UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
GRADO DE INGENIERÍA INFORMATICA

Aplicación de la Inteligencia Artificial Geoespacial a la epidemiología medio ambiental

Application of Geospatial Artificial Intelligence to environmental epidemiology

Realizado por
Martín Ignacio Mazzola Ortega

Tutorizado por
Francisco López Valverde

Departamento
Lenguajes y Ciencias de la Computación

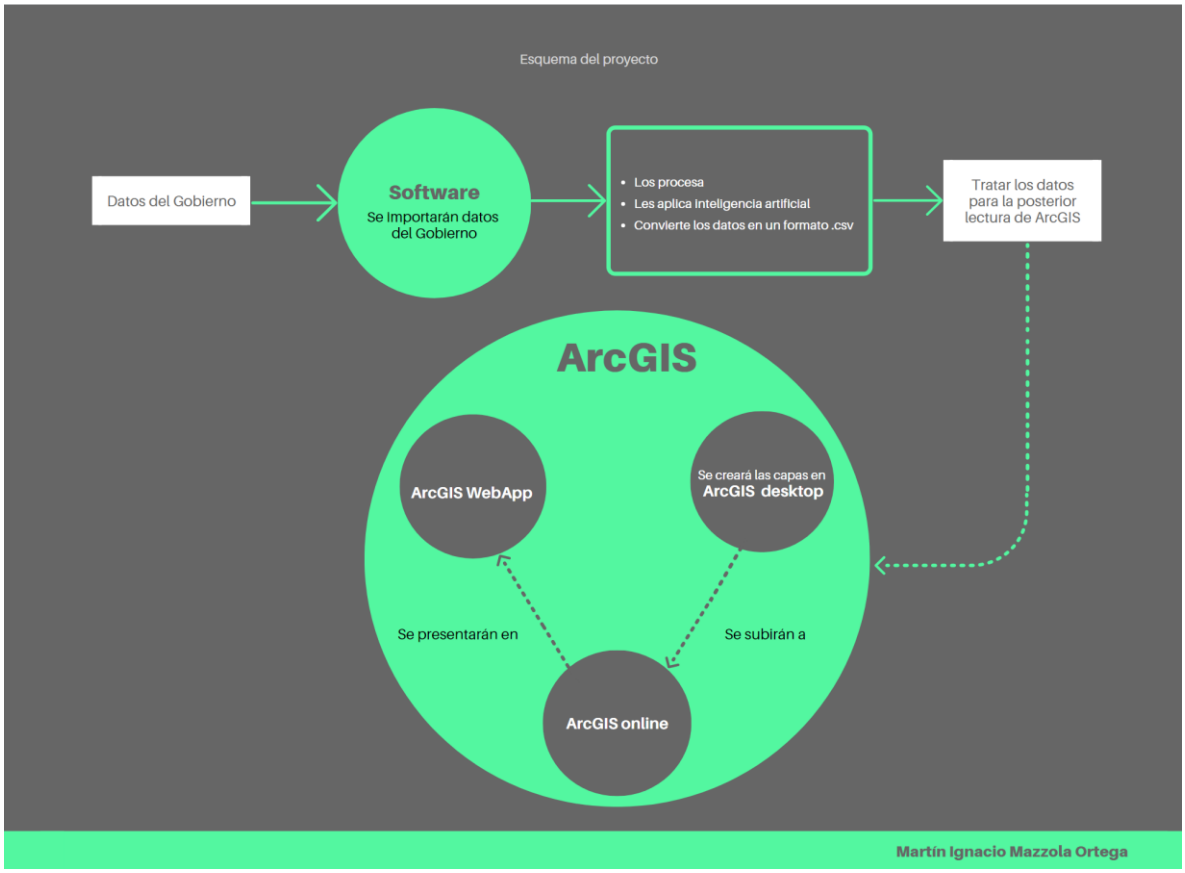
UNIVERSIDAD DE MÁLAGA

MÁLAGA, DICIEMBRE DE 2020

Fecha defensa: 11 de diciembre de 2020

Resumen

En este trabajo se ha buscado el análisis de los datos mediante la Inteligencia artificial, para aplicarlo a la epidemiología medio Ambiental. En este trabajo, se ha utilizado un programa desarrollado en Python (que utiliza librerías de Pytorch para poder aplicar inteligencia artificial a los datos), que recoge datos de un repositorio con datos oficiales, se los descarga los procesa y los deja en el formato necesario para poder subirlo a una plataforma de visualización y tratamiento de mapas, como es ArcGIS. Una vez tenemos los datos procesados en ArcGIS los exponemos en una aplicación web que es facilitada por la propia empresa, para poder realizar estadísticas, gráficos y búsquedas de una forma mucho más visual. Para que así un exporto pueda hacer un uso adecuado de dicha información.



Ref. 44

Palabras clave: COVID-19, ArcGIS, Pytorch, IA.

Índice

1

Introducción

1.1 Motivación

Desde que ha empezado esta pandemia, he buscado en mil páginas la respuesta a mi pregunta, y he leído todo tipo de cosas. La única certeza que encontré fue que en realidad nadie está seguro y los gobiernos ponen medidas que creen que son correctas dados los indicios de anteriores pandemias, de ahí viene la verdadera razón, de este trabajo. Quiero aportar mi grano de arena haciendo uso de mis conocimientos informáticos, en el estudio de la propagación del COVID-19. La idea de poder usar la inteligencia artificial para el estudio de epidemias pasadas, presentes y futuras me parece una de las funcionalidades más importantes que le podemos dar a la tecnología y así poder volver a vivir una situación como la que estamos viviendo hoy en día.

1.2 Objetivos

El objetivo de este TFG es la creación de una aplicación de inteligencia artificial que, integrará un motor de inteligencia artificial para el análisis de datos geoespaciales que utilice métodos de geo AI para abordar problemas relacionados con la salud humana (en nuestro caso se estudiará el COVID-19).

Se construirá un modelo para integrar el geo AI con la epidemiología ambiental para llevar a cabo una modelización más precisa y altamente resuelta de las exposiciones ambientales, lo que a su vez conduciría a una evaluación más precisa de los factores ambientales a los que estamos expuestos y, por tanto, a una mejor comprensión de las posibles asociaciones entre las exposiciones ambientales y la enfermedad en la epidemiología.

Se van a desarrollar bases de datos, con mapas interactivos, gracias a la aplicación ArcGIS.

Por otro lado, se diseñará una interfaz web para usuario el usuario que use nuestro motor. Dicha interfaz pretende facilitar, agilizar y hacer agradable (fácil de usar) el uso de nuestro software de manera que el operador necesita proporcionar una entrada mínima para conseguir la salida deseada, y también que el programa minimice las salidas no deseadas para el ser humano.

Para resumir, nuestro objetivo es crear una tabla en Excel que podamos usar en ArcGIS, que nos proporcione un informe del crecimiento de casos que se prevé hasta el siguiente informe. Los datos se descargarán desde el repositorio de Carlos III y posteriormente serán tratados mediante inteligencia artificial usando Pytorch.

1.4 Desarrollo

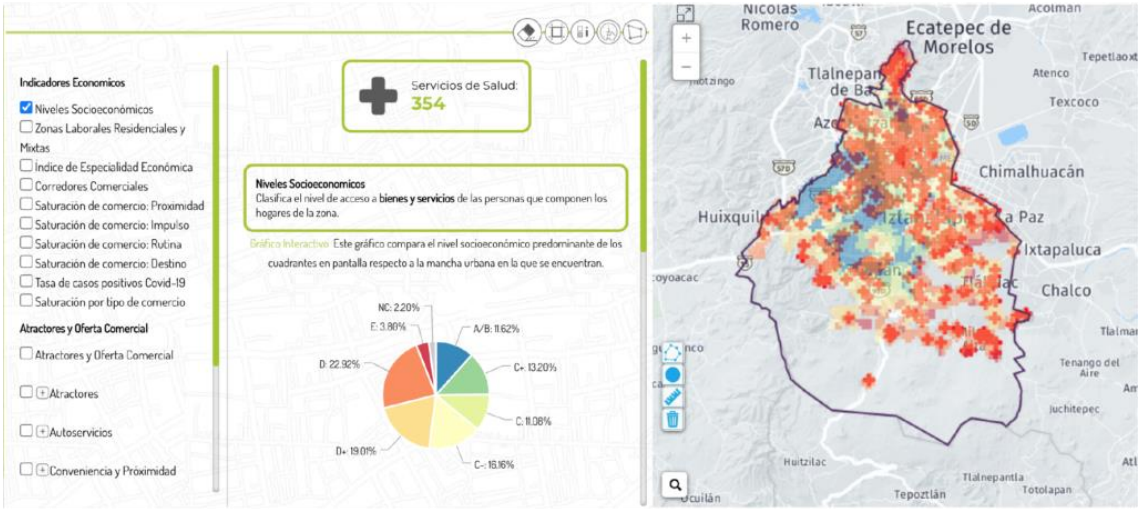
Normalmente, un sistema de IA es capaz de analizar datos en grandes cantidades (big data), identificar patrones y tendencias y, por lo tanto, formular predicciones de forma automática, con rapidez y precisión. Para nosotros, lo importante es que la IA permite que nuestras experiencias cotidianas sean más inteligentes. ¿Cómo? Al integrar análisis predictivos (hablaremos sobre esto más adelante) y otras técnicas de IA en aplicaciones que utilizamos diariamente.
La mayoría de nosotros tenemos un concepto de la Inteligencia artificial alimentado por las películas de Hollywood. Exterminadores, robots con crisis existenciales y píldoras rojas y azules. De hecho, la IA ha estado en nuestra imaginación y en nuestros laboratorios desde 1956, cuando un grupo de científicos inició el proyecto de investigación “Inteligencia artificial” en Dartmouth College en los Estados Unidos.
La Inteligencia artificial es el campo científico de la informática que se centra en la creación de programas y mecanismos que pueden mostrar comportamientos considerados inteligentes. En otras palabras, la IA es el concepto según el cual “las máquinas piensan como seres humanos”.
Una de las principales dudas que nos puede surgir es, ¿Por qué es la inteligencia artificial tan importante?

Para dar respuesta se describen las siguientes características:

- La inteligencia artificial automatiza el aprendizaje y descubrimiento repetitivos a través de datos.
- IA agrega inteligencia a productos existentes.
- La inteligencia artificial analiza más datos y datos más profundos utilizando redes neurales que tienen muchas capas ocultas.
- La inteligencia artificial logra una precisión increíble a través de redes neurales profundas – lo cual antes era imposible.
- La inteligencia artificial saca el mayor provecho de los datos.
- En este proyecto nos hemos centrado en este último punto de las características, ya que se va a realizar una investigación de la inteligencia artificial geoespacial. Lo que significa una enorme cantidad de datos y una red neuronal capaz de procesar esta información.

La inteligencia artificial geoespacial (geo AI) es una disciplina científica emergente que combina innovaciones en la ciencia espacial, métodos de inteligencia artificial en el

aprendizaje automático (por ejemplo, el famoso “Deep Learning”), la minería de datos y la computación de alto rendimiento para extraer conocimientos de los grandes datos espaciales. En la epidemiología ambiental, la modelización de la exposición es un método comúnmente utilizado para realizar evaluaciones de la exposición a fin de determinar la distribución de las exposiciones en las poblaciones estudiadas. Las tecnologías geo AI ofrecen importantes ventajas para la modelización de la exposición en la epidemiología ambiental, entre ellas la capacidad de incorporar grandes cantidades de datos espaciales y temporales de gran tamaño en diversos formatos; la eficiencia computacional; la flexibilidad en los algoritmos y flujos de trabajo para dar cabida a las características pertinentes de los procesos espaciales (ambientales), incluida la no estacionalidad espacial; y la posibilidad de ampliar la escala para modelizar otras exposiciones ambientales en diferentes zonas geográficas. Los objetivos de este comentario son proporcionar una visión general de los conceptos clave en torno al campo evolutivo e interdisciplinario del geo AI, incluida la ciencia de los datos espaciales, el aprendizaje automático, el aprendizaje profundo y la minería de datos; las aplicaciones recientes del geo AI en la investigación; y las posibles direcciones futuras del geo AI en la epidemiología ambiental.



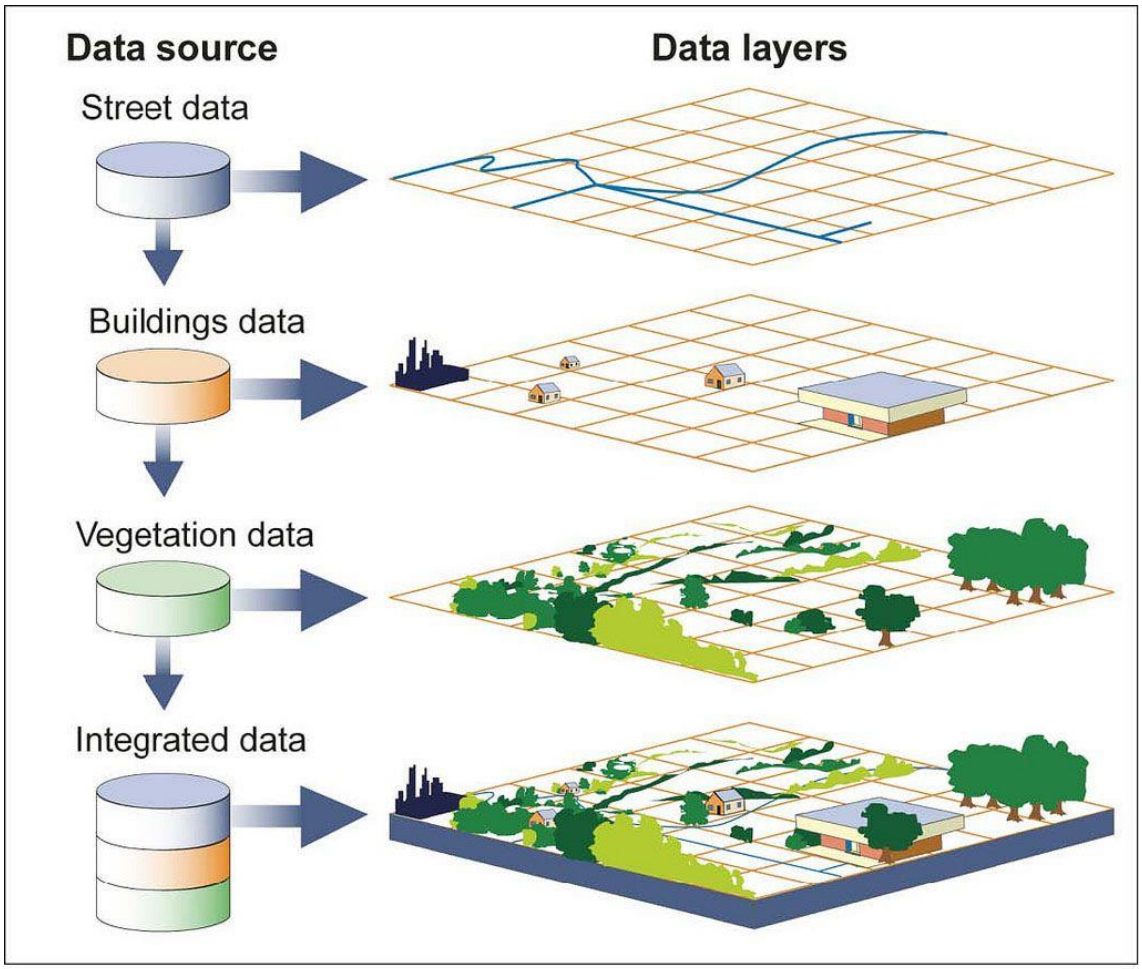
Ref. 48

Este campo, aún está en pleno desarrollo. Por eso cuando decidimos entrar en este mundo la información que se nos presenta es abrumadora. Tenemos muchísimos lenguajes con los que aprender ya sea C++, java, Python… Es cierto que también existe una pequeña predominancia de Python, pero con muchísimas variantes. También existen muchísimos cursos tanto gratuitos como de pagos para aprender a usarlos. Pero… ¿Cómo podemos saber cuál es el que mejor se ajusta a nuestro perfil?

Bueno, este es uno de los puntos que trataremos más adelante en este trabajo, ya que existen muchas variantes muy buenas y es muy difícil decantarse tanto por una como por otra.

Por otro lado, tenemos las tecnologías GIS. En sistema de información geográfica (GIS) es un sistema diseñado para capturar, almacenar, manipular, analizar, gestionar y presentar todo tipo de datos geográficos. La palabra clave de esta tecnología es "geografía", lo que significa que una parte de los datos son espaciales. En otras palabras, datos que de alguna manera se refieren a lugares de la tierra.

Junto con estos datos, normalmente hay datos tabulares conocidos como datos de atributo. Los datos de atributo pueden definirse generalmente como información adicional sobre cada una de las características espaciales. Un ejemplo de esto sería las escuelas. La ubicación real de las escuelas son los datos espaciales. Los datos adicionales como el nombre de la escuela, el nivel de educación impartido, la capacidad de los estudiantes constituiría los datos de atributos.
Es la asociación de estos dos tipos de datos lo que permite que el GIS sea una herramienta tan efectiva para la resolución de problemas a través del análisis espacial.
El GIS es más que un simple software. Las personas y los métodos se combinan con software y herramientas geoespaciales, para permitir el análisis espacial, gestionar grandes conjuntos de datos y mostrar la información en forma de mapa/gráfico.



Source: GAO.
Ref. 43

Como podemos observar en las imágenes (1) y (2) tenemos dos grandes recursos. La inteligencia artificial geoespacial y las tecnologías GIS. Si nos paramos a mirar, enseguida nos daremos cuenta de que estamos ante dos campos que pueden llegar a ir unidos de la mano, ya que, si conseguimos que la inteligencia artificial geoespacial tenga una visualización clara e intuitiva, estamos haciendo de esto, una herramienta para la ayuda de toma de decisiones, que pueden llegar a niveles gubernamentales con el correcto equipo de expertos

2

Background

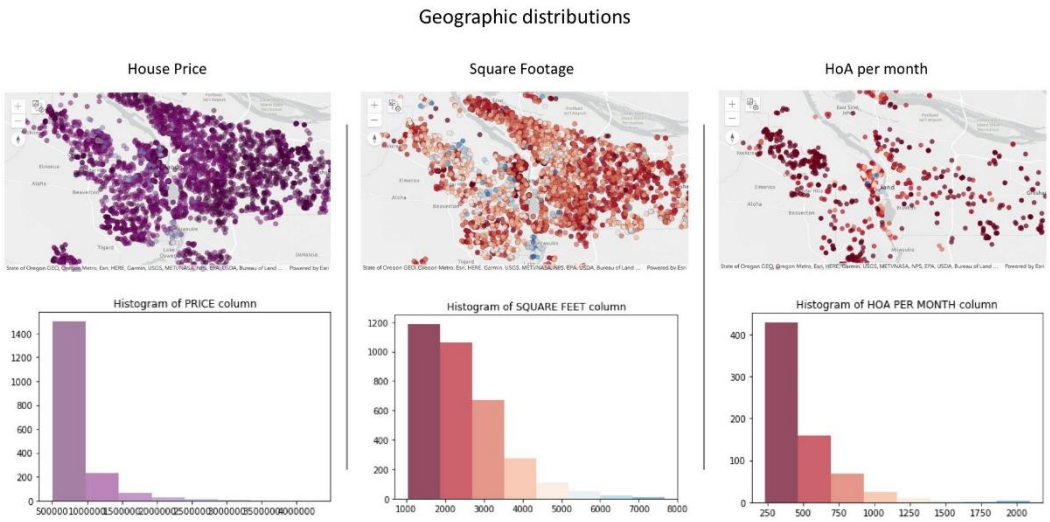
2.1 Inteligencia Artificial Geoespacial

La ciencia espacial, también conocida como ciencia de la información geográfica, juega un papel importante en muchas disciplinas científicas ya que busca entender, analizar y visualizar los fenómenos del mundo real de acuerdo con su ubicación. Los científicos espaciales aplican tecnologías como los sistemas de información geográfica (SIG) y la teleobservación a los datos espaciales (por ejemplo, georreferenciados) para lograr esos objetivos, es decir, para identificar y dar sentido a las pautas en el espacio. Vinculada a la actual era de los grandes datos está la generación en tiempo real de grandes datos espaciales, que han pasado a estar disponibles de forma ubicua desde los mensajes de los medios sociales geoetiquetados en Twitter hasta los sensores ambientales que recogen información meteorológica [1]. Se ha sugerido que al menos el 80% de todos los datos son de naturaleza geográfica, ya que la mayoría de la información que nos rodea puede ser georreferenciada [1]. En esta medida, el 80% de los 2,5 exabytes (2.500.000.000 gigabytes) de grandes datos generados diariamente es geográfico [2]. La ciencia de los datos, y por extensión la ciencia de los datos espaciales, son todavía campos en evolución que proporcionan métodos para organizar cómo pensamos y nos acercamos a la generación de nuevos conocimientos a partir de grandes datos (espaciales).

El campo científico de la inteligencia artificial geoespacial (geoAI) se formó recientemente a partir de la combinación de innovaciones en la ciencia espacial con el rápido crecimiento de los métodos de la inteligencia artificial (IA), en particular el aprendizaje automático (por ejemplo, el aprendizaje profundo), la minería de datos y la informática de alto rendimiento para obtener información significativa a partir de grandes datos espaciales. geoAI es altamente interdisciplinario, y constituye un puente entre muchos campos científicos, entre ellos la informática, la ingeniería, la estadística

y la ciencia espacial. La innovación del geoAI radica en parte en sus aplicaciones para abordar problemas del mundo real. En particular, las aplicaciones de geoAI se presentaron en el Taller Internacional sobre GeoAI: IA y Aprendizaje Profundo para el descubrimiento de conocimientos geográficos (el comité directivo estuvo dirigido por el Instituto de Dinámica Urbana del Laboratorio Nacional de Oak Ridge del Departamento de Energía de los Estados Unidos), que incluyó avances en la clasificación de imágenes de teleobservación y en el modelado predictivo del tráfico. Además, la aplicación de las tecnologías de la IA para el descubrimiento de conocimientos a partir de datos espaciales refleja una tendencia reciente, como se ha demostrado en otras comunidades científicas, incluido el Simposio Internacional sobre Bases de Datos Espaciales y Temporales.

Estos novedosos métodos geoAI pueden utilizarse para abordar problemas relacionados con la salud humana, por ejemplo, en la epidemiología ambiental [3]. En particular, las tecnologías geoAI están empezando a utilizarse en el campo de la modelización de la exposición ambiental, que se utiliza comúnmente para llevar a cabo la evaluación de la exposición en estos estudios [4]. En última instancia, uno de los objetivos generales para integrar la geoAI con la epidemiología ambiental es llevar a cabo un modelado más preciso y de mayor resolución de las exposiciones ambientales (en comparación con los enfoques convencionales), lo que a su vez conduciría a una evaluación más precisa de los factores ambientales a los que estamos expuestos y, por lo tanto, a una mejor comprensión de las posibles asociaciones entre las exposiciones ambientales y la enfermedad en los estudios epidemiológicos. Además, el geoAI proporciona métodos para medir nuevas exposiciones que antes eran difíciles de captar.



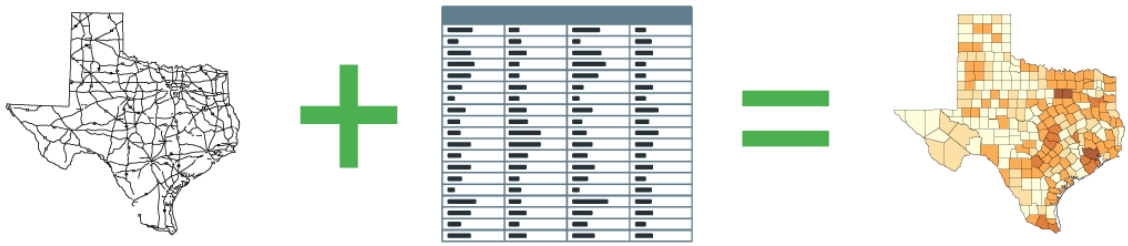
Ref. 52

Estos novedosos métodos geo AI pueden utilizarse para abordar problemas relacionados con la salud humana, por ejemplo, en la epidemiología ambiental. En particular, las tecnologías geo AI están empezando a utilizarse en el campo de la modelización de la exposición ambiental, que se utiliza comúnmente para llevar a cabo la evaluación de la exposición en estos estudios. En última instancia, uno de los objetivos generales para integrar la geo AI con la epidemiología ambiental es llevar a cabo un

modelado más preciso y de mayor resolución de las exposiciones ambientales (en comparación con los enfoques convencionales), lo que a su vez conduciría a una evaluación más precisa de los factores ambientales a los que estamos expuestos y, por lo tanto, a una mejor comprensión de las posibles asociaciones entre las exposiciones ambientales y la enfermedad en los estudios epidemiológicos. Además, el geo AI proporciona métodos para medir nuevas exposiciones que antes eran difíciles de captar.

2.2 Sistema de información geográfica (GIS)

GIS significa "Sistema de Información Geográfica". GIS es un término muy amplio, y tratar de obtener una definición consistente puede ser difícil. Puede que en algunas webs de habla castellana encuentren las siglas GIS, esto ocurre porque así son las siglas en castellano.
Se podría argumentar que cualquier dato digital que contenga información basada en la localización es de hecho un GIS. Esta información de localización en la industria de los GIS se llama "datos espaciales" y podría ser una dirección, unas coordenadas que contengan la latitud y la longitud o una compleja geometría tridimensional.
La verdad es que las herramientas GIS pueden hacer una gran variedad de cosas, pero aquí está la respuesta corta: Un GIS nos permite visualizar los datos que se recogen en una base de datos como un mapa.
Somos criaturas visuales que poseen una habilidad innata para visualizar patrones. Los patrones que pueden llevarnos horas identificar en una hoja de cálculo pueden ser a menudo identificados en un instante cuando se muestran en un formato más visualmente atractivo como un gráfico, un cuadro o, en este caso, un mapa.
Hay muchas formas innovadoras de que los datos pueden ser mostrados en un mapa. Puede ser trazando marcadores, codificando con colores las ubicaciones basadas en un valor de datos o usando mapas térmicos para identificar los cúmulos y patrones en sus datos, las posibilidades y las perspectivas potenciales son literalmente infinitas.
Los datos pueden ser visualizados de infinitas maneras, además, los sistemas GIS no son estáticos. Nos permiten hacer preguntas complejas, o "consultas" como se llaman en el lenguaje de los GIS, en cualquier momento que queramos. Un sistema GIS puede responder a estas preguntas instantáneamente modificando colores, formas o resaltando lugares en el mapa.



Ref. 54

La aplicación de un GIS es una tarea compleja y no debe subestimarse; sin embargo, los beneficios que se pueden obtener son importantes. Entre ellos figuran

- Una cartografía más clara
- Establecer conjuntos de datos geográficos de referencia
- Manipulación más fácil de los datos
- Almacenamiento de datos más conveniente
- Capacidad de establecer relaciones entre las variables geográficas
- Aumento de la capacidad para la adopción de decisiones
- Mayor capacidad de comunicación con los interesados

También hay una serie de limitaciones, o problemas potenciales, que pueden socavar la calidad de cualquier análisis de GIS. Entre ellas se incluyen:

- La calidad de los datos
- Disponibilidad de los datos
- El costo del hardware y el software
- El costo de los datos
- Nivel de formación del usuario

La implementación de un GIS para ayudar en la toma de decisiones debe llevarse a cabo cuidadosamente para asegurar que se mantenga la máxima calidad de los datos a lo largo del ciclo de decisión. Es importante tener en cuenta las siguientes cuestiones:
La complejidad de la realidad que se está mapeando
Consideraciones organizativas, incluidas las necesidades de personal y de capacitación

- Calidad de los datos
- Costo
- Datos de la fuente
- Tiempo Consideraciones específicas relacionadas con el medio ambiente costero

2.2.1 Distinción entre: lo espacial en los grandes datos y la ciencia de los datos

Varios conceptos clave están actualmente en la vanguardia de la comprensión de la gran revolución de los datos geoespaciales. Los grandes datos, como las historias clínicas electrónicas y las transacciones con los clientes, se caracterizan generalmente por un gran volumen de datos; una gran variedad de fuentes, formatos y estructuras de datos; y una gran velocidad de creación de nuevos datos [5,6,7]. En consecuencia, los grandes datos requieren métodos y técnicas especializadas para su procesamiento y análisis. La ciencia de los datos se refiere en general a los métodos para proporcionar nuevos conocimientos a partir del análisis riguroso de grandes datos, integrando métodos y conceptos de disciplinas como la informática, la ingeniería y la estadística [8, 9]. El flujo de trabajo de la ciencia de los datos generalmente se asemeja a un proceso iterativo de importación y procesamiento de datos, seguido por la limpieza, la transformación, la visualización, el modelado y, finalmente, la comunicación de los resultados [10].

