

Desarrollo y comparación de modelos de aprendizaje automático para la predicción del ciclo de vida útil restante (RUL) en motores turbofán: integración de Graph Neural Networks (GNN) y XGBoost.

Nombre del equipo:

Turbinas ML

Integrantes:

Julio César Posada Torres, David Álvarez Grisales, Sara Illatopa Marín, Jose Manuel López Arias

Marco teórico

El mantenimiento predictivo constituye un enfoque avanzado frente al mantenimiento preventivo y correctivo. Se basa en el análisis de datos de sensores para predecir el Remaining Useful Life (RUL) de un sistema antes de que ocurra una falla. Este paradigma busca anticipar fallos en componentes críticos, aumentando la seguridad y reduciendo costos asociados a paradas no planificadas.

La literatura científica muestra que técnicas como regresión multivariable, random forests, support vector machines (SVM) y redes neuronales recurrentes (LSTM, GRU) han alcanzado resultados destacados en la estimación del RUL de motores aeronáuticos y otros sistemas complejos. Estas metodologías permiten capturar relaciones no lineales y dependencias temporales en series multivariadas de sensores, lo cual es fundamental para modelar procesos de degradación progresiva.

El dataset NASA C-MAPSS (Commercial Modular Aero-Propulsion System Simulation), desarrollado por el Ames Research Center, se ha convertido en un referente académico y de investigación para el benchmarking de modelos de mantenimiento predictivo. Este dataset simula trayectorias de motores turbofán bajo diferentes condiciones de operación y modos de falla, proporcionando un escenario realista para el desarrollo de algoritmos de pronóstico.

Diversos estudios previos han trabajado con este conjunto de datos y han aportado enfoques relevantes:

Saxena & Goebel (2008) introdujeron el dataset C-MAPSS y demostraron su uso para la validación de algoritmos de pronóstico, destacando la importancia de métricas estandarizadas en la predicción de RUL.

Zheng et al. (2017) evaluaron modelos basados en long short-term memory (LSTM) y confirmaron que estas arquitecturas superan a métodos tradicionales al capturar dependencias de largo plazo en los datos de sensores.

Li et al. (2019) propusieron enfoques híbridos que combinan extracción manual de características con redes neuronales profundas, logrando mejoras en la precisión de las predicciones del RUL en distintos subsets de C-MAPSS.

Babu et al. (2016) aplicaron técnicas de deep learning end-to-end para estimación del RUL y mostraron que la reducción de preprocesamiento manual incrementa la robustez del modelo frente a ruido en los datos.

En conjunto, estos antecedentes demuestran que el problema de predicción de vida útil restante en motores turbofan ha sido abordado con éxito mediante técnicas de machine learning y deep learning, y que el dataset NASA C-MAPSS es un estándar consolidado para evaluar la efectividad de dichas aproximaciones.

Este marco teórico sustenta la viabilidad técnica del presente proyecto y refuerza su relevancia en aplicaciones de seguridad crítica en la industria aeronáutica y en otros sectores que dependen de la confiabilidad de sistemas complejos.

Descripción general de una industria, negocio o problema

- La industria aeronáutica es altamente regulada y dependiente de la **fiabilidad de motores a reacción**.
- El mantenimiento preventivo tradicional (basado en horas de vuelo) genera **altos costos y posibles reemplazos innecesarios**.
- El mantenimiento correctivo (esperar a que falle) es inviable por **riesgo de seguridad y costos de parada no planificada**.
- Se requiere **mantenimiento predictivo**: anticipar el fallo a partir de sensores y condiciones operativas, optimizando costos y mejorando seguridad.

Problema a resolver

Problema:

Estimar el **Remaining Useful Life (RUL)** de motores turbofán a partir de series temporales multivariadas de sensores.

Importancia:

- Incrementa la seguridad operacional al anticipar fallos.
- Optimiza costos de mantenimiento y logística.
- Permite maximizar la vida útil de componentes críticos sin comprometer la seguridad.
- Caso aplicable no solo en aeronáutica, sino en cualquier industria con sistemas complejos (energía, transporte, manufactura).

