CAMPUS DIGITAL FP
2024 - 2025

Curso de especialización
Inteligencia Artificial y Big Data

INFORME DE INVESTIGACIÓN

ERRORES DE COMUNICACIÓN EN DISPOSITIVOS IOT

1. Introducción

La comunicación en los dispositivos IoT es fundamental para poder garantizar el correcto funcionamiento de los mismos, así como para mantener un registro de datos consistente a lo largo del tiempo. El uso de tecnologías de comunicación móviles como 2G (segunda generación de redes móviles) y NB-IoT (Narrowband Internet of Things) como medio para esta comunicación permite que los dispositivos mantengan una conexión estable incluso en entornos con poca cobertura o con restricciones energéticas.

Estas tecnologías ofrecen varias ventajas frente a otras opciones como la comunicación por Wi-Fi o por radiofrecuencia de corto alcance, particularmente en el sector agrícola donde no siempre hay acceso a infraestructura de red fija. En concreto, el 2G proporciona una cobertura amplia y confiable incluso en zonas remotas, gracias a su extensa infraestructura móvil. Por su parte, la tecnología NB-IoT resulta especialmente adecuada para este entorno debido a su bajo consumo energético, su buena penetración en áreas con obstáculos naturales y su capacidad para mantener conectividad en ubicaciones con cobertura limitada.

Conocer si las características del dispositivo pueden provocar algún fallo de comunicación resulta esencial para poder minimizar la aparición de estos errores y garantizar una transmisión de datos robusta y fiable. Es igualmente importante analizar cómo evolucionan estos errores a lo largo del tiempo para lograr detectar patrones, tendencias o anomalías en la frecuencia de fallos.

En este contexto, este proyecto emplea distintas técnicas de minería de datos e inteligencia artificial para evaluar tanto la influencia de las características del dispositivo como la dinámica temporal de los errores, con el objetivo de construir un sistema de comunicación más resiliente y eficiente.

2. Objetivos e Hipótesis

Los dispositivos IoT analizados realizan publicaciones periódicas cuya frecuencia varía en función del modo de energía utilizado o la interacción del usuario. Cada publicación incluye diversos campos de información, entre ellos un identificador denominado *Data*, que indica el tipo de publicación. En el contexto de este documento, se considera como error toda publicación cuyo campo *Data* tiene el valor "*First Connect*", ya que suele estar asociada a eventos anómalos como reinicios del dispositivo, baja intensidad de señal o intentos de conexión a redes no autorizadas, entre otras posibles causas.

Esta investigación se articula en torno a dos líneas de análisis. La primera de ellas busca confirmar la existencia de una relación entre las características de los dispositivos y los errores de comunicación, identificando los atributos principales de dicha relación en caso de que exista. La segunda línea se centra en el estudio del comportamiento temporal de estos errores, con el objetivo de detectar patrones, tendencias o anomalías a lo largo del tiempo.

Los objetivos principales de este proyecto son:

1. Confirmar la existencia de una relación entre las características estáticas de los dispositivos y los errores de comunicación.
2. Identificar patrones y anomalías a través del análisis temporal de los errores de comunicación.

Dada la naturaleza del primer objetivo (confirmar o negar la existencia de dicha relación), se formulan las siguiente hipótesis para su estudio:

- H_0 (Hipótesis nula): No existe una relación significativa entre las características del dispositivo y la probabilidad de error en la comunicación.
- H_1 (Hipótesis alternativa): Existe una relación significativa entre al menos una característica del dispositivo y la probabilidad de error en la comunicación.

Adicionalmente se formulan objetivos secundarios para cada una de las líneas de investigación:

1.1. Identificar los atributos principales de dicha relación, es decir, aquellas características del dispositivo que están relacionadas con la aparición de dichos errores.
1.2. Crear un modelo de errores esperados dadas las características de un dispositivo, con el fin de identificar dispositivos con más fallos de lo habitual.
2.1. Crear un modelo de errores esperados a lo largo del tiempo, con el fin de identificar en tiempo real periodos críticos.
2.2. Estudiar la periodicidad de los errores a lo largo del tiempo para identificar tendencias.

3. Metodología y herramientas

El desarrollo del presente proyecto se ha basado en una metodología de análisis de datos que incluye diversas fases claramente diferenciadas: recopilación y preparación de los datos, exploración y visualización, análisis y modelado. Para cada una de estas fases se han empleado herramientas y bibliotecas específicas que han permitido maximizar la eficiencia y la precisión de los análisis.

3.1 Preparación y procesamiento de datos

Para la recopilación, limpieza y transformación de los datos se ha empleado principalmente el lenguaje Python, utilizando bibliotecas como Pandas para la manipulación de estructuras tabulares y Dask para el procesamiento distribuido de grandes volúmenes de datos, especialmente útil en contextos donde el tamaño del dataset supera la capacidad de memoria de un único dispositivo.

3.2 Exploración y visualización

Durante la fase exploratoria, se ha recurrido a librerías como Plotly o ggplot para generar visualizaciones que permitan identificar patrones, valores atípicos y relaciones entre variables. Este análisis exploratorio ha sido clave para orientar las decisiones posteriores en cuanto a la selección de variables y métodos de modelado.

3.3 Modelado predictivo y aprendizaje automático

El modelado ha sido realizado mediante bibliotecas como scikit-learn (sklearn), TensorFlow y Keras, permitiendo la construcción de modelos tanto clásicos como basados en redes neuronales. Estas herramientas han facilitado tanto el entrenamiento como la validación cruzada de modelos. Además, se ha utilizado Google Colab como entorno de ejecución para aprovechar recursos de computación avanzados sin necesidad de infraestructura local.

4. Análisis por características del dispositivo

4.1 Conjunto de datos

Se ha armado un conjunto de datos denominado “caracteristics_with_age.csv”, que reúne las características más significativas de cada dispositivo. En este conjunto de datos se han incluido aquellas características que no dependen de una variable temporal, como por ejemplo, el modelo del dispositivo, constante a lo largo del tiempo, y no el modo de energía, variable.

Este conjunto de datos contiene un único registro para cada dispositivo, identificado por su Imei cifrado.

4.1.1 Obtención del conjunto de datos

Para la construcción de este conjunto de datos se han cruzado dos tipos de archivos:

- deviceInfo.parquet: contiene características físicas del dispositivo, como el modelo, el módem o la versión de firmware.

- log_files.parquet: un conjunto de archivos de registros de publicaciones de los dispositivos que contiene, entre otros datos, el tiempo de vida del dispositivo, medido en días.

Para combinar estos archivos se ha utilizado como clave el identificador del dispositivo. El resultado de dicha unión ha sido procesado para generar el conjunto de datos final.

4.1.2 Descripción del conjunto de datos

El conjunto de datos contiene las siguientes columnas:

Nombre de columna	Tipo de dato	Descripción
Imei	Categórico	Identificador del dispositivo
InternalName	Categórico	Modelo del dispositivo
DID	Numérico	Identificador del distribuidor
LID	Numérico	Identificador del lote
Firmware	Categórico	Versión del firmware del dispositivo
Latitude	Categórico	Grupo* al que pertenece el dispositivo en función de la latitud en la que se encuentra
Longitude	Categórico	Grupo* al que pertenece el dispositivo en función de la longitud en la que se encuentra
TimeZone	Categórico	Zona horaria del dispositivo
Age	Numérico	Edad del dispositivo, en días
NumFirstConnect	Numérico	Cantidad de errores registrados por el dispositivo
CatFirstConnectEqual	Categórico	Categoría** a la que pertenece el dispositivo en función de la cantidad de errores que ha registrado
CatFirstConnectExtreme	Categórico	Categoría** a la que pertenece el dispositivo en función de la cantidad de errores que ha registrado.