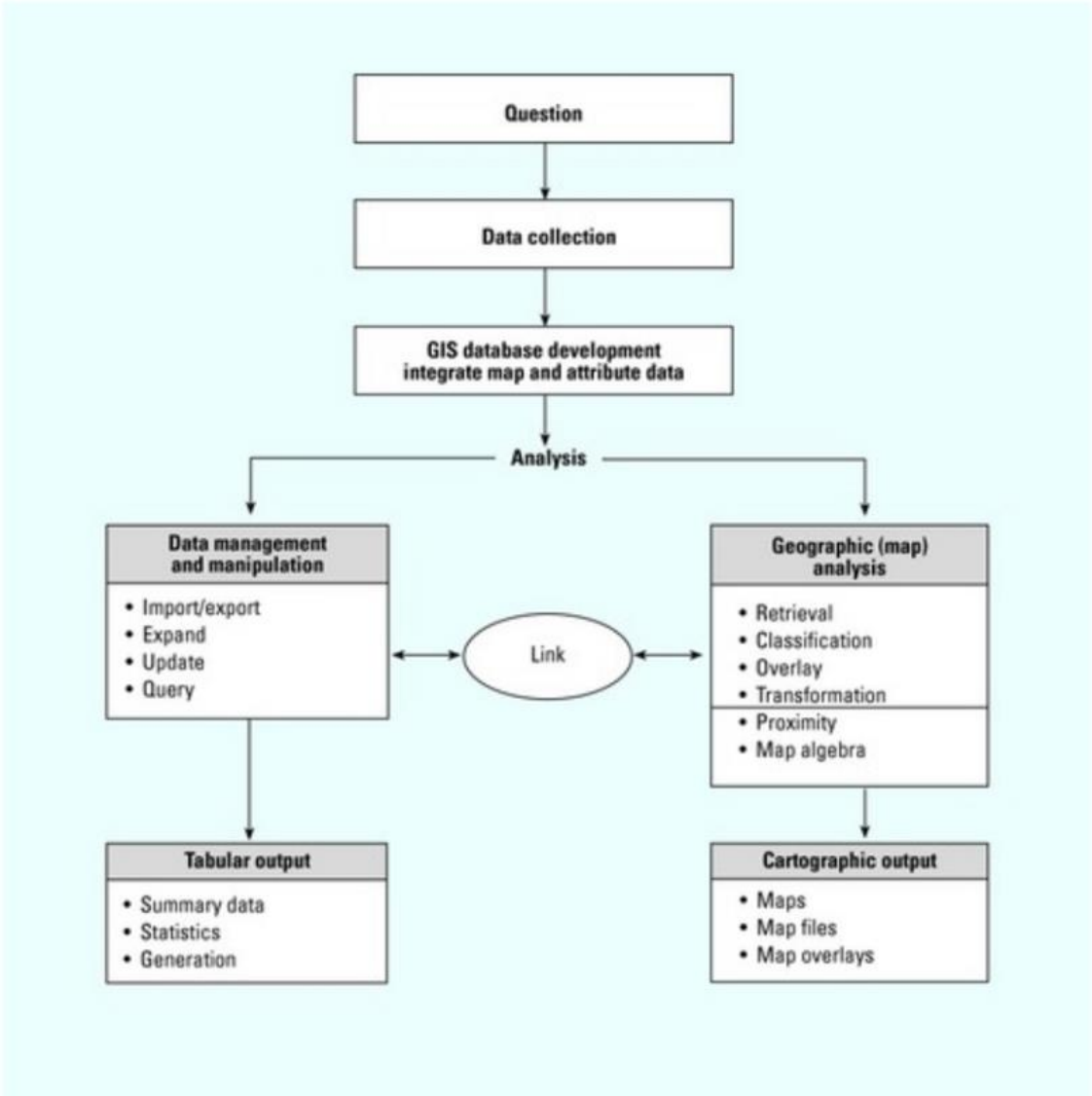
La ciencia de los datos espaciales es un nicho y un campo en formación que se centra en los métodos para procesar, gestionar, analizar y visualizar grandes datos espaciales, proporcionando oportunidades para derivar conocimientos dinámicos de fenómenos espaciales complejos [11]. Los flujos de trabajo de la ciencia de los datos espaciales se componen de pasos para la manipulación de datos, la integración de datos, el análisis exploratorio de datos, la visualización y el modelado, y se aplican específicamente a los datos espaciales, a menudo utilizando software especializado para formatos de datos espaciales [12].
Por ejemplo, un flujo de trabajo de la ciencia de los datos espaciales puede incluir la discusión de datos utilizando soluciones de código abierto como la Biblioteca de Abstracción de Datos Geoespaciales (GDAL), scripting en R, Python, y Spatial SQL para análisis espaciales facilitados por la computación de alto rendimiento (por ejemplo, la consulta de grandes datos almacenados en una infraestructura de datos distribuidos a través de plataformas de computación en nube como Amazon Web Services para el análisis; o análisis de grandes datos espaciales realizados en una supercomputadora), y la geo visualización utilizando 3D. La síntesis de datos espaciales se considera un reto importante en la ciencia de los datos espaciales, que incluye cuestiones relacionadas con la agregación de datos espaciales (de diferentes escalas) y la integración de datos espaciales (armonización de diversos tipos de datos espaciales relacionados con el formato, la referencia, la unidad, etc.) [11]. Los avances en el ciber gigante (definido como el SIG basado en la ciber infraestructura avanzada y la ciber ciencia) -y, más en general, las capacidades informáticas de alto rendimiento para datos de gran dimensión- han desempeñado un papel integral en la transformación de nuestra capacidad para manejar grandes datos espaciales y, por lo tanto, para las aplicaciones de la ciencia de los datos espaciales. Por ejemplo, en 2014 se creó una supercomputadora cibernética de SIG apoyada por la Fundación Nacional de la Ciencia llamada ROGER, que permite la ejecución de aplicaciones geoespaciales que requieren una ciber infraestructura avanzada a través de la computación de alto rendimiento (por ejemplo, > 4 petabytes de almacenamiento persistente de alta velocidad), la computación acelerada por unidad de procesamiento gráfico (GPU), los grandes subsistemas de datos intensivos que utilizan Hadoop y Spark, y la computación en nube Openstack [11, 13].

A medida que la ciencia de los datos espaciales continúa evolucionando como disciplina, los grandes datos espaciales se expanden constantemente, con dos ejemplos destacados que son la información geográfica voluntaria (VGI) y la teledetección. El término IGV encapsula el contenido generado por el usuario con un componente de localización [14]. En el último decenio, la IGV ha experimentado una explosión con el advenimiento y la continua expansión de los medios sociales y los teléfonos inteligentes, en los que los usuarios pueden publicar y, por tanto, crear tweets geoetiquetados en Twitter, fotos de Instagram, vídeos de Snapchat y reseñas de Yelp [15]. El uso del IGV debe ir acompañado de una conciencia de las posibles cuestiones jurídicas, incluidas, entre otras, la propiedad intelectual, la responsabilidad y la privacidad del operador, el colaborador y el usuario del IGV [16]. La teleobservación es otro tipo de grandes datos espaciales que captan las características de los objetos a distancia, como las imágenes de los sensores de los satélites [17]. Según el sensor, los grandes datos espaciales de la teleobservación pueden ser extensos tanto en su cobertura geográfica (que abarca todo el globo) como en su cobertura temporal (con frecuentes tiempos de revisión). En los

últimos años, hemos visto un enorme aumento de los grandes datos de teledetección por satélite a medida que las empresas privadas y los gobiernos siguen lanzando satélites de mayor resolución. Por ejemplo, DigitalGlobe recoge más de 1.000 millones de km2 de imágenes de alta resolución cada año como parte de su constelación de satélites comerciales, incluidas las naves espaciales WorldView y GeoEye [18]. El Servicio Geológico de los Estados Unidos y el programa Landsat de la NASA han lanzado continuamente satélites de observación de la Tierra desde 1972, con resoluciones espaciales tan finas como 15 m y una resolución espectral cada vez mayor con cada misión subsiguiente de Landsat (por ejemplo, la cámara terrestre operativa Landsat 8 y el sensor térmico infrarrojo lanzados en 2013 constan de 9 bandas espectrales y 2 bandas térmicas) [19].

2.2.2 ¿Cómo funciona el GIS?

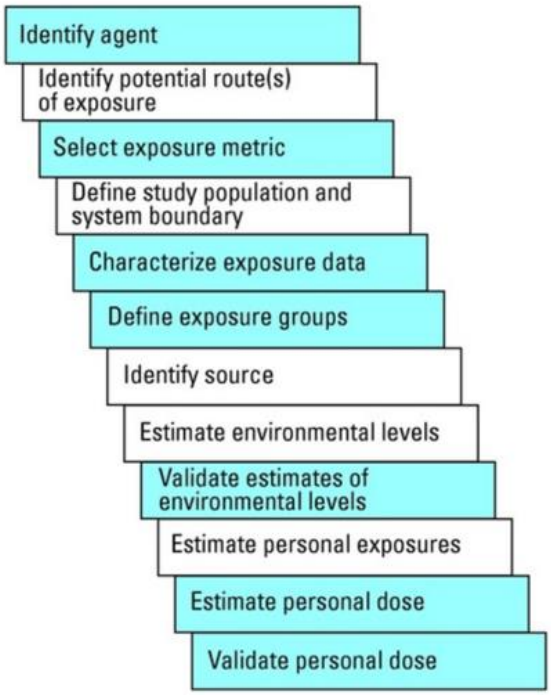
Así pues, el estudio de las interacciones entre los seres humanos y su medio ambiente requiere información y análisis espaciales. El software del sistema de información geográfica (GIS) permite almacenar, analizar y mostrar espacialmente los datos ambientales y epidemiológicos. La recolección de datos se puede lograr importando datos tabulares o digitales que son referenciados con coordenadas de mapa que definen su posición geográfica. Los datos se introducen en una base de datos donde se almacenan como un mapa con un tema específico (denominado "capa de datos"). Con cada capa de datos se pueden almacenar datos tabulares (de atributos) correspondientes al tema. Las funciones analíticas del software pueden utilizarse para procesar y manipular tanto los datos del mapa como los de los atributos mediante los vínculos establecidos en el GIS. Son comunes dos tipos de resultados: tabulares (datos de resumen, estadísticas, informes) y cartográficos (mapas, archivos de mapas y superposiciones de mapas).



Ref. 39

Los GIS se han utilizado en diferentes niveles de sofisticación en los estudios de epidemiología ambiental. Esos usos van desde la simple localización de la población de estudio mediante la geo codificación de direcciones (asignación de coordenadas cartográficas) hasta el uso de la proximidad a la fuente de contaminantes como sustituto de la exposición, pasando por la integración de los datos de vigilancia ambiental en el análisis de los resultados sanitarios. Sin embargo, la mayoría de estos últimos estudios han tenido un diseño ecológico; relativamente pocos estudios han utilizado los GIS para estimar los niveles ambientales de un contaminante a nivel individual. Aunque todavía no se ha aplicado en el contexto de un análisis epidemiológico, varios estudios han investigado el uso de los GIS para estimar los patrones de actividad de la población estudiada para su posible vinculación con los datos ambientales a fin de perfeccionar las estimaciones de exposición personal. Del mismo modo, el uso de los GIS en las estadísticas espaciales para vincular los datos de exposición y salud en el contexto del análisis epidemiológico es un campo de investigación cada vez más amplio. En este trabajo se examinan también los fundamentos de las disciplinas científicas necesarias para utilizar los GIS en la evaluación de la exposición en los estudios epidemiológicos y se explora la forma en que un GIS puede utilizarse para realizar varios pasos en el

proceso de evaluación de la exposición (los que aparecen sombreados en azul en la figura 2). Específicamente estos pasos son a) definir la población de estudio, b) identificar la fuente y las rutas potenciales de exposición, c) estimar los niveles ambientales de los contaminantes objetivo y destimar las exposiciones personales.



Ref. 39

2.2.3 Oportunidades para el geo AI en la epidemiología ambiental

Dados los avances y las capacidades que se han puesto de manifiesto en las investigaciones recientes, podemos empezar a conectar los puntos relativos a la forma en que las tecnologías geoAI pueden aplicarse específicamente a la epidemiología ambiental. Para determinar los factores a los que podemos estar expuestos y que, por lo tanto, pueden influir en la salud, los epidemiólogos ambientales aplican métodos directos de evaluación de la exposición, como la biomonitorización (por ejemplo, medida en orina), y métodos indirectos, como la modelización de la exposición. La modelización de la exposición supone la elaboración de un modelo para representar una variable ambiental concreta utilizando diversas entradas de datos (como las mediciones ambientales) y métodos estadísticos (como la regresión del uso de la tierra y los modelos mixtos aditivos generalizados) [23]. La modelización de la exposición es un enfoque eficaz en función de los costos para evaluar la distribución de las exposiciones en poblaciones de estudio particularmente grandes en comparación con la aplicación de métodos directos [23]. Los modelos de exposición incluyen medidas básicas basadas en

la proximidad (por ejemplo, topes y distancia medida) a modelos más avanzados como el kriging [3]. La ciencia espacial ha sido fundamental en la elaboración de modelos de exposición para los estudios epidemiológicos durante los dos últimos decenios, lo que ha permitido a los epidemiólogos ambientales utilizar las tecnologías de los SIG para crear y vincular modelos de exposición con datos sobre los resultados de la salud utilizando variables geográficas (por ejemplo, direcciones geo codificadas) para investigar los efectos de factores como la contaminación atmosférica en el riesgo de desarrollar enfermedades como las cardiovasculares [22, 24].

Los métodos geoAI y las grandes infraestructuras de datos (por ejemplo, Spark y Hadoop) pueden aplicarse para hacer frente a los problemas que rodean a la modelización de la exposición en la epidemiología ambiental -incluida la ineficiencia en el procesamiento computacional y el tiempo (en particular cuando se combinan grandes datos con grandes zonas de estudio geográfico) y las limitaciones relacionadas con los datos que afectan a la resolución espacial y/o temporal. Por ejemplo, los esfuerzos anteriores de modelización de la exposición se han asociado a menudo con resoluciones espaciales gruesas, lo que afecta a la medida en que el modelo de exposición es capaz de estimar con precisión la exposición a nivel individual (es decir, el error de medición de la exposición), así como las limitaciones en la resolución temporal que pueden dar lugar a que no se capturen las exposiciones durante las ventanas de tiempo pertinentes para el desarrollo de la enfermedad de interés [23]. Los avances en el geoAI permiten una modelización precisa y de alta resolución de la exposición para estudios epidemiológicos ambientales, especialmente en lo que respecta a la computación de alto rendimiento para manejar grandes datos (grandes en el espacio y el tiempo; espacio-temporales), así como el desarrollo y la aplicación de algoritmos de máquina y de aprendizaje profundo y grandes infraestructuras de datos para extraer las piezas más significativas y pertinentes de información de entrada para, por ejemplo, predecir la cantidad de un factor ambiental en un momento y lugar determinados.

Un ejemplo reciente de geoAI en acción para la evaluación de la exposición ambiental fue un método basado en datos desarrollado para predecir la contaminación del aire por materia particulada < 2,5 µm de diámetro (PM2,5) en Los Ángeles, CA, EE. UU. [4]. Esta investigación utilizó la investigación pediátrica utilizando la infraestructura del Centro de Coordinación e Integración de Datos y Software (DSCIC) de los Sistemas Integrados de Monitoreo de Sensores (PRISMS) [4, 31]. Se desarrolló un enfoque de minería de datos espaciales utilizando el aprendizaje por máquina y los grandes datos espaciales de OpenStreetMap (OSM) para permitir la selección de las características geográficas más importantes de OSM (por ejemplo, el uso de la tierra y las carreteras) que predicen las concentraciones de PM2,5. Este enfoque de extracción de datos espaciales aborda cuestiones importantes en la modelización de la exposición a la contaminación atmosférica en lo que respecta a la variabilidad espacial y temporal del "vecindario" pertinente dentro del cual determinar cómo y qué factores influyen en las exposiciones previstas (la no estacionalidad espacial se examina más adelante). Utilizando millones de características geográficas disponibles en OSM, el algoritmo para crear el modelo de exposición a PM2.5 identificó primero las estaciones de monitoreo del aire de la Agencia de Protección Ambiental de los Estados Unidos (EPA) que exhibían patrones temporales similares en concentraciones de PM2.5. A continuación, el algoritmo entrenó un modelo de bosque aleatorio (un método popular de aprendizaje

de máquinas que utiliza árboles de decisión para la clasificación y el modelado de regresión) para generar la importancia relativa de cada característica geográfica de OSM. Esto se realizó determinando el geo-contexto, o qué características del OSM y dentro de qué distancias (por ejemplo, topes de radio de 100 m vs. 1000 m) se asocian con las estaciones de vigilancia del aire (y sus niveles medidos de PM2,5) caracterizadas por un patrón temporal similar. Por último, el algoritmo entrenó un segundo modelo de bosque aleatorio utilizando los geo-contextos y midió las PM2,5 en las estaciones de vigilancia del aire para predecir las concentraciones de PM2,5 en lugares no medidos (es decir, la interpolación). Los errores de predicción se redujeron al mínimo mediante la incorporación de la temporalidad de las concentraciones de PM2,5 medidas en cada etapa del algoritmo, aunque la modelización se habría mejorado con información variable en el tiempo sobre los predictores.

Por último, el algoritmo entrenó un segundo modelo de bosque aleatorio utilizando los geo-contextos y midió las PM2,5 en las estaciones de vigilancia del aire para predecir las concentraciones de PM2,5 en lugares no medidos (es decir, la interpolación). Los errores de predicción se redujeron al mínimo mediante la incorporación de la temporalidad de las concentraciones de PM2,5 medidas en cada etapa del algoritmo, aunque la modelización se habría mejorado con información variable en el tiempo sobre los predictores. El rendimiento predictivo del modelo utilizando niveles medidos de PM2,5 en las estaciones de vigilancia del aire de la EPA como patrón de oro mostró una mejora en comparación con el uso de la ponderación de distancia inversa, un método de interpolación espacial comúnmente utilizado [4].
Mediante este innovador enfoque, se desarrolla un algoritmo flexible basado en la minería de datos espaciales que elimina la necesidad de seleccionar a priori los predictores para la modelización de la exposición, ya que los predictores importantes pueden depender de la zona de estudio específica y de la hora del día, dejando esencialmente que los datos decidan lo que es importante para la modelización de la exposición [4].

2.3 Epidemiología y SARS-CoV-2

2.3.1 Información sobre el SARS-COV-2

El 2019-nCoV, (ahora rebautizado como SARS-CoV-2 desde que descubrimos que pertenece a la misma familia que el SARS) es un tipo de coronavirus detectado por primera vez en humanos en diciembre de 2019 en la localidad china de Wuhan; es también conocido simplemente como 'nuevo coronavirus'.
La enfermedad respiratoria de este coronavirus, etiquetada a partir del 11 de febrero como COVID-19, fue designada emergencia sanitaria global por la Organización Mundial de la Salud (OMS) el 30 de enero. No obstante, aunque la OMS reconoce que tiene potencialidad de pandemia, todavía no se ha declarado oficialmente como tal.
Los síntomas del COVID-19 incluyen tos, estornudos, fiebre y dificultad para respirar; estos pueden aparecer entre dos y catorce días después de la exposición al virus. El contagio se produce por la inhalación de gotitas minúsculas que se emiten de la persona afectada a la persona sana a través de expectoraciones como tos y estornudos, en un contacto interpersonal de uno o dos metros de distancia.

No hay día en el que no se publique o difunda, a través de las redes sociales, alguna noticia falsa que, sin duda, despierta el nerviosismo y el malestar en muchas de las personas que lo leen. De hecho, son varias las instituciones que han avisado de que estas noticias falsas pueden provocar que el miedo aumente entre la población. Lo cierto es que, teniendo en cuenta la crisis sanitaria que vive el mundo entero, es esencial leer información veraz y formarse sobre lo que realmente es el COVID-19, es decir, un virus. Y como virus, tiene sus características, sus cuidados, medidas terapéuticas y, por supuesto, sus complicaciones.
La cuestión es ir más allá de lo que cuentan los medios de comunicación o el propio Gobierno, que, principalmente, se limitan a dar datos, y conocer de primera mano qué es exactamente el COVID-19. Es por esto por lo que vamos a cruzar herramientas muy punteras en sus campos. Queremos dar algunos datos claves que sirvan como base para los expertos y a partir de estos obtener respuestas.
Pero primero debemos saber cómo se estudian estas epidemias.

2.3.2 Características del estudio de la Epidemiología

2.3.2.1 Según la temporalidad de la novela:

- Estudio retrospectivo: es un estudio longitudinal en el tiempo que se analiza en el presente, pero con datos del pasado. Su inicio es posterior a los hechos estudiados.
- Estudio transversal: es un estudio que se realiza con los datos obtenidos en un momento puntual como el estudio de prevalencia.
- Estudio prospectivo: es un estudio longitudinal en el tiempo que se diseña y comienza a realizarse en el presente, pero los datos se analizan transcurrido un determinado tiempo, en el futuro.

2.3.2.2 Según el tipo de resultado que se obtenga en el estudio:

- Estudio descriptivo, que es un tipo de metodología a aplicar para deducir un bien o circunstancia que se esté presentando
- Estudio analítico. Según si existe intervención, los estudios analíticos se clasifican en:
- Estudio observacional: El investigador no interviene. Se limita a observar y describir la realidad.
- Estudio de intervención: El investigador introduce variables en el estudio, interviniendo en la realidad y desarrollo de este.

2.3.2.3 Dependiendo de si existe aleatorización o no:

- Estudios cuasiexperimentales: Son estudios en los que existe intervención, pero los sujetos participantes no son aleatorizados.
- Estudios experimentales: Los sujetos participantes han sido incluidos de forma aleatoria (ensayo clínico, ensayo comunitario, o de laboratorio). Un ensayo clínico es un estudio prospectivo, analítico y de intervención con aleatorización.

2.3.2.4 Según la unidad de estudio:

- Estudio ecológico o de correlación: La unidad de estudio es la población.

- Estudios en los que los individuos son las unidades del estudio: Comunicación de un caso, estudio de serie de casos, estudio transversal, estudio longitudinal.

2.3.3 Cómo se propaga el COVID-19

Se cree que el COVID-19 se propaga principalmente a través del contacto cercano de persona a persona. Algunas personas que no presentan síntomas pueden propagar el virus. Todavía seguimos aprendiendo acerca de cómo se propaga el virus y sobre la gravedad de la enfermedad que causa.

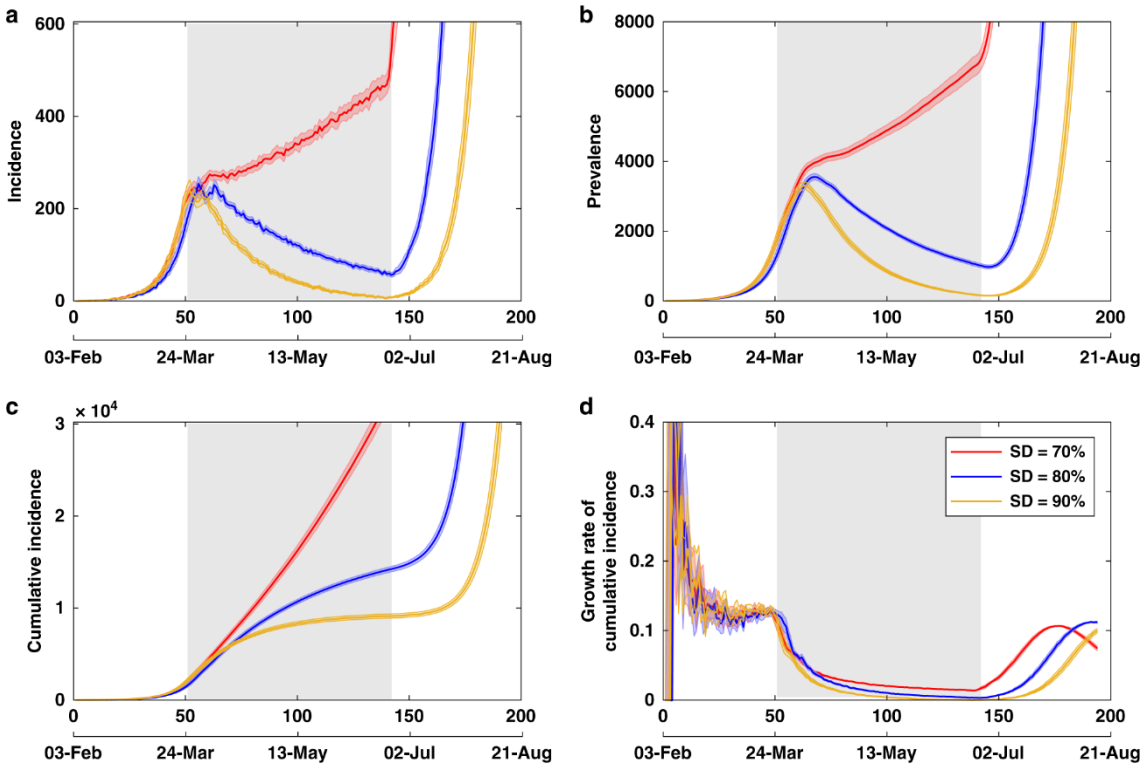
2.3.3.1 Propagación de persona a persona

Otra parte importante es el conocimiento de las características de la propagación entre personas, debido a que es un factor que se puede tener en cuenta a la hora de hacer un estudio con inteligencia artificial.

- Entre personas que están en contacto cercano (a una distancia de hasta aproximadamente 6 pies).
- A través de gotitas respiratorias que se producen cuando una persona infectada tose, estornuda o habla.
- Estas gotitas pueden terminar en la boca o en la nariz de quienes se encuentran cerca o posiblemente ser inhaladas y llegar a los pulmones.
- Las personas sin síntomas pueden propagar el COVID-19.

Estas características proporcionan que el virus se propague con asombrosa facilidad, adicionalmente hay que añadir ciertos factores que son importantes en el estudio de la propagación entre personas:

- La facilidad con la que el virus se propaga de persona a persona puede variar. Algunos virus son muy contagiosos, como el del sarampión, mientras que otros virus no se propagan tan fácilmente. Otro factor que hay que tener en cuenta es si la propagación es sostenida, es decir, se propaga de manera continua de persona a persona.
- El virus que causa el COVID-19 se propaga muy fácilmente y de manera continua entre las personas. La información sobre la pandemia en curso del COVID-19 sugiere que este virus se propaga de manera más eficiente que el virus de la influenza, pero no tan eficientemente como el del sarampión, que es un virus altamente contagioso. En general, cuanto más cercana y prolongada sea la interacción entre las personas, mayor es el riesgo de propagación del COVID-19.



Ref. 46

Estos datos tomados de Australia nos muestran como el fuerte cumplimiento del distanciamiento social (al 80% o más) controla eficazmente la enfermedad durante el período de supresión, mientras que los niveles más bajos de cumplimiento (al 70% o menos) no tienen éxito durante ninguna duración de la supresión. Una comparación de las estrategias de distanciamiento social, junto con el aislamiento de casos, la cuarentena domiciliaria y las restricciones a los viajes internacionales, a través de diferentes niveles de cumplimiento (70, 80 y 90%). La duración de cada estrategia de distanciamiento social (DS) se establece en 91 días (13 semanas), que se muestran como un área sombreada en gris entre los días 51 y 142 (los días de inicio y fin de la DS variaron a través de recorridos estocásticos: para la DS del 70% el último día de la supresión fue de 141,4 en promedio; para la DS del 80% fue de 144,2; y para la DS del 90% fue de 141,5, véase el archivo de datos de la fuente). El aislamiento de los casos, la cuarentena domiciliaria y las restricciones a las llegadas internacionales durarán hasta el final de cada escenario. Los rastros incluyen a la incidencia, b la prevalencia, c la incidencia acumulada y d la tasa de crecimiento diario de la incidencia acumulada $\dot{C}$, mostrados como perfiles de promedio (sólido) e intervalo de confianza del 95% (sombreado), a lo largo de 20 corridas. Los intervalos de confianza del 95% se construyen a partir de las distribuciones bootstrap corregidas por sesgo. La alineación entre los días simulados y las fechas reales puede diferir ligeramente entre las distintas ejecuciones.[]

2.3.3.2 Propagación de animal a persona

Como bien sabemos, el virus no solo está presente en humanos. Existen diversas noticias sobre el origen animal del COVID-19. Las principales características que se tratarán en el estudio son:

- Podría ser posible que una persona se infecte por el COVID-19 al tocar una superficie u objeto que tenga el virus y luego se toque la boca, la nariz o los ojos. No se cree que esta sea la principal forma de propagación del virus, pero aún estamos aprendiendo acerca de cómo se propaga el virus.
- Puede existir una propagación entre los animales y las personas, pero por el momento, el riesgo de propagación del COVID-19 de animales a personas se considera bajo.
- Al parecer el virus que causa el COVID-19 puede propagarse de personas a animales en ciertas situaciones. Los CDC tomaron conocimiento de una pequeña cantidad de notificaciones de mascotas en todo el mundo, incluidos gatos y perros, infectadas con el virus que causa el COVID-19, principalmente después de haber estado en contacto cercano con personas con COVID-19.

2.3.4 Formas de prevenir la enfermedad

- Respetar las medidas de distanciamiento social (aproximadamente 6 pies). Es muy importante para prevenir la propagación del COVID-19.
- Lavarse las manos con frecuencia con agua y jabón. Si no dispone de agua y jabón, use algún desinfectante de manos que contenga al menos un 60 % de alcohol.
- Limpie y desinfecte de manera rutinaria las superficies que se tocan con frecuencia.
- Cúbrase la boca y la nariz con una cubierta de tela para la cara cuando está rodeado de otras personas.

2.3.5 Diferenciación geográfica

Si bien los médicos e investigadores aún no han establecido un "cuadro epidemiológico completo" del COVID-19, los investigadores han elaborado hasta ahora cuatro teorías probables que podrían explicar por qué el nuevo coronavirus ha devastado algunos países, pero no otros.

2.3.6 El virus en la población joven

Los investigadores han observado que muchos de los países que experimentan tasas comparativamente más bajas de casos de COVID-19 también tienen poblaciones comparativamente más jóvenes y, por lo tanto, han teorizado que los países con residentes más jóvenes en promedio tienen menos probabilidades de ver brotes generalizados de COVID-19.

Esta teoría podría explicar por qué África, que tiene la población más joven en promedio cuando se la compara con todos los demás continentes, ha notificado hasta ahora unos 45.000 casos de COVID-19 entre unos 1.300 millones de personas, mientras que Italia, que tiene una edad media nacional de más de 45 años, se encuentra entre los países que se han visto más afectados por COVID-19.
Según Robert Bollinger, profesor de enfermedades infecciosas de la Escuela de Medicina de Johns Hopkins, los pacientes más jóvenes suelen ser menos propensos a tener problemas de salud subyacentes como diabetes e hipertensión, que pueden causar complicaciones potencialmente mortales en los pacientes de COVID-19.
Además, Josep Car, experto en población y salud mundial de la Universidad Tecnológica de Nanyang, en Singapur, dijo que los pacientes más jóvenes suelen tener sistemas inmunológicos más fuertes que los pacientes de más edad, lo que puede hacerlos más propensos a experimentar casos más leves de COVID-19.
Sin embargo, se describe que la teoría general de los investigadores sobre la edad media de un país que afecta a la intensidad del impacto del nuevo coronavirus tiene algunas excepciones, como Japón, que tiene la población media más vieja del mundo pero que hasta ahora ha informado de menos de 520 muertes relacionadas con el COVID-19.

2.3.7 Distancia cultural

La gente en algunos países tiende a estar más distante socialmente que en otros, lo que los investigadores teorizan que podría ayudar a prevenir la propagación del nuevo coronavirus, escribe Beech.
Por ejemplo, en la India y Tailandia, las personas suelen saludarse a distancia poniendo sus propias palmas juntas, un saludo que no requiere que las personas toquen a otras. Estos países han notificado un número relativamente bajo de casos de COVID-19. Asimismo, en Japón y Corea del Sur que también han informado de un número relativamente bajo de casos de COVID-19- las personas suelen hacer una reverencia para saludarse y tienden a usar mascarillas en público cuando no se sienten bien, .
"El distanciamiento nacional", o el aislamiento de otras naciones, también parece haber sido beneficioso para algunos países. Por ejemplo, los lugares del Pacífico Sur que han tenido comparativamente menos visitantes del extranjero que otros países han informado de un número comparativamente menor de casos de COVID-19.
Pero también hay excepciones a esta teoría. Por ejemplo, señalan, hay "muchas partes de Oriente Medio, como Iraq y los países del Golfo Pérsico, [donde] los hombres a menudo se abrazan o se dan la mano al reunirse, pero la mayoría no se enferman".

2.3.8 Medio Ambiente

Los investigadores han observado que el nuevo coronavirus parece ser el que se ha propagado más rápidamente en países con entornos templados como los Estados Unidos. Observaron que, cuando los países empezaron a notificar casos de COVID-19, los países con climas más cálidos no estaban notificando muchos casos. Las observaciones llevaron a los investigadores a preguntarse si el nuevo coronavirus luchaba por sobrevivir en el calor.
Además, se observan que un estudio realizado por modalizadores ecológicos de la Universidad de Connecticut ha descubierto que los rayos ultravioletas podrían inhibir el

nuevo virus, lo que sugiere que "las superficies en lugares soleados podrían tener menos probabilidades de permanecer contaminadas".
Pero de nuevo, como los graves brotes que se han visto en lugares tropicales como la región del Amazonas en Brasil.
"La mejor conjetura es que las condiciones del verano ayudarán, pero es poco probable que por sí solas conduzcan a una significativa desaceleración del crecimiento o a una disminución de los casos", dijo Marc Lipsitch, director del Centro de Dinámica de Enfermedades Contagiosas de la Universidad de Harvard.

