IEBS Business School

Análisis predictivo de fallos en inversores fotovoltaicos utilizando técnicas de Machine Learning semi supervisado

---

VERSIÓN RESUMIDA

01/10/2024

1. RESUMEN

Este trabajo aborda el problema de la detección y prevención de fallos en inversores fotovoltaicos mediante el análisis predictivo utilizando técnicas de machine learning semi-supervisado. Los inversores fotovoltaicos, elementos esenciales para la conversión de energía solar, pueden presentar fallos que disminuyen su eficiencia y disponibilidad, lo que impacta negativamente en la producción energética. El objetivo principal es predecir fallos y anomalías en estos inversores utilizando datos operativos y meteorológicos.

La solución planteada consiste en desarrollar un modelo de predicción de fallos basado en redes neuronales LSTM y técnicas de detección de anomalías como Isolation Forest. Se procesaron más de 2 millones de datos operativos de cinco inversores ubicados en una planta fotovoltaica, integrando además información meteorológica obtenida mediante la API Open-Meteo. Las técnicas de reducción de dimensionalidad y normalización permitieron optimizar el rendimiento de los modelos.

Esta solución es adecuada porque, a diferencia de enfoques convencionales, logra predecir fallos de manera anticipada, lo que permite tomar medidas preventivas y reducir el tiempo de inactividad. Además, el uso de modelos basados en series temporales permite capturar patrones más complejos y específicos, mejorando la precisión de las predicciones.

Los resultados obtenidos demuestran que el modelo LSTM predijo anomalías con una precisión considerable, aunque existe margen de mejora en términos de anticipación temporal. Este trabajo sienta las bases para futuras investigaciones más detalladas en la implementación de estos modelos en entornos reales y su optimización para mejorar aún más la eficiencia de las plantas fotovoltaicas.

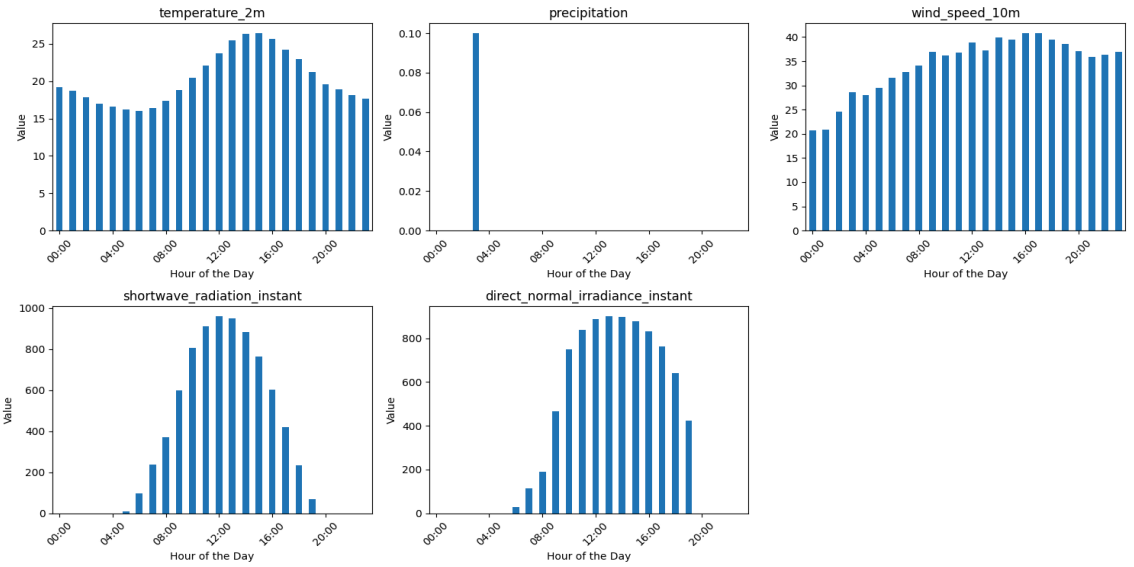
2. EJECUCIÓN DE LA SOLUCIÓN

2.1 Creación y unión de los datasets

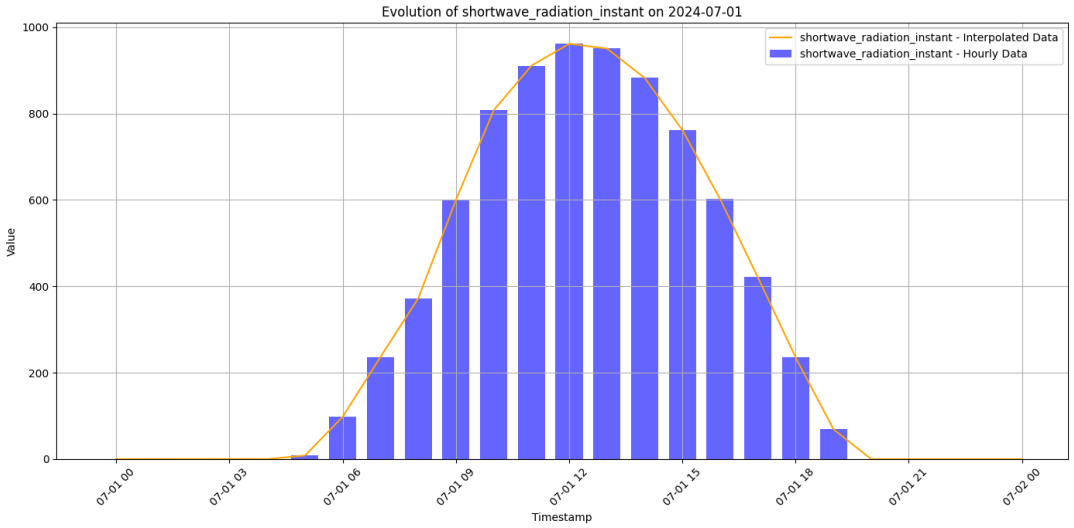
2.1.1 Carga del dataset de la API Open-Meteo

Los datos meteorológicos, recolectados mediante la API de Open-Meteo, incluyeron datos sobre la meteorología durante los meses de Julio y Agosto en la ubicación exacta de la instalación fotovoltaica. Para nuestro caso de estudio son especialmente interesantes los valores de las siguientes métricas:

- **Shortwave Solar Radiation GHI:** Es la medida directa de la radiación solar que llega a la superficie, crucial para la generación de energía solar.
- **Direct Normal Irradiance DNI:** Indica la radiación solar directa, que es fundamental para el rendimiento óptimo de los paneles solares.
- **Temperature:** Es muy relevante ya que la temperatura del aire afecta la eficiencia del sistema solar.
- **Precipitation (rain + showers + snow):** La precipitación puede afectar la acumulación y el rendimiento de los paneles solares, especialmente en términos de limpieza y acumulación.
- **Wind Speed (10 m):** La velocidad del viento puede influir en la acumulación de polvo y en la eficiencia de los paneles solares.



Cómo se puede observar en la figura anterior, Open-Meteo registra las métricas meteorológicas en intervalos horarios. Para nuestro proyecto se requiere de datos por intervalo de segundos así que se realiza una interpolación lineal de los datos horarios obtenidos de la API.



2.1.2 Unión de los datasets

En esta fase se combinan los dataframes de los inversores y los datos meteorológicos en base a la columna Timestamp. Se utiliza el método 'merge' con el atributo 'left' para mantener los timestamps de las observaciones del dataset de los inversores y descartar el resto de timestamps del dataset de Open-Meteo.