*La agrupación de la latitud y longitud se ha realizado de forma que cada grupo abarque el mismo número de grados, dando como resultado 7 grupos a partir de la latitud y 6 empleando la longitud.

**La categorización de los dispositivos en función de la cantidad de errores registrados se ha realizado de dos formas distintas, generando como resultado dos columnas de datos distintos:

- Agrupaciones equitativas (CatFirstConnectEqual): Se han creado cuatro grupos (muy bajo, bajo, alto, muy alto), con aproximadamente el mismo número de dispositivos en todos ellos, en función de la cantidad de errores de cada uno de ellos.

- Identificación del 5% con más errores (CatFirstConnectExtreme): Se han creado dos grupos (bajo, alto) en los que se dividen a los dispositivos con la mayora cantidad de errores registrados y el resto.

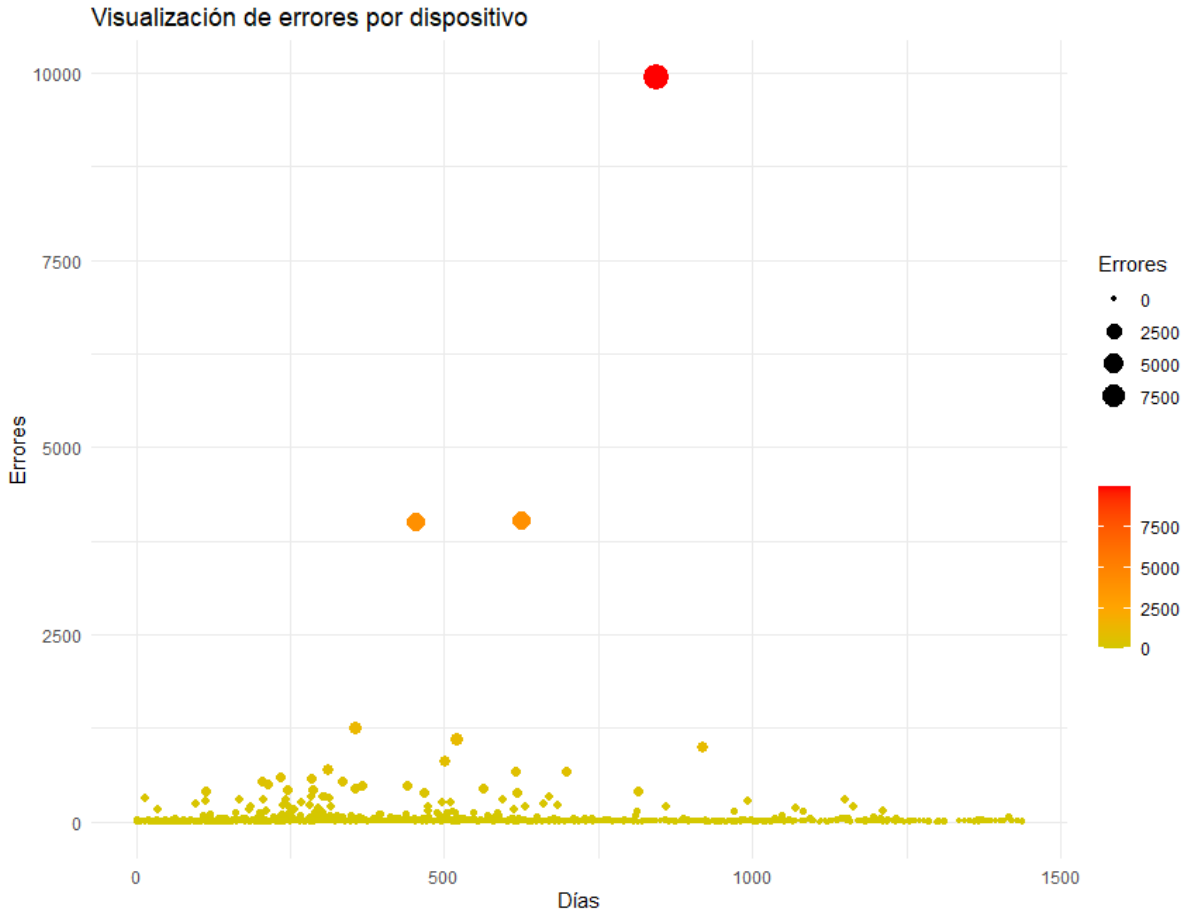
4.2 Análisis exploratorio de datos

Antes de adentrarnos en el análisis estadístico, es fundamental tener una visión clara de la estructura y comportamiento de los datos. A través de la visualización de datos se puede identificar tendencias, detectar valores atípicos y evaluar la calidad de la información disponible.

Este proceso nos ayuda a examinar cada variable de manera individual y en conjunto, permitiendo comprender las interacciones dentro del conjunto de datos. Para ello, se emplean herramientas gráficas y estadísticas que facilitan la interpretación de los datos y aseguran su fiabilidad antes de aplicar metodologías más complejas.

A través de la visualización de la edad del dispositivo junto con la cantidad de errores podemos observar que la mayor concentración de dispositivos con un alto número de errores se encuentra en los dispositivos más recientes, con menos de 2 años de vida.

Además, se pueden distinguir 3 dispositivos con una cantidad de errores extrema, muy por encima del resto de dispositivos.



Al visualizar las características de los tres dispositivos anómalos, se identifican dos características comunes entre ellos, el modelo del dispositivo (ATLAS_2S) y el lote (7).

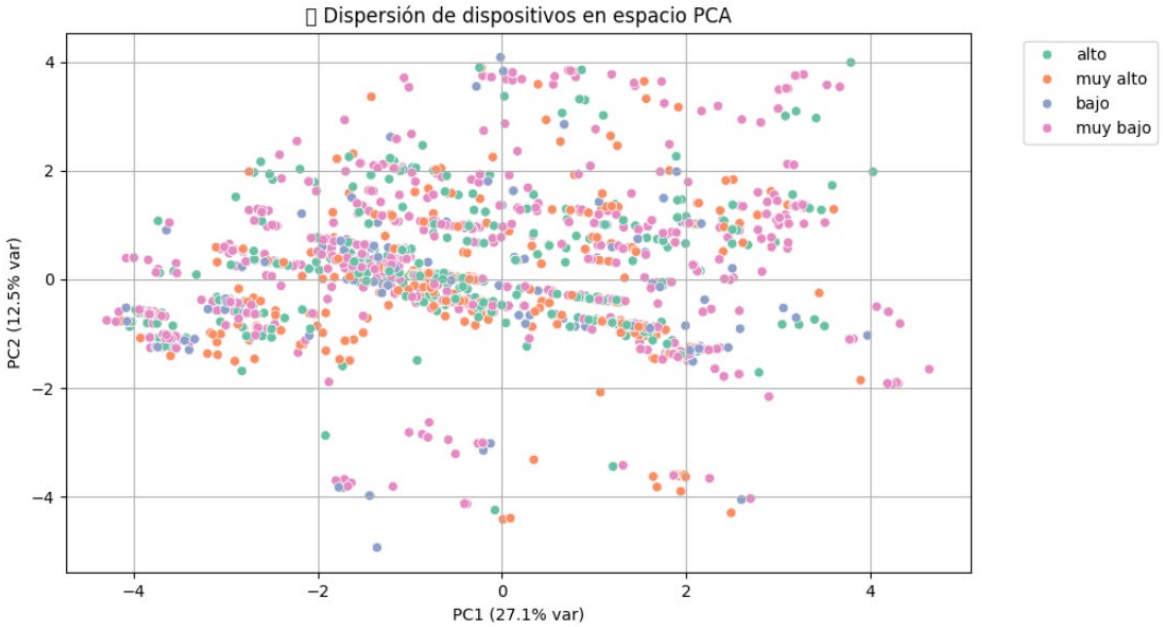
Variables	**Dispositivo 1**	**Dispositivo 2**	**Dispositivo 3**
Imei	1ee30dad47941e6f5dbf09039287b6da	44a89ea60c690d234a2323ab867cc92a	29d6ac808d9f113a4e652ca52fa29f1a
InternalName	ATLAS_2S	ATLAS_2S	ATLAS_2S
DID	2	20	20
LID	7	7	7
FID	1828	1924	1887
Firmware	0.1.21MBomba	0.1.341V7AUS	0.1.15AUS
Modem	BC66NBR01A11	BC66NBR01A10	BC66NBR01A11
Operator	Vodafone	Telstra	Telstra
Country	Spain	Australia	Australia
Latitude	Grupo 3	Grupo 5	Grupo 5
Longitude	Grupo C	Grupo E	Grupo F

