

Master de Ciencia de Datos de Kschool

“Reducción del desvío de producción en una instalación de generación termosolar”

Marzo 2024

Autores:

Carlos Velázquez del Burgo

Jesús Villamuelas Álvarez

Joan Pasqual Llop Suñé

Abstract

El presente Trabajo de Fin de Máster (TFM) se centra en la aplicación de técnicas avanzadas de ciencia de datos para mejorar la precisión en la previsión de generación de energía en una central termosolar. Este proyecto aborda el reto de minimizar la discrepancia entre la generación de energía prevista y la real, crucial para optimizar el rendimiento y la eficiencia de la planta.

En primera instancia, se realiza un análisis exhaustivo de los datos históricos de la planta termosolar, incluyendo variables como la radiación solar, temperatura, presión, y producción energética pasada. Basándose en este análisis, se seleccionan y preparan los datos pertinentes para el modelado predictivo.

Posteriormente, se implementan y comparan varios modelos de machine learning, incluyendo redes neuronales, árboles de decisión y algoritmos de regresión. Estos modelos se entrenan y validan utilizando un conjunto de datos históricos, con un enfoque en la optimización de su capacidad para predecir con precisión la generación de energía en diferentes condiciones ambientales y operativas.

El proyecto también contempla la integración de técnicas de análisis de series temporales para mejorar las previsiones, teniendo en cuenta la naturaleza dinámica y a menudo no lineal de la generación de energía termosolar. Además, se exploran estrategias para el tratamiento de datos faltantes y la gestión de variables influyentes.

Los resultados obtenidos demuestran una mejora significativa en la precisión de las previsiones, reduciendo el desvío entre la generación de energía prevista y la real. Esta mejora representa un avance importante para la gestión eficiente de la central termosolar, permitiendo una planificación más precisa y una mejor respuesta a las fluctuaciones en la demanda y oferta de energía.

Finalmente, el estudio concluye con recomendaciones para la implementación práctica de los modelos desarrollados, así como sugerencias para futuras investigaciones que podrían extender y profundizar los hallazgos de este trabajo.

Índice

[**Introducción 4**](#_g8wog22dg0xk)

[**Objetivos 5**](#_tm6xx4g1b8j3)

[**Datos 6**](#_xt6zrthopzgh)

[**Previsión de energía generada por instalación termosolar 6**](#_z9stg2ogbl02)

[**Energía generada por la instalación 8**](#_bxmr11a63754)

[**Desvío de generación 10**](#_dfrzb6oilxb4)

[**Tecnología 12**](#_8lj4y1x3owx0)

[**Procesamiento y Análisis de Datos: 12**](#_mpt6to3ijwk)

[**Desarrollo de Modelos de Machine Learning: 12**](#_dnvtrujvefn4)

[**Visualización de Datos: 12**](#_3gofysv4snxv)

[**Control de Versiones y Colaboración: 12**](#_5kwpv5zbitnb)

[**Lenguajes de Programación 12**](#_82nz8e2mdaeg)

[**Modelización 13**](#_eshv41ifsyrz)

[Regresión Lineal simple 13](#_g1mupmniipm1)

[Random Forest 15](#_u4uzxfjg4qec)

[Gradient Boost Regressor 17](#_2vsvv2sh55tz)

[**Resultados 19**](#_mc0p4mptu45t)

[**Despliegue 20**](#_tcr6whowazf9)

[**Puesta en valor 21**](#_iqbov3w1rkke)

[**Conclusiones 22**](#_y8iigeaj33o5)

[**Autores y contribución 23**](#_9szi4jtlw7o7)

[**Bibliografía y Recursos 24**](#_21azfp1axom5)

[**Anexo 25**](#_hffymau5gvxg)

# Introducción

Las centrales solares tienen, al igual que el resto de instalaciones de generación eléctrica, conectadas a la red de transporte de España y participan en el mercado eléctrico español, la obligación de informar en el día D-1 la producción prevista dividida en los 24 periodos del día.

Con el fin de la correcta programación de la energía generada y consumida ( no es posible un exceso o un defecto de la misma en el sistema ) se establece un sistema de penalizaciones económicas, en función del desvío de la instalación frente al desvío resultante del conjunto del sistema eléctrico.

El hecho anterior hace necesaria que la previsión de generación sea lo más precisa a la generación real de energía, reduciendo de esta forma el impacto económico en la compañía propietaria de la instalación de generación.

El presente proyecto va a desarrollar una serie de modelos basados en ciencia de datos, que buscan mediante varias herramientas, corregir la previsión recibida y reducir el desvío resultante.

Este proyecto es un reto importante ya que la previsión de generación es calculada por empresas con gran experiencia en el sector y un número importante de inputs. Aun con ello, el know how del que se dispone y la posibilidad de alimentar el modelo con datos solo accesibles por el Centro de Control asociado a la instalación, dan la posibilidad de conseguir dicha mejora.

La motivación principal de realizar este modelo es fundamentalmente conseguir una solución óptima para este caso y fácilmente extrapolable a otras instalaciones de similar tecnología, pudiendo así dar una gran utilidad a la empresa propietaria de la instalación o al Centro de Control asociado, que hace las veces de representante de mercado y tiene la potestad de modificar la previsión de generación enviada al gestor del sistema eléctrico.

# Objetivos

El objetivo final de este proyecto es la reducción en la medida de lo posible, del desvío producido entre la previsión de generación eléctrica, provista por una empresa externa y la energía producida resultante medida en el contador de una instalación termosolar.

El objetivo empresarial se basa en la disminución del coste de sobrecoste del desvío resultante de la diferencia entre previsión y energía real.

El análisis de los diferentes comportamientos históricos de la instalación ante periodos concretos y estacionalidad de dicho comportamiento, son los puntos de partida de análisis.

Una reducción de esta magnitud por leve que parezca en el resultado inicial, puede traducirse en un impacto económico considerable para el negocio, al estar vinculado a un precio del sobrecoste que puede llegar a ser muy elevado. Esta reducción si es mantenida en el tiempo por el modelo, conlleva un ahorro de la penalización significativa.