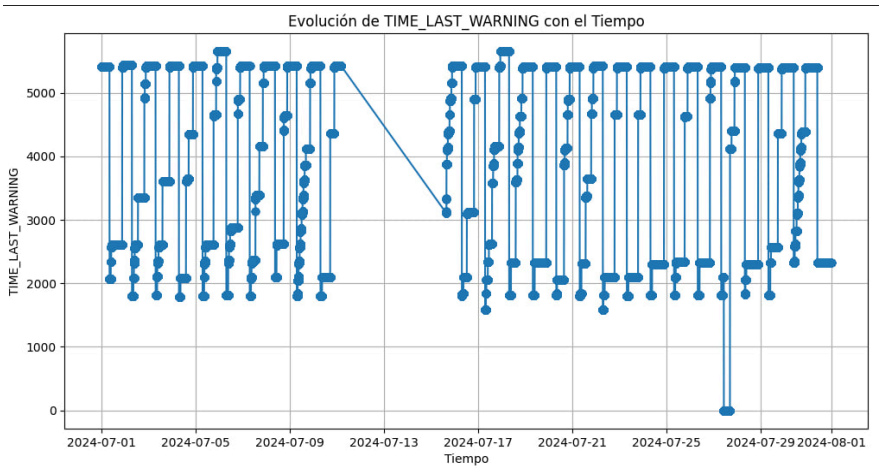
2.2 Preprocesamiento y limpieza de los datos

2.2.1 Eliminación de columnas inconsistentes

Gracias a la exploración inicial realizada se observa que el dataset contiene ciertas variables con valores inconsistentes. El primer grupo de estas variables son las que tienen un sólo valor constante a lo largo de toda su columna lo cual confirma su eliminación debido a que no aportan información relevante para nuestro objetivo.

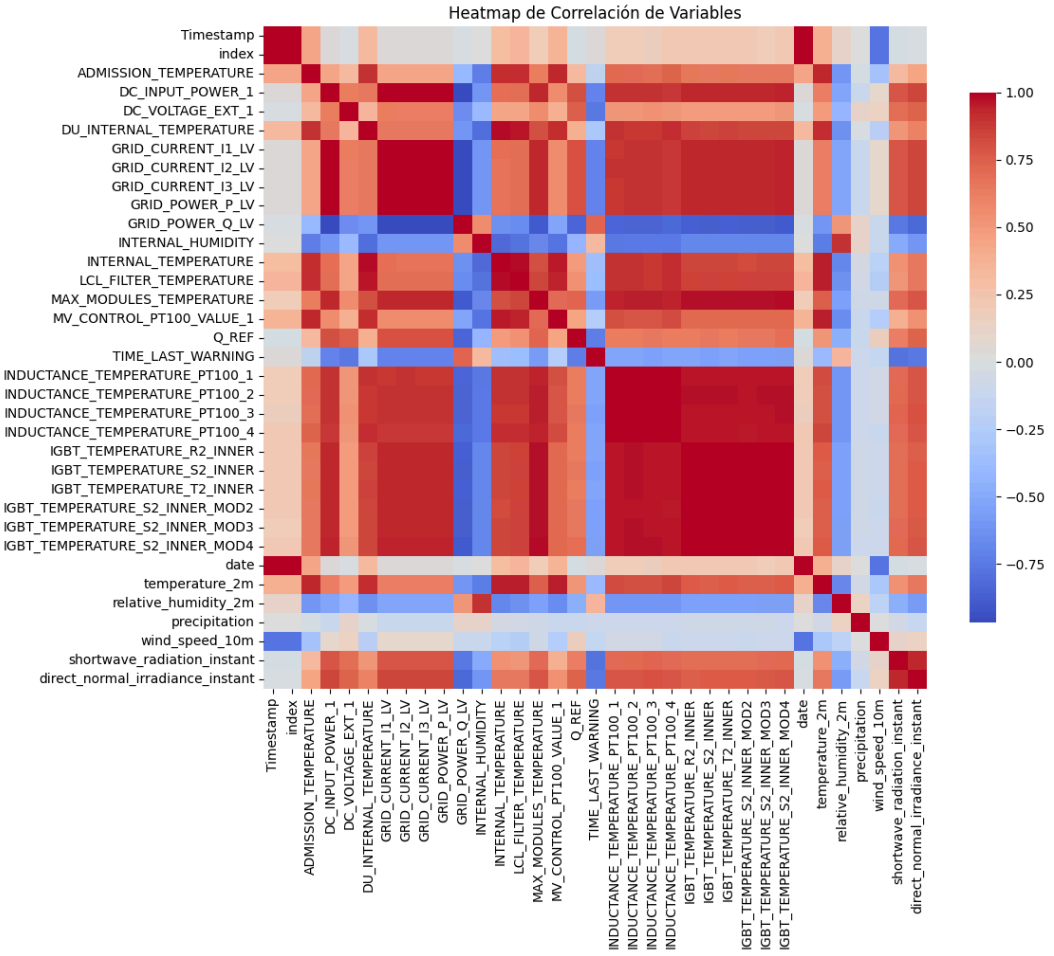
DataFrame	Columnas con Valores Únicos
df1_7	[CONTROL_MODE_Q, COS_PHI_REF_SIGNED, DIAGNOSIS_WATCHDOG_LAST_UP_MASK, EFICIENCY, IQ_REF, LIMIT_Q, PAC_REF]

Asimismo, se eliminan las variables 'LAST_WARNING' y 'DATE_LAST_WARNING' debido a las inconsistencias detectadas en sus datos. La variable 'LAST_WARNING' solo indica el tipo del último aviso registrado, lo que limita su utilidad para el análisis. Por otro lado, 'DATE_LAST_WARNING' presenta patrones temporales ilógicos, lo que sugiere posibles errores en la captura o registro de la información.



2.2.2 Correlación de variables

Tras una limpieza de las variables claramente descartables, el siguiente paso es profundizar en cómo interactúan entre sí. Para esto, resulta esencial examinar las correlaciones entre las diferentes variables del dataset, lo que nos permitirá identificar patrones de dependencia o redundancia que podrían influir en la detección de fallos. Este análisis es crucial para optimizar la selección de variables y mejorar la precisión de los modelos predictivos.



Se puede observar que hay una gran diversidad de comportamientos respecto a la correlación entre las variables del conjunto de datos. Algunas variables presentan una correlación muy baja con respecto a otras en el dataset, lo cual sugiere que su impacto sobre las métricas analizadas podría ser limitado. Paralelamente hay ciertos grupos de variables que tienen una correlación muy elevada, lo cual indica que hay cierta redundancia entre ellas.

