UNIVERSIDAD DE ANTIOQUIA

1 8 0 3

Predicción de las variables del APDIL y ODT con base en las variables climáticas asociadas al proyecto MINFECLIMA

María Isabel Aristizábal Alzate

Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Instituto de Matemáticas
Medellín, Colombia
2022

Predicción de las variables del APDIL y ODT con base en las variables climáticas asociadas al proyecto MINFECLIMA

María Isabel Aristizábal Alzate

Trabajo de grado presentado como requisito parcial para optar al título de:
Matemático

María Eugenia Castañeda López
Orientador Interno, Instituto de Matemáticas

Esteban Marín Ramírez
Orientador Externo, Sistemas Inteligentes en Red

Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Instituto de Matemáticas
Medellín, Colombia
2022

A mis padres, Luis y Martha.

> (...) el hombre no tiene sino sus dos pies, su corazón, y un camino que no conduce a ninguna parte - Gonzalo Arango

Agradecimientos

Agradezco en primer lugar, a mis padres, Luis Aristizábal y Martha Alzate, por no escatimar esfuerzos, cariño y amor, para conmigo y nuestra familia.

A la Universidad de Antioquia, gratitud por siempre, al darme la oportunidad de formarme en un ambiente integral.

Para la profesora María Eugenia Castañeda del Instituto de Matemáticas, mi más sentido agradecimiento y aprecio por su valioso aporte a la realización de este trabajo y acompañamiento durante este periodo.

Gracias a la empresa SIER y en especial a Esteban Marín y Juan Manuel Restrepo, por darme la oportunidad de realizar mis prácticas académicas en un lugar lleno de aprendizaje y acogida. Un agradecimiento especial nuevamente, a Juan Manuel Restrepo, por haber sido un maestro en todo el sentido de la palabra.

Agradezco a mi abuela Amparo Cano, por sus cuidados y amor durante estos años de universidad. A mis hermanos Ferney, Sebastian, Camilo y Samuel Aristizábal, por apoyar siempre la educación. A mi hermano Camilo Aristizábal, médico de esta misma Universidad, gracias por su compañía, amistad y escucha.

Por último, un gracias y cariño por siempre, a los amigos tan especiales y apreciados que me dejo la carrera: Yessenia Álvarez, Carolina Lopera, Jose Manuel Jaramillo y Paola Velásquez.

¡Gracias Universidad Pública, Gracias Universidad de Antioquia!

Resumen

Una de las principales causas de daño de la vía férrea es el cambio climático es por esto que se hace necesario conocer periódicamente su estado de funcionamiento para observar su evolución e intervenir únicamente cuando sea necesario. Buscando contribuir a un mejor mantenimiento de la infraestructura ferroviaria para adaptarla al cambio climático, la empresa SIER junto con otras entidades, está desarrollando el proyecto MINFECLIMA en España. En este trabajo se presenta el desarrollo para lograr predicciones del comportamiento de la estructura férrea, considerando los datos de las estaciones del clima a través de un desarrollo de series de tiempo con los modelos de Random Forest y Ridge. De esta manera se pueden obtener predicciones diarias del comportamiento de la estructura férrea a través de las variables del clima proporcionadas por AEMET y AVAMET.

Palabras clave: AEMET, AVAMET, Random Forest, Ridge, Series de tiempo, SIER, Vía férrea.

Abstract

One of the main causes of damage to the railway is climate change, which is why it is necessary to know periodically its operating status to observe its evolution and intervene only when necessary. Looking to contribute to a better maintenance of the railway infrastructure to adapt it to climate change, the SIER company, and other entities, are developing the MINFECLIMA project in Spain. This work presents the development to achieve predictions of the behavior of the iron structure, considering the data of the weather stations through a development of time series with the Random Forest and Ridge models. This way, daily predictions of the behavior of the iron structure can be obtained through the climate variables provided by AEMET and AVAMET.

Keywords: AEMET, AVAMET, Railway, Random Forest, Ridge, SIER, Time series.

Predicción de las variables del APDIL y ODT con base en las variables climáticas asociadas al proyecto MINFECLIMA

María Isabel Aristizábal Alzate [*]

2 de abril de 2022

[*]E-mail: maria.aristizabala@udea.edu.co, Instituto de Matemáticas, Universidad de Antioquia, Medellín, Colombia.

Resumen

Una de las principales causas de daño de la vía férrea es el cambio climático es por esto que se hace necesario conocer periódicamente su estado de funcionamiento para observar su evolución e intervenir únicamente cuando sea necesario. Buscando contribuir a un mejor mantenimiento de la infraestructura ferroviaria para adaptarla al cambio climático, la empresa SIER junto con otras entidades, está desarrollando el proyecto MINFECLIMA en España. En este trabajo se presenta el desarrollo para lograr predicciones del comportamiento de la estructura férrea considerando los datos de las estaciones del clima, a través de un desarrollo de series de tiempo con los modelos de Random Forest y Ridge. De esta manera se pueden obtener predicciones diarias del comportamiento de la estructura férrea a través de las variables del clima proporcionadas por AEMET y AVAMET.

Palabras clave: AEMET, AVAMET, Random Forest, Ridge, Series de tiempo, SIER, Vía férrea

Contenido

1. Introducción

En la actualidad la tendencia en el mantenimiento de instalaciones, bienes y equipos se basa en el llamado mantenimiento predictivo, es por esto, que se hace necesario conocer periódicamente su estado de funcionamiento para observar su evolución e intervenir únicamente cuando sea necesario. En cuanto a las vías férreas, una de las principales causas de daño y sus componentes tiene que ver con el comportamiento dinámico de la misma. Por esto, se hace necesario conocer cómo se deteriora la vía en función del tráfico que atiende y sus características. Además, el cambio climático y el aumento de los fenómenos climáticos extremos, supone un gran desafío para los medios de transporte, en este caso en particular, el ferrocarril. Son diversos los impactos que el clima puede tener sobre la infraestructura ferroviaria, como fallos por pandeo debido a las olas de calor, o inestabilidad de taludes debido a las lluvias torrenciales. Siendo esto así, resulta necesario adaptar la infraestructura ferroviaria a las condiciones cambiantes del clima, por lo que el proyecto MINFECLIMA (Mantenimiento inteligente de infraestructuras ferroviarias con base en el tratamiento integral de datos ante nuevos escenarios climáticos) tiene como objetivo contribuir a un mejor mantenimiento de la infraestructura ferroviaria para adaptarla al cambio climático.

El proyecto MINFECLIMA, fruto de la colaboración entre AZVI, SIER y la Universidad Politécnica de Valencia, busca contribuir a un mejor mantenimiento de la infraestructura ferroviaria de España, para adaptarla al cambio climático. Una de las principales tareas del proyecto se centra en monitorizar durante un largo periodo de tiempo ciertos elementos de la vía especialmente críticos y vulnerables, tales como desvíos o aparatos de dilatación, para poder recopilar un gran volumen de datos que permita determinar con mayor precisión los impactos provocados por los fenómenos climáticos extremos. Este plan de monitorización incluye la instalación de sensores en un Aparato de Dilatación (APDIL) y una Obra de Drenaje Transversal (ODT) que se encuentran en la vía para determinar su comportamiento frente a los cambios de temperatura.

El objetivo principal del proyecto es proponer un procedimiento de análisis masivo de datos y detección automática de patrones y tendencias de fallo y riesgo. Esto, desde la creación de un sistema de alarma para detectar anomalías climáticas, anomalías en los datos del APDIL y la ODT, y la detección de patrones de fallo en combinación con las variables climáticas.
El objetivo principal desarrollado durante la práctica académica fue realizar un análisis del Aparato de Dilatación (APDIL), de la Obra de Drenaje Transversal (ODT) y de las variables climáticas, con el fin de proponer una predicción de las variables en cuestión.