TimeZone	Europe/Madrid	Australia/Perth	Australia/Adelaide
Age	844	625	454
NumFirstConnect	9933	4016	4007
CatFirstConnectEqual	muy alto	muy alto	muy alto
CatFirstConnectExtreme	alto	alto	alto

4.2.3 Análisis de componentes principales

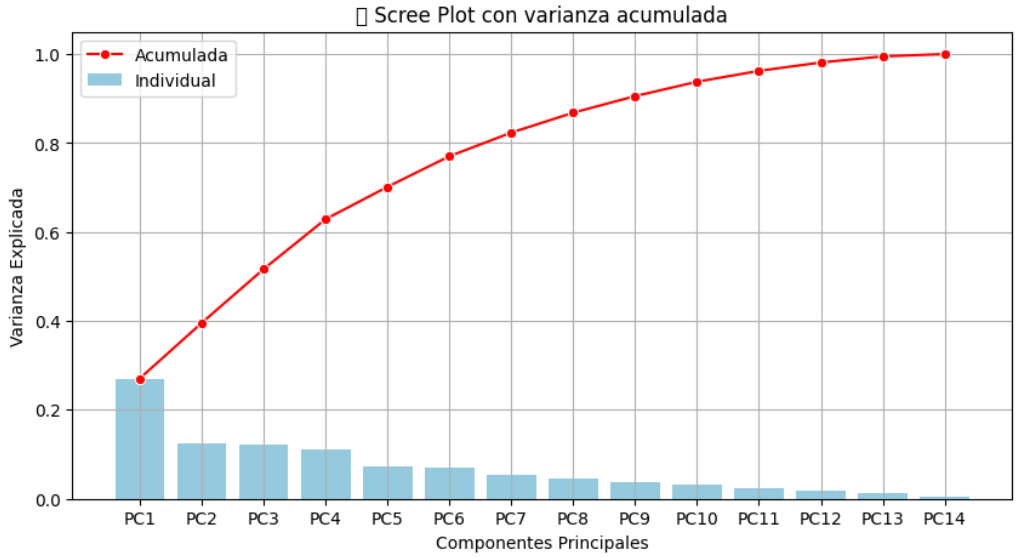
El análisis de componentes principales (PCA) es un algoritmo que permite reducir la dimensionalidad de un conjunto de datos manteniendo la variación de los mismos.[1] PCA, por tanto, identifica nuevas variables, que son combinaciones lineales de las variables originales. A través de la visualización de los datos generados se pueden observar similitudes y diferencias entre ellos y determinar si pueden ser agrupados.

En este caso no se aprecian grupos claramente diferenciados, lo que no permite agrupar los datos.



Este mismo hecho se verifica a través de la visualización de la varianza acumulada por cada componente, que permite determinar la cantidad de información original retenida por cada nuevo componente. En este caso se puede apreciar que el conjunto de datos original contiene una gran variedad de datos, ya que de las 16 variables originales, únicamente ha sido posible reducirlo a 14, sin que ninguna de las nuevas variables abarque la mayoría de datos.

[1] Ringnér, M. (2008). *What is principal component analysis?*. Nature biotechnology, 26(3), 303-304.



Por tanto, el resultado del análisis de componentes principales no ha arrojado resultados concluyentes que permitan aceptar la hipótesis alternativa H_1.

4.3 Análisis estadístico

4.3.1 Correlación entre variables

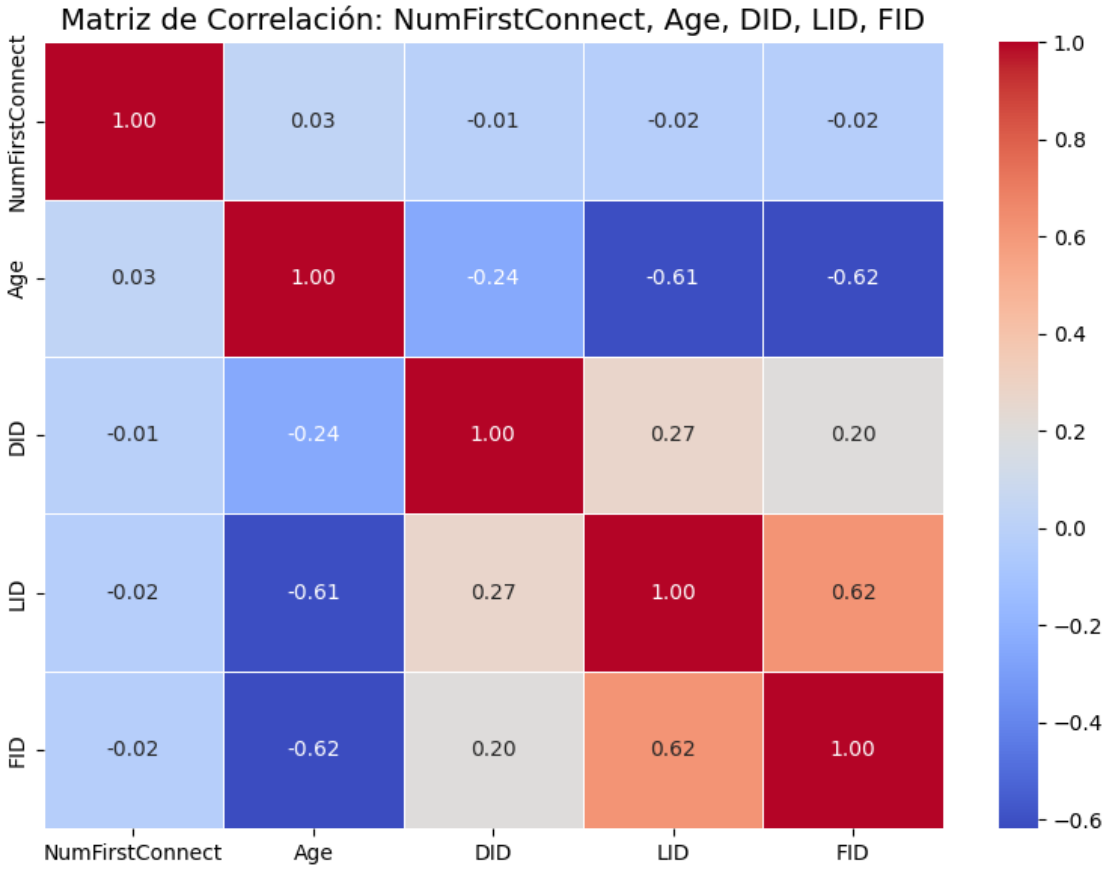
Evaluar la correlación entre variables cuantitativas es fundamental en el análisis de datos, ya que nos permite identificar relaciones y dependencias entre distintos factores del conjunto de información. A través de este análisis, podremos determinar cómo una variable influye en otra y si existe una asociación estadísticamente significativa entre ellas.