2.3.9 Respuesta de los gobiernos ante la pandemia

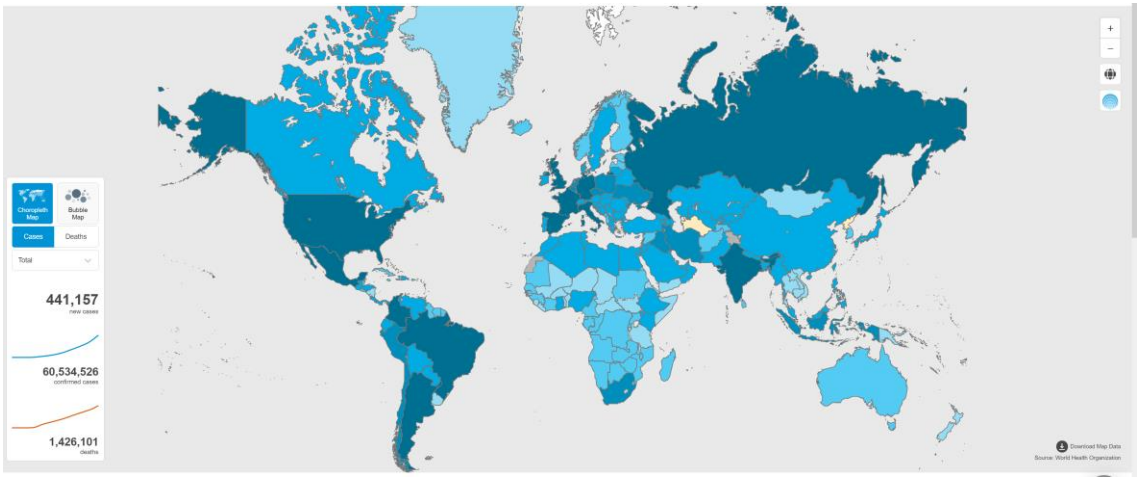
Los países que promulgaron políticas estrictas de distanciamiento social y de permanencia en el hogar al principio de sus brotes, en su mayoría experimentaron brotes más leves en general.
Señalan que los países de África que experimentaron anteriormente brotes de VIH y Ébola "reaccionaron rápidamente" y pudieron promulgar ciertas medidas de mitigación con mayor rapidez que otros países. Por ejemplo, los empleados de los aeropuertos de Sierra Leona y Uganda llevaban máscaras y pedían a los viajeros detalles de contacto mucho antes de que algunos países occidentales promulgaran precauciones similares. Sierra Leona también volvió a utilizar los protocolos de rastreo de enfermedades que los funcionarios habían utilizado durante el brote de Ébola en África occidental en 2014. Hasta ahora, el país ha notificado sólo 155 casos confirmados de COVID-19.

2.3.10 Extensión de la pandemia

Finalmente, la mayoría de los expertos concuerdan en que quizás no exista una razón en particular por la que algunos países hayan sido afectados y otros no. Es probable que la respuesta sea una combinación de los factores mencionados con anterioridad y otro al que aluden los expertos: la suerte. Países con la misma cultura y el mismo clima podrían tener resultados muy diferentes cuando una persona infectada asiste a un evento social muy concurrido y lo convierte en lo que los expertos califican como un evento de altísimo contagio.
Eso sucedió cuando un pasajero infectó a 634 personas en el crucero Diamond Princess en la costa de Japón, cuando un invitado infectado asistió a un gran funeral en Albany, Georgia, y cuando una mujer de 61 años fue a la iglesia en Daegu, Corea del Sur, propagando la enfermedad a cientos de fieles y luego a miles de otros coreanos. Debido a que la persona infectada puede no experimentar síntomas durante una semana o más, si acaso los siente, la enfermedad se propaga de forma desapercibida, de manera exponencial y aparentemente aleatoria. Si la mujer en Daegu se hubiera quedado en casa aquel domingo de febrero, el brote en Corea del Sur podría haber sido menos de la mitad de lo que es.

Algunos países que deberían haber estado inundados de casos no lo están, dejando a los investigadores rascándose la cabeza. Tailandia reportó el primer caso confirmado de coronavirus fuera de China a mediados de enero, de un viajero de Wuhan, la ciudad china donde se cree que comenzó la pandemia. En esas semanas críticas, Tailandia continuó recibiendo una afluencia de visitantes chinos. Por alguna razón, esos turistas no provocaron una transmisión local exponencial.

Tampoco se sabe qué sucede cuando los países no hacen bien las cosas y, sin embargo, al final no les va tan mal con el virus como se esperaría.



Ref. 51

3

Análisis de Requisitos

3.1 Softwares necesarios

Se van a necesitar:

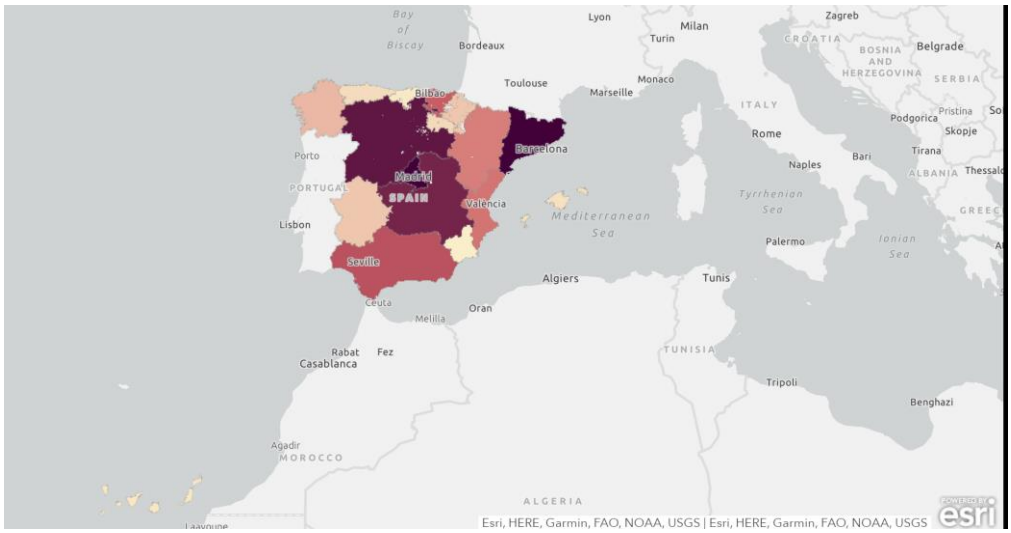
- Microsoft Excel: Para el tratamiento de los datos que se subirán en la base de datos de ArcGIS
- ArcGIS Desktop: Para la creación de las capas que se exportaran a ArcGIS Online.
- ArcGIS Online: Para el tratamiento de la visualización que tendrán las capas creadas anteriormente.
- ArcGIS WebApp: Necesario para poder dar resultados a la persona que observe el mapa y que quiera utilizar estadísticas en ellos.
- Python: Lenguaje en el que se desarrollará el código que utilizará la inteligencia artificial.
- Pytorch: Repositorio con el que se escribirán las funciones que se encarga de crear las funciones necesarias para aplicar inteligencia artificial a partir de los datos obtenidos.
- GitHub: Donde se guardarán los repositorios tanto de la inteligencia artificial como la aplicación Web, para que se pueda desplegar de forma local. También se descargarán datos de repositorios oficiales.
- Pdfaid: Se encargará de transformar los ficheros pdf que nos provee el gobierno, para cambiarlos a un formato Excel, legible en ArcGIS.

3.2 Entregables

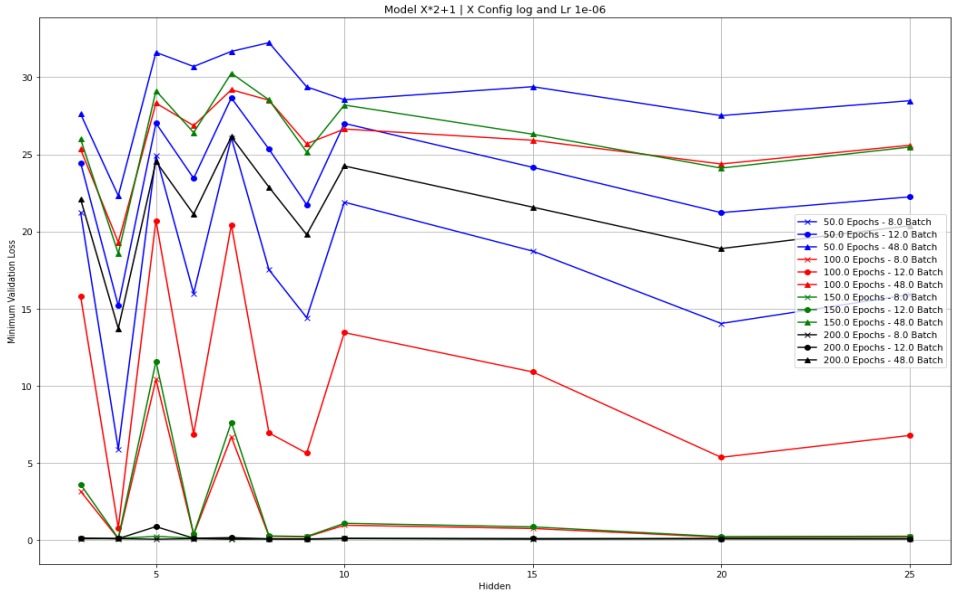
Para poder completar esta tarea necesitamos tres entregables fundamentales:

- Base de datos ArcGIS.
- Código en Pytorch capaz de procesar datos.
- ArcGIS WebApp para poder ver los resultados obtenidos.

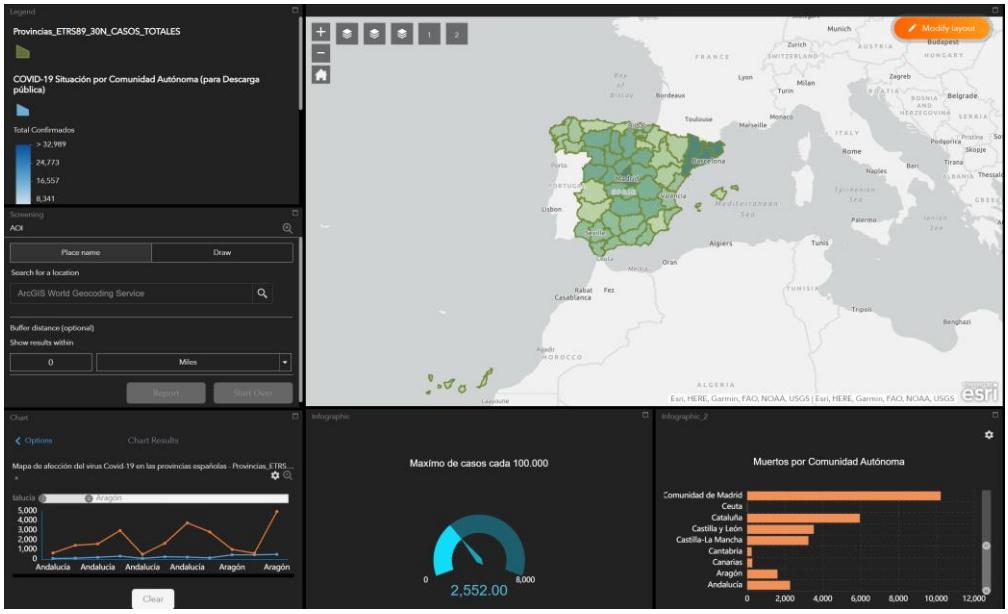
En la base de datos de ArcGIS, vamos a tener que incluir mucha información aparte del mapa generado a partir del código Pytorch, ya que un experto necesita toda la información posible para poder contrastar y dar con una predicción lo más precisa posible. Por lo que vamos a añadir mapas con recuento de casos totales, casos por comunidad autónoma, casos por provincia, mapas con índices de población más alta y con diferenciación entre jóvenes y mayores, ya que estamos hablando del COVID-19 y como hemos escuchado mil y una vez, esto no afecta igual a las personas mayores que a los jóvenes.



Para el código Pytorch, necesitamos que el código sea capaz de leer un repositorio donde tengamos información actualizada y con una fuente oficial, ya que los datos deben respetarse al máximo. Estos datos deben ser procesados mediante inteligencia artificial con un entrenamiento escaso, ya que, si queremos dar datos realmente y fiables requeriríamos de una red neuronal muy grande, de una función óptima y ordenador superponte, que sea capaz de procesar grandes cantidades de datos.



Por último, pero no menos importante, necesitamos darle visibilidad a nuestro proyecto. Para esto vamos a usar ArcGIS webApp, que nos brinda un excelente servicio de creaciones de aplicaciones webs. Vamos a crear un mapa con todas las capas que hemos descrito anteriormente, y vamos a dejar cinco espacios para poner las aplicaciones para ver las estadísticas que se necesiten. Dedicaremos una parte especial de este mapa para poder ver la capa que se crea con la IA, y así poder combinarla con las otras capas y dar todas las posibles funcionalidades que se necesitan.



3.3 El tiempo en la elaboración del estudio

Estamos trabajando sobre una pandemia que nos está afectando a todos y ha cambiado nuestra percepción de los riesgos que corremos hoy en día respecto a lo que a salud se refiere. Y es que, aunque estemos en España que es un país del primer mundo, este virus ha tenido un impacto brutal sobre nuestro país. Pero no importa donde miremos, ya sean países ricos, países pobres, países con alta densidad de población, países con baja densidad de población, este virus ha afectado al mundo entero y no siempre de la manera que se espera. Y es por eso por lo que cuanto antes podamos dar esta ayuda a la comunidad, podemos dar más información e ideas de como poder tratar los datos del COVID-19 y así mejorar nuestra reacción ante este.

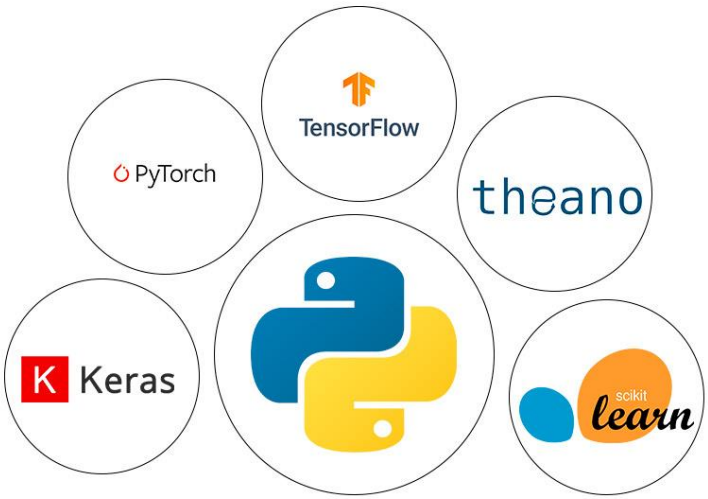
4

Diseño

4.1 Estudios de las Tecnologías

4.1.1 Tecnologías para la Inteligencia Artificial

Para el desarrollo de este proyecto, es necesario encontrar el software que mejor se adapte a nuestro perfil, que es el de una persona con conocimientos no muy avanzados, pero tampoco es básico. Y la correcta elección del software va a ser vital para poder desempeñar el trabajo de la mejor manera posible. Para la elección del software nos hemos basado en los siguientes criterios: Bibliotecas, lenguaje de programación, entorno gráfico, capacidad y alcance. Se han estudiado todas las posibilidades que nos brindaban estos softwares



Ref. 45

4.1.1.1 Microsoft Azure machine Learning Studio

Microsoft Azure ofrece Azure Machine Learning como un servicio de pago. Usando Azure ML, los negocios no requieren de la instalación de complejos o la compra de grandes equipos o software. Sólo se necesita comprar los servicios que creamos necesarios y podremos empezar a desarrollar las aplicaciones de Machine Learning inmediatamente.

Es una herramienta colaborativa que, con la herramienta drag-and-drop, se puede utilizar para crear, probar y desplegar soluciones predictivas de ML. Azure publica modelos como servicios web que pueden ser usados fácilmente por aplicaciones personalizadas o soluciones de software. La interfaz principal de Microsoft Azure ML Studio presenta un entorno de pruebas gráfico donde los usuarios pueden combinar diferentes módulos predefinidos para crear experimentos destinados a analizar datos y realizar diferentes tipos de predicciones y deducciones. La colección de módulos predefinidos es bastante extensa, ofrece funcionalidades para la conversión de datos, transformación y selección, y algoritmos de machine learning para diferentes tareas como clustering, clasificación o regresión. Usando estos módulos, un usuario casi sin experiencia puede crear un proyecto relativamente sofisticado de inteligencia artificial o data science.

Desde el punto de vista del desarrollo, las principales ventajas del Microsoft Azure Machine Learning Studio son su exhaustividad, facilidad de uso y la versatilidad de la experiencia requerida. Un usuario con experiencia limitada en proyectos ML puede desarrollar soluciones bastante complejas en poco tiempo utilizando esta herramienta. El único inconveniente de esto son las restricciones en la entrada y la salida de los módulos de scripts (solo las colecciones de datos almacenadas en dataframes) y el menor soporte para frameworks de ML típicos como TensorFlow y Keras. Con respecto a este último aspecto, el framework admite por defecto Microsoft Cognitive Toolkit de Microsoft (CNTK) que proporciona funcionalidades similares a TensorFlow.

4.1.1.2 Amazon Machine Learning

Machine Learning es adecuado para el investigador de datos, el investigador de Machine Learning o el desarrollador. AWS ofrece servicios de aprendizaje automático y herramientas adaptadas para satisfacer nuestros deseos dependiendo del nivel de experiencia.

Amazon Machine Learning es una herramienta visual que ayuda a previsualizar los datos para garantizar la calidad. Una vez construido el modelo, el usuario puede utilizar las herramientas de aprendizaje automático de AWS para evaluarlas y ajustarlas. Después de esto, el modelo está listo para las predicciones posteriores. Estas aplicaciones también pueden llamar a la API de lotes para las predicciones. Además, la API en tiempo real puede utilizarse para generar predicciones a petición. Con Amazon ML el usuario puede crear datos a partir de grandes conjuntos de datos, generar miles de millones de predicciones y servir estas predicciones en tiempo real y con un alto rendimiento. No hay costos iniciales para AWS ML, sólo que el usuario tiene que pagar por lo que ha utilizado. Esto beneficia de tal manera que el usuario puede iniciar una aplicación a pequeña escala a medida que el negocio crece. Además, si tiene alguna duda, no dude en preguntar en el cuadro de comentarios.

Los servicios de Amazon Machine Learning están disponibles en dos niveles: análisis predictivo con Amazon Machine Learning y la herramienta SageMaker para científicos de datos.

Amazon Machine Learning para análisis predictivo es una de las soluciones más automatizadas del mercado y la mejor opción para operaciones sensibles a plazos.
Todas las operaciones de preprocesamiento de datos se realizan automáticamente: el servicio identifica qué campos son categóricos y cuáles son numéricos, y no le pide a un usuario que lo haga.El usuario no está obligado a conocer ningún método de aprendizaje automático, ya que Amazon los elige automáticamente después de ver los datos proporcionados.
Este alto nivel de automatización actúa como una ventaja y una desventaja para el uso de Amazon Machine Learning. Si necesita una solución totalmente automatizada pero limitada, el servicio puede satisfacer sus expectativas. Si no, hay SageMaker.

4.1.1.3 Tensor Flow

TensorFlow es una biblioteca de código abierto que se basa en un sistema de redes neuronales. Esto significa que puede relacionar varios datos en red simultáneamente, de la misma forma que lo hace el cerebro humano. Por ejemplo, puede reconocer varias palabras del alfabeto porque relaciona las letras y fonemas. Otro caso es el de imágenes y textos que se pueden relacionar entre sí rápidamente gracias a la capacidad de asociación del sistema de redes neuronales. En el programa, se almacenan todas las pruebas y experimentos que se realizaron para el desarrollo de programas y aplicaciones.
TensorFlow proporciona una sintaxis accesible y legible que es esencial para facilitar el uso de estos recursos de programación. La compleja sintaxis es lo último que los desarrolladores necesitan saber dada la naturaleza avanzada del aprendizaje de las máquinas.
TensorFlow también proporciona excelentes funcionalidades y servicios cuando se compara con otros populares marcos de deep learning . Estas operaciones de alto nivel son esenciales para llevar a cabo complejos cálculos paralelos y para construir modelos avanzados de redes neuronales. Hay que tener en cuenta que es una biblioteca de bajo nivel que proporciona más flexibilidad, pero a su vez dificultad. Se, puede definir sus propias funcionalidades o servicios para nuestros modelos, este es un parámetro muy importante para los investigadores porque les permite cambiar el modelo en función de las necesidades cambiantes de los usuarios.
Por último, proporciona más control de la red. De esta manera permite a los desarrolladores e investigadores entender cómo se implementan las operaciones a través de la red. Siempre pueden hacer un seguimiento de los nuevos cambios realizados a lo largo del tiempo.

4.1.1.4 Keras

Keras fue creado para ser fácil de usar, modular, fácil de extender y para trabajar con Python. Las capas neuronales, las funciones de coste, los optimizadores, los esquemas de inicialización, las funciones de activación y los esquemas de regularización son todos módulos independientes que pueden combinarse para crear nuevos modelos. Los nuevos módulos son simples de agregar, como nuevas clases y funciones. Los modelos se definen en código Python, no en archivos de configuración de modelos separados.
Las mayores razones para usar Keras proviene principalmente de su facilidad a la hora de usarlo y su gran versatilidad. Más allá de la facilidad de aprendizaje y la facilidad de

construcción de modelos, Keras ofrece las ventajas de una amplia adaptación y soporte para una amplia gama de opciones de despliegue de producción, integración con al menos cinco motores de back-end (TensorFlow, CNTK, Theano, MXNet, y PlaidML), y un fuerte soporte para múltiples GPUs y entrenamiento distribuido. Además, Keras está respaldado por Google, Microsoft, Amazon, Apple, Nvidia, Uber, y otros.
Con Keras, podemos construir redes neuronales simples o muy complejas en pocos minutos. Keras fue desarrollado de tal manera que debe ser más fácil de usar. El modularidad, es un elegante principio de Keras ya que todo puede ser representado como módulos y estos a su vez pueden ser combinados según los requerimientos del usuario. Aunque Keras ha sido diseñado de tal manera que puede implementar casi todo lo que desee, no llegan al nivel que tienen bibliotecas de bajo nivel, ya que estas proporcionan más flexibilidad (Tensor Flow). Puedes ajustar TF mucho más en comparación con Keras. Aunque Keras proporciona nuevos módulos que son fáciles de agregar (como nuevas clases y funciones), y los módulos existentes brindan amplios ejemplos, que al ser tan fáciles de crear permite una expresividad total, lo que hace que Keras sea adecuado para la investigación avanzada TensorFlow ofrece operaciones más avanzadas en comparación con Keras, pero a su vez requiere un conocimiento más avanzado y detallado. Los modelos se describen en el código de Python, que es compacto, más fácil de depurar y permite la extensibilidad. Esto es muy útil si estás haciendo una investigación o desarrollando algún tipo especial de modelos de aprendizaje profundo.

4.1.1.5 Caffe

Caffe (Convolutional Architecture for Fast Feature Embedding) es un marco de código abierto y de alto rendimiento para el desarrollo de modelos de aprendizaje de máquinas.
Caffe es un marco popular debido a su velocidad. El marco puede procesar más de 60 millones de imágenes por día con una sola GPU de alto rendimiento, como la Nvidia Tesla K40. El marco toma sólo un milisegundo por imagen para la inferencia y cuatro milisegundos por imagen para el aprendizaje, también soporta muchos tipos de modelos de aprendizaje profundo y está especializado en la segmentación y clasificación de imágenes. Los tipos soportados incluyen redes neuronales convolucionales (CNN), redes neuronales recurrentes (RNN), memoria a largo plazo y corto plazo (LSTM) y diseños de redes neuronales completamente conectadas. El marco soporta aceleración de CPU Intel y GPGPU Nvidia junto con implementaciones de tarjetas multigráficas. Caffe2 soportará AMD OpenCL, FPGAs y aceleradores AI. El programa está codificado en C++ con una interfaz Python y está disponible bajo una licencia BSD. Como uno de sus principales desarrolladores, Facebook anunció Caffe2 en abril de 2017.
En Caffe, no tenemos ningún método sencillo para desplegar. Necesitamos compilar todos y cada uno de los códigos fuente para desplegarlos, lo cual es un inconveniente, además, no tiene API de alto nivel, por lo que será difícil experimentar con Caffe, la configuración de una manera no estándar con API de bajo nivel. El enfoque de Caffe de las API de nivel medio a bajo proporciona poco apoyo de alto nivel y una configuración profunda limitada. La interfaz de Caffe es más de C++, lo que significa que los usuarios necesitan realizar más tareas manualmente, como la creación de archivos de configuración, etc.

El marco de trabajo de Caffe es más adecuado para el despliegue del borde de producción. Mientras que ambos marcos tienen un conjunto diferente de usuarios objetivo. Caffe apunta a los teléfonos móviles y a las plataformas con restricciones computacionales.

4.1.1.6 Pytorch

PyTorch es una librería open source basada en Python, enfocada a la realización de cálculos numéricos mediante programación de tensores, lo que facilita su aplicación al desarrollo de aplicaciones de aprendizaje profundo. La sencillez de su interfaz, y su capacidad para ejecutarse en GPUs (lo que acelera el entrenamiento de los modelos), lo convierten en la opción más asequible para crear redes neuronales artificiales.
Originalmente desarrollado por FAER (siglas de Facebook AI Research), PyTorch ha sido a su vez una pieza fundamental en el desarrollo de relevantes aplicaciones de inteligencia artificial, como el Autopilot de Tesla y el Pyro de Uber.
PyTorch soporta gráficos de computación dinámicos que permiten cambiar el comportamiento de la red sobre la marcha, a diferencia de los gráficos estáticos que se utilizan en marcos como Tensorflow.
Desde su lanzamiento en enero de 2016, muchos investigadores han seguido adoptando cada vez más PyTorch. Se ha convertido rápidamente en una biblioteca de consulta debido a su facilidad para construir redes neuronales extremadamente complejas. Le está dando una dura competencia a TensorFlow, especialmente cuando se usa para trabajos de investigación. Sin embargo, todavía queda algo de tiempo antes de que sea adoptado por las masas debido a sus todavía "nuevas" y "en construcción" etiquetas.
Los creadores de PyTorch imaginaron que esta biblioteca era altamente imperativa, lo que les permitiría ejecutar todos los cálculos numéricos rápidamente. Esta es una metodología ideal que encaja perfectamente con el estilo de programación de Python. Ha permitido a los científicos de aprendizaje profundo, a los desarrolladores de aprendizaje de máquinas y a los depuradores de redes neuronales ejecutar y probar parte del código en tiempo real. Así no tienen que esperar a que se ejecute todo el código para comprobar si funciona o no.
Siempre puedes usar tus paquetes Python favoritos como NumPy, SciPy y Cython para extender las funcionalidades y servicios de PyTorch cuando sea necesario. Ahora te preguntarás, ¿por qué PyTorch? ¿Qué tiene de especial usarlo para construir modelos de aprendizaje profundo?
La respuesta es bastante simple, PyTorch es una biblioteca dinámica (muy flexible y que se puede usar según los requerimientos y cambios) que actualmente es adoptada por muchos de los investigadores, estudiantes y desarrolladores de inteligencia artificial. En la reciente competición de Kaggle, la biblioteca de PyTorch fue utilizada por casi todos los 10 primeros clasificados.
En pytorch es destacable:
Interfaz simple: Ofrece una API fácil de usar, por lo que es muy sencillo de operar y ejecutar como Python.

- De naturaleza pitónica: Esta biblioteca, al ser python, se integra sin problemas con la pila de datos científicos de Python. Por lo tanto, puede aprovechar todos los servicios y funcionalidades que ofrece el entorno Python.
- Gráficos computacionales: Además de esto, PyTorch proporciona una excelente plataforma que ofrece gráficos computacionales dinámicos, por lo que puede

cambiarlos durante el tiempo de ejecución. Esto es muy útil cuando no se tiene idea de cuánta memoria será necesaria para crear un modelo de red neuronal.

- Una colección versátil de módulos: PyTorch viene con varios módulos especialmente desarrollados como torchtext, torchvision y torchaudio para trabajar con diferentes áreas de aprendizaje profundo como la PNL, la visión por ordenador y el procesamiento del habla.
- Amigable con los números: PyTorch trabaja con NumPy como estructuras tensoriales para sus cálculos que son todos compatibles con la GPU.
- Fácil de implementar la retropropagación: PyTorch soporta la autodiferenciación, es decir, simplifica enormemente la forma en que se manejan los cálculos complejos como la retropropagación, registrando las operaciones realizadas en una variable y ejecutándolas hacia atrás. Esto demuestra ser eficaz para ahorrar tiempo y también quita la carga de la espalda de los programadores.
- Más pitónico: PyTorch es considerado más pythonico por varios desarrolladores ya que soporta hacer cambios dinámicos en su código.
- Depuración flexible y sin dolor: PyTorch no requiere que definas todo el gráfico a priori. Funciona con un paradigma imperativo, lo que significa que cada línea de código añade un determinado componente al gráfico, y cada componente puede ejecutarse, probarse y depurarse independientemente de la estructura completa del gráfico, lo que lo hace muy flexible.

Por último, haciendo una pequeña comparación con Tensorflow, ya que vemos que, aunque Tensorflow ya es un marco ML/DL bien establecido con varios fieles partidarios, PyTorch ha encontrado su fortaleza gracias a su enfoque gráfico dinámico y su flexible estrategia de depuración. PyTorch tiene varios investigadores que lo apoyan activamente por estas razones. En el año 2018-19, se observó que los trabajos de investigación que mencionan a PyTorch se han duplicado en número.
Tensorflow 2.0 ha introducido un paradigma de ejecución ávido de definiciones de gráficos dinámicos en líneas similares a PyTorch. Sin embargo, los recursos para ayudar a aprender esta característica son todavía escasos. Aunque Tensorflow es a menudo promocionado como la librería ML/DL más fuerte de la industria, PyTorch sigue creciendo, debido a sus curvas de aprendizaje más suaves para los recién llegados.
Estas redes neuronales se han convertido, quizá, en la rama más prometedora de la inteligencia artificial, siendo la base de otras tecnologías como los sistemas de traducción automática, de reconocimiento de imágenes, facial, de voz...
Con el tiempo, y gracias a una facilidad de uso no reñida con su uso en el ámbito industrial, PyTorch se ha convertido en uno de los frameworks de Deep Learning más populares del mundo, al que sólo hacen sombra Tensorflow y Keras, ambos respaldados por el patrocinio de Google.

4.1.1.7 Resultado del estudio

Como podemos observar, una gran variedad de tecnologías que podemos usar tanto para el machine learning como para el Deep learning (y eso que nos hemos dejado muchas interesantes en el tintero), pero debemos intentar elegir la que más se adapte a nuestro perfil. Personalmente, creo que mi perfil es de una persona con conocimientos medios, ya que tengo algunas nociones sobre todo lo que abarca la inteligencia artificial,

pero no puedo considerar que mi perfil sea avanzado, ya que son escasas horas las que yo le he dedicado al uso de estas tecnologías.
Después de este análisis que hemos realizado de las diferentes tecnologías, he decidido usar Pytorch por varios motivos.

