UAB
Universitat Autònoma
de Barcelona

Dipòsit digital
de documents
de la UAB

This is the **published version** of the bachelor thesis:

Muro Gómez, Adrià. *Detecció i classificació d'anomalies en maquinària rotativa industrial per a manteniment predictiu.* Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337321

under the terms of the BY-NC-ND license.

Detecció i classificació d'anomalies en maquinària rotativa industrial per a manteniment predictiu

Adrià Muro Gómez

Resum— Es presenta un sistema de detecció i classificació d'anomalies per al **manteniment predictiu** de maquinària rotativa industrial a partir de dades de sensors d'alta precisió. Les senyals de vibració es representen mitjançant vectors de característiques espectrals i estadístiques derivades de les mesures en posició i freqüència. El dataset proporcionat per l'empresa AEInnova comprèn registres reals de dispositius de la família **NOD-0007**. En una primera etapa, s'entrena un model **Isolation Forest** amb normalització robusta i ajust del paràmetre de contaminació per detectar comportaments anòmals. En una segona etapa, les anomalies detectades es processen mitjançant tècniques de clustering **(K-Means)** per identificar patrons naturals en les dades, i s'entrena un model supervisat de **Gradient Boosting** capaç de classificar-les en diferents tipologies de fallo amb coherència física. El pipeline complet —preprocessament, modelatge i generació d'alertes— es desplega a **AWS SageMaker** amb emmagatzematge centralitzat a **Amazon S3** i visualització mitjançant un dashboard interactiu desplegat a **Amazon EC2**. El sistema permet no només detectar anomalies sinó també proporcionar una classificació interpretable orientada al diagnòstic industrial.

Paraules clau— manteniment predictiu, classificació de fallos, detecció d'anomalies, Isolation Forest, Gradient Boosting, anàlisi de vibració, AWS SageMaker, sensors industrials.

Abstract— We present a system for *anomaly detection* and fault classification for **predictive maintenance** in industrial rotating machinery using high-precision sensor data. Vibration signals are represented as feature vectors combining spectral and statistical characteristics —RMS, energy, axis dominance, and energy ratios between domains. The dataset provided by AEInnova consists of real-world records from **NOD-0007** devices. In the first stage, an **Isolation Forest** model is trained with robust normalization and contamination tuning to detect anomalous behavior. In the second stage, detected anomalies are processed using **K-Means** clustering to uncover natural patterns, and a **Gradient Boosting** classifier is trained to assign them to different fault types with physical meaning. The full pipeline —preprocessing, modelling, and alert generation— is deployed on **AWS SageMaker**, with centralized storage on **Amazon S3** and visualization through an interactive dashboard deployed on **Amazon EC2**. The system not only detects anomalies but also provides interpretable fault classification for industrial diagnostics.

Index Terms— predictive maintenance, fault classification, *anomaly detection*, Isolation Forest, Gradient Boosting, vibration analysis, AWS SageMaker, industrial sensors.

1 INTRODUCCIÓ – CONTEXT DEL TREBALL

El **manteniment predictiu** representa un dels reptes tecnològics més rellevants en la indústria moderna. La creixent complexitat de les instal·lacions i l'augment del nombre de màquines rotatives en operació contínua incrementen la probabilitat d'avaries no planificades, amb un impacte directe en la productivitat, l'eficiència energètica i els costos operatius. Segons informes del sector [1], aproximadament un **30% de les avaries** en equips rotatius podrien evitar-se mitjançant un sistema predictiu adequat, amb un estalvi potencial d'entre el **15% i el 20% en costos** de manteniment anuals.

Els sistemes de monitorització basats en vibració permeten capturar informació sobre l'estat dinàmic dels equips de forma no invasiva i en temps real. L'espectre de vibració conté informació directa sobre possibles defectes mecànics —desequilibris, desalineacions, afluixaments o degradació de rodaments— que es manifesten com a components característics en determinades bandes de freqüència. Tanmateix, l'anàlisi manual d'aquest tipus de senyals és inviable a escala industrial, on el volum de dades generat per múltiples dispositius en continu fa necessari l'ús de tècniques automàtiques de detecció d'anomalies. En aquest context, les tècniques d'aprenentatge automàtic permeten automatitzar no només la detecció de comportaments anòmals, sinó també la seva interpretació. Tot i que molts sistemes es limiten a identificar desviacions respecte a un comportament normal, en entorns industrials és especialment rellevant anar un pas més enllà i proporcionar una classificació del

E-mail de contacte: Adria.Muro@autonoma.cat
Treball tutoritzat per: Raul Aragonés Ortíz (Departament de Microelectrònica i Sistemes Electrònics)
Curs 2025/26

tipus de fallo, facilitant la presa de decisions per part del personal tècnic.

Aquest projecte es desenvolupa en col·laboració amb AEInnova, empresa especialitzada en la monitorització industrial, que proporciona dades reals de sensors instal·lats en màquines rotatives. Les dades comprenen espectres de vibració i velocitat de rotació recollits per dispositius de la família **NOD-0007**, distribuïts en diversos equips i configuracions de mesura.

El sistema proposat combina tècniques d'aprenentatge no supervisat i supervisat en un pipeline estructurat. En una primera etapa, s'identifiquen anomalies mitjançant models basats en aïllament. Posteriorment, aquestes anomalies es processen per descobrir patrons estructurats en les dades i entrenar un model capaç de classificar-les en diferents tipologies de fallo amb significat físic. Finalment, el pipeline complet es desplega en un entorn cloud basat en **AWS SageMaker**, amb emmagatzematge a **Amazon S3** i visualització mitjançant **Streamlit**, permetent la seva aplicació en entorns industrials reals.

2 ESTAT DE L'ART

2.1 Detecció d'anomalies en maquinària rotativa

La detecció d'anomalies en senyals de vibració industrial ha estat àmpliament estudiada en la literatura científica. Els primers enfocaments es basaven en llindars estàtics sobre paràmetres estadístics del senyal —**RMS, curtosi o factor de cresta**—, però la variabilitat de les condicions operatives limita la seva aplicabilitat en entorns reals [9]. Posteriorment, tècniques basades en densitat com **LOF (*Local Outlier Factor*)** [5] van permetre detectar anomalies de forma no supervisada comparant la densitat local de cada observació amb la dels seus veïns, sense assumir cap distribució concreta de les dades.

