Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

Trabajo de Fin de Grado

Grado en Ingeniería Informática

Aplicación práctica de los datos abiertos: actualización y predicción.

Practical application of open data: updating and forecasting.

Brian Samir Santamaría Valero

La Laguna, 30 de junio de 2021

D. **José Luis Roda García**, con N.I.F. 43.356.123-L profesor Titular de Universidad adscrito al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como tutor

D. **Ginés León Rodríguez**, con DNI 78.408-712-X responsable del Departamento de Big data y Data Science de TITSA, como co-tutor

C E R T I F I C A N

Que la presente memoria titulada:

“Aplicación práctica de los datos abiertos: actualización y predicción.”

ha sido realizada bajo su dirección por D.**Brian Samir Santamaría Valero**,

con N.I.F. 79.08.44.93-M

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 30 de junio de 2021.

Agradecimientos

La experiencia adquirida durante la realización del proyecto ha sido de gran ayuda para mi formación. Por tanto, agradezco a todas las personas que me han acompañado durante este periodo.

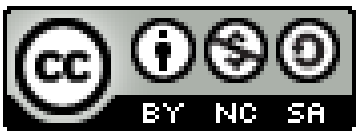
En especial a Jose Luis Roda García, tutor del proyecto el cual ha estado siempre dispuesto a ayudar de buena gana.

También he de agradecer a Ginés León Rodríguez, Responsable del Departamento de Big Data y Data Science en la empresa Titsa, ya que con su gran experiencia hemos conseguido marcar unos objetivos reales en el proyecto, y he podido aprender mucho de él tanto a nivel profesional como personal.

He de hacer una mención especial a Carlos Domínguez García, ex-alumno de la Universidad de La Laguna, y actualmente colaborador en la Cátedra Cajasiete de Big Data, Open Data y Blockchain, el cual me ha apoyado durante la elaboración del Trabajo.

Por supuesto, tampoco puedo olvidarme de mi familia y resto de amigos siempre apoyando en los momentos de dificultad.

Licencia



Resumen

Cada vez son más las administraciones públicas y empresas que publican datos abiertos para su posterior reutilización por la ciudadanía. Su utilidad es evidente, por la variedad de datos que existen, y por la facilidad de acceso, ya que estos datos pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, siempre que se respete la veracidad de la información.

Por esa razón, el objetivo del Trabajo de Fin de Grado ha sido aprender a desenvolverse en los procesos de análisis de datos, principalmente de Canarias, realizando un seguimiento del dato y determinando su actualización y persistencia. Para que los datos obtenidos sean más entendibles por los usuarios, se utilizan herramientas que ayudan a mejorar la visualización de los datos, como Power Bi. Además, los datos obtenidos de portales abiertos han servido para realizar predicciones de series temporales sobre la COVID-19 en Prophet y Neural Prophet.

Palabras clave: Datos abiertos, análisis del dato, predicción, Covid-19, meteorología.

Abstract

More and more public administrations and companies are starting to publish open data to be used by the general public. The usefulness of open data is evident not only due to the variety of the information, but also due its accessibility, since these resources can be used, reused, and redistributed freely by anyone, so long as the veracity of the information is respected.

For this reason, the objective of this Final Degree Project has been to learn how to manage the processes of data analysis, mainly in the Canary Islands, by monitoring the data and determining its updating and persistence. In order to make the data more comprehensible for users, tools, such as Power Bi, are used that help to improve its visualisation. In addition, data obtained from open portals have been used to perform time series predictions on COVID-19 in Prophet and Neural Prophet.

Keywords: Open data, data analysis, forecasting, Covid-19, meteorology.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1 Justificación

Cada vez es más común encontrarnos con portales de datos abiertos en internet. Con estos portales se pretende que todos los datos disponibles puedan hacerse públicos y accesibles tanto para la ciudadanía como para las empresas. De esta forma, los datos pueden ser consultados, reutilizados y redistribuidos libremente por cualquier persona, siempre que se respete la veracidad de la información.

La utilidad que presentan los portales de datos abiertos es evidente, prueba de ello es el crecimiento que se ha experimentado en España, pasando de 11 portales operativos y una oferta aproximada de 5.000 datasets en el año 2013, a más de 45.000 conjuntos de datos en la actualidad.

Sin embargo, existen portales de datos abiertos que no se actualizan periódicamente, o sus datos presentan errores de persistencia. Por esta razón, es de vital importancia saber analizar los datos de manera correcta para determinar la fiabilidad a la hora de obtener el dato, y mejorar la visualización de forma que sean entendibles por cualquier usuario. De esta forma, se ha elaborado un semáforo para determinar a tiempo real la calidad del dato de la API meteorológica de GRAFCAN [1]. Este semáforo nos ofrecerá métricas de la fiabilidad de actualización del dato.

Además, en el estado actual de pandemia, el uso de los datos abiertos ha sido indispensable para numerosas empresas y ciudadanos, los cuales podían saber la incidencia acumulada de casos (entre otros datos de interés) que había en su zona en tiempo real. Debido a la importancia que ha tenido este dato en la sociedad actual, se ha decidido realizar predicciones sobre la incidencia acumulada de la COVID-19 en Tenerife, a partir de los datos abiertos obtenidos.

1.2 Antecedentes

Actualmente, con el auge que han experimentado los portales abiertos de datos públicos, muchas empresas, usuarios y las propias Administraciones Públicas hacen uso de estos portales para sus procesos y negocios. Un estudio [2] realizado por la Comisión Europea y publicado en el portal europeo de datos [3], analiza la madurez de los diferentes países en la política del Open Data, los portales abiertos, el impacto que suponen, y su calidad.

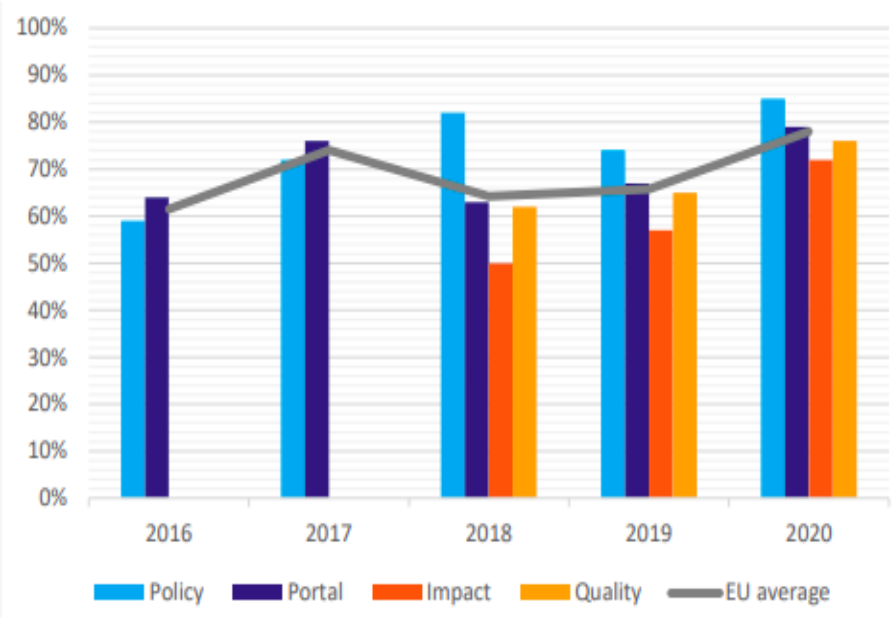


Figura 1.1: Madurez de los datos abiertos en los países Europeos. **Fuente:** Portal Europeo de Datos [3]

Como se puede observar en la Figura 1.1, la Comisión Europea ha añadido en los últimos años un parámetro para determinar la calidad del dato. Esto se debe a la importancia de que los datos tengan un seguimiento y estén actualizados. Por esta razón, cada vez son más las empresas que realizan el proceso de análisis del dato antes de usar un portal abierto. Sin embargo, cada portal de datos que se investiga difiere de los demás.

En lo que respecta a realizar predicciones sobre la COVID-19 con datos abiertos, ha sido un tema recurrente en la actualidad debido a la importancia que presenta en la sociedad. Existen diversos estudios realizados con datos abiertos, basados en modelos matemáticos que ayudan a pronosticar el número de personas infectadas, hospitalizadas e ingresadas en UCI hasta mayo de 2021. Un ejemplo es el caso de La Universidad de Granada [4].

En julio de 2020, en la Conferencia Internacional sobre Tecnologías de la Computación, la Comunicación y las Redes (ICCCNT), se presentó un estudio [5] que analiza la situación de la COVID-19 en los 10 países que más impacto ha dejado la pandemia, y con Prophet y ARIMA (Modelos de predicción basados en Series Temporales) realiza predicciones, determinando que modelo es más eficaz para dicho caso.

Aun así, no se han publicado estudios relacionados con predicción acerca de la COVID-19 en la isla de Tenerife.

1.3 Objetivos

1.3.1 Objetivo General

Como se mencionó anteriormente, existen portales de datos abiertos que no se actualizan periódicamente, o sus datos presentan errores de fiabilidad. El uso de estos portales pueden conllevar a errores drásticos.

Por esta razón, el objetivo en este periodo de elaboración del Trabajo de Fin de Grado, es analizar detenidamente la API de las estaciones meteorológicas de GRAFCAN, determinando la fiabilidad en la obtención del dato y su persistencia de forma automática, haciéndolo entendible para los usuarios.

Además, a partir de datos públicos de la COVID-19 obtenidos de portales abiertos oficiales, se pretende realizar predicciones sobre la incidencia acumulada en la isla de Tenerife.

1.3.2 Objetivos específicos

Los objetivos marcados que derivan del general son:

- Analizar de forma manual la API de GRAFCAN de las estaciones meteorológicas, entendiendo su funcionamiento.

- Elaborar diferentes scripts para almacenar de forma automática en local los datos de la API y determinar la fiabilidad de actualización.

- Crear un semáforo con la herramienta Power BI, de tratamiento y mejora de visualización de datos, que ayude a comprender el estado de la API a tiempo real para que los datos sean entendibles por personas no expertas en la materia.

- Recopilar datos relacionados con la COVID-19, que puedan ser de interés en la época actual de pandemia.

- Con los datos obtenidos, realizar predicciones sobre la incidencia acumulada en Tenerife con Prophet y Neural Prophet.

1.4 Alcance

Para cumplir los objetivos marcados el TFG se divide en dos partes relacionadas entre sí por la obtención de datos abiertos.

La primera parte de este proyecto se basa en la elaboración de un semáforo de la API meteorológica de GRAFCAN, que pueda indicar a un usuario del portal si los datos están disponibles y actualizados. Para ello, se calcula la calidad de obtención del dato y la persistencia del mismo. El proceso mencionado se lleva a cabo durante las primeras 10 semanas. El proceso empleado servirá como base para cualquier conjunto de datos a tratar.

Este proyecto se ha realizado en época de COVID-19, donde el Open Data ha sido de gran ayuda para tener un seguimiento de la pandemia a tiempo real. Como ya existen portales de visualización sobre este tema de interés a tiempo real, actualizados diariamente, no tiene mayor relevancia realizar un análisis de portales con un porcentaje tan elevado de fiabilidad. Por esta razón, la segunda parte de este proyecto se centra en la elaboración de predicciones con los datos obtenidos de estos portales abiertos de máxima fiabilidad. En concreto, se mide la incidencia acumulada a posteriori en la isla de Tenerife, con diferentes modelos. El tratamiento, la normalización de los datos necesarios, y la elaboración de los modelos de predicción se realizan aproximadamente en 8 semanas.

Capítulo 2

Fundamentos Teóricos

2.1 Datos abiertos

Los Datos abiertos (más conocido por su nombre en inglés Open Data) es una iniciativa que pretende que determinados datos estén disponibles de forma libre para cualquier usuario, sin restricciones de derechos de autor. Es decir, datos que sean accesibles y reutilizables, sin exigencia de permisos específicos.

Para que un conjunto de datos pueda ser considerado "abierto", necesita cumplir con algunas características básicas:

- **Disponibilidad y acceso**: los datos deben estar disponibles de manera integral y en una forma que no genere costos exorbitantes para la parte interesada en copiarlos. El escenario más favorable es poner los datos a disposición para que sean descargados a través de Internet en diferentes portales, bases de datos, etc. Además, los datos deben ofrecerse en un formato conveniente y que sea modificable por el usuario. Los formatos más extendidos actualmente son:
    - **Comma-separated values (CSV)**: es el formato más simple para el almacenamiento de datos, ya que se basa en ficheros de texto plano. Se utiliza frecuentemente para el intercambio de información entre distintos programas de hoja de cálculo que utilizan formatos propietarios (como **XLS de Excel**).
    - **JavaScript Object Annotation (JSON**): es un formato ligero para el intercambio de datos basado en la notación literal de objetos de JavaScript.
- **Reutilización y redistribución**: los datos deben ponerse a disposición bajo términos de uso que permitan su reutilización y redistribución, e incluso que puedan entrecruzarse con otros datos.
- **Participación universal:** cualquier persona debe poder usar, reutilizar y redistribuir los datos. Es decir, no debe existir ningún tipo de restricción contra campos de actuación, individuos o grupos.

Como se ha mencionado anteriormente, el Open Data ha presentado un auge extremo en los últimos años debido a su gran utilidad. En España, podemos localizar fuentes de datos abiertos de la Administración del Estado para facilitar su reutilización. Actualmente posee un portal de datos abiertos oficial que recopila más de 45.000 conjuntos de datos, convirtiéndose en un país puntero en el Open data como se puede apreciar en la Figura 2.1

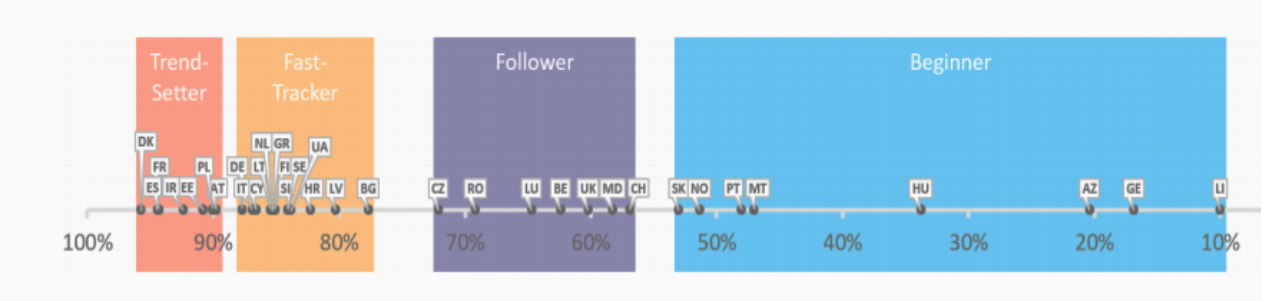


Figura 2.1 : Uso del Open Data en Países Europeos (2020). **Fuente:** Portal Europeo de Datos

En Canarias más concretamente, el Gobierno ha estrenado el nuevo Portal de Datos Abiertos, alojado en datos.canarias.es, con más de 7.500 conjuntos, convirtiéndose en el punto de acceso con más datos de información pública registrado en toda España, formando casi el 20% del catálogo del país.

Otro ejemplo del auge mencionado lo demuestra GRAFCAN S.A, empresa pública de la Comunidad Autónoma de Canarias, que realiza actividades de producción, mantenimiento y gestión de la información geográfica y territorial de la Comunidad Autónoma de Canarias. Entre otros servicios, oferta una API de datos meteorológicos de las islas, que es objeto de estudio en este proyecto.

2.2 Business Intelligence

Business Intelligence hace referencia al uso de diferentes estrategias y herramientas para transformar información en conocimiento, con el objetivo de mejorar el proceso de toma de decisiones.

Estas estrategias y herramientas acceden y analizan conjuntos de datos, presentando hallazgos analíticos en informes, paneles, gráficos, y cuadros de mando, para proporcionar a los usuarios inteligencia detallada sobre el asunto en cuestión.

Tienen en común las siguientes características:

- Fácil accesibilidad a la información. Los datos son la fuente principal de este concepto. Por tanto, se debe garantizar el acceso de los usuarios a los datos con independencia de la procedencia de estos.

- Apoyo en la toma de decisiones. Requiere más que la presentación de la información, de manera que los usuarios tengan acceso a herramientas de análisis que les permitan seleccionar y manipular sólo aquellos datos que les interesen.

- Orientación al usuario final. Tiene que existir cierta independencia entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.

2.2.1 ETL

Un proceso ETL (Del inglés: Extract, Transform, Load) tiene como objetivo producir de manera automática datos limpios y accesibles que puedan utilizarse de manera eficaz. Consta de tres fases:

- **Extraer** datos en bruto desde los sistemas de origen: Un requerimiento que se debe exigir a la tarea de extracción es que ésta cause un impacto mínimo en el sistema origen. Si los datos a extraer son muchos, el sistema de origen se podría ralentizar, provocando que este no pueda utilizarse con normalidad para su uso cotidiano. Por esta razón, en sistemas grandes las operaciones de extracción suelen programarse en horarios o días donde el impacto sea mínimo.
- **Transformar** los datos: En esta fase se produce la operación con mayor importancia del proceso. Se deben aplicar una serie de normas que modifiquen los datos en bruto obtenidos previamente para que presenten el formato idóneo de uso. Estas normas suelen ser: normalización, eliminación de duplicados, verificación de la veracidad de los datos, otras normas que sean de interés para ese proceso en concreto.
- **Cargar** los datos: La última fase consiste en cargar esos datos extraídos y transformados a un nuevo destino.