- Su gran modularidad, es muy ventajoso para un perfil como el mío, ya que la comunidad es inmensa y al tener algún problema que no se resolver puedo acudir a los módulos de la comunidad que me pueden servir de ayuda.
- La curva de aprendizaje no es tan elevada como la de Tensor Flow.
- Es una tecnología muy nueva y puntera que está ganando cada vez más seguidores.
- Se trabaja con Python, esto es una de las ventajas que más me convencen, ya que hoy en día Python es uno de los lenguajes de programación más utilizados y que quiero aprender por la utilidad y versatilidad que tiene.
- Tiene un gran alcance, lo que significa que se puedo desarrollar mi proyecto estando seguro de que Pytorch no será un límite.
- Es gratuito

- Tensor Flow es el siguiente paso, una vez aprenda a usar Pytorch considero que podría llegar al siguiente nivel y empezar a trabajar con Tensor Flow, pero ahora mismo con mis conocimientos, creo que Tensor Flow me llevaría demasiado tiempo y no conseguiría una mejora significativa.

A continuación, dejo esta pequeña imagen, para que se vea de una forma más representativa de cuáles son las librerías más buscadas. El estudio ha sido realizado por "The Data Incubator" en el cual tienen subido en GitHub todo el código que se ha utlizado para la realización de dicha tabla.

Library	Rank	Overall	Github	Stack Overflow	Google Results
tensorflow	1	10.87	4.25	4.37	2.24
keras	2	1.93	0.61	0.83	0.48
caffe	3	1.86	1.00	0.30	0.55
theano	4	0.76	-0.16	0.36	0.55
pytorch	5	0.48	-0.20	-0.30	0.98
sonnet	6	0.43	-0.33	-0.36	1.12
mxnet	7	0.10	0.12	-0.31	0.28
torch	8	0.01	-0.15	-0.01	0.17
cntk	9	-0.02	0.10	-0.28	0.17
dlib	10	-0.60	-0.40	-0.22	0.02
caffe2	11	-0.67	-0.27	-0.36	-0.04
chainer	12	-0.70	-0.40	-0.23	-0.07
paddlepaddle	13	-0.83	-0.27	-0.37	-0.20
deeplearning4j	14	-0.89	-0.06	-0.32	-0.51
lasagne	15	-1.11	-0.38	-0.29	-0.44
bigdl	16	-1.13	-0.46	-0.37	-0.30
dynet	17	-1.25	-0.47	-0.37	-0.42
apache singa	18	-1.34	-0.50	-0.37	-0.47
nvidia digits	19	-1.39	-0.41	-0.35	-0.64
matconvnet	20	-1.41	-0.49	-0.35	-0.58
tflearn	21	-1.45	-0.23	-0.28	-0.94
nervana neon	22	-1.65	-0.39	-0.37	-0.89
opennn	23	-1.97	-0.53	-0.37	-1.07

Ref. 36

4.1.2 Tecnologías para GIS

4.1.2.1 Bentley Systems

Bentley Systems se compromete a avanzar BIM y GIS a través de gemelos digitales de ingeniería de infraestructura 4D para ciudades digitales. Ingenieros, profesionales geoespaciales y propietarios-operadores de infraestructuras se benefician de aplicaciones y servicios de nubes gemelas digitales que hacen avanzar el modelado de la realidad (ContextCapture y OrbitGT); la planificación, diseño y operaciones de sistemas de agua, aguas residuales y aguas pluviales, y la resistencia a las inundaciones (OpenFlows); la planificación y visualización geoespacial urbana lista para la ingeniería (OpenCities Map y OpenCities Planner); la gestión de información geotécnica (OpenGround); y la simulación y análisis de movilidad (LEGION y CUBE).
Tanto en 2018 como en 2019, Microsoft nombró a Bentley Systems como socio del año en su categoría CityNext. En 2019, las herramientas de diseño de ingeniería para plantas, infraestructuras y el estudio de mercado BIM del ARC Advisory Group clasificaron a Bentley como el número 1 en distribución de agua y aguas residuales.

Systems, Incorporated proporciona varios programas y servicios para el diseño, la construcción y el desarrollo de infraestructuras. En el mercado de los GIS, algunos de los productos clave que ofrece la empresa son Bentley Map, Bentley Map PowerView, Bentley Map Enterprise y Bentley Map Mobile. Bentley Map Enterprise es un GIS intrínsecamente tridimensional que permite la creación, el mantenimiento y el intercambio de datos geoespaciales en 2D y 3D. Bentley Map Enterprise es compatible con las bases de datos espaciales de Oracle y Graph, y con las bases de datos espaciales de Microsoft SQL Server.
Bentley Systems es el principal proveedor mundial de soluciones de software para ingenieros, arquitectos, profesionales geoespaciales, constructores y propietarios-operadores para el diseño, la construcción y el funcionamiento de infraestructuras, incluyendo obras públicas, servicios públicos, plantas industriales y ciudades digitales. Las aplicaciones de modelado abierto basadas en MicroStation de Bentley, y sus aplicaciones de simulación abierta, aceleran la integración del diseño; sus ofertas ProjectWise y SYNCHRO aceleran la entrega de proyectos; y sus ofertas AssetWise aceleran el rendimiento de los activos y de la red. Abarcando la ingeniería de infraestructuras, los servicios iTwin de Bentley están avanzando fundamentalmente BIM y GIS a los gemelos digitales 4D.
Bentley Systems emplea a más de 3.500 personas, genera unos ingresos anuales de 700 millones de dólares en 170 países y ha invertido más de 1.000 millones de dólares en investigación, desarrollo y adquisiciones desde 2014. Desde sus inicios en 1984, la compañía ha permanecido mayoritariamente en manos de sus cinco hermanos fundadores de Bentley. www.bentley.com

4.1.2.2 ENVI

El software de análisis de imágenes ENVI es utilizado por los profesionales de GIS, científicos de teledetección y analistas de imágenes para extraer información significativa de las imágenes para tomar mejores decisiones. ENVI puede ser desplegado y accedido desde el escritorio, en la nube, y en dispositivos móviles, y puede ser personalizado a través de una API para cumplir con los requisitos específicos del proyecto.
La familia de productos ENVI proporciona una variedad de soluciones de software para el procesamiento y análisis de imágenes geoespaciales utilizadas por los científicos, investigadores, analistas de imágenes, y profesionales de GIS en todo el mundo. Las soluciones ENVI combinan la última tecnología de procesamiento de imágenes espectrales y análisis de imágenes con una interfaz intuitiva y fácil de usar para ayudarle a obtener información significativa de las imágenes.

Profesionales de diversas industrias y disciplinas, tales como defensa e inteligencia, planificación urbana, minería, geología y ciencias del espacio, y ciencias de la tierra utilizan las soluciones ENVI para obtener respuestas rápidas y precisas para ayudarles a tomar decisiones. La familia de productos ENVI ofrece un robusto conjunto de herramientas de procesamiento y análisis de imágenes para apoyar sus flujos de trabajo de explotación de imágenes, e integrarse con el software de GIS más popular.

Y, todas las soluciones ENVI están construidas en IDL, un poderoso lenguaje de programación, permitiendo una fácil personalización de características y

funcionalidades para satisfacer sus necesidades únicas. Los productos ENVI hacen más fácil que nunca leer, explorar, preparar, analizar y compartir información de imágenes.

La familia de productos ENVI incluye:

- ENVI
- ENVI para ArcGIS® Server
- Módulos ENVI
- ENVI EX
- SARScape

4.1.2.3 QGIS

QGIS una aplicación de sistema de información geográfica de escritorio (GIS) libre y de código abierto que soporta la visualización, edición y análisis de datos geoespaciales.
Características principales
QGIS es capaz de:

- Soportar tanto las capas ráster [imágenes] como las capas de vectores
- Integrar con otros paquetes GIS de código abierto, incluyendo PostGIS, GRASS GIS, y MapServer.
- Soportar archivos shape, coberturas, bases de datos geológicos personales, dxf, MapInfo, PostGIS, y otros formatos.
- Importar diferentes formatos de archivo, como una exportación KML de la Plataforma TravelTime

QGIS también es capaz de soportar un gran número de plugins. Los plugins escritos en Python o C++ amplían las capacidades de QGIS. Los plugins pueden:

- Realizar funciones de geoprocesamiento, que son similares a las herramientas estándar encontradas en ArcGIS
- Interfaz con las bases de datos PostgreSQL/PostGIS, SpatiaLite y MySQL

Mientras que ArcGIS sigue siendo el estándar, QGIS es una alternativa cada vez más popular a las opciones de software GIS comercial. Muchas organizaciones públicas y privadas han adoptado QGIS, incluyendo el estado austriaco de Vorarlberg, y los cantones suizos de Glarus y Solothurn.

A un nivel básico, los usuarios necesitarán las siguientes habilidades:

- Conocimientos informáticos
- La comprensión de los conceptos del GIS, como las coordenadas, las capas y las proyecciones

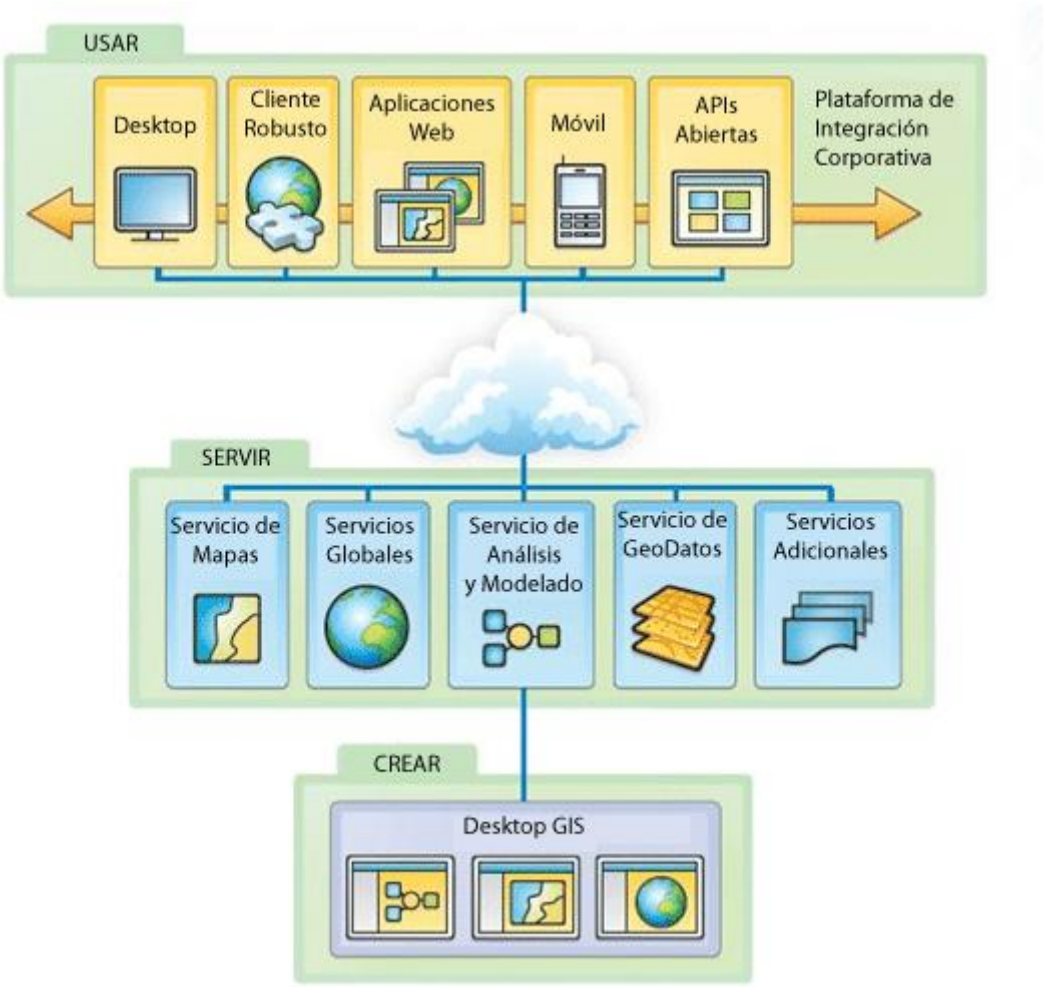
Los usuarios más avanzados también podrían tener conocimiento de:

- Gestión de bases de datos
- Cartografía
- Análisis espacial

4.1.2.4 ArcGIS

ArcGIS es el nombre de un conjunto de productos de software en el campo de los Sistemas de Información Geográfica o GIS. Producido y comercializado por ESRI, agrupando bajo el nombre genérico ArcGIS varias aplicaciones para la captura, edición, análisis, tratamiento, diseño, publicación e impresión de información geográfica. Estas aplicaciones se engloban en familias temáticas como ArcGIS Server, para la publicación y gestión web, o ArcGIS Móvil para la captura y gestión de información en campo.
A medida que el mundo cambió a la cartografía basada en la computadora, la prioridad se convirtió en la eficiencia y el poder analítico. Las herramientas informáticas para hacer hermosos mapas no existieron durante un tiempo, y durante un tiempo después de eso, fueron difíciles de encontrar y utilizar.
ArcGIS Desktop, la familia de aplicaciones GIS de escritorio, es una de las más ampliamente utilizadas, incluyendo en sus últimas ediciones las herramientas ArcReader, ArcMap, ArcCatalog, ArcToolbox, ArcScene y ArcGlobe, además de diversas extensiones. ArcGIS for Desktop se distribuye comercialmente bajo tres niveles de licencias que son, en orden creciente de funcionalidades (y coste): ArcView, ArcEditor y ArcInfo. La licencia que nos fue proporcionado por la UMA fue ArcEditor.
Actualmente ARCGIS no es sólo una tecnología para elaborar mapas, sino que es también una infraestructura basada en la nube que posibilita la colaboración y el uso compartido de la información geográfica. Así pues, ARCGIS ha evolucionado desde una única herramienta para el análisis y el procesamiento de datos espaciales a todo un conjunto de aplicaciones relacionadas entre sí destinadas al manejo y el tratamiento de la información geográfica. Lo cual la hace una herramienta muy potente y que se adapta muy bien al proyecto que se plantea en este trabajo.
Aunque normalmente se asocie ArcGIS con ArcGIS desktop, el software de escritorio, la familia ArcGIS es muy extensa:

- ArcGIS Pro
- ArcGIS Online
- ArcGIS Server
- ArcGIS Mobile



Ref. 47

4.1.2.4.1 Orígenes y desarrollo temprano

Esri, anteriormente conocido como Environmental Systems Research Institute, fue fundado en 1968 como una firma consultora bajo el mando de Jack y Laura Dangermond. Comenzaron a producir su software actualizado basado en GIS alrededor de 1997, estrenando ArcMap en 1999. ArcMap fue finalmente reemplazado por ArcGIS en 2000. Aunque comúnmente se le llama el grupo de aplicaciones y herramientas de escritorio utilizadas para el mapeo de GIS, ahora consiste en una plataforma o infraestructura completa de software y servicios que es totalmente móvil y puede estar basado en la nube.

Entre sus aspectos más recientemente innovadores de integración de plataformas, los usuarios pueden incorporar la modelización en 3D, el análisis espacial, las visualizaciones de mapas complejos, la navegación de datos y la recopilación de información geográfica en tiempo real, todo ello mediante herramientas de gestión y recopilación de datos. Como tal, donde las anteriores encarnaciones del programa se limitaban a la cartografía de información geográfica, se ha convertido en una plataforma líder en la industria con capacidades de nube de vanguardia.

4.1.2.4.2 ¿Cómo funciona?

El ArcGIS utiliza el concepto de Sistema de Información Geográfica (GIS) para construir mapas en los que cada categoría de característica espacial es una capa separada. Las capas están "registradas" espacialmente, de modo que cuando el usuario las superpone,

el programa puede alinearlas correctamente para construir un mapa. Hay varios tipos de capas, y el usuario tiene muchas opciones en cuanto a cómo representarlas. Las tres primeras se llaman "capas de vectores" o "capas de características" y contienen características individuales que el programa puede distinguir.

- Punto (por ejemplo, edificios, puntos de referencia). Cero dimensiones.
- Línea o arco (por ejemplo, carreteras y calles, arroyos, ferrocarriles, líneas eléctricas). Unidimensional.
- Polígono (por ejemplo, entidades políticas, geografías de censo como tramos). Dos dimensiones.
- Imágenes rasterizadas (por ejemplo, una fotografía aérea, un mapa topográfico escaneado o un modelo de elevación). En contraste con la capa vectorial basada en características, se trata de imágenes basadas en una cuadrícula de celdas X por Y, cada una de las cuales tiene un valor que representa algo como la elevación, la clasificación del uso de la tierra o el valor del color.
- Los datos pueden asociarse con las características espaciales, y ser mapeados o analizados:
- Puede haber atributos, o datos tabulares, asociados con cada característica en una capa (por ejemplo, datos demográficos para cada Tramo de Censo).
- Se pueden añadir ("unir") tablas de datos (por ejemplo, base de datos o archivos de hojas de cálculo) a una capa si hay un campo con valores comunes (por ejemplo, el número de la zona de censo).

El programa también puede cartografiar los archivos de datos con referencia espacial en algunos formatos de hoja de cálculo y base de datos (por ejemplo, si un campo contiene coordenadas de latitud/longitud). Las tablas que contienen datos de direcciones pueden "geocodificarse" para cartografiar las ubicaciones basándose en una capa de calles. Los usuarios pueden abrir una imagen raster no registrada y georreferenciarla utilizando las funciones del programa, o vectorizar los rasgos a partir de una imagen raster.
También puede añadir su propia información a un mapa con herramientas de dibujo y escritura.

Este software le permite producir rápidamente asombrosos mapas y modelos visuales, incluyendo representaciones en 3D y mapas de flujo de población. Con la función de arrastrar y soltar, las hojas de cálculo de los datos se pueden cargar rápidamente en la nube y visualizar. También hay una herramienta de mapeo inteligente que sugiere las mejores clasificaciones, estilos y colores que se ajustan a sus datos.
Las imágenes están disponibles en alta resolución, tomadas tanto de fuentes históricas como recientes en todo el mundo, lo que permite la construcción de mapas históricos, así como observaciones de datos demográficos recientes. Los fenómenos de la superficie, como la temperatura, la elevación, la precipitación, etc., también pueden integrarse plenamente en estos mapas visuales y modelos con herramientas únicas para el análisis de la superficie.

4.1.2.4.3 ¿Dónde se utiliza?

Como plataforma líder en la industria, el conjunto de herramientas y aplicaciones centrales de este programa es utilizado por la gran mayoría de las instituciones, empresas y departamentos que se ocupan del análisis de datos geográficos. Sin embargo, la simplicidad de su interfaz en línea también ha visto aumentar su valor en el uso periodístico y mediático.

El software Esri tiene una sólida historia y reputación. Este hecho lo convierte en un software básico para muchas empresas que se dedican a los sistemas de información geográfica. En particular, es utilizado por los gobiernos locales y estatales de todo el mundo, incluso en los Estados Unidos de América.

4.1.3 Evaluación ArcGIS y QGIS

Para el desarrollo de la conclusión de nuestro estudio, es necesario que se tengan en cuenta estas comparaciones, ya que en la comunidad GIS, es un debate que tiene innumerables hilos especificando porqué uno es mejor que el otro.

Ambos han estado disponibles durante años y tienen una fuerte base de fans. El uso de cada uno de este software genera una dualidad o polarización de las preferencias y opiniones sobre las herramientas de los Sistemas de Información Geográfica (GIS). Este artículo no trata de mostrar las ventajas de las deficiencias de cualquiera de ellas, sino de mostrar las razones por las que QGIS es una buena opción.

QGIS es versátil, está a la vanguardia cuando se trata de consumir datos ya que usa la biblioteca GDAL/OGR para leer y escribir formatos de datos GIS. Más de 70 formatos de vectores son soportados.

ArcGIS está hecho para trabajar con cualquier tipo de archivo: ENC, shapefile, geodatabase, formatos de MapInfo, formatos de archivos de Microstation, AutoCAD DXF, SpatiaLite, Oracle Spatial, MSSQL Spatial databases, WellKnownText (WKT)...

4.1.3.1 Documentación

Cuando se trata de documentación, ArcGIS es insuperable. Proporciona una documentación muy elaborada sobre cómo utilizar las herramientas, con ejemplos incluidos, lo que ayuda a una mayor comprensión, tanto a nivel desktop como desarrollo. Esto es diferente en QGIS, ya que la documentación es insuficiente.

4.1.3.2 Sistema operativo

QGIS se puede instalar en varios sistemas operativos, como mac, linux o windows. Sin embargo, ArcGIS sólo puede instalarse en windows. La posibilidad multiplataforma que ofrece QGIS lo convierte en una gran alternativa respecto a ArcGIS.

4.1.3.3 Licencia para geoprocesar

ArcGIS tiene muy buenas herramientas de geoprocesamiento, sólidas y extensas. Sin embargo, el nivel de licencia del que dispongas determina qué herramientas se pueden utilizar en ArcGIS. Una licencia básica da acceso a un gran número de herramientas de gran alcance, pero una licencia avanzada te da acceso a todo.

En QGIS no hay licencias básicas o avanzadas. Al tratarse de un software libre, no limita las herramientas que se pueden utilizar. En QGIS están disponibles todo tipo de

herramientas, lo que le diferencia de ArcGIS, en cual no las tiene disponibles a no ser que sea con licencia.

4.1.3.4 Plugins

Una de las grandes ventajas que tiene QGIS es la biblioteca de plugin. Se trata de una amplia biblioteca de desarrollo libre. Por ejemplo, la instalación, procesamiento y post-procesamiento Landsat, Sentinel imagery, Semi-Automatic Classficatinn Plugin, etc. De hecho, QGIS permite construir un plugin propio.
En ArcGIS, hablamos de un add-in en lugar de un plugin. Es una personalización, como una colección de herramientas en la barra de herramientas, que se conecta a una aplicación ArcGIS for Desktop para proporcionar funcionalidad suplementaria para realizar determinadas tareas. Estos complementos se crean utilizando .NET, Java o Python. Hay soluciones pagadas (y gratuitas) para casi cualquier problema espacial que se pueda imaginar.
Aunque ArcGIS también tiene complementos, no se pueden comparar en número con el de QGIS, que tiene una gran cantidad de complementos gratuitos, sin embargo, no tiene la variedad de herramientas especializadas disponibles en ESRI.
La geocodificación es el proceso de asignar coordenadas geográficas a puntos del mapa, que luego se emplearán para localizar dicho punto en un GIS. Una opción para geocodificar es emplear ArcGIS Online Geocoding, servicio de pago que requiere créditos para su uso. Por su parte, ArcGIS Desktop permite añadir una ubicación mediante su barra de herramientas Geocoding.
Por otro lado, QGIS pone a nuestra disposición dos herramientas para geocodificación. En primer lugar, el plugin MMQGIS (permite agregar a la vista de mapa direcciones como puntos a partir de CSV y, en segundo lugar, el plugin GeoCoding que requiere una dirección como entrada.
Este inconveniente de los costos adicionales por el pago de créditos hace inclinar la balanza hacia QGIS.

4.1.3.5 Desarrollo

El proceso de desarrollo de QGIS es libre, en base a los estándares que se proponen para poder unificar código. Por lo general, el coste del desarrollo está sufragado por organizaciones no financieras, pero con el apoyo de las empresas comerciales que aportan donaciones al proyecto.
En cambio, el desarrollo de ArcGIS está hecho íntegramente por el equipo de desarrollo de Esri. Aunque las funcionalidades mejoradas que se agregan surgen como resultado de la valoración de los usuarios del software.

4.1.3.6 Model Builder

Model Builder de ArcGIS es la forma más intuitiva, sólida y pragmática de automatizar los trabajos de geoprocesamiento, ya que permite encadenar conjuntos de herramientas para automatizar procesos. Tiene dos iteradores para hacer bucles «for» y «while», también te da la posibilidad de exporta tu modelo y compartirlo con los demás, o exportarlo como un script de Python y personalizarlo.

Las secuencias de comandos de ArcGIS se ejecutan casi por completo a través del módulo Arcpy, el cual es muy fácil de usar, porque casi todas las herramientas en ArcGIS

tienen una herramienta de secuencias de comandos del mismo nombre ya creado (que se puede copiar y pegar fácilmente del sitio web de Esri).
Para QGIS tenemos PyQGIS, aunque no hay un módulo definitivo, lo que te proca que se acaben usando otros módulos para diferentes proyectos y puede ser difícil averiguar qué usar y dónde está todo.

4.1.3.7 Topología

La topología explica las representaciones espaciales entre las distintas funciones vectoriales en la misma capa de datos o capas de datos diferentes. Las reglas topológicas se utilizan para la detección de errores y corregir la digitalización.
La topología en ArcGIS es genial. Si tus datos GIS tienen errores (superposiciones, lagunas), ArcGIS admite comprobaciones de errores complejas con sus herramientas de topología. Puedes inspeccionar la topología con más de 30 reglas con el inspector de errores, y resolver problemas de topología con arreglos automáticos o manuales. QGIS ofrece un par de reglas para la topología: «debe contener», «no debe tener duplicados», «no debe tener lagunas», «no debe tener geometrías inválidas», «no debe tener geometrías de varias partes», «no debe superponerse» y «no debe superponerse con «. Eres tú quien valida su geometría basada en estas reglas. La fijación de topología de ArcGIS es interactiva. Uno a uno, puede pasar por errores y arreglarlos. La edición de topología es un punto fuerte en ArcGIS con un montón de opciones para corregir errores de edición.

4.1.3.8 Creación de simbología

ArcGIS tiene una simbología predeterminada muy potente (transporte, raíces, suelos, clima, etc.), es ideal para estilos de puntos, líneas y polígonos. En definitiva, la simbología existente en ArcMap es útil y abundante.
Por el contrario, QGIS no tiene una simbología preexistente tan atractiva. La vida sería más fácil en QGIS si estuviera equipada con simbología por defecto, aunque si te da la opción de descargarlos y cargarlos en tu paleta de simbología. También te da la opción de realizar mezclas con la simbología para aligerar pantalla, esquivar, añadir, oscurecer, multiplicar, quemar, superposición, luz suave, luz dura y la diferencia. Los rellenos degradados hacen de QGIS un paraíso de cartógrafos permitiéndote crear gradientes simples con dos o varios colores.

4.1.3.9 Etiquetas y anotaciones

Aunque ArcGIS carece de estas herramientas del etiquetado, si cuenta con un etiquetado avanzado que permite establecer la ubicación de la etiqueta y la dependencia de la escala. El etiquetado curvo y paralelo es fácil en ArcGIS. Es inteligente. La barra de herramientas de dibujo es cómo se controlan los grupos de anotación en ArcGIS. No es intuitivo. Pero con un poco de práctica se puede controlar a qué anotación pertenecen los grupos de etiquetas. Las etiquetas en QGIS permiten más versatilidad. Sin embargo, las propiedades de ubicación y la anotación son mejores en ArcGIS.

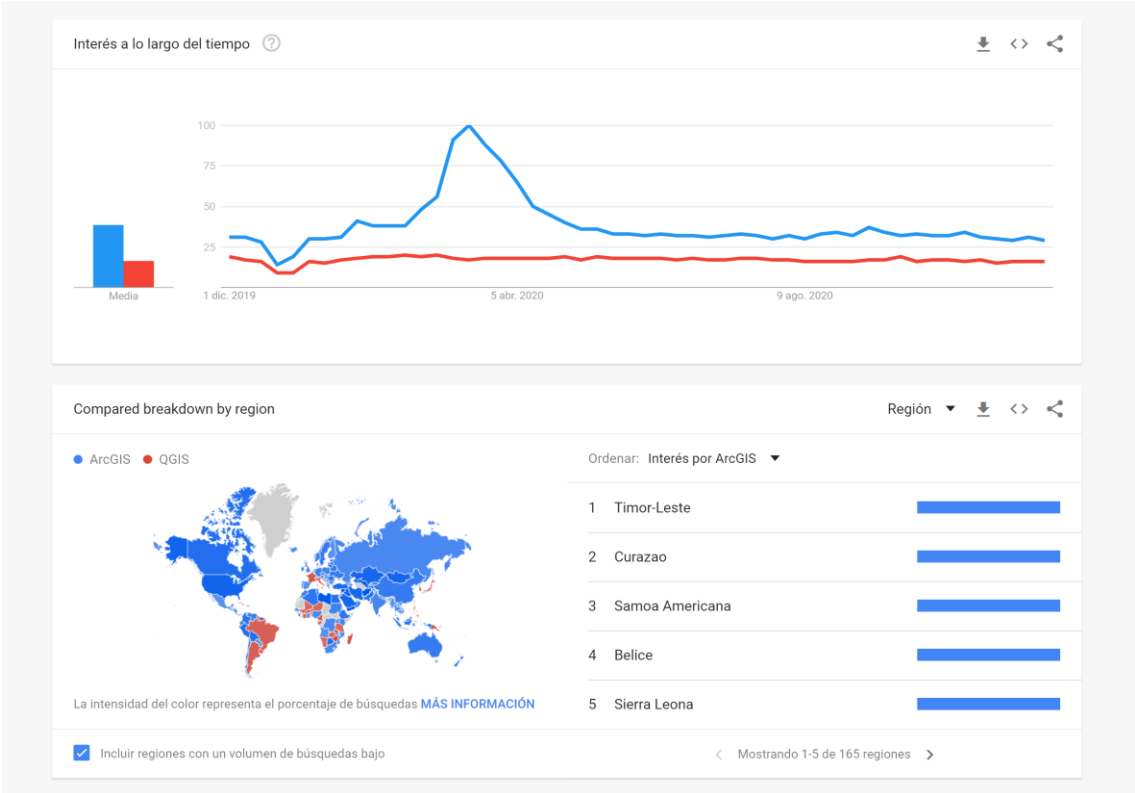
4.1.3.10 Diseño de mapas web

Los Web maps son tendencia y la industria de noticias, los gobiernos y las empresas están utilizando web maps porque son muy visuales e intuitivos. La asignación de páginas web es fácil en ArcGIS, los datos se envían a la web a través de ArcGIS Online,

que es donde se encuentran los mapas online para ESRI. Los Story Maps también son muy interesantes, los puedes aprovechar para contar una historia a través de mapas.
QGIS Server proporciona un servicio de mapas web (WMS) que utiliza las bibliotecas QGIS. Los mapas y plantillas de impresión creados en el escritorio QGIS se pueden publicar como mapas web simplemente copiando el archivo de proyecto QGIS en el directorio del servidor.

4.1.3.11 Resultados del estudio

Todos los softwares desarrollados anteriormente, tienen un alcance casi inimaginable, destacando a QGIS y ArcGIS. Hubo un largo periodo de investigación y aprendizaje de las nociones básicas de ambas tecnologías para poder enfocarlo.
Gracias a esta investigación nos dimos cuenta de que ArcGIS es la herramienta más potente del mercado, pero cuesta dinero y eso equilibró mucho la balanza a QGIS, pero la universidad de málaga nos proporcionó una licencia de estudiante para este software, por lo que finalmente disponíamos de la herramienta más potente del mercado a nuestro alcance (Más adelante de desarrollará este punto).
Como este es un proyecto que pretende ser ambicioso y llegar a ser útil a los mayores expertos, escogimos ArcGIS la herramienta GIS por excelencia y más utilizada en todo el planeta.