Correlación de variables cuantitativas

Para evaluar la correlación entre variables cuantitativas podemos emplear una matriz de correlación, que nos proporcionará una visión global de las interacciones entre múltiples variables, ayudándonos a detectar patrones y posibles relaciones lineales.

Al realizar analizar la matriz no se observa ningún resultado significativo sobre la variable NumFirstConnect. Todos los valores obtenidos se encuentran en el rango -0.5 < x < 0.5, insuficiente para considerar que hay una correlación entre las variables. Sobre los valores restantes, se observan los resultados esperados: relación entre el número de lote y el tiempo, relación entre el número de lote y la finca y relación entre el identificador de finca y el tiempo.[2]

[2] N.A. Todas estas variables son incrementales, por lo que están estrechamente ligadas con la variable temporal.



Correlación de variables categóricas

A través del test de Chi-cuadrado se puede evaluar la posible relación estadística entre dos variables de tipo categórico.[3] En este caso, se va a estudiar la relación entre la variable CatFirstConnect y el resto de atributos de los dispositivos. El resultado del test se mide a través del valor p-value. Si dicho valor es inferior a 0.05 (p-value < 0.05) se acepta la hipótesis alternativa: "Las variables examinadas tienen una relación significativa".

El test de Chi-cuadrado, sin embargo, no mide la intensidad de dicha relación, por lo que se puede dar el caso de que dicho test arroje un resultado positivo sin que posteriormente se pueda identificar dicha relación a través de otro tipo de pruebas, como los árboles de decisión. Es por ello que como herramienta adicional se ha empleado el test de Cramér's V que mide la intensidad de la relación entre dos variables y que permite categorizar dicha relación en función del resultado obtenido.[4]

Variable	Chi²	p-value	DOF	Cramér's V	Fuerza	Relacionado

[3] McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, *23*(2), 143-149.
[4] IBM (2025) *Cramér's V*. Recuperado de: https://www.ibm.com/docs/en/cognos-analytics/12.1.0?topic=SSEP7J_12.1.0/com.ibm.swg.ba.cognos.ug_ca_dshb.doc/cramersv.htm

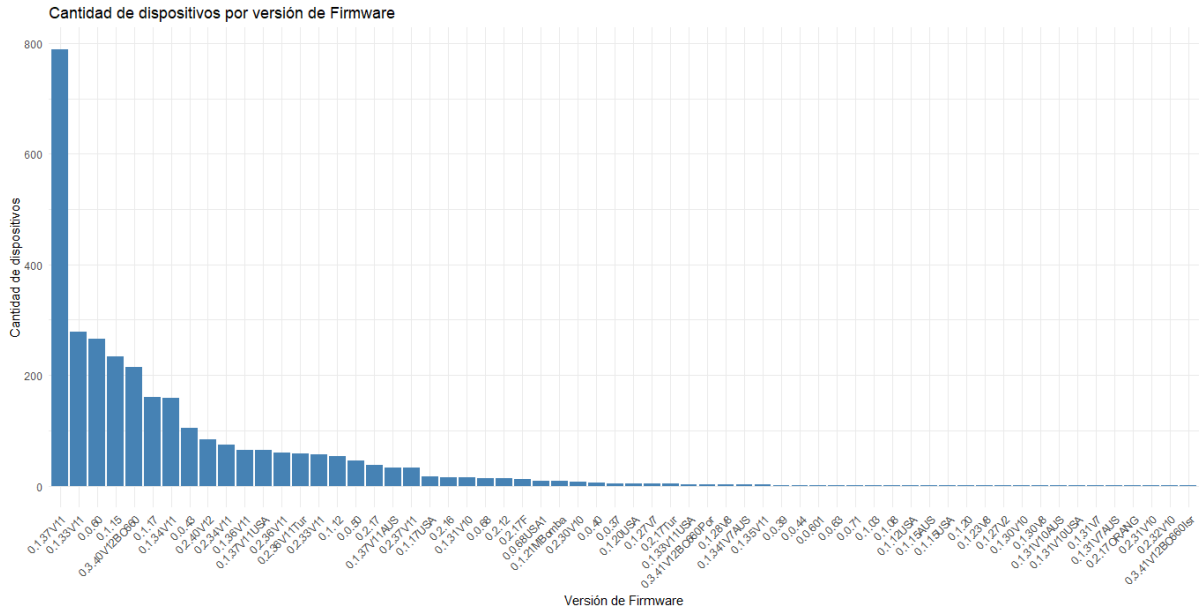
Firmware	764.1420	5.234e-72	183	0.2896	Moderada	Sí
TimeZone	257.9219	4.252e-12	120	0.1682	Débil	Sí
Modem	248.5975	1.573e-42	18	0.1652	Débil	Sí
Country	244.5249	3.831e-13	105	0.1638	Débil	Sí
Operator	164.7986	2.698e-27	15	0.1345	Débil	Sí
InternalName	119.8025	4.573e-17	18	0.1147	Débil	Sí
Longitude	81.0949	4.402e-11	15	0.0943	Nula o muy débil	Sí
Latitude	33.0463	1.311e-04	9	0.0602	Nula o muy débil	Sí

Analizando los resultados obtenidos se observa que pese a que existen relaciones entre las variables, la intensidad de dichas relaciones es débil, muy débil o nula, por lo que no se puede aceptar la hipótesis alternativa H_1.

4.3.2 Regresión lineal

En este estudio se ajustó un modelo de regresión lineal múltiple para explicar la variable NumFirstConnect en función de las características del dispositivo. Una primera aproximación que emplea todas las variables disponibles revela que las variables numéricas lid, did, fid no muestran asociaciones relevantes con la variable de interés, por lo que se descartan para el modelo.

En la creación de un nuevo modelo excluyendo dichas variables, la versión de firmware fue la única variable con un impacto estadísticamente significativo en el modelo ($F = 15.77$, $p < 2\times10^{-16}$). En particular, ciertas versiones, como 0.1.15AUS, 0.1.21MBomba y 0.1.341V7AUS, presentaron coeficientes muy elevados y p-value < 0.001. Esta variable, sin embargo, tampoco resulta significante, debido a que hay multitud de versiones de firmware con un número muy reducido de dispositivos.



El resto de variables incluidas (como operador, geografía, huso horario y antigüedad) no aportan información adicional.

El modelo final es estadísticamente significativo (p-value < 2.2e-16) pero su valor r-squared es muy bajo, 0.25, por lo que no explica la variabilidad en NumFirstConnect. Además, el error estándar de los residuos y su amplio rango sugiere la presencia de valores atípicos o de una alta variabilidad no explicada por el modelo. Por tanto, pese a que el modelo tiene validez estadística, su capacidad predictiva es limitada.

4.3.3 Modelos de árboles de decisión

Los árboles de decisión son modelos de predicción o clasificación que permiten dividir el conjunto inicial de datos en grupos cada vez más pequeños a partir de las características más relevantes de cada uno de esos grupos.[5] Una de las principales ventajas de los árboles de decisión frente a otros modelos es la facilidad para comprender el funcionamiento del modelo final gracias a la visualización de sus criterios.

El objetivo al emplear este tipo de algoritmos sobre el conjunto de datos es encontrar y visualizar aquellas características o conjuntos de ellas que tuvieran relevancia sobre la cantidad de errores de cada dispositivo.

DecisionTreeClassifier (CatFirstConnectExtreme)

Se ha entrenado un árbol de decisión con Python utilizando el algoritmo *DecisionTreeClassifier* para predecir la columna CatFirstConnectExtreme. Después de experimentar con varias configuraciones se ha observado que el mejor resultado utiliza el valor “entropy” para el parámetro “criterion”. Además, debido a que el número de datos está

[5] De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448-455.

desbalanceado (95% de los registros pertenecen a la clase “bajo”) se emplean pesos para penalizar con más fuerza los errores de clasificación en la categoría “alto” (de otro modo el modelo podría predecir el 100% de los registros como “bajo” y acertar el 95% del tiempo). Por último se ha limitado la profundidad del árbol a 10 para evitar el sobreajuste del modelo.

