORTANCIA PREDICTORES(rmse)																								
	CIERRE_ HOSTELERIA	CIERRE_ GIMNASIOS	CIERRE_ C_ COMERCIALES	REST_ ACCESO _ CCAA	REST_ ACCESO _ZS	TOQUE_ NOCTURNO	LIMIT_ REUN_ 10	LIMIT_ REUN_ 6	RED_ PERS_ SACTV_ 50	PROHIB_ CONSUM _ BARRA	AFORO _ HOST_ 75	AFORO _ HOST_ 50	LIMIT _ MESA _ 10	LIMIT _ MESA _ 6	CIERRE _ PEÑAS	CIERRE _ DISCO	LIMIT_ OC_ NOCT_ 23 PM	LIMIT_ OC_ NOCT_ 1AM	PROHIB _ FUMAR	PROHIB_ VISIT_ RESI
SEGOVIA I	0.020			0.668	0.004		0.001		0.306												
SEGOVIA II	0.025			0.156	0.023		0.112		0.558												0.123
SEGOVIA III	0.333			0.138	0.060	0.046	0.124		0.118												0.177

Figura 9.21: Importancia predictores zonas test

IMPORTANCIA MEDIA POR PERMUTACIÓN(rmse)

	CIERRE_ HOSTELERIA	CIERRE_ GIMNASIOS	CIERRE_ C_ COMERCIALES	REST_ ACCESO_ CCAA	REST_ ACCESO _ZS	TOQUE_ NOCTURNO	LIMIT_ REUN_10	LIMIT_ REUN_6	RED_ PERS_ SACTV_ 50	PROHIB_ CONSUM _ BARRA	AFORO _ HOST_ 75	AFORO _ HOST_ 50	LIMIT _ MESA _10	LIMIT _ MESA _6	CIERRE_ PEÑAS	CIERRE_ DISCO	LIMIT_ OC_ NOCT_ 23 PM	LIMIT_ OC_ NOCT_ 1AM	PROHIB _ FUMAR	PROHIB_ VISIT_ RESI
SEGOVIA I	0.082			0.413	0.042		0.0009		0.282											
SEGOVIA II	0.002			0.060	0.012		0.048		0.201											0.093
SEGOVIA III	0.123			0.190	0.045	0.020	0.197		0.261											0.100

Figura 9.22: Importancia predictores por permutación zonas test

Al igual que en el entrenamiento, podemos ver en las tablas generadas que la importancia de los predictores (Figura 9.21) coincide con la obtención de un valor distinto de 0 en la importancia en cada iteración (Figura 9.22), esto nos permite validar que aquellos predictores con gran importancia son usados en el proceso de aprendizaje del modelo y por tanto pueden ser considerados como medidas eficaces en la zona a la que corresponde el modelo.

Resultados finales:

Al igual que en el apartado 8.5.3, se analizarán los resultados y conclusiones obtenidos, exponiendo por cada zona de test aquellas medidas más eficaces y las distintas razones que validen dichos resultados. Se utilizará toda la información obtenida en este capitulo, en cuanto a análisis y aplicación de clustering sobre las zonas estudiadas.

Segovia I:

MEDIDA	**IMPORTANCIA PREDICTOR(rmse)**
Restricción de acceso a la CCAA	0.668
Reducción de personas en los sectores de actividad a 50 %	0.306

Tabla 9.1: Medidas más efectivas en la zona de salud de Segovia I

Segovia II:

MEDIDA	**IMPORTANCIA PREDICTOR(rmse)**
Reducción de personas en los sectores de actividad a 50 %	0.558
Restricción de acceso a la CCAA	0.156
Prohibida visita a residencias	0.123
Limite de reuniones a 10 personas	0.112

Tabla 9.2: Medidas más efectivas en la zona de salud de Segovia II

Segovia III:

MEDIDA	IMPORTANCIA PREDICTOR(rmse)
Cierre de Hostelería	0.333
Prohibida visita a residencias	0.177
Restricción de acceso a CCAA	0.138
Limite de reuniones a 10 personas	0.124
Reducción de personas en los sectores de actividad a 50 %	0.118

Tabla 9.3: Medidas más efectivas en la zona de salud de Segovia III

En las tres zonas usadas como test, podemos apreciar como la medida de restricción de acceso a la CCAA ha sido calificada como más efectiva en la lucha contra la pandemia. Podemos encontrar una justificación remontándonos a los análisis previos de movilidad realizados en las zonas de test, donde hemos visto como la aplicación de esta medida reducía drásticamente el flujo de entradas tanto a las zonas, como al municipio y provincia. Otra medida con un gran impacto en la movilidad, como es el cierre de la hostelería, ha sido obtenida como efectiva en la zona de salud de Segovia III (zona con el mayor número de entradas de las estudiadas).

Medidas relacionadas con la reducción de la interacción social o las reuniones sociales también han sido eficaces en las zonas evaluadas, especialmente las medidas de reducción del número de personas reunidas a 10 y el limite al 50 % de personas en sectores de actividad. Otra de las medidas detectada como eficaz ha sido la prohibición de visitas a residencias.

Medidas ineficaces:

Una vez obtenidas las medidas más eficaces en cada una de las zonas de test estudiadas, al igual que en las zonas de entrenamiento, podemos observar como algunas medidas de las estudiadas han tenido una eficacia muy baja o nula en la mayoría de las zonas. A continuación se indican que medidas no han sido obtenidas como no eficaces.