Ref. 42

5

Implementación

5.1 ArcGIS

En nuestro caso nos hemos centrado en dos tecnologías en concreto que son ArcGIS Desktop y ArcGIS Online, ya que la licencia de estudiante solo cubría ArcGIS Desktop y esta herramienta es un poco limitada a la hora de hacer una presentación más visible. Es por esto por lo que se ha usado una cuenta de desarrollador en ArcGIS Online, ya que nos permite poder hacer mapas de una manera mucho más vistosa y con más posibilidades a la hora de visualizar distintas capas. Pero, esta licencia tiene también una cierta cantidad de restricciones, ya que estamos hablando de licencias gratuitas, por lo que solo se nos otorgan unos 50 “créditos”, que se van consumiendo a medida que se desarrollan mapas en ArcGIS Online, es esta la razón por lo que hemos tenido que hacer un mix entre ambas tecnologías. Más adelante se explicará cómo se hizo este desarrollo.
Las ventajas del programa en línea incluyen el intercambio de contenidos tanto durante como fuera de su organización. Los grupos pueden acceder a mapas personales en una invitación, permitiendo la colaboración. Otras partes de la plataforma de software incluyen aplicaciones, como herramientas de navegación, recolección y topografía, así como un explorador rápido y herramientas de trabajo para el trabajo de campo coordinado.
Como muchos programas de GIS, ArcGIS crea mapas que requieren categorías organizadas en capas. Cada capa se registra espacialmente de modo que cuando se superponen una sobre otra, el programa las alinea adecuadamente para crear un mapa de datos complejos. La capa base es casi siempre un mapa geográfico, extraído de una gama de fuentes dependiendo de la visualización necesaria (satélite, mapa de carreteras, etc.). Este programa tiene muchos de ellos disponibles para los usuarios y también contiene capas de alimentación en vivo que incluyen detalles del tráfico.

Las tres primeras capas se denominan capas de características o vectoriales, y cada una de ellas contiene funciones individuales que se distinguen a través de la plataforma. Estas son:

- puntos (como puntos de referencia, edificios)
- líneas (como carreteras y otros esquemas 1D)
- polígonos (como la información política y el censo geográfico, llamados datos 2D)
- imágenes raster (una capa de vector base como una imagen aérea)

Los datos pueden correlacionarse con al menos una de estas capas espaciales y pueden ser tanto mapeados como analizados, ya sea a través de características como los cambios demográficos, o a través de tablas de datos.
Sin embargo, lo que diferencia a este método de sus competidores es la compleja plataforma a través de la cual se puede realizar el mapeo y los datos. Por lo tanto, es un programa de gran alcance sujeto a las últimas mejoras y actualizaciones. Actualmente está disponible en los escritorios de Microsoft Windows, aunque el programa en línea es accesible en muchos sistemas operativos. Dado que funciona como una plataforma, los usuarios no deberían vadear páginas de información y datos; se dispone de recursos para disminuir y extraer información específica de conjuntos de datos geográficos mucho más grandes. En resumen, es una solución integral para la gestión y el análisis de datos, filtrada a través de la construcción de mapas.

5.1.1 Cualidades de la herramienta

Esri está trabajando para hacer que el GIS sea más accesible a más personas, menos intimidante técnicamente, y por lo tanto una herramienta más poderosa que pueda ser utilizada por cualquiera y¸ no sólo por un grupo de expertos de élite. ArcGIS Desktop fue diseñado con esto en mente.
Primero, debemos hace un repaso a algunas de la cantidad de posibilidades que nos ofrece tanto ArcGIS Online, como ArcGIS Desktop:

- Crear datos geográficos con digitalización asistida.
- Dibujar y editar entidades en un mapa.
- Trabajar con dispositivos móviles actualizando los datos en tiempo real.
- Sintetizar datos de diferentes fuentes.
- Almacenar la información en una base de datos geográficos.
- Realizar operaciones de análisis espacial.
- Diseñas y calcular redes.
- Automatizar geo procesos.
- Crear visualizaciones de propiedades espaciales en 2D y 3D.
- Maquetar mapas y controlar la salida de datos.
- Publicar la información geográfica para que esté accesible para cualquier usuario.

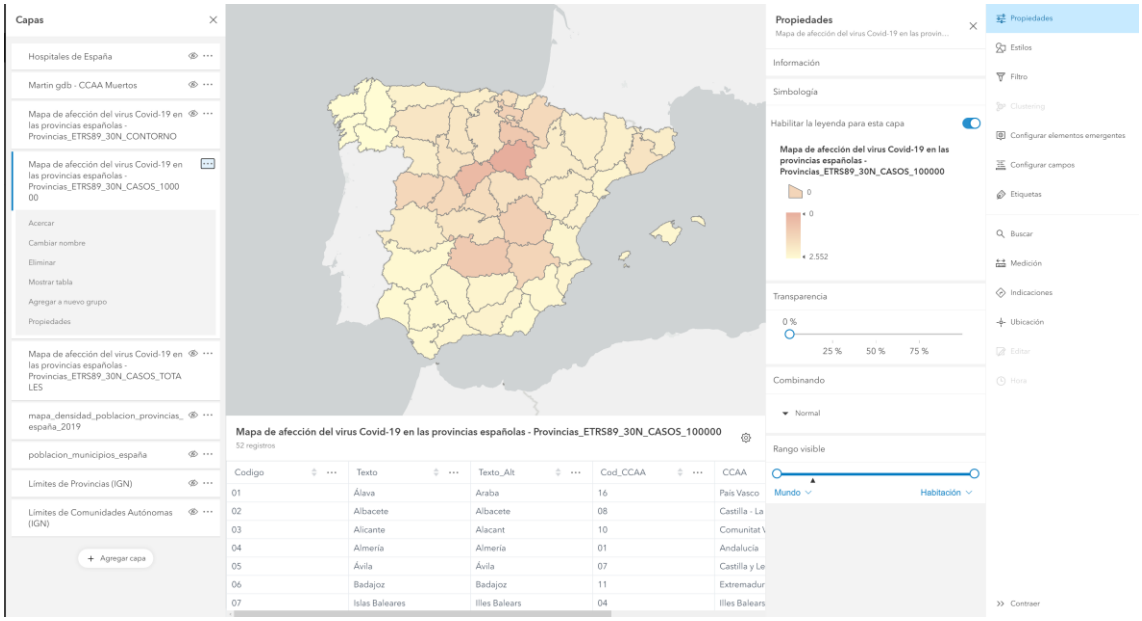
La pestaña de análisis es donde se encuentran las herramientas más usadas o personales, así como el botón "Herramientas", que abre el panel de geoprocesamiento. Allí podemos navegar por todas las cajas de herramientas, buscar herramientas y rellenar los parámetros para ejecutarlas.

Todo sucede en el panel de Geoprocesamiento que a medida que se configura una herramienta, no se tiene que hacer un seguimiento de múltiples ventanas emergentes o perder de vista la tabla de atributos o mapa como puede llegar a pasar en otras herramientas de geoprocesamiento similar. También se puede acceder al historial de geoprocesamiento desde esta pestaña, lo que permite revisar para refinar los flujos de trabajo y volver a ejecutar las herramientas con los mismos parámetros.
Es importante destacar que navegar por una interfaz de usuario que sigue la última tendencia del software hace que ArcGIS Online sea una interfaz navegable y mucho más amigable que ArcGIS Desktop, que aún no dispone de una interfaz tan amigable e incluso primitiva se podría decir.
En ArcGIS Online, no tienes que iniciar y detener el editor cada vez que quieras editar una tabla o una función. Esencialmente puedes editar en cualquier momento sin hacer nada especial o interferir con ninguna de las otras funciones del programa.
ArcGIS Pro nos permite compartir su mapa en línea con facilidad, ya sea que se quiera crear un mapa web o un mapa de la historia que se pueda ver públicamente, o que se necesite colaborar remotamente en un proyecto. Todo esto se hace a través de un portal que le conecta a ArcGIS Online. Este está diseñado para compartir, lo que va de la mano con el rápido aumento de la accesibilidad de GIS que mencioné anteriormente. Compartir sus datos en un formato interactivo y atractivo lo hace más poderoso. Utilizar los mapas compartidos para crear presentaciones o proyectos impresionantes para la clase o el trabajo.
Es muy difícil ser capaz de hacer atractivos los gráficos y tablas directamente, pero desde ArcGIS hace que tus proyectos y presentaciones sean mucho más fáciles. No es necesario que nos peleemos con tablas de atributos en Excel o R ya que si una visualización espacial de los datos (es decir, un mapa) no es suficiente, esta capacidad es enorme.
Utilizar favoritos del proyecto para ahorrar tiempo si se utiliza a menudo la misma carpeta, para que esto ocurra hay que haber usado un cierto número de veces las bases de datos y conexiones de servidor para que estos elementos favoritos estén disponibles en la pestaña Favoritos del panel Catálogo y en la vista Catálogo.
ArcGIS Online también permite tener muchas plantillas de proyecto que se pueden preconfigurar para tareas específicas como la edición o el análisis. Las plantillas de proyecto crean proyectos preconfigurados guardando las modificaciones de la interfaz, las cajas de herramientas y las conexiones necesarias para una tarea específica.

5.1.2 Descripción implantación en ArcGIS

En este proyecto, lo que se ha intentado conseguir una recreación del mapa de España, con los datos reales que se manejan hoy en día frente al COVID-19 y añadiendo los datos que hemos considerado como posibles factores que pueden ayudar a la propagación del virus y el aumento de los contagios a lo largo de estos meses que hemos sufrido.
Para comenzar vamos a ver una representación del mapa de España vacío que presta a sus usuarios ArcGIS Online.

Como puede verse en esta captura, existen varias capas con información detallada en tablas sobre los datos que hemos mencionado anteriormente. Cada capa tiene la información escrita de con la siguiente estructura.



Gracias a esta captura, se puede observar a simple vista, el potencial de la herramienta que nos presenta ArcGIS. El mapa a simple vista da mucha información y cualquier persona ajena a trabajar con esta tecnología puede navegar e interactuar con cada parte visible del mapa. El proceso de creación de cada mapa se detallará en los siguientes puntos.

Ahora, nos centraremos en el contenido de las tablas y su repercusión en el coloreado del mapa. Como se puede observar en la imagen siguiente tenemos tres tablas con el mismo nombre, cada tabla contiene:

1. Contagios Totales
2. Contagios cada 100mil habitantes
3. Contagios Activos

Esta información, ya de por sí es bastante útil por separado, pero ArcGIS Online, nos permite solaparlas y obtener de ellas un resultado, que aparentemente parecía distinto. Un ejemplo:

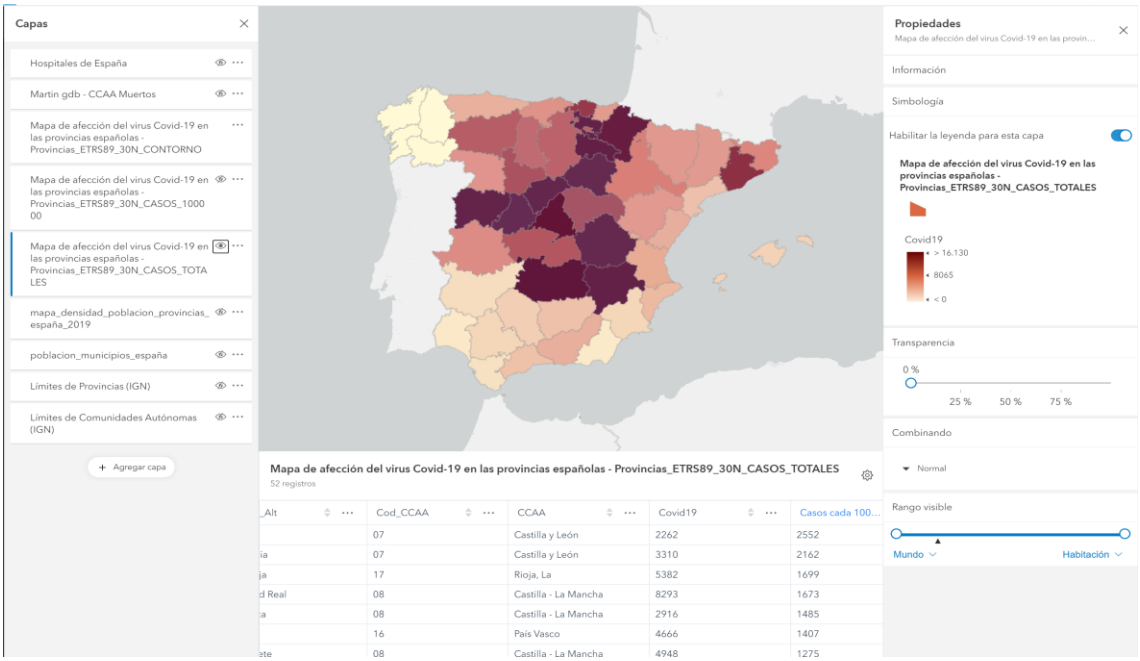


Ilustración 1

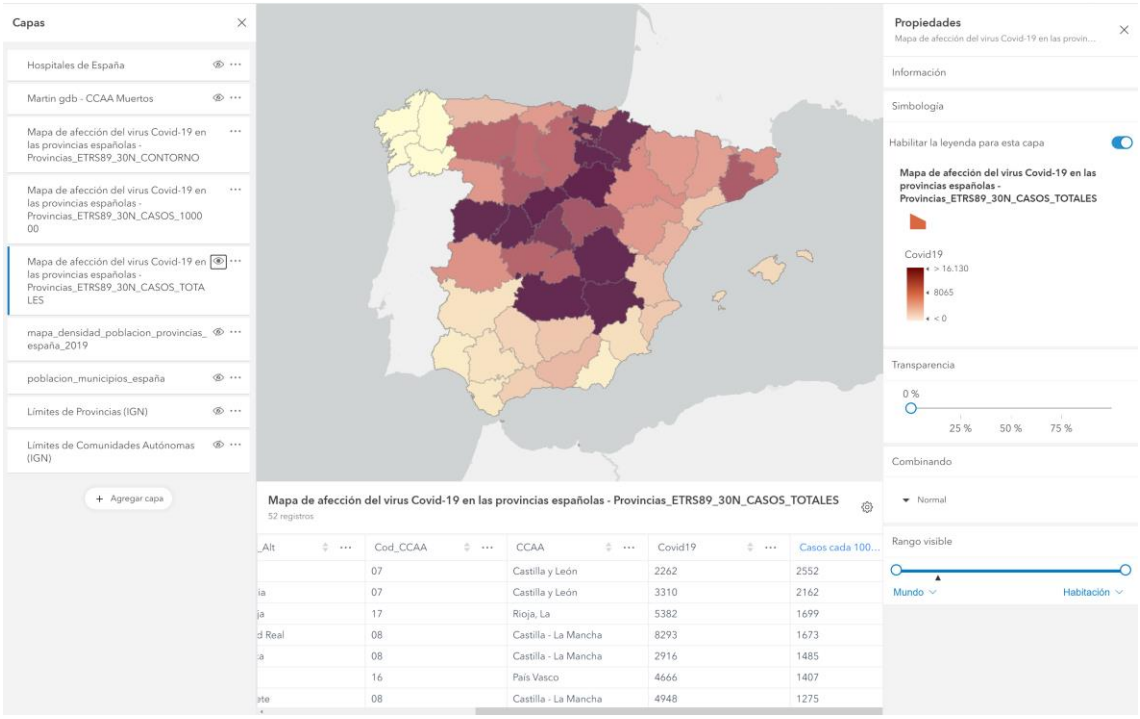


Ilustración 2

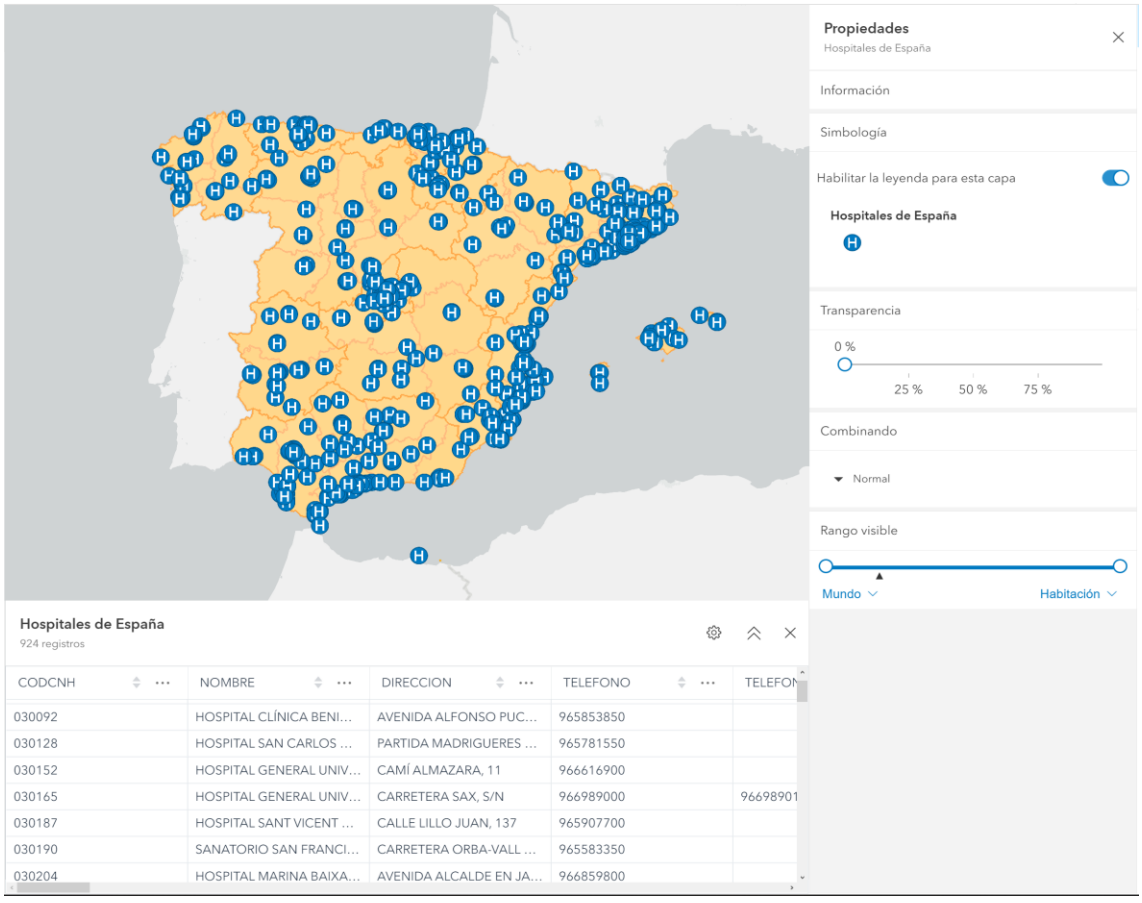
En estas dos capturas, se ha jugado un poco con las distintas capas de nuestro proyecto, para obtener una información muy interesante. Si nos fijamos en la captura (2) vemos

como Madrid, parece que no está en una grave situación ya que su color es pálido y no tan oscuro como puede llegar a estarlo otras comunidades cercanas a ella. Pero, sin embargo, cuando añadimos la otra capa (imagen 3) que añade la información de contagios activos, nos damos cuenta de que en este aspecto Madrid sí está en una situación grave, pero aun así podemos darnos cuenta de que no es la que peor está al añadir las siguientes capas a nuestro mapa base y que hay zonas de España que están peor ya que la relación de casos cada 100.000 habitantes es preocupante en muchas otras zonas que no son Madrid.

A continuación, se van a mostrar las capas restantes, para poder hacer un resumen de que queremos conseguir con dicha información:

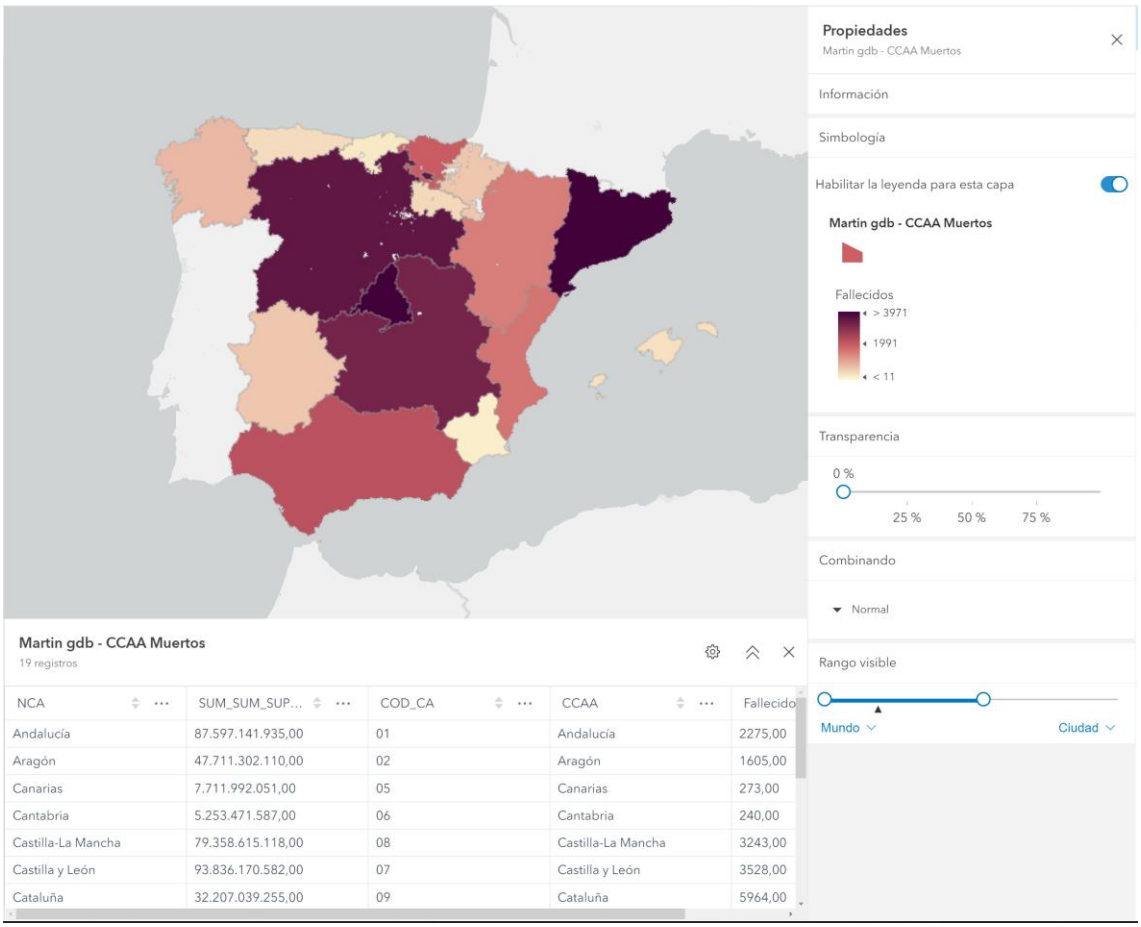
- Hospitales de España

Este mapa es interesante a la hora del estudio del número de muertos respecto al número de casos activos. También nos permite comprobar, las zonas que en caso de tener un alto número de casos, correrían riesgos de saturación en los hospitales.



- Muertos

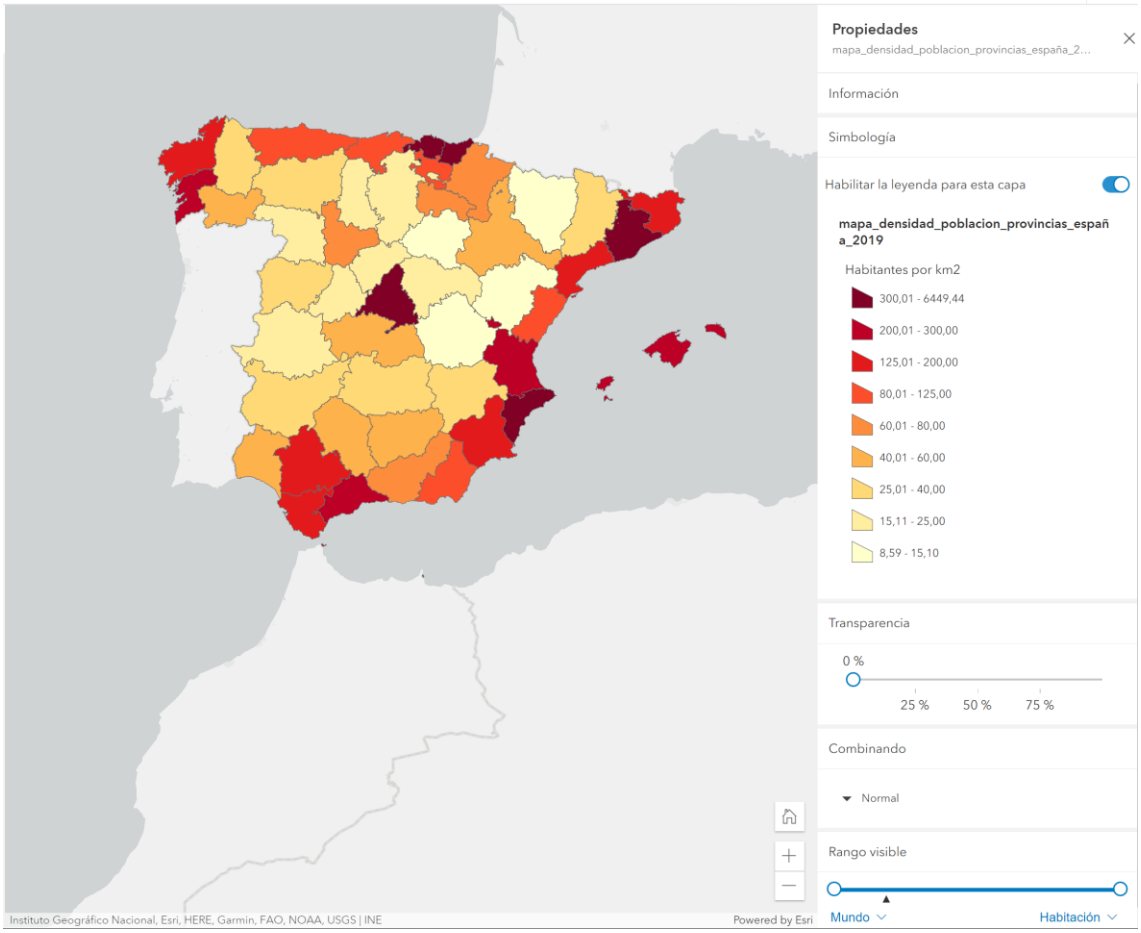
Estadística super importante teniendo en cuenta la situación crucial que se vive en España debido a lo realmente importante, que es el número de muertos que hay en este país. Y también motivo principal de este trabajo, que es que este número sea lo menor posible y brindar todas las herramientas posibles para ayudar a combatir esta pandemia.



NCA	SUM_SUM_SUP...	COD_CA	CCAA	Fallecido
Andalucía	87.597.141.935,00	01	Andalucía	2275,00
Aragón	47.711.302.110,00	02	Aragón	1605,00
Canarias	7.711.992.051,00	05	Canarias	273,00
Cantabria	5.253.471.587,00	06	Cantabria	240,00
Castilla-La Mancha	79.358.615.118,00	08	Castilla-La Mancha	3243,00
Castilla y León	93.836.170.582,00	07	Castilla y León	3528,00
Cataluña	32.207.039.255,00	09	Cataluña	5964,00

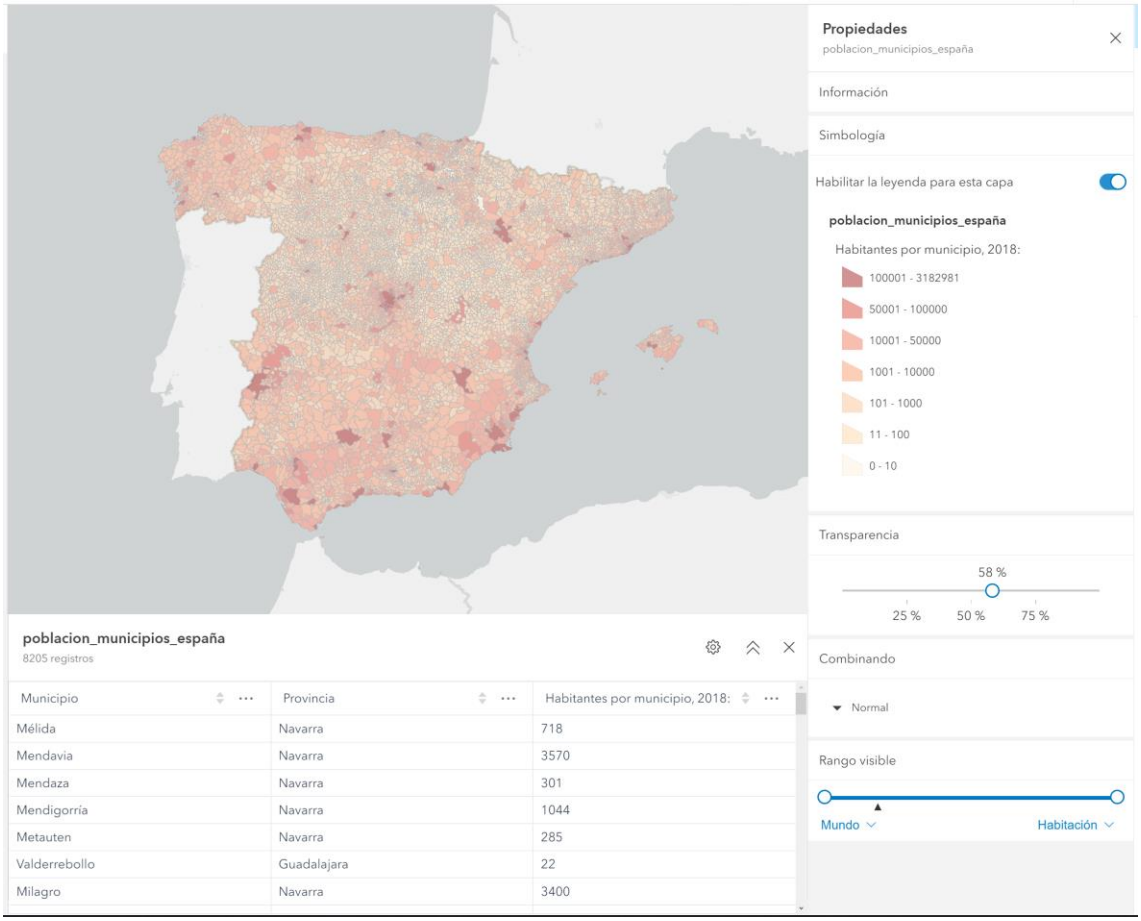
- Densidad de población por provincia

Estadística importante, para poder hacer predicciones en base a la relación número de casos cada 100.000 personas y ver si realmente los núcleos de población son los que reflejan más casos.