L'algorisme **Isolation Forest** [4] representa un canvi de paradigma: en lloc de modelar el comportament normal, aïlla directament les anomalies aprofitant que les observacions atípiques són més fàcils de separar en particions aleatòries de l'espai de característiques. Gràcies a la seva eficiència computacional i robustesa davant de conjunts d'alta dimensionalitat, s'ha consolidat com un dels mètodes de referència en aplicacions industrials d'*anomaly detection* [6].

2.2 Representació espectral i extracció de característiques

El processament de senyals de vibració en el domini de la freqüència permet extreure informació sobre l'estat mecànic dels equips de forma més robusta que en el domini temporal. La transformada de Fourier discreta converteix el senyal en una representació freqüència–amplitud sobre la qual es poden calcular indicadors com **l'energia per bandes, la freqüència dominant o el factor de cresta** [9]. Aquests vectors redueixen la dimensionalitat i constitueixen l'entrada habitual dels models en manteniment predictiu

2.3 Classificació de tipologies de fallada

La classificació de fallos a partir de senyals vibracionals ha estat abordada amb una gran varietat de tècniques supervisades. Mètodes clàssics com les màquines de vectors de suport (**SVM**) i els arbres de decisió han mostrat bons resultats en condicions controlades, però requereixen dades etiquetades, que rarament estan disponibles en entorns industrials reals [6]. Les tècniques d'*ensemble* com el **Gradient Boosting** han demostrat una major robustesa davant de dades heterogènies i desequilibris de classe, mantenint un bon rendiment fins i tot amb conjunts d'entrenament de mida moderada [7].

En absència d'etiquetes, els mètodes de clustering —especialment **K-Means**— permeten descobrir patrons naturals en les anomalies detectades i construir categories de fallada amb coherència estadística, que posteriorment poden ser interpretades físicament per experts en maquinària [8]. Aquest enfocament híbrid, que combina detecció no supervisada i classificació posterior, és el que ha guanyat més rellevància en aplicacions industrials reals on l'etiquetatge manual és inviable.

2.4 Desplegament en entorns cloud

El desplegament de sistemes de manteniment predictiu en infraestructura cloud ha experimentat un creixement sostingut en els darrers anys. Plataformes com **AWS SageMaker** [2] ofereixen serveis gestionats per a totes les etapes del pipeline de machine learning —processament, entrenament, inferència i monitorització— sense necessitat de gestionar infraestructura de còmput. L'emmagatzematge centralitzat a **Amazon S3** [3] facilita la traçabilitat dels experiments i la reutilització dels artefactes entre etapes, mentre que la integració nativa amb eines de visualització permet construir sistemes d'explotació accessibles per a perfils no tècnics.

3 MOTIVACIONS

3.1 Motivació Industrial

En entorns industrials, la **detecció d'anomalies** en maquinària rotativa és clau per reduir avaries no planificades i optimitzar els costos de manteniment. No obstant això, detectar una anomalia no sempre és suficient si no es pot interpretar correctament. El personal tècnic necessita informació addicional sobre la possible causa del problema per poder actuar de manera eficient.

L'empresa AEInnova disposa d'una infraestructura de sensors capaç de generar dades contínues sobre l'estat dels equips. Tanmateix, el volum i la complexitat d'aquestes dades fan inviable la seva anàlisi manual. En aquest context, es fa necessari un sistema automatitzat que no només detecti comportaments anòmals, sinó que també

sigui capaç **de classificar-los en tipologies de fallo** interpretables.

Aquest enfocament permet millorar la presa de decisions en manteniment predictiu i aporta un valor afegit respecte als sistemes basats únicament en detecció.

3.2 Motivació Tècnica

Des del punt de vista tècnic, el projecte representa una intersecció especialment rica entre diverses disciplines de l'Enginyeria de Dades. Treballar amb senyals de vibració en el domini de la freqüència implica dominar tècniques d'anàlisi espectral —transformades de Fourier, extracció de característiques per bandes— que van més enllà del processament de dades tabular convencional.

A més, la detecció d'anomalies en un entorn no supervisat, on no existeixen etiquetes de fallada, planteja reptes metodològics rellevants: la definició de normalitat, el control del paràmetre de contaminació i la reducció de falsos positius mitjançant filtres temporals. Finalment, el desplegament del pipeline a **AWS SageMaker** introdueix la dimensió d'enginyeria de sistemes al núvol, des de la gestió de dades a **Amazon S3** fins a l'automatització de jobs d'entrenament i inferència, competències directament aplicables en entorns professionals reals.

4 OBJECTIUS

• Construir un pipeline complet de detecció i classificació de fallos sobre dades reals de vibració industrial, des de la ingesta dels registres JSON fins a la generació d'alertes estructurades.

• Desenvolupar un model híbrid basat en **Isolation Forest**, **K-Means** i **Gradient Boosting** capaç de detectar anomalies i classificar-les en tipologies de fallo amb coherència física.

• Desplegar el pipeline complet a **AWS SageMaker** amb emmagatzematge centralitzat a **Amazon S3**, automatització dels processos de preprocessament i entrenament, i visualització de resultats mitjançant Streamlit.

• Avaluar el comportament del sistema sobre dades reals d'AEInnova i analitzar les limitacions derivades d'un entorn sense etiquetes de fallada.

5 CONJUNT DE DADES

5.1 Origen i abast del dataset

Aquest treball es basa en dades reals de monitorització industrial proporcionades per l'empresa AEInnova, amb l'objectiu d'analitzar el comportament vibracional de maquinària rotativa i detectar anomalies associades a possibles fallades.

Tot i que el repositori global inclou diferents tipologies de dispositius i formats de dades, s'ha definit un abast específic per garantir la coherència de l'anàlisi. L'estudi es centra en els registres del **DatasetI**, que conté dades de vibració consistents per a l'anàlisi espectral.

Durant el procés de desenvolupament s'ha identificat l'existència de dos formats de representació de senyal dins del dataset. En aquesta segona versió del sistema, s'ha prioritzat el format *readings*, ja que proporciona una representació directa en el domini de la freqüència i permet una descripció més rica del comportament vibracional. Això ha permès ampliar significativament el volum de dades utilitzades i millorar la representativitat dels models.

5.2 Estructura de les dades

Les dades es presenten en format JSON per línies, amb una estructura equivalent a exportacions de bases de dades tipus DynamoDB. Cada registre conté:

- Identificador del dispositiu (**dev_eui**)
- Mode de captació (**mode**)
- Eix de mesura (**axis**)
- Marca temporal (**timestamp**)
- Temperatura (**T1**)
- Espectre de vibració (***readings***)

La variable *readings* representa el senyal en el domini de la freqüència com una col·lecció de parelles freqüència–amplitud. Aquesta representació permet analitzar la distribució d'energia espectral i extreure característiques més informatives que en el format de canals separats.