# Datos

El presente proyecto va a tratar fundamentalmente dos datos principales. **Previsión de generación** (MWh) y **Energía real medida en contador de central** (MWh). En primera instancia serán estas dos variables las que compondrán el dataset original del estudio.

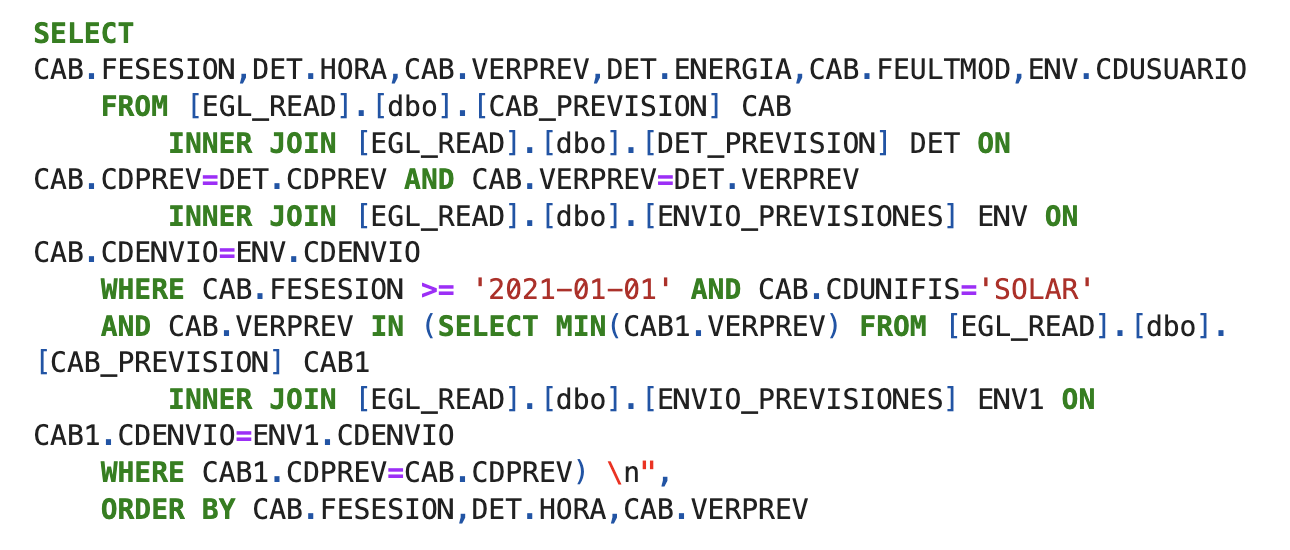
Estas dos variables son lo suficientemente importantes y relacionadas con el desvío objetivo del proyecto, como para componer la base del proyecto.

Como posible mejora del proyecto a futuro será la inclusión de las variables meteorológicas o eventos adversos, que por conocimiento de negocio se observan a priori, muy vinculadas al comportamiento final de la instalación de generación. Por complejidad y acceso de pago al dato, se descartan en primera instancia a incluir en el modelo, quedando para versión posterior, si se consiguiera la aplicación en negocio real que pudiera suplir el coste de estos datos.

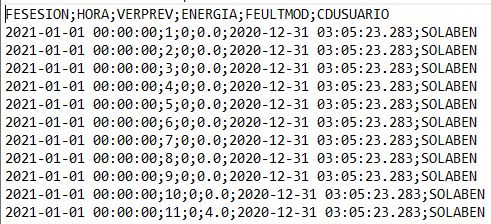
## **Previsión de energía generada por instalación termosolar**

Los datos de previsión de generación son proporcionados por la empresa **Meteológica**. Estas previsiones son insertadas de manera diaria en la base de datos de Axpo, empresa que colabora con la realización de este TFM.

La obtención de los datos se realiza de la base de datos SQL de **Axpo** que actúa como concentrador de previsiones de las diferentes instalaciones de su cartera de representación.

A través de la siguiente query se obtienen los datos de previsión horaria para la UFI que representa la unidad de generación objeto de este estudio. Por LPD se utiliza un Alias en sustitución de la Unidad Física que es el código de identificación de la instalación en el sistema.

El resultado de esta query se exporta en archivo csv separado por comas

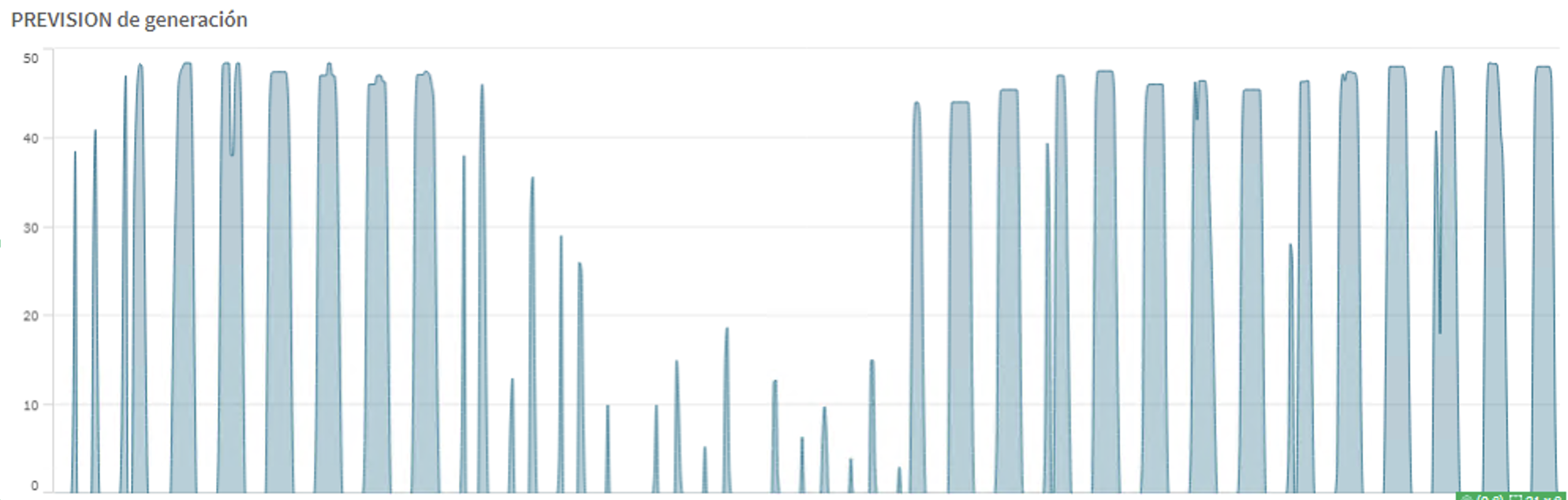


Los datos obtenidos conforman un dataset de 25272 registros divididos en los siguientes campos:

* **FESESION**: Será la fecha relativa a la previsión insertada. El tipo de dato es de “small datetime” (a pesar de ser siempre 00:00:00), con el formato YYYY-MM-DD hh:mm:ss.

En el presente estudio la fecha partirá del 01-01-2021 hasta el 17-11-2023

* **HORA**: Corresponde al periodo eléctrico en el que se divide el día. Corresponde a las horas del día, yendo de 1 a 24 (23 o 25 en los días de cambio de hora). El tipo de dato es un “int”.
* **VERPREV**: indica que versión de la previsión es. Esto es, las previsiones habitualmente se suelen actualizar por parte de la empresa proveedora del servicio. Ante un cambio de las condiciones climatológicas o por indisponibilidad técnica de la instalación, la previsión se actualiza incrementando su versión. En este caso se obtiene la última previsión válida para el estudio. El valor de la versión no se incorporará al dataset del proyecto al carecer de importancia. El tipo de dato es “int”.
* **ENERGIA:** Muestra la energía prevista a producir. Su magnitud será de MWh. Será una de las dos principales medidas del dataset final. El tipo de dato es “float”. Los valores esperados en esta magnitud son mayores a 0 (no se contemplan valores negativos o de consumo) hasta 49.9 MWh que sería la energía máxima que la instalación puede producir.

*Ejemplo del dato de previsión para un mes*

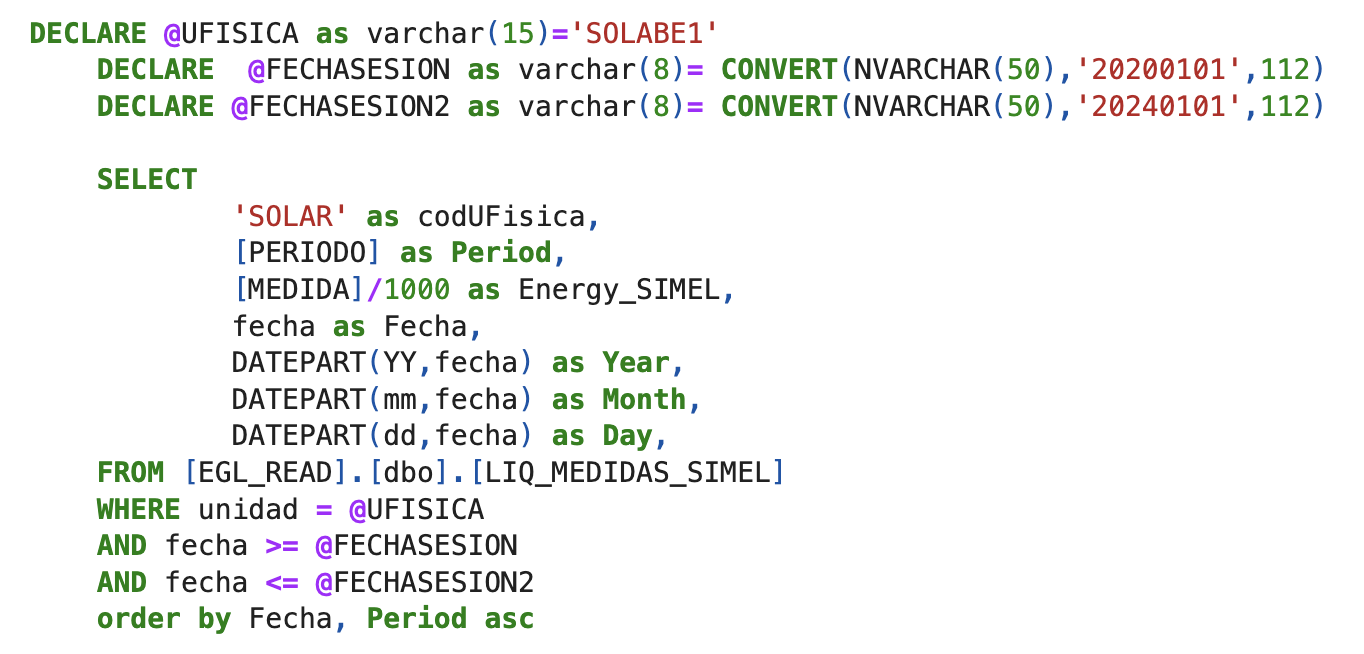
* **FEULTMOD:** Hace referencia a la fecha en la que se insertó la previsión en la base de datos. Vinculado a la variable VERPREV. Su formato es “smalldatetime”. Se obviara en el dataset al no tener valor para el estudio, siendo únicamente informativo de cara a auditar y validar el proceso de envío de previsiones por parte del proveedor.
* **CDSUARIO:** Por LOPD se obvia este campo ya que hace referencia al usuario que insertó la previsión en la base de datos.

## **Energía generada por la instalación**

Los datos de energía de generación se obtienen originariamente de la lectura telemática del contador instalado en el punto frontera de la instalación. El sistema eléctrico español obliga a disponer de un contador bidireccional y consultable de forma telemática a través de una IP y clave de lectura.