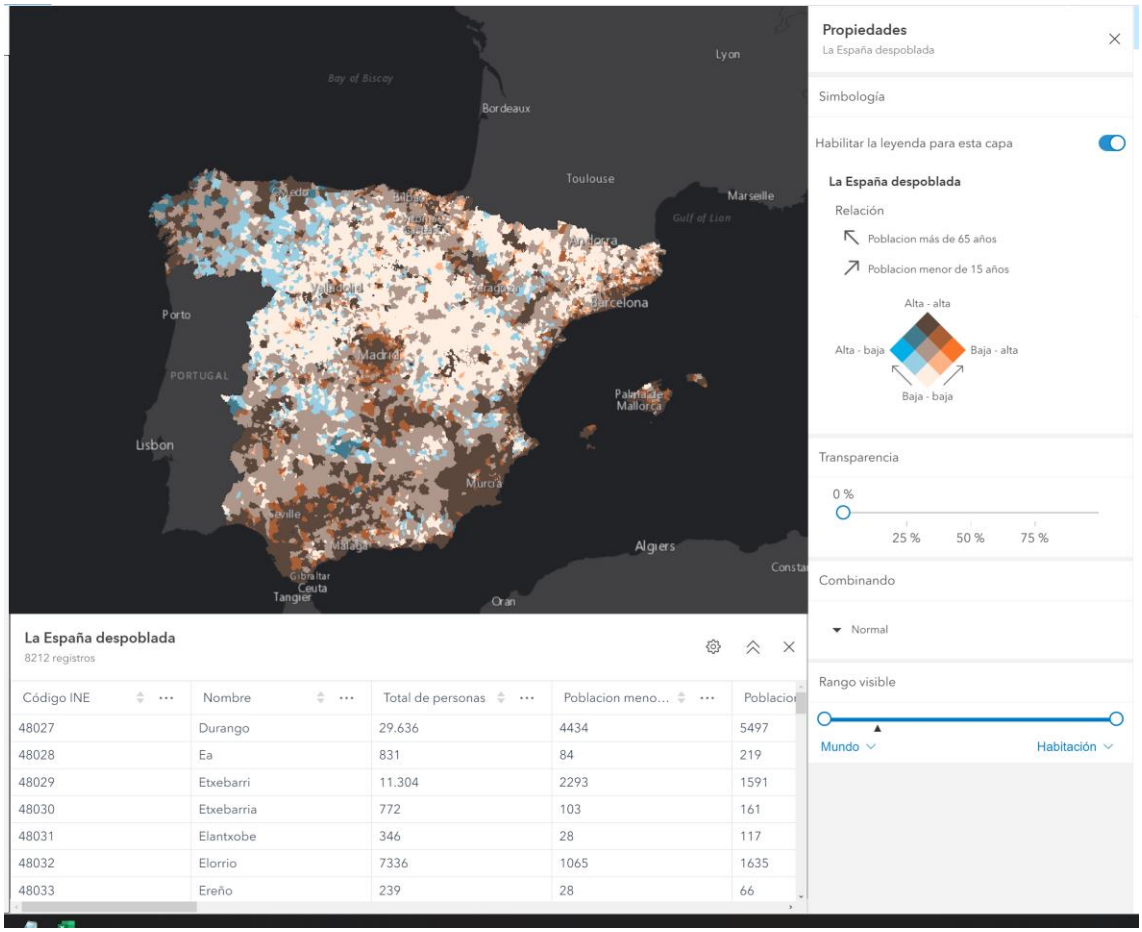
- Densidad de población por municipios

Exactamente la misma función que el mapa anterior pero a una escala mucho menor para hacer un estudio más profundo de los hechos.



- Relación de habitantes menores de 15 años y mayores de 65

Importantísimo dato y en este caso más, ya que como sabemos el COVI-19 no afecta de la misma manera a jóvenes y a adultos, por lo que puede ser combinado con otros mapas y reflejar un información muy relevante. También nos da información sobre las zonas despobladas de España y las zonas rurales.



Código INE	Nombre	Total de personas	Poblacion meno...	Poblacio
48027	Durango	29.636	4434	5497
48028	Ea	831	84	219
48029	Etxebarri	11.304	2293	1591
48030	Etxebarria	772	103	161
48031	Elantxobe	346	28	117
48032	Elorrio	7336	1065	1635
48033	Ereño	239	28	66

- Aeropuertos en España

Gracias a este mapa también podemos y relacionar la cantidad de aeropuertos con el aumento de casos de COVID-19 ya que se intentará relacionar un mayor número de casos a un mayor número de aeropuertos cercanos.

- Existen otros mapas que se han dejado sin explicación ya que estos ofrecen datos a nivel global, por si la persona que este observando quiere hacer una comparación entre países.

5.1.3 ¿Cómo se han procesado los mapas?

Viendo los resultados obtenidos, parece que es un proceso fácil, ya que es subir una tabla y poner las propiedades que quieres resaltar.
Nada más lejos de la realidad. ArcGIS funciona como una BD, ya que los datos y tablas que se suben tienen que estar relacionadas de alguna manera, ya que si no tendríamos datos puesto sobre el mapa de España sin sentido.
Para entender el proceso que se ha seguido vamos a mostrar los pasos para la creación de cualquier mapa y después vamos a detallar que cosas hemos tenido que ingeniar para poder sacar algunos resultados.

1º El primer paso cuando se va a desarrollar un mapa es definir una base, sobre la cual se va a hacer un desarrollo. En el caso de este estudio, es un mapa de España con las divisiones de las comunidades autónomas y las ciudades ya que en este estudio se va a tener datos a nivel de comunidad autónoma y provincial, por lo que nos es necesario dicha separación (Si se han fijado en detalle de los mapas anteriores, habrán visto que se detalla incluso a nivel municipal, pero eso se desarrollará más adelante). Pero hay que tener que existen muchos tipos diferentes tipos de mapas e incluso lo podríamos tener de relieve y con coordenadas geográficas, pero los datos del COVID-19 no se expresan de esa forma.

2º Lo siguiente que se ha realizado es una búsqueda de la información que se necesita ajustar a al diseño del mapa de España, como es bien sabido por cualquier persona que esté haciendo un estudio, este punto es bastante peligroso, ya que si no se obtienen los datos de fuentes fiables podemos correr el riesgo de enseñar al usuario información inválida y que lo podría llevar a observar datos y soluciones que no son adecuadas, por lo que es obligatorio sacarlo de fuentes oficiales del estado(y aun así tenemos que coger esos datos con pinzas, ya que cualquier informe puede contener erratas), por lo que siempre si se puede, hay que contrastar con varias fuentes fiables.

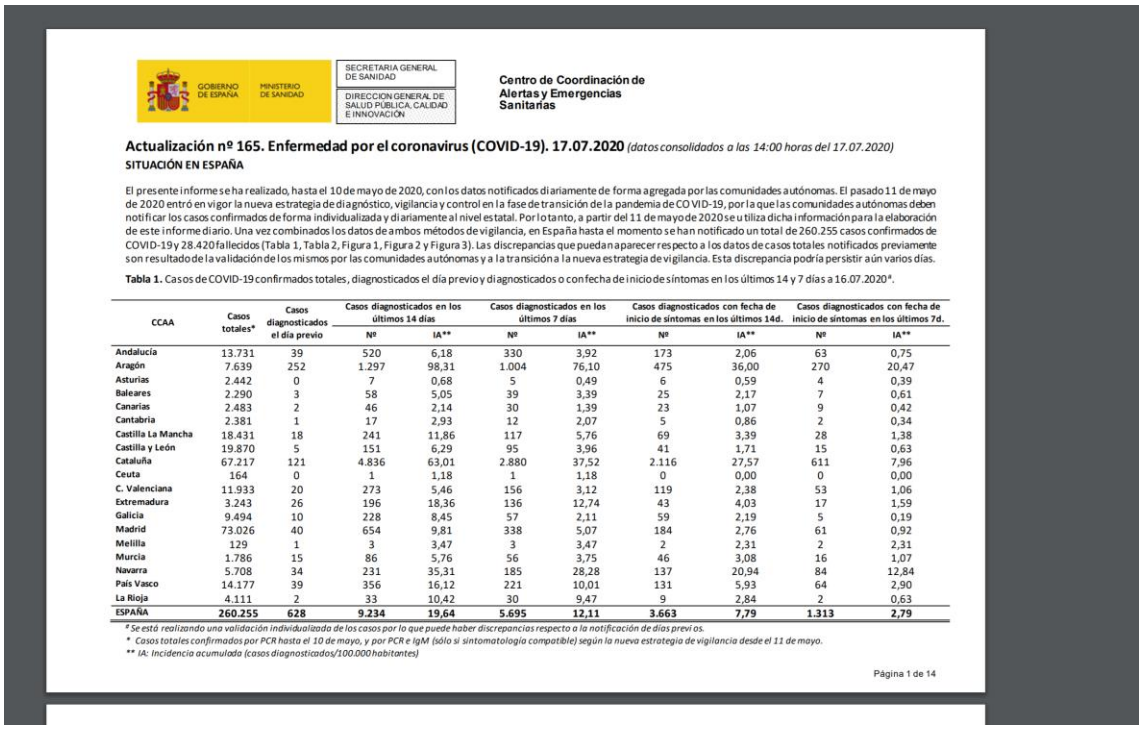
3º En este paso hablaremos del tratamiento de datos. Normalmente los datos vienen en .csv, pero este no siempre es el caso. Como hemos mencionado anteriormente no especificaré en esta descripción los pasos a seguir en casos específicos. Cuando uno busca información nacional, vienen tablas gigantescas en Excel, con todo tipo de información, nuestra misión en este apartado es segregar, ordenar, normalizar y desechar la duplicación de los datos., para así obtener los datos limpios y relevantes para el estudio del COVID-19 en España. Puede que el paso de "normalización" suene algo extraño, ya que parece que estamos hablando de una base de datos, pero es estrictamente necesario, ya que si por ejemplo queremos relacionar los mapas de Andalucía con Málaga y con el número de casos cada mil personas que tenemos en málaga, es

necesario que estos tres elementos estén relacionados y normalizados como si de una base de datos se tratase.

4º Por último, estos datos deben ser muy visuales, por lo que primero se deben seguir una serie de pasos (propios de ArcGIS), para poder añadir un mapa como una capa y después asignarle un tipo de visualización a los datos. Podemos hacerlo punto a punto, por colores, por formas, de calor… etc. Pero algunos son más sencillos que otros, por ejemplo, si quiero hacer una visualización por puntos tengo que seleccionarla y ya está, pero si hablamos de hacer un mapa de calor, ya es otra cosa, ya que tenemos que definir centroides y tener muchos datos muy juntos unos a otros y por todas partes del mapa sin que falte ninguno y que casi se sobrepongan unos a otros, ya que si falta algún valor el mapa de calor puede mostrar datos falsos, después de comprobar esto, simplemente tenemos que seleccionar los datos, hacerles un clusttering y crear el centroide que tendrá nuestro mapa.

Estos son los pasos que se deben seguir normalmente para la creación de un mapa, pero hay que tener en cuenta que estamos intentando obtener datos que afectan al gobierno actual y que son de tratamiento sensible. Sí, es posible conseguir casi cualquier dato que se quiera obtener del COVID-19, pero parece que los han puesto de tal manera que poca gente se quiera aventurar a obtenerlos. ¿Cómo se puede hacer difícil el trabajo de alguien que quiera obtener los datos al instante y representarlo? Muy fácil, poniéndolo en PDF.

Existen varias tablas de las que se han obtenido aquí que vienen en archivos del gobierno, que se mostrará a continuación.



GOBIERNO DE ESPAÑA · MINISTERIO DE SANIDAD · SECRETARIA GENERAL DE SANIDAD · DIRECCION GENERAL DE SALUD PÚBLICA, CALIDAD E INNOVACIÓN

Centro de Coordinación de Alertas y Emergencias Sanitarias

Actualización nº 165. Enfermedad por el coronavirus (COVID-19). 17.07.2020 *(datos consolidados a las 14:00 horas del 17.07.2020)*

SITUACIÓN EN ESPAÑA

El presente informe se ha realizado, hasta el 10 de mayo de 2020, con los datos notificados diariamente de forma agregada por las comunidades autónomas. El pasado 11 de mayo de 2020 entró en vigor la nueva estrategia de diagnóstico, vigilancia y control en la fase de transición de la pandemia de COVID-19, por la que las comunidades autónomas deben notificar los casos confirmados de forma individualizada y diariamente al nivel estatal. Por lo tanto, a partir del 11 de mayo de 2020 se utiliza dicha información para la elaboración de este informe diario. Una vez combinados los datos de ambos métodos de vigilancia, en España hasta el momento se han notificado un total de 260.255 casos confirmados de COVID-19 y 28.420 fallecidos (Tabla 1, Tabla 2, Figura 1, Figura 2 y Figura 3). Las discrepancias que puedan aparecer respecto a los datos de casos totales notificados previamente son resultado de la validación de los mismos por las comunidades autónomas y a la transición a la nueva estrategia de vigilancia. Esta discrepancia podría persistir aún varios días.

Tabla 1. Casos de COVID-19 confirmados totales, diagnosticados el día previo y diagnosticados o con fecha de inicio de síntomas en los últimos 14 y 7 días a 16.07.2020[¥].

CCAA	Casos totales*	Casos diagnosticados el día previo	Casos diagnosticados en los últimos 14 días		Casos diagnosticados en los últimos 7 días		Casos diagnosticados con fecha de inicio de síntomas en los últimos 14d.		Casos diagnosticados con fecha de inicio de síntomas en los últimos 7d.	
			Nº	IA**	Nº	IA**	Nº	IA**	Nº	IA**
Andalucía	13.731	39	520	6,18	330	3,92	173	2,06	63	0,75
Aragón	7.639	252	1.297	98,31	1.004	76,10	475	36,00	270	20,47
Asturias	2.442	0	7	0,68	5	0,49	6	0,59	4	0,39
Baleares	2.290	3	58	5,05	39	3,39	25	2,17	7	0,61
Canarias	2.483	2	46	2,14	30	1,39	23	1,07	9	0,42
Cantabria	2.381	1	17	2,93	12	2,07	5	0,86	2	0,34
Castilla La Mancha	18.431	18	241	11,86	117	5,76	69	3,39	28	1,38
Castilla y León	19.870	5	151	6,29	95	3,96	41	1,71	15	0,63
Cataluña	67.217	121	4.836	63,01	2.880	37,52	2.116	27,57	611	7,96
Ceuta	164	0	1	1,18	1	1,18	0	0,00	0	0,00
C. Valenciana	11.933	20	273	5,46	156	3,12	119	2,38	53	1,06
Extremadura	3.243	26	196	18,36	136	12,74	43	4,03	17	1,59
Galicia	9.494	10	228	8,45	57	2,11	59	2,19	5	0,19
Madrid	73.026	40	654	9,81	338	5,07	184	2,76	61	0,92
Melilla	129	1	3	3,47	3	3,47	2	2,31	2	2,31
Murcia	1.786	15	86	5,76	56	3,75	46	3,08	16	1,07
Navarra	5.708	34	231	35,31	185	28,28	137	20,94	84	12,84
País Vasco	14.177	39	356	16,12	221	10,01	131	5,93	64	2,90
La Rioja	4.111	2	33	10,42	30	9,47	9	2,84	2	0,63
ESPAÑA	**260.255**	**628**	**9.234**	**19,64**	**5.695**	**12,11**	**3.663**	**7,79**	**1.313**	**2,79**

[¥] Se está realizando una validación individualizada de los casos por lo que puede haber discrepancias respecto a la notificación de días previos.

** Casos totales confirmados por PCR hasta el 10 de mayo, y por PCR e IgM (sólo si sintomatología compatible) según la nueva estrategia de vigilancia desde el 11 de mayo.*

*** IA: Incidencia acumulada (casos diagnosticados/100.000 habitantes)*

Página 1 de 14

Ref. 43

Esta imagen pertenece al informe diario que da el gobierno a diario en su reporte del medio día, como se puede apreciar es una tabla con unos valores bastante interesantes a la hora del estudio del COVID-19 en España. Pero, ArcGIS no funciona con archivos pdf, solo con archivos soportados por Excel.

Para solventar dicha carencia por parte del gobierno, en un principio hemos tenido que crear una macro de Excel, que pase dicho archivo de pdf a Excel, ya que, si alguien quiere obtener dichos datos de manera actualizada, tiene que utilizar esta macro. Esto se hizo para las muertes que había por comunidad autónoma, pero como no era la única tabla y las macros llevan bastante tiempo, se acabó usando la aplicación pdfaid que pasa de pdf a .txt y después se pueden coger las tablas y pasarlas a pdf sin problemas a un archivo .csv y ya ahí empezar el método de ajuste.

5.2 Pytorch

El desarrollo en Pytorch es una de las partes más importantes de este proyecto, para la realización de este cogido se ha tomado como fuente de información el código creado previamente por un grupo de ingenieros[].

5.2.1 Descarga de datos

Como base de datos para COVID-19 en nuestro análisis estamos usando el repositorio de Instituto de Salud Carlos III. Estos datos se almacenan en la Carpeta COVID-19, donde replicamos el contenido de los informes de las series temporales. Los datos en sí son descargados por el script main.py cada vez que se ejecuta, lo que provoca que solo se guarde una copia en el repositorio.
Como datos de densidad de población estamos buscando los últimos datos disponibles en la revisión de la población en España. Como estos datos son anuales, los mantenemos almacenados en el repositorio sin ningún cambio en la Carpeta de Densidad de Población Bruta.
Como máscaras los datos de uso han creado una lista hecha a mano usando conocimiento de dominio público.
A medida que los gobiernos contrarrestan los datos, los hemos recopilado de diferentes fuentes. Dado que esta es la parte más subjetiva, no ha habido ninguna transformación de datos, sino directamente características que estarán en la siguiente sección.
Los datos de COVID-19 han sido procesados y guardados desde la Carpeta de COVID-19 a la Carpeta de COVID-19 de.
Los datos de densidad de población han sido procesados y guardados desde la Carpeta de Densidad de Población Bruta a la Carpeta de Densidad de Población de los Accidentes.
El uso de máscaras ha sido procesado y guardado desde la carpeta de uso de máscaras crudas a la carpeta de uso de máscaras de características.
El riesgo de población ponderado que hemos calculado también ha sido procesado y guardado desde la Carpeta de Riesgo de Población Bruta a la Carpeta de Riesgo de Población de Características.
Las medidas de los gobiernos que hemos recogido como se explica en la historia de Medium, han sido copiadas directamente en la Carpeta de Medidas de los Gobiernos.
Lo que su código fuente hacía era tomar datos de otros repositorios de GitHub y transformaban un csv a u panda data Frame, lo que nos permite utilizar Pytorch más adelante.

5.2.2 Objetivos con este código

Pero como puede haber pensado con seguridad, esto no es suficiente. No podemos comparar o correlacionar la tasa de propagación o la cantidad de casos confirmados en diferentes tipos de ambiente tan fácilmente. Debemos tener en cuenta principalmente los siguientes factores:
La densidad de población: Los casos de COVID-19 no van a crecer al mismo ritmo si hay 1 persona cada km² en lugar de mil, ¿verdad?
El uso de la máscara a diario. Tenga en cuenta que estamos suponiendo aquí que tendrá un impacto en los casos, pero no sabemos cuánto, eso lo decidirán tanto los datos como el modelo.

Grupo étnico: especialmente, asiático. Esto está relacionado con la presencia de ACE2 en nuestras células, como dice la ECDC.
Las contramedidas tomadas por los gobiernos.
Siendo el último factor una secuencia de diferentes medidas que estos gobiernos son capaces de tomar, y cuando lo hicieron aplicar estas medidas. Mientras que en los casos de COVID-19 utilizamos fuentes externas, los otros han sido obtenidos directamente por nosotros de la siguiente manera:
Densidad de población: hemos tomado los últimos datos disponibles en el World Population Review para cada lugar necesario.
Uso de máscaras: para los occidentales es raro ver a gente usando máscaras en la vida diaria, pero en algunos países orientales como Japón su uso es algo común que puede ayudar a frenar la propagación del virus. Las fuentes sobre este asunto se pueden encontrar en nuestro Léame de Datos de Máscaras específico en el repositorio.
Riesgo poblacional (Grupo étnico): hemos realizado una clasificación basada en la siguiente imagen (a decir verdad, no conozco la fuente específica, es decir, el conjunto de datos base o publicación donde se publicó originalmente como lo pongo aquí), que está elaborada a partir del Proyecto 1000 Genomas como se especifica en la leyenda al final del mismo.
Contramedidas: estos datos han sido elaborados por nosotros mismos en base a las medidas tomadas por los gobiernos para frenar la propagación. Hay que tener en cuenta que debido a la falta de datos para la formación, no deberíamos tener una gran cantidad de características, o los modelos no podrán inferir cómo cada una de ellas afecta a la propagación del virus.
La lista de contramedidas es la siguiente, y encontrar en nuestro repositorio Raw Government Measures Readme las fuentes de donde hemos encontrado esta información tanto para Lockdown como para las fronteras.

- Confinamiento en casa (Lockdown)
- Las fronteras del país están cerradas

Los posibles valores de estas medidas se discutirán en la sección Definición de características.

5.2.3 Preprocesamiento de datos

Antes de utilizar los datos que hemos reunido, es importante que tanto el proceso de regularización como el algoritmo de ML tengan los datos más limpios posibles. Esto implica que es necesario tener en cuenta algunas consideraciones.
Los casos de COVID-19 son variables entero para cada ubicación, pero incluso siendo valores enteros, sus posibles valores son tan amplios que cualquier sistema ML encontraría extremadamente difícil de generalizar en estas condiciones, por lo que necesitamos aplicar algunas restricciones a estos valores. Si imaginamos un país con 100 millones de ciudadanos, podríamos tener cualquier número de casos confirmados entre 0 y 100 millones. Eso no se puede manejar para cualquier sistema de machine learning, al menos si no tenemos suficientes muestras de entrenamiento. Y estamos tratando un nuevo virus, así que por supuesto, no lo tenemos. Para resolverlo, vamos a discretizar estos valores:

- Los casos de COVID-19.
- Para los valores de densidad de población, vamos a hacerlos también un múltiplo de 5 (el valor en ciudadanos/km²). De esta manera reducirá el rango de posibilidades y facilitará la convergencia de los modelos. Básicamente pretendemos corregir algunos datos como por ejemplo la densidad de población española, ya que la mayoría de la población vive en un área reducida frente a la superficie total del país. Este tipo de corrección se aplicará como un factor x3 sobre la densidad de población original dada, debido al enfoque de la población en una fracción de la tierra del país. Finalmente vamos a reducir la característica aún más. Nos proponemos establecer tres niveles de densidad: baja, media y alta. Diseñaremos el umbral apropiado para cada grupo y luego usaremos como característica sólo una de estas tres categorías para cada lugar o país.
- Las máscaras serán un valor verdadero/falso por país.
- Los datos de riesgo de la población sufrirán un enfoque similar al aplicado a la densidad de población (tres niveles).

Mediante la combinación de estos modificadores, aprovecharemos nuestros resultados facilitando que nuestros algoritmos de ML encuentren patrones comunes entre nuestro reducido conjunto de datos de entrenamiento. Por último, no estamos hablando de Contramedidas Gubernamentales porque las hemos convertido directamente en características (CSV hecho a mano para las características).

5.2.4 Definición de las características

Basándonos en los puntos anteriores, lo único que queda es especificar cómo se alimentarán estos rasgos en nuestros algoritmos. Necesitamos en un proceso para calcularlos y aplicando una normalización a cada característica al final, por lo que los valores están entre 0 y 1. Vamos a ir uno por uno:

1. Los casos de COVID-19. Esta es la única característica que no sufrirá una normalización entre 0 y 1, sino una normalización logarítmica. De esta manera podemos evitar problemas de gradiente que desaparecen debido a que los casos bajos están demasiado cerca de 0. Recuerde que esta característica se procesa así, pero la salida esperada se trataría de igual manera ya que es la misma unidad/magnitud.
2. Densidad de población: clasificada como baja/media/alta. Está clasificación no está puesta al azar, más adelante se desarrollará el porqué de este proceso.
3. Hemos tomado la densidad de población, como se especifica en las secciones anteriores, de los datos en bruto almacenados en un archivo CSV.
4. Hacemos que estos valores sean múltiplos de 5, y aplicamos las desventajas (en realidad sólo para España, ya que su vasta tierra no utilizada distorsiona la densidad de población en las zonas pobladas).
5. Finalmente aplicamos una transformación a esta clasificación: la densidad de población es un valor continuo que puede hacer casi imposible que el modelo aprenda su efecto con nuestro muy limitado conjunto de datos de entrenamiento. Lo simplificamos agrupándolos en tres categorías: baja (1), media (2) o alta (3) densidad.

6. ¿Pero cómo poner la frontera? No queremos poner a mano el límite de cada categoría, así que estamos aplicando aquí nuestro primer algoritmo: K-Means. Recordemos que esta técnica tratará de separar nuestros datos en cúmulos de K, así que lo estamos aplicando para K=3.

Esta agrupación da como resultado una clasificación para cada ciudad con sus valores desde 1 la densidad más baja, hasta 3 la más alta.

- Máscaras: clasificadas como no utilizadas (0) o utilizadas (1).
- Riesgo poblacional (Grupo étnico): clasificado como bajo (1), medio (2) o alto (3).

Hemos diseñado el siguiente proceso para clasificar la población de cada ciudad en un grupo:

- Calcular el riesgo ponderado para cada país, basado en la imagen anterior del Proyecto de los 1000 Genomas. Para las ciudades con mucha gente extranjera como Madrid, donde viven muchos inmigrantes chinos, los hemos tenido en cuenta.
- Si no hay datos sobre un país que presente casos, hemos fijado su valor de riesgo en un 50% como valor mínimo, lo que se puede observar en la tabla para los que son caucásicos. Estos dos primeros pasos se han realizado para obtener un riesgo ponderado en una hoja de Excel (nótese que en la carpeta de datos Raw del enlace hay dos archivos: el propio libro de Excel con el cálculo, y el archivo CSV con los resultados que se utilizarían en los siguientes pasos).
- Finalmente aplicamos una transformación a esta clasificación: los porcentajes son valores continuos que pueden hacer casi imposible que el modelo aprenda su efecto con nuestro muy limitado conjunto de datos de entrenamiento. Lo simplificamos agrupándolos en tres categorías: riesgo bajo (1), medio (2) o alto (3) debido al riesgo ponderado.
- Pero ¿a dónde pertenece la población de cada ciudad? No queremos poner a mano el límite de cada categoría, así que estamos aplicando de nuevo el K-Means. Recordemos que esta técnica tratará de separar nuestros datos en grupos K, así que lo estamos aplicando para K=3.

Esta agrupación da como resultado una clasificación para cada país que va de 1 el riesgo más bajo a 3 el más alto.

5.2.5 Clasificación por países - Descripción general

Por último, tenemos como características las medidas del gobierno. No se han explicado en la sección anterior de Procesamiento de Datos ya que lo hemos recogido y etiquetado en niveles directamente.

- Bloqueo: nivel de no bloqueo (0) a bloqueo estricto (2), como se hace en Madrid.
- Cierre de fronteras: nivel desde ninguna aplicación de la contramedida (0) hasta el cierre completo del país o hacerlo extremadamente severo (2), como se hace en España o Francia.
- Recuerde que después de todo esto viene una normalización de las características por lo que todos los valores están entre 0 y 1, excepto para los casos que serán normalizados logarítmicamente.

5.2.6 Datos utilizados para el entrenamiento

Después de todo lo que hemos explicado, tenemos que especificar brevemente qué datos vamos a utilizar para entrenar nuestro sistema. Sí, incluso antes de decir qué modelo o técnica de aprendizaje automático vamos a utilizar. Puede que se pregunte... ¿Por qué?

Bueno, teniendo en cuenta que es un tema de actualidad (la cuestión de COVID-19 en su conjunto), por lo que los datos se actualizan todos los días, añadiendo nuevos países afectados y así sucesivamente. Si fuéramos a entrenar usando los últimos datos, necesitaríamos también actualizar todos los datos relacionados: actualización de los datos de las ciudades que implican más densidad de población para recuperar y añadir a nuestras características, continuar el seguimiento de la aplicación de las contramedidas gubernamentales, y así sucesivamente. Incluso el análisis de la clasificación de las características debería hacerse de nuevo

Ahora que hemos explicado todo esto podemos decir que vamos a tomar como datos desde el principio de los informes (22/01/2020), hasta el 31 de marzo (31/03/2020).

Hay que tener en cuenta que hay ciudades en el conjunto de datos que empezaron a tener casos confirmados más tarde que otros. Eso se traduce en un grupo de ciudades que tienen un grupo de días seguidos con un valor de 0 casos (puedes comprobarlo en el repositorio de origen del que tomamos los datos). Tenemos que eliminarlos de nuestras muestras. Si no, el sistema no aprenderá correctamente, ya que verá tanto ejemplos que dicen que con una cantidad de 0 casos un día al día siguiente volverá a ser 0, como otros ejemplos que le dirán a la red que con una cantidad de 0 un día, al día siguiente sube (punto de partida de los casos confirmados en este país). En resumen: considerar como primera muestra para cada ciudad el primer día con los casos de COVID-19.

5.2.7 El modelo y el marco

Comencemos por definir lo que será una muestra para la red. En las secciones anteriores hemos revisado las diferentes características y sus posibles valores, ¿verdad? Bueno, ahora es el momento de reunir todos estos datos.

Como entrada para nuestra red estamos usando el número de casos de ese día específico, como pueden ver, estamos construyendo nuestro modelo sólo para los casos confirmados, nada de muertes ni de casos recuperados. Tiene sentido ya que estos casos no podrían seguir el mismo comportamiento que el primero, por lo que no deberíamos mezclarlos.

$$[Cases_d, Popden, Masks, Poprisk, Lockdown_d, Borders_d]_c$$

Donde el subíndice d representa el día específico, y la c un país específico. Pero estamos haciendo un aprendizaje supervisado aquí, así que sabemos que necesitamos entrenar nuestro modelo dando un ejemplo con el resultado deseado. Entonces, ¿cuál es el resultado de esta muestra? La predicción que esperamos son los casos confirmados, pero del día siguiente.

$$[Cases_d, Popden, Masks, Poprisk, Lockdown_d, Borders_d]_c \rightarrow Cases_{d+1}$$

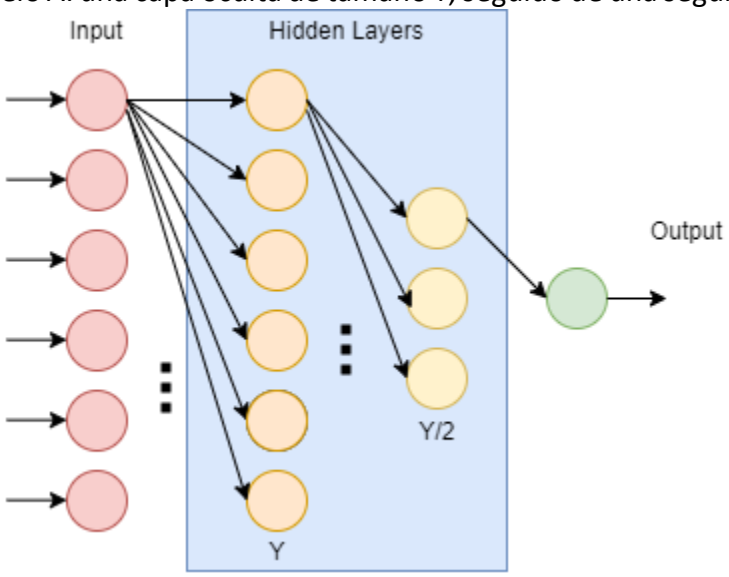
Al notar que tenemos 6 entradas, el primer parámetro de nuestro modelo es claro. Ahora es el momento de elegir qué tipo de modelo estamos construyendo
Sabemos que los modelos estadísticos para las epidemias tienen la forma de una función gaussiana en caso de casos activos. Aquí vamos a trabajar con los casos confirmados, que es la cantidad total de casos, no restamos los casos recuperados ni los muertos. Para esta métrica, sigue una forma sigmoide: los casos normalizados pueden ir de 0 a 1, lo que significa que se ha informado de que toda la población ha sufrido el virus. De esta manera podemos tener un enfoque aproximado de la complejidad que nuestro modelo va a tener que resolver, por lo que es un buen punto de partida para el diseño del modelo. Apilando suficientes (y no más, debido a los pocos datos disponibles y evitando el sobreajuste) capas ocultas con un tamaño apropiado, nuestro problema tiene algunas grandes posibilidades de ser resuelto o aproximado con suficiente precisión.
A partir de esta base hemos construido varias redes, y las hemos entrenado con diferentes hiperparámetros buscando una mejor predicción. Al final, como elección para la función de activación, LeakyReLU se ha aplicado a todas las capas ocultas. Como función de pérdida hemos utilizado el error medio cuadrado para penalizar los errores más grandes.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

MSE Error Formula — From Wikipedia

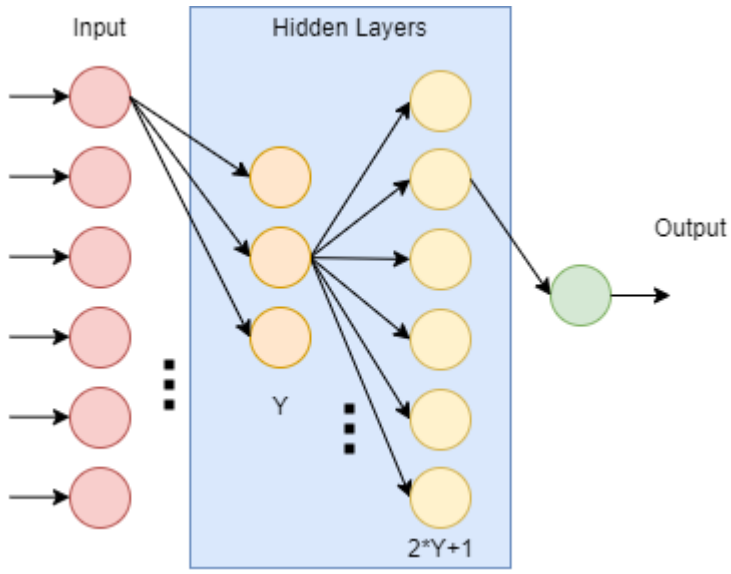
Se implementará en PyTorch usando MSELoss y aplicando una reducción media para cada lote. Para aquellos que necesiten una rápida revisión de algunas posibles métricas, aquí tienen una rápida referencia en los documentos de Microsoft.
Hemos estudiado tres modelos y buscamos el tamaño correcto de ellos en nuestro bucle de entrenamiento que veremos más adelante:

Modelo A: una capa oculta de tamaño Y, seguido de una segunda capa oculta de tamaño Y/2.