Con el fin de optimizar la reducción de la dimensionalidad del conjunto de datos y mejorar el rendimiento de los modelos de detección de anomalías, se tomaron las siguientes medidas:

- **Eliminación de variables con correlación despreciable**: Se identificaron variables que no mostraban una correlación significativa con las demás variables del conjunto de datos. Estas variables se consideran irrelevantes para el análisis y se eliminaron del dataset. Las variables eliminadas por este criterio son:
    - *GRID_VOLT_RS_LV*
    - *GRID_VOLT_ST_LV*
    - *GRID_VOLT_TR_LV*
    - *LIMIT_P_DISCHARGE*
    - *MV_CONTROL_PT100_VALUE_2*
    - *MV_CONTROL_PT100_VALUE_3*

- **Eliminación de variables meteorológicas con baja correlación**: Las variables meteorológicas añadidas al dataset también fueron evaluadas. Se observó que algunas de ellas no tenían una correlación suficiente con el resto de las variables del conjunto de datos, lo que sugiere que su impacto en el análisis de anomalías sería mínimo. Por este motivo, se decidió descartarlas. Las variables meteorológicas eliminadas son:
    - *wind_speed_10m*
    - *precipitation*

2.3 Reducción de la dimensionalidad:

- **Normalización de los datos**
  Previamente a aplicar los métodos de reducción de la dimensionalidad o a implementar algoritmos de aprendizaje no supervisado se requiere una normalización de los datos en caso de disponer de datos con diferentes escalas y unidades. El uso de **StandardScaler** asegura que todas las variables tengan una media de 0 y una desviación estándar de 1.

- **Aplicación de técnica de reducción de la dimensionalidad a los grupos de variables altamente correlacionadas**
  Como se observa en el mapa de correlaciones de la Figura anterior, hay tres grupos de variables que muestran alta correlación interna dentro de cada grupo. Estas variables son agrupadas y procesadas mediante **PCA** (Análisis de Componentes Principales) para reducir su dimensionalidad de manera que se preserve la mayor cantidad posible de varianza, y se combine la información en componentes principales que represente cada grupo.

  Los grupos son los siguientes:

  I. **Grupo de temperaturas y sensores de inductancia**:
    - *MAX_MODULES_TEMPERATURE*
    - *INDUCTANCE_TEMPERATURE_PT100_1*
    - *INDUCTANCE_TEMPERATURE_PT100_2*
    - *INDUCTANCE_TEMPERATURE_PT100_3*

- *INDUCTANCE_TEMPERATURE_PT100_4*
- *IGBT_TEMPERATURE_S2_INNER*
- *IGBT_TEMPERATURE_S2_INNER_MOD2*
- *IGBT_TEMPERATURE_S2_INNER_MOD3*
- *IGBT_TEMPERATURE_S2_INNER_MOD4*

La varianza explicada por el componente principal de este grupo es **97.88%**.

II. **Grupo de corriente y potencia**:
- *GRID_CURRENT_I1_LV*
- *GRID_CURRENT_I2_LV*
- *GRID_CURRENT_I3_LV*
- *GRID_POWER_P_LV*
- *GRID_POWER_Q_LV*

La varianza explicada por el componente principal de este grupo es **99.03%**.

III. **Grupo de temperatura interna y humedad**:
- *DU_INTERNAL_TEMPERATURE*
- *INTERNAL_HUMIDITY*
- *INTERNAL_TEMPERATURE*
- *LCL_FILTER_TEMPERATURE*

La varianza explicada por el componente principal de este grupo es **91.97%**

- **Aplicación de PCA a todo el dataset:** Finalmente se aplica PCA a la totalidad del dataset para obtener 2 dimensiones y poder visualizar la distribución de los puntos la aplicación de PCA a cada grupo, los componentes principales obtenidos se incorporarán al conjunto de datos. Esto reducirá la cantidad de variables redundantes, lo que optimizará tanto el análisis de anomalías como la eficiencia computacional.

La varianza explicada por el componente principal al aplicar PCA a todo el dataset es **91.42%**.

2.4 Detección de anomalías con IsoForest

Tras completar el preprocesamiento del dataset, se procede a la fase de creación de modelos de detección de anomalías. El objetivo principal de esta fase es identificar patrones inusuales en los datos que puedan indicar situaciones anómalas. Para ello, se generarán etiquetas para cada observación del dataset: **0** para las situaciones consideradas normales y **1** para aquellas consideradas anómalas.

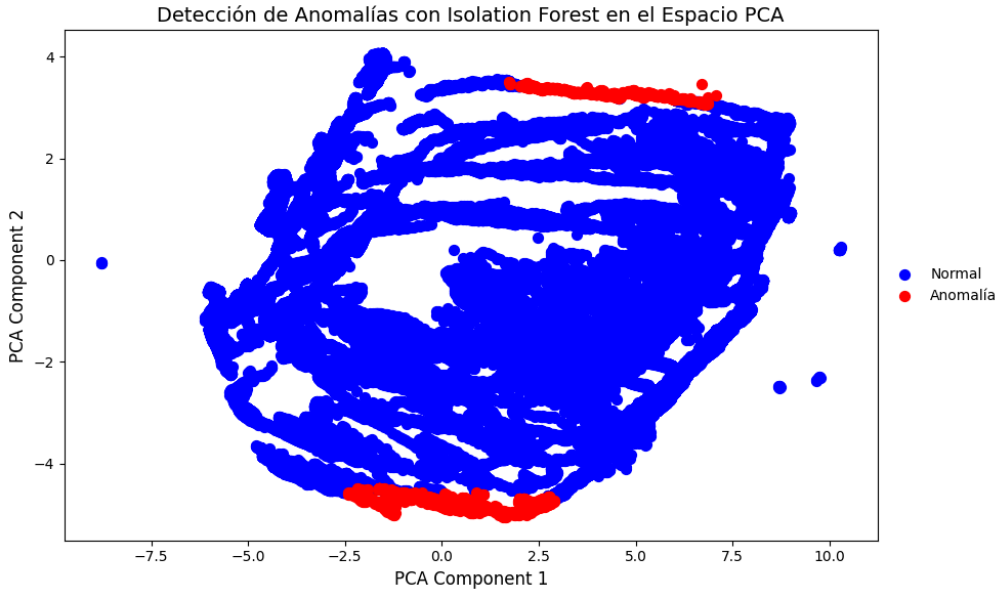
Se pretende experimentar con un algoritmo de aprendizaje no supervisado, Isolation Forest, para identificar posibles anomalías en los datos. Las fases de esta etapa son las siguientes:

- **Normalización de datos:** Antes de aplicar estos algoritmos, se normalizan los datos para garantizar que todas las variables contribuyan de manera equilibrada al proceso de detección de anomalías. Concretamente, se usa **StandardScaler** para escalar las características del dataset, lo que asegura que cada variable tenga una media de 0 y una desviación estándar de 1. Esto es especialmente importante cuando las variables originales tienen diferentes unidades o escalas, ya que permite que cada una tenga el mismo peso en los algoritmos de detección.

- **Reducción de dimensionalidad:** Para facilitar el procesamiento, la visualización y evaluación de los resultados, se aplica reducción de dimensionalidad a todo el dataset mediante técnicas como PCA (Análisis de Componentes Principales). Esto permite proyectar los datos en un espacio de menor dimensión, donde las anomalías pueden ser visualizadas y analizadas de forma más clara.

- **Implementación de IsoForest:** Con los datos normalizados se implementa el algoritmo de **IsoForest**. Este modelo se entrena en el dataset completo, y cada observación es evaluada para determinar si es una posible anomalía. Los resultados se almacenan en forma de etiquetas que indican la condición de cada observación (normal o anómala).

- **Visualización de Anomalías en las Componentes de PCA**: Para verificar si el modelo de Isolation Forest ha captado correctamente las anomalías, se visualiza un gráfico de dispersión de las primeras dos componentes de PCA. Las anomalías se resaltan de forma distintiva para observar si están separadas del resto de los datos.