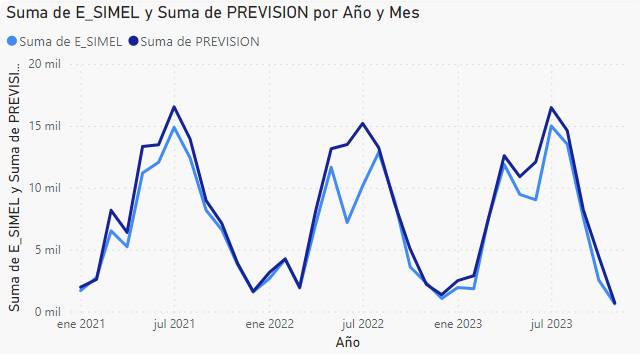
Esta lectura la realiza el organismo **SIMEL** de forma diaria. Este organismo es el encargado de la lectura y agregación de las curvas de potencia de energía generada de todos los productores eléctricos.

Los representantes de mercado tienen la potestad de consultar estos datos mediante un acceso privado. En el caso que nos ocupa **Axpo** como representante de mercado de la instalación objeto del estudio, es quien recopila estos registros de forma periódica para importarlos en base de datos interna, con el fin de usar los mismos, para liquidaciones y otras funciones. **Axpo** como colaborador de este proyecto, nos permite la consulta de los históricos de producción siempre que se mantenga el anonimato de la instalación.

La obtención de los datos se realiza de la base de datos SQL de **Axpo** siguiendo la siguiente consulta en SQL Server.

*Ejemplo del dato de energía para un mes*

*Análisis gráfico de la energía y la previsión*



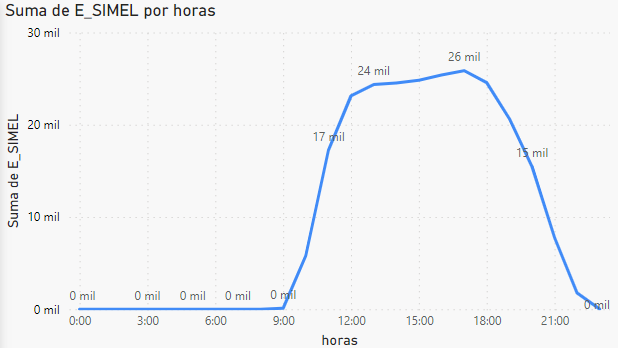
Se puede observar como ambas series temporales tienen un comportamiento muy similar y con datos muy parejos, destacar que hay momentos en los que se evidencia los malos ajustes de la previsión, que suele ser en los meses de julio, principalmente en julio de 2022.

Lo que apoya nuestro objetivo de mejorar la previsión.

En cuanto a la serie temporal se observa que es no estacionaria, debido a que presenta, estacionalidad, tendencia y heterocedasticidad.

Tiene claros máximos en los meses de verano ya que la temperatura es mayor en la latitud en la que se encuentra la planta.

*Análisis gráfico de la energía por horas*

**

En el gráfico de la energía por horas, graficado de manera agregada, también observamos una estacionalidad, debido a la naturaleza de la planta, en las horas nocturnas no hay producción, también se deduce que las horas con mayor producción y por ello mayor incidencia de radiación solar son el periodo del mediodía entre las 12.00 horas y las 17.00 horas.

## **Desvío de generación**

Será el objeto del proyecto y su reducción la finalidad de los modelos a desarrollar.

Se trata de un dato calculado a partir de las dos medidas anteriores, siendo la diferencia entre ambas su resultado.

DESVIO = PREVISION - ENERGIA SIMEL

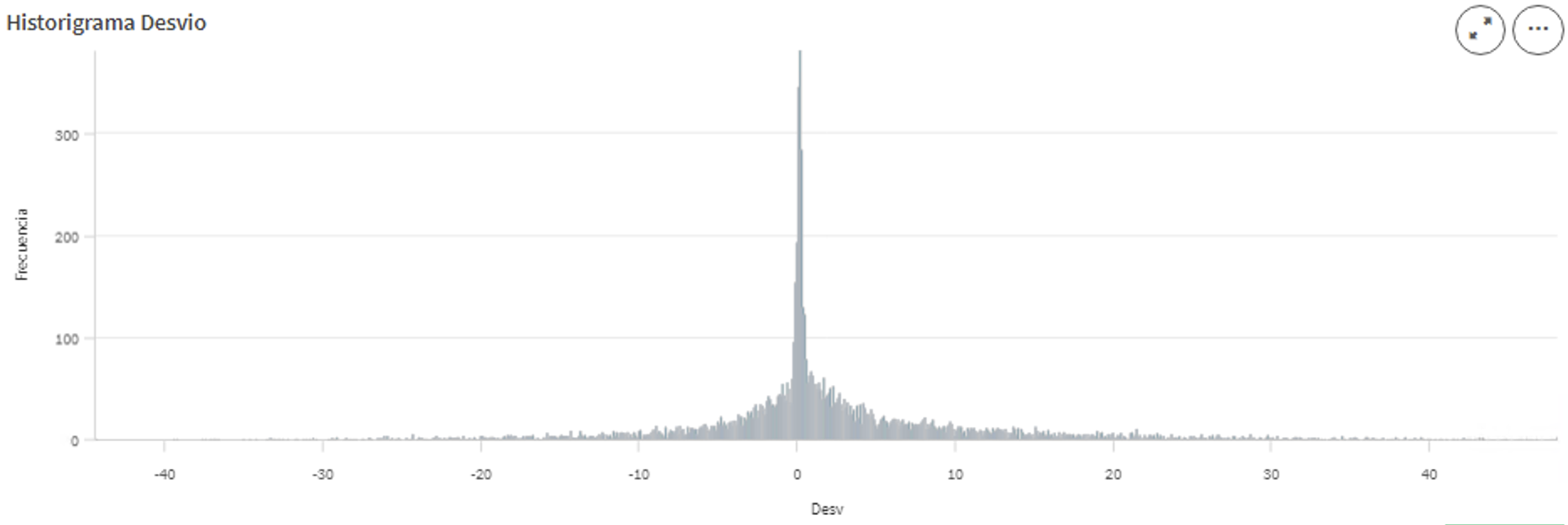
De esta manera el criterio de signos será: (+) Exceso de previsión (-) Exceso de energía

El resultado del desvío será esencial ya que es directamente proporcional a la penalización económica asociada.