Para llevar a cabo el presente trabajo se tuvieron en cuenta estrategias de solución tales como: entendimiento del proyecto MINFECLIMA, visualización, exploración y agrupación de los datos, análisis de correlación y exploratorio de modelos, y por último la elaboración del modelo final. Entre las limitaciones ocurridas durante la realización del proyecto, se encuentra el daño inesperado de los sensores de los aparatos en medición, por lo que el análisis de datos logro capturar datos de solo dos meses.

El presente informe cuenta con el siguiente esquema. En la sección 2, se encuentra el marco teórico donde se presentan los fundamentos técnicos y teóricos del sistema de trenes y mantenimiento ferroviario, la revisión de las nubes de almacenamiento y las metodologías estadísticas implementadas. En la sección 3, se presenta la metodología empleada en el estudio, en el que da cuenta de los detalles de la preparación de los datos y la aplicación de los modelos de predicción. En la sección 4, se presentan los resultados de los modelos de predicción implementados para los datos a futuro. Finalmente, en la sección 5, se presentan las conclusiones y recomendaciones del estudio.

2. Marco Teórico

2.1. Revisión del sistema de trenes y mantenimiento ferroviario

La entidad pública empresarial española encargada de la construcción de líneas de ferrocarril y gestión de su explotación es **Adif** (Administrador de Infraestructuras Ferroviarias). En relación al clima, Adif tiene implementado un plan de contingencias, enfocado fundamentalmente a prevenir los efectos de los temporales de lluvia, establecer pautas de actuación para todos los actores que intervienen en la Red gestionada por Adif (inclusive las empresas de mantenimiento), y proteger la seguridad de viajeros y mercancías, así como la integridad de las instalaciones.

La información de base para la prevención y valoración de los riesgos viene determinada por las diferentes predicciones emitidas por la Agencia Estatal de Meteorología (AEMET), datos publicados en su web y prestaciones específicas según el convenio vigente entre Adif y AEMET. Los recursos y actuaciones se disponen teniendo en cuenta tanto las medidas preventivas que difunde el Centro de Gestión de la Red 24 Horas como la experiencia en las alertas y emergencias anteriores.

El Plan tiene establecidos 3 niveles de alerta:
- Nivel 0 (sin alerta): precipitación menor de 20 mm en el intervalo de 6 horas considerado.
- Nivel 1: precipitación entre 20 y 40 mm en el intervalo de 6 horas considerado.
- Nivel 2: precipitación entre 40 y 80 mm en el intervalo de 6 horas considerado.
- Nivel 3: precipitación mayor de 80 mm en el intervalo de 6 horas considerado.

La identificación de riesgos derivados para el ferrocarril son los siguientes:
- Interrupción de la circulación por superar el nivel de agua de la cota del carril.
- Ripados (desplazamientos laterales) de vía al ser arrastrada por las aguas.
- Desprendimientos sobre la vía o sus proximidades, interceptando el gálibo.
- Posible arrastre de líneas aéreas (catenaria) por debilitación de la cimentación.
- Interrupciones en el suministro eléctrico por derivaciones o cortocircuitos.
- Inundaciones de pasos inferiores en estaciones y zonas de vía soterradas.

En cada línea ferroviaria, se establece el listado con los puntos de riesgo. Dichos puntos aparecen marcados en la vía con un cartelón específico que muestra la letra T. A cada punto se le asigna un nivel de deficiencia (alto, medio-alto y medio).

2.2. Revisión de las nubes de almacenamiento y arquitectura de cómputo

Para realizar la ingesta u operación de Extracción, Transformación y Carga se adquirieron los servicios de Microsoft Azure. La Factoría de datos (V2) (Azure Data Factory), es una solución de integración de datos sin servidor, administrada para la ingesta, preparación y transformación de datos a gran escala, con funciones de copiar, hacer flujo de datos, buscar, obtener metadatos y eliminar actividades. Azure data factory dispone de mecanismos de conexión con 90 tipos distintos de fuentes de datos.
Se creo el pipeline SQL-parseddatav3, donde se procede a hacer la copia histórica de los datos a PostgreSQL.
Por otro lado, se trabajó en la máquina virtual vm-minfeclima la cual tiene como objetivo la instalación de múltiples contenedores utilizando la tecnología Docker, para prestar algunos servicios al presente proyecto.
La figura 1 permite observar gráficamente la arquitectura de cómputo implementada en el proyecto:

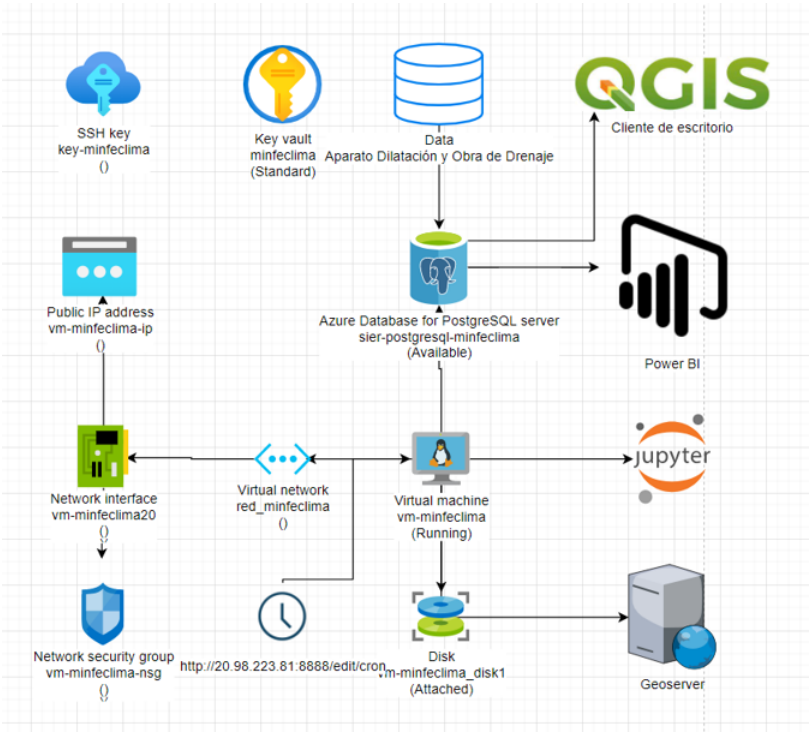

Figura 1: Arquitectura MINFECLIMA

2.3. Metodologías estadísticas

Para resolver el problema en cuestión, se utilizaron metodologías de Machine Learnig. En cuanto a las metodologías estadísticas implementadas se realizó un análisis de correlación, regresión lineal con características polinómicas y métodos de ensamble, y series de tiempo con los métodos Ridge y Random Forest. Finalmente, estas metodologías fueron implementadas en Python junto con las librerías principales de Machine Learning.

2.3.1. Correlación

Dos variables están asociadas cuando una variable proporciona información acerca de la otra. Por el contrario, cuando no existe asociación, el aumento o disminución de una variable no dice nada sobre el comportamiento de la otra variable. Dos variables se correlacionan cuando muestran una tendencia creciente o decreciente. La correlación permite medir las fuerzas de asociación entre dos variables. El valor del coeficiente de correlación varía entre $+1$ y -1. Los coeficientes de correlación usados fueron: Pearson, Kendall y Spearman.