Tras implementar el algoritmo **Isolation Forest** combinado con una reducción de dimensionalidad mediante **PCA**, se observa que los resultados obtenidos no reflejan una clara diferenciación entre los puntos normales y las anomalías. En lugar de identificar patrones anómalos específicos, el modelo tiende a marcar como anomalías los puntos que se encuentran en los extremos de la nube de datos, sin una justificación sólida en términos de la naturaleza del problema. Esto sugiere que el modelo está capturando únicamente outliers por posición, más que verdaderas anomalías significativas.

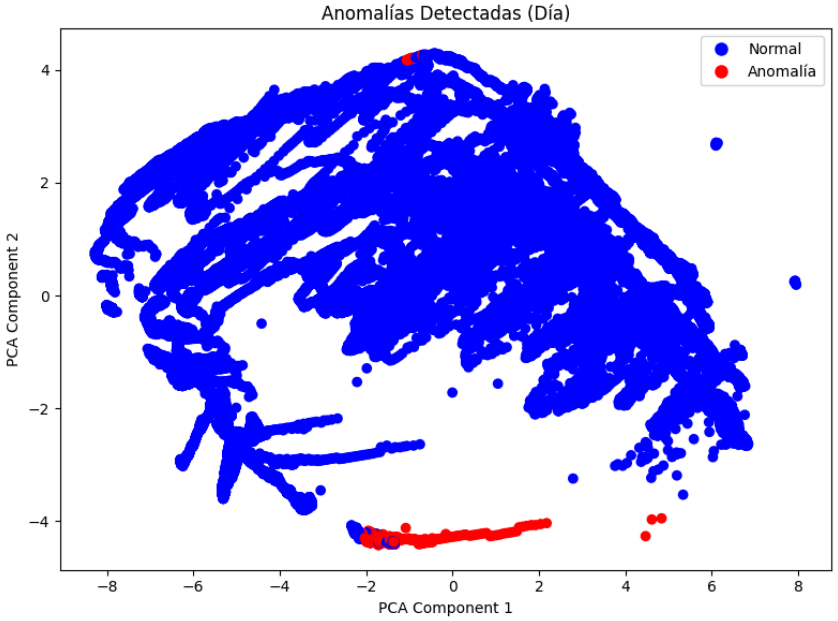
Esta observación nos lleva a reflexionar sobre la **naturaleza intrínseca de los datos**, en los que es evidente que existen patrones temporales significativos, como la **dependencia del momento del día** en el que se toman las mediciones. Por ejemplo:

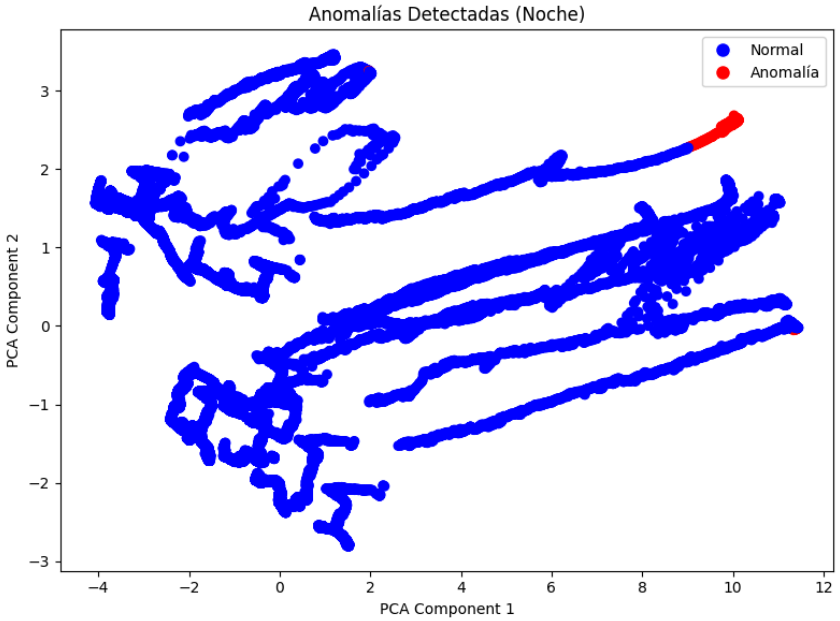
- **De día**, valores cercanos a **0** en la columna de potencia de la red ("grid power") podrían ser indicativos de una anomalía, ya que en condiciones normales debería haber actividad.
- **De noche**, esos mismos valores no son necesariamente anómalos, ya que es esperado que la potencia se reduzca o incluso caiga a cero.

El algoritmo en su estado actual no es capaz de capturar estas dependencias **temporales** y **contextuales**, lo que genera falsos positivos, es decir, marca como anomalías valores que son normales en ciertos momentos del día.

2.5 División Temporal en Día y Noche

Para mejorar la precisión del modelo en la detección de anomalías, se propone dividir el conjunto de datos en dos subconjuntos basados en el momento del día: los datos diurnos y los datos nocturnos. Los datos diurnos abarcan las horas del día en las que se espera una actividad significativa en variables clave, como la potencia de la red, donde los valores cercanos a cero o inactivos pueden ser indicativos de anomalías. Por otro lado, los datos nocturnos suelen presentar valores bajos o nulos, que son esperables y no deben ser considerados como anomalías, lo que permite al modelo centrarse en identificar patrones anómalos más relevantes en este contexto. Esta división temporal ayudará a reducir falsos positivos y a mejorar la capacidad del modelo para detectar anomalías verdaderas.





Se puede observar que el modelo parece mejorar un poco, especialmente en los valores diurnos, ya que se puede diferenciar un cluster a la parte inferior del gráfico que podría pertenecer a un grupo de anomalías similares. Por otra parte, en la visualización de los datos nocturnos se puede observar que se continúan identificando cómo anómalos los valores extremos de un patrón contínuo, lo cual no parece muy adecuado.

Por ejemplo, en los inversores fotovoltaicos, los valores de producción de energía durante el día pueden variar considerablemente en comparación con la noche debido a la disponibilidad de luz solar. Un valor elevado de corriente o voltaje puede ser completamente **normal** en horas diurnas cuando los inversores están operando a plena capacidad. Sin embargo, ese mismo valor observado durante la noche, cuando no hay producción solar, debería ser considerado **anómalo**. Dado que IsoForest no tiene en cuenta este tipo de variaciones temporales, clasifica los valores de manera aislada, sin diferenciar entre contextos temporales.

La falta de integración de la información temporal en IsoForest resulta en **falsos positivos** o **falsos negativos**, ya que el algoritmo trata los datos de forma independiente de su contexto. Por esta razón, en conjuntos de datos donde el **comportamiento normal varía con el tiempo**, como es el caso en aplicaciones industriales o energéticas, **los métodos de detección de anomalías basados en series temporales o modelos secuenciales** como LSTM (Long Short-Term Memory) pueden ser más efectivos al capturar y analizar la evolución de las tendencias a lo largo del tiempo.

2.6 Detección de anomalías con LSTM

En el contexto de la detección de anomalías, las técnicas tradicionales, como el algoritmo IsoForest, han demostrado limitaciones, especialmente en conjuntos de datos que presentan una fuerte dependencia temporal. Estas limitaciones se evidencian en la tendencia del modelo a clasificar outliers como anomalías, sin considerar el comportamiento subyacente de los datos a lo largo del tiempo. Para abordar este desafío, se propone la implementación de redes neuronales LSTM (Long Short-Term Memory), las cuales son especialmente adecuadas para el análisis de series temporales.