5.3 Qualitat de les dades i preprocessament

Abans del modelatge, s'ha realitzat una fase de validació per garantir la qualitat del dataset. Els principals criteris aplicats han estat:

- **Eliminació de registres** sense espectre vàlid
- **Filtrat de valors** físicament inconsistents o buits
- **Conversió i normalització** dels espectres de freqüència
- **Ordenació i estructuració** homogènia dels vectors de característiques

Aquest procés assegura que les dades utilitzades siguin consistents i representatives del comportament real dels sistemes.

5.4 Limitacions

No tots els grups disposen d'un volum de dades suficient per a l'entrenament. Per això, s'estableix un llindar mínim de mostres per considerar un grup com a vàlid. A més, cal tenir en compte que el sistema es basa en dades de vibració i, per tant, no pot capturar totes les tipologies de fallada possibles. El model s'ha d'interpretar com una eina de detecció d'anomalies i suport a la decisió, més que no pas com un sistema de diagnosi exhaustiva.

Finalment, el fet de treballar en un entorn no supervisat implica que no es disposa d'etiquetes de fallada reals, per la qual cosa la validació es basa en coherència estadística, estabilitat del model i interpretabilitat física de les característiques extretes.

6 METODOLOGIA I MODELATGE

6.1 Flux general del sistema

El sistema desenvolupat implementa un pipeline complet de detecció d'anomalies i classificació de fallades sobre dades de vibració industrial. Aquest pipeline cobreix totes les etapes del procés, des de la ingestió de dades fins a la generació de prediccions interpretables. El flux de processament es pot resumir en les següents etapes:

1 Lectura i validació de dades en format JSON
2 Extracció de l'espectre de vibració (format *readings*)
3 Generació de característiques espectrals
4 Detecció d'anomalies mitjançant **Isolation Forest**
5 Agrupació de patrons anòmals amb **K-Means**
6 Classificació final amb **Gradient Boosting**
7 Generació de sortides interpretables

Aquest pipeline s'ha implementat inicialment en entorn local i posteriorment s'ha adaptat a **AWS SageMaker** per al seu desplegament escalable.

6.2 Extracció de característiques espectrals

Cada espectre de vibració es transforma en un vector de **17 característiques** que resumeixen el seu comportament. Les principals característiques extretes són:

- Amplitud màxima
- Amplitud mitjana
- Amplitud pic a pic
- Desviació estàndard
- Valor RMS
- Freqüència dominant
- Energia total
- Factor de cresta
- Temperatura
- Mode operatiu (mode)
- Eix físic de mesura (axis)

El mode operatiu i l'eix físic s'inclouen perquè el rang de freqüències mesurat varia en funció del mode; obviar-los podria induir el model a confondre patrons espectrals de modes d'operació diferents.

El valor RMS, on x_i és l'amplitud de la mostra i i N és el nombre total de mostres de l'espectre, es defineix com:

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$$

i l'energia espectral, on x_i és l'amplitud de cada component freqüencial de l'espectre discret, es calcula com:

$$E = \sum_{i=1}^{N} x_i^2$$

A més, es calcula la distribució de l'energia en tres bandes de freqüència:

- Baixa: $f < 30$Hz
- Mitjana: $30 \leq f < 70$Hz
- Alta: $f \geq 70$Hz

També s'inclouen les proporcions relatives d'energia per cada banda, normalitzades respecte al total.

Aquest conjunt de característiques permet representar cada observació com un vector numèric compacte i informatiu.

6.3 Model de detecció d'anomalies

Per a la detecció d'anomalies s'utilitza l'algorisme **Isolation Forest**, adequat per a problemes no supervisats. El model assumeix que les observacions anòmales són més fàcils d'aïllar mitjançant particions aleatòries de l'espai de característiques. El model s'entrena de forma global sobre totes les mostres i retorna una predicció binària: **normal (0) o anòmala (-1)**.

El paràmetre de **contaminació** s'ajusta de forma adaptativa en funció del volum de dades, mantenint un rang aproximat del **10%–20%**.

6.4 Agrupació de patrons anòmals

Un cop detectades les anomalies, aquestes es modelen com a subconjunt independent del dataset. Per identificar tipologies de comportament dins de les anomalies, s'aplica un algorisme de clustering (**K-Means**).

El nombre de clústers s'ha fixat en **K=4** per dues raons complementàries. Des del punt de vista del domini, la literatura de diagnosi vibracional estableix **quatre tipologies principals de fallada** en maquinària rotativa: **desequilibri, desalineació, afluixament mecànic i degradació de rodaments** [9]. Valors superiors de K —com K=6— tendeixen a fragmentar patrons que responen a un mateix fenomen físic, produint clústers solapats sense interpretació tècnica clara. Aquesta segmentació és clau per passar de detecció d'anomalies a classificació de tipus de fallada.

6.5 Classificació de fallades

Sobre les anomalies agrupades s'entrena un model supervisat de tipus **Gradient Boosting Classifier**, que permet inferir una possible tipologia de fallada associada a cada patró detectat.

Per reduir falsos positius, el sistema incorpora un mecanisme implícit d'estabilitat basat en la consistència de

les prediccions sobre el conjunt de dades. En entorns industrials, aquest tipus de filtres són importants per evitar que variacions puntuals del senyal generin alarmes innecessàries.

6.6 Interpretació dels resultats

Les classes generades pel model es relacionen amb els patrons espectrals identificats pel procés de clustering. Segons la literatura de diagnosi vibracional [9], cada tipologia constitueix una hipòtesi de diagnòstic compatible amb els patrons vibracionals observats.

- **Alta desviació estàndard d'amplitud i factor de cresta elevat, energia dominant en alta freqüència (>70 Hz)** → compatible amb inestabilitat estructural per impactes o afluixament mecànic.
- **Energia dominant en banda mitjana (30–70 Hz), freqüència dominant ~52 Hz, valors globals baixos** → compatible amb desalineació, component a 2× RPM.
- **Energia concentrada exclusivament en alta freqüència (>70 Hz), freqüència dominant >1.000 Hz** → compatible amb modes operatius d'alt rang freqüencial, registres capturats en modes d'adquisició d'alt rang freqüencial.
- **Energia predominant en banda mitjana (~63 Hz) amb components menors en baixa i alta freqüència** → compatible amb degradació general o comportament vibracional mixt sense una causa única dominant.