- **Coeficiente de Pearson γ**
  El coeficiente de correlación de Pearson es una medida de dependencia lineal entre dos variables aleatorias cuantitativas. Se usa para medir el grado de relación de dos variables siempre y cuando ambas sean cuantitativas y continuas.
  Dados n pares de datos $\{(x_i, y_i)\}_{i=1}^{n}$, se define el coeficiente de correlación muestral de Pearson como:

$$\gamma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

donde

- n es el tamaño de la muestra,
- x_i, y_i son puntos muestrales individuales indexados con i,
- $\bar{x}$ denota la media muestral definida por $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (análogamente para $\bar{y}$)

- **Coeficiente de Kendall τ**
  El coeficiente de correlación de rango de Kendall, es una estadística utilizada para medir la asociación ordinal entre dos cantidades medidas. Intuitivamente, la correlación de Kendall entre dos variables será alta cuando las observaciones tengan un rango similar.
  Sea $(x_1, y_1), ..., (x_n, y_n)$ un conjunto de observaciones de las variables aleatorias conjuntas X e Y, de modo que todos los valores de (x_i) y (y_i) son únicos. Cualquier par de observaciones (x_i, y_i), (x_j, y_j) dónde $i < j$, se dice que son un par concordante si el orden de clasificación de (x_i, y_i), (x_j, y_j) está de acuerdo: es decir, si ambos $x_i > x_j$ e $y_i > y_j$ o ambos $x_i < x_j$ e $y_i < y_j$; de lo contrario se dice que son discordantes.

  El coeficiente τ de Kendall se define como:

$$\tau = \frac{(\text{número de pares concordantes}) - (\text{número de pares discordantes})}{C_{(n,2)}}$$

  donde $C_{(n,2)} = \frac{n(n-1)}{2}$ es el coeficiente binomial para la cantidad de formas de elegir dos elementos de n elementos.

- **Coeficiente de Spearman ρ**
  El coeficiente de correlación de Spearman, es una medida de la correlación entre dos variables aleatorias, tanto continuas como discretas. Para calcular ρ, los datos son ordenados y reemplazados por su respectivo orden.

  El estadístico ρ viene dado por la expresión:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

  donde D es la diferencia entre los correspondientes estadísticos de orden de $x - y$ y N es el número de parejas de datos.

2.3.2. Detección de anomalías

Existen actualmente una gran cantidad de softwares para detección de anomalías en series de tiempo. En este trabajo se utilizó el modelo **ADTK** (Anomaly Detection Toolkit), el cual es un paquete de Python que permite la detección de anomalías en series de tiempo no supervisadas o basadas en reglas. Este paquete ofrece un conjunto de detectores, transformadores y agregadores comunes con APIs unificadas, así como clases pipeline (tuberías) que conectan los modelos. También proporciona algunas funciones para procesar y visualizar series temporales y eventos de anomalías. De este modelo se aplican dos métodos: ThresholdAD e InterQuartileRangeAD. El método ThresholdAD, compara cada valor de serie

de tiempo con umbrales dados. Los umbrales usados son la media más una desviación estándar y la media menos una desviación estándar. Para el caso de la ODT, se considera como umbral una distancia entre 30 cm y 100 cm. El InterQuartileRangeAD se basa en el rango intercuartílico (IQR). Cuando un valor está fuera del rango definido por

$$[Q1 - c \times IQR,\ \ Q3 + c \times IQR] \tag{1}$$

donde $IQR = Q3 - Q1$ es la diferencia entre los cuantiles del 25 % y el 75 %, se considera un punto de anomalía.

2.3.3. Regresión lineal

La librería SKLEARN, se usa para aplicar Machine Learning. Para hacer predicción de las variables correlacionadas usando regresión lineal, se usan dos metodos de ensamble que permiten combinar varios modelos de Machine Learnig para producir una única predicción en la que se presentan mejores desempeños que en los modelos individuales. Estos metodos son Random Forest y Gradient Boosting.

2.3.4. Modelo de regresión lineal con Gradient Boosting

Los modelos simples son eficientes y de rápido entrenamiento, pero tienen poco poder explicativo sobre datos observados (subajuste). Por otro lado, los modelos muy complejos tienen mucho poder explicativo sobre un conjunto de entrenamiento, pero poca habilidad predictiva para datos no observados (sobreajuste).

Los ensambles combinan varios modelos de Machine Learning para producir una única predicción y pueden presentar mejor desempeño que los modelos individuales.

El Boosting engloba a una familia de algoritmos cuya idea general es tomar modelos sencillos, por lo general árboles de decisión, y mejorar sus predicciones de manera secuencial. Para mejorar esas predicciones el algoritmo entrena cada modelo secuencialmente con todos los datos y, para cada nuevo modelo, se le da más peso a los datos que no fueron bien clasificados o cuyo error en regresión sea más alto. Finalmente, la predicción será un promedio ponderado de todos los clasificadores base. El bossting es secuencial y dependiente, es decir, el modelo en la iteración actual depende de las predicciones en la iteración anterior.

El Gradient boosting es un algoritmo que generaliza la idea del boosting para tratarlo como un problema de optimización que se puede solucionar para diferentes funciones de pérdida y con un método similar al descenso por el gradiente. En lugar de darle más peso directamente a los datos con el mayor error, el gradient boosting entrena el siguiente modelo para que minimice el residual (la diferencia entre las etiquetas y las predicciones del ensamble actual) o para que se ajuste al gradiente de la pérdida.

2.3.5. Series de tiempo con Random Forest

Una serie temporal es una sucesión de datos ordenados cronológicamente, espaciados a intervalos iguales o desiguales. El proceso de Forecasting consiste en predecir el valor futuro de una serie temporal, bien modelando la serie temporal únicamente en función de su comportamiento pasado (autorregresivo) o empleando otras variables externas a la serie temporal.

Supongamos que tenemos una sucesión aleatoria $(X_t, Y_t)_{t \in \mathbb{Z}} \in X \times Y$ tal que

$$Y_t = f(X_t) + \varepsilon_t$$

Y el error ε_t es tal que $\mathbb{E}[\varepsilon_t | X_t] = 0$. El propósito de random forest es estimar, con solo observar una muestra de entrenamiento $D_n = ((X_1, Y_1), ..., (X_n, Y_n))$ la función de regresión

$$\forall x \in \chi,\ f(x) = \mathbb{E}[Y_t | X_t = x]$$

Un modelo Random Forest está formado por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria extraída de los datos de entrenamiento originales mediante bootstrapping. Esto implica que cada árbol se entrena con unos datos ligeramente distintos. En cada árbol individual, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

La predicción de un modelo Random Forest es la media de las predicciones de todos los árboles que lo forman.

2.3.6. Criterios de evaluación del módelo

En este estudio, el 80 % de los datos se seleccionaron como datos de entrenamiento y el 20 % restante se reservó como datos de prueba. Para evaluar la calidad de los modelos, se calcularon el MAE (error absoluto medio) y el R^2 (coeficiente de determinación).

El MAE es el valor absoluto del error medio (diferencia media entre el original y los valores predichos). Un MAE más bajo corresponde a un mayor rendimiento del modelo. El R^2 mide qué tan bien los valores pronosticados coinciden con los valores originales. Su valor oscila de 0 (sin correlación entre los valores reales y predichos) a 1 (correlación perfecta entre los dos valores). En conclusión, un buen modelo de predicción se caracteriza por un bajo MAE y alto R^2. Las medidas se expresan respectivamente como sigue:

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

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

donde n es el número total de datos de prueba, y_i y $\hat{y}_i$ son los datos medidos y los valores predichos por el modelo, respectivamente, y $\bar{y}_i$ es el valor de salida promedio de los datos de prueba.