Las LSTM son capaces de capturar patrones temporales complejos y relaciones a largo plazo en los datos, lo que permite una identificación más precisa de anomalías en función del contexto temporal. Esta metodología no solo mejora la capacidad del modelo para diferenciar entre comportamientos normales y anómalos, sino que también considera variaciones en las observaciones a diferentes horas del día. A través de esta aproximación, se espera lograr una detección más robusta y confiable de anomalías, mejorando así la calidad del análisis en el conjunto de datos en estudio.

2.6.1 Implementación de LSMT para detección de anomalías

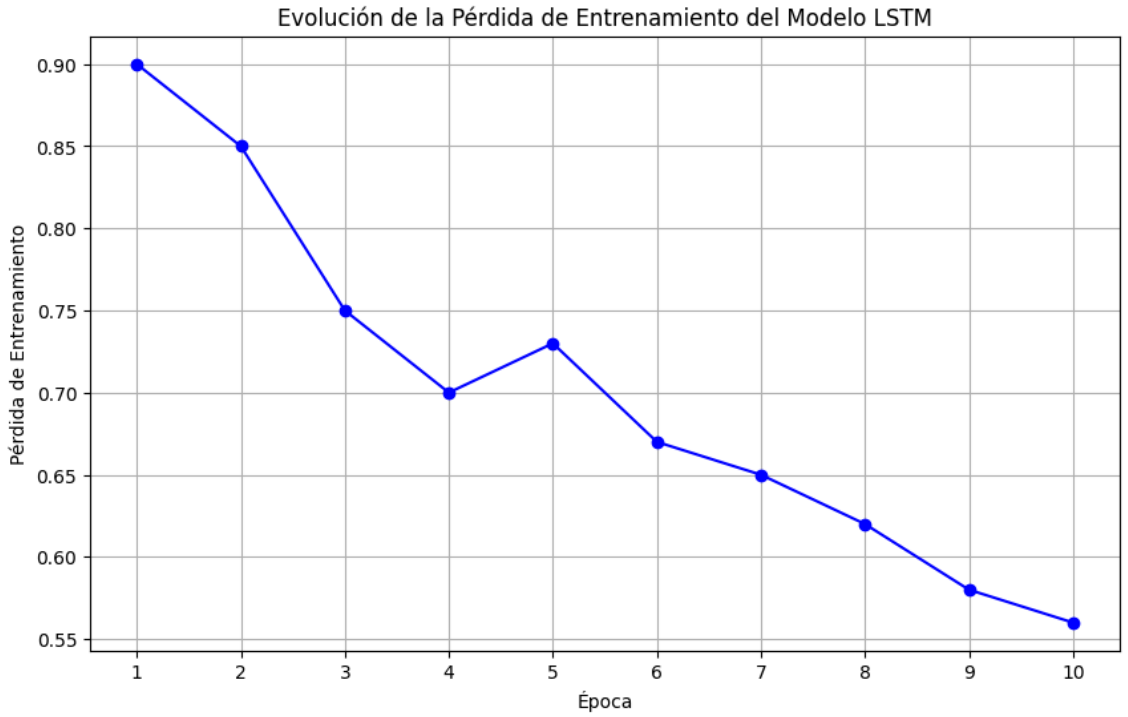
Se utilizan los datos anteriormente procesados con el método de reducción de la dimensionalidad PCA. Posteriormente, se crean ventanas de tiempo que permiten estructurar los datos de forma adecuada para el modelo LSTM.

Los datos son divididos en dos subconjuntos: uno para entrenamiento y otro para prueba, utilizando un 80% de los datos para el entrenamiento y el 20% restante para la prueba. Esta separación se realiza considerando las ventanas de tiempo, lo que permite evaluar la capacidad de generalización del modelo de manera efectiva, simulando un escenario real donde el modelo predice sobre datos futuros a partir de patrones aprendidos en datos pasados.

La arquitectura del modelo incluye capas LSTM para capturar patrones en la secuencia de datos, acompañadas de capas de Dropout para prevenir el sobreajuste. Finalmente, el modelo se compila y se entrena utilizando una función de pérdida adecuada, lo que le permite aprender de los datos de entrada y mejorar la precisión en la detección de anomalías.

2.6.2 Gráfica de pérdida

Para optimizar el rendimiento del modelo LSTM, es crucial ajustar el número de épocas de entrenamiento, y una herramienta fundamental para este propósito es la gráfica de pérdida. Esta gráfica permite visualizar cómo evoluciona la pérdida del modelo a lo largo de las épocas, proporcionando información valiosa sobre su proceso de aprendizaje.

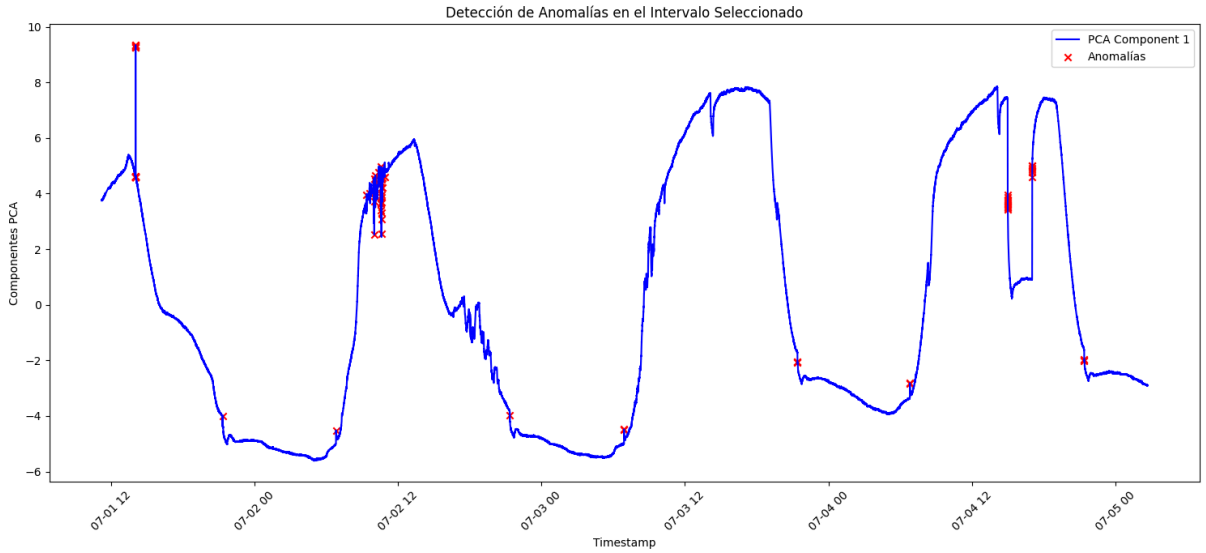


La gráfica presenta la evolución de la **pérdida de entrenamiento** a lo largo de las **épocas** del modelo LSTM utilizado para la generación de etiquetas de anomalías.