2.2.2 Cuadros de mando

Para las empresas y los usuarios que trabajan con datos, es de vital importancia conocer lo que sucede en tiempo real, ya que facilita la toma de decisiones. Un cuadro de mando realiza dicha función, al monitorizar todos los parámetros marcados, y disponer de una gráfica a tiempo real de los datos.

Cada proyecto o empresa decide cuál es la estructura más favorable para sus intereses y qué tipo de indicadores es apropiado incluir en él. Suele poner de manifiesto aquellas áreas más relevantes dentro del proyecto.

Una característica que posee es la capacidad de mostrar el camino a cuestiones y decisiones que ni siquiera se han planteado con anterioridad por el costo de llegar a ese dato cruzando múltiples informes. Además, un cuadro de mandos se puede exportar a diferentes formatos (hojas de cálculo, documentos electrónicos, etc.), posibilitando también hacer trazabilidad.

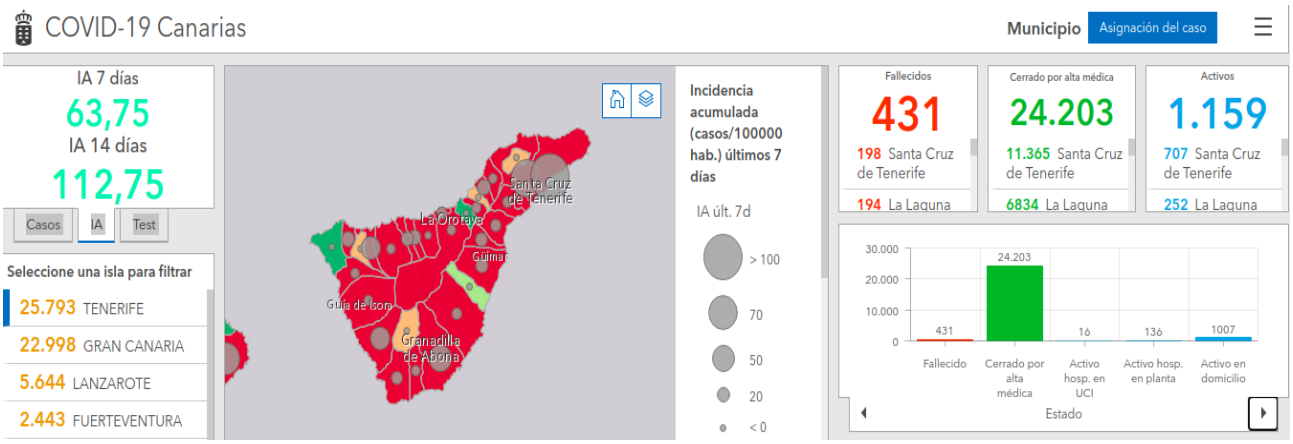


Figura 2.2: Ejemplo de cuadro de mando sobre la COVID-19. **Fuente:** Gobierno de Canarias [6]

2.3 Series Temporales

Una serie temporal se define como una colección de observaciones de una variable recogidas secuencialmente en el tiempo. Estas observaciones se suelen recoger en instantes de tiempo con la misma periodicidad, o bien se acumulan los valores sobre el mismo intervalo de tiempo.

Las series temporales presentan varias características, denominadas componentes, que ayudan a explicar su comportamiento en el tiempo. Estas características son:

- **Tendencia** (componente tendencial): Se refiere a su comportamiento a largo plazo. Indica si una serie temporal es creciente, decreciente, o constante.

- **Componente cíclica**: Refleja comportamientos recurrentes, aunque no tienen por qué ser exactamente periódicos, en intervalos que suelen superar el año.

- **Componente estacional**: recoge el crecimiento o decrecimiento en los valores de la serie que se producen en una determinada época. Puede darse por dos situaciones:

  - Físico-naturales: tiempo meteorológico, ciclos biológicos.
  - Institucionales: vacaciones escolares, etc.

- **Componente irregular** o “Ruido”: busca alteraciones de la serie, sin una pauta periódica ni tendencia reconocible, que en principio tienen poca incidencia.

2.4 Redes Neuronales

Las redes neuronales se inspiran en el funcionamiento del cerebro humano. Están formadas por un conjunto de nodos conocidos como neuronas artificiales, que se conectan entre sí y transmiten señales hasta generar una salida. Estas redes reciben una serie de valores de entrada que llegan a un nodo o neurona. Cada una de las neuronas de la red posee a su vez un valor, con el que modifica la entrada recibida. Los nuevos valores obtenidos salen de las neuronas y continúan su camino por la red.

Las neuronas están agrupadas en capas que forman la red neuronal. Existe la capa de entrada, la capa de salida, y en medio de ambas, pueden haber capas ocultas añadidas que hacen más compleja la red neuronal, y por tanto, las funciones que pueden realizar.

El objetivo que tienen las redes neuronales es aprender modificándose automáticamente a sí mismas. De esta forma, pueden llegar a realizar tareas complejas que serían imposible de ejecutar por un programa basado en reglas.

2.4.1 Redes Neuronales FeedForward

Las redes neuronales feedforward fueron el primer tipo de red neuronal artificial y son unas de las más simples. En este tipo de redes neuronales artificiales, las conexiones entre neuronas no forman un ciclo, como se observa en la Figura 2.3. Se denominan feedforward porque la información solo viaja hacia adelante en la red (sin bucles), primero a través de los nodos de entrada, luego a través de los nodos ocultos (si están presentes) y finalmente a través de los nodos de salida.

Toda red neuronal prealimentada (feedforward) ha de tener una capa de entrada y una capa de salida. El número de neuronas de estas capas varía dependiendo del problema a resolver: la capa de entrada tendrá tantos elementos como atributos tengan los datos que se introduzcan al modelo. Y la capa de salida tendrá tantas neuronas como salidas queramos que tenga el modelo.

A diferencia de las capas de entrada y de salida, el número de capas internas, y el número de neuronas en estas capas son decisiones libres. Es decir, son hiperparámetros que no se establecen durante el proceso de aprendizaje.

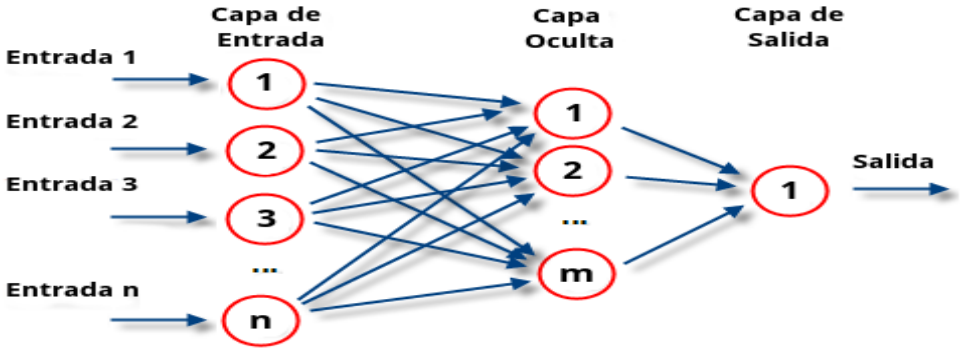


Figura 2.3: Diagrama de una red neuronal feedforward.

2.4.2 AR-Net

En un modelo auto-regresivo clásico, se pronostica la variable de interés usando una combinación lineal de valores pasados de la propia variable. Es decir, la variable se vuelve a comparar consigo misma, utilizando valores pasados de la serie temporal como predictores. Sin embargo, cuando se trata de dependencias de largo alcance, con muchos elementos, los modelos autorregresivos pueden resultar intratables.

Para solucionar dicho problema surge AR-Net [7]. Este modelo imita el proceso autorregresivo tradicional con una red neuronal. Está diseñado de tal manera que los parámetros de su primera capa son equivalentes a los coeficientes AR. Además, se puede ampliar opcionalmente con capas ocultas para lograr una mayor precisión de pronóstico, sin embargo aumenta de manera significativa su complejidad y tiempo de ejecución.

AR-Net permite ajustar un modelo más amplio con coeficientes autorregresivos dispersos, eliminando así la suposición de que los coeficientes deben seguir un orden exacto. Esto hace posible el aprendizaje de dependencias de largo alcance sin sobreajuste. Lo consigue añadiendo un valor de regularización al error que se está minimizando.

2.5 Validación de modelos de predicción

Los métodos de validación, son estrategias que permiten estimar la capacidad predictiva de los modelos cuando se aplican a nuevas observaciones, haciendo uso de los datos de entrenamiento. Es decir, el modelo se ajusta empleando un subconjunto de observaciones (los datos de entrenamiento), y se evalúa con las observaciones restantes.

Existen diversas metodologías, como son la comparación de los parámetros obtenidos con los adquiridos mediante la predicción, utilizar conjunto de datos adicionales para comparar con los obtenidos, o el uso de validación cruzada. Las dos versiones de validación cruzada más utilizada son :

- **Hold-out**: separa el conjunto de datos en dos subconjuntos, uno utilizado para entrenar el modelo y generar datos de salida, luego compararlos con el segundo subconjunto para realizar la validación. Los datos obtenidos representan la validez del modelo en términos del error.

- **K-fold**: divide el total de datos en k subconjuntos de manera de aplicar el método hold-out k veces, utilizando cada vez un subconjunto distinto para validar el modelo entrenado con el histórico restante. El error promedio obtenido nos indica la validez del modelo.

2.5.1 Métricas de Validación

En cuanto a las métricas, sirven para evaluar el rendimiento de los modelos, es decir, permiten estimar cuantitativamente el error obtenido. Las métricas más habituales son:

Error cuadrático medio (MSE): Mide el error cuadrado promedio de las predicciones. Es decir, para cada punto, calcula la diferencia cuadrada entre las predicciones y el objetivo y luego promedia esos valores. Esta métrica potencia las diferencias mayores.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Raíz cuadrada del error cuadrático medio (RMSE): Es la raíz cuadrada de MSE. Hace que la escala de los errores sea igual a la escala de los objetivos

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

Error absoluto medio (MAE) : el error se calcula como un promedio de diferencias absolutas entre los valores objetivo y las predicciones.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

R^2 o Coeficiente de determinación: está relacionada con la MSE, pero tiene en su favor la ventaja de estar libre de escala. No importa si los valores de salida son muy grandes o muy pequeños,ya que siempre estará entre -∞ y 1, siendo 1 el valor óptimo.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y_i})^2}$$

Capítulo 3

Infraestructuras y Tecnologías empleadas

3.1 IAAS de la ULL

Durante la ejecución del proyecto, fueron surgiendo necesidades de software. El sistema operativo empleado es Ubuntu, distribución de Linux basada en Debian de código abierto. Sin embargo, ciertas herramientas como Power BI Desktop, en su versión gratuita, solo permite descarga en Windows. Por tanto, la solución fue utilizar el servicio IAAS [8] que ofrece la ULL a sus estudiantes. Este servicio comenzó con usos puntuales en algunas áreas, y actualmente es de vital importancia por situaciones de semipresencialidad en la Universidad.

Como resultado, se obtuvo una máquina virtual con Sistema operativo Windows Server 2012 R2, con 8 Gb de Memoria Ram,y 32 Gb de memoria en disco duro. De esta manera se pudo instalar Power BI Desktop, después de realizar las configuraciones y actualizaciones pertinentes.

Además, al estar 24 horas en servicio, se decidió automatizar los scripts en esta plataforma para ahorrar en recursos.

3.2 Herramientas

3.2.1 Visual Studio Code

El editor utilizado para la elaboración de los scripts, así como los modelos de predicción, es Visual Studio Code [9]. Este editor está escrito totalmente en Electron, un framework utilizado para unir Chromium y Node.js en forma de aplicación de escritorio. Cuenta con una gran cantidad de opciones de depuración de código para encontrar errores, y optimizarlo.Tiene compatibilidad con Git.

Para poder programar en Python, se instalaron las extensiones correspondiente del lenguaje que incluye un verificador de código fuente, errores y calidad denominado Pylint.

3.2.2 Postman

Inicialmente para analizar, observar los datos y realizar pruebas en la API meteorológica de GRAFCAN se emplea Postman [10]. Dicha herramienta permite realizar peticiones HTTP a cualquier API. Además, guarda y agrupa conjuntos de solicitudes en carpetas de distintos niveles que organizan las peticiones HTTP de forma ordenada.

Una de las características principales es que se puede definir el tipo de petición (GET, POST, etc.), los tokens de autenticación, las cabeceras asociadas a la petición, todo de una manera muy intuitiva. Además, las peticiones almacenadas pueden ser exportadas a múltiples lenguajes.

3.2.3 Microsoft Power BI

Microsoft Power BI [11] es el nombre colectivo para una variedad de aplicaciones y servicios basados en la nube que ayudan a las organizaciones a administrar y analizar datos a través de una interfaz fácil de usar.

Este es un concepto bastante amplio ya que puede referirse a una aplicación de escritorio de Windows llamada Power BI Desktop, un servicio en línea SaaS llamado Power BI Service, o aplicaciones móviles de Power BI disponibles en teléfonos y tablets con Windows, iOS y Android.

La principal función de esta herramienta es proporcionar visualizaciones interactivas y capacidades de business intelligence a los usuarios. Posee una interfaz simple, que permite transformar los datos introducidos, y crear informes y paneles con gráficas más entendibles.

La herramienta más utilizada del paquete de Power BI, es Power BI Desktop. Esta herramienta es útil por diversas razones:

- Permite acceder a datos de cientos de orígenes, tanto locales como basados en la nube.
- Ofrece la posibilidad de modelar los datos.
- Se pueden realizar análisis avanzados con funciones de medidas rápidas, agrupación, previsión, la agrupación en clústeres, etc.
- Permite crear informes interactivos más entendibles por el usuario.

Además, proporciona a los usuarios un control total sobre el modelo mediante el lenguaje de fórmulas DAX [12].

3.3 Librerías de tratamiento de datos y visualización

Para el tratamiento de datos en los diversos scripts realizados en Python se han utilizado las siguientes librerías:

- **Numpy [13]**: Proporciona un objeto de matriz multidimensional de alto rendimiento y herramientas para trabajar con estas matrices. Además, ofrece funciones matemáticas completas, y generadores de números aleatorios entre otras funciones. Se considera una gran librería por la buena manipulación de datos que permite realizar.
- **Pandas [14]:** Es una librería escrita como extensión de NumPy, la cual carga y soporta manipulación de datos de una manera muy eficiente. En concreto, ofrece estructuras de datos, y operaciones para manipular tablas y series temporales. La característica de la librería más usada en este proyecto es leer y escribir datos en archivos csv, así como la mezcla y unión de datasets.
- **Matplotlib [15]:** es una librería para crear visualizaciones en Python. Es decir, permite desarrollar gráficos de calidad con pocas líneas de código.

3.4 Librerías para los modelos de predicción

Los modelos predictivos desarrollados se han elaborado a partir de las siguientes librerías:

- **Prophet [16]:** Empleada para el modelo de Prophet.
- **NeuralProphet [17]:** Librería basada en **Pytorch [18]** para el modelo de Neural Prophet.

Lo que respecta a las métricas para la validación de los modelos predictivos, se usa, aparte de la validación cruzada integrada en Prophet, **sklearn.metrics [19]**, que implementa funciones que evalúan el error de predicción para propósitos específicos

3.4.1 Prophet

Prophet es una librería de código abierto disponible en Python y R, que ofrece un modelo para pronosticar nuevos elementos en una serie temporal. Este procedimiento funciona mejor con series temporales que tienen fuertes variaciones estacionales y varias temporadas de datos históricos. Es robusto ante los cambios tendenciales, y por norma general, maneja bien los valores atípicos.

Una de sus características principales es la sencillez que presenta en su uso, ya que proporciona una manera fácil y personalizada de ajustar los hiperparámetros, siendo entendible incluso por personas que carecen de experiencia en modelos de pronóstico. Sin embargo, una de las características que limitan su uso es que realmente solo permite usar una variable de entrada (sin contar los regresores), para predecir valores futuros de esa misma variable.

En esencia, este modelo consiste en la suma de tres funciones de tiempo más un término de error:

$$y(t) = g(t) + s(t) + h(t) + e(t)$$

- **g(t)**: curva de crecimiento logístico o lineal para modelar la tendencia en las series de tiempo.
- **s(t)**: modela las variaciones estacionales, es decir, lo cambios periódicos (por ejemplo, estacionalidad semanal / anual)
- **h(t)**: el efecto que tienen las vacaciones (proporcionadas por el usuario).
- **e(t)**: el término de error da cuenta de cualquier cambio inusual no acomodado por el modelo.