2.4. QGIS

QGIS es un Sistema de Información Geográfica de software libre y de código abierto. Permite manejar formatos vectoriales, así como la extensión espacial de PostgreSQL. Dentro de sus complementos se encuentra el **NnJoin**, que permite combinar registros de múltiples tablas, dando como resultado una nueva tabla con una nueva columna que contiene la menor distancia entre puntos y vectores.

3. Metodología

Para llevar a cabo el presente trabajo se tuvieron en cuenta cinco etapas. La primera es el entendimiento del proyecto Minfeclima. La segunda, corresponde a la visualización y exploración de los datos de las variables a estudiar. En la tercera etapa se hace una agrupación de los datos, para más adelante, realizar un análisis de correlación entre las variables de los aparatos de la ODT y APDIL y las variables climáticas. Posteriormente, se lleva a cabo un análisis exploratorio de modelos de detección de anomalías y de predicciones, y finalmente, en la última etapa, realizar un modelo de predicción para cada una de las variables analizadas.

3.1. Entendimiento del proyecto, visualización y exploración de los datos

En el marco del proyecto MINFECLIMA, se cuenta con dos puntos de medición en la vía: el Aparato de dilatación-APDIL y la Obra de Drenaje Transversal-ODT.
La figura 2 corresponde al APDIL, en esta se observa la junta de dilatación y el carril. En este punto se instalaron ocho sensores de diferente tipo (ver figura 3).

Figura 2: APDIL

Variable	Sensores
Temperatura en carril (ºC). Tipo: **temp**	4 sensores de temperatura en alma de carril (T)
Separación entre estribo y tablero del puente (cm). Tipo: **desp**	2 potenciómetros de desplazamiento en junta de dilatación (P)
Desplazamiento en puntas del aparato de dilatación (cm). Tipo: **dist**	2 sensores de distancia por ultrasonidos en las puntas del aparato (U)

Figura 3: Variables en el APDIL

La figura 4 permite identificar los puntos donde están instalados estos sensores del APDIL:

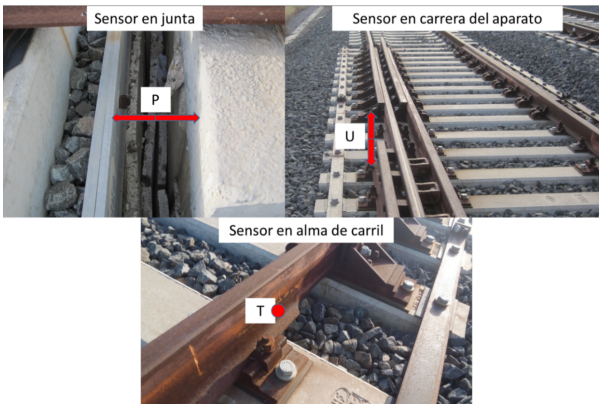

Figura 4: Puntos de medición en el APDIL

De igual manera, la figura 5 corresponde a la ODT, donde se instalaron dos sensores de distancia (ver figura 6).

Figura 5: ODT

Variable	**Sensores**
Distancia desde la clave de la ***ODT*** hasta la lámina de agua, en centímetros (cm). Tipo: **dist**	2 sensores de distancia por ultrasonidos

Figura 6: Variables en la ODT

Por otra parte, se cuenta con las datos de las climatológicas diarias reportadas por **AEMET** (Agencia Estatal de Metereología de España) y **AVAMET**(Asociación Valenciana de Metereología Josep Peinado), ver figura 7. En estas, se identificaron 188 estaciones del clima que reportaban datos en las fechas del análisis.

Figura 7: AEMET y AVAMET

La figura 8 corresponde al mapa por tramos del ferrocarril de España, donde las líneas rojas son cada uno de los tramos del ferrocarril y los puntos azules son las estaciones climatológicas proporcionadas por AVAMET y AEMET. Además, los dos puntos verdes corresponden a los puntos de medición en la vía, el APDIL y la ODT.

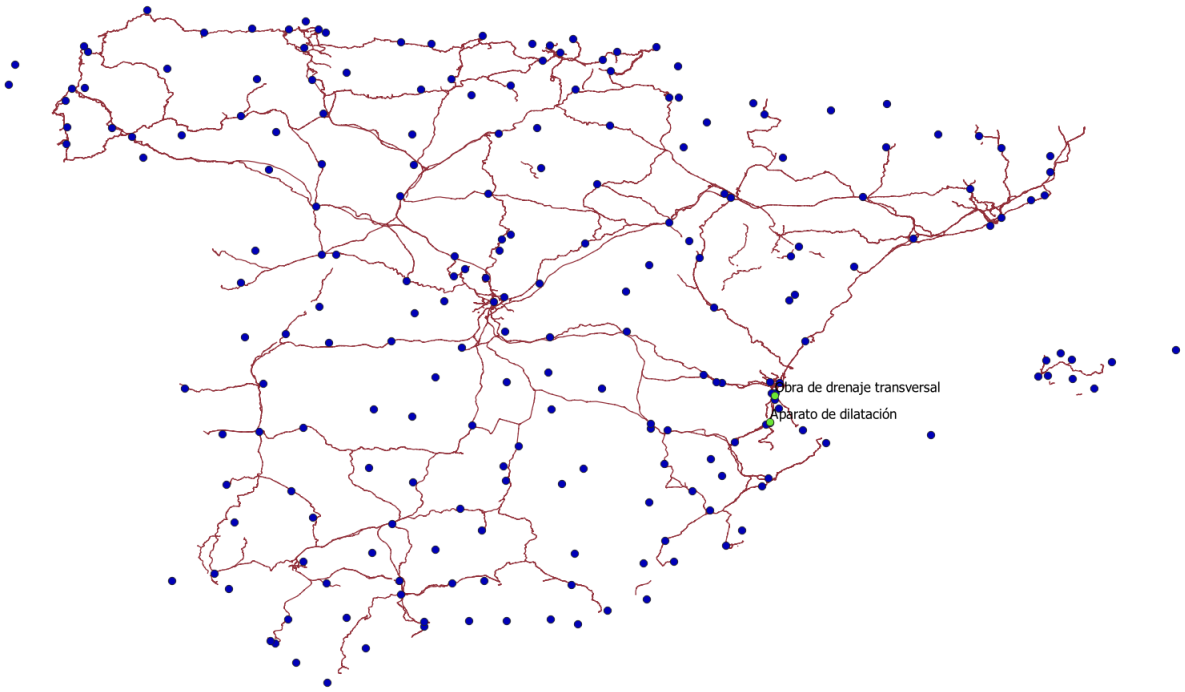

Figura 8: Mapa por tramos del ferrocarril y estaciones climatológicas de España

El objetivo durante esta etapa fue empezar con el análisis y filtrado de los datos de los 10 sensores registrados. También poder visualizar por medio de gráficas el comportamiento y tendencia de estos datos, a la vez identificar posibles fallas en los sensores, al reportar datos atípicos y salidos del promedio. De igual manera, se identificó cuáles eran las estaciones del clima que reportaban datos para las fechas del análisis.

Un primer desarrollo se hizo en Python usando la librería pandas y matplotlib.

La segunda parte se realizó en PowerBi, haciendo un dashboard que permitió la visualización gráfica inmediata de los sensores asociados a la ODT o APDIL por día, intervalo de días, semana, según la preferencia o necesidad. Las fechas analizadas en el dashboard van desde el 7 de julio de 2021 hasta el 27 de julio de 2021.