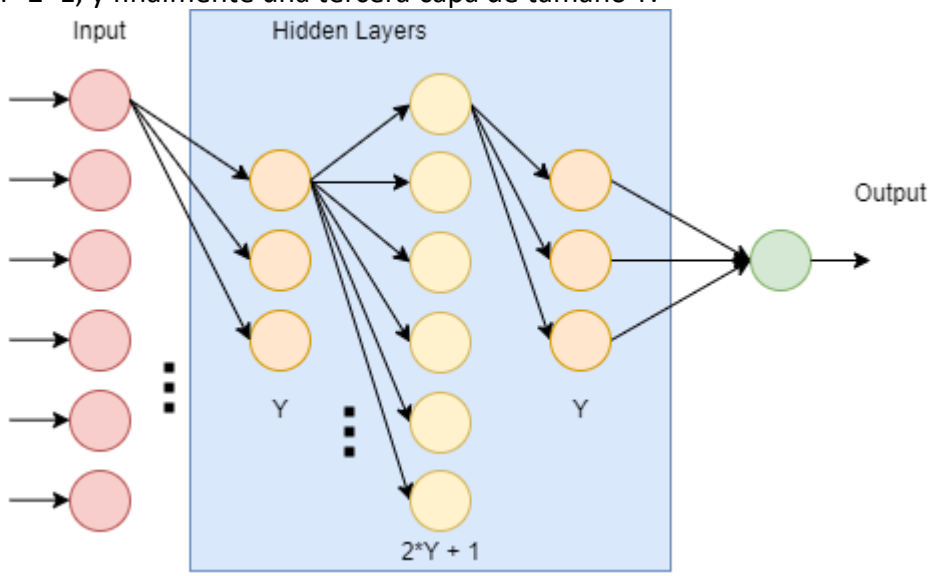
Esquema del Modelo A

Modelo B: una capa oculta de tamaño Y, seguida de una segunda capa oculta de tamaño Y*2+1.



Esquema del Modelo B

Modelo C: una capa oculta de tamaño Y, seguida de una segunda capa oculta de tamaño Y*2+1, y finalmente una tercera capa de tamaño Y.



Esquema del Modelo C

Y, por último, pero no menos importante, noten que vamos a implementarlo usando Pytorch, lo que nos permitirá (entre otros beneficios) tener un buen rendimiento en la CPU. No es realmente algo de lo que preocuparse con los pocos datos disponibles, ¿verdad? Pero recuerden que todavía queremos implementar algún tipo de bucle de entrenamiento.

5.2.8 Entrenamiento

Ahora que tenemos todo definido, sólo queda el entrenamiento para lograr nuestros objetivos. Así que aquí vamos a especificar nuestros parámetros de entrenamiento:

- Conjunto de datos de entrenamiento: se están tomando el 90% de los datos disponibles para el entrenamiento. Es decir: 90% de lo que tenemos desde el principio de los datos hasta el 31 de marzo (después de la limpieza de preprocesamiento).

Muestras utilizables: 4395
Muestras hasta el 31 de marzo: 3278
Muestras después del 31 de marzo: 1117
Tasa de aprendizaje: 0.00001
Lote: estamos tomando 48 como tamaño de lote.
Época: revisaremos todo el conjunto de datos de entrenamiento 50 veces.
Tamaño de la capa oculta: sí, sabemos que justo arriba dijimos que iba a ser un modelo secuencial, con 6 y luego 3 neuronas, pero como no sabemos si es la mejor opción luego se escogerán 12 (luego 6 en la segunda capa) sólo para empezar.
Con esta configuración, podemos realizar un entrenamiento regular que resultaría en algo como la siguiente imagen:

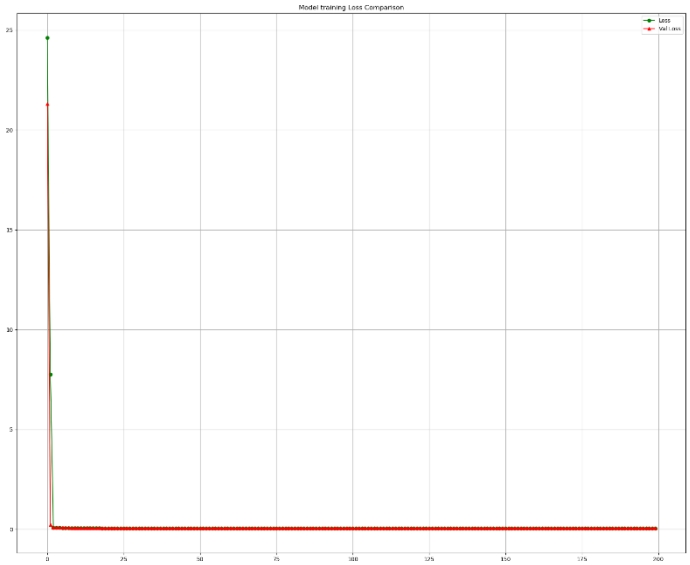
Colocar la imagen del resultado poniendo estos datos

Pero de esta manera sufriríamos mucho para saber si nuestro modelo es el mejor que podemos conseguir. Así que vamos a añadir algunas líneas de código para permitir definir:

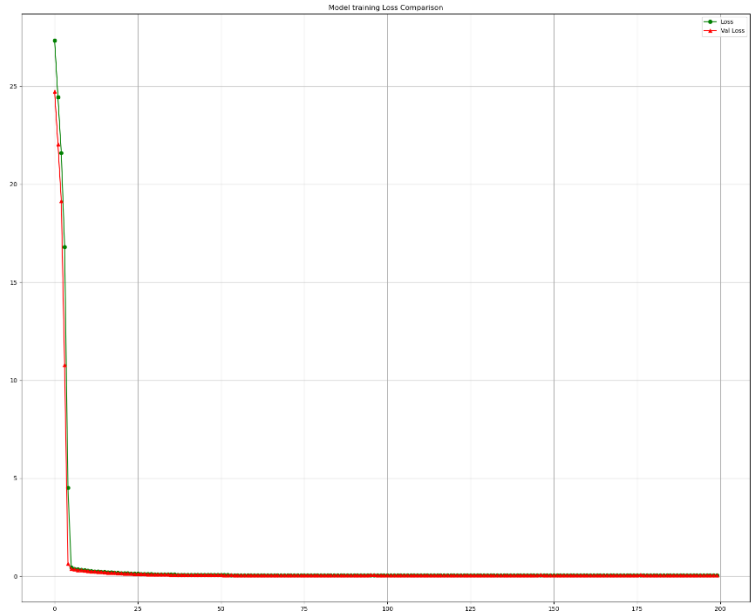
- Un conjunto de tasas de aprendizaje: [0.0001, 0.00001, 0.000001]
- Un conjunto de tamaño de lote: [8, 12, 48]
- Un conjunto de época: [50, 100 ,150, 200]
- Un modelo de matriz de tamaño de capa oculta: [3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25]

De esta manera, vamos a generar varios gráficos de comparación para tratar de llegar al mejor modelo. Sin embargo, no hay que olvidar que el entrenamiento tendrá resultados aleatorios, por lo que estamos haciendo un bucle también 3 veces cada caso para obtener la media para un modelo específico, es decir: para una tasa de aprendizaje específica, tamaño del lote, época y tamaño de la capa oculta, entrenamos 3 veces y calculamos la media de la misma, por lo que podemos comprobar, independientemente de la aleatoriedad de la ejecución, cómo la variación de los parámetros afecta al entrenamiento.
Podríamos dibujar mil gráficos diferentes aquí, mostrando diferentes relaciones entre estos parámetros. Pero se van a elegir algunos de los más interesantes, y luego se indicará nuestro mejor modelo.
Primero se va a mostrar algunas imágenes de pérdida de validación comparando los resultados del bucle de entrenamiento de cada tipo de modelo. En cada imagen se muestra el modelo específico que se está entrenando en el título, y también la tasa de aprendizaje.
Poner las 2 imágenes que obtendremos

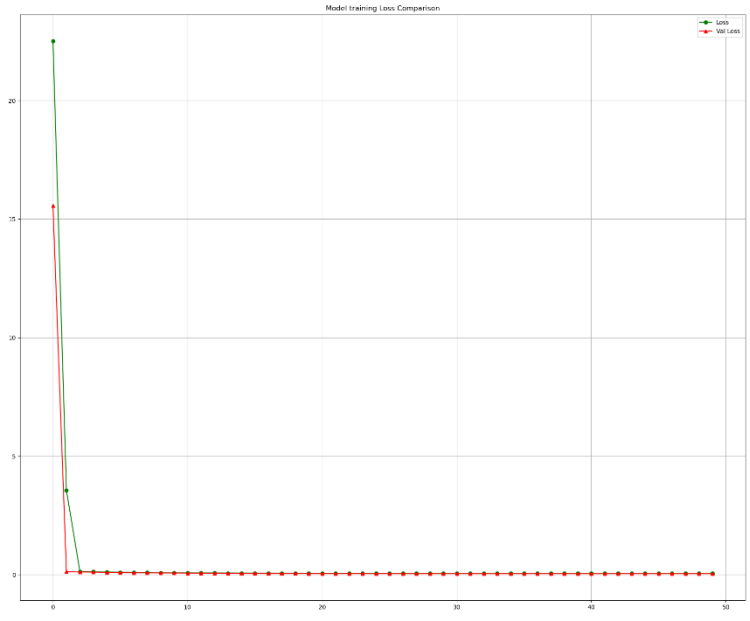
Como podemos ver, el comportamiento de los modelos entrenados tiene sentido, y podemos ver en estas imágenes relaciones como menos etapas y lotes más grandes que resultan en un aprendizaje más lento, y por lo tanto un peor resultado.
Ahora vamos a buscar el mejor modelo, y a echar un breve vistazo a cómo fue entrenado. Para ello estamos trazando la pérdida de validación para cada época durante el entrenamiento. Tenemos que darnos cuenta de que encontrar la mejor pérdida de validación para todos los modelos entrenados podría no ser suficiente, y tenemos que comprobar sus comportamientos en diferentes situaciones como el número de casos bajos, o el número de casos altos.
Hemos entrenado a miles de modelos debido a nuestro bucle de entrenamiento, por lo que vamos a atenernos a estos mejores modelos de pérdida de validación para ver cómo se comportan: hay que tener en cuenta que probablemente no estemos dando con la solución óptima.
Modelo A: 200 etapas, tamaño del lote 8, lr 1e-4, tamaño oculto Y = 25
Pérdida de validación final: 0.06039683411193148



Pérdida vs. Validación Pérdida para el entrenamiento específico del modelo A
Modelo B: 200 épocas, tamaño del lote, 12, lr 1e-4, tamaño oculto Y = 3
Pérdida de validación final: 0.05095066449471882



Pérdida vs Validación Pérdida para el entrenamiento específico del modelo B
Modelo C: 50 épocas, tamaño del lote 8, lr 1e-4, tamaño oculto Y = 20
Pérdida de validación final: 0.04294525007376584



Pérdida vs Validación Pérdida para el entrenamiento específico del modelo C
No son gráficos muy interesantes, ¿verdad? Debemos tener en cuenta dos cosas que están sucediendo aquí.
La primera puede ser la escala del eje Y que empieza con un valor alto y no nos permite ver las pequeñas mejoras en épocas posteriores.

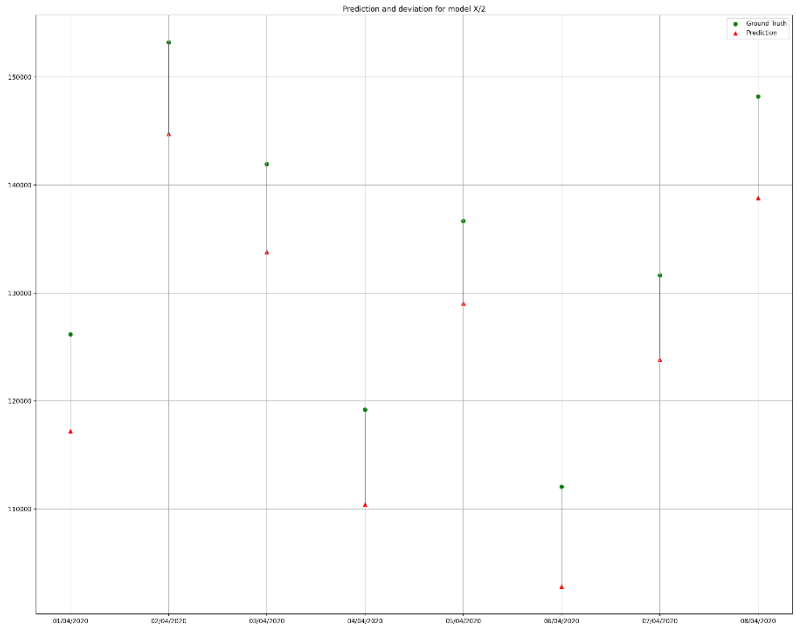
La segunda es ¿Por qué no hay alguna mejora? Podemos pensar en las causas fundamentales de este problema, pero la simplicidad de nuestro modelo y sus características pueden indicar que el modelo ha aprendido todo lo que está disponible. Dado que el modelo en sí mismo carece de complejidad, no es capaz de aprender más (al menos, sin olvidarse algo), por lo que este ajuste nunca se produce. Esto significaría que tal vez podamos hacer más grande nuestro modelo para tratar de obtener mejores resultados.

5.2.9 Predicción

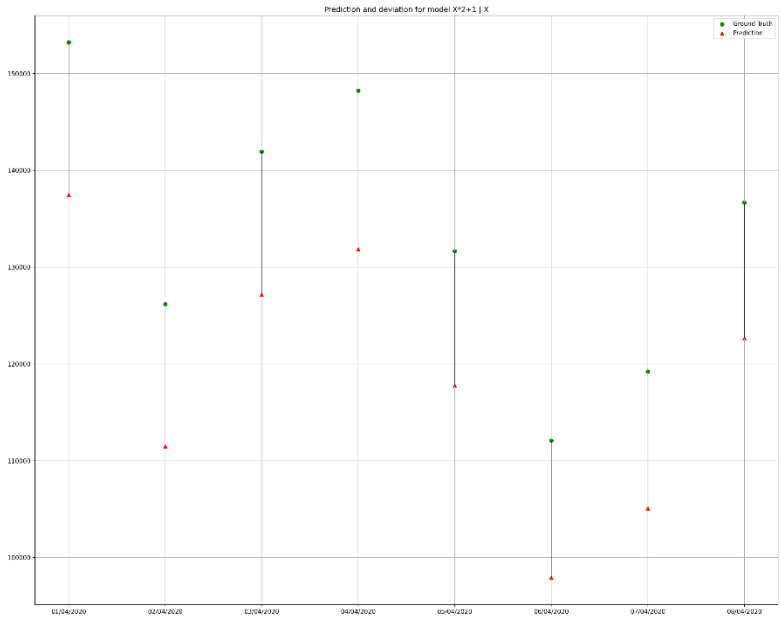
Ahora es el momento de comprobar si nuestro modelo puede decirnos cuántos casos totales tenemos al día siguiente dados los datos del anterior. Tomamos los modelos con la mejor pérdida de validación, teniendo en cuenta que aún no sabemos cómo se traduce eso en un error de predicción.
Se va a comprobar de tres maneras diferentes:
Como ya se mencionó anteriormente, estábamos usando datos hasta el 31 de marzo. Bien, empezaremos nuestra predicción y validaciones desde este mismo punto.

- Vamos a dibujar los casos confirmados reales reportados por cada país desde el 31 de marzo hasta el 6 de abril. ¿Por qué? Porque hemos llenado nuestros datos hasta el día de hoy, por lo que tenemos las características necesarias para predecir en estos días.
- Entonces, vamos a predecir con los datos de cada día (del 1 de abril al 6 de abril) la predicción para el día siguiente y la comparamos con la anterior. De esta manera, podemos ver el error o desviación que sufren nuestros modelos contra los datos reportados de los países.

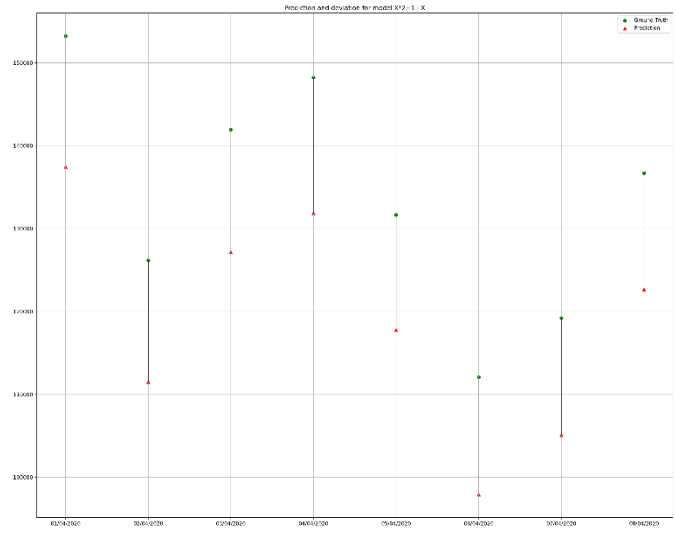
Estos dos se van a graficar juntos en el mismo valor del eje x, y se van a unir por una línea para que podamos tener una representación clara de la desviación. Tengan en cuenta los valores del eje ya que son fechas, y los valores del eje y. Hemos descomprimido la salida logarítmica de la red de vuelta a los casos. No se establecerán los límites del eje Y, debido a que se aplican valores diferentes para cada país, lo que haría que una escala común no fuera útil para la comparación.



Modelo A - Predicciones para España en abril



Modelo B - Predicciones para España en abril



Modelo C - Predicciones para España en abril

A partir de estos tres modelos podemos ver de nuevo que el modelo B es el que mejor se desempeña, presentando desviaciones realmente bajas en las predicciones de algunos días.

5.2.9 Conclusión

A través de esta presentación hemos pasado por todos los pasos necesarios para aplicar las técnicas de Aprendizaje Automático a un caso de uso real, desde la obtención de los datos en bruto hasta la predicción de los casos.

Hemos recogido, aplicado un proceso de regularización y definido un conjunto de características que nos han permitido entrenar varios modelos simples con una precisión limitada en sus predicciones.

Los resultados son mejorables, y el proceso permite estas mejoras tanto en los datos como en el modelo. Incluso podemos mejorar la selección del mejor modelo, ya que se tienen miles de modelos entrenados y no estamos seguros de cuáles son los mejores con sólo mirar la pérdida de validación. De esta manera, podemos dividir nuestro problema de decisión aquí en dos partes: los resultados de los modelos en sí mismos y la selección de modelos de la formación que hemos realizado.

Como decíamos al principio de la historia, fue una tarea difícil, pero permítanos recordar que esto fue un trabajo realizado para un TFG y que hacer un buen modelo es casi imposible.

Un punto importante para destacar es que, si un nuevo virus o pandemia surge con características similares, podríamos usar este mismo modelo y ejecutar más entrenamiento o ajuste sobre él para reutilizarlo con el nuevo y obtener valiosos datos de comportamiento. Otro factor a tener en cuenta y que es incluso más importante es que las correcciones de datos que podrían ser publicadas por los gobiernos en el recuento de casos y así sucesivamente podrían también mejorar la entrada de datos para el modelo para que éste aprenda mejor el patrón de propagación.

5.3.4 Proyección del estudio en un futuro

1. Sobre los datos:
    - Cada día hay más datos que podrían ser utilizados para entrenar el modelo.
    - Si hay más datos que entrenar, podría conducir a un modelo más complejo con nuevas características que podrían añadir valor al modelo.
    - Aplicar más/mejorar las desventajas o ajustar sus valores para la característica de densidad de población, si es necesario.
2. Sobre nuestra predicción o modelo:
    - Como se ha comentado antes de mostrar algunas predicciones, parece que los modelos más complejos (es decir, más grandes, más neuronas, más capas) podrían aprender más de nuestro conjunto de datos.
    - Hemos realizado el estudio sobre los casos confirmados. Podría aplicarse a las muertes y a los casos recuperados (también, añadiendo una nueva característica de edad, por ejemplo).
    - También es posible probar otro tipo de red, como los RNNs comentados anteriormente.
3. Hay margen para mejoras técnicas:
    - Mejorar el código de rendimiento.
    - Añadir más parámetros configurables para el guion de entrenamiento.
    - Otras consideraciones

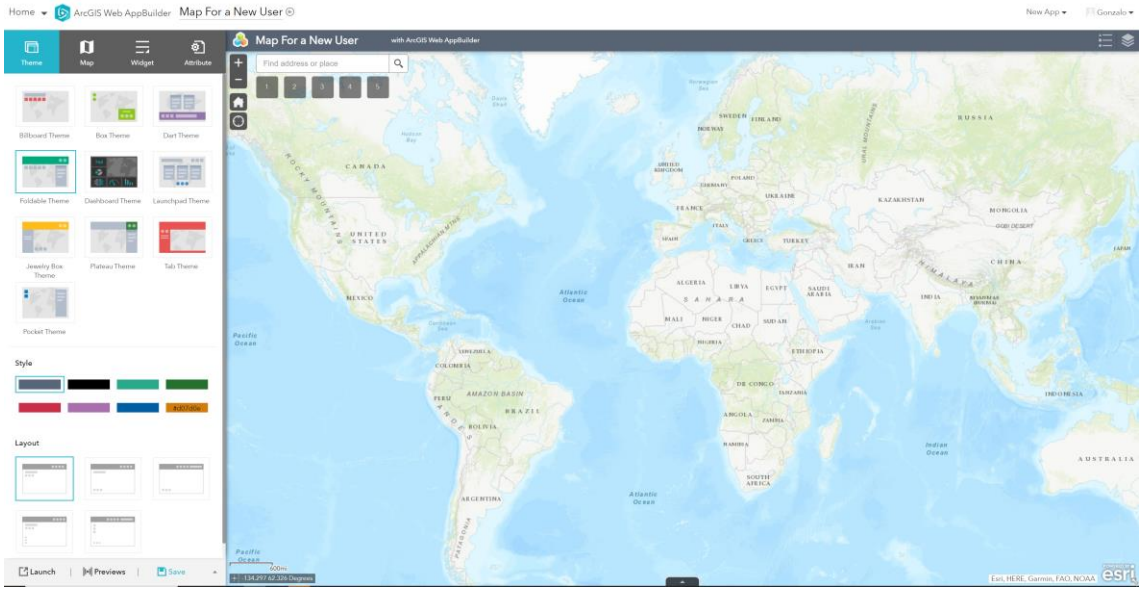
Somos conscientes de que algunos de los datos utilizados pueden ser imprecisos: densidad de población, informes de casos confirmados (debido a los informes de los gobiernos...) y así sucesivamente. Hemos hecho todo lo posible por utilizar un terreno de datos consistente para el ejercicio, pero nos damos cuenta y reconocemos que es imposible modelarlo de la forma más precisa posible.

5.3 ArcGIS WebApp

Por último, queda explicar qué se va a ver de cara al usuario y que posibilidades tiene este si quiere cambiar parámetros y estadísticas.

5.3.1 Empezando con la Interfaz Vacía

El usuario si así lo desea, puede crearse un perfil de desarrollador en ArcGIS y descargarse nuestras capas y crear la aplicación a su antojo. De ser así el usuario vería esto.



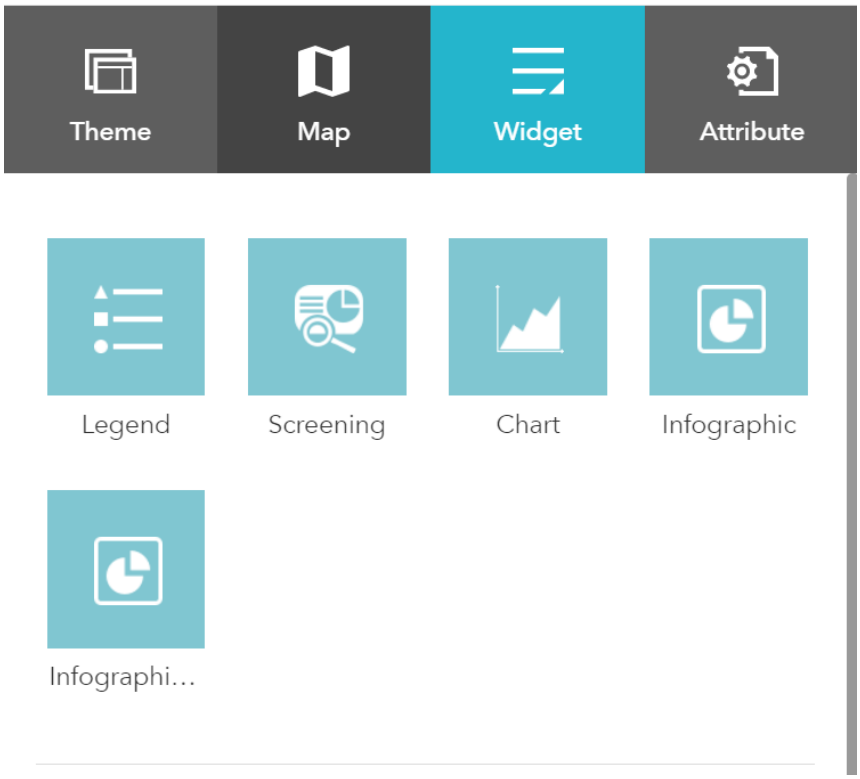
Esta es la interfaz completamente vacía y lista para que un usuario avanzado se ponga a elegir todos los parámetros que desea.
En este punto no nos vamos a dedicarle mucho más, ya que se necesita de una persona que sepa adentrarse en las configuraciones y elegirlas a su gusto.

5.3.2 Usando nuestra interfaz

Para las personas menos experimentas no es necesario tener conocimientos avanzados de ArcGIS, ya que todas las capas están subidas y listas para usarse. Vamos a ir haciéndoles un pequeño recorrido para que sepan todas las posibilidades que le ofrece nuestra interfaz.

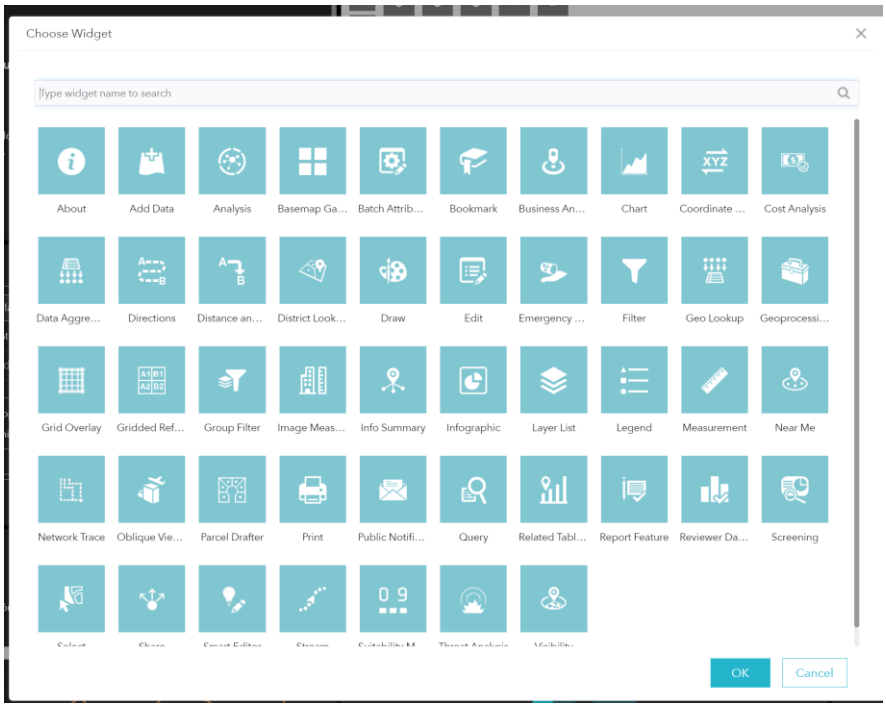
5.3.3 Configuraciones de los Widgets

Los widgets son pequeñas aplicación o programas, usualmente presentado en archivos o ficheros pequeños que son ejecutados por un Widget Engine y entre sus objetivos están dar fácil acceso a funciones frecuentemente usadas y proveer de información visual.
ArcGIS nos da muchísimos widgets los cuales tienen sus propias configuraciones y también nos da diferentes opciones de configuraciones dentro de ellas. Los widgets se encuentran en la esquina superior derecha y tienen esta forma:



Si hacemos clic en alguno de los 5 widgets se verán estas posibilidades.