3.4.2 Neural Prophet

Neural Prophet es una librería hecha sobre PyTorch (biblioteca de aprendizaje automático de código abierto) que incluye las características de Prophet, y añade las ventajas que poseen las redes neuronales. Este “avance” de un modelo a otro está determinado por las siguientes características:

- Modelado de auto-regresión de series temporales usando AR-Net
- Modelado de regresores rezagados utilizando una red neuronal Feed-Forward
- Poder añadir capas ocultas configurables.
- Métricas personalizadas.

Además, se desarrolla en una arquitectura completamente modular que lo hace escalable. Es decir, Neural Prophet, que puede producir pronósticos tanto de un solo paso como de varios, ayuda a construir modelos de pronóstico para escenarios donde existen factores externos, los cuales pueden impulsar el comportamiento de la serie a lo largo del tiempo. El uso de dicha información externa puede mejorar los resultados de la predicción notablemente.

Capítulo 4

Desarrollo

4.1 Parte I: Semáforo de actualización

4.1.1 Introducción

GRAFCAN posee una API, con datos registrados por las estaciones meteorológicas de Canarias. En este proceso se determinará la fiabilidad de actualización del dato, desarrollando un prototipo de observación de su persistencia y calidad de actualización.

4.1.2 Compatibilidad con Sitcan

Antes de comenzar a automatizar el proceso, debe realizarse un análisis manual para esclarecer qué datos son de interés, y qué estructura presentan. En este periodo se pudo observar que el historial de los datos que ofrece la API de Grafcan se recopilan en Sitcan [20] (Sistema de Información Territorial de Canarias) mediante dos datasets:

Estaciones Meteorológicas: Contiene el número de estaciones que se encuentran en servicio actualmente, el nombre de la zona a la que pertenece, las coordenadas, y la fecha en la que se instauró la estación (Figura 4.1). Actualmente cuenta con 46 estaciones, y está creciendo de manera masiva en los últimos meses. Este dataset se actualiza en Sitcan diariamente alrededor de las 2 am.

thing_id	thing_name	location_id	location_description	location_coordinates	date_from
2	MTD3016CP (SN: 0379)	1	Centro C♦vico Social El Risco (El Risco) en Agaete (Gran Car	POINT(-15.7256321875752 28.0461647859535)	2018-12-17 00:00:00+00
3	MTD3016CP (SN: 0380)	3	Casa Juventud Tamaimo (Tamaimo) en Santiago del Teide (Ten	POINT(-16.8193037932352 28.2682964694257)	2019-03-12 00:00:00+00
4	MTD3016CP (SN: 0381)	4	Bodegas Tenegu♦a (Los Canarios) en Fuencaliente (La Palma)	POINT(-17.8473463183208 28.4927577357052)	2019-01-11 00:00:00+00
5	MTD3016CP (SN: 0382)	5	Centro Ciudadano Las Carboneras (Las Carboneras) en Santa	POINT(-16.2770308401483 28.5539678026471)	2019-03-12 00:00:00+00
6	MTD3016CP (SN: 0383)	6	Centro Socio Cultural Caleta Famara (Famara) en Teguise (Lan	POINT(-13.5668375027642 29.1177392752713)	2019-01-23 00:00:00+00
7	MTD3016CP (SN: 0384)	7	Comisar♦a Polic♦a Local Gu♦a en Santa Mar♦a de Gu♦a (G	POINT(-15.6323335157924 28.1405949790433)	2018-12-12 00:00:00+00
8	MTD3016CP (SN: 0385)	8	Comisar♦a Polic♦a Local Teguise en Teguise (Lanzarote)	POINT(-13.555249574747 29.0614678040927)	2019-01-25 00:00:00+00
9	MTD3016CP (SN: 0386)	9	Centro Ciudadano La Cuesta (La Cuesta) en San Crist♦bal de	POINT(-16.2923370717827 28.4671736674559)	2019-03-29 00:00:00+00
10	MTD3016CP (SN: 0387)	10	Protecci♦n Civil Moya en Moya (Gran Canaria)	POINT(-15.5841565416324 28.1074099195746)	2018-12-13 00:00:00+00
11	MTD3016CP (SN: 0388)	11	Oficina Municipal Costa Teguise (Costa Teguise) en Teguise (La	POINT(-13.4900866758893 29.0006945117489)	2019-01-29 00:00:00+00
12	MTD3016CP (SN: 0389)	12	Policia Local Costa Calma (Costa Calma) en P♦jara (Fuertever	POINT(-14.2250280324208 28.166335915468)	2019-02-05 00:00:00+00
13	MTD3016CP (SN: 0390)	13	Biblioteca P♦blica Centro Cultural de Frontera en Frontera (El	POINT(-18.0082353205858 27.7540825469866)	2019-03-13 00:00:00+00
14	MTD3016CP (SN: 0391)	14	Bilbioteca de La Isleta (La Isleta) en Las Palmas de Gran Can	POINT(-15.4206499098891 28.1507154731195)	2018-12-28 00:00:00+00
15	MTD3016CP (SN: 0392)	15	Nuevo Tanatorio (Fataga) en San Bartolom♦ de Tirajana (Gran	POINT(-15.563562172392 27.8867972261262)	2019-01-10 00:00:00+00
16	MTD3016CP (SN: 0393)	16	Hogar del Pensionista en P♦jara (Fuerteventura)	POINT(-14.1084130492027 28.3502473220667)	2019-01-30 00:00:00+00
17	MTD3016CP (SN: 0394)	17	Biblioteca-Centro Cultural Betancuria en Betancuria (Fuertever	POINT(-14.0555729026257 28.424160729349)	2019-02-07 00:00:00+00
18	MTD3016CP (SN: 0395)	18	Centro Cultural La Monta♦eta (La Monta♦eta) en Garachico (T	POINT(-16.7568942931336 28.3403618233128)	2019-02-21 00:00:00+00
19	MTD3016CP (SN: 0396)	19	Centro Ocupacional Los Portales (Los Portales) en Arucas (Gr	POINT(-15.5344265980733 28.0956641103915)	2019-04-24 00:00:00+01
20	MTD3016CP (SN: 0397)	20	Piscina Municipal de Mog♦n en Mog♦n (Gran Canaria)	POINT(-15.7223706023427 27.8831282879566)	2018-12-20 00:00:00+00
21	MTD3016CP (SN: 0398)	21	Centro Cultural Tiscamanita (Tiscamanita) en Tuineje (Fuertever	POINT(-14.0371818183331 28.3511910643586)	2019-01-30 00:00:00+00
22	MTD3016CP (SN: 0399)	22	Centro Usos M♦ltiples Chasna (Camino de Chasna) en La Oro	POINT(-16.5286733217941 28.3671533885171)	2019-06-13 00:00:00+01
23	MTD3016CP (SN: 0400)	23	Centro de Arte, Cultura y Turismo Arrecife (Barrio La Destila) en	POINT(-13.5546816882957 28.9610344943019)	2019-02-28 00:00:00+00
24	MTD3016CP (SN: 0401)	24	Centro Cultural de Tigalate (Tigalate) en Mazo (La Palma)	POINT(-17.8003837390828 28.5541323139345)	2019-01-11 00:00:00+00
25	MTD3016CP (SN: 0402)	25	IES Vega de San Mateo (Las Monta♦etas) en Vega de San M	POINT(-15.5407190121133 28.0030399658059)	2019-07-17 00:00:00+01
26	MTD3016CP (SN: 0407)	26	Malpaso en El Pinar (El Hierro)	POINT(-18.0407378430818 27.7288973490526)	2019-10-01 00:00:00+01
27	MTD3016CP (SN: 0426)	27	IES Josefina de la Torre (Vecindario) en Santa Luc♦a de Tiraja	POINT(-15.4440829548081 27.8576724010744)	2020-03-05 00:00:00+00
28	MTD3016CP (SN: 0425)	28	CEIP La Escalona en Vilaflor (Tenerife)	POINT(-16.6663293485423 28.1187389411721)	2020-09-28 00:00:00+01
29	MTD3016CP (SN: 0404)	29	El Cedro en Hermigua (La Gomera)	POINT(-17.213267646031 28.1363151234253)	2020-10-08 00:00:00+01

Figura 4.1: Dataset de las estaciones meteorológicas de Canarias. **Fuente:** Sitcan

Observaciones Meteorológicas: Recopila el historial de las observaciones meteorológicas desde mitad del año 2019 hasta la actualidad. Se divide en 3 datasets (filtrando los datos por año) en los que se incluye la estación que recoge la observación, la fecha en la que se observó dicho suceso, el fenómeno observado (temperatura del aire, precipitaciones, humedad, etc), el valor del fenómeno acompañado de la medida correspondiente (temperatura -> ºC, lluvia -> milímetros recogidos, etc), y la validación del dato, como se aprecia en la Figura 4.2. Estos datasets se actualizan diariamente en Sitcan a las 2 am aproximadamente. Contienen más de 26 300 000 observaciones.

thing_id	observed_property	phenomenon_time_begin	phenomenon_time_end	result_time	result	unit_of_measurement	observation
23	Relative humidity	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	63.8800	percent	valid
23	Rain	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	0.0000	millimeter	valid
23	Air temperature	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	17.4389	degree celsius	valid
23	Air temperature	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	17.7000	degree celsius	valid
23	Air temperature	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	17.3000	degree celsius	valid
23	Wind direction	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	13.4931	degree angle	valid
23	Wind direction	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	354.0000	degree angle	valid
23	Wind direction	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	20.4426	degree angle	valid
23	Atmospheric pressure	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	1017.8880	millibar	valid
23	Dew point	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	10.6271	degree celsius	valid
23	Precipitation intensity	2020-12-31 23:00:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	0.0000	millimeter per hour	valid
23	Wind speed	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	4.1081	meter per second	valid
23	Wind speed	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	7.7200	meter per second	valid
23	Wind speed	2020-12-31 23:50:00+00	2021-01-01 00:00:00+00	2021-01-01 00:00:00+00	1.2482	meter per second	valid
23	Relative humidity	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	64.3717	percent	valid
23	Rain	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	0.0000	millimeter	valid
23	Air temperature	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	17.2939	degree celsius	valid
23	Air temperature	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	17.4000	degree celsius	valid
23	Air temperature	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	17.2000	degree celsius	valid
23	Wind direction	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	20.1066	degree angle	valid
23	Wind direction	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	24.0000	degree angle	valid
23	Wind direction	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	18.2269	degree angle	valid
23	Atmospheric pressure	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	1017.7650	millibar	valid
23	Dew point	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	10.5859	degree celsius	valid
23	Precipitation intensity	2020-12-31 23:10:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	0.0000	millimeter per hour	valid
23	Wind speed	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	3.4279	meter per second	valid
23	Wind speed	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	6.6900	meter per second	valid
23	Wind speed	2021-01-01 00:00:00+00	2021-01-01 00:10:00+00	2021-01-01 00:10:00+00	1.0579	meter per second	valid
23	Relative humidity	2021-01-01 00:10:00+00	2021-01-01 00:20:00+00	2021-01-01 00:20:00+00	60.6333	percent	valid

Figura 4.2. Conjunto de datos con observaciones meteorológicas de Canarias. **Fuente**: Sitcan

Como se ha mencionado anteriormente, estos datos son accesibles mediante la API meteorológica de GRAFCAN al instante. Se pueden obtener los datos de manera totalmente libre, con una autenticación momentánea que te proporciona la API, pero para una mayor facilidad, se recomienda solicitar una Key. De esta forma, se puede acceder de forma continuada y realizar las peticiones automáticamente mediante Scripts.

4.1.3 Automatización y tratamiento de datos

Para poder automatizar el proceso de recopilación de datos en local, se han realizado diversos scripts en Python. Este es un lenguaje de programación que destaca por su fácil comprensión, velocidad en el procesamiento de datos, y por contener un gran número de bibliotecas que le permiten ser una herramienta de gran ayuda en diferentes campos, como el machine learning o el Open Data.

Los scripts que realizan la ETL (Extract, Transform and Load) de los datos obtenidos de Grafcan para determinar la fiabilidad del portal son:

- El script denominado ‘registros.py’, genera un archivo CSV (Figura 4.3) con diferente información de interés recogida de la API de GRAFCAN que ayuda a determinar la actualización y persistencia del dato. El archivo se denomina ‘datos finales.csv’ y contiene:

    1. El tipo de dato almacenado (temperatura o precipitaciones).
    2. Los registros recibidos ese día de las más de 40 estaciones meteorológicas.
    3. El estado: nos indica si el dato ha sido modificado posteriormente. Si no ha sido modificado se determina mediante la cadena de texto ‘No ha cambiado’. Si el dato se ha actualizado, indica la fecha en la que se produjo el cambio.
    4. La fecha de la primera ejecución para el día en cuestión.
    5. La fecha de la última ejecución.
    6. La fecha del dato.
    7. La fecha de validez del dato.
    8. La columna denominada ‘¿Cuántas veces ha cambiado?’ que nos indica el número de ocasiones que ha sido modificado un registro para ese día en concreto.

	Tipo de Dato	Registros	Estado	Fecha primer	Fecha última	Fecha del Dato	Fecha validez del dato	Cuantas vece
2	Temperatura del aire	6084	No ha cambiado	2021-03-17	2021-04-14	2021-03-15	Validado permanentemente	0
3	Temperatura del aire	6090	No ha cambiado	2021-03-17	2021-04-13	2021-03-14	Validado permanentemente	0
4	Temperatura del aire	6088	No ha cambiado	2021-03-17	2021-04-12	2021-03-13	Validado permanentemente	0
5	Temperatura del aire	6089	No ha cambiado	2021-03-17	2021-04-11	2021-03-12	Validado permanentemente	0
6	Temperatura del aire	6086	No ha cambiado	2021-03-17	2021-04-10	2021-03-11	Validado permanentemente	0
7	Temperatura del aire	6088	No ha cambiado	2021-03-17	2021-04-09	2021-03-10	Validado permanentemente	0
8	Temperatura del aire	6087	No ha cambiado	2021-03-17	2021-04-08	2021-03-09	Validado permanentemente	0
9	Temperatura del aire	6006	No ha cambiado	2021-03-17	2021-04-07	2021-03-08	Validado permanentemente	0
10	Temperatura del aire	5938	No ha cambiado	2021-03-17	2021-04-06	2021-03-07	Validado permanentemente	0
11	Temperatura del aire	5943	No ha cambiado	2021-03-17	2021-04-05	2021-03-06	Validado permanentemente	0
12	Temperatura del aire	5944	No ha cambiado	2021-03-17	2021-03-30	2021-03-05	Validado permanentemente	0
13	Temperatura del aire	5856	No ha cambiado	2021-03-17	2021-03-30	2021-03-04	Validado permanentemente	0
14	Temperatura del aire	5798	No ha cambiado	2021-03-17	2021-03-30	2021-03-03	Validado permanentemente	0
15	Temperatura del aire	5800	No ha cambiado	2021-03-17	2021-03-30	2021-03-02	Validado permanentemente	0
16	Temperatura del aire	5798	No ha cambiado	2021-03-17	2021-03-30	2021-03-01	Validado permanentemente	0
17	Temperatura del aire	5800	No ha cambiado	2021-03-17	2021-03-30	2021-02-28	Validado permanentemente	0
18	Temperatura del aire	5797	No ha cambiado	2021-03-17	2021-03-24	2021-02-27	Validado permanentemente	0
19	Temperatura del aire	5798	No ha cambiado	2021-03-17	2021-03-24	2021-02-26	Validado permanentemente	0
20	Temperatura del aire	5800	No ha cambiado	2021-03-17	2021-03-24	2021-02-25	Validado permanentemente	0

Figura 4.3: Dataset con los datos recopilados de Grafcan tras realizar la ETL. **Fuente:** Diseño propio

La validación del dato se contempla porque en un análisis llevado a cabo sobre la API, se identificó que en algunas ocasiones parte de los datos que se esperan recibir, no llegan al instante. Estos datos se quedan almacenados en un buffer hasta que la estación recupera la conexión, o soluciona el problema detectado, y puede enviar dichos datos correctamente.

Cabe destacar que para validar un dato de forma permanente se ha establecido un periodo de 30 días. Es decir, si el dato recopilado para un día 'x', no se modifica en ese periodo de tiempo, el dato estará validado de forma indefinida, ya que no se espera una actualización con un retraso de tiempo tan elevado. Durante ese periodo de 30 días en el que se lleva a cabo el análisis diario, si el dato no se modifica se describe como 'Válido actualmente'. En cambio, si se actualiza el dato, se genera una fila nueva en el dataset con el nuevo registro, que será la que prosiga con el análisis, y la fila obsoleta se marca como 'No válida'.

- El segundo script realizado permite entrar más en detalle con cada estación meteorológica de forma individual. De esta forma, si se encuentra alguna anomalía en el resultado del primer script, este sirve de apoyo para poder entender en qué estación meteorológica se produjo el error. Es el único script que no se ha automatizado, ya que realiza numerosas peticiones a la API para obtener el valor medio de la temperatura y el máximo de las precipitaciones. Por esta razón, se tomó la decisión de ejecutar manualmente en caso de necesidad. Genera un archivo CSV (Figura 4.4) en el que se puede visualizar:

  1. El tipo de dato a observar (temperatura del aire y/o precipitaciones) por cada una de las estaciones. En periodo de ejecución, se puede definir exportar ambos fenómenos meteorológicos al archivo, o elegir el que sea necesario visualizar, disminuyendo así el tiempo de ejecución.
  2. El valor medio de ese dato durante el día que se ha seleccionado. Es decir, la media de la temperatura obtenida, o el valor máximo de milímetros de agua recolectados durante el día.