Se analizaron los datos del APDIL y de la ODT, desde el 15 de julio, hasta el 8 de septiembre, fecha en la que los sensores empiezan a presentar fallas.

Durante esta etapa se da un primer acercamiento al proyecto MINFECLIMA, después de un estudio

teórico del proyecto y del manejo de las bases de datos que se utilizan.

3.2. Agrupación de los datos

Partiendo de que los datos de los 10 sensores tienen una frecuencia de diez minutos y las estaciones climáticas de AEMET y AVAMET tiene una frecuencia de una hora; se hace necesario integrarlos a una misma frecuencia para realizar un análisis de correlación entre las variables del APDIL, de la ODT y de las variables climáticas.

Además, se requiere contar con los datos de las estaciones climáticas más cercanas al APDIL y a la ODT, por lo que se consideran las estaciones que estén a distancias menores o iguales a 40 kilómetros de distancia de uno o de otro aparato. Se nota que son muchas estaciones las que cumplen las condiciones, pero no todas reportan datos en las fechas analizadas, solo cinco de estas: *OLIVA, XÁTIVA , POLINYA , VALENCIA/AEROPUERTO, VALENCIA DT.*

Se considerará una solo tabla que recoja los datos de los 10 sensores con una frecuencia de cada hora, tomando el dato mínimo, máximo y medio para cada intervalo. Así, estos datos se pueden integrar con los datos de las estaciones del clima mencionadas.

3.3. Análisis de correlación

Después de la agrupación de datos realizada, se procede con el análisis de correlación de las variables ejecutando los coeficientes de correlación de Pearson, Kendall y Spearman. Se consideran los valores de correlación tal que su valor absoluto sea mayor o igual que 0.5 ó 0.7 para identificar las variables altamente relacionadas.

De esta manera se puede inferir que las variables altamente relacionadas son:

- El sensor de temperatura del APDIL $\leftrightarrow$ La temperatura de las estaciones climáticas.
- El sensor de temperatura del APDIL $\leftrightarrow$ El sensor de desplazamiento del APDIL.
- Los sensores de la ODT $\leftrightarrow$ La temperatura de las estaciones climáticas.
- La precipitación de las estaciones climáticas $\leftrightarrow$ Los sensores del APDIL y la ODT.

La figura 9, corresponde a la gráfica de la matriz de correlación con el coeficiente de Pearson para la estación climatológica XÁTIVA. En ella se puede observar que entre más oscuro sea el color, el coeficiente de correlación es más alto, así, los azules oscuros corresponden a coeficientes con correlación positiva alta, y los colores rojos, corresponden a coeficientes con correlación negativa.

3.4. Análisis de anomalías y valores atípicos del APDIL y la ODT

Durante esta etapa se realizó una identificación de anomalías y valores atípicos de los datos de los 10 sensores en el estudio.

Para la identificación de anomalías, sin pérdida de generalidad, se consideran solo los datos de la estación XÁTIVA, ya que al hacer una box-plot de las cinco estaciones consideradas desde un comienzo, se observó que no existen grandes diferencias en sus distribuciones como se muestra en la figura 10.

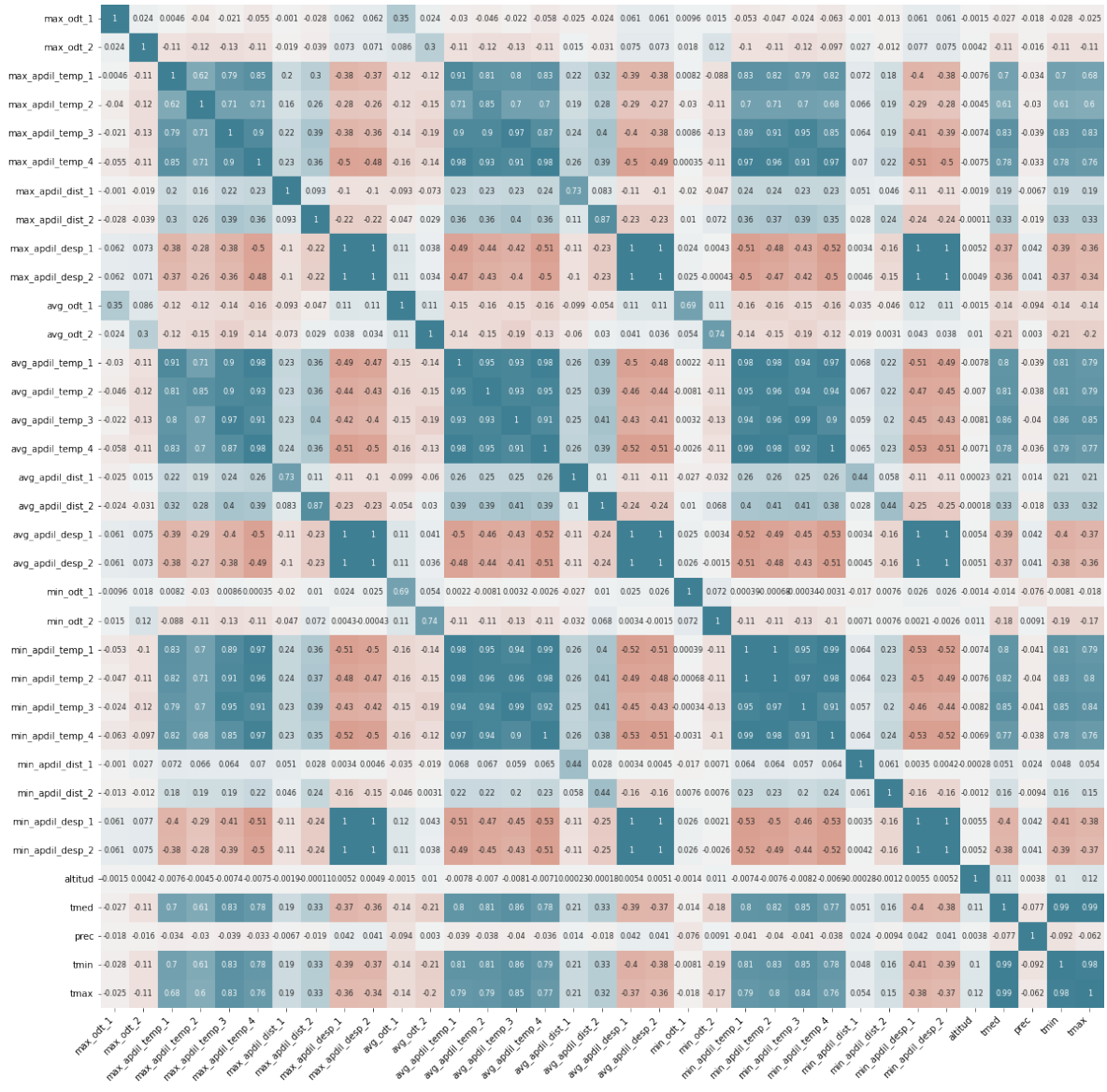

Figura 9: Matriz de correlación

Al graficar la temperatura de las estaciones climáticas junto con las temperaturas de los cuatro sensores, se observa la correlación positiva identificada en la matriz de correlación, (ver figura 11) y desde el 8 de septiembre se nota un daño en los cuatro sensores al estar reportando datos bastante anómalos (ver figura 12), es por esto, que para posteriores análisis se hace necesario omitir estos valores.