Con todo esto se ha logrado un modelo con una precisión aproximada del 80%. Sin embargo, al examinar la precisión específica de cada grupo se aprecia que aunque la categoría “bajo” la clasifica correctamente el 98% de las ocasiones, la categoría “alto” únicamente la clasifica correctamente en el 10% de los casos.

Por tanto, el modelo obtenido no tiene una base estadística sólida como para aceptar sus clasificaciones y, en consecuencia, no permite aceptar la hipótesis alternativa H_1.

```
Accuracy: 0.7976973684210527
              precision    recall  f1-score   support

           0       0.10      0.54      0.17        24
           1       0.98      0.81      0.88       584

    accuracy                           0.80       608
   macro avg       0.54      0.67      0.53       608
weighted avg       0.94      0.80      0.86       608

Confusion Matrix:
 [[ 13  11]
 [112 472]]
```

DecisionTreeClassifier (CatFirstConnectEqual)

Usando el mismo algoritmo, *DecisionTreeClassifier*, pero cambiando la columna a predecir a CatFirstConnnectEqual obtenemos una precisión global inferior, aunque la precisión de cada grupo aumenta.

De nuevo, los resultados obtenidos por este modelo no permiten aceptar la hipótesis alternativa H_1.

```
Accuracy: 0.5016447368421053
              precision    recall  f1-score   support

           0       0.37      0.39      0.38       148
           1       0.37      0.08      0.13        86
           2       0.42      0.37      0.39       119
           3       0.60      0.77      0.68       255

    accuracy                           0.50       608
   macro avg       0.44      0.40      0.39       608
weighted avg       0.47      0.50      0.47       608

Confusion Matrix:
 [[ 57   5  36  50]
 [ 28   7  11  40]
 [ 32   3  44  40]
 [ 39   4  15 197]]
```

DecissionTreeRegressor

Se ha entrenado también un árbol de decisión en Python utilizando el algoritmo *DecisionTreeRegressor* para predecir la columna NumFirstConnect. Pese a haber experimentado con distintas configuraciones, no se ha logrado alcanzar un resultado significativo. El mejor modelo, que emplea el valor *absolute_error* como *criterion* y limita la profundidad del árbol a 10, alcanzando un error medio absoluto del 13% pero con un error medio cuadrado de 4299 y un valor r-squared 0.226. Esto indica que el modelo comete errores significativos y que no explica la variabilidad de los datos.

```
Mean Squared Error: 4299.165296052632
Mean Absolute Error: 13.085526315789474
R2 Score: 0.22632573919053178
```

4.3.4 Algoritmos de reglas de asociación

Se ha empleado el algoritmo *a priori* cuyo funcionamiento se basa en la identificación de los subconjuntos frecuentes de mayor tamaño dentro del conjunto de datos, lo que permite determinar reglas de asociación para identificar tendencias en el conjunto de datos.[6]

En este caso se ha empleado específicamente para determinar cualquier regla en el conjunto de datos que sirva para pronosticar la categorización del número de errores del dispositivo.

[6] Agrawal, R & Srikant, R. (1994, 12 de septiembre) *Fast Algorithms for mining Association Rules.* IBM Almadem Research Center

Las reglas identificadas por el algoritmo no obtienen un índice suficiente como para considerarse relevante, teniendo como máximo un índice de 0.65, muy por debajo del 0.8 necesario, lo que no permite aceptar la hipótesis H_1.

4.3.5 Análisis de subgrupos

Dada la observación de que los tres dispositivos con mayor cantidad de errores comparten dos características, el modelo y el número de lote, se ha realizado un análisis para determinar si el grupo de los dispositivos con estas características tiene alguna diferencia significativa con respecto al resto.

Para ello se ha empleado el test de hipótesis Welch, que permite determinar si dos grupos de datos difieren estadísticamente de forma significativa.[7] El primer resultado obtenido no ha sido concluyente, obteniendo un p-value de 0.07246, cercano al valor necesario para aceptar la hipótesis alternativa pero sin alcanzarlo.

Con el objetivo de lograr un valor concluyente y dado que el conjunto de datos empleado incluye los tres dispositivos con valores anómalos (lo que afecta a los valores obtenidos), se han realizado dos pruebas adicionales:

En primer lugar se ha repetido el test excluyendo los tres dispositivos con valores anómalos. Este nuevo test da un resultado (p-value) de 0.7984, lo que no permite aceptar la hipótesis alternativa.

En segundo lugar se ha empleado el test Wilcoxon-Mann-Whitne que permite comparar las distribuciones o medianas de dos conjuntos de datos.[8] Por sus características, este test es menos sensible a valores anómalos. El resultado de este test (p-value) es 1, lo que tampoco permite aceptar la hipótesis alternativa.

Los resultados obtenidos no permiten aceptar la hipótesis alternativa, no hay diferencias significativas entre el grupo de dispositivos con del mismo lote y con el mismo modelo que los tres dispositivos con más errores y el resto de los dispositivos.

5. Análisis temporal de errores

5.1 Conjunto de datos

Se han creado dos conjuntos de datos, uno para cada línea de investigación. El primer conjunto de datos, denominado "*errors_with_disp.csv*", incluye la cantidad de errores registrada en un día junto con la cantidad de dispositivos únicos que han realizado

[7] West, R. M. (2021). Best practice in statistics: Use the Welch t-test when testing the difference between two groups. *Annals of clinical biochemistry*, *58*(4), 267-269.
[8] Turcios, R. S. (2015). Prueba de Wilcoxon-Mann-Whitney: mitos y realidades. *Rev Mex Endocrinol Metab Nutr*, *2*, 18-21.

publicaciones dicho día. El segundo conjunto de datos "*monthly_mtbf.csv*" contiene la cantidad de errores mensuales junto con la cantidad de publicaciones realizadas.

Los datos abarcan un periodo de 4 años, desde febrero de 2021 hasta febrero de 2025.

5.1.1 Obtención del conjunto de datos

Se ha sumado la cantidad de errores registrados cada día en los archivos de registro *"log_files.parquet"*. También se han contabilizado los identificadores únicos diarios.