3. La cantidad de registros obtenidos por cada estación.

Fecha	Estación	Tipo de dato	Dato	Cantidad de registros
2021-04-05	Gran Canaria Agaete El Risco (109 m)	Temperatura del aire (avg)	17.89784660339355	145
2021-04-05	Tenerife -Santiago del Teide- Tamaimo (574 m)	Temperatura del aire (avg)	13.955753962198887	145
2021-04-05	La Palma- Fuencaliente- Los Canarios (691 m)	Temperatura del aire (avg)	14.18992602030436	145
2021-04-05	Tenerife- Santa Cruz de Tenerife- Las Carboneras (628 m)	Temperatura del aire (avg)	12.926195208231594	145
2021-04-05	Lanzarote- Teguise- Caleta Famara (9 m)	Temperatura del aire (avg)	18.192132695515934	145
2021-04-05	Gran Canaria- Santa María de Guía- Guía (170 m)	Temperatura del aire (avg)	17.196510060628256	145
2021-04-05	Lanzarote- Teguise- Villa de Teguise (323 m)	Temperatura del aire (avg)	14.854120635986325	145
2021-04-05	Tenerife- San Cristóbal de La Laguna- La Cuesta (350 m)	Temperatura del aire (avg)	15.242326609293599	145
2021-04-05	Gran Canaria- Moya- Villa de Moya (486 m)	Temperatura del aire (avg)	14.759116045633942	145
2021-04-05	Lanzarote- Teguise- Costa Teguise (13 m)	Temperatura del aire (avg)	17.69038416544597	145
2021-04-05	Fuerteventura- Pájara- Costa Calma (18 m)	Temperatura del aire (avg)	18.418000666300443	145
2021-04-05	El Hierro- Frontera- Frontera (266 m)	Temperatura del aire (avg)	15.934473927815752	145
2021-04-05	Gran Canaria- Las Palmas de Gran Canaria- La Isleta (6 m)	Temperatura del aire (avg)	19.058618672688805	145
2021-04-05	Gran Canaria- San Bartolomé de Tirajana- Fataga (602 m)	Temperatura del aire (avg)	14.370521418253578	145
2021-04-05	Fuerteventura- Pájara- Pájara (201 m)	Temperatura del aire (avg)	17.187885284423825	145
2021-04-05	Fuerteventura- Betancuria- Betancuria (409 m)	Temperatura del aire (avg)	14.81301256815592	145
2021-04-05	Tenerife- Garachico- La Montañeta (922 m)	Temperatura del aire (avg)	7.991214370727539	145
2021-04-05	Gran Canaria- Arucas- Los Portales (450 m)	Temperatura del aire (avg)	14.769722048441563	145
2021-04-05	Gran Canaria- Mogán- Mogán (281 m)	Temperatura del aire (avg)	17.59277458190916	145
2021-04-05	Fuerteventura- Tuineje- Tiscamanita (244 m)	Temperatura del aire (avg)	16.60986194610596	145
2021-04-05	Tenerife- La Orotava- Camino de Chasna (812 m)	Temperatura del aire (avg)	0	0
2021-04-05	Lanzarote- Arrecife- Arrecife (4 m)	Temperatura del aire (avg)	18.06020393371582	145
2021-04-05	La Palma- Mazo- Tigalate (664 m)	Temperatura del aire (avg)	12.850883293151867	145

Figura 4.4. Conjunto de datos obtenido mediante una ETL con la temperatura media del día . **Fuente:** Diseño propio

- El tercer script elaborado transforma los datos obtenidos de GRAFCAN dejando un porcentaje, el cual determina la cantidad de datos que fueron publicados en el periodo esperado. De esta forma, podemos determinar cual es el valor entre 0-100%. La información que posee el archivo se puede observar en la Figura 4.5. Incluye:

    1. Los datos que se esperan recibir entre todas las estaciones meteorológicas.
    2. Los datos que se han recibido para el día en cuestión, durante la primera ejecución.
    3. El porcentaje entre datos esperados y datos recibidos, tanto para la temperatura del aire, como para las precipitaciones.

Fecha	Datos esperados	Datos recibidos (aire)	porcentaje calidad (aire)	Datos recibidos (pre)	porcentaje calidad (pre)
2021-06-27	6670	6380	95.7	6380	95.7
2021-06-26	6670	6380	95.7	6380	95.7
2021-06-25	6670	5942	89.1	5942	89.1
2021-06-24	6670	6044	90.6	6045	90.6
2021-06-23	6670	6334	95.0	6334	95.0
2021-06-22	6670	6230	93.4	6230	93.4
2021-06-21	6670	6227	93.4	6227	93.4
2021-06-20	6670	6335	95.0	6335	95.0
2021-06-19	6670	6380	95.7	6380	95.7
2021-06-18	6670	6106	91.5	6105	91.5
2021-06-17	6670	6386	95.7	6386	95.7
2021-06-16	6670	6524	97.8	6525	97.8
2021-06-15	6670	6522	97.8	6621	99.3
2021-06-14	6670	6523	97.8	6670	100.0
2021-06-13	6670	6234	93.5	6233	93.4
2021-06-12	6670	6232	93.4	6233	93.4
2021-06-11	6670	6293	94.3	6293	94.3
2021-06-10	6670	6379	95.6	6379	95.6
2021-06-09	6670	6234	93.5	6235	93.5
2021-06-08	6670	6310	94.6	6308	94.6
2021-06-07	6670	6438	96.5	6493	97.3
2021-06-06	6670	6476	97.1	6616	99.2
2021-06-05	6670	6523	97.8	6665	99.9

Figura 4.5. Conjunto de datos obtenido mediante una ETL con los datos esperados y recibidos . **Fuente:** Diseño propio

- El script denominado ‘persistencia.py’, transforma los datos del archivo principal ‘datos finales.csv’ para visualizar de forma más clara la persistencia de la API. De esta forma, será mucho más sencillo introducir los datos en herramientas de visualización posteriormente. La información que aporta dicho script es observable en la Figura 4.6, e indica :

    1. Las veces que se ha modificado el dato a posterior para un día en concreto. Como se mencionó, esto suele ocurrir cuando el sistema se queda sin conexión y no puede enviar todos los datos, o cuando se añade una nueva estación que todavía está en periodo de pruebas. Con lo cual, los datos se quedan almacenados en un buffer hasta que pueden ser enviados.
    2. El valor cuantitativo de dicho cambio. Es decir, el número de datos que se recibieron después de la fecha esperada. Esto se aplica tanto a la temperatura del aire como a las precipitaciones.

Fecha	Veces que ha cambiado (aire)	Valor del cambio(aire)	Veces que ha cambiado (pre)	Valor del cambio (pre)
2021-06-27	0	0	0	0
2021-06-26	0	0	0	0
2021-06-25	1	435	1	435
2021-06-24	2	332	2	332
2021-06-23	1	38	1	38
2021-06-22	1	145	1	145
2021-06-21	2	138	2	138
2021-06-20	1	45	1	45
2021-06-19	0	0	0	0
2021-06-18	1	273	1	273
2021-06-17	0	0	0	0
2021-06-16	0	0	0	0
2021-06-15	0	0	0	0
2021-06-14	0	0	0	0
2021-06-13	1	290	1	435
2021-06-12	1	290	1	435
2021-06-11	1	230	1	375
2021-06-10	1	145	1	290
2021-06-09	2	290	2	435
2021-06-08	3	292	3	436

Figura 4.6. Conjunto de datos obtenido mediante una ETL con la persistencia. **Fuente:** Diseño propio

4.1.4 Visualización en Power BI

A la hora de utilizar datos es de vital importancia que sean entendibles para el usuario. Mejorar la visualización puede parecer un tema simplemente de diseño, pero a la larga se obtienen ganancias de tiempo y de comprensión en la interpretación de los datos. En este caso, para la elaboración del cuadro de mandos se ha utilizado la herramienta Power Bi Desktop, creando un semáforo que indica el estado de la API meteorológica de GRAFCAN, basándose en los datos obtenidos de forma automática mediante los scripts mencionados anteriormente.

Para comenzar con la elaboración del semáforo, se cargan los archivos con extensión CSV, que están automatizados en el IAAS. Estos archivos poseen los datos necesarios ya transformados. Sin embargo, se debe hacer alguna modificación para que la lectura sea adecuada:

- Se indica que la primera fila del conjunto de datos es el encabezado, de esta manera no tomará los nombres de las columnas como datos.

- Se debe indicar el tipo de dato de las columnas que no detecta la herramienta. En determinadas ocasiones, las fechas, los porcentajes, y los datos decimales no son detectados de forma automática al cargar el dataset.

- Se marca la opción de 'Eliminar filas vacías', ya que puede que el dataset contenga filas sin valor que solo aportarían problemas al cuadro de mando.

Una vez cargados los distintos datasets, se encuentran de forma independiente en la herramienta. Existen algunos datos que necesitan estar relacionados entre sí, por lo que se deben indicar las dependencias que existen entre ellos. En este caso, la fecha de todos los datasets es el mismo dato, por tanto, se crea una relación de tipo uno a uno bidireccional.

El último paso antes de realizar las visualizaciones trata sobre DAX. Este lenguaje de fórmulas ayuda a crear información nueva a partir de datos ya incluidos en nuestro modelo. De esta forma, se han elaborado las siguientes medidas:

- **Datos Esperados**: se obtiene el valor total de los datos que se deben haber recibido por la API en el periodo de tiempo indicado. Esto se realiza sumando los valores de la columna 'Datos esperado', indicando con la función 'Datesbetween()' si se requieren los últimos 15 días, o el historial que se dispone de análisis.

- **Datos Recibidos:** Para saber los datos que se han recibido desde que se empezó el análisis, se suman los valores de la columna 'Datos recibidos'. También con la función 'Datesbetween()', se especifica el valor para el historial y para los 15 últimos días. Esta medida se crea tanto para las precipitaciones como para la temperatura.

- **Datos que faltan:** Esta medida sirve para conocer cuántos valores ha dejado de publicar la API, tanto para el apartado de temperatura como el de precipitaciones. Para ello, al valor esperado se le restan los datos recibidos y los datos de persistencia que se han añadido posteriormente. Se crean dos medidas, una para el historial y otra para los últimos 15 días.

- **Suma porcentaje de calidad:** Esta medida suma todos los valores de la columna de porcentajes, creando un valor total para la temperatura y otro para las precipitaciones. Además, se genera una medida para contabilizar los últimos 15 días, y otra para el historial.

- **Porcentaje total:** Se obtiene el valor que se mostrará visualmente en las gráficas. Para crear dicha medida, se divide la suma del porcentaje de calidad hallado anteriormente, entre el número de días establecido (15 días o el historial). Además, se le suma en porcentaje el valor de datos de persistencia. Obteniendo el total de los datos recibidos en porcentaje.

Una vez creadas las medidas principales, se empieza a elaborar el cuadro de mandos. Este se dividirá en tres páginas que indican la calidad de obtención en el momento esperado, la persistencia, y el resultado global de la API. Cada página tendrá en su parte izquierda los datos de la temperatura, y en la derecha los valores relacionados con las precipitaciones, mostrando las mismas gráficas.

La página principal muestra un estado global de la API (temperatura y precipitaciones), midiendo la calidad de obtención del dato y su persistencia. De esta manera, un usuario que quiera saber si puede confiar en el portal de GRAFCAN, tan solo tendrá que observar dicha página. Como se puede apreciar en la Figura 4.7 **,** las gráficas superiores nos dan información del histórico, es decir, desde que se empezó a realizar el análisis hasta 15 días atrás.

Para la temperatura, a día 21 de junio se muestra que un 98.91% de los datos han sido publicados. Además, en la tabla inferior se puede observar un desglose de los datos, sabiendo que se deberían haber publicado 670 915 registros, se han obtenido 663 379 en total, y faltan 7 536.

En el caso de las precipitaciones, el 21 de junio se obtuvo que un 98.98% de los registros se han publicado. En el desglose de los datos se puede apreciar que se han recibido 663 733 registros de los 670 915 que se esperaban, faltando 7 182 valores.

Figura 4.7: Histórico de datos almacenados en Grafcan desde que se inició el estudio . **Fuente:** Diseño propio

La parte inferior de la página muestra el porcentaje global de la API (temperatura y precipitaciones) obtenido en los últimos 15 días como se analiza en la Figura 4.8. En el lado izquierdo podemos observar que a día 21 de junio, se ha recibido un 97.555% de los datos de temperatura, siendo estos 97 608 de los 100 050 esperados.

A su vez, en la parte derecha se observa que para el mismo día, se ha obtenido un 98.944%, es decir, solo faltan 1 046 datos de precipitaciones

Las gráficas se muestran verde si el porcentaje es superior al 90%, amarillo si está entre el 75% y el 90%, y rojo cualquier otro valor inferior. Además cabe destacar que desde esta página principal del informe, se puede acceder a la página de calidad de publicación del dato en el momento esperado , y persistencia del dataset, 'clickando' en el icono de información.

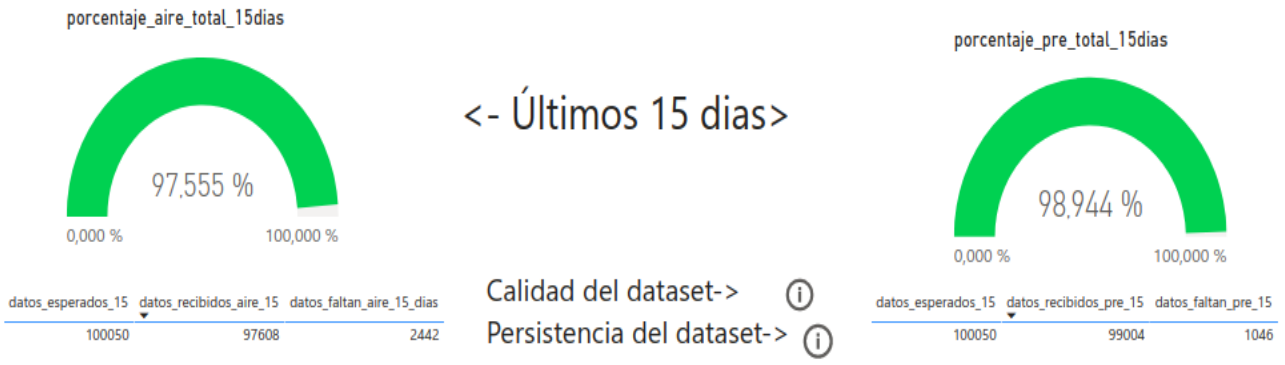


Figura 4.8: Datos almacenados en Grafcan los últimos 15 días . **Fuente:** Diseño propio

En cuanto a la página de calidad del dato, se centra en la cantidad de valores que se han obtenido en el día esperado ,ya que muchas veces se reciben valores a posterior. Esta página se divide igual que la principal, la parte izquierda está dedicada a la temperatura, y la parte derecha a las precipitaciones. Se dispone de tres gráficas por fenómeno observado (temperatura del aire y lluvia).

En la parte superior de la página hay dos gráficas para la temperatura del aire y dos para las precipitaciones, como se observa en la Figura 4.9. En la primera gráfica se miden los datos obtenidos por el conjunto de todas las estaciones para cada día, en su primera publicación. Es decir, sin modificaciones, ni actualizaciones posteriores, son los datos que se publicaron para un día concreto al instante.

La segunda gráfica, denominada gráfica de líneas, indica visualmente la diferencia que existe entre los valores que se esperaban recibir por primera vez, y los que se recibieron. Como se puede observar, los valores esperados aumentan cada vez que se añade una estación a la API. Además, los valores recibidos aunque no suelen ser el total de los esperados en la primera publicación, se suele acercar bastante, lo que muestra la calidad de la API.