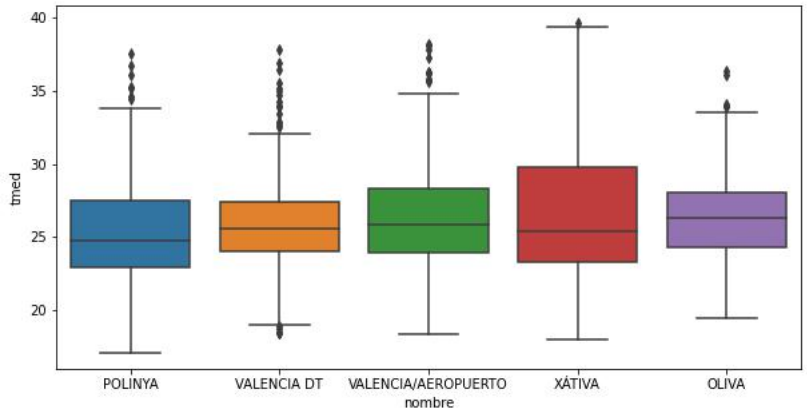

Figura 10: Box-plot de las estaciones cercanas a los puntos de medición

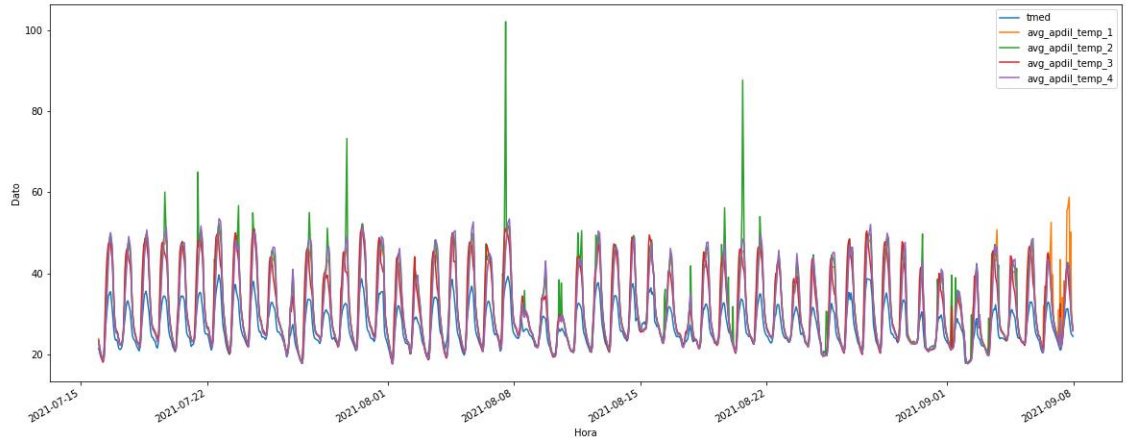

Figura 11: Identificación de correlación positiva entre tmed de Xátiva y los sensores de temperatura

Algunas observaciones del análisis de anomalías:

- En los sensores de temperatura del APDIL desde el 8 de septiembre de 2021 se identifican anomalías, con datos muy extremos, por lo que para posteriores análisis de predicción se hará necesario eliminar estos datos que presentan fallas para evitar errores de predicción.
- Para los datos que en su descripción estadística cuentan con una desviación estándar se hace necesario considerar el método InterQuartileRangeAD para su graficación de anomalías. Para los demás sensores, es indiferente el método utilizado.
- Para la ODT se conoce de antemano que los valores entre 30 y 100 cm son datos que se consideran alarma, por lo que es más adecuado usar el método ThresholdAD que permite ingresar los

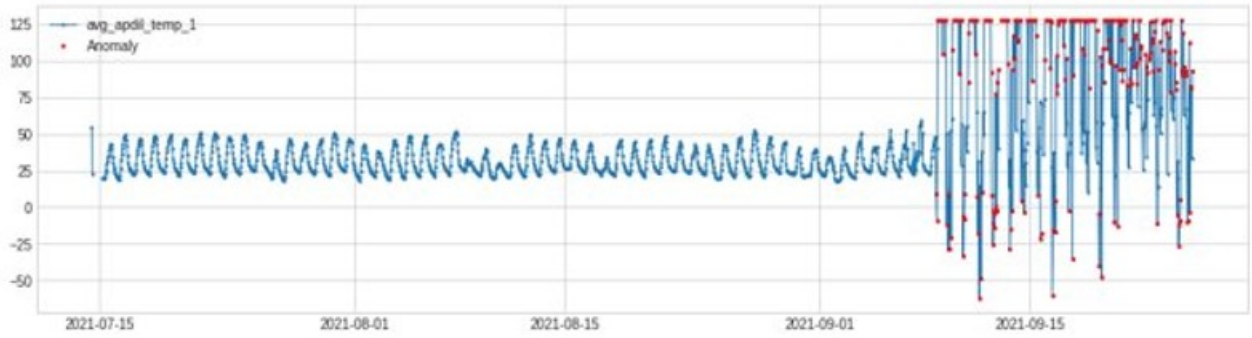

Figura 12: MÉTODO ThresholdAD, compara cada valor de serie de tiempo con umbrales dados.

umbrales deseados.

- Pese a que se identificó correlación entre la temperatura de las estaciones, la precipitación y los sensores de la ODT, gráficamente esta correlación no se logra identificar.

3.5. Análisis exploratorio de modelos de predicción

En esta etapa se hace un estudio de los modelos de predicción que se ajustan a los datos. Siendo esto así, la predicción de los datos de las variables correlacionadas se realiza por medio de regresión lineal y series de tiempo. Para la regresión lineal, se realizan dos predicciones: la primera utiliza regresión lineal con características polinómicas y la segunda, regresión lineal con características polinómicas y métodos de ensamble.

Después de analizar los respectivos errores de predicción al ejecutar cada modelo, se concluye que los mejores modelos para implementar y replicar son:

- Regresión lineal: con características polinómicas y métodos de ensamble, con el modelo gradient boosting.
- Series de tiempo: con el modelo Ridge o Random Forest.

Al realizar varias predicciones y comparar sus respectivos errores de predicción, se obtiene que estos son muy similares. Siendo esto así, el modelo que se implemento fue series de tiempo al resultar más adecuado en su aplicación.

3.6. Elaboración del modelo predictivo

Uno de los primeros objetivos en esta etapa era tomar las vías del tren, y a cada tramo de estas (44107 tramos en total), asociarle la estación climática más cercana, y posteriormente obtener una predicción de temperatura de cada tramo de la vía. Este procedimiento se realizó en QGIS, instalando el complemento NNjoin, que permite hacer cálculos entre vectores. Así, se cuenta con una nueva tabla que tiene todos los tramos de la línea junto con el nombre de la estación más cercana a cada tramo. Después se procede en Jupyter, a realizar la predicción en series de tiempo para el sensor de temperatura del APDIL, para cada tramo de la línea.