Aquesta interpretació connecta els resultats del model amb fenòmens físics reals de la maquinària industrial.

7 VALIDACIÓ EXPERIMENTAL EN ENTORN LOCAL

7.1 Validació del pipeline

El pipeline s'ha executat sobre el dataset complet en entorn local, cobrint totes les etapes definides a la Secció 6. Els resultats generats permeten validar tant la funcionalitat computacional del sistema com la coherència física dels patrons detectats.

7.2 Anàlisi exploratòria i visualització de dades

En el marc del desenvolupament del pipeline en entorn local, s'ha realitzat una anàlisi exploratòria de les dades amb l'objectiu de validar la qualitat del conjunt de dades, entendre el comportament dels senyals de vibració i verificar la coherència física de les característiques espectrals utilitzades pel model.

Aquesta etapa té un caràcter estrictament de validació preliminar i no forma part del sistema final de desplegament, sinó que permet assegurar que la representació dels senyals és adequada per a la detecció d'anomalies.

El conjunt de dades analitzat està format per registres industrials reals en format *readings*, on cada observació conté un espectre de vibració representat per parelles freqüència–amplitud.

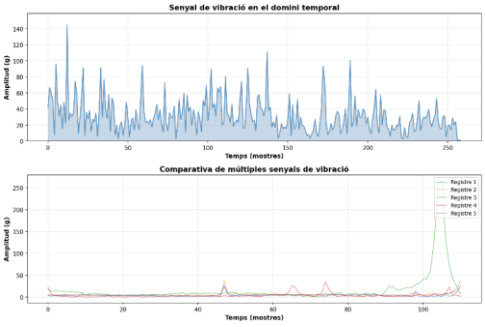


Fig. 1. Senyal de vibració en el domini temporal i comparativa de múltiples registres.

En la primera part de la figura es mostra l'evolució temporal del senyal de vibració d'un registre individual. Aquest gràfic permet observar la variabilitat de l'amplitud al llarg del temps i identificar possibles fluctuacions o inestabilitats en el comportament del sistema.

En la segona part es presenta una comparativa de cinc registres diferents superposats, la qual permet analitzar la consistència del senyal entre diferents observacions. Aquesta visualització és útil per identificar variabilitat entre condicions operatives i detectar patrons recurrents o desviacions respecte del comportament habitual.

Comentado [RA1]: Quines freqüències associes a cad fenòmen? Fes referència a alguna bibliografia.

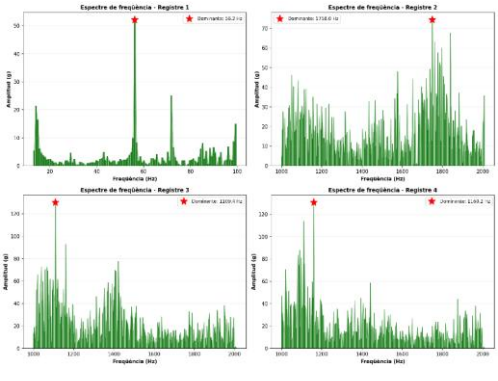


Fig. 2. Espectre de freqüència de diferents registres amb identificació de la freqüència dominant.

Aquesta figura mostra l'espectre de freqüència de quatre registres diferents de vibració. En cada cas es

representa la distribució d'amplitud en funció de la freqüència. S'hi destaca la freqüència dominant, que correspon al punt de màxima amplitud dins de cada espectre. Aquest element és especialment rellevant, ja que pot estar associat a fenòmens físics concrets com desequilibris, desalineacions o ressonàncies mecàniques.

La comparació entre espectres permet observar diferències significatives en la distribució d'energia, fet que justifica l'ús d'aquest tipus de representació per a la detecció d'anomalies.

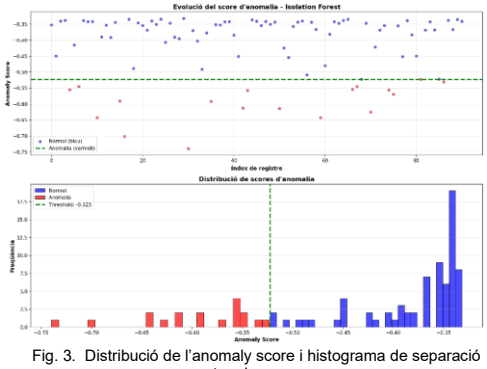


Fig. 3. Distribució de l'anomaly score i histograma de separació entre classes.

Aquesta figura està composta per dues visualitzacions complementàries. En la part superior es mostra la distribució dels registres segons l'*anomaly score* assignat pel model Isolation Forest. Els punts blaus representen observacions normals, mentre que els punts vermells representen anomalies detectades. La línia verda indica el llindar (*threshold*) utilitzat pel model per separar ambdós comportaments.

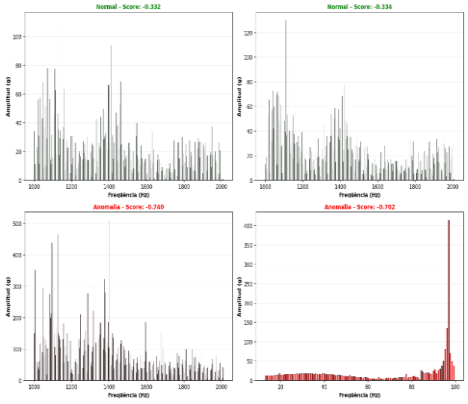


Fig. 4. Exemple de senyals classificats com a normals i anòmals.

En la part inferior es presenta un histograma de distribució dels anomaly scores per a ambdues classes. Aquesta visualització permet observar de manera clara la separació entre el comportament normal i anòmal, evidenciant que el model aconsegueix diferenciar adequadament entre ambdós grups.

Aquesta figura mostra exemples representatius de senyals classificats pel sistema com a normals (verd) i anòmals (vermell).

Els casos classificats com a normals presenten espectres més estables i amb una distribució d'energia més homogènia. En canvi, els casos anòmals mostren una major irregularitat en la distribució espectral, amb concentracions d'energia més marcades o comportaments menys uniformes.

Aquesta diferència qualitativa reforça la coherència del model, ja que les decisions no es basen únicament en criteris estadístics, sinó en patrons físicament interpretables del senyal de vibració.