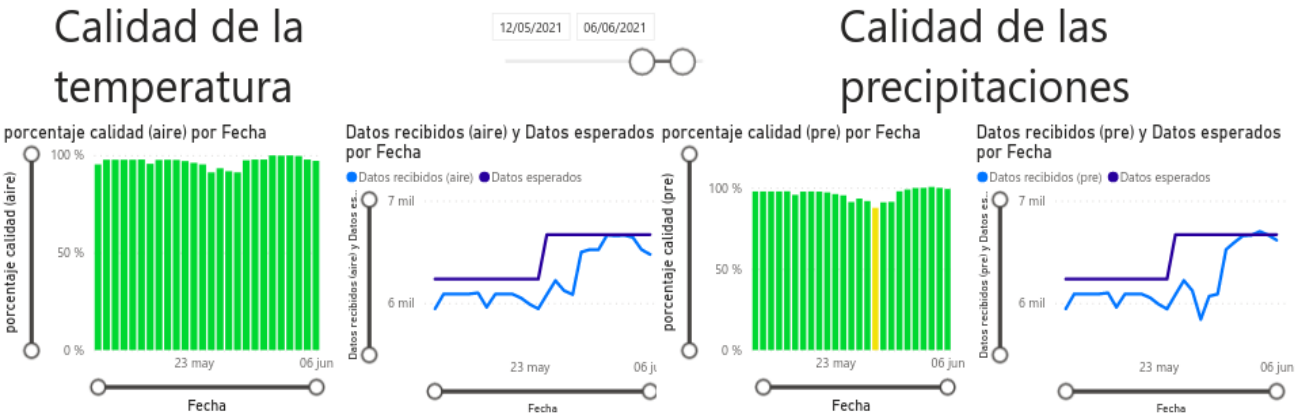


Figura 4.9: Gráficas del semáforo que miden los datos publicados en el periodo esperado . **Fuente:** Diseño propio

En la parte inferior, como se aprecia en la Figura 4.10, se crea una tabla para precipitaciones y otra para la temperatura, en la que se puede entrar más en detalle a analizar los datos diarios. En esta tabla se indica:

- La fecha en formato 'Dia de la semana, Dia del mes, Mes, Año'
- El valor esperado por el conjunto de las estaciones para dicho día
- Los datos recibidos ese día
- El porcentaje de calidad, dividiendo los valores recibidos entre los valores esperados.
- Un indicador con el color del semáforo, siendo verde si el porcentaje es superior al 90%, amarillo si está entre el 75 % y el 90 %, y rojo cualquier otro valor inferior.

Cabe destacar que en la página se añadió un selector de fechas, gracias al cual se puede seleccionar el periodo de fecha que se desea analizar, ampliando las gráficas y marcando los valores elegidos.

Fecha	Datos esperados	Datos recibidos (pre)	porcentaje calidad (pre)	Lightpre
miércoles, 12 de mayo de 2021	6235	5945	95,30 %	
jueves, 13 de mayo de 2021	6235	6090	97,70 %	
viernes, 14 de mayo de 2021	6235	6089	97,70 %	
sábado, 15 de mayo de 2021	6235	6090	97,70 %	
domingo, 16 de mayo de 2021	6235	6090	97,70 %	
lunes, 17 de mayo de 2021	6235	6100	97,80 %	
martes, 18 de mayo de 2021	6235	5960	95,60 %	
miércoles, 19 de mayo de 2021	6235	6088	97,60 %	
jueves, 20 de mayo de 2021	6235	6090	97,70 %	
viernes, 21 de mayo de 2021	6235	6088	97,60 %	

Figura 4.10: Tabla detallada de precipitaciones que muestra los datos recibidos. **Fuente:** Diseño propio

Por último, la página de persistencia se centra en los valores que se añaden posteriormente para un día concreto. Es decir, son valores que deberían haber sido almacenados desde un principio, pero se han ido agregando con el paso de los días. Por tanto, aunque la API presente un gran porcentaje de datos recolectados como se analizó en la página principal, algunos de esos datos se envían posteriormente, y no se encuentran disponibles en el momento inicial. Esta pagina presenta el mismo formato que las anteriores, es decir, el lado izquierdo se basa en la temperatura, y el lado derecho en las precipitaciones.

La parte superior contiene una gŕafica para la temperatura y otra idéntica para las precipitaciones, como se puede observar en la Figura 4.11. Esta gráfica mide el valor del cambio/actualización de cada día. Es decir, los valores que se han ido añadiendo después de que se publicara por primera vez en la API un valor inicial para dicho día. Si el valor del cambio es inferior a 100 se marca en verde, si se encuentra entre 100 y 200 valores añadidos en amarillo, y cualquier valor superior se identifica como rojo. Los días que no muestran ningún valor en la gráfica, indican que no se han actualizado posteriormente, es decir, se mantiene el valor inicial.

Cabe destacar que, aunque se observan muchos valores del semáforo en color rojo o amarillo, esto se debe a que el umbral se estableció en un valor pequeño. Cada estación aporta una media de 145 datos al día, por lo que en conjunto, dependiendo del número de estaciones activas del momento, suelen recopilar entre 6 000 y 7 000 datos. En comparación con dicho valor, que se actualicen alrededor de 200 valores posteriormente no es excesivo. Sin embargo, se tomó el umbral en base a que se actualice una cantidad de datos superior a una estación meteorológica.

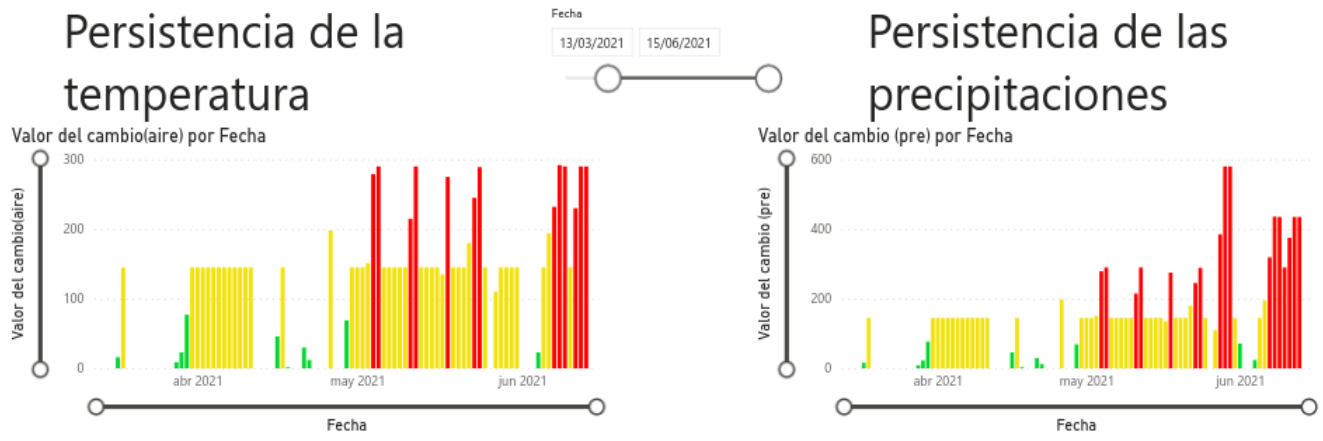


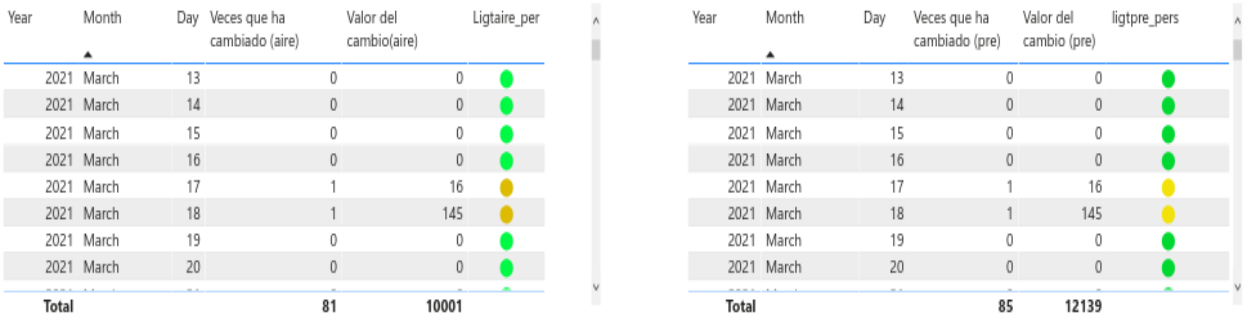
Figura 4.11: Gráficas del semáforo que mide la persistencia. **Fuente:** Diseño propio

En la parte inferior se elaboró una tabla que nos muestra los valores de persistencia de una forma más cuantitativa, se puede analizar en la Figura 4.12. Existe

una tabla tanto para precipitaciones como para la temperatura del aire. Los datos recogidos son:

- La fecha en formato ‘año, mes, día del mes’ .
- El número de veces que ha cambiado el valor. Es decir, el número de actualizaciones recibidas en días posteriores a su primera publicación en la API.
- El valor del cambio, que se obtiene restando al último registro de datos recibidos almacenado, el primer registro publicado para dicho día.
- Un indicador de color, que en esta ocasión indica el numero de veces que se ha actualizado el dato. Si el número de actualizaciones es 0, es decir, desde la primera publicación se han aportado los datos necesarios, la persistencia es correcta y se pone el indicador de color verde. Si se actualiza una vez dicho valor, se pone el semáforo en amarillo. Cualquier otro valor superior se indica como rojo.

Se debe apuntar que, al igual que en la página anterior, existe un selector de fecha para seleccionar los datos indicados.



Year	Month	Day	Veces que ha cambiado (aire)	Valor del cambio(aire)	Ligtaire_per
2021	March	13	0	0	
2021	March	14	0	0	
2021	March	15	0	0	
2021	March	16	0	0	
2021	March	17	1	16	
2021	March	18	1	145	
2021	March	19	0	0	
2021	March	20	0	0	
Total			81	10001	

Year	Month	Day	Veces que ha cambiado (pre)	Valor del cambio (pre)	ligtpre_pers
2021	March	13	0	0	
2021	March	14	0	0	
2021	March	15	0	0	
2021	March	16	0	0	
2021	March	17	1	16	
2021	March	18	1	145	
2021	March	19	0	0	
2021	March	20	0	0	
Total			85	12139	

Figura 4.12: Tablas detalladas del semáforo que miden las veces que ha cambiado un dato. **Fuente:** Diseño propio

Otra ventaja que presenta el cuadro de mando se hace visible a la hora de seguir la trazabilidad del dato. De esta forma, cuando se selecciona una fecha concreta en cualquiera de las gráficas, no solo se amplía el dato en dicha gráfica, sino que se marca dicha fecha en todas las gráficas de la página (tanto para la temperatura como para las precipitaciones), dejando los diferentes parámetros de análisis al descubierto. Este hecho se puede comprobar en la Figura 4.13, y se debe a la relación existente entre los Datasets. Es una característica que facilita la observación, y a la larga, permite ahorrar tiempo y recursos en el análisis.

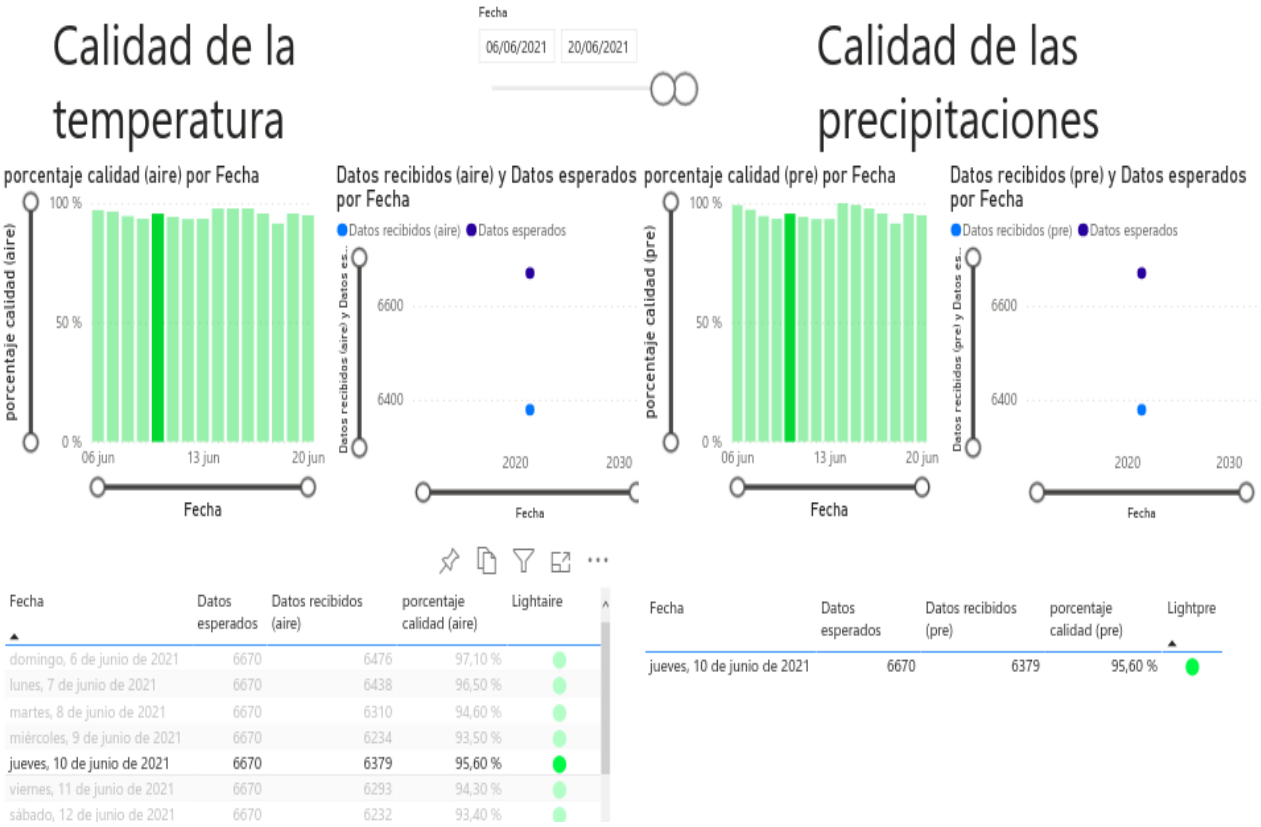


Figura 4.13: Sencillez en la búsqueda de datos entre gráficas. **Fuente:** Diseño propio

4.1.5 Resultados del Análisis

Después de estudiar los resultados obtenidos en este periodo de tiempo, se puede afirmar que la API meteorológica de GRAFCAN, en lo que respecta a los fenómenos atmosféricos de temperatura y precipitaciones, es altamente fiable.

Por un lado, se recoge que el conjunto de los datos obtenidos en su totalidad es muy elevado. Como se mencionó anteriormente, a día 21 de junio, el historial muestra que se han recogido un 98.91% de los datos esperados en caso de la temperatura, y un 98.98% en las precipitaciones. Es decir, de los 670 915 datos que se debían recibir durante estos meses, se han recibido 663 379 registros en el caso de la temperatura y 663 733 de precipitaciones, faltando únicamente 7 536 y 7 182 respectivamente. Estos valores son extremadamente pequeños cuando se trata de un periodo tan extenso de tiempo, y con tantos datos recibidos.

En cuanto a la calidad de obtención del dato en su primera publicación, aunque por lo general no es tan elevado como en los datos totales, ya que suele estar entre 92% y 97%, ha recogido máximas del 100%, su mínimo ha sido de 87.6% en el caso de las precipitaciones, y 91.2% en el caso de la temperatura. Lo que demuestra que la mayoría de los datos de la API son publicados el mismo día en el que se registra el fenómeno observado.

La persistencia nos permite observar la cantidad de datos que se han añadido posteriormente, y las veces que ha sido modificado. La realidad es que aunque se le haya puesto un mínimo al indicador del semáforo en Power BI tan bajo, los datos que ofrece son favorables. Una gran cantidad de días en el histórico no arroja datos, ya que los registros publicados por primera vez suman el total de los datos esperados. Por lo general, el valor de persistencia cuando existe suele ser de 145 registros (número de datos que se espera de una estación al día), lo que indica que alguna estación meteorológica no pudo enviar los datos a tiempo, y fueron publicados posteriormente. Presenta máximas de 580 datos.

Otro característica a favor de GRAFCAN, es que los datos que no se publican en el momento inicial, por lo general se publican uno o dos días después de lo previsto, por lo que trabajan con rapidez para solucionar los posibles problemas que van surgiendo.

Por tanto, los resultados muestran que se pueden usar los datos con tranquilidad, que la gran mayoría de estos datos se publican en el periodo previsto, y que se actualiza con eficacia.

4.2 Parte II: Predicción con datos COVID-19

4.2.1 Introducción

El Open Data deja de tener utilidad si a los ciudadanos o las empresas que usan dichos datos, no se le soluciona ningún problema o se le aporta conocimiento. Gracias al desarrollo tecnológico y la importancia que está cobrando esta filosofía, cada vez se recopilan datos de más ámbitos, lo que proporciona nuevas utilidades a sus “clientes”.

Como se mencionó anteriormente, ya existen portales de visualización con datos de la COVID-19 que se actualizan diariamente y presentan una fiabilidad muy elevada, por tanto no tiene sentido realizar un análisis más profundo de dichos portales. Por el contrario, elaborar predicciones con estos datos de interés obtenidos de portales abiertos es un campo que se ha explorado en menor medida. Aunque se han realizado predicciones en diferentes comunidades y países con diversos métodos, en Tenerife es un campo por explorar, y es lo que se va a investigar a lo largo de este apartado. Se realizarán predicciones sobre la incidencia acumulada.

4.2.2 Datos

Antes de comenzar a realizar predicciones sobre la COVID-19 con diferentes modelos, se necesita recopilar los datos de interés en un dataset. Después de estudiar con detenimiento qué datos pueden llegar a ser efectivos, se ha decidido recopilar de diferentes portales abiertos oficiales:

Incidencia Acumulada: Se ha realizado un script para obtener la incidencia acumulada de 7 y 14 días en la isla de Tenerife. Se dispone de un archivo con extensión CSV del Gobierno de Canarias con cada caso de COVID-19 contabilizado en las islas.