El primer modelo de predicción logrado tiene como resultado una función cuyo único parámetro a ingresar es el indice del tramo para el que se quiere la predicción:

```
info_com(64)
```

	tb_hora	avg_apdil_temp_1	avg_apdil_temp_1 (prediction)	join_indicativo	join_nombre	MAE_avg_apdil_temp_1	id_tramo	codigo_tramo	nombre_tramo
0	2021-08-28 01:00:00	26.281167	26.384106	8293X	XÀTIVA	1.83814	4.600000e+11	033000250	300 - MADRID CHAMARTÍN-VALENCIA-NORD
1	2021-08-28 02:00:00	24.854000	25.126508	8293X	XÀTIVA	1.83814	4.600000e+11	033000250	300 - MADRID CHAMARTÍN-VALENCIA-NORD
2	2021-08-28 03:00:00	23.676833	24.398964	8293X	XÀTIVA	1.83814	4.600000e+11	033000250	300 - MADRID CHAMARTÍN-VALENCIA-NORD
3	2021-08-28 04:00:00	22.781000	23.597780	8293X	XÀTIVA	1.83814	4.600000e+11	033000250	300 - MADRID CHAMARTÍN-VALENCIA-NORD
4	2021-08-28 05:00:00	21.999714	22.906217	8293X	XÀTIVA	1.83814	4.600000e+11	033000250	300 - MADRID CHAMARTÍN-VALENCIA-NORD
...	...	...	...	...	...	...	...	...	...
247	2021-09-07 19:00:00	37.447667	35.291657	8293X	XÀTIVA	1.83814	4.600000e+11	033000250	300 - MADRID CHAMARTÍN-VALENCIA-NORD
248	2021-09-07 20:00:00	50.229000	31.416863	8293X	XÀTIVA	1.83814	4.600000e+11	033000250	300 - MADRID CHAMARTÍN-VALENCIA-NORD
249	2021-09-07 21:00:00	29.822667	28.943089	8293X	XÀTIVA	1.83814	4.600000e+11	033000250	300 - MADRID CHAMARTÍN-VALENCIA-NORD
250	2021-09-07 22:00:00	27.947833	28.179043	8293X	XÀTIVA	1.83814	4.600000e+11	033000250	300 - MADRID CHAMARTÍN-VALENCIA-NORD
251	2021-09-07 23:00:00	26.541334	26.712953	8293X	XÀTIVA	1.83814	4.600000e+11	033000250	300 - MADRID CHAMARTÍN-VALENCIA-NORD

252 rows × 9 columns

Figura 13: Predicción para el tramo 4.6 e+11

La figura 13 muestra la ejecución de la predicción para uno de los tramos, cuya estación cercana es XÁTIVA, y permite comparar los datos reales con los datos que predice el modelo. Por otro lado la columna *MAE_avg_apdil_temp_1*, corresponde al error absoluto medio de la predicción de ese tramo en específico.

En la figura 14, se observa gráficamente los resultados de la predicción de la figura 13, donde, la línea azul corresponde al dato del sensor monitoreado, mientras que la línea de color naranja corresponde a la predicción del sensor realizada por medio del modelo. Se puede observar que el modelo es bueno y da una buena predicción que se ajusta a los datos.

Ahora, después de obtener un modelo de predicción que se ajusta a lo que busca el proyecto y que su respectivo error de predicción es bajo para la frecuencia de los datos (ver figura 13, columna *MAE_avg_apdil_temp_1*), procedemos a hacer las predicciones de las fechas de las que no se tiene registro de los sensores, es decir, desde el 8 de septiembre de 2021. La figura 15, muestra la ejecución de la predicción de 48 fechas hacia adelante del tramo que tiene indice cero, y en la figura 16 se puede ver gráficamente esta predicción.

Debido al tiempo de ejecución del modelo, se hace necesario considerar las predicciones sobre el número de estaciones climáticas y no sobre el número de tramos, dado que muchos de estos pueden tener la misma estación cercana, y es sobre ésta que se realiza la predicción.

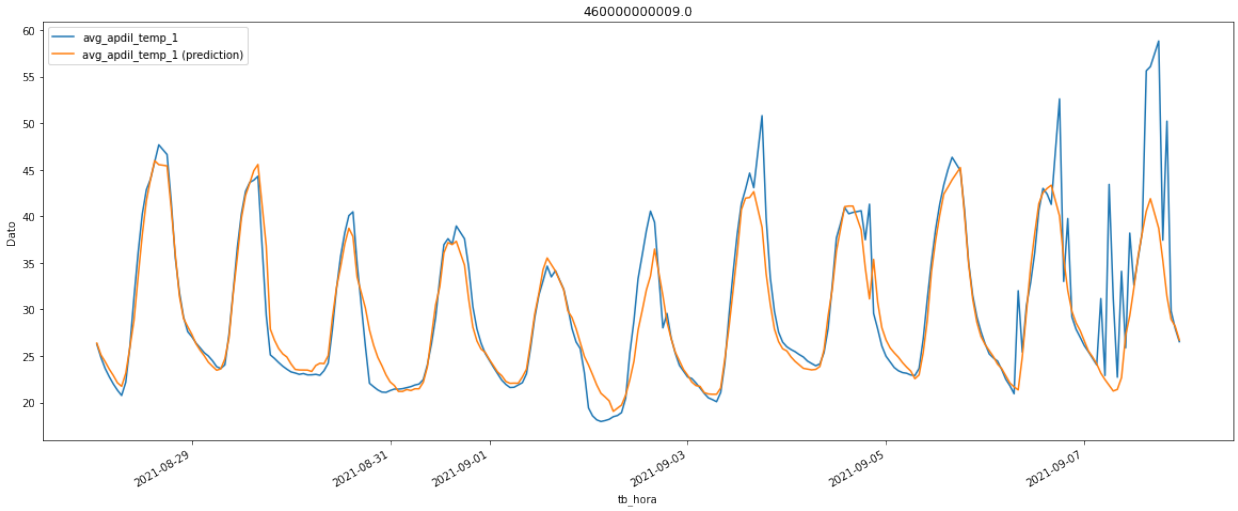

Figura 14: Comparativo de la predicción para el tramo 4.6 e+11

```
info_comp_tramo(0,48,'2021-09-08 00:00:00','2021-09-09 23:00:00')
```

	tb_hora	avg_apdil_temp_1 (prediction)	join_indicativo	join_nombre	id_tramo	codigo_tramo	nombre_tramo
0	2021-09-08 00:00:00	26.730926	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
1	2021-09-08 01:00:00	26.026349	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
2	2021-09-08 02:00:00	25.338336	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
3	2021-09-08 03:00:00	24.800602	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
4	2021-09-08 04:00:00	24.247269	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
5	2021-09-08 05:00:00	23.586076	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
6	2021-09-08 06:00:00	23.158865	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
7	2021-09-08 07:00:00	23.614746	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
8	2021-09-08 08:00:00	27.074020	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
9	2021-09-08 09:00:00	30.780869	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
10	2021-09-08 10:00:00	35.491196	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
11	2021-09-08 11:00:00	38.996866	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
12	2021-09-08 12:00:00	42.089134	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN
13	2021-09-08 13:00:00	43.943051	9001D	NESTARES	3.900000e+11	087900070	790 - ASUNCION UNIVERSIDAD-ARANGUREN

Figura 15: Predicción del tramo 3.90e+11 entre el ’2021-09-08 00:00:00’ y ’2021-09-09 23:00:00’

Es necesario mencionar, que antes de realizar la integración de las tablas y ejecución de las predicciones de cada día, se debe verificar que las estaciones climatológicas que se estén usando para las predicciones sí estén reportando datos en las fechas a predecir, de lo contrario es necesario quitarlas de las tablas. Es por esta razón que cuando inicialmente se cargan las tablas, se seleccionan 179 estaciones. Otra información que se tiene hasta la fecha es que solo se tienen datos de las estaciones climatológicas hasta el 21 de diciembre de 2021, por lo que hasta ese día se han realizado las predicciones correspondientes.