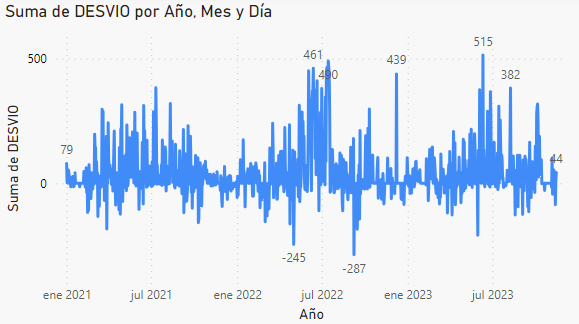
También su signo determinará si existe penalización o no, ya que en caso que el signo del desvío sea diferente al resultante del desvío general del sistema eléctrico español, no conlleva una penalización

Para los efectos del proyecto nos basaremos en la reducción absoluta del desvío, no contemplando el factor de mercado.

En la exploración y análisis del dataset, vamos a obviar los resultados de desvío 0 cuando las variables de previsión y energía SIMEL son ambas cero. Esto es, por tener una vinculación directa a las horas solares, existen numerosos períodos que cumplen esta premisa, resultando un gran sesgo en el dataset.



*Análisis gráfico del desvío*



Con la gráfica del desvío se refuerza nuestro objetivo para mejorar la previsión, aunque no es muy grande, hay periodos como los meses de julio, citados anteriormente en los que el mismo aumenta.

# Tecnología

## Procesamiento y Análisis de Datos:

* QlikSense Server (versión 14.113.12): Utilizado para el procesamiento eficiente de grandes conjuntos de datos.
* PowerBI Free Desktop: Utilizado para la visualización de grandes conjuntos de datos
* Pandas (versión 1.2.4): Empleado para la manipulación y análisis de datos.
* NumPy (versión 1.20.3): Utilizado para cálculos numéricos avanzados.

## Desarrollo de Modelos de Machine Learning:

TensorFlow (versión 2.4.1): Usado para construir y entrenar modelos de aprendizaje profundo.

Scikit-learn (versión 0.24.1): Utilizado para algoritmos de aprendizaje automático y herramientas de modelado estadístico.

## Visualización de Datos:

* Matplotlib (versión 3.3.4) y Seaborn (versión 0.11.1): Empleados para la creación de gráficos y visualizaciones de datos.

## Control de Versiones y Colaboración:

* GitHub: Utilizado para control de versión y repositorio de código colaborativo.

## Lenguajes de Programación

* Python (versión 3.8.5): Lenguaje principal para el desarrollo de scripts y análisis de datos.
* SQL para la adquisición de datos procedentes de base de datos.

# Modelización

Se van a presentar los tres modelos que han dado los resultados positivos más consistentes para nuestro objetivo.

Todos los modelos utilizados se encuentran en nuestro repositorio de Github.

<https://github.com/jesusvillaalvarez/TFM_KSCHOL/blob/main/4.MODELO/4.2_MODELOS_PRUEBA/MODELO_REGRESION_LINEAL_2.ipynb>

## Regresión Lineal simple

Presentamos primero un modelo de Regresión Lineal simple. Partimos de nuestra variable objetivo E\_SIMEL y utilizamos para el modelo la variable PREVISION que es la que tiene mayor correlación con la variable objetivo:

MATRIZ DE CORRELACIÓN

| DESVIO | -0.081421 |
| --- | --- |
| Es\_fin\_semana | -0.003424 |
| Dia\_Semana | -0.000052 |
| Mes | 0.016508 |
| f\_PREV\_LOW | 0.435761 |
| f\_PREV\_HIGH | 0.492146 |
| f\_RUN | 0.803932 |
| PREVISION | 0.888352 |
| E\_SIMEL | 1.000000 |

Para este modelo dividimos nuestro conjunto de datos de entrenamiento y prueba a partir de las fechas; cogeremos el primer 85% de los datos según las fechas para el conjunto de entrenamiento y el 15% restante para la prueba del modelo.

Evaluamos el resultado de las predicciones con el mse (error cuadrático medio), el r2 (coeficiente de determinación), el coeficiente para la PREVISON y el intercepto para la variable E\_SIMEL:

| Coeficiente | Intercepto | mse | r2 |
| --- | --- | --- | --- |
| 0.8161717082413648 | 0.4560723343705586 | 56.85411270876022 | 0.8346287013233246 |

El resultado del coeficiente para la PREVISION nos indica que por cada unidad que incrementa la PREVISIÓN, se esperará que la variable E\_SIMEL incremente en 0.816 lo que nos demuestra la relación directa entre estas dos variables.

El resultado del intercepto nos indica el valor esperado de la variable E\_SIMEL cuando la/las variables independientes, en este caso PREVISION sean 0. Es un dato interesante y veremos cómo evolucionan las predicciones.

A continuación, cogemos los datos de los días de noviembre de 2023 que tenemos disponibles y hacemos la predicción de la variable objetivo, día a día, y con que tenemos los resultados reales, podemos hacer la comparación de los resultados.



Podemos apreciar en este gráfico como la predicción del modelo se acerca más a la producción real de energía, que los datos de PREVISION.

Los resultados numéricos son los siguientes:

Suma de los valores en la columna E\_SIMEL: 622.953

Suma de las predicciones: 667.0568361849755

Suma de las previsiones: 696.6

Diferencia entre predicciones totales y E\_SIMEL total: 44.103836184975535

Diferencia entre previsiones y E\_SIMEL total: 73.64700000000005

**La predicción es MEJOR** que la previsión en: **29.543** unidades, por lo tanto, cumplimos nuestro objetivo de mejorar la PREVISIÓN.