1. **Tendencia General a la Baja:** Se observa una tendencia general a la disminución de la pérdida de entrenamiento, lo que indica que el modelo está aprendiendo de manera efectiva a capturar las dinámicas de los datos. Este comportamiento es esperado, ya que una disminución en la pérdida sugiere que el modelo mejora su capacidad para ajustar las predicciones a los datos de entrada.
2. **Comportamiento en Épocas Específicas:** Aunque la pérdida de entrenamiento disminuye en general, se notan ciertos picos en la época 5 donde la pérdida presenta un ligero aumento. Esto podría sugerir una variabilidad en el aprendizaje del modelo, lo que es normal en entrenamientos de redes neuronales. Sin embargo, se observa que la pérdida regresa a niveles más bajos en las épocas siguientes, lo que indica que el modelo continúa refinando su aprendizaje.
3. **Pérdida Estable en las Épocas Finales:** A partir de la época 6, la pérdida de entrenamiento se estabiliza y muestra una leve tendencia a la baja, lo que sugiere que el modelo ha encontrado un equilibrio en su aprendizaje. Este comportamiento es prometedor, ya que indica que el modelo ha logrado aprender de los patrones en los datos, lo que es crucial para la posterior generación de etiquetas de anomalías.

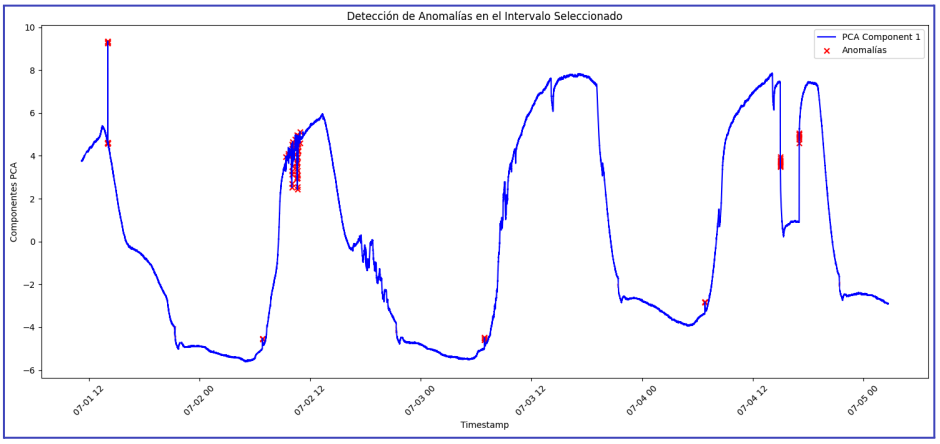
5.8.3 Visualización de la detección de anomalías

Para evaluar la efectividad del modelo LSTM en la detección de anomalías, se visualizan los datos en un intervalo específico donde se ha observado un número notable de anomalías. Esta visualización permite observar cómo el modelo responde ante situaciones anómalas en los datos reales.



Se puede observar que el modelo responde cómo se espera. En el momento que se dan patrones anómalos en la primera componente principal (PCA1) se etiqueta una anomalía. Pese a un rendimiento satisfactorio se puede identificar que en ciertos puntos de cambio de tendencia el modelo identifica la muestra cómo anomalías, lo cual se debería ajustar.

Se incrementa el umbral de detección de anomalías de 2 a 3 desviaciones estándar para reducir la cantidad de falsos positivos y pulir el error de clasificar cómo anomalías ciertos cambios de tendencia que aparentemente son normales. Visualizamos el comportamiento del modelo tras este ajuste, lo cual puede conllevar el riesgo de pasar por alto algunas anomalías menos pronunciadas:



Al aumentar el umbral para la designación de anomalías, aquellos puntos dudosos que anteriormente podían haberse clasificado como anómalos debido a fluctuaciones leves o cambios de tendencia en comportamientos aparentemente normales, ahora se consideran dentro del rango aceptable. Este ajuste permite que el modelo se enfoque en detectar anomalías más significativas, evitando falsos positivos en situaciones donde las variaciones pueden ser parte del comportamiento natural del sistema.

Este resultado es satisfactorio para los objetivos actuales del proyecto, que busca establecer una base sólida para la detección de anomalías en inversores fotovoltaicos. A pesar de que se espera seguir perfeccionando el modelo en el futuro, este enfoque inicial ofrece una plataforma robusta para continuar aprendiendo y mejorando el rendimiento en próximas fases, donde se podrán afinar aún más los umbrales y técnicas de predicción para obtener resultados más precisos y adaptados a las necesidades operativas del sistema.

2.7 Predicción de anomalías con LSTM

En esta fase final de la ejecución de la solución, nos enfocaremos en la predicción de anomalías utilizando la arquitectura de redes neuronales LSTM. Al implementar el modelo LSTM, utilizamos los datos previamente procesados y etiquetados, permitiendo que la red aprenda las dinámicas subyacentes de los sistemas de inversión de energía. A través de este proceso, buscamos no solo identificar las anomalías presentes en los datos, sino también entender los factores que influyen más en la aparición de anomalías e anticipar futuras anomalías, mejorando así la capacidad de respuesta y la eficiencia operativa de la planta fotovoltaica.

2.7.1 Implementación del modelo

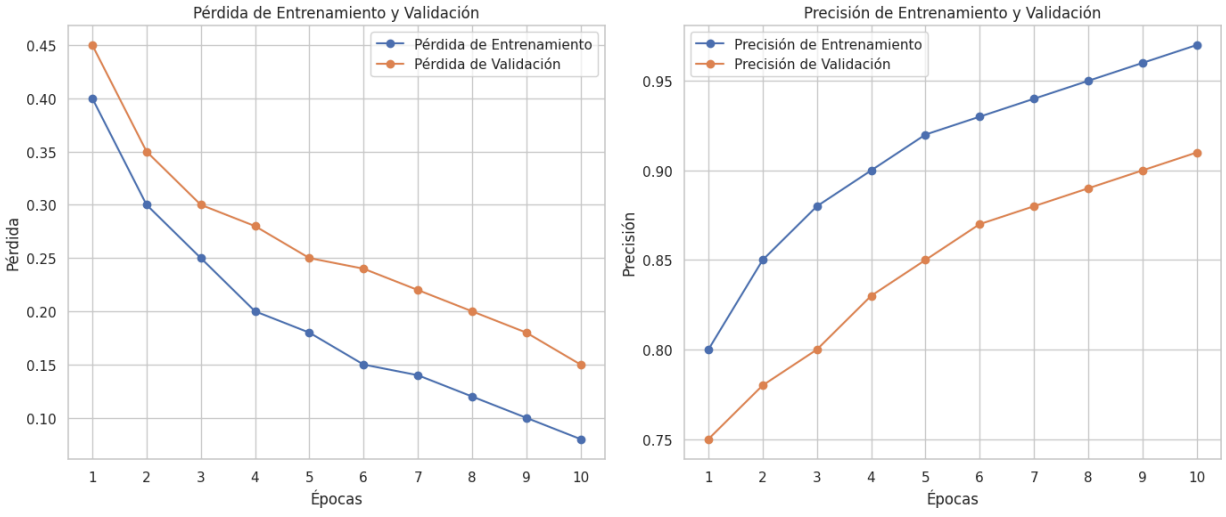
Se utilizan los datos procesados en apartados anteriores, en particular, aquellos en los que se han eliminado las variables con baja correlación y se han agrupado las variables con alta correlación mediante Análisis de Componentes Principales (PCA). Este enfoque permite al modelo captar de manera más efectiva los patrones presentes en los datos, evitando la sobrecarga de variables y optimizando su rendimiento.