7.3 Avaluació quantitativa del sistema

Per complementar la validació qualitativa del pipeline, s'ha realitzat una anàlisi quantitativa sobre el conjunt complet de registres processats en entorn local. Tot i treballar en un context no supervisat i sense etiquetes de fallada reals, aquestes mètriques permeten avaluar la consistència del sistema i la separació dels patrons detectats.

El model **Isolation Forest** s'ha executat sobre un total de **21.789 registres** de vibració, dels quals **17.431 (80,0%)** han estat classificats com a **normals** i **4.358 (20,0%)** com a **anòmals**. Aquesta distribució és coherent amb el rang de contaminació configurat durant l'entrenament del model i mostra un comportament estable sobre el conjunt global de dades.

Les anomalies detectades s'han agrupat posteriorment mitjançant **K-Means** amb **K = 4** per identificar patrons diferenciats de comportament vibracional. Per avaluar la qualitat de la segmentació s'ha calculat el **silhouette score** sobre el subconjunt d'anomalies i les etiquetes de clúster generades. Aquesta mètrica mesura, per a cada observació, com de semblant és al seu propi clúster en comparació amb el clúster veí més proper, i pren valors en el rang **[−1, +1]**: un valor proper a +1 indica que l'observació està ben assignada i clarament separada dels altres clústers, un valor proper a **0** indica que es troba en la frontera entre dos clústers, i un valor negatiu indica una possible assignació incorrecta. El model obté:

$$Silhouette\ Score = 0.4212$$

Aquest valor indica una separació moderada però consistent entre els clústers detectats, fet rellevant tenint en compte la variabilitat pròpia de les dades industrials i l'absència d'etiquetes de fallada reals.

La distribució dels clústers mostra una presència predominant de patrons compatibles amb **desalineació (31,2%)** i **degradació general (29,7%)**, mentre que els patrons relacionats amb **modes operatius d'alta freqüència (13,4%)** i **inestabilitat estructural (25,7%)** apareixen amb menor freqüència. Aquesta distribució és coherent amb el comportament esperat en maquinària rotativa industrial, on les alteracions en alineació i estabilitat mecànica acostumen a generar una part important de les anomalies vibracionals.

Pel que fa al model de classificació, les característiques amb més pes en el procés de decisió són la **desviació estàndard** de l'amplitud **(36,9%)**, **l'energia en alta freqüència (26,5%)** i **l'energia en banda mitjana (16,6%)**. Aquests resultats reforcen la coherència física del sistema, ja que la variabilitat del senyal i la distribució espectral de l'energia són indicadors directament relacionats amb processos de degradació mecànica i comportaments vibracionals anòmals.

Cal tenir en compte que el model **Gradient Boosting** s'entrena sobre les etiquetes generades pel propi procés de clustering. Per tant, els nivells elevats de confiança en la classificació reflecteixen principalment la coherència interna del pipeline i no una validació independent basada en etiquetes verificades per experts industrials.

Per contrastar la idoneïtat de l'algorisme escollit, s'ha aplicat **LOF (Local Outlier Factor)** [5] sobre el mateix conjunt de dades, mantenint idèntic paràmetre de contaminació (20,0%). Atès que el sistema opera en un entorn no supervisat sense etiquetes de fallada verificades, no és possible avaluar la detecció amb mètriques supervisades clàssiques (precisió, *recall* o F1).

En el seu lloc, s'utilitza el *silhouette score* del clustering posterior (**K-Means, K=4**) com a mètrica proxy indirecta: si les anomalies detectades per un algorisme formen clústers més ben separats, es pot inferir que l'algorisme ha identificat patrons estructuralment més coherents i diferenciats, la qual cosa és un indicador indirecte de millor qualitat de detecció en absència de *ground truth*. Sota aquest criteri, **LOF** obté un *silhouette score* de **0,23**, notablement inferior als **0,4212** d'**Isolation Forest**.

Cal subratllar que es tracta d'una comparativa indirecta, i que la seva validesa recolza en l'assumpció que anomalies ben separades en l'espai de característiques corresponen a patrons físics diferenciats; una validació independent amb etiquetes d'expert superaria l'abast d'aquest treball.

Aquests resultats son coherents amb la interpretació física descrita a la secció 6.6, on la variabilitat d'amplitud s'associa a inestabilitat estructural i l'energia en banda mitjana a possibles desalineacions, la tipologia de fallo més prevalent en el dataset.

8 DESPLEGAMENT DEL SISTEMA EN AWS SAGEMAKER

8.1 Arquitectura general del sistema en entorn cloud

El sistema de detecció i classificació d'anomalies es desplega en un entorn cloud utilitzant *Amazon Web Services* (AWS), amb **AWS SageMaker** com a plataforma principal de processament i **Amazon S3** com a repositori centralitzat de dades i models.

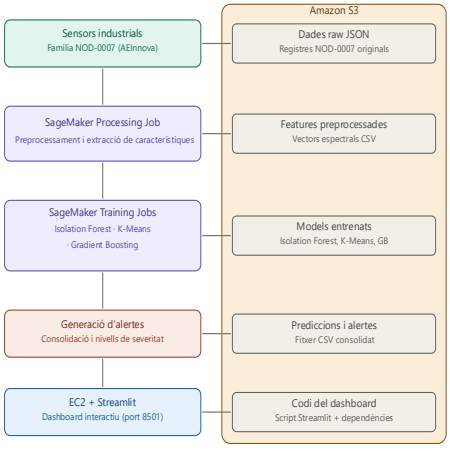


Fig. 5. Arquitectura del pipeline desplegat a AWS

L'arquitectura implementada manté la mateixa estructura funcional desenvolupada en entorn local, però adaptada a un esquema distribuït orientat a la gestió automatitzada del pipeline. El sistema s'organitza en **quatre blocs** principals:

- Emmagatzematge de dades i resultats mitjançant **Amazon S3**
- Preprocessament i execució del pipeline amb **AWS SageMaker**
- Generació i consolidació d'alertes industrials
- Visualització i explotació de resultats mitjançant un dashboard interactiu desplegat a **Amazon EC2**

Aquest enfocament permet separar les diferents etapes del sistema i facilita la traçabilitat dels experiments, la reutilització dels models i la reproducció dels resultats.

8.2 Organització de dades a Amazon S3

Totes les dades del sistema es centralitzen a Amazon S3, que actua com a repositori únic del pipeline. Les dades originals, els fitxers preprocessats, els models entrenats i les prediccions finals es mantenen separats mitjançant una estructura jeràrquica de directoris.