MEDIDA	IMPORTANCIA TOTAL
Cierre de gimnasios e instalaciones deportivas	0
Cierre de centros comerciales y grandes establecimientos	0
Prohibido el consumo en barra y de pies en recintos de hostelería	0
Limitación al 75 % del aforo permitido en el interior delos recintos de hostelería	0
Limitación al 50 % del aforo permitido en el interior delos recintos de hostelería	0
Limite de reuniones a 10 personas máximo	0
Limite de reuniones a 6 personas máximo	0
Limite a 10 personas en mesas de hostelería	0
Limite a 6 personas en mesas de hostelería	0
Cierre de peñas y organizaciones con motivos festivos	0
Cierre de discotecas y locales de ocio nocturno	0
Limite de apertura de locales de ocio nocturno a las 23PM	0
Limite de apertura de locales de ocio nocturno a las 1AM	0
Prohibición de fumar sin distancia mínima	0

Tabla 9.4: Medidas no efectivas en zonas de salud test

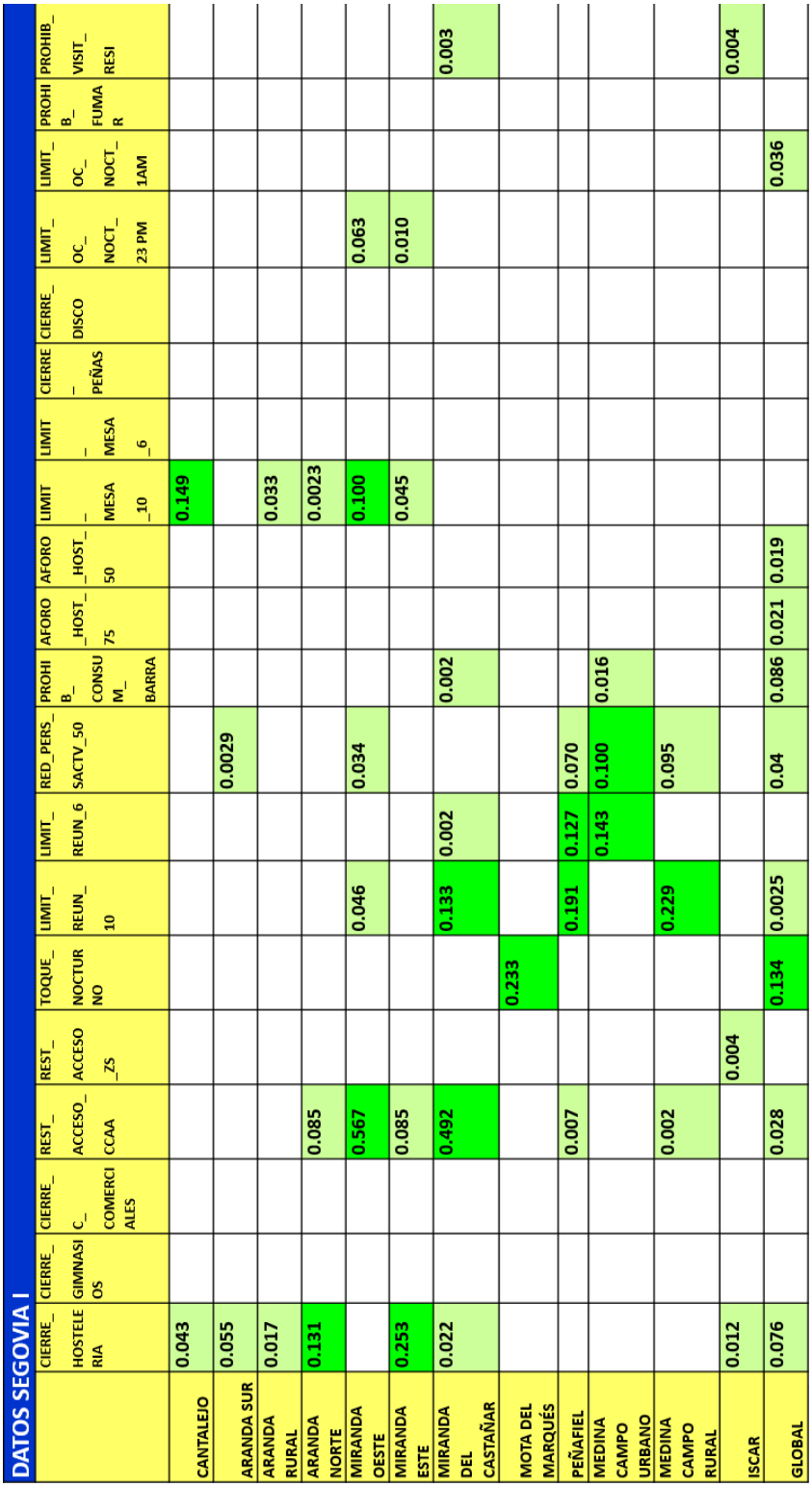
Como conclusión, podremos afirmar que aquellas medidas con poca o nula eficacia en las zonas de evaluación estudiadas podrían ser descartadas o aplicadas durante un menor tiempo, de cara a una menor restricción de la población e impacto económico.

9.3.3. Aplicación de modelos de zonas de entrenamiento

Como fase final de este proceso de evaluación, se ha aplicado a los datos de las zonas de salud de Segovia capital, los modelos de aprendizaje creados (modelos XGBT) de las zonas usadas como entrenamiento.

La aplicación de estos modelos nos permite saber por cada zona de entrenamiento, cuales de las medidas aplicadas en dichas zonas serían más efectivas (de aplicarse), en el contexto epidemiológico dado por las zonas de test.

Para este proceso se emplean los datasets individuales de cada una de las zonas de salud de Segovia capital, usando en cada dataset cada uno de los modelos finales ya entrenados de las zonas de entrenamiento (evaluando el rmse obtenido y la importancia por permutación de cada predictor usado). Nos basaremos únicamente en la importancia por permutación debido a que al usar un modelo ya entrenado, la importancia por predictor queda fija para poder realizar el proceso de aprendizaje con los datos de las zonas de test.



DATOS SEGOVIA I	CIERRE_ HOSTELERIA	CIERRE_ GIMNASIOS	CIERRE_ C_ COMERCIALES	REST_ ACCESO_ CCAA	REST_ ACCESO _ZS	TOQUE_ NOCTURNO	LIMIT_ REUN_ 10	LIMIT_ REUN_6	RED_PERS_ SACTV_50	PROHIB_ CONSUM_ BARRA	AFORO _HOST_ 75	AFORO _HOST_ 50	LIMIT _ MESA _10	LIMIT _ MESA _6	CIERRE _ PEÑAS	CIERRE_ DISCO	LIMIT_ OC_ NOCT_ 23 PM	LIMIT_ OC_ NOCT_ 1AM	PROHIB_ FUMAR	PROHIB_ VISIT_ RESI
CANTALEJO	0.043												0.149							
ARANDA SUR	0.055								0.0029											
ARANDA RURAL	0.017												0.033							
ARANDA NORTE	0.131			0.085									0.0023							
MIRANDA OESTE				0.567			0.046		0.034				0.100				0.063			
MIRANDA ESTE	0.253			0.085									0.045				0.010			
MIRANDA DEL CASTAÑAR	0.022			0.492			0.133	0.002		0.002										0.003
MOTA DEL MARQUÉS						0.233														
PEÑAFIEL				0.007			0.191	0.127	0.070											
MEDINA CAMPO URBANO								0.143	0.100	0.016										
MEDINA CAMPO RURAL				0.002			0.229		0.095											
ISCAR	0.012				0.004															0.004
GLOBAL	0.076			0.028		0.134	0.0025		0.04	0.086	0.021	0.019						0.036		