Una vez que los datos han sido procesados y optimizados, se procede al entrenamiento del modelo LSTM. Este modelo, basado en redes neuronales recurrentes, es particularmente adecuado para el análisis de series temporales, ya que puede aprender patrones secuenciales en los datos. Durante el entrenamiento, se utilizan las etiquetas generadas previamente, lo que permite al modelo aprender a identificar y clasificar anomalías en el comportamiento de los inversores fotovoltaicos.

El modelo se entrena utilizando un conjunto de datos de entrenamiento, donde las pérdidas se monitorean a lo largo de múltiples épocas. La función de pérdida utilizada permite evaluar la efectividad del modelo en la predicción de las etiquetas, y se aplica un optimizador para minimizar esta pérdida.

2.7.2 Evaluación del modelo

El rendimiento del modelo se evalúa en función de su capacidad para generalizar a un conjunto de validación, que se mantiene separado del conjunto de entrenamiento para evitar sobreajuste.



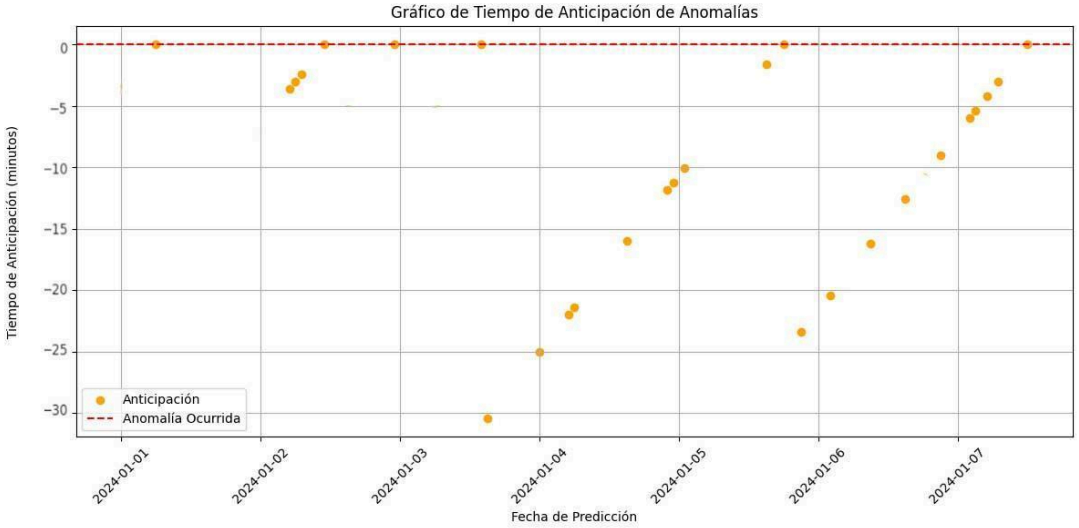
Los resultados del entrenamiento que se visualizan mediante gráficas anteriores, muestran la pérdida de entrenamiento y validación, así como la precisión alcanzada a lo largo de las épocas. Estas gráficas indican un rendimiento óptimo del modelo, con una disminución constante en la pérdida y un aumento en la precisión.

- Pérdida de Entrenamiento y Validación: La gráfica muestra que la pérdida de entrenamiento disminuye de manera constante, alcanzando valores mínimos, lo que sugiere un buen ajuste a los datos. Por otro lado, la pérdida de validación también se mantiene en un rango bajo, indicando una buena capacidad de generalización.
- Precisión de Entrenamiento y Validación: La precisión de entrenamiento inicia en un 80% y se incrementa hasta un 97%, lo que refleja una adecuada adaptación del modelo a los datos. La precisión de validación también muestra un aumento considerable, comenzando en un 75% y alcanzando un 91%, lo que respalda la efectividad del modelo en el contexto de datos no vistos.

2.7.3 Análisis de la capacidad predictiva del modelo LSTM para la detección temprana de anomalías

El principal objetivo de este proyecto, además de realizar un análisis exploratorio y de ganar entendimiento de cómo influyen las diferentes métricas a la aparición de anomalías, es generar las bases y la comprensión del flujo de trabajo para llegar a crear un modelo con la capacidad de predecir anomalías de forma precoz. A través de la implementación de un enfoque de ventana deslizante, se generan conjuntos de datos que permiten al modelo aprender de las condiciones previas a los eventos anómalos.

Se visualizan los resultados en un gráfico de tiempo de anticipación de las anomalías detectadas para evaluar la capacidad de predicción



En el gráfico anterior se puede observar la capacidad del modelo de predecir anomalías con antelación. Los puntos naranjas representan anomalías detectadas en una ventana temporal y su posición en el eje horizontal indica el momento real de ocurrencia de dicha anomalía. La

distancia vertical de cada punto naranja desde la línea horizontal que marca el tiempo de ocurrencia real representa el tiempo de anticipación que el modelo logró.

El gráfico refleja la capacidad limitada del modelo LSTM para predecir anomalías con anticipación ya que el modelo puede detectar anomalías con un máximo de 30 minutos de antelación. A pesar de que el modelo puede identificar eventos anómalos, la escasa distancia entre los puntos naranjas y la línea de ocurrencia real indica que las predicciones se realizan solo unos minutos antes de que las anomalías se manifiesten. Esto sugiere que, aunque el modelo está funcionando, su rendimiento en términos de anticipación temporal es aún insuficiente. Para mejorar esta capacidad, sería necesario explorar ajustes en la arquitectura del modelo, la selección de características, y quizás la incorporación de datos adicionales que puedan proporcionar información más relevante para la detección temprana de anomalías.

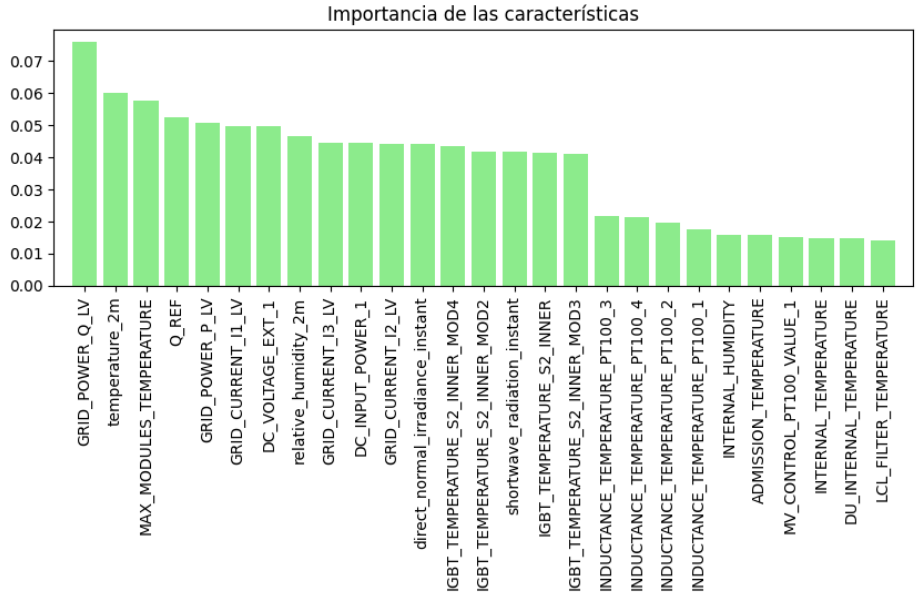
3. RESULTADOS

En el presente apartado se presentan los resultados obtenidos a partir del modelo de predicción de anomalías entrenado. Se destacan tres aspectos clave: el análisis de la importancia de las variables, el diagrama de árbol de decisión y un análisis de series temporales de anomalías. Cada uno de estos elementos proporciona una perspectiva valiosa sobre el funcionamiento del modelo y permite extraer conclusiones significativas que pueden guiar la operación y el mantenimiento de los inversores fotovoltaicos.