Aquesta organització permet conservar de forma independent cada etapa del processament i facilita tant la reutilització dels artefactes generats com la reproducció dels experiments.

Les principals categories d'emmagatzematge corresponen a les dades originals provinents dels sensors IoT, les dades preprocessades amb les característiques espectrals calculades, els models entrenats, les prediccions finals generades pel sistema i el codi del dashboard per al seu desplegament.



Fig. 6. Estructura jeràrquica del repositori central del projecte a Amazon S3

8.3 SageMaker Processing Jobs

La fase de preprocessament del dataset s'executa mitjançant **SageMaker Processing Jobs**. Aquesta etapa permet processar les dades directament des d'**Amazon S3** i generar un conjunt estructurat de característiques preparades per al modelatge.

El job de processament llegeix el dataset *raw* des de **S3**, calcula les característiques espectrals de cada registre (RMS, energia per bandes de freqüència, *crest factor*, freqüència dominant, entre d'altres) i emmagatzema el resultat en format CSV de nou al bucket. El dataset resultant conté els vectors de característiques que alimenten posteriorment els models d'aprenentatge automàtic.

L'ús de **Processing Jobs** permet desacoblar el preprocessament de la resta del pipeline i facilita l'execució automatitzada sobre grans volums de dades industrials.

8.4 SageMaker Training Jobs i execució del pipeline

Un cop preprocessades les dades, el sistema executa el pipeline de modelatge definit a les seccions anteriors. Aquesta etapa inclou l'entrenament dels models de detecció d'anomalies (**Isolation Forest**), clustering per tipologia de fallada (**K-Means**) i classificació supervisada (**Gradient Boosting**).

Un cop entrenats, els artefactes es guarden al bucket per a la seva reutilització en futures execucions. Aquest esquema permet mantenir separades les diferents etapes del sistema i facilita l'automatització completa del pipeline.

Cal tenir en compte que, per limitacions de quota de la plataforma AWS en comptes noves, l'entrenament i la inferència s'han executat en entorn local mantenint el codi i els artefactes completament compatibles amb **SageMaker Training Jobs** i **Batch Transform**. Els models resultants es pugen directament al bucket S3 i el pipeline és desplegable en entorn cloud complet un cop s'obtinguin les quotes corresponents.

8.5 Generació d'alertes industrials

A partir de les prediccions generades pel pipeline, el sistema construeix alertes estructurades orientades a entorns de manteniment predictiu. El procés d'inferència llegeix els models entrenats i el conjunt de dades preprocessat, genera una predicció per a cada registre i consolida els resultats en un únic fitxer emmagatzemat al bucket.

Cada alerta inclou l'identificador del dispositiu, el timestamp de detecció, el mode i eix de mesura, el tipus de fallada detectada amb la seva etiqueta descriptiva, el score de confiança de la classificació i el nivell de severitat derivat.

Addicionalment, el sistema aplica un postprocessament sobre les prediccions per reduir deteccions aïllades i consolidar els resultats per dispositiu i eix.

8.6 Desplegament del dashboard a Amazon EC2

Per fer els resultats accessibles de forma permanent, s'ha desplegat un dashboard interactiu sobre una instància Amazon EC2 de tipus **t3.micro** amb **Amazon Linux 2023**, a la mateixa regió que la resta de la infraestructura. El procés de desplegament s'ha automatitzat mitjançant un *script* Python que gestiona tots els passos necessaris sense intervenció manual.

En primer lloc, el codi del dashboard i les seves de pendències es pugen al bucket perquè la instància els pugui descarregar en l'arrencada.

S'ha optat per un rol IAM amb permisos de lectura mínims sobre el bucket, evitant credencials explícites al codi i seguint el principi de **mínim privilegi**.

L'accés al dashboard es restringeix al port de **Streamlit** a través d'un Security Group, sense exposar serveis addicionals de la instància. Les dependències s'aïllen en un entorn virtual per desacoblar l'aplicació dels paquets gestionats pel sistema operatiu, i Streamlit s'executa com a servei del sistema per garantir disponibilitat sense supervisió manual: la instància es recupera automàticament en cas de reinici o fallada.

Un cop completat el desplegament, el dashboard queda accessible des del navegador sense cap configuració addicional. La instància llegeix les dades directament des de **S3** en cada càrrega, de manera que qualsevol execució nova del pipeline es reflecteix al dashboard sense haver de redesplegar res.

9 RESULTATS

Els resultats del pipeline es presenten mitjançant un dashboard interactiu desenvolupat amb Streamlit i Plotly, desplegat sobre Amazon EC2 i connectat directament al bucket d'Amazon S3.

En cada càrrega, el dashboard descarrega el dataset de característiques i el conjunt de prediccions, i els combina en un únic *dataframe* de treball sobre el qual s'apliquen els filtres de l'usuari.



Fig. 7. Tauler de KPIs principals del dashboard.

La capçalera mostra cinc indicadors principals: el total de registres analitzats, el nombre i percentatge d'anomalies detectades, la distribució de severitat (HIGH / MEDIUM / LOW), el nombre de dispositius afectats i el recompte de tipologies de fallada identificades. Totes les mètriques es recalculen dinàmicament en funció dels filtres actius del sidebar.



Fig. 8. Distribució per tipologia de fallada, distribució per nivell de severitat i severitat per tipus de fallada.

La distribució per tipologia de fallada es presenta en un gràfic de donut que permet identificar quins patrons de degradació predominen en la flota analitzada.

La distribució de severitat, en barres verticals, reflecteix la jerarquia de risc calculada a partir del factor de cresta: els llindars absoluts —inferiors a **5,0** per a **LOW**, entre **5,0** i **8,0** per a **MEDIUM**, i iguals o superiors a **8,0** per a **HIGH**— s'han calibrat a partir dels percentils reals del conjunt d'anomalies detectades (**p50** i **p80**).

El gràfic de barres apilades de severitat per tipologia permet comparar quins tipus de fallada concentren major risc, creuant ambdues dimensions en una única visualització.

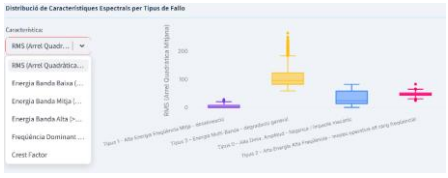


Fig. 9. Distribució de característiques espectrals per tipus de fallada mitjançant *boxplot* interactiu.