Figura 9.23: Resultados aplicación modelos zonas de entrenamiento sobre segovia I

DATOS SEGOVIA II																						
	CIERRE_ HOSTELERIA	CIERRE_ GIMNASIOS	CIERRE_ C_ COMERCIALES	REST_ ACCESO_ CCAA	REST_ ACCESO_ ZS	TOQUE_ NOCTURNO	LIMIT_ REUN_ 10	LIMIT_ REUN_ 6	RED_PERS _ SACTV_50	PROHIB_ CONSUM_ BARRA	AFORO _HOST _ 75	AFORO _HOST _ 50	LIMIT _ MESA _10	LIMIT _ MESA _6	CIERRE_ PEÑAS	CIERRE_ DISCO	LIMIT_ OC_ NOCT_ 23 PM	LIMIT_ OC_ NOCT_ 1AM	PROHIB _ FUMAR	PROHIB_ VISIT_ RESI
CANTALEJO				0.006			0.022						0.154							
ARANDA SUR 6	0.001						0.069		0.0023				0.010							0.008
ARANDA RURAL													0,091							
ARANDA NORTE				0.048			0.216						0.003 9							0.008
MIRANDA OESTE				0.024	0.0031		0.049		0.007				0.045				0.061			
MIRANDA ESTE				0.0078			0.0104						0.043				0.026			
MIRANDA DEL CASTAÑAR				0.116		0.035	0.285													
MOTA DEL MARQUÉS	0.028			0.120		0.130	0.108													0.033
PEÑAFIEL				0.049				0.077	0.040											
MEDINA CAMPO URBANO								0.152	0.176		0.016									0.062
MEDINA CAMPO RURAL	0.044					0.109	0.0382		0.221											0.028
ISCAR					0.006				0.032											
GLOBAL						0.066	0.048	0.034	0.076	0.06 8	0.018			0.023			0.041	0.059		0.014

Figura 9.24: Resultados aplicación modelos zonas de entrenamiento sobre segovia II

DATOS SEGOVIA III	CIERRE_ HOSTELERIA	CIERRE_ GIMNASIOS	CIERRE_ C_ COMERCIALES	REST_ ACCESO_ CCAA	REST_ ACCESO _ 25	TOQUE_ NOCTURNO	LIMIT_ REUN_ 10	LIMIT_ REUN_6	RED_PERS_ SACTV_50	PROHIB_ CONSUM_ BARRA	AFORO _HOST_ 75	AFORO _HOST_ 50	LIMIT _ MESA _10	LIMIT _ MESA _6	CIERRE_ PEÑAS	CIERRE_ DISCO	LIMIT_ OC_ NOCT_ 23 PM	LIMIT_ OC_ NOCT_ 1AM	PROHIB _ FUMAR	PROHIB_ VISIT_ RESI
CANTALEJO	0.058						0.047						0.147							
ARANDA SUR	0.075								0.004				0.013							
ARANDA RURAL	0.021												0.065							
ARANDA NORTE	0.182			0.071			0.194													
MIRANDA OESTE	0.0073			0.328	0.003		0.060		0.036				0.113				0.066			
MIRANDA ESTE	0.359			0.066			0.031						0.113				0.034			
MIRANDA DEL CASTAÑAR	0.027			0.332		0.019	0.246													0.0084
MOTA DEL MARQUÉS				0.011		0.122	0.035													
PEÑAFIEL	0.0061						0.068	0.0942	0.047											
MEDINA CAMPO URBANO								0.069	0.105		0.002									
MEDINA CAMPO RURAL						0.081	0.016		0.114											
ISCAR	0.015				0.003				0.020											0.006
GLOBAL	0.088			0.024	0.011	0.080	0.031		0.059	0.071		0.0023	0.0002	0.0023			0.0015	0.053		

Figura 9.25: Resultados aplicación modelos zonas de entrenamiento sobre segovia III

Análisis de resultados obtenidos con aplicación de modelos de zonas de entrenamiento

Una vez aplicados los modelos de las zonas de entrenamiento en las zonas de test, y obtenidas aquellas medidas más efectivas, a través de una serie de tablas veremos la eficacia que podrían tener las medidas de otras zonas en Segovia I, Segovia II y Segovia III, en función a sus estados epidemiológicos (variable de resultados clustering **MOV_-CLUSTER**).

Cada tabla corresponderá a un modelo aplicado y solo se indicarán aquellas medidas con una importancia por permutación **superior a 0.1**, correspondiente a la valoración de **medida efectiva** dada por la tabla de referencia de la Figura 8.18.

Aplicación modelo Cantalejo:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	Limite de 10 personas en mesas de hostelería.
SEGOVIA II	Limite de 10 personas en mesas de hostelería.
SEGOVIA III	Limite de 10 personas en mesas de hostelería.

Tabla 9.5: Resultados aplicación modelo Cantalejo.

Aplicación modelo Aranda Sur:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	N/A
SEGOVIA II	N/A
SEGOVIA III	N/A

Tabla 9.6: Resultados aplicación modelo Aranda Sur.

Aplicación modelo Aranda Rural:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	N/A
SEGOVIA II	N/A
SEGOVIA III	N/A

Tabla 9.7: Resultados aplicación modelo Aranda Rural.

Aplicación modelo Aranda Norte:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	Cierre de hostelería
SEGOVIA II	Limite de reuniones a 10 personas
SEGOVIA III	Cierre de Hostelería y limite de reuniones a 10 personas

Tabla 9.8: Resultados aplicación modelo Aranda Norte.