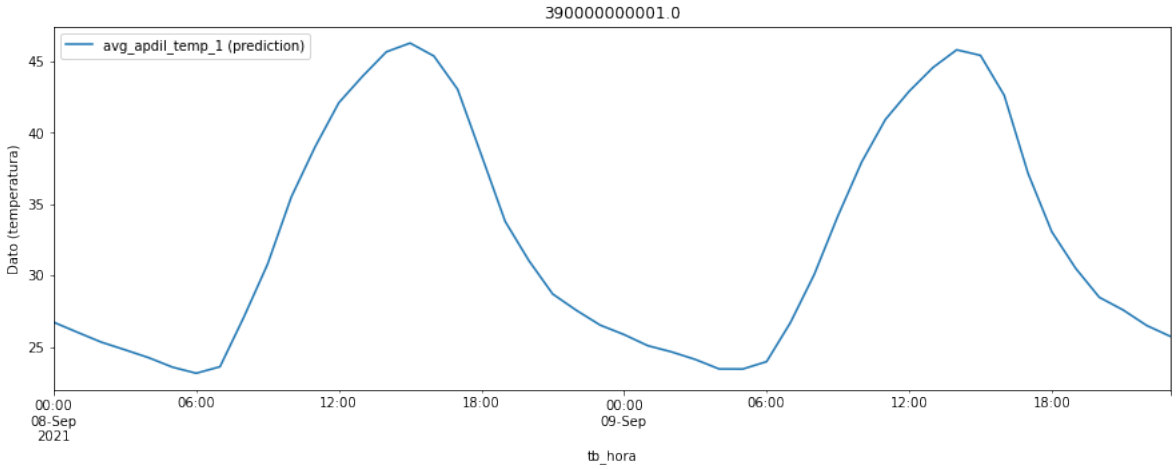

Figura 16: Predicción del tramo 3.90e+11 entre el ’2021-09-08 00:00:00’ y ’2021-09-09 23:00:00’

4. Resultados

Para el modelo final se tiene en cuenta que para cada día que se realice la predicción, está se debe integrar a la tabla inicial, de tal manera que se cuente con una tabla con datos al día. Este modelo construye una función que toma como parámetros, el número de estaciones de las que se quiere realizar la predicción, el intervalo entre la fecha inicial y final que se quisiera predecir, y da como resultado una tabla integrada con los datos de la predicción realizada y los datos anteriores a la fecha de la predicción.

El modelo final de predicción se realizó para cada tipo y número de sensor. Este modelo cuenta con dos versiones finales. La versión 1 es más fácil de ejecutar pues cuenta con un parámetro menos, pero a medida que van aumentando los datos, aumenta también el tiempo de ejecución. La versión 2, cuenta con un parámetro adicional (d), el cual permite que, para la predicción ejecutada, el número de datos se mantenga igual, por ejemplo, sí la primera predicción es la del día 8 de septiembre, para esta, d=0 pues no se elimina ningún número de datos, sin embargo, para la siguiente predicción, la del día 9 de septiembre, d=1, es decir, que se eliminan los datos de la primera fecha que se tienen registros. De esta manera se garantiza que el tiempo de ejecución para las predicciones de todos los días sea el mismo, el cual es aproximadamente de 2 minutos, 14 segundos.

La figura 17 muestra los parámetros que se necesitan para ejecutar la versión 1 del modelo final de predicción, para el sensor de temperatura del carril dado la temperatura de las estaciones climatológicas. De igual manera, la figura 18 para la versión 2 del modelo.

```
def pre_nuevas_fechas(i,n,s,e):   #s: star  e:end , fecha inicio y fecha final
    # i: valor puede variar entre  [0,187]  , que son los indices de la tabla df2
    # n es el numero de steps que quiero predecir, por ejemplo steps=24, nos dara las 24 horas siguientes.
```

Figura 17: Parámetros versión 1

```
def pre_nuevas_fechas(i,n,s,e,d):   #s: star  e:end , fecha inicio y fecha final
    # i: valor puede variar entre  [0,187]  , que son los indices de la tabla df2
    # n es el numero de steps que quiero predecir, por ejemplo steps=24, nos dara las 24 horas siguientes.
    # d: numero de predicciones, d=0 sera la primera predicción realizada
```

Figura 18: Parámetros versión 2

En la figura 19 se observa una ejecución del modelo para calcular la predicción del día 21 de octubre del 2021. Se puede notar que da como resultado una tabla integrada con los datos desde la primera fecha que se tienen registros hasta el día de la ejecución.

```
%%time
df=historico_estaciones(180,24,'2021-10-21 00:00:00','2021-10-21 23:00:00')
```

```
CPU times: user 3min 48s, sys: 232 ms, total: 3min 48s
Wall time: 3min 48s
```

```
df
```

	tb_hora	avg_apdil_temp_1	indicativo	nombre	tmed
0	2021-07-15 04:00:00	20.093500	0016A	REUS/AEROPUERTO	18.65
1	2021-07-15 05:00:00	19.604000	0016A	REUS/AEROPUERTO	17.65
2	2021-07-15 06:00:00	19.385333	0016A	REUS/AEROPUERTO	18.70
3	2021-07-15 07:00:00	19.562167	0016A	REUS/AEROPUERTO	21.60
4	2021-07-15 08:00:00	21.093500	0016A	REUS/AEROPUERTO	23.90
...	...	...	...	...	...
400798	2021-10-21 19:00:00	24.219007	B691Y	SA POBLA SA CANOVA	19.50
400799	2021-10-21 20:00:00	24.249297	B691Y	SA POBLA SA CANOVA	19.40
400800	2021-10-21 21:00:00	24.283195	B691Y	SA POBLA SA CANOVA	19.50
400801	2021-10-21 22:00:00	24.307274	B691Y	SA POBLA SA CANOVA	19.60
400802	2021-10-21 23:00:00	24.307274	B691Y	SA POBLA SA CANOVA	18.95

400803 rows × 5 columns

Figura 19: Predicción para el día 21 octubre del 2021 con la versión 1 del modelo final

5. Conclusiones y Recomendaciones

- Para la visualización y exploración de los datos se elaboró un Dashboard en PowerBi, que le permitiera al usuario interactuar con los datos, fechas y sensores para observar gráficamente el comportamiento de las variables.
- Se realizó un análisis del APDIL y de la ODT, con el fin de proponer un modelo predictivo de estos aparatos.
- Se construyeron dos modelos de predicción, teniendo en cuenta el personal encargado de realizar las predicciones diarias, esto por recomendaciones de AZVI.
- Finalmente, los dos modelos propuestos se diferencian en el tiempo de ejecución y en los parámetros de las funciones a ejecutar. Ambos modelos cuentan con errores de predicción similares, de lo que se puede concluir que no hay diferencia entre la calidad de los modelos.
- Lo que se pretende en un futuro es que las predicciones de todos los sensores se integren en un solo archivo, de tal manera que se puedan ejecutar desde una sola función.

Referencias

[1] Salvador, P., Insa, R., Viñas, V., Pineda, J., Martinez, P., Villalba, I. (2021). *Monitorización y análisis del comportamiento de aparatos de vía frente a eventos climáticos extremos*.

[2] Sistemas Inteligentes en Red, AZVI. (2019). *Bilateral Technological Cooperation Project Proposal Certification and Monitoring by CDTI (Spain)*.

[3] Rodrigo, J. (Octubre 2021). *Random Forest con Python*. Ciencia de Datos, Estadística, Machine Learning y Programación. `https://www.cienciadedatos.net/documentos/py08_random_forest_python.html`

[4] Rodrigo, J. (Marzo 2022). *Skforecast: forecasting series temporales con Python y Scikit-learn*. Ciencia de Datos, Estadística, Machine Learning y Programación. `https://www.cienciadedatos.net/documentos/py27-forecasting-series-temporales-python-scikitlearn.html`