## Random Forest

El modelo de Random Forest con imputación de la variable f\_RUN.

<https://github.com/jesusvillaalvarez/TFM_KSCHOL/blob/main/4.MODELO/4.2_MODELOS_PRUEBA/MODELO_RANDOM_FOREST_IMPUTACION_f_RUN.ipynb>

Partimos de la premisa que nos marca la matriz de correlación ya que la variable f\_RUN es la segunda variable con más peso respecto la variable objetivo. Aunque la variable f\_RUN es un dato que no se tiene disponible para hacer las predicciones ya que es un dato que se da a posteriori, decidimos utilizar un imputador para imputar valores a esta columna para así poder utilizarla para la predicción.

Para este modelo utilizamos todas las variables disponibles menos las variables que, como f\_RUN, se dan pasado el día de predicción: DESVIO, f\_PREV\_HIGH y f\_PREV\_LOW.

Dividimos el Dataset en dos conjuntos. El primero que recoge los datos des del inicio hasta el 31 de octubre de 2023 y el segundo conjunto recoge los datos del mes de noviembre de 2023, que al igual que en el modelo de Regresión Lineal, los utilizaremos para ir haciendo las predicciones día a día y poder ver la comparación de los resultados respecto las previsiones.

Para empezar con el modelado, utilizamos un modelo de RandomForestRegressor con un 80% de datos para el entrenamiento y un 20% para la prueba, del primer conjunto de datos.

Utilizamos un GridSearchCv para la búsqueda de los mejores hiperparámetros que le estamos pasando para así configurar el modelo.

También utilizamos el primer conjunto de datos para entrenar el imputador MICE con un modelo de RandomForestRegressor. Para entrenar el imputador utilizamos todas las variables disponibles menos la variable objetivo.

Una vez tenemos los dos modelos entrenados, procedemos a hacer los pasos para la predicción. Primero cogemos los datos para el primer día del mes de noviembre que queremos hacer la predicción y utilizamos el imputador MICE para dar datos a la variable f\_RUN. Una vez hecho este paso podemos proceder a hacer la predicción.

Una vez hecha la predicción y ya con la disponibilidad de los datos reales del día 5 de noviembre, actualizamos el primer conjunto de datos con los datos reales del día 5 para reentrenar el modelo de predicción y el imputador.

Y partir de los reentrenos volvemos a repetir el proceso con la carga de datos del día 6 para hacer la imputación de la variable f\_RUN y a hacer la predicción. Este proceso lo repetimos hasta que tenemos todas las predicciones de los días de noviembre que tenemos disponibles y pasamos a presentar los resultados numéricos:

Suma de los valores en la columna E\_SIMEL: 622.953

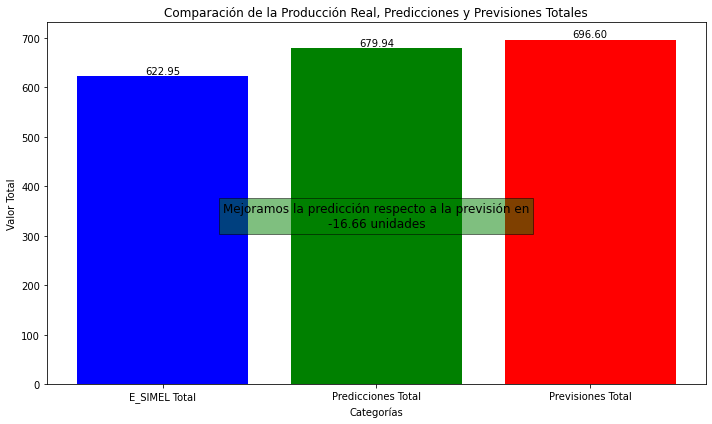
Suma de las predicciones: 679.93608

Suma de las previsiones: 696.6

Diferencia entre predicciones totales y E\_SIMEL total: 56.98307999999997

Diferencia entre previsiones y E\_SIMEL total: 73.64700000000005

**La predicción es MEJOR** que la previsión en: **16.663920000000076**, por lo tanto, cumplimos nuestro objetivo de mejorar la PREVISIÓN.



Como podemos ver, con este modelo también mejoramos los datos de la Previsión, aunque no son tan buenos como el de Regresión Lineal.

## Gradient Boost Regressor

Modelo que nos ha dado resultados mejores y consistentes. GradientBoostingRegressor con imputación de la variable f\_RUN.

<https://github.com/jesusvillaalvarez/TFM_KSCHOL/blob/main/4.MODELO/MODELO_GRADIENT_BOOSTING_IMPUTACI%C3%93N_f_RUN.ipynb>

El modelado de este modelo es exactamente igual que el modelo de Random Forest:

* Nos quedamos con la variable f\_RUN para hacer la imputación
* Entrenamos un modelo de GradientBoostingRegressor, antes con la búsqueda de mejores hiperparámetros con el GridSearchCV.
* Entrenamos el imputador MICE con un modelo de RandomForestRegressor.
* Una vez entrenados los modelos con el primer conjunto de datos pasamos a las predicciones de los datos de noviembre.
* Utilizamos el imputador para dar valores a la variable f\_RUN.
* Hacemos las predicciones de la variable objetivo E\_SIMEL.
* Alimentamos el primer conjunto de datos con los datos del primer día de predicción.
* Y volvemos a reentrenar los modelos de predicción y de imputación.
* Repetimos los procesos para todos los días del mes de noviembre de que disponemos datos.