Aplicación modelo Miranda Oeste:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	Restricción de acceso a la CCAA y limite de personas a 10 en mesa de hostelería
SEGOVIA II	N/A
SEGOVIA III	Restricción de acceso a la CCAA y limite de personas a 10 en mesa de hostelería

Tabla 9.9: Resultados aplicación modelo Miranda Oeste.

Aplicación modelo Miranda Este:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	Cierre de hostelería
SEGOVIA II	N/A
SEGOVIA III	Cierre de hostelería y limite de personas a 10 en mesa de hostelería

Tabla 9.10: Resultados aplicación modelo Miranda Este.

Aplicación modelo Miranda del Castañar:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	Restricción de acceso a la CCAA y limite de reuniones a 10 personas
SEGOVIA II	Restricción de acceso a la CCAA y limite de reuniones a 10 personas
SEGOVIA III	Restricción de acceso a la CCAA y limite de reuniones a 10 personas

Tabla 9.11: Resultados aplicación modelo Miranda del Castañar.

Aplicación modelo Mota del Marqués:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	Toque de queda nocturno
SEGOVIA II	Restricción de acceso a la CCAA, toque de queda nocturno y limite de reuniones a 10 personas.
SEGOVIA III	Toque de queda nocturno

Tabla 9.12: Resultados aplicación modelo Mota del Marqués.

Aplicación modelo Peñafiel:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	Limite de reuniones a 10 personas y limite de reuniones a 6 personas
SEGOVIA II	N/A
SEGOVIA III	N/A

Tabla 9.13: Resultados aplicación modelo Peñafiel.

Aplicación modelo Medina del Campo Urbano:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	Limite de reuniones a 6 personas y reducción de personas en sectores de actividad al 50 %
SEGOVIA II	Limite de reuniones a 6 personas y reducción de personas en sectores de actividad al 50 %
SEGOVIA III	Reducción de personas en sectores de actividad al 50 %

Tabla 9.14: Resultados aplicación modelo Medina del Campo Urbano.

Aplicación modelo Medina del Campo Rural:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	Limite de reuniones a 10 personas
SEGOVIA II	Restricción de acceso a la zona de salud y reducción de personas en sectores de actividad al 50 %
SEGOVIA III	Reducción de personas en sectores de actividad al 50 %

Tabla 9.15: Resultados aplicación modelo Medina del Campo Rural.

Aplicación modelo Íscar:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	N/A
SEGOVIA II	N/A
SEGOVIA III	N/A

Tabla 9.16: Resultados aplicación modelo Íscar.

Aplicación modelo Global:

ZONA DE SALUD	MEDIDAS MÁS EFECTIVAS
SEGOVIA I	Toque de queda nocturno
SEGOVIA II	N/A
SEGOVIA III	N/A

Tabla 9.17: Resultados aplicación modelo Global.

A través de las diferentes tablas creadas observamos como en las zonas de salud de Aranda Sur (Tabla 9.6), Aranda Rural (Tabla 9.7), Miranda Este (Tabla 9.10), Peñafiel (Tabla 9.13) e Íscar (Tabla 9.16), no se obtiene ninguna medida como eficaz para su aplicación en las zonas de test de Segovia, esto puede deberse a que las medidas contra el COVID-19 fueron aplicadas en estos territorios en un periodo donde los contagios eran muy altos, mientras que en Segovia eran muy bajos, por lo que no tendría sentido aplicar restricciones en las zonas de salud de Segovia capital. En el caso del modelo global (Tabla 9.17) obtenemos únicamente como medida efectiva, en la zona test de Segovia I, el toque de queda nocturno.

Para el resto de modelos aplicados se detectan como medidas más eficaces aquellas relacionadas con la movilidad (restricción de acceso a CCAA) o la limitación de reuniones sociales.

Capítulo 10

Conclusiones y trabajo futuro

En este capitulo final se incluyen las conclusiones alcanzadas una vez acabado el proyecto y cumplidos los objetivos marcados (ver apartado 1.3.1). También se exponen las posibles ampliaciones del proyecto con vistas al futuro.

10.1. Conclusión

Gracias a este proyecto se ha obtenido una forma alternativa de lucha contra el COVID-19 desde el campo de la informática y la Inteligencia Artificial. Como se ha presentado en el desarrollo del mismo, se han utilizado distintos métodos y modelos de Machine Learning que, a partir de unos datos ya procesados, pudieran dar un contexto epidemiológico de cada zona para la obtención de las medidas y restricciones más efectivas a aplicar.

Las conclusiones obtenidas son que aquellas medidas con un gran impacto en la movilidad exterior (entradas desde otras CCAA) e interior (movilidad entre zonas de salud) han sido detectadas (por las técnicas de Machine Learning empleadas) como las más efectivas, destacando el cierre de la hostelería y la restricción de acceso a la CCAA de Castilla y León y zonas de salud en las que se divide. Dichos resultados demuestran que toda medida que cause impacto en el número de entradas a cualquier territorio resulta eficaz para la bajada de contagios.

Otro resultado obtenido sobre medidas efectivas en las zonas estudiadas han sido aquellas relacionadas con el limite de personas en reuniones y sectores de actividad o la prohibición de visitas a residencias de ancianos. Este tipo de medidas tienen como objetivo el control de la propagación del virus en entornos externos al ámbito doméstico o familiar, donde se ha podido comprobar que la probabilidad de infección es muy alta. La medida de prohibición de visitas a residencias de ancianos ha sido estimada como muy eficaz en la

mayoría de territorios debido al gran número de contagios y muertes provocados en estos centros. En la Figura 10.1 se muestran aquellas medidas más eficaces en todas las zonas estudiadas.