Para obtener el histórico de la incidencia acumulada, se accede a dicho archivo y se contabiliza para cada día, los casos obtenidos durante un intervalo de 7 y 14 días en la isla de Tenerife. Como la incidencia acumulada se expresa en 'número de casos por cada 100.000 habitantes', el dato obtenido no es el dato final. Se debe dividir la población de la isla de Tenerife (929.470) entre 100 000 habitantes, con lo cual se obtiene 9,29 aproximadamente en la actualidad. El número de casos dividido entre dicho valor devuelve la Incidencia Acumulada de la Isla.

Pasajeros llegados a Tenerife: Un dato que puede estar relacionado con el crecimiento de la curva es la llegada de turistas a las islas. Por tanto, se ha decidido obtener los datos de los pasajeros procedentes de Alemania, Gran Bretaña, Nacionales, Canarias, u otro lugar del extranjero. El portal del cual se obtienen los datos es [21]

Fase actual de la isla: Este dato engloba el porcentaje de aforo en interior y exterior, el número máximo de personas habilitados por el gobierno para reunirse, y el horario inicial y final del toque de queda.

Días festivos: Se ha creado un dataset con los días festivos regionales, y nacionales, así como periodos de mayor movilidad en la isla como Navidad, Carnavales, Semana Santa o Vacaciones de Verano, ya que son épocas en las que la curva de Incidencias puede aumentar.

En primera instancia se ha decidido llevar a cabo predicciones con periodos semanales, sujeto a posibles cambios para poder mejorar el modelo en uso. Los datos, al provenir de diferentes conjuntos, tienen de forma predeterminada periodos de tiempo diversos. Por tanto, se ha realizado un script que almacena cada lunes desde enero de 2020, con los datos extraídos de los conjuntos anteriores, y se recopilan todos en un mismo dataset mediante el uso de Pandas. Esto se puede observar en la Figura 4.14

Fecha	aforo	aforo_ext	aforo_com	reuniones	ini_toque_Q	fin_toque_Q	festi	ia7	semana	numero de semana	procedentes gb	procedentes de	correo	mercancias
2020-01-13	100 %	100 %	100 %	No hay maximo	No habia	No habia	0	0.0	s2-1-2020	1	422.2	3061.8	38808.0	201018.2
2020-01-20	100 %	100 %	100 %	No hay maximo	No habia	No habia	0	0.0	s3-1-2020	2	422.2	3061.8	38808.0	201018.2
2020-01-27	100 %	100 %	100 %	No hay maximo	No habia	No habia	0	0.0	s4-1-2020	3	422.2	3061.8	38808.0	201018.2
2020-02-03	100 %	100 %	100 %	No hay maximo	No habia	No habia	0	0.0	s1-2-2020	4	422.2	3061.8	38808.0	201018.2
2020-02-10	100 %	100 %	100 %	No hay maximo	No habia	No habia	0	0.0	s2-2-2020	5	163.75	2700.0	51552.75	247215.25
2020-02-17	100 %	100 %	100 %	No hay maximo	No habia	No habia	0	0.0	s3-2-2020	6	163.75	2700.0	51552.75	247215.25
2020-02-24	100 %	100 %	100 %	No hay maximo	No habia	No habia	3	0.0	s4-2-2020	7	163.75	2700.0	51552.75	247215.25
2020-03-02	100 %	100 %	100 %	No hay maximo	No habia	No habia	0	0.54	s1-3-2020	8	163.75	2700.0	51552.75	247215.25
2020-03-09	100 %	100 %	100 %	No hay maximo	No habia	No habia	0	1.08	s2-3-2020	9	103.8	2808.0	41640.2	239604.4
2020-03-16	0 %	0 %	0 %	0	Todo el dia	Todo el dia	0	8.18	s3-3-2020	10	103.8	2808.0	41640.2	239604.4
2020-03-23	0 %	0 %	0 %	0	Todo el dia	Todo el dia	0	29.69	s4-3-2020	11	103.8	2808.0	41640.2	239604.4
2020-03-30	0 %	0 %	0 %	0	Todo el dia	Todo el dia	0	50.57	s5-3-2020	12	103.8	2808.0	41640.2	239604.4
2020-04-06	0 %	0 %	0 %	0	Todo el dia	Todo el dia	2	26.9	s1-4-2020	13	103.8	2808.0	41640.2	239604.4
2020-04-13	0 %	0 %	0 %	0	Todo el dia	Todo el dia	0	18.51	s2-4-2020	14	118.25	2472.25	42030.0	215379.5
2020-04-20	0 %	0 %	0 %	0	Todo el dia	Todo el dia	0	9.36	s3-4-2020	15	118.25	2472.25	42030.0	215379.5
2020-04-27	0 %	0 %	0 %	0	Todo el dia	Todo el dia	1	6.89	s4-4-2020	16	118.25	2472.25	42030.0	215379.5

Figura 4.14: Recopilación de datos de interés mediante Pandas para los modelos de predicción. **Fuente:** Diseño propio

En algunos casos -como los pasajeros llegados a Tenerife- al proporcionar la API el dato de forma mensual, se tuvo que dividir entre el número de semanas que tiene el mes, para tener una aproximación semanal. Otro caso es el de los días festivos, ya que se tuvo que crear una columna con la suma de los días no lectivos que incluye la semana, a partir del dataset inicial de festividades.

4.2.3 Modelo Prophet

Una vez recopilados los datos que pueden llegar a ser necesarios, se comienzan a elaborar los modelos de predicción. El primer modelo se realiza con Prophet, mencionado anteriormente en el apartado 3.4.1.

En Prophet, realmente solo se permite usar una variable de entrada. Se pueden añadir días festivos, y regresores, pero el tipo de regresor que admite debe ser un dato conocido tanto para el historial como para fechas futuras. Por tanto, no podemos añadir como regresor el número de pasajeros, ni el aforo de los establecimientos, ya que no conocemos su valor futuro.

En primera instancia, se introdujeron datos semanales de la incidencia acumulada para entrenar el modelo.

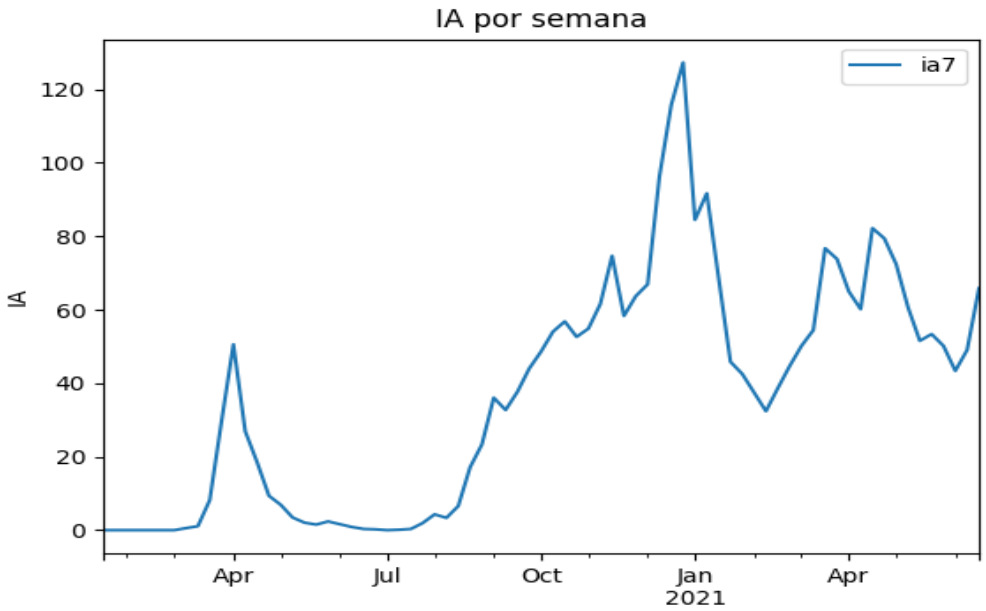


Figura 4.15: Incidencia acumulada semanal desde enero de 2020. **Fuente:** Diseño propio

Prophet contiene una gran cantidad de hiperparámetros que pueden modificar el resultado del modelo drásticamente. En el proceso de entrenamiento se fue probando con diferentes hiperparametros para ver los efectos resultantes. Sin embargo, los datos de la incidencia acumulada desde enero de 2020 hasta agosto de 2020 son muy poco precisos, como se puede apreciar en la Figura 4.15 La gran mayoría de estos valores, exceptuando el pico de abril, son 0. Esto se debe a que en dichos meses no se llevaba un control tan exhaustivo en la contabilización de los datos por el temor en la sociedad al virus. Además, fue una época en la cual la sociedad española estaba en confinamiento, por lo que no variaron las fases del estado de alarma y no se tenía en cuenta la incidencia acumulada sino el número de casos diarios. En el modelo se pudo ver reflejado dicho suceso, arrojando resultados poco precisos como se aprecia en la Figura 4.16. Las métricas que figuran en la Tabla 4.1 están elaboradas con la función de validación cruzada incluida en Prophet. Sin embargo, no se realiza con k-pliegues, debido a la escasez de datos de entrenamiento. Por tanto, se utiliza el último pliegue para la validación, simulando un entrenamiento con el 90% del conjunto de datos, y el 10% final destinado a validar el modelo.

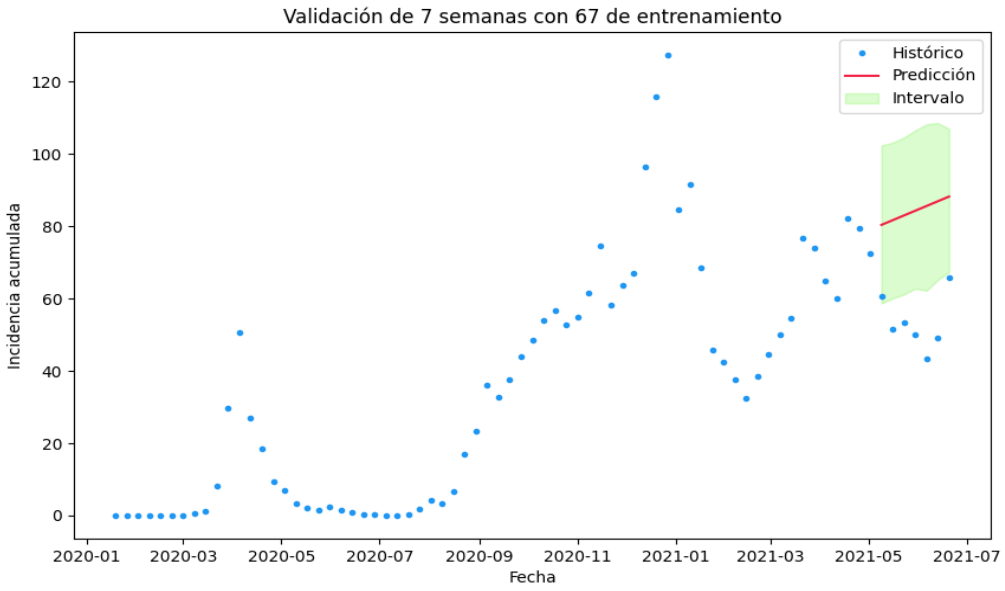


Figura 4.16: Validación del modelo con datos semanales de incidencia acumulada. **Fuente:** Diseño propio

MAE	MSE	RMSE
30.4285	998.9928	30.8114

Tabla 4.1: Métricas de Prophet con la IA semanal.

Para mejorar el modelo, se decidió eliminar los datos entre enero y agosto de 2020, y como consecuencia de que este algoritmo solo permite una variable de entrada a predecir, se cogieron los datos pertenecientes a todos los días de la semana (Figura 4.17), no solo un valor semanal como se había planificado inicialmente. De esta manera, se aumenta la cantidad de datos, que aun así, siguen siendo escasos para realizar predicciones.

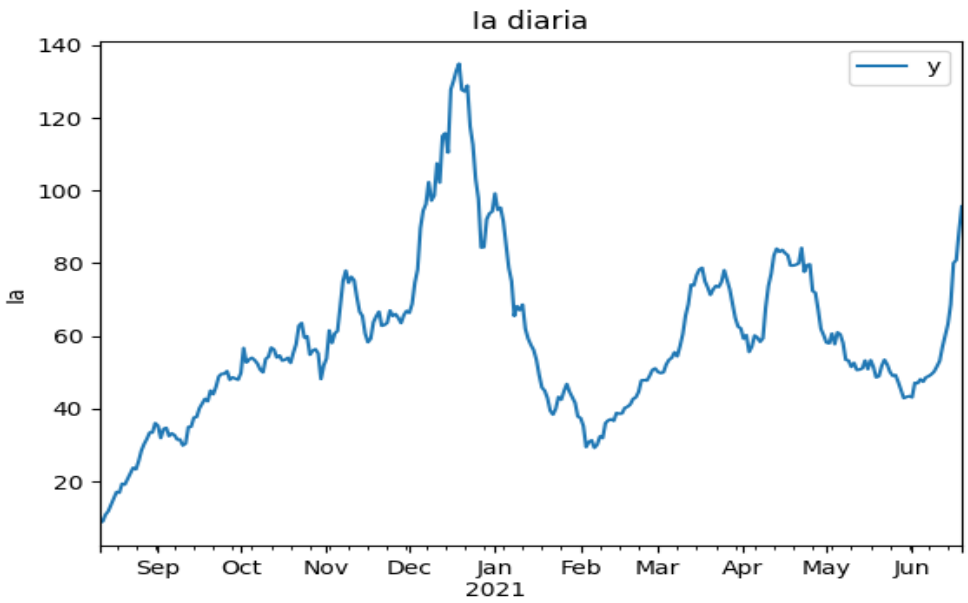


Figura 4.17: Incidencia acumulada diaria desde agosto de 2020. **Fuente:** Diseño propio

Con estos cambios, el modelo fue mejorando incluso con los valores predeterminados de hiperparámetros. Además, se incluyeron los días festivos de la isla de Tenerife. En una primera aproximación se añadieron estrictamente dichos días, pero al comprobar que no tenían demasiada incidencia en el modelo, se optó por añadir períodos no lectivos, ya que son épocas señaladas que por lo general suele haber más movilidad y por tanto, más contagios. Con esta apreciación, los resultados presentan cierta mejoría (Figura 4.18).

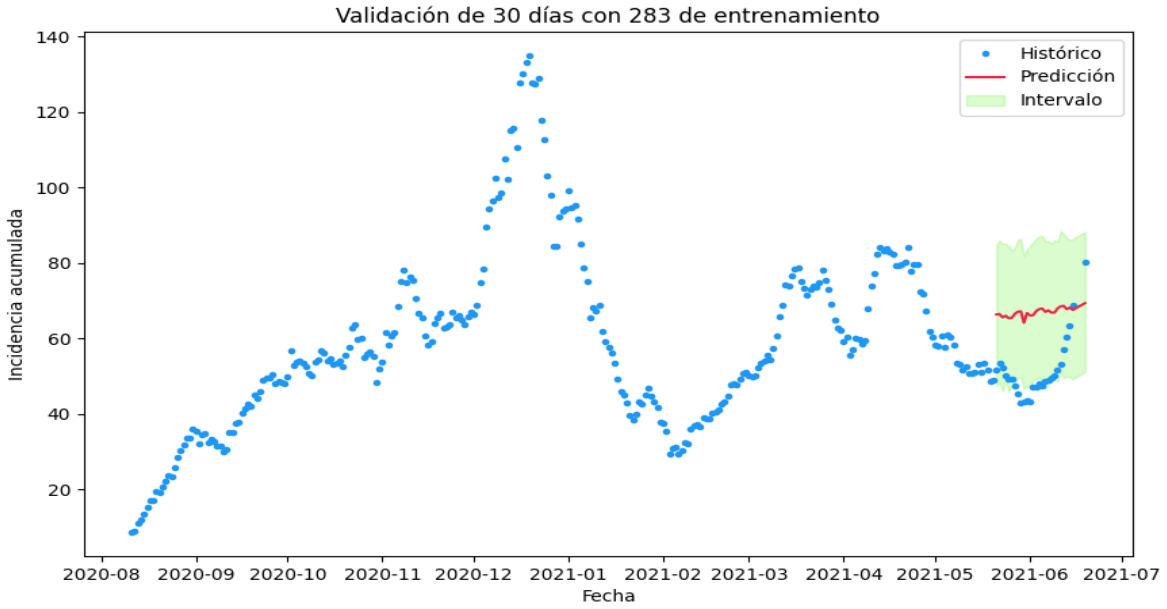


Figura 4.18: Validación del modelo con la incidencia acumulada diaria y los días festivos incluidos. **Fuente:** Diseño propio

MAE	MSE	RMSE
15.94	307.467	16.3

Tabla 4.2: Métricas de Prophet con la IA diaria y los días festivos.