5.1.2 Descripción del conjunto de datos

errors_with_disp.csv

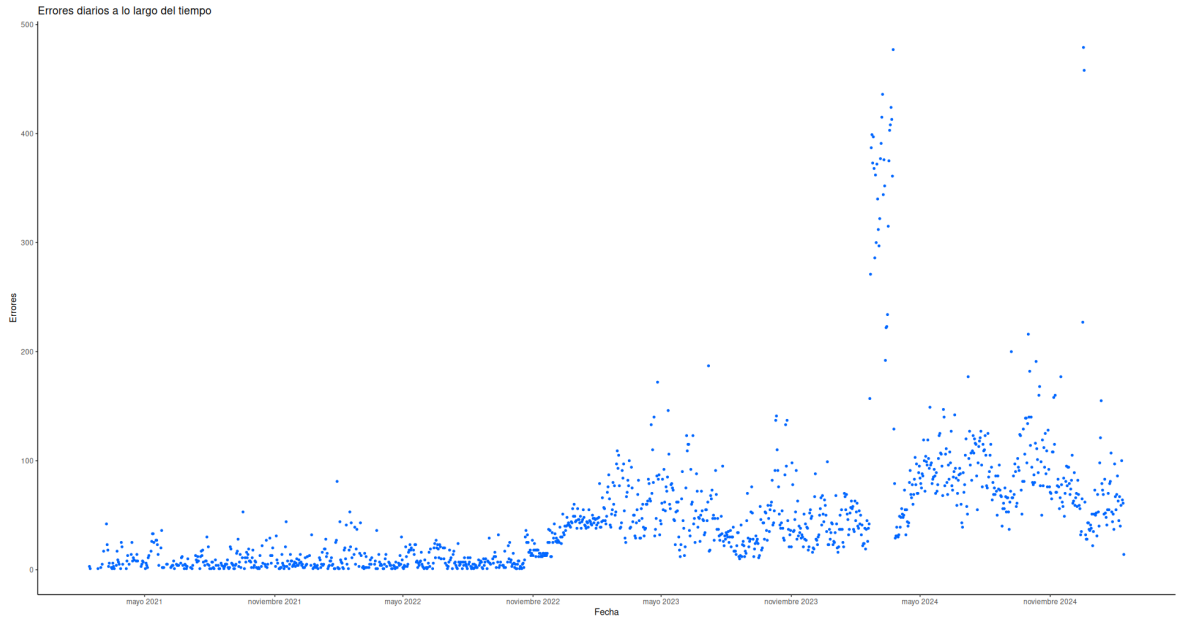
Nombre de columna	Tipo de dato	Descripción
Fecha	Fecha	Fecha
Errores	Numérico	Cantidad de errores registrada en esa fecha
Dispositivos	Numérico	Cantidad de dispositivos que publicaron datos en esa fecha

monthly_mtbf.csv

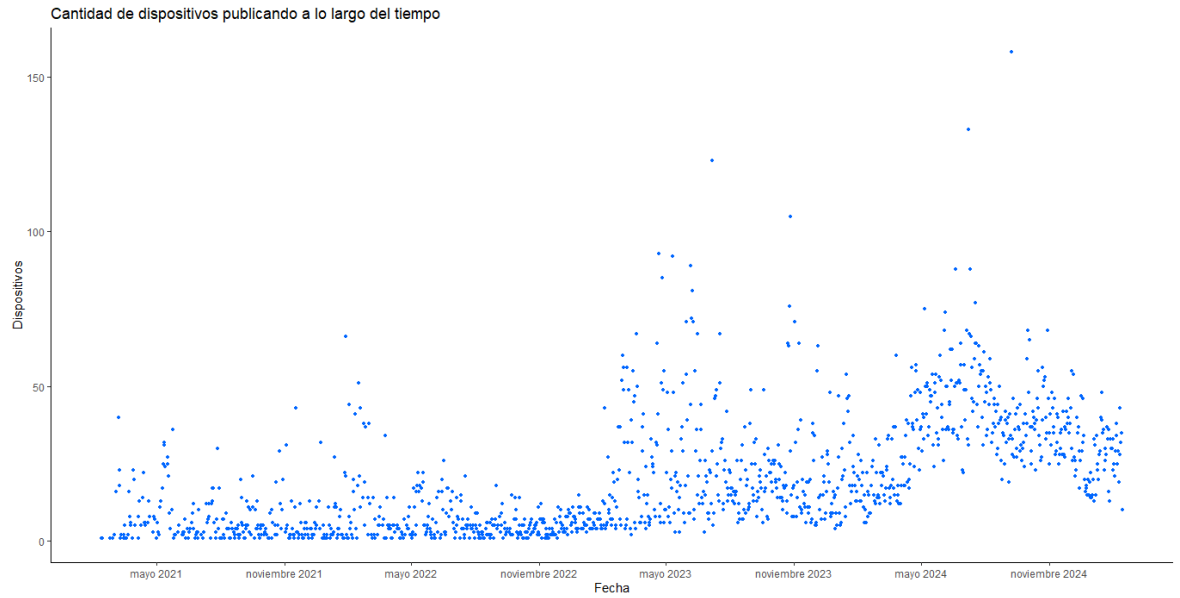
Nombre de columna	Tipo de dato	Descripción
Mes	Fecha	Fecha
CFC	Numérico	Cantidad de errores registrada ese mes
DisImei	Numérico	Cantidad de publicaciones realizadas ese mes

5.2 Análisis exploratorio de datos

En la visualización de la cantidad de errores a lo largo del tiempo, se observa que la mayoría de los puntos se concentran en valores bajos, lo que sugiere un comportamiento relativamente estable y predecible. Asimismo, se detectan agrupaciones esporádicas de puntos con valores significativamente más altos, que podrían corresponder a eventos anómalos o condiciones específicas.

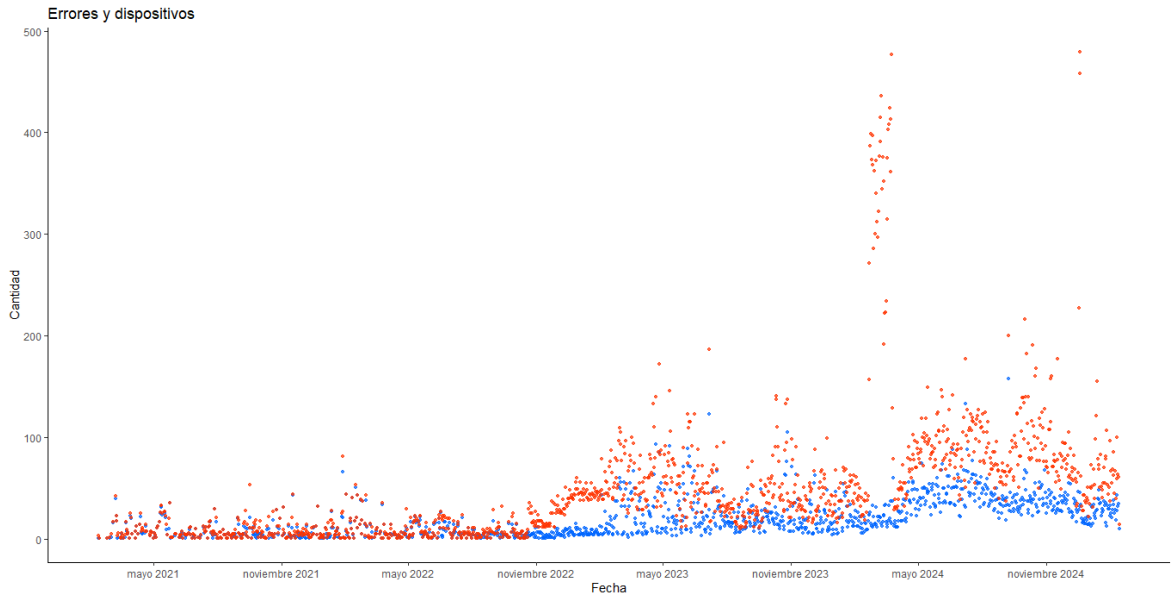


La visualización de la cantidad de dispositivos publicando en un día a lo largo del tiempo también muestra un alto grado de concentración en los valores bajos, con un crecimiento progresivo hacia el final. Este aumento es coherente con el contexto del proyecto, ya que refleja el aumento de dispositivos totales en funcionamiento. La tendencia sugiere que el número total de dispositivos en funcionamiento influye directamente en la cantidad de datos generados (publicaciones diarias).