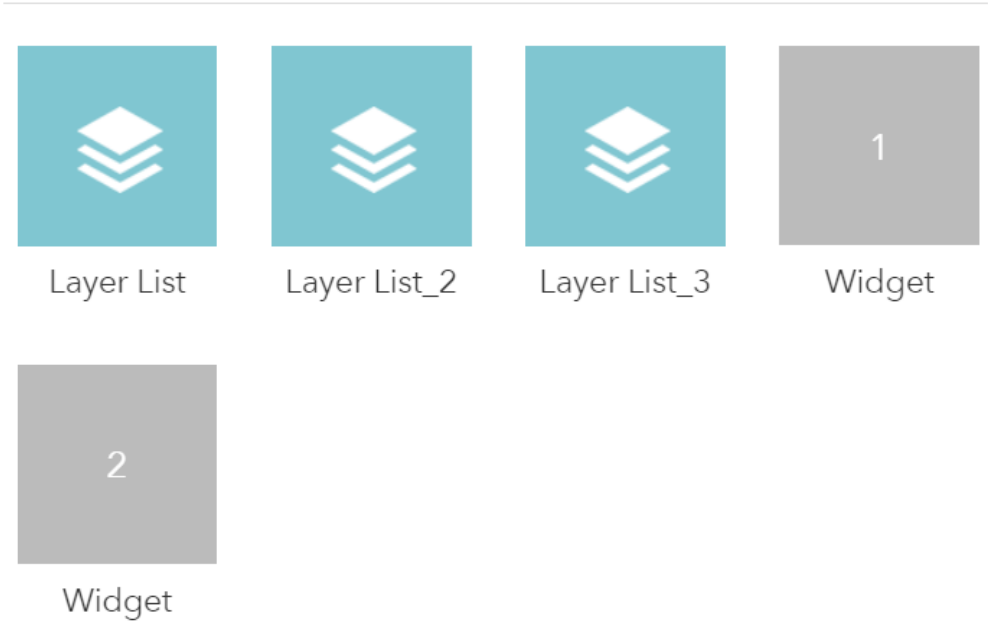
Arriba a la izquierda tenemos añadir más widgets para poder combinar resultados, arriba a la derecha nos deja eliminarlo para dejar espacio a un nuevo widget y abajo a la derecha modificarlo para ajustarlo a nuestro gusto.
Los widgets que están disponibles son estos:



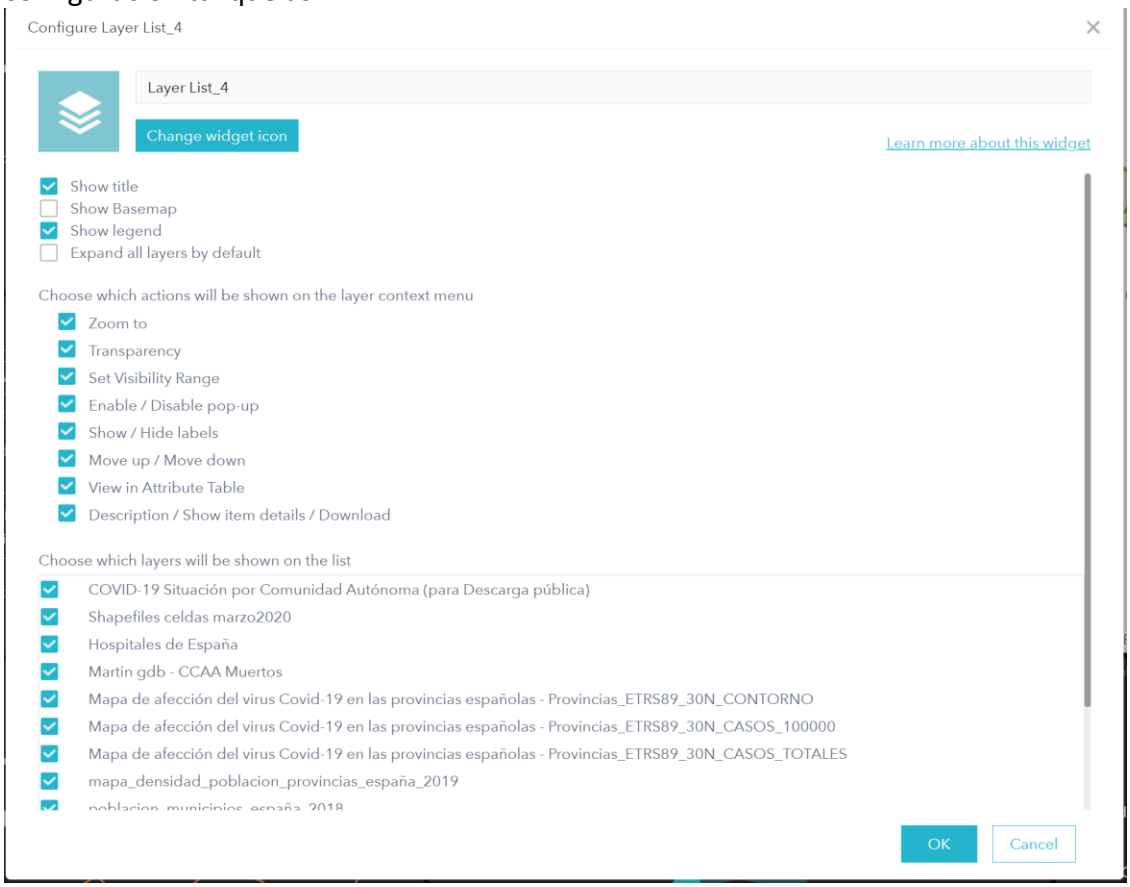
Como se puede observar hay muchísimas diferentes, por lo que es normal querer cambiar de widgets y aplicarle uno propio. También como es de entender no se puede ir explicando uno a uno, ya que no es la proyección de este trabajo el enseñar todas las funcionalidades de ArcGIS WebApp.

5.3.4 Configuraciones de las capas

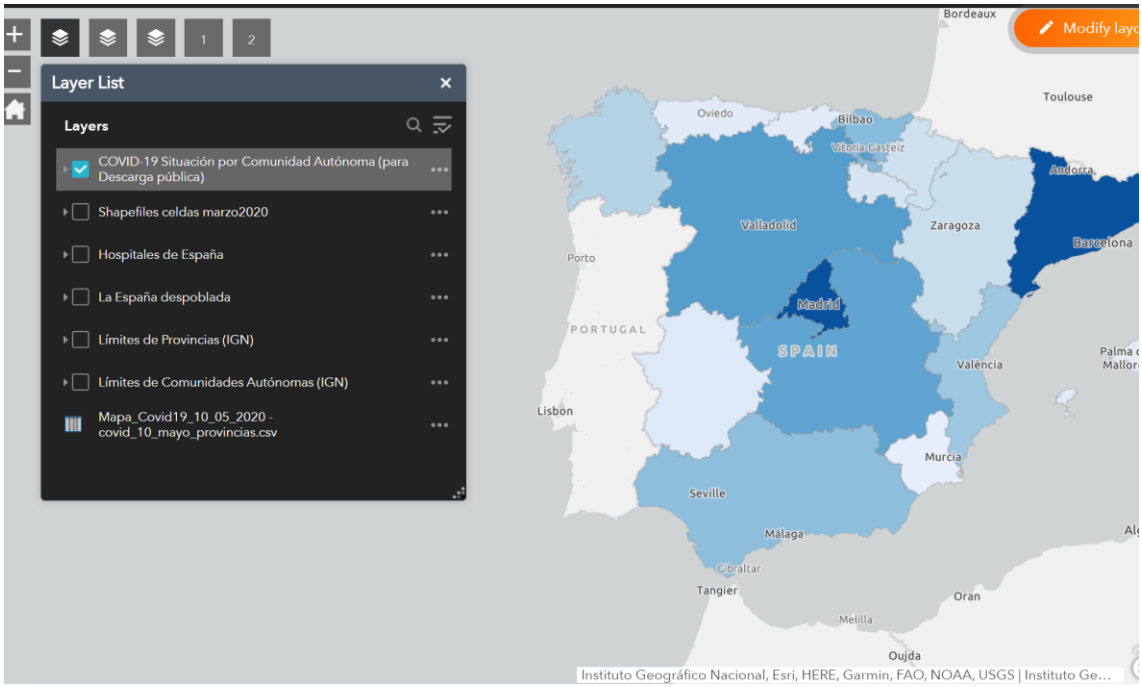
Para configurar las distintas combinaciones de capas que podemos ofrecer simplemente hay hacer clic en la parte inferior izquierda que tiene un menú tal que así:



Como pueden ver nosotros hemos dejado dos listas de capas a configurar por el usuario. Normalmente con una lista de capas es suficiente, pero esto se puede modificar al gusto del consumidor, solo tiene que pinchar en el hueco vacío. Apareciéndole una configuración tal que así.



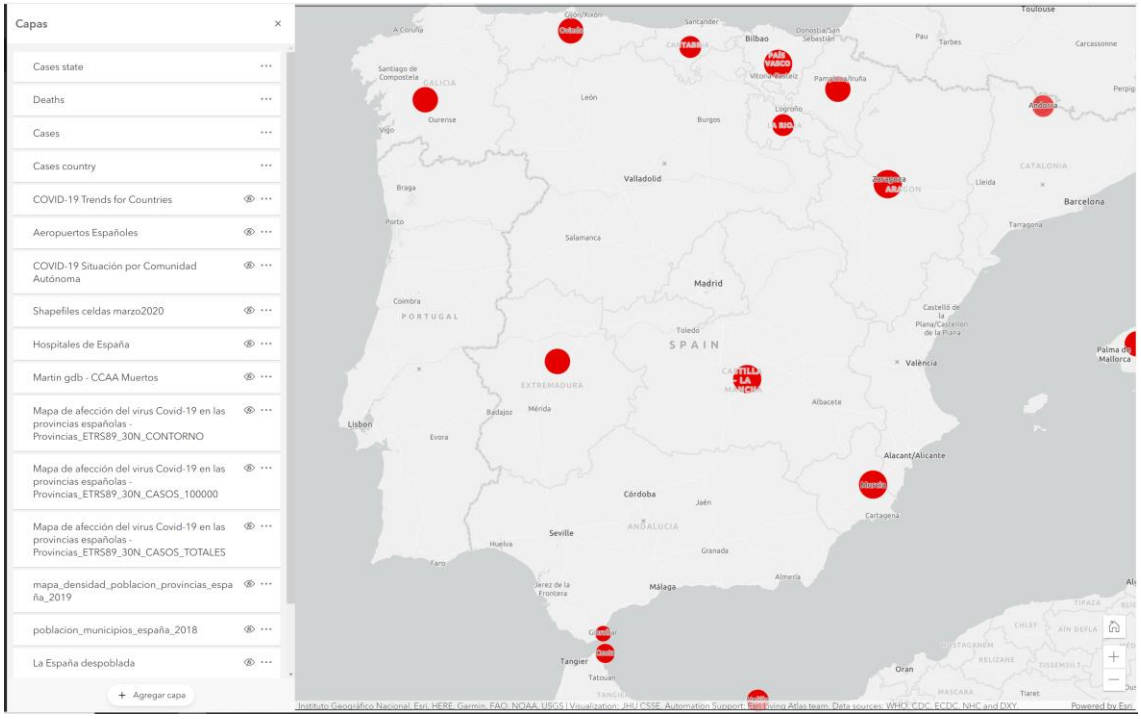
Si se seleccionan todas las capas como aparece en la imagen, se mostrará una imagen muy similar a esta.



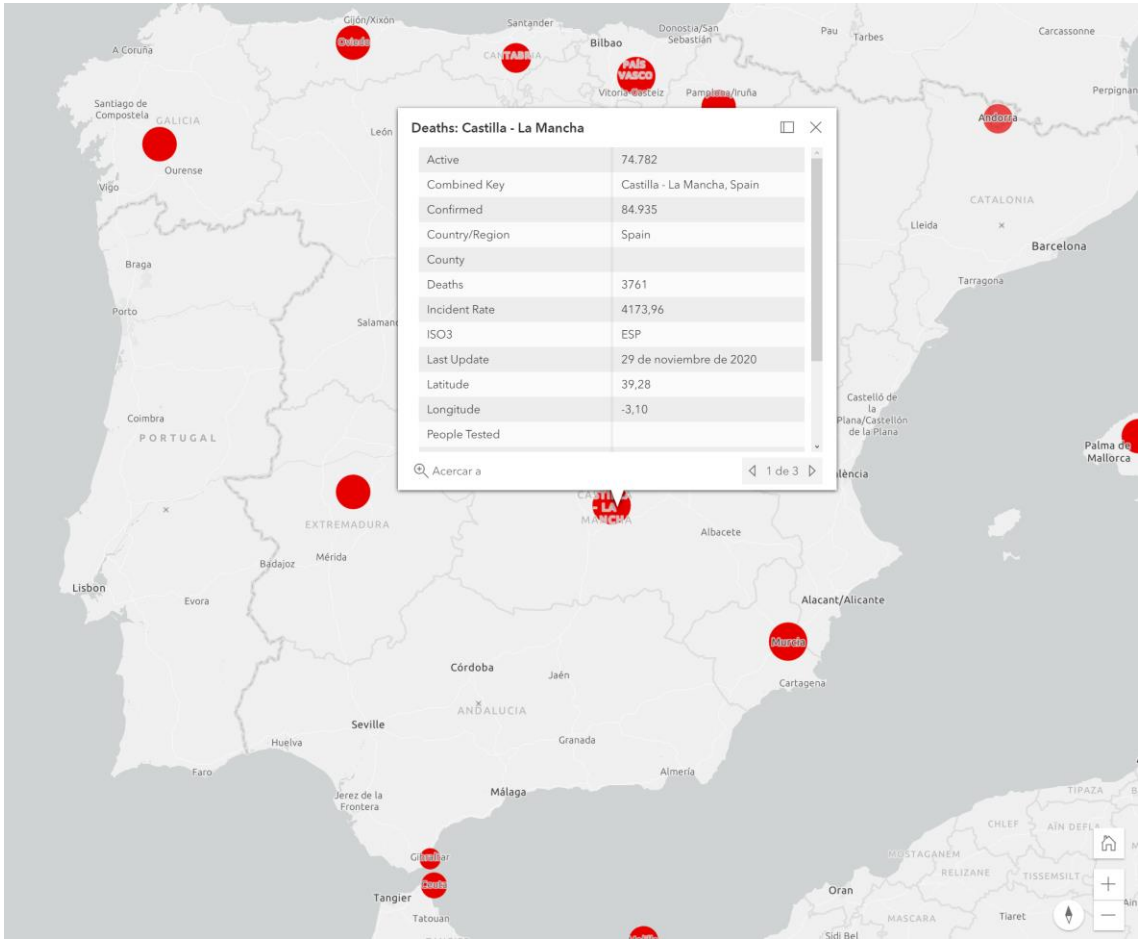
6

Resultados

Los resultados obtenidos a primera vista deberían verse divididos en dos, uno para ArcGIS y otro para la IA. Pero esto no es así, los resultados son mucho más interesantes cuando los vemos en conjunto.
Como podemos observar en los modelos, los datos no son tan dispares como los que hemos proyectado nosotros para esas fechas, lo que supone un gran avance, teniendo en cuenta que esto ha sido desarrollado por gente que no se dedica en exclusiva a esto.
Un ejemplo de esto es el mapa que se presenta a continuación.



Como podemos observar vemos distintos puntos rojos marcando zonas provinciales en las que se cree que el virus tendrá un desarrollo mucho mayor que en las zonas done no se encuentra dicho punto.
Si acercamos el ratón a cualquiera de estos puntos nos encontraremos con esta información.



Donde nos hace un informe entero de esa comunidad autónoma, para que se pueda investigar y se trate de paliar un posible incremento de los casos
Estas imágenes es por la que se ha realizado este proyecto. Una imagen donde podemos ver en donde creemos que aumentarán los casos. Se da por hecho que estos datos no son exactos, ya que es una herramienta en una fase de pruebas muy primitiva, pero en la que, si se dispusiese de un potente equipamiento y un grupo de personas con conocimientos informáticos medios, se podrían obtener resultados mucho más preciosos. Lo que ayudaría a una entidad tan importante como es el gobierno español, a tomar unas medidas más prudentes en los lugares donde creemos que se desarrollará mucho más este virus. El potencial de este trabajo, si se llega a desarrollar por expertos de ArcGIS y IA el alcance sería mucho mayor, y puede hacer que un país se plantea una estrategia con datos reales y con mucha mayor fiabilidad. Esto es un avance increíble y que puede salvar muchas vidas, pero requiere de una investigación mucho más larga y duradera.

7

Proyección a futuro

Al principio esta idea, era una aplicar la inteligencia artificial, como de costumbre, haciendo una base de datos y entrenándola a partir de ahí. Pero cuando nos metemos en un tema tan actual como este, nos damos cuenta de que hay muy pocos datos y muchos de ellos guardados de tal manera que hace más complicado el estudio. También en el momento de la creación de este proyecto, encontrábamos estadísticas básicas (solo casos de infectados y posteriormente de muertes), pero no existía una forma de hacer ver una interactivamente los datos como los hemos puesto nosotros.

Gracias al uso de herramientas tan novedosas y con tanto camino por recorrer, creemos que este proyecto podrá servir de referencia durante por lo menos unos cuantos años, que como ya sabemos la informática es un mundo en constante cambio y las herramientas se quedan obsoletas de un día para otro, por lo que una duración de más de cuatro años es un tiempo bastante aceptable.

La intención de este trabajo de investigación es aclarar ciertas bases que requieren de muchas horas de lectura y navegación por internet sobre las herramientas GIS, la inteligencia artificial y la combinación de ambas, para que alguien, si vuelve a ocurrir una epidemia parecido pueda leer este trabajo y tener ciertas nociones de cómo aplicar la IA y el GIS juntos.

Para ello se dejarán todos los repositorios de forma pública y con su descripción correspondiente (tanto del código con comentarios como de los mapas y tablas usados para ArcGIS), para intentar hacer de estos lo más accesibles que se pueda Incluso creemos haber dejado una huella en ArcGIS España, ya hay usuarios usando nuestras capas y dando datos relevantes sobre el COVID-19.

8

Conclusión

Una pandemia no es nueva en la historia de la humanidad. Pero lo que hace especial a la pandemia COVID-19 es que tiene lugar en un contexto sin precedentes, cuando se creía que la interconexión e interdependencia entre las personas, entre los países y entre los continentes era muy profunda, esta pandemia nos ha demostrado que aún nos falta mucho camino por recorrer y que las inversiones en la comunidad científica pueden llegar a salvar países de desastres como este

Ha quedado muy claro, que para evitar crisis como la que estamos viviendo a día de hoy hay muchísimos factores que tenemos que cambiar, por ejemplo:

- La comunidad científica debe revisar la forma en que se relaciona con el conjunto de la sociedad. Ahora más que nunca existe una necesidad patente de transparencia y de utilización de un lenguaje accesible a todos.
- Las publicaciones científicas tienen ahora la oportunidad de revisar su modelo de negocio y analizar la forma en la que este configura la producción académica y la investigación en general. Ha llegado el momento de dejar atrás los vicios adquiridos durante las épocas en que las comunicaciones escritas eran la norma.
- En paralelo, las redes sociales deben desintoxicar sus algoritmos para que reduzcan la presencia de desinformación, los grupos de páginas y los dominios pertenecientes a los aceleradores de desinformación, y mantener el contenido perjudicial alejado de su tráfico.
- Todos debemos contribuir a generar y difundir información de calidad, evitando los rumores y chismes que solo contribuyen a la desinformación.

Estos son sólo alguno de los ejemplos de las dificultades a las que la comunidad científica se enfrenta a la hora de dar soluciones contra la pandemia.

Pero no todo es malo, ya que existen muchas herramientas que se usaban enfocaban para usos comerciales o totalmente diferentes a las necesidades que surgen hoy en día, pero que son igualmente aplicables y útiles, como puede ser la inteligencia artificial geoespacial.

Si bien los datos geoespaciales satelitales de alto detalle han estado disponibles durante décadas, los avances geoespaciales contemporáneos de la inteligencia artificial recién nos acaban de permitir desbloquear su verdadero potencial y gracias a esta pandemia la comunidad científica se ha puesto manos a la obra con el desarrollo de estas tecnologías. Por ejemplo, ArcGIS ha desarrollado una campaña para que todos sus usuarios suban mapas (a la que nos hemos unido nosotros también por supuesto), para así tener una base de datos lo más grande posible con todos los datos relevantes para el estudio del COVID-19.

Tenemos que admitir que la pandemia de COVID-19 nos ha mostrado ejemplos que carecen de humanitarismo. Esto puede deberse al caos causado por la propagación de la amenaza. Sin embargo, esa falta de humanitarismo parece estar muy arraigada. Esto se debe al egoísmo incurable de algunos países y sus elites gobernantes. Los que se proclaman líderes morales con tradiciones democráticas no unieron a todas las partes para buscar el entendimiento mutuo. En su lugar, comenzaron a actuar de acuerdo con la ley de la selva, independientemente de las normas de etiqueta y las limitaciones éticas.

Nuestro propósito en este trabajo es dar nuestro grano de arena como mucha gente ha hecho y empezar a dar soluciones para la lucha contra el crecimiento de esta pandemia. Pero no solo nos queríamos quedar en ese punto, nuestra idea va un poco más allá. La idea de que esto esté en un repositorio público con toda la información y explicaciones estén a disposición de cualquier persona, para que, si algún día vuelve a ocurrir esto, se tengan referencia de que se puede hacer para empezar a estudiar el crecimiento y dar indicios de lo que pueda pasar en un futuro.

9

Referencias

1. Li S, Dragicevic S, Castro FA, Sester M, Winter S, Coltekin A, Pettit C, Jiang B, Haworth J, Stein A. Geospatial big data handling theory and methods: a review and research challenges. ISPRS J Photogramm Remote Sens. 2016;115:119–33.

2. IBM. Industry Insights: 2.5 quintillion bytes of data created every day. How does CPG & Retail manage it? https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/. Accessed 30 Oct 2017.

3. Baker D, Nieuwenhuijsen MJ. Environmental epidemiology: study methods and application. New York: NY: Oxford University Press.

4. Lin Y, Chiang Y-Y, Pan F, Stripelis D, Ambite JL, Eckel SP, Habre R. Mining public datasets for modeling intra-city PM2.5 concentrations at a fine spatial resolution. In: Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems. Los Angeles area, CA: ACM; 2017. p. 1–10.

5. Dietrich D. Data science & big data analytics: discovering, analyzing, visualizing and presenting data. Indianapolis, IN: John Wiley & Sons, Inc; 2015.

6. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health information science and systems. 2014;2(1):3.

7. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev. 2012;90(10):60–8.

8. Dominici F, Parkes D. Harvard in Allston: data science: SoundCloud. Harvard University podcast; 2017. https://soundcloud.com/harvard/harvard-in-allston-data-science?in=harvard/sets/harvard-in-allston

9. Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. Big Data. 2013;1(1):51–9.

10. Wickham H, Grolemund G. R for data science. Sebastopol, Canada: O'Reilly Media, Inc.; 2016.

11. Wang S. CyberGIS and spatial data science. GeoJournal. 2016;81(6):965–8.

12. Anselin L. Spatial data, spatial analysis and spatial data science. The University of Chicago: the Center for Spatial Data Science 2016.

13. University of Illinois Urbana-Champaign. ROGER: The CyberGIS Supercomputer. https://wiki.ncsa.illinois.edu/display/ROGER/ROGER%3A+The+CyberGIS+Supercomputer. Accessed 30 Oct 2017.

14. Goodchild MF. Citizens as sensors: the world of volunteered geography. GeoJournal. 2007;69(4):211–21.

15. Senaratne H, Mobasheri A, Ali AL, Capineri C, Haklay M. A review of volunteered geographic information quality assessment methods. Int J Geogr Inf Sci. 2017;31(1):139–67.

16. Scassa T. Legal issues with volunteered geographic information. Can Geogr. 2013;57(1):1–10.

17. Ma Y, Wu H, Wang L, Huang B, Ranjan R, Zomaya A, Jie W. Remote sensing big data computing: challenges and opportunities. Futur Gener Comput Syst. 2015;51:47–60.

18. DigitalGlobe. The DigitalGlobe Constellation. https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/223/Constellation_Brochure_forWeb.pdf. Accessed 30 Oct 2017.

19. U.S. Geological Survey. Landsat. https://landsat.usgs.gov/. Accessed 30 Oct 2017.

20. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: The MIT Press; 2016.

21. Nieuwenhuijsen MJ. Exposure assessment in environmental epidemiology. 2nd ed. New York, NY: Oxford University Press; 2015.

22. Nuckols JR, Ward MH, Jarup L. Using geographic information systems for exposure assessment in environmental epidemiology studies. Environ Health Perspect. 2004;112(9):1007–15.

23. Nieuwenhuijsen MJ. Exposure assessment in environmental epidemiology. 2nd ed. New York, NY: Oxford University Press; 2015.

24. Stripelis D, Ambite JL, Chiang Y-Y, Eckel SP, Habre R. A scalable data integration and analysis architecture for sensor data of pediatric asthma. In: Data Engineering (ICDE), 2017 IEEE 33rd International Conference on: IEEE; 2017. p. 1407–8.

25. *5 Reasons why Azure ML for Machine Learning solutions*. (s. f.). SaviantConsulting.com. Recuperado 10 de octubre de 2020, de https://www.saviantconsulting.com/blog/5-reasons-azureml-for-machine-learning-solutions.aspx

26. *A remote sensing and GIS-assisted landscape epidemiology approach to West Nile virus*. (2013, 1 diciembre). ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0143622813002361

27. Ai, B. (2019, 4 agosto). *Top 5 de las mejores librerias de Inteligencia Artificial*. Medium. https://bootcampai.medium.com/top-10-de-las-mejores-librerias-de-inteligencia-artificial-aebaa62513dc

28. Bermejo, E. (2020, 22 noviembre). *¿QUÉ ES LA TECNOLOGÍA ARCGIS?* Territorio Geoinnova - SIG y Medio Ambiente. https://geoinnova.org/blog-territorio/que-es-la-tecnologia-ArcGIS/

29. colaboradores de Wikipedia. (2020, 10 octubre). *Estudio epidemiológico*. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/Estudio_epidemiol%C3%B3gico

30. Data, S. B. (2019, 3 junio). *Amazon Machine Learning, Azure, Cloud AI y Watson*. sitiobigdata.com. https://sitiobigdata.com/2019/01/31/mlaas-amazon-machine-learning/

31. *Estadísticas - Aeropuertos Españoles - aena.es*. (s. f.). AENA. Recuperado 25 de noviembre de 2020, de http://www.aena.es/csee/Satellite?pagename=Estadisticas/Home
32. *GIS - Practice - Exercise*. (s. f.). BCC. Recuperado 19 de noviembre de 2020, de http://www.biodiversity.ru/coastlearn/gis-eng/exercise.html
33. Grandio, X. (2017, 14 julio). *Qué es Tensor Flow: aplicaciones del sistema de inteligencia artificial de Google*. Marketing 4 Ecommerce. https://marketing4ecommerce.net/tensorflow-que-es-y-sus-aplicaciones/
34. Heller, M. (2019, 28 enero). *What is Keras? The deep neural network API explained*. InfoWorld. https://www.infoworld.com/article/3336192/what-is-keras-the-deep-neural-network-api-explained.html
35. J. (2018, 14 febrero). *Todo lo que necesitas saber sobre TensorFlow, la plataforma para Inteligencia Artificial de Google – Puentes Digitales*. Puentes Digitales. https://puentesdigitales.com/2018/02/14/todo-lo-que-necesitas-saber-sobre-tensorflow-la-plataforma-para-inteligencia-artificial-de-google/
36. Li, M. (2018, 18 julio). *Ranking Popular Deep Learning Libraries for Data Science*. The Data Incubator. https://blog.thedataincubator.com/2017/10/ranking-popular-deep-learning-libraries-for-data-science/
37. Moreno-Altamirano, A. (2000, 1 agosto). *SciELO - Saúde Pública - Principales medidas en epidemiología Principales medidas en epidemiología*. SciELO. https://www.scielosp.org/article/spm/2000.v42n4/337-348/es/
38. Nain, A. (2018, 18 julio). *TensorFlow or Keras? Which one should I learn? - Imploding Gradients*. Medium. https://medium.com/implodinggradients/tensorflow-or-keras-which-one-should-i-learn-5dd7fa3f9ca0

39. Nuckols, J. R., Ward, M. H., & Jarup, L. (2004, 1 enero). *Using Geographic Information Systems for Exposure Assessment in Environmental Epidemiology Studies*. EHP. https://ehp.niehs.nih.gov/doi/full/10.1289/ehp.6738

40. Patil, A. (2020, 24 abril). *TensorFlow vs Caffe*. EDUCBA. https://www.educba.com/tensorflow-vs-caffe/

41. Singh, Y. (2020, 14 enero). *What is ArcGIS*. GeoSpatial World. https://www.geospatialworld.net/blogs/what-is-ArcGIS/

42. VoPham, T., Hart, J. E., Laden, F., & Chiang, Y. (2018, 17 abril). *Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology*. Environmental Health. https://ehjournal.biomedcentral.com/articles/10.1186/s12940-018-0386-x?optIn=true

43. *Ministerio de Sanidad, Consumo y Bienestar Social - Profesionales - Enfermedad por nuevo coronavirus, COVID-19*. (s. f.). Ministerio de Sanidad, Consumo y Bienestar Social. Recuperado 28 de noviembre de 2020, de https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/

44. *Amazingly Simple Graphic Design Software – Canva*. (s. f.). Canva. Recuperado 28 de noviembre de 2020, de https://www.canva.com

45. Bdm, R. (2020, 16 junio). *Estas son algunas de las librerías de Python que necesitas conocer*. Big Data Magazine. https://bigdatamagazine.es/estas-son-algunas-de-las-librerias-de-python-que-necesitas-conocer

46. Chang, S. L., Harding, N., Zachreson, C., Cliff, O. M., & Prokopenko, M. (2020, 11 noviembre). *Modelling transmission and control of the COVID-19*

pandemic in Australia. Nature Communications. https://www.nature.com/articles/s41467-020-19393-6?error=cookies_not_supported&code=efdf0a61-4a1d-4dce-9644-2f56ab502f63

47. *Crear gráficos de diagrama de caja—Ayuda | ArcGIS for Desktop*. (s. f.). ArcGIS. Recuperado 16 de septiembre de 2020, de https://desktop.arcgis.com/es/arcmap/10.3/map/graphs/creating-box-plot-graphs.htm
48. editor. (2020, 12 septiembre). *¿Qué es la inteligencia geoespacial o geointeligencia?* Nave. https://aceleradoranave.com.mx/inteligencia-geoespacial/#:%7E:text=La%20inteligencia%20geoespacial%20busca%20entender,bases%20de%20datos%20e%20informaci%C3%B3n.
49. Li, M. (2018, 18 julio). *Ranking Popular Deep Learning Libraries for Data Science*. The Data Incubator. https://blog.thedataincubator.com/2017/10/ranking-popular-deep-learning-libraries-for-data-science/
50. Mazzola, M. I. (2020, noviembre). *ArcGis vs QGIS*. Google Trends. https://trends.google.es/trends/explore?q=%2Fm%2F082gc5,%2Fm%2F0ct9z5
51. *WHO Coronavirus Disease (COVID-19) Dashboard*. (s. f.). WHO. Recuperado 27 de noviembre de 2020, de https://covid19.who.int
52. Mani, A. (2020, 18 agosto). *House hunting — the data scientist way - GeoAI*. Medium. https://medium.com/geoai/house-hunting-the-data-scientist-way-b32d93f5a42f

53. National Geographic Society. (2012, 9 octubre). *GIS (Geographic Information System)*. https://www.nationalgeographic.org/encyclopedia/geographic-information-system-gis/

54. *The Beginner's Guide to GIS - What is GIS?* (2017, 12 enero). Mango. https://mangomap.com/what-is-gis

UNIVERSIDAD DE MÁLAGA | uma.es

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga

E.T.S. DE INGENIERÍA INFORMÁTICA