Los resultados numéricos de este modelo:

Suma de los valores en la columna E\_SIMEL: 622.953

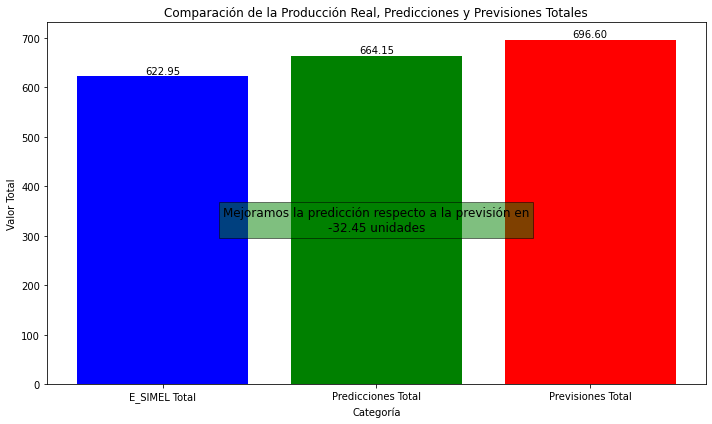
Suma de las predicciones: 664.1521659123504

Suma de las previsiones: 696.6

Diferencia entre predicciones totales y E\_SIMEL total: 41.199165912350395

Diferencia entre previsiones y E\_SIMEL total: 73.64700000000005

**La predicción es MEJOR** que la previsión en: **32.44783408764965**, por lo tanto, cumplimos nuestro objetivo de mejorar la PREVISIÓN.



Como podemos observar es el resultado más positivo de los tres modelos que hemos visualizado y que a la vez también es consistente con 32.44 unidades a favor de la predicción respecto la Previsión.

Siguiendo el mismo enfoque de imputación de variables que no tendríamos disponibles a la hora de hacer las predicciones, hemos explorado la imputación de más variables como la variable DESVIO y la variable f\_PREV\_HIGH, con modelos de GradientBoosting cambiando los hiperparámetros del GridSearchCV haciendo los modelos un poco más complejos.

Aunque en algunos modelos hemos obtenido resultados más positivos que en los anteriores también se ha dado el caso de obtener resultados negativos ejecutando el mismo modelo lo que nos indica que no son resultados consistentes y que no nos podemos fiar de las predicciones.

También hemos explorado modelo de Deep Learning, concretamente 3 modelos en que hemos ido incrementando la complejidad en la arquitectura de los modelos, como por ejemplo añadiendo unidades en las capas ocultas, añadir regularizaciones L2, normalización por lotes para estabilizar y acelerar el entrenamiento, añadir Dropout en la salida de la segunda capa oculta y también añadiendo earlystopping y ReduceLROnPlateau como CallBacks en el. fit().

Los resultados de los dos modelos más sencillos de DeepLearning siempre nos han dado resultados negativos para nuestro propósito y el modelo más complejo que hemos utilizado, si que en algunas ejecuciones nos ha dado resultados positivos, pero también de negativos.

# Resultados

Como se detalla en el apartado anterior se obtienen resultados de mejora absoluta en el desvío obtenido.

Esta mejora trasladada a negocio se traduce en un ahorro económico sea cual fuere el escenario de precios de sobrecoste. El impacto real en lo económico se debe trasladar junto con el resto de la cartera de representación de mercado ya que la liquidación del sobrecoste del desvío se realiza a nivel de cartera, no siendo posible mostrar un número definitivo.

El resultado es óptimo y mejorable añadiendo módulos extra al modelo que ofrece los mejores resultados, de todos los entrenados. Quedando como futuros puntos de mejora, la detección de eventos climatológicos extraordinarios o un módulo de detección de no arranque de la instalación.

# Despliegue

La puesta en producción del presente modelo, se realizará en un servidor Windows Server 2020 virtualizado en un DataCenter, siguiendo la actual estructura de servidores de la empresa colaboradora del proyecto.

En primer lugar la adquisición del dato se desarrolla mediante script basado en Python que realizará las consultas a base de datos SQL.

Para ello se ejecutarán los archivos .py mediante el programador de tareas de Windows, a la hora que se asegure que los datos se encuentran importados. Existe la posibilidad de correr el proceso de ingesta de datos de forma manual, ejecutando los proyectos en el programador de tareas de Windows si se necesitará de forma puntual.

Los archivos csv generados se guardarán en el propio servidor quedando accesibles tanto de forma local para el / los modelos como en red para una posible consulta por Power BI o QlikSense, actuales herramientas de BI de la empresa.

Siguiendo la automatización de procesos, la ejecución del modelo se realizará mediante una tarea condicionada a la previa adquisición del dato, de esta forma siempre se ejecutará una vez se dispongan de los datos actualizados.

El resultado del modelo será exportado en un archivo csv dentro del repositorio general del modelo.

Desde el departamento de Advance Analytics consultarán en la H23 el resultado y tomarán la decisión basada en la estrategia del departamento, de incorporar el resultado del modelo.

# Puesta en valor

La aplicación en un hipotético caso real es inmediata. Esto es, basta con automatizar la consulta a las bases de datos SQL Server para obtener los datos de previsión para el día D+1. A su vez el dataset deberá también ser completado con la energía SIMEL que se carga de forma diaria en la base de datos SQL utilizada en el punto “Datos”.

Esta automatización hará una carga incremental diaria almacenando el resultado en el dataset en formato csv.

Posterior a la actualización del dato se realizará la tarea de limpieza y homogeneización similar a la realizada en el proyecto, verificando que no existan datos fuera de escala.

Tras esta verificación se incorporará al dataset definitivo, como fuente principal para la ejecución del modelo.

Este modelo realizará el cálculo con los datos actualizados, devolviendo una previsión para el día D+1.

Será atribución del equipo de Advance Analytics de la empresa Axpo, quien opte por insertar esta previsión en el portal web del operador del sistema REE.

Una posterior puesta en producción del modelo puede ser el establecer un peso determinado a la previsión del modelo y otro a la previsión del proveedor Meteorológica, a partir de un estudio continuo, que haga tener mayor presencia a una fuente u otra, actuando como contrapeso mutuo.

# Conclusiones

Después de explorar diferentes modelos regresión como son los de Regresión Lineal, Random Forest, Gradient Boosting, además de modelos de Deep Learning y de Series Temporales, según nuestros resultados llegamos a la conclusión, con los datos que teníamos disponibles, que los modelos que han dado los mejores y más consistentes resultados, son los más sencillos:

Regresión Lineal, con resultados consistentes de 29.54 unidades favorables para nuestro objetivo.