El *boxplot* interactiu mostra la distribució de cada característica espectral desglossada per tipologia de fallada, facilitant la interpretació física dels patrons detectats i la connexió amb els fenòmens mecànics subjacents. Un filtre interactiu permet seleccionar la característica a visualitzar, adaptant el gràfic dinàmicament a l'anàlisi d'interès.



Fig. 10. Heatmap d'anomalies per dispositiu i tipus de fallada

El *heatmap* dispositiu × tipologia de fallada creua els identificadors dels dispositius monitoritzats amb les tipologies detectades, revelant quins equips presenten major concentració d'anomalies i de quina naturalesa. Aquesta visualització és especialment útil per prioritzar les intervencions de manteniment sobre la flota.

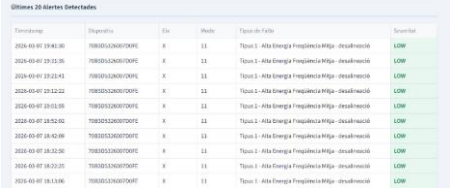


Fig. 11. Taula de les últimes 20 alertes detectades

La taula d'últimes alertes presenta les deteccions més recents amb codificació de color per nivell de severitat, oferint una vista operativa immediata per al personal tècnic. El *sidebar* incorpora filtres interactius per nivell de severitat i tipologia de fallada que afecten totes les visualitzacions de forma simultània.

10 CONCLUSIONS

Aquest treball ha desenvolupat un sistema complet de manteniment predictiu basat en dades de sensors IoT vibracionals, des del preprocessament del senyal fins al desplegament d'una eina de visualització en entorn cloud, cobrint totes les etapes del pipeline sense necessitat d'intervenció manual en la seva execució.

Els objectius s'han assolit parcialment dins les limitacions derivades de l'absència d'etiquetes reals. S'ha construït un pipeline automatitzat sobre **AWS SageMaker** capaç d'ingerir registres reals en format JSON, calcular característiques espectrals rellevants i entrenar models de detecció d'anomalies i classificació de fallades de forma desatesa.

L'aproximació híbrida adoptada —**Isolation Forest** per a la detecció, **K-Means** per a la segmentació de patrons i **Gradient Boosting** per a la classificació final— ha permès construir el sistema sense necessitat de dades etiquetades prèvies. Aquesta restricció, derivada de la naturalesa del dataset proporcionat per AEInnova, ha condicionat el rol del sistema: l'eina resultant actua com a **suport a la decisió**, identificant i classificant comportaments anòmals amb coherència física, però sense la precisió que requeriria un sistema de decisió autònom en un entorn industrial crític.

Des del punt de vista tecnològic, el desplegament sobre AWS ha demostrat la viabilitat d'un sistema d'aquest tipus en infraestructura cloud de cost reduït. El pipeline s'executa sobre instàncies **SageMaker** sota demanda i el dashboard opera sobre una instància **EC2** en funcionament continu, amb una arquitectura prou modular per incorporar futures millores sense redissenyar el sistema des de zero.

Finalment, el projecte ha posat de manifest la dificultat inherent a la validació de sistemes no supervisats en entorns industrials: en absència d'etiquetes reals, la qualitat dels resultats s'ha d'avaluar a partir de la coherència estadística dels clústers, l'estabilitat dels models entre execucions i la interpretabilitat física de les classificacions generades. Superar aquesta limitació és, precisament, el repte principal que orienta el treball futur.

11 TREBALL FUTUR

11.1 Incorporació de dades etiquetades

La limitació més significativa del sistema actual és **l'absència d'etiquetes de fallada reals**. Treballar amb un conjunt de registres on cada anomalia estigui associada a una fallada identificada i validada per un tècnic permetria substituir l'aproximació no supervisada per models supervisats entrenats sobre veritat de camp. Amb etiquetes reals, el sistema podria superar el rol d'eina de suport i convertir-se en una eina de decisió, emetent alertes amb un grau de confiança suficient per desencadenar accions de manteniment de forma autònoma. Aquesta transició requeriria establir un **procés de captura i validació d'etiquetes en col·laboració amb els tècnics d'AEInnova**, integrat en el flux operatiu existent.

11.2 Ingesta de dades en temps real

El sistema actual processa dades en *batch* sobre conjunts de registres acumulats. Una evolució natural seria incorporar un sistema d'ingesta en temps real que processi cada lectura dels sensors en el moment en què arriba, reduint la latència entre l'aparició d'un problema i la generació de l'alerta. Tecnologies com **Amazon Kinesis Data Streams** permetrien adaptar el pipeline actual a aquest esquema sense canvis estructurals profunds.

11.3 Sistema de notificació proactiva

Comentado [RA2]: Unificar tipus de lletra a tot el do

El dashboard actual requereix que l'usuari accedeixi activament a la interfície per detectar anomalies noves. Un sistema de notificació que enviés alertes per correu electrònic o integrat en plataformes de comunicació corporativa en el moment de detectar una anomalia d'alta severitat incrementaria significativament el valor operatiu de l'eina, especialment per a flotes amb molts dispositius distribuïts geogràficament.

12 REFERÈNCIES

[1] *Sensors*, "Predictive Maintenance in Motors," vol. 25, no. 8, art. 2590, 2025. Disponible a: https://www.mdpi.com/1424-8220/25/8/2590 (consultat el 2 de març de 2026)

[2] Amazon Web Services, "Amazon SageMaker Documentation," AWS Documentation. Disponible a: https://docs.aws.amazon.com/sagemaker/ (consultat el 5 d'abril de 2026)

[3] Amazon Web Services, "Amazon S3 User Guide," AWS Documentation. Disponible a: https://docs.aws.amazon.com/s3/ (consultat el 4 d'abril de 2026)

[4] Scikit-learn Developers, "Isolation Forest," Scikit-learn Documentation. Disponible a: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html (consultat el 3 d'abril de 2026)

[5] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," *Proc. ACM SIGMOD Int. Conf.*, pp. 93–104, 2000. https://doi.org/10.1145/342009.335388 (consultat el 15 d'abril de 2026)

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Comput. Surv.*, vol. 41, no. 3, 2009. https://doi.org/10.1145/1541880.1541882 (consultat el 20 d'abril de 2026)

[7] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001. https://doi.org/10.1214/aos/1013203451 (consultat el 21 d'abril de 2026)

[8] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, Wiley, 2015. (consultat el 21 d'abril de 2026)

[9] R. B. Randall, *Vibration-based Condition Monitoring: Industrial, Aerospace and Automotive Applications*, Wiley, 2011. (consultat el 22 d'abril de 2026)

APÈNDIX A — CARACTERITZACIÓ ESPECTRAL DELS CLÚSTERS DE FALLADA

La taula següent recull els valors mitjans de les principals característiques espectrals per a cadascuna de les quatre tipologies de fallada identificades pel sistema. Aquests valors complementen la interpretació física descrita a la secció 6.6 i evidencien la separació estadística entre els clústers generats pel model K-Means.