En última instancia, para conocer el mejor valor de los hiperparametros, se ha realizado la técnica 'Grid search' mediante validación cruzada. Esta técnica permite ajustar los siguientes hiperparámetros:

- **'Changepoint priors scale':** Determina la flexibilidad de la tendencia y, en particular, cuánto cambia dicha tendencia en los puntos de cambio. Si el valor es demasiado pequeño, la tendencia no se ajustará. Si es demasiado grande, la tendencia se sobre ajustará. Suele estar en el rango [0.005 , 0.5]
- **'Seasonality prior scale':** Este parámetro controla la flexibilidad de la estacionalidad. De tal forma que, un valor grande permite que la estacionalidad se ajuste a grandes fluctuaciones, y un valor pequeño reduce la magnitud de la estacionalidad.
- **'Holidays prior scale':** Regula la flexibilidad para adaptarse a los efectos de los días festivos.

De esta forma, se prueban todas las opciones posibles con los parámetros introducidos, como se puede apreciar en la Figura 4.19

```
param_grid = {
    'changepoint_prior_scale': [0.001, 0.01, 0.1, 0.5],
    'seasonality_prior_scale': [0.01, 0.1, 1.0, 10.0],
    'holidays_prior_scale': [0.01,0.1,1,10]
}

# Generamos todas las combinaciones posibles entre parametros
all_params = [dict(zip(param_grid.keys(), v)) for v in itertools.product(*param_grid.values())]
rmses = []  # almacenamos el rmses

# Usamos cross validation para cada parametros
for params in all_params:
    m = Prophet(**params).fit(df)  # Fit model with given params
    df_cv = cross_validation(m,initial= '260 days', horizon='20 days', parallel="processes")
    df_p = performance_metrics(df_cv, rolling_window=1)
    rmses.append(df_p['rmse'].values[0])
```

Figura 4.19: Detección de los mejores hiperparámetros para prophet. **Fuente:** Diseño propio

Tras el ajuste mencionado, los mejores hiperparametros encontrados fueron:

- **Changepoint priors scale**: 0.01 para la flexibilidad de la tendencia
- **Seasonality prior scale**: 0.1 reduciendo la magnitud de la estacionalidad
- **Holidays prior scale**: 10, de tal forma que los días festivos tienen máxima incidencia en el modelo

Así, el modelo obtenido finalmente después de realizar diferentes pruebas ha tenido mejoras desde que se inició el proceso. El valor de 'y' que se puede visualizar en la Tabla 4.3 es el dato real de la incidencia acumulada para dicha fecha, y el valor que predice el modelo en la validación viene definido en 'yhat'. El rango superior e inferior de predicción se denomina 'yhat _upper' y 'yhat_lower' respectivamente.

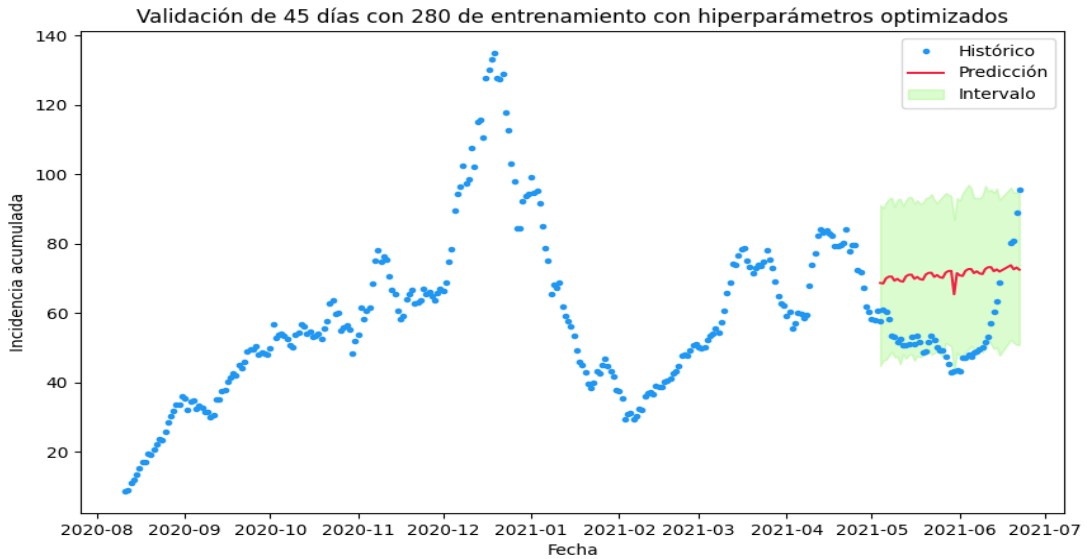


Figura 4.20: Validación del modelo con la incidencia acumulada diaria, los días festivos, y los mejores hiperparámetros. **Fuente:** Diseño propio

Fecha	Yhat	Yhat_lower	Yhat_upper	Y	Cutoff
2021-05-09	69.481333	46.994166	91.477260	53.26	2021-05-08
2021-05-10	69.907679	45.776096	92.084411	51.53	2021-05-08
2021-05-11	68.969421	45.116870	92.628125	52.61	2021-05-08
2021-05-12	68.976747	45-648429	91.215959	50.67	2021-05-08
…….	……..	…….	…….	…….	……...
2021-06-19	73.377669	48.592179	96.247077	80.05	2021-05-08
2021-06-20	72.802115	49.459365	96.248903	80.91	2021-05-08
2021-06-21	73.228461	49.542605	96.289434	88.76	2021-05-08

Tabla 4.3: Diferencia diaria entre la predicción y el valor real.

Las métricas son observables en la Tabla 4.4. El RMSE tiene un valor medio durante el periodo de validación en torno a 17 unidades por encima de los datos reales. Ese valor se mantiene constante excepto en el periodo final de validación, donde existe una ligera variación de las 17 unidades. Esto se puede explicar visualmente en la Figura 4.20, donde se observa que el valor esperado del modelo es robusto a cambios repentinos de tendencia y en la realidad se observa este tipo de cambios. Sin embargo, la parte positiva es que, considerando también el intervalo de predicción, el modelo obtenido podría ser de utilidad para tener una estimación de la incidencia acumulada con unos días de margen.

MAE	MSE	RMSE
14.95	223.5	15.3

Tabla 4.4: Métricas de Prophet con la IA diaria, los días festivos y los mejores hiperparámetros.

Además, se realiza una predicción de 90 días a futuro (Figura 4.21), utilizando el modelo validado anteriormente, con todos los datos destinados al entrenamiento, para obtener un resultado más preciso de predicción. Se puede apreciar que en septiembre de 2021 se prevé un descenso radical en la incidencia acumulada, no obstante, este fenómeno se debe a que los datos de septiembre de 2020 eran valores muy pequeños, y al no poseer un histórico mayor, sigue la tendencia que tenía entonces.

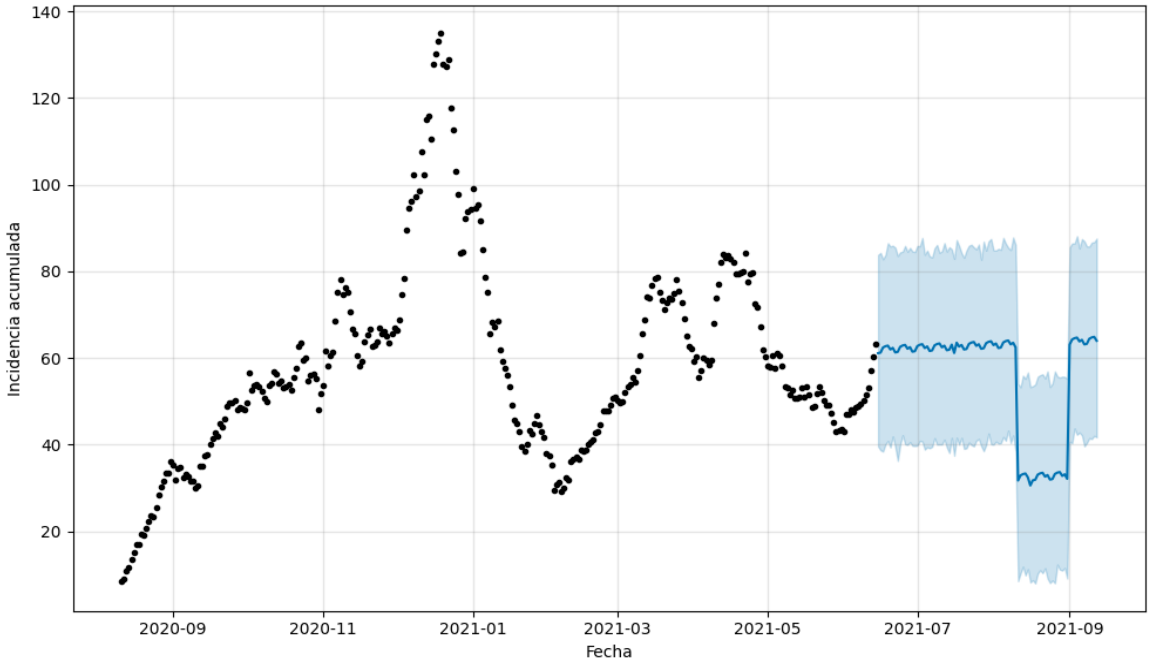


Figura 4.21: Predicción de 90 días con el modelo validado anteriormente. **Fuente:** Diseño propio

Cabe destacar que Prophet asume que la frecuencia y la magnitud media de los cambios de tendencia en el futuro, serán las mismas que las que observamos en el histórico. El problema principal es que en los datos, los cambios de tendencia son muy poco regulares, siendo imposible detectar un patrón a simple vista. Esto sumado a la escasez de datos disponibles hace que el modelo, aún siendo optimizado, no sea del todo fiable.

4.2.4 Modelo Neural Prophet

El siguiente modelo se realiza con Neural Prophet. Como se mencionó anteriormente, esta librería mantiene las características en la sencillez de uso de Prophet y, a su vez, añade las ventajas que poseen las redes neuronales. Cabe destacar que -aunque tiene rasgos que muestran su gran potencial- fue lanzada en octubre de 2020, y está aún en fase beta, por lo que carece de ciertas singularidades que pueden ser de interés en el proyecto.

En primera instancia, se introducen los datos de incidencia acumulada diarios, desde agosto de 2020 hasta la actualidad. Además, se le añaden los días festivos al modelo. El conjunto de datos se divide en aproximadamente un 90% para el entrenamiento, y un 10% del dataset está destinado para la validación. Además, se ha establecido el valor de ‘epochs’ en 150, siendo este el número de veces en las que todos los datos de entrenamiento han pasado por la red neuronal.

El resultado obtenido no deja unas métricas desorbitadas, sin embargo se puede observar en la Figura 4.22 que tiene una tendencia decreciente, cuando la incidencia acumulada está creciendo de forma exponencial durante los días de validación del modelo.

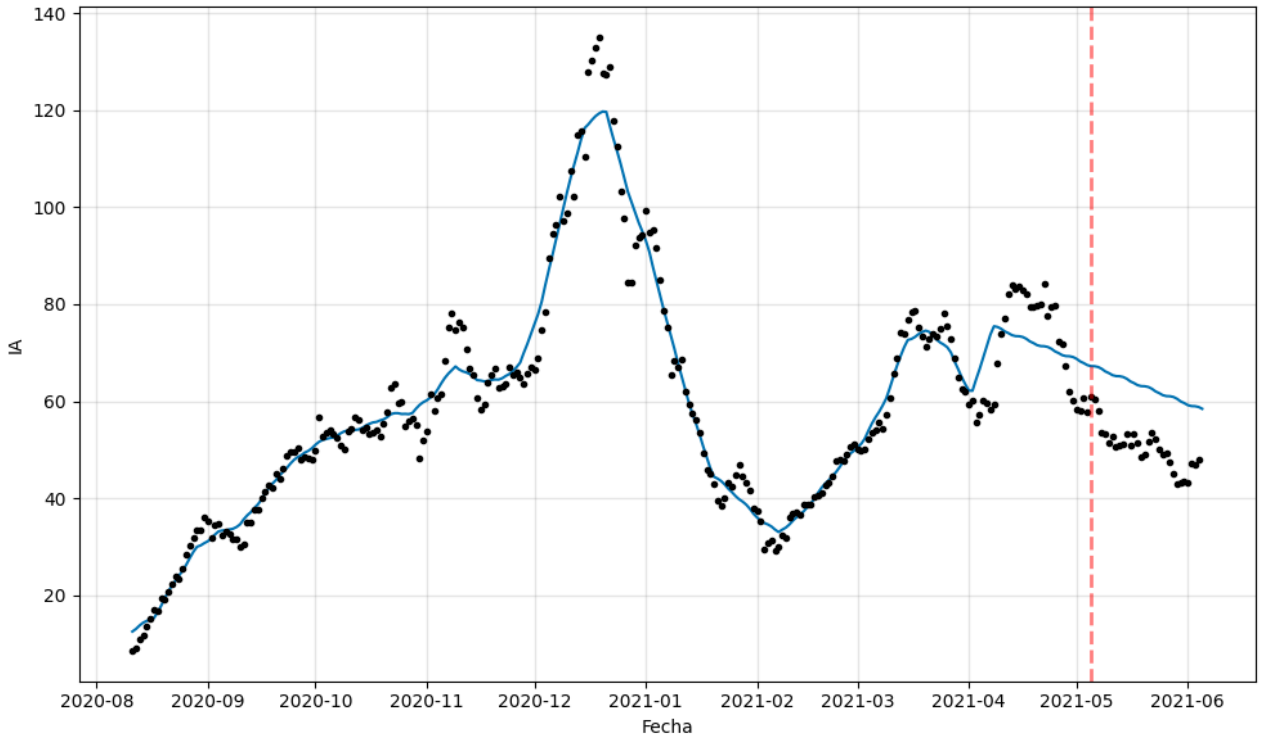


Figura 4.22: Validación del modelo con la incidencia acumulada diaria en Neural Prophet **Fuente:** Diseño propio

MAE	MSE
12.763045	168.794373

Tabla 4.5: Métricas de Neural Prophet con la IA diaria.

Con el objetivo de intentar mejorar el modelo, el siguiente paso a realizar es modificar los hiperparámetros. Una de las capacidades que posee Neural Prophet es AR-Net, que se puede activar añadiendo un valor correcto a 'n_lags' y 'n_forecasts'. Este valor indica hasta qué punto en el pasado deben considerarse las dependencias autorregresivas.

Después de realizar varias pruebas para encontrar el mejor valor en los hiperparámetros, el resultado obtenido de validación del modelo es el que se muestra en la Figura 4.23. Se ha establecido un valor de 15 'n_lags' y 15 'n_forecasts'. Al añadir una cifra a dichos hiperparámetros, se activan las dependencias autorregresivas, y realiza un número de predicciones acorde al valor indicado. Además, se ha ajustado el modelo con 200 'epochs' en base a las métricas obtenidas durante las pruebas realizadas, ya que con 150 epochs el error de validación seguía disminuyendo junto al error de entrenamiento. Con un número de 'epochs' mayor de 200 se generaba un sobreajuste en el modelo.

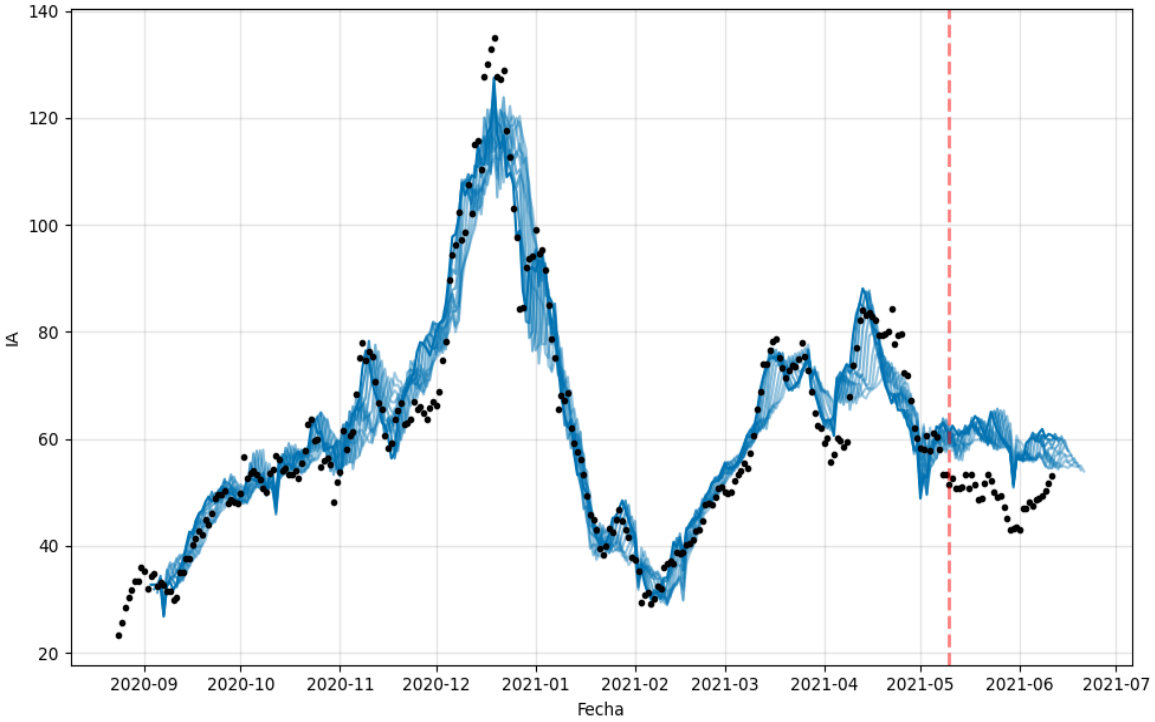


Figura 4.23: Validación del modelo con la incidencia acumulada diaria activando las dependencias autorregresivas. **Fuente:** Diseño propio

MAE	MSE
8.208212	74.516548

Tabla 4.6: Métricas de Neural Prophet con la IA diaria activando las dependencias autorregresivas.