3.1 Importancia de las variables al detectar anomalías

El análisis de la importancia de las variables es fundamental para comprender cuáles características son determinantes en la detección de anomalías. Al identificar las variables más influyentes, se pueden realizar ajustes operativos y configuraciones del sistema de manera más efectiva, lo que contribuye a mejorar el rendimiento del modelo y a optimizar el monitoreo de los inversores fotovoltaicos. Esta comprensión no solo ayuda a priorizar recursos, sino que también proporciona información valiosa para la implementación de estrategias más robustas.

Para llevar a cabo este análisis, se utiliza Random Forest, aprovechando las etiquetas generadas anteriormente por el modelo LSTM no supervisado. Aunque Random Forest tiene limitaciones al no estar diseñado específicamente para series temporales, su técnica de **feature_importances** permite evaluar el impacto de cada variable en el rendimiento del modelo. Esta evaluación facilita la identificación de las características que requieren mayor atención, ofreciendo así una base sólida para la toma de decisiones informadas en la gestión de anomalías.



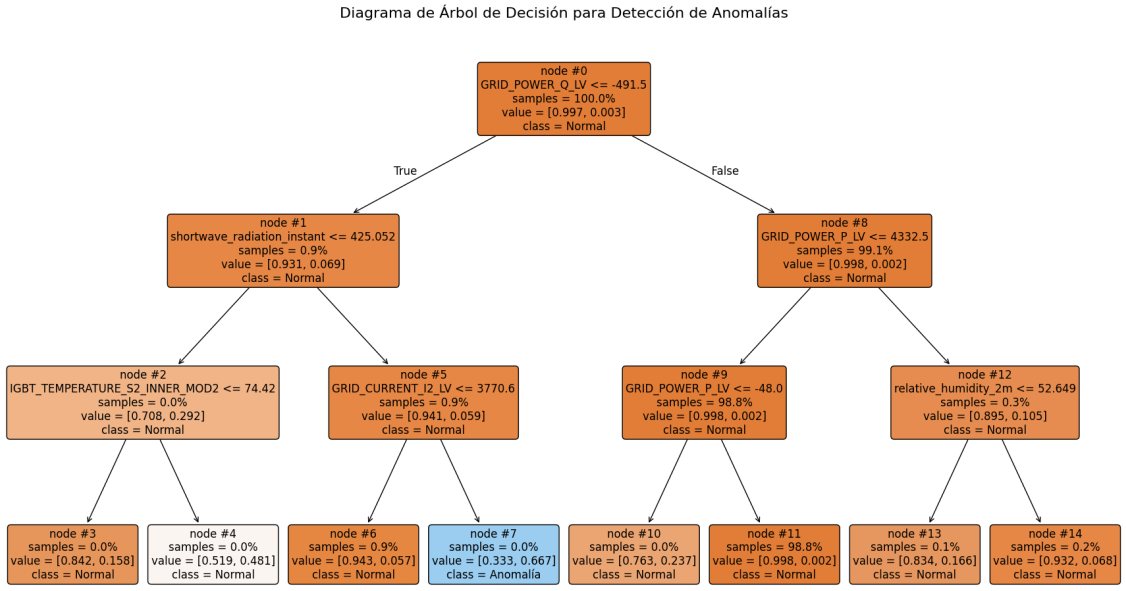
Tal y como se puede observar en la gráfica anterior, *GRID_POWER_Q_LV* es la variable, con una significativa diferencia, más importante en la detección de anomalías. Esto indica que las fluctuaciones en la potencia de red es muy relevante al clasificar datos entre anómalos o normales.

Otras variables importantes son *temperature_2m* y *MAX_MODULES_TEMPERATURE*, sugiriendo que la relación entre la temperatura ambiental y la temperatura de los módulos y el funcionamiento del sistema es especialmente relevante, ya que las altas temperaturas pueden afectar la eficiencia de conversión y aumentar la probabilidad de fallos.

La variable *Q_ref* también se destaca como una de las más importantes en el análisis de la detección de anomalías, lo cual indica que algunas anomalías en la producción pueden manifestarse a través de variaciones en esta variable, la cual registra la producción esperada del sistema fotovoltaico. Esto implica que si hay discrepancias significativas entre la producción real y la *Q_ref*, es probable que se estén presentando fallos o malfuncionamientos en los componentes del sistema. Por lo tanto, el monitoreo constante de *Q_ref* es crucial para identificar desviaciones que podrían indicar problemas en la generación de energía.

3.2 Diagrama de árbol de decisión

El diagrama de árbol de decisión ofrece una representación visual de las reglas de clasificación empleadas por el modelo Isolation Forest en la detección de anomalías. Esta visualización permite desglosar el proceso de toma de decisiones del modelo, mostrando cómo las diferentes variables interactúan para clasificar los datos. Comprender las reglas subyacentes a la clasificación ayuda a interpretar mejor el comportamiento del modelo y proporciona un marco para evaluar las decisiones operativas basadas en estas clasificaciones. Con la finalidad de una interpretabilidad más precisa se usan los datos previos a su normalización.



El análisis del árbol de decisión revela que tal y cómo con la importancia de variables a partir del método **feature_importances** la variables con más impacto es **GRID_POWER_Q_LV**, reafirmando que las fluctuaciones en la potencia de red son especialmente relevantes al clasificar datos entre anómalos o normales.

Otras variables que destacan por su importancia en el diagrama del árbol de decisión son **INDUCTANCE_TEMPERATURE_PT100_4**, **DU_INTERNAL_TEMPERATURE** y **IGBT_TEMPERATURE_S2_INNER**, las cuales en el gráfico de importancia anterior del método **feature_importances** no parecían tener un impacto tan significativo al identificar anomalías.

Esto sugiere que, aunque estos atributos no son las variables más influyentes según la técnica de **feature_importance**, su rol en el árbol de decisión es notable en ciertas interacciones y divisiones del modelo. La discrepancia entre ambos análisis puede indicar que su efecto no sea evidente de forma aislada, pero en combinación con otras variables, su presencia se vuelve crítica para la detección efectiva de anomalías. Esto resalta la importancia de utilizar múltiples métodos de evaluación para obtener una comprensión completa de cómo las variables contribuyen al rendimiento del modelo en diferentes contextos.

En general, este análisis resalta la importancia de controlar tanto las condiciones eléctricas como térmicas para prevenir fallos operativos. Las conclusiones obtenidas ofrecen una base sólida para implementar estrategias de monitoreo más efectivas, optimizando así la detección de anomalías y mejorando la seguridad y eficiencia del sistema. Es recomendable continuar investigando las relaciones entre estas variables en condiciones operativas reales para validar y ajustar los umbrales de detección.