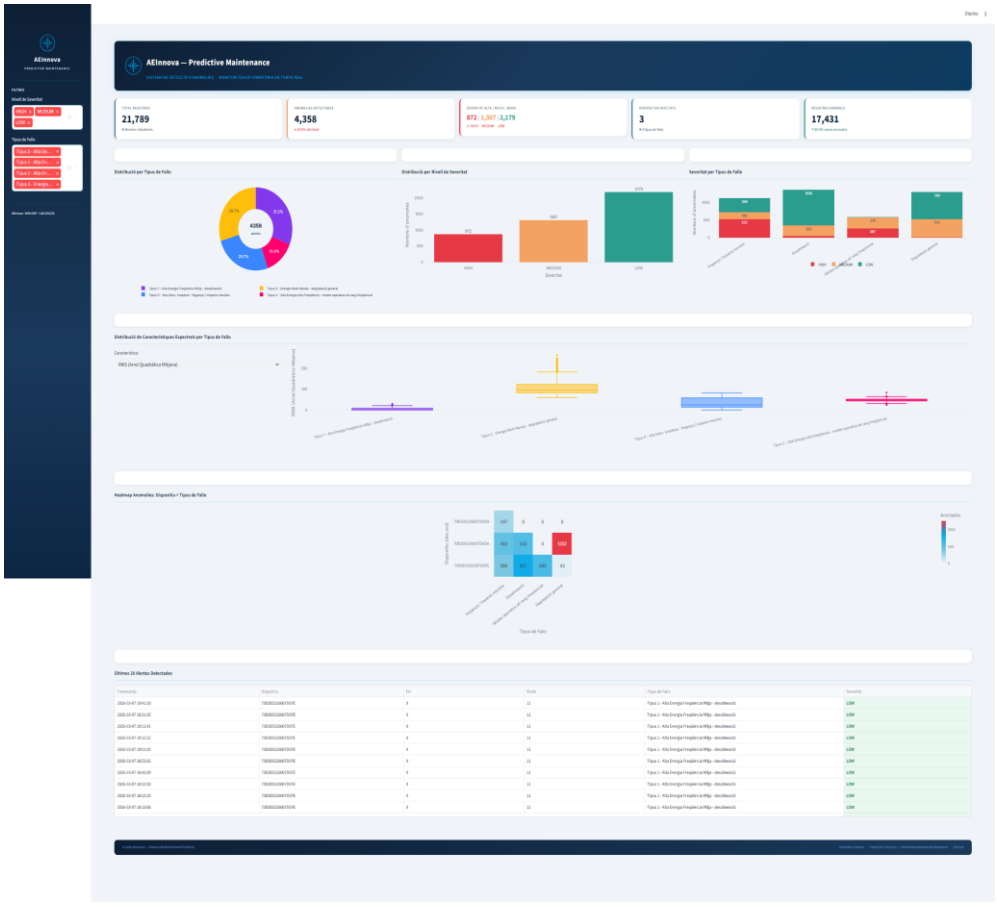
Característica	Tipus 0	Tipus 1	Tipus 2	Tipus 3
RMS	33,98	4,48	108,58	47,72
Desv. estàndard amplitud	27,22	3,55	70,39	43,49
Energia banda baixa	1.619	597	0	3.461
Energia banda mitja	2.495	4.477	0	242.578
Energia banda alta	397.652	1.195	3.045.770	13.003
Factor de cresta	8,33	3,79	4,72	8,34
Freqüència dominant (Hz)	1.077	52	1.585	63

- El Tipus 0 destaca per l'elevada desviació estàndard i el factor de cresta alt, compatibles amb una hipòtesi d'**inestabilitat estructural** per impactes.
- El Tipus 1 presenta l'energia en banda mitja més elevada i la freqüència dominant més baixa (~52 Hz), compatible amb una hipòtesi de **desalineació**.
- El Tipus 2 presenta valors nuls a les bandes baixa i mitja, amb energia concentrada exclusivament en alta freqüència, la qual cosa reflecteix el comportament de registres capturats en **modes operatius d'alt rang freqüencial**.
- El Tipus 3 concentra la major part de l'energia en la banda mitja amb distribució multibanda, compatible amb una hipòtesi de **degradació general**.

Comentado [RA3]: Ordena'ls

APÈNDIX B — DASHBOARD INTERACTIU DE VISUALITZACIÓ DE RESULTATS

El dashboard desenvolupat amb **Streamlit** i **Plotly** es desplega sobre una instància **Amazon EC2** i carrega les dades directament des d'**Amazon S3**. La figura següent mostra la interfície completa, que inclou KPIs, distribució de fallades i severitat, *boxplot* interactiu, heatmap per dispositiu i taula d'alertes recents, amb filtres laterals que afecten totes les visualitzacions simultàniament.



APÈNDIX C — RENDIMENT COMPUTACIONAL DEL PIPELINE

Per complementar l'avaluació funcional del sistema, s'ha analitzat el comportament computacional del pipeline sobre el conjunt de dades complet. La taula mostra el temps d'execució de cada etapa, mesurat en entorn local sobre els 21.789 registres del dataset.

Etapa	Temps d'execució
Preprocessament i extracció de característiques	4,9 s
Detecció d'anomalies (Isolation Forest)	0,2 s
Clustering (K-Means)	1,4 s
Classificació (Gradient Boosting)	6,0 s
Pipeline complet	12,5 s

El sistema és capaç de processar els **21.789 registres** en menys de **13 segons** en execució local. La major part del cost computacional es concentra a les etapes d'extracció de característiques espectrals **(39%)** i d'entrenament del Gradient Boosting **(48%)**, mentre que la detecció d'anomalies amb Isolation Forest representa únicament el **2%** del temps total gràcies a la seva eficiència en espais d'alta dimensionalitat. El clustering K-Means ocupa el **11%** restant.

Aquests resultats confirmen que l'arquitectura és compatible amb aplicacions reals de manteniment predictiu en modalitat *batch*. A més, el desplegament sobre AWS SageMaker permet paral·lelitzar el preprocessament i escalar el còmput a volums de dades superiors sense modificacions al codi del pipeline.