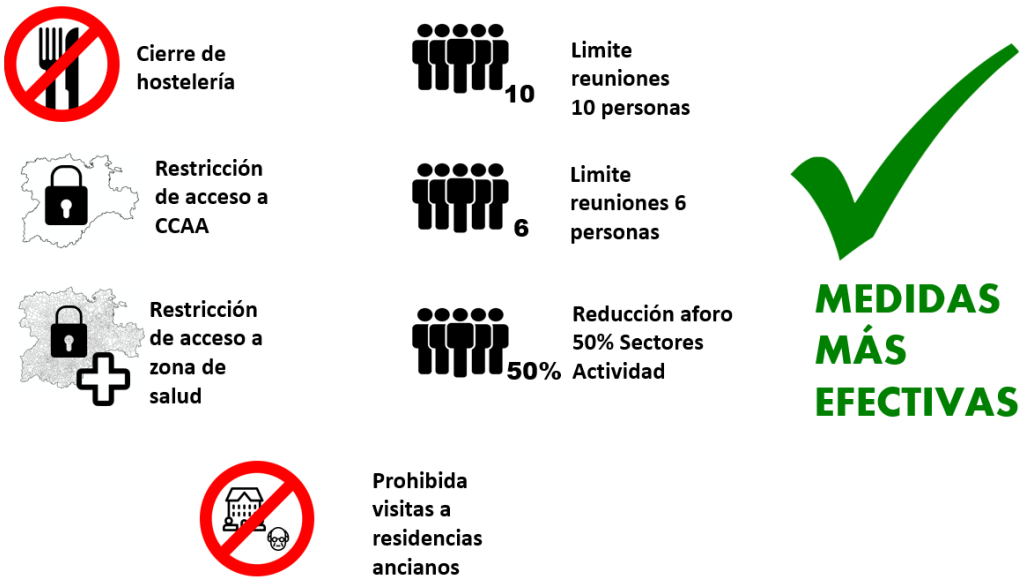


Figura 10.1: Medidas más efectivas zonas de salud estudiadas.

Por último, se ha obtenido que el cierre de gimnasios y centros comerciales han sido medidas ineficaces en la lucha contra la pandemia. Estas restricciones, que han tenido un impacto económico negativo en los negocios del sector, son detectadas como ineficaces probablemente debido a los protocolos adoptados contra la COVID-19 y la reducida interacción entre los clientes de estos establecimientos en comparación con otros negocios como la hostelería, cuyas medidas de cierre han sido estimadas como más efectivas. En la Figura 10.2 se muestran las medidas ineficaces aplicadas en las zonas usadas como caso de estudio.

Figura 10.2: Medidas ineficaces en zonas de salud estudiadas.

10.2. Experiencias y aprendizajes personales

A nivel personal, este proyecto ha supuesto un gran reto debido a factores como la situación provocada por el COVID-19 (tema abordado en el proyecto), que ha modificado la forma de trabajo y estudio, eliminando las ventajas de una comunicación en persona o presencial con mis tutores, por una telemática. El tema escogido para este trabajo ha provocado también una gran fatiga psicológica debido a su presencia tanto en las horas de trabajo, como en las horas de desconexión y descanso. Otro de los factores que ha aumentado este reto ha sido el poco conocimiento previo en los campos de Inteligencia Artificial empleados, que han requerido un gran estudio y dedicación para su correcta comprensión. Entre todas las complicaciones señaladas, además cabe destacar la compatibilidad entre el tiempo dedicado al desarrollo del TFG y las prácticas extracurriculares, algo que ha supuesto mayor carga de trabajo y conocimientos en el día a día.

En cuanto al aprendizaje adquirido con el desarrollo del TFG, me gustaría destacar todos los conocimientos y técnicas usados, las cuales me han aportado una experiencia muy beneficiosa de cara a mi introducción profesional en campos de la Inteligencia Artificial y el procesamiento de datos, sectores a los que me gustaría dedicarme en un futuro. A su vez, este proyecto me ha permitido dar mis primeros pasos en el sector de la investigación, lo que me ha enseñado a valerme por mi mismo y valorar aún más los resultados satisfactorios conseguidos.

10.3. Trabajo futuro

A continuación se exponen las distintas ampliaciones del proyecto ya finalizado.

Ampliación del alcance del proyecto: Tal y como se ha visto, el proyecto expuesto anteriormente se ha centrado en la Comunidad Autónoma de Castilla y León para su desarrollo, limitándose su uso a aquellas zonas de salud dentro de ella. Por tanto, y de cara a implementaciones futuras, se propone ampliar a toda España el marco de trabajo y estudio para la obtención de las medidas más efectivas en cada una de las Comunidades y zonas que componen el país.

Uso de datos de vacunación y nuevas cepas del virus: Se propone el uso de los datos referentes a la vacunación realizada contra el COVID-19 y las nuevas cepas surgidas del virus en los métodos y modelos creados.

Predicción de las subidas y bajadas en cada zona de salud: Debido a la ausencia de información actual en el comportamiento estacional del virus, ha sido imposible poder desarrollar modelos de aprendizaje basados en series temporales para la predicción de las distintas subidas y bajadas en cada zona. Es por ello por lo que se plantea la posibilidad futura de realizar dichos modelos una vez se disponga de toda la información.

Creación de una interfaz de usuario: Debido al tiempo dado para el desarrollo del Trabajo de Fin de Grado, se ha hecho imposible la creación de una interfaz o aplicación que integre los modelos desarrollados para un uso sencillo y accesible. Así se plantea la posibilidad de crear una interfaz intuitiva que permita a cada territorio obtener de manera sencilla las medidas más efectivas.

Bibliografía

[1] Yogesh Agrawal. "K-Means Clustering on Cars Dataset using Seaborn Visualization". En: *kaggle* (2019). URL: https://www.kaggle.com/yugagrawal95/k-means-clustering-using-seaborn-visualization.

[2] Gemma Saura Barcelona. "¿Por qué Italia y España?" En: *La Vanguardia* (2020). URL: https://www.lavanguardia.com/internacional/20200405/48314311965/por-que-espana-italia.html.

[4] Fernando Berzal. "Clustering basado en particiones". En: *DECSAI, Universidad de Granada* (2015). URL: https://elvex.ugr.es/idbis/dm/slides/41\%20Clustering\%20-\%20Partitional.pdf.

[6] Anibal Bregón. "Clustering". En: *Universidad de Valladolid* (2019).

[7] Anibal Bregón. "Árboles y ensembles". En: *Universidad de Valladolid* (2019).

[10] Jefatura del Estado. "Ley Orgánica 3/2018, de 5 de diciembre de Protección de Datos Personales y garantía de los derechos digitales." En: *BOE* (2018). URL: https://www.boe.es/buscar/doc.php?id=BOE-A-2018-16673.

[11] Manuel Trigas Gallego. "Metodología Scrum". En: (2012). URL: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/17885/1/mtrigasTFC0612memoria.pdf.