Random Forest con imputación de la variable f\_RUN, variable con peso en nuestro DataFrame, con resultados de 16.66 unidades positivas para nuestro objetivo.

Gradient Boosting con imputación de la variable f\_RUN, con resultados de 32.44 unidades positivas para nuestro objetivo.

Los modelos más complejos que también hemos explorado, aunque alguno de ellos nos han dado resultados positivos, incluso a veces mejores que los anteriores, no son consistentes pudiendo dar alguna vez resultados negativos.

De todas formas, creemos haber cumplido nuestro objetivo de mejorar las previsiones, lo que implica un ahorro importante, en cuestión de penalizaciones para la planta termosolar en concreto. Es un proyecto que da solución, en parte, a un problema empresarial que, en caso de implantarlo, la planta termo solar obtendría mejores resultados que con las previsiones que utilizaba hasta ahora.

También cabe comentar, que disponíamos de pocas variables para hacer el trabajo y que creemos que, con la incorporación de datos meteorológicos a los modelos, podríamos conseguir mejores resultados. Se podría plantear a la dirección de la planta termosolar, que incorpórase, como mínimo, una estación meteorológica en la central. Estos datos fiables y de primera mano, seguramente ayudarían a mejorar las previsiones, con la disminución de las correspondientes penalizaciones.

Con la exploración de todos los modelos, lo que hemos aprendido es que la posibilidad de ir probando diferentes estrategias de modelado y de tratamiento de variables, dan muchas opciones para seguir trabajando en mejorar las predicciones.

# Autores y contribución

* + Conceptualización del caso de Uso: JVA
  + Extracción de Datos: JVA
  + Tratamiento y Limpieza de Datos: JVA, JPS
  + Plataforma Tecnológica: JVA, JPS, CVB
  + Desarrollo de algoritmos: JPS
  + Visualización y maquetado de resultados en BI: CVB
  + Elaboración de Memoria: JVA
  + Elaboración de Video TFM: JPS, CVB, JVA

# Bibliografía y Recursos

* Libros
  + von Meier, A. (2006). Electric Power Systems: A Conceptual Introduction. John Wiley & Sons.
  + Donovan, C. W. (2019). Renewable Energy Finance: Funding The Future Of Energy. Imperial College Press.
  + Chase, J. (2019). Solar Power Finance Without the Jargon. Springer.
  + Lovegrove, K., & Stein, W. (Eds.). (2012). Concentrating Solar Power Technology: Principles, Developments and Applications. Woodhead Publishing.
* Articulos científicos
  + Bacher, P., Madsen, H., & Nielsen, H. A. (2009). A review of solar forecasting techniques and their application to the electricity markets. Journal of Renewable and Sustainable Energy, 1(5), 051404.
  + Caldes, N., Varela, M., Santamaria, M., & Saez, R. (2010). Techno-economic analysis of solar thermal power plants. Energy Procedia, 2(1), 69-80.
  + Weron, R. (2006). Modeling and forecasting electricity loads and prices: A statistical approach. Wiley.
  + Blanco, M., & Ramírez Santigosa, L. (Eds.). (2016). Advances in Concentrating Solar Thermal Research and Technology. Woodhead Publishing.
* Recursos web
  + Agencia Estatal de Meteorología (AEMET). (2023). <http://www.aemet.es/>
  + Sistema de Información del Mercado Eléctrico (SIMEL). (2024). https://www.ree.es/es/clientes/generador/gestion-medidas-electricas
  + Web de Clientes AXPO Iberia. (2024). https://www.despachoaxpo.com
* Otros recursos
  + Formación sobre mercado eléctrico. Previsiones y Desvíos. Autor: Axpo Iberia 2023

# Anexo

En este anexo se detalla la forma de trabajo seguido para la elaboración del presente proyecto.

El punto de partida es de tres colaboradores con pleno derecho de edición tanto en documentos como código. Geográficamente a distancia por lo que se ha optado por una modalidad de trabajo a distancia, utilizando herramientas de almacenamiento cloud que sirviera de repositorio a los participantes.

Estas soluciones han sido fundamentalmente:

GitHub:

<https://github.com/jesusvillaalvarez/TFM_KSCHOL>

como repositorio colaborativo de código y entorno de ejecución de código a la vez que versionado de mejoras y aportaciones individuales.

Se utiliza como fuente final de entrega del proyecto incluyendo código, datos y cuadernos explicativos de procesos.

La estructura es la siguiente:

1. DEFINICIÓN DEL PROBLEMA
2. RECOPILACIÓN DE DATOS
   1. EXTRACCIÓN DE DATOS
   2. REPOSITORIO DATOS BRUTOS
   3. REPOSITORIO DATOS PROCESADOS
   4. TRATAMIENTO DE DATOS
3. EXPLORACIÓN DEL DATO EDA
4. MODELO

El modelo con una mejor optimización del desvio se almacena en esta ruta. Será el *MODELO\_GRADIENT\_BOOSTING\_IMPUTACIÓN\_f\_RUN.ipynb*

* 1. MODELOS PRUEBA

Será el repositorio de los diferentes modelos no óptimos que se han descartado en el presente proyecto tanto por no ofrecer una mejora o por ser modelos realizados sin resultado significativo.

1. ARCHIVOS

Repositorio de archivos de entrenamiento de modelos.

1. VIDEO

Repositorio de video de presentación.

Google Drive:

Se utiliza la herramienta Google Drive como repositorio colaborativo. Durante la elaboración del proyecto se suben documentos, proyectos, archivos de datos para consulta de los integrantes del proyecto. Tras la homogeneización de los archivos, se replica la estructura en github para acceso a la evaluación del proyecto.

Google Meet:

Durante la elaboración del proyecto se realiza un seguimiento periodico cada 15 dias por parte de los integrantes del proyecto, utilizando Google Meet para programar dichas reuniones de seguimiento.

Youtube:

Video de presentación del proyecto alojado en el siguiente enlace como video oculto.

<https://youtu.be/PRHlmshjmqM>