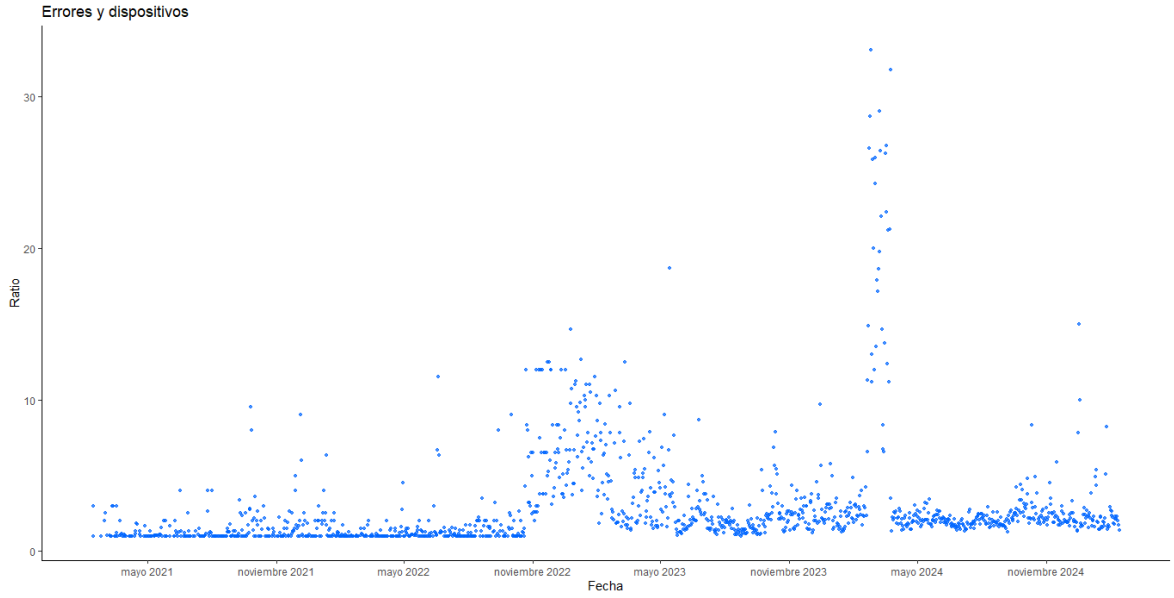


Para examinar la posible relación entre el número de dispositivos activos y los errores de comunicación diarios, se incluye un gráfico conjunto que muestra ambas variables en el

tiempo. Visualmente, se puede apreciar una correspondencia notable entre los errores y la cantidad de dispositivos, si bien los picos de errores quedan fuera de dicha relación. Esto abre la posibilidad de modelar dicha relación para identificar periodos anómalos, con un mayor número de errores de lo esperado.



Otra forma de visualizar la relación entre la cantidad de dispositivos y los errores diarios es calcular la relación entre ambas variables. Esto permite identificar periodos anómalos, aquellos que se alejan de la media.



5.3 Análisis estadístico

El análisis de los errores de comunicación a lo largo del tiempo se ha planteado desde dos perspectivas diferentes. Por un lado, se han empleado regresiones (lineales y logarítmicas) con el objetivo de clasificar la cantidad de errores diaria. Por otro lado, se ha llevado a cabo un análisis sobre la periodicidad de errores y su evolución a lo largo del tiempo.

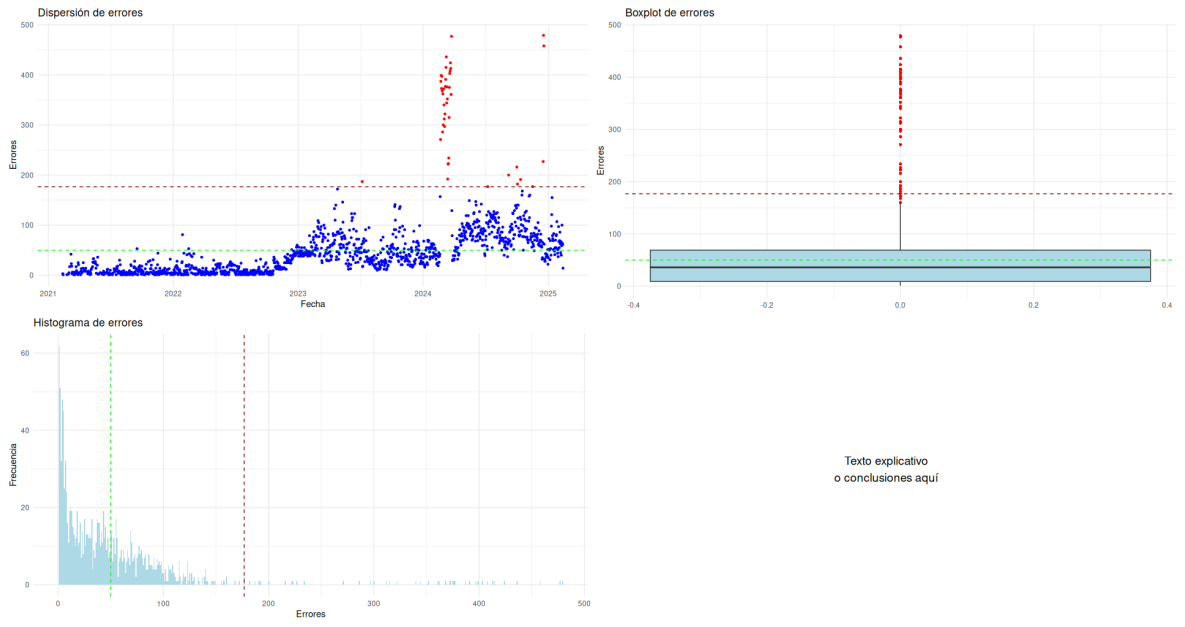
5.3.1 Clasificación de errores diarios

Las regresiones iniciales, lineal y logarítmica, utilizadas para calcular los errores en función de la cantidad de dispositivos, arrojaron una correlación significativa (coeficientes) entre las variables independientes y la variable dependiente. El p-value de ambos modelos están por debajo del límite (p-value: < 2.2e-16), por lo que son estadísticamente significativos. Sin embargo, el valor r-squared, que mide la proporción de la variabilidad de la variable dependiente, es demasiado bajo (0.26 en la regresión lineal y 0.48 en la regresión logarítmica), por lo que las regresiones iniciales no son suficientes para explicar la variabilidad de los datos.

Para lograr mejorar el valor r-squared se añadieron nuevas columnas que añaden contexto temporal a los datos, como el día de la semana (lunes, martes, miércoles…), el mes del año (enero, febrero) y la semana del año (1, 2, …, 53). Estas nuevas variables permiten capturar patrones estacionales que puedan influir en la aparición de errores.