[13] Andrés González. "Mejora de las predicciones con Ensembles (conjuntos de modelos)". En: *cleverdata* (2020). URL: https://cleverdata.io/mejorar-predicciones-ensembles-bigml/.

[14] J. S. Hicks M. Foster. "Adapting Scrum to Managing a Research Group". En: (2010a). URL: https://drum.lib.umd.edu/handle/1903/10743.

[15] J. S. Hicks M. Foster. "SCORE: agile research group management". En: *Communications of the ACM* (2010b). URL: https://dl.acm.org/doi/10.1145/1831407.1831421.

[17] National Center for Immunization y Division of Viral Diseases Respiratory Diseases (NCIRD). "Science Brief: Options to Reduce Quarantine for Contacts of Persons with SARS-CoV-2 Infection Using Symptom Monitoring and Diagnostic Testing". En: *Centers for Disease Control and Prevention* (2020). URL: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html.

[21] Equipo editorial de LabMedica. "Herramienta estadística predice los picos de COVID-19 en todo el mundo". En: *LabMedica.es* (2020). URL: https://www.labmedica.es/covid-19/articles/294782855/herramienta-estadistica-predice-los-picos-de-covid-19-en-todo-el-mundo.html.

[26] Daniele Palumbo y David Brown Lora Jones. "Coronavirus: 8 gráficos para entender cómo la pandemia ha afectado a las mayores economías del mundo". En: *BBC News* (2021). URL: https://www.bbc.com/mundo/noticias-55802814.

[27] A. Taufiq Asyhari Fadi Al-Turjman Md.Zakirul Alam Bhuiyan M.F.Zolkipli Md.Arafatur Rahmanan Nafees Zamanb. "Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices". En: *ScienceDirect* (2020). URL: https://www.sciencedirect.com/science/article/abs/pii/S221067072030593X.

[28] Ricardo Moya. "Selección del número óptimo de Clusters". En: *Jarroba* (2016). URL: https://jarroba.com/seleccion-del-numero-optimo-clusters/.

[29] Eric Niiler. "An AI Epidemiologist Sent the First Warnings of the Wuhan Virus". En: *WIRED* (2020). URL: https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings/.

[30] Kira Paulin. "Clustering". En: *SlidePlayer* (2015). URL: https://slideplayer.com/slide/2735334/.

[31] Relaciones con las Cortes y Memoria Democrática Ministerio de la Presidencia. "Real Decreto 926/2020, de 25 de octubre por el que se declara el estado de alarma para contener la propagación de infecciones causadas por el SARS-CoV-2." En: *BOE* (2020). URL: https://www.boe.es/buscar/doc.php?id=BOE-A-2020-12898.

[33] Joseph Rocca. "Ensemble methods: bagging, boosting and stacking". En: *towards data science* (2019). URL: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205.

[34] Joaquín Amat Rodrigo. "Gradient Boosting con Python". En: *cienciadedatos* (2020). URL: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html.

[35] Motaz Saad. "Stacking vs Bagging vs Boosting". En: *Motaz Saad* (2019). URL: https://mksaad.wordpress.com/2019/12/21/stacking-vs-bagging-vs-boosting/.

[37] Imperial College COVID-19 Response Team. "Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand". En: *Imperial College* (2020). URL: https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf.

Webgrafía

[3] Hospital Clínic Barcelona. *¿Qué es el Coronavirus SARS-CoV-2?* URL: https://www.clinicbarcelona.org/asistencia/enfermedades/covid-19/definicion (visitado 13-04-2021).

[5] EPICALSOFT INSTANCE BLOG. *[Azure Machine Learning] La dicotomía Varianza-sesgo (Bias-Variance).* URL: http://epicalsoft.blogspot.com/2019/02/azure-machine-learning-la-dicotomia.html (visitado 13-04-2021).

[8] Instituto de salud Carlos III. *Coronavirus: términos epidemiológicos más utilizados.* URL: https://cutt.ly/PzTojdY (visitado 03-11-2020).

[9] Centros para el Control y la Prevención de Enfermedades. *Cómo se propaga el COVID-19.* URL: https://espanol.cdc.gov/coronavirus/2019-ncov/transmission/index.html (visitado 13-04-2021).

[12] glassdoor. *Sueldos para Analista Junior.* URL: https://www.glassdoor.es/Sueldos/analista-junior-sueldo-SRCH_KO0,15.htm (visitado 08-02-2021).

[16] IArtificial.net. *Ensembles: voting, bagging, boosting, stacking.* URL: https://bit.ly/2QwBx9C (visitado 11-04-2021).

[18] INE. *Aportación del turismo a la economía española.* URL: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736169169&menu=ultiDatos&idp=1254735576863#:~:text=El\%20peso\%20del\%20turismo\%20alcanz\%C3\%B3,\%2C9\%25\%20del\%20empleo\%20total. (visitado 01-02-2021).

[19] INE. *Sección prensa de PIB en España.* URL: https://www.ine.es/prensa/pib_prensa.htm (visitado 01-02-2021).

[20] Ingur. *Clustering example.* URL: https://i.stack.imgur.com/cIDB3.png (visitado 20-02-2021).

[22] Scikit learn. *HistGradientBoostingClassifier.* URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html (visitado 04-01-2021).

[23] Junta de Castilla y León. *Análisis de datos abiertos JCyL.* URL: https://analisis.datosabiertos.jcyl.es/explore/?sort=modified (visitado 01-10-2020).

[24] Junta de Castilla y León. *Población de Tarjeta Sanitaria(TSI).* URL: https://www.saludcastillayleon.es/transparencia/es/transparencia/informacion-datos-publicos/datos-interes/poblacion-tsi (visitado 03-10-2020).

[25] Junta de Castilla y León. *Portal Boletín Oficial de Castilla y León(BOCYL).* URL: https://bocyl.jcyl.es/ (visitado 16-10-2020).

[32] Reuters. *World coronavirus tracker and maps.* URL: https://graphics.reuters.com/world-coronavirus-tracker-and-maps/es/ (visitado 15-02-2021).

[36] Synapptica. *Dar los primeros pasos en SCRUM.* URL: https://synapptica.net/metodologia-scrum.html (visitado 17-02-2021).