Es preciso añadir que se realizaron pruebas añadiendo capas ocultas en todo el proceso, con el dataset de incidencia acumulada diaria. Sin embargo, no aportaron una mejora, más bien, se generó un ligero aumento de tiempo de cálculo.

Cabe destacar que Neural Prophet permite añadir otras variables como regresores rezagados, a diferencia de Prophet, en los que no se tiene por qué conocer el valor futuro de dicha variable. Al haber recopilado datos semanales de variables que pueden ayudar a realizar mejores predicciones, pasamos a realizar un modelo con datos de incidencia acumulada semanal, añadiendo la llegada de pasajeros extranjeros a Tenerife, con procedencia de Alemania y Gran Bretaña. Tras realizar ajustes en los hiperparámetros, el mejor resultado obtenido se muestra en la Figura 4.24. En este caso se añadieron dos capas ocultas a la red neuronal mediante el hiper parámetro 'num_hidden_layers', ya que mejoraba considerablemente el resultado de la validación.

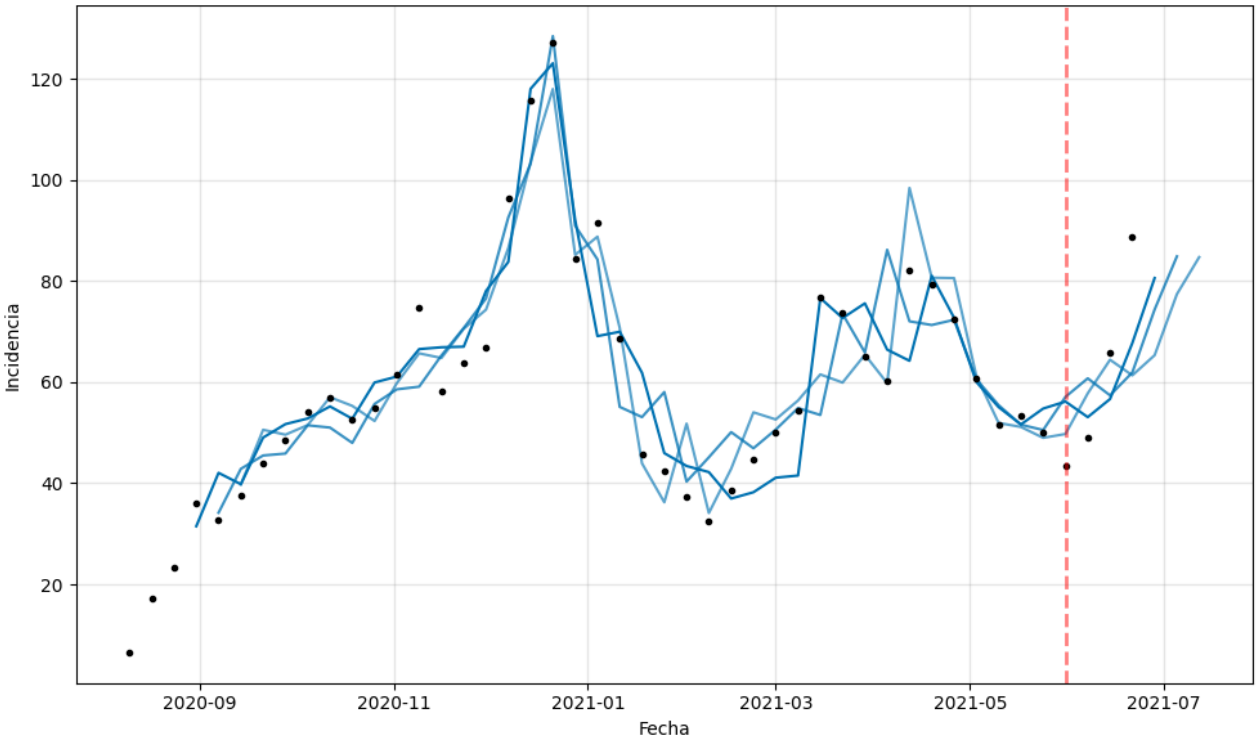


Figura 4.24: Validación del modelo con la incidencia acumulada semanal y los pasajeros como regresores (2 capas ocultas). **Fuente:** Diseño propio

MAE	MSE
13.325213	281.283539

Tabla 4.7: Métricas de Neural Prophet con la IA semanal y los pasajeros como regresores (2 capas ocultas).

A su vez, el análisis extraído de este modelo muestra que un cambio menor en el valor de los hiperparámetros, modifican radicalmente la tendencia. Si bien es cierto que los resultados de la métrica (Tabla 4.7) no mejoran al modelo obtenido sin los regresores rezagados, se puede observar que la tendencia de la curva la predice con cierta efectividad. Además, cabe destacar que al utilizar el dataset de incidencia acumulada semanal, el número de datos empleados para el entrenamiento no supera 50 semanas, con lo cual, siguen siendo muy pocos datos para definir con exactitud la veracidad de un modelo.

4.2.5 Comparación entre modelos

Para comparar el modelo de Prophet y Neural Prophet, se procede a medir el mejor resultado obtenido de cada uno de ellos con las mismas condiciones. En el caso de Prophet, esto se consiguió con los hiperparámetros seleccionados en el último modelo del apartado 4.2.3. Con Neural Prophet -para medir en igualdad de condiciones- no se añaden los regresores rezagados ya que estos no están disponibles en Prophet.

El conjunto de entrenamiento utilizado para realizar la comparación es de 90% de entrenamiento, 10% para la validación.

MODELOS	MAE	MSE	RMSE
Prophet	17.1512	301.9312	17.3762
Neural Prophet	8.5295	116.1253	10.7766

Tabla 4.8: Métricas de la comparación Prophet vs Neural Prophet con el mejor modelo hallado.

Si bien es cierto que con Prophet se realizó 'Grid Search' con validación cruzada para determinar los hiperparámetros, con Neural Prophet no se determinó de la misma manera, ya que el número de hiperparámetros es considerablemente mayor, lo que aumenta exponencialmente el tiempo de ejecución. Las métricas obtenidas de la validación que se observan en la Tabla 4.7, nos muestran que Neural Prophet ha obtenido mejores resultados que Prophet. Estas se han obtenido con la librería Sklearn mencionada anteriormente en el apartado 3.4. La más útil en este caso en concreto es RMSE, ya que los errores grandes son particularmente indeseables, y esta los penaliza.

Por tanto, se puede concluir que, aunque Neural Prophet ha obtenido mejores resultados globales que Prophet, ambos modelos poseen muy pocos datos para el entrenamiento, siendo imposible saber con seguridad si los modelos van a predecir bien el futuro.

Capítulo 5

Conclusiones y líneas futuras

5.1 Conclusiones

Este trabajo ha estado motivado por la importancia que presenta el Open Data en la actualidad. Su auge en los últimos años ha permitido que, cada vez en más ámbitos, los usuarios hagan uso de los portales de datos abiertos. La relevancia de estos datos hace que sea vital llevar un control del portal del que se extrae la información. Por esta razón, se llevó a cabo el análisis de la API meteorológica de GRAFCAN.

Para ello, se realizaron diferentes ETL con el objetivo de automatizar la transformación y recopilación de los datos. Posteriormente, se creó un cuadro de mando que denominamos “semáforo” que indica de forma gráfica el porcentaje de fiabilidad que tiene la API en cuanto a las características de actualización de datos y persistencia. Tras realizar un análisis del semáforo elaborado, se puede concluir que la API meteorológica de GRAFCAN, en lo que respecta a la temperatura del aire y las precipitaciones, proporciona un porcentaje muy elevado de los datos esperados, sobre el 98%. Además, el análisis muestra que la publicación de los datos se realiza en un plazo corto de tiempo, y presenta una persistencia más que aceptable. Estos resultados mejoran incluso las expectativas positivas que se tenía de la API antes del estudio.

En el proyecto se pudo haber continuado con el análisis de otros portales de datos abiertos oficiales. Sin embargo, para ampliar el aprendizaje y experimentar la diversidad de usos que presenta el Open Data, se decidió proseguir con la elaboración de predicciones a partir de datos obtenidos de portales abiertos de máxima fiabilidad.

Por esta razón, se emplearon dos librerías para la elaboración de los modelos predictivos sobre la incidencia acumulada en la isla de Tenerife: Prophet, y Neural Prophet. Los resultados obtenidos muestran que el histórico de datos es demasiado corto para extraer resultados totalmente concluyentes. Por tanto, aunque se ha conseguido elaborar modelos en los que el error -teniendo en cuenta la escasez de datos- no es extremadamente elevado, se deben seguir realizando pruebas cuando se obtengan más datos.

5.2 Líneas Futuras

El proyecto realizado hasta ahora analiza la API meteorológica de GRAFCAN, mostrando el porcentaje de actualización del dato y la persistencia. Además, realiza modelos de predicción, con Prophet y Neural Prophet, a partir de datos obtenidos de portales abiertos.

Los pasos a seguir en el futuro pueden ser muy diversos. A continuación se enumeran algunas de estas posibilidades existentes:

- Usar más datos relacionados con la COVID-19 en los modelos de predicción que admiten regresores, no solo los pasajeros que llegan a la isla, sino también el aforo que permite los locales, el toque de queda, etc.

- Probar otros modelos de Redes Neuronales, para poder realizar una comparación más exhaustiva entre modelos.

- Predecir, además de la incidencia acumulada, los cambios de fase de la COVID-19 con métodos que admitan 2 o más variables a predecir. En su defecto, se puede puede plantear detectar un cambio de fase como una detección de outliers. Es decir, entrenar el modelo para que muestre cómo debería evolucionar la incidencia acumulada y, si se detecta que la incidencia se sale del rango de predicción, entonces decir que se puede venir un cambio de fase.

- Crear un cuadro de mando con Power BI, que permita visualizar mejor los datos de los modelos, comparando los datos reales obtenidos de los portal abiertos, con las predicciones que se hayan realizado.

- Analizar otros portales de datos que sean de interés con una metodología similar

Capítulo 6

Conclusions and Future research lines

6.1 Conclusions

This project has been motivated by the current importance of Open Data. Its boom in recent years has allowed users to utilize open data portals in an increasing variety of areas. The relevance of this data makes it vital to keep track of the portal from which the information is extracted. That’s the reason why this analysis of GRAFCAN's meteorological API was carried out.

For this purpose, the team carried out different ETLs in order to automate the conversion and collection of the data. Subsequently, a scorecard was created that is known as "traffic light", which graphically indicates the percentage of reliability of the API in terms of data update and persistence characteristics. After an analysis of the traffic light, it can be concluded that the GRAFCAN weather API, in regards to air temperature and precipitation, provides a very high percentage of the expected data, around 98%. In addition, the analysis shows that the publication of the data takes place within a short period of time and has a more than acceptable persistence. These results improve even the positive expectations of the IPA prior to the study.

The project could have continued with the analysis of other official open data portals. However, in order to broaden the learning and experience the diversity of uses of Open Data, the team decided to move forward with the development of predictions based on data obtained from the most reliable open portals.

For this reason, two libraries were used for the elaboration of the predictive models on the accumulated incidence on the island of Tenerife: Prophet and Neural Prophet. The results obtained show that the data history is too small to extract entirely conclusive results. Therefore, although it has been possible to develop models in which the error - taking into account the scarcity of data - is not extremely high, further tests should be carried out when more data are available.

6.2 Future research lines

The project so far has analysed the meteorological API from GRAFCAN, showing the update rate of the data, as well as its persistence. In addition, it performs prediction models from data obtained from open portals using Prophet and Neural Prophet.

Future steps can go in a variety of directions. Some of the possibilities are listed in the points below:

- Use more COVID-19 related data in the prediction models that support regressors: not only passengers arriving on the island, but also venue capacity, curfew, etc.

- Test other neural network models, in order to make a more exhaustive comparison between models.

- Predict, in addition to the cumulative incidence, the COVID-19 phase changes with methods that allow the prediction of 2 or more variables. Failing that, one can consider detecting a phase shift as an outlier detection. That is, train the model to show how the cumulative incidence should evolve and, if it is detected that the incidence is out of the prediction range, assume that a phase change may be coming.

- Create a dashboard with Power BI, which allows for an improvement on the visualisation of the models' data, to compare the actual data obtained from the open portals with the predictions that have been made.

- Analyse other portals with a similar methodology.

Capítulo 7

Presupuesto

Este capítulo recoge los costes estimados del proyecto. Para el cálculo del coste total se tuvo en cuenta 2 factores: El hardware empleado en el proyecto, y los recursos humanos.

7.1 Coste del Hardware

Tipos	Descripción	Precio
Portatil Asus	Portatil desde el que se realizó el proyecto.	959 €

Tabla 7.1: Coste del Hardware

7.2 Coste de Recursos Humanos

Tipos	Coste por hora	Precio
480 h	20 €/h	9 600 €

Tabla 7.2: Coste de Recursos Humanos

7.3 Coste Total

Recursos Humanos	Hardware	Total
9 600 €	959 €	10 559 €

Tabla 7.3: Coste Total

Bibliografía

[1] Sensores de GRAFCAN. Sistema de observación meteorológico de Canarias. Dirección: https://sensores.grafcan.es/ (visitado el 30-06-2021)

[2] Laura van Knippenberg, Daphne van Hesteren, Raymonde Weyzen, Eline Lincklaen Arriens, Marit Blank. Open Data Maturity. Dirección: https://data.europa.eu/es/dashboard/2020 (visitado el 26-02-2021)

[3] Portal Europeo de datos.The official portal for European data. Dirección: https://data.europa.eu/es (visitado el 15-02-2021)

[4] Jose Manuel Garrido, David Martinez-Rodriguez, Fernando Rodriguez-Serrano, Jose Miguel Perez-Villares, Andrea Ferreiro-Marzal, Maria del Mar Jimenez-Quintana, Grupo de Estudio COVID-19 Granada, Rafael Jacinto Villanueva. Mathematical model optimized for prediction and health care planning for COVID-19, 2020. arXiv:2012.05804 [math.DS]

[5] Naresh Kumar, Seba Susan. COVID-19 Pandemic Prediction using Time Series Forecasting Models, 2020. arXiv:2009.12176 [physics.soc-ph]

[6] Gobierno de Canarias. Datos COVID-19. Dirección: https://www.gobiernodecanarias.org/principal/coronavirus/acceso_datos.html (visitado el 29-06-2021)

[7] Oskar Triebe, Nikolay Laptev, Ram Rajagopal, *AR-Net: A simple Auto-Regressive Neural Network for time-series*, 2019. arXiv:1911.12436 [cs.LG] (visitado el 15-05-2021)

[8] IAAS. Servicio de Tecnologías de la Información y las Comunicaciones, Universidad de La Laguna. Dirección: https://www.ull.es/servicios/stic/category/iaas/ (visitado el 20-06-2021)

[9] Visual Studio Code. Code Editing. Dirección: https://code.visualstudio.com/

[10] Postman. The Collaboration Platform for API Development. Dirección: https://www.postman.com/ (visitado el 20-03-2021)

[11] Power BI. Visualización de datos. Dirección: https://powerbi.microsoft.com/es-es/ (visitado el 25-04-2021)

[12] Dax. Expresiones de análisis de datos . Dirección: https://docs.microsoft.com/es-es/dax/ (visitado el 17-03-2021)

[13] Numpy. Software de análisis numérico. Dirección: https://numpy.org/ (visitado el 06-05-2021)

[14] Pandas. Python Data Analysis Library. Dirección: https://pandas.pydata.org/

(visitado el 09-06-2021)

[15] Matplotlib. Visualization with Python. Dirección : https://matplotlib.org/ (visitado el 11-05-2021)

[16] Prophet. Forecasting at scale. Facebook Open Source. Dirección: https://facebook.github.io/prophet/ (visitado el 15-06-2021)

[17] Neural Prophet. Facebook Open Source. Dirección: https://neuralprophet.com/ (visitado el 02-06-2021)

[18] Pytorch. Library for deep learning. Dirección: https://pytorch.org/ (visitado el 16-06-2021)

[19] Sklearn. Machine Learning in Python. Dirección: https://scikit-learn.org/stable/ (visitado el 17-06-2021)

[20] Portal de Datos Abiertos SITCAN. Sistema de Información Territorial de Canarias. Dirección .https://opendata.sitcan.es/ (visitado el 19-02-2021)

[21] Portal: Datos abiertos de Canarias. Datos del sector público canario en formatos abiertos, gratuitos y reutilizables. Dirección: https://datos.canarias.es/ (visitado el 11-04-2021)