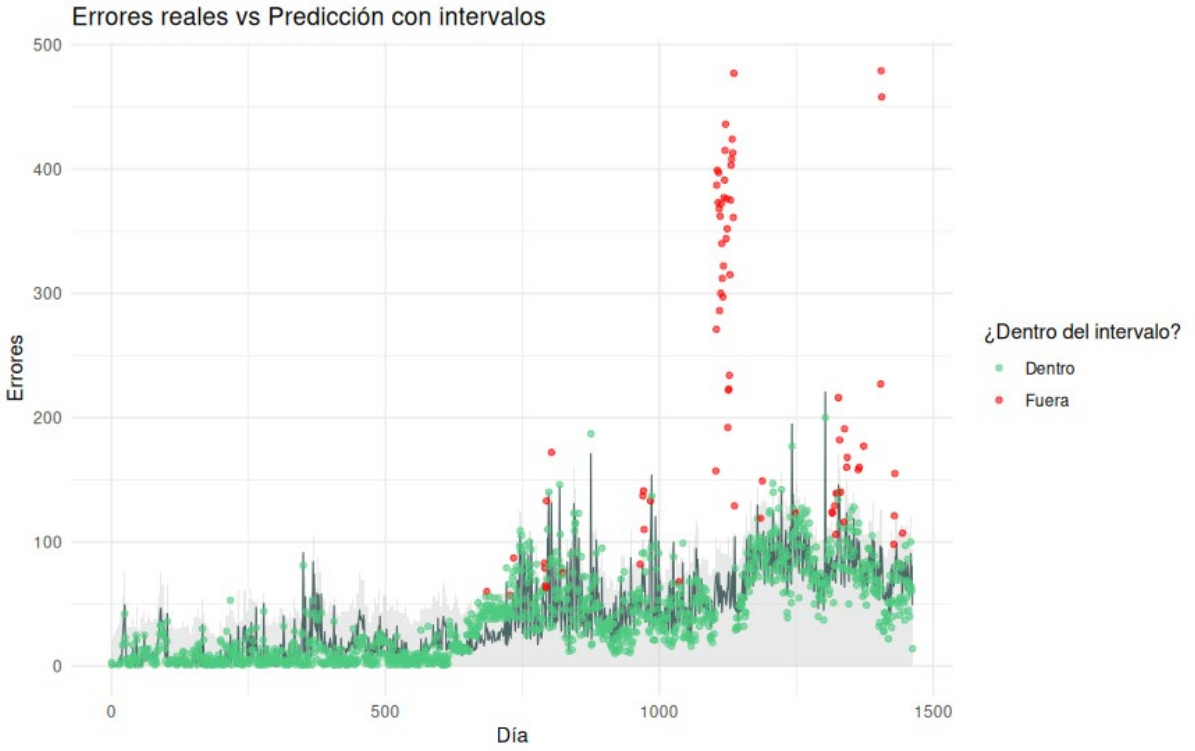
Al repetir las regresiones con estas nuevas variables, se observa una mejora en el ajuste de los modelos, reflejado en los nuevos valores r-squared (0.54 en la regresión lineal y 0.73 en la logarítmica), aunque no logran alcanzar el umbral para realizar predicciones con seguridad.

Finalmente, se realizó una limpieza al conjunto de datos para eliminar valores atípicos, ya que pueden distorsionar el ajuste de los modelos. Para identificar dichos valores, se utilizaron varias técnicas de visualización, optando finalmente por emplear el gráfico de cajas para excluir valores.



Al reconstruir los modelos tras la limpieza de datos, el valor r-squared alcanzó un valor significante en la regresión lineal (0.83), mientras que la regresión logarítmica no obtuvo mejoría.

Por tanto el modelo obtenido clasifica de forma precisa, mediante el intervalo de confianza, la cantidad de errores producidos en un día como normal o anómalo.



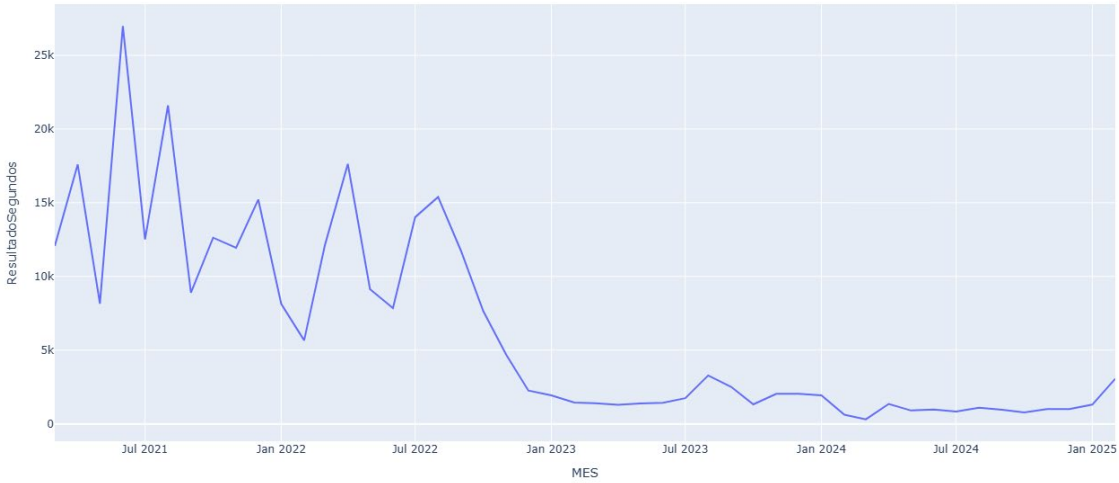
5.3.2 Tiempo medio entre fallos

El tiempo medio entre fallos (MTBF, por sus siglas en inglés *Mean Time Between Failures*) representa el tiempo promedio de funcionamiento continuo entre un error y el siguiente, lo que permite estimar la estabilidad operativa del dispositivo.[9] En este análisis, el MTBF se calculó mes a mes para identificar tendencias temporales y comparar el desempeño de las comunicaciones, facilitando la detección de patrones recurrentes y oportunidades de mejora en la disponibilidad del sistema.

Al visualizar el MTBF mensual, se observa que en los meses iniciales el tiempo medio entre errores era elevado, lo que sugiere un mayor grado de fiabilidad en las comunicaciones de los dispositivos. Desde septiembre de 2022, sin embargo, disminuye el MTBF significativamente, hasta situarse alrededor de 1000 segundos de intervalo. Por último, se

[9] Duer, S., Woźniak, M., Paś, J., Zajkowski, K., Bernatowicz, D., Ostrowski, A., & Budniak, Z. (2023). Reliability testing of wind farm devices based on the mean time between failures (MTBF). *Energies*, *16*(4), 1659.

aprecia una disminución adicional en los meses de febrero y marzo de 2024, coincidente con el periodo anómalo identificado anteriormente.



6. Conclusiones

Tras haber realizado un análisis exhaustivo de los errores de comunicación de los dispositivos en función de las características físicas de los mismos, no hay una evidencia que permita aceptar la hipótesis alternativa H_1. Es decir, no se puede afirmar que exista una relación entre las características estáticas de los dispositivos, como su modelo, su tecnología de comunicación o su localización y la cantidad de errores emitidos.

Por otro lado, el análisis temporal de los errores ha arrojado un resultado significativo, permitiendo construir un modelo que clasifica la cantidad de errores diarios en función de variables como el contexto temporal y la cantidad de dispositivos publicando datos. Este modelo permite reaccionar en tiempo real ante situaciones potencialmente problemáticas, permitiendo la toma de decisiones de negocio inteligentes y mejorando la eficiencia del sistema. A través del estudio del tiempo medio entre errores se han observado cambios en la tendencia de los dispositivos, con una diferencia clara antes y después de finales de 2022.

Además, la realización de este informe ilustra las ventajas del uso de la Inteligencia Artificial y el Big Data en entornos de IoT. Estas tecnologías permiten abordar los grandes volúmenes de datos generados por los dispositivos, identificar patrones temporales y comportamientos anómalos con alta precisión, y construir modelos predictivos que mejoran significativamente la capacidad de respuesta ante fallos. De este modo, se facilita una gestión más eficiente, proactiva y automatizada de las redes de dispositivos, contribuyendo a una mayor fiabilidad del sistema y a una mejor toma de decisiones operativas y estratégicas.

6.1 Líneas de investigación futuras

Durante la realización de este informe han surgido ciertas preguntas que no han sido resueltas y que podrían ser objeto de investigación en el futuro. La creación de un modelo

capaz de detectar eventos anómalos en tiempo real, si bien aporta un valor fundamental, no ofrece respuestas sobre los motivos de dichos eventos. Esto hace que ante sucesos como el periodo entre febrero y marzo de 2024, con una cantidad de errores irregular, sea necesario realizar una nueva investigación que explique el motivo. Además, el estudio de los errores de comunicación se ha realizado únicamente sobre dispositivos Atlas 2, por lo que sería útil ampliar dicho estudio a otros modelos y tipos de dispositivos. Por último, hay multitud de características adicionales que podrían ser investigadas además de los errores de comunicación, como el consumo de batería de los dispositivos (y la detección de anomalías de consumos energéticos) o la creación de modelos que mejoren las prestaciones de los mismos.