[38] Ministerio de transportes movilidad y agenda urbana. *Estudio de movilidad con Big Data.* URL: https://www.mitma.es/ministerio/covid-19/evolucion-movilidad-big-data (visitado 14-12-2020).

[39] Ministerio de transportes movilidad y agenda urbana. *Open Data Movilidad.* URL: https://www.mitma.es/ministerio/covid-19/evolucion-movilidad-big-data/opendata-movilidad (visitado 14-12-2020).

[40] Wikipedia. *Boosting.* URL: https://bit.ly/3xeXACi (visitado 14-04-2021).

Siglas

BOCYL Boletín Oficial de Castilla y León. 47

CA Comunidad Autónoma. 7, 165

CCAA Comunidades Autónomas. 7, 149

COVID-19 Coronavirus disease 19. 3

GBT Gradient Boosting Tree. 137, 145

IA Incidencia Acumulada. 34, 81

PCR Polymerase chain reaction. 35

PT Punto de Tarea. 20

RMSE Root Mean Squared Error. 148, 149

SARS-CoV-2 Severe Acute Respiratory Sindrome - CoronaVirus-2. 3

SCORE SCRUM for Research. 11

Parte II

Apéndices

Apéndice A

Contenido adjunto

Junto a la memoria del proyecto desarrollado se adjuntarán los siguientes directorios:

- **DATASETS**: Carpeta con todos los datasets usados en el proyecto. Se incluyen los modelos XGBT de todas las zonas estudiadas (entrenamiento y test).
- **IMPLEMENTACION:**Carpeta que contendrá todos los ficheros o notebooks con el formato .ipynb, en los cuales se encontrará todo el código correspondiente a los métodos y modelos de aprendizaje usados en el proyecto (ya ajustados).
- **IMAGENES-ALTA-DEFINICION:**Carpeta con imágenes en alta definición, correspondientes a las gráficas mostradas tanto en el Capitulo 6 como en el Capitulo 9, cuya apreciación puede hacerse compleja para el lector.

Estos directorios se compartirán y almacenarán a través del repositorio en Onedrive proporcionado por la Escuela de Ingeniería Informática de Segovia.

A continuación se mostrará la estructura de cada uno de estos directorios:

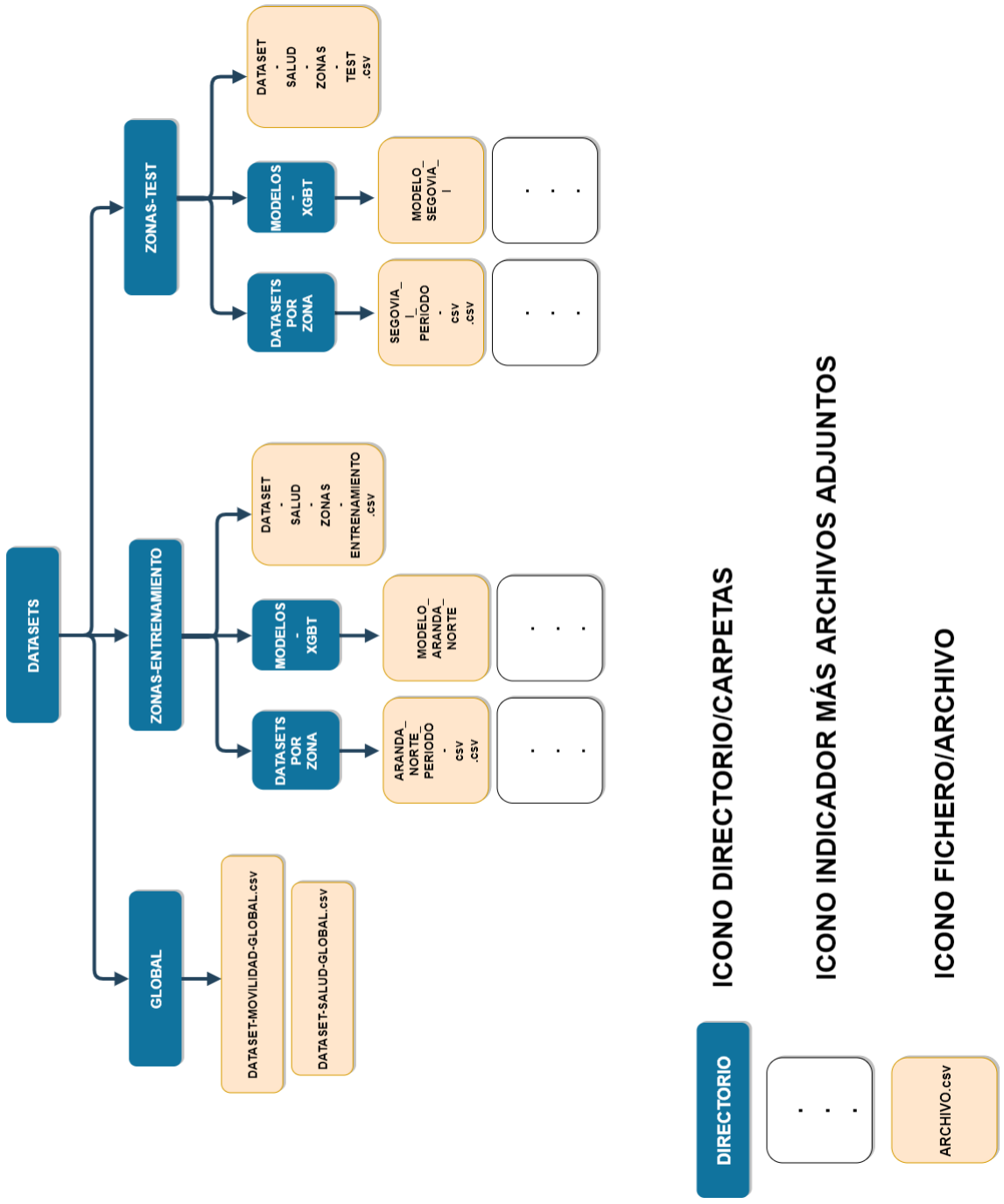


Figura A.1: Diagrama de árbol de carpeta DATASETS

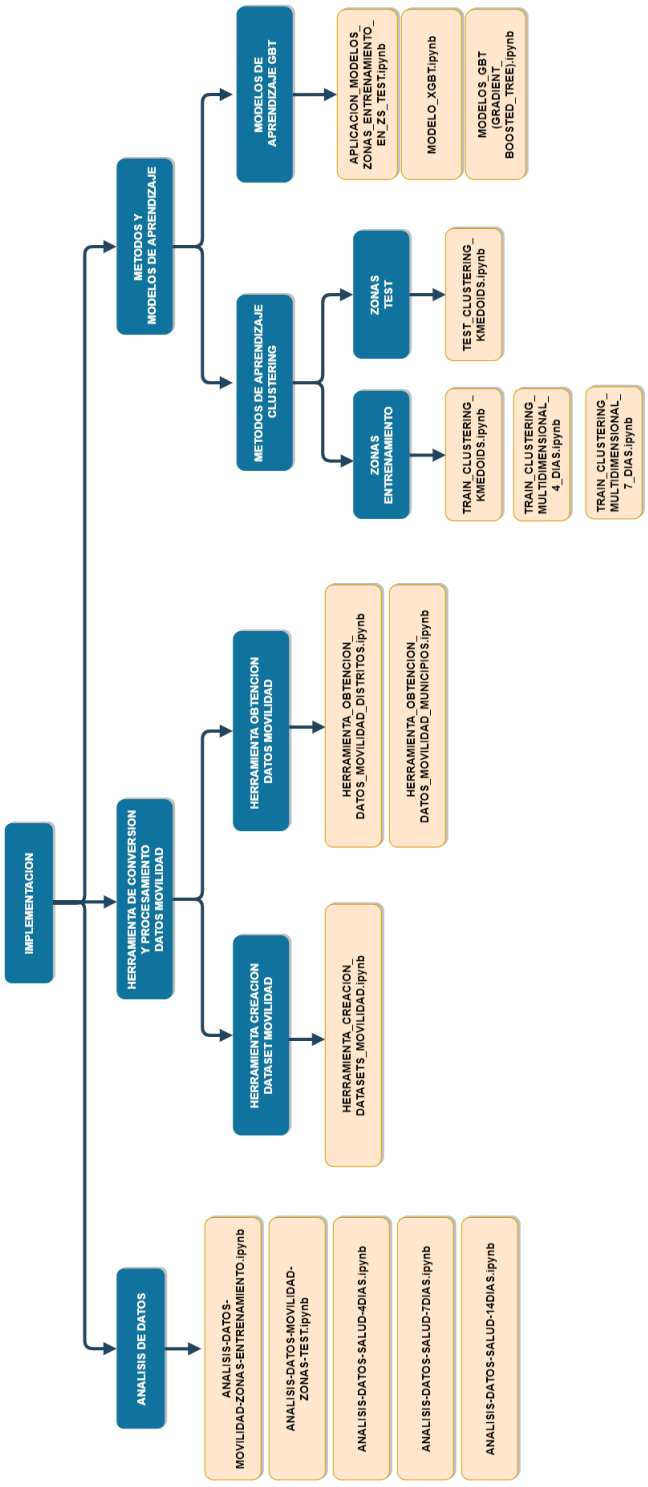


Figura A.2: Diagrama de árbol de carpeta IMPLEMENTACION

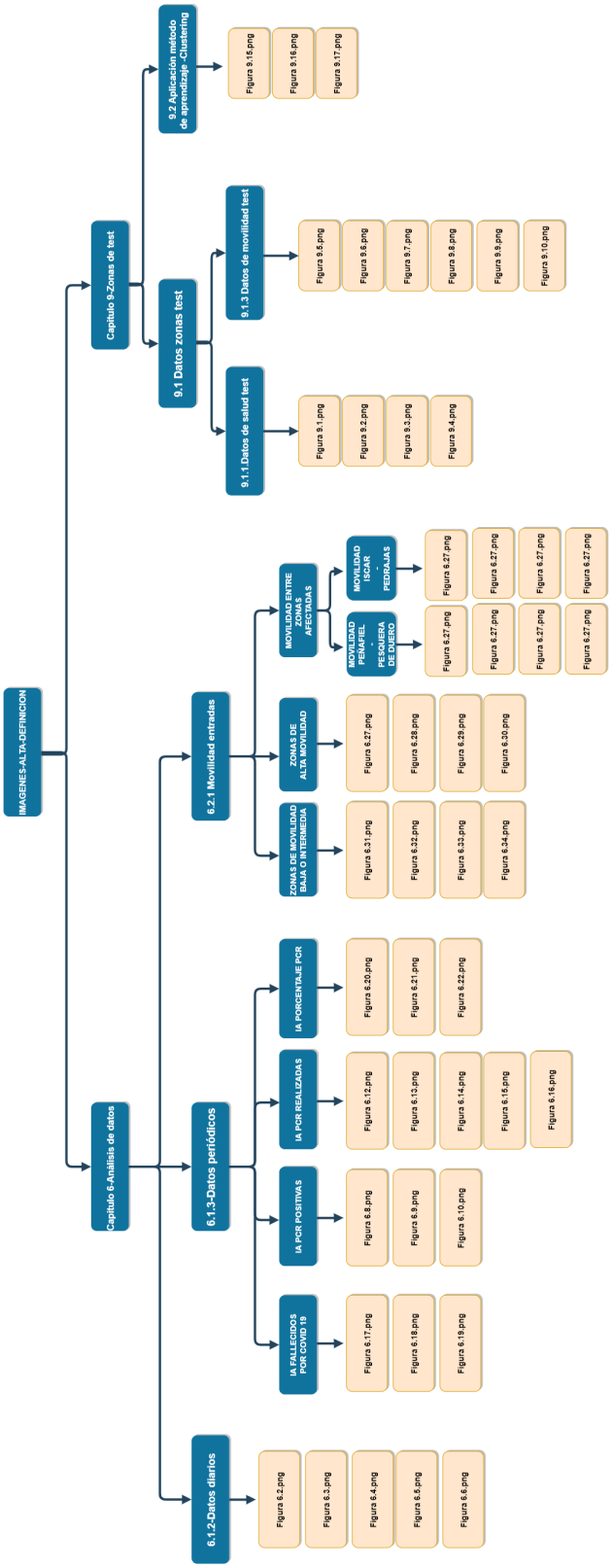


Figura A.3: Diagrama de árbol de carpeta IMAGENES-ALTA-DEFINICION