Impacto de la solución

La implementación de un modelo de predictivo para estimar la vida útil restante (RUL) en motores turbofán tiene un impacto significativo, específicamente en las siguientes dimensiones:

Seguridad operacional: anticipar fallas antes de que ocurran reduce la probabilidad de incidentes en aeronaves.

Optimización de costos: Se minimizan reemplazos prematuros y paradas no programadas, logrando un mejor aprovechamiento de componentes críticos.

Eficiencia logística: Permite planificar de manera precisa la adquisición de repuestos y la programación de mantenimientos.

Conjuntos de datos a utilizar

Fuente: [*NASA C-MAPSS Jet Engine Simulated Data*](#) (Propulsion CoE, Ames Research Center).

Este dataset simula la degradación de motores turbofan bajo diferentes condiciones operativas y modos de falla. Cada subset contiene trayectorias multivariadas de motores, divididas en entrenamiento (train) y prueba (test).

Subsets disponibles:

- **FD001**
 - 100 motores (train), 100 motores (test)
 - 1 condición operativa (nivel del mar)
 - 1 modo de falla: degradación del compresor de alta presión (HPC)
- **FD002**
 - 260 motores (train), 259 motores (test)
 - 6 condiciones operativas
 - 1 modo de falla: HPC
- **FD003**
 - 100 motores (train), 100 motores (test)
 - 1 condición operativa
 - 2 modos de falla: HPC y fan

- **FD004**
 - 248 motores (train), 249 motores (test)
 - 6 condiciones operativas
 - 2 modos de falla: HPC y fan

Estructura de los datos

Cada archivo contiene **series temporales multivariadas** de motores, con 26 columnas:

- 1. unit: ID del motor
- 2. time: Ciclo de operación
- 3–5. Variables de configuración operacional (ej. altitud, Mach, temperatura ambiente)
- 6–26. Sensores de turbina (temperatura, presión, rpm, flujo, etc.)

Características:

- Cada fila representa un ciclo de un motor.
- Los motores inician en estado “sano” y progresivamente desarrollan una falla.
- Los datos incluyen **ruido simulado**, replicando condiciones reales de sensores.

Archivos principales

- **train_FD00X.txt**
Trayectorias completas de motores hasta el fallo → se usan para entrenar el modelo.
- **test_FD00X.txt**
Trayectorias truncadas (antes del fallo real) → se usan para evaluar el modelo.
- **RUL_FD00X.txt**
Vector con el RUL verdadero de cada motor en test → sirve como ground truth para comparar predicciones.

Estado de apertura de los datos y licencias

El conjunto de datos **C-MAPSS Jet Engine Simulated Data** es **abierto y de acceso público** (*accessLevel: public*), mantenido por el **Propulsion CoE (PCoE) de NASA** y disponible en el portal oficial [NASA Open Data](https://data.nasa.gov/dataset/cmapss-jet-engine-simulated-data).

No requiere registro ni permisos especiales para acceder o descargarlo.

NASA, C-MAPSS Jet Engine Simulated Data, Propulsion CoE, 2008–2024.

Disponible en: <https://data.nasa.gov/dataset/cmapss-jet-engine-simulated-data>.

Historial de actualizaciones:

- Creado: **31 de marzo de 2025**
- Última modificación registrada: **May 29, 2025, 3:34 PM (UTC-05:00)**
- Última actualización visible en el portal: **hace 4 meses (2025)**

Lista de actividades

- Definición del problema y objetivos de negocio.
- Adquisición y comprensión de los datos NASA C-MAPSS.
- Análisis exploratorio de datos (EDA).
- Ingeniería de características y preprocesamiento.
- Selección y entrenamiento de modelos (Modelo clásico vs. Modelo novedoso)
- Comparación de modelos (Modelo clásico vs. Modelo novedoso)
- Validación y pruebas de desempeño.
- Visualización de resultados.
- Documentación y comunicación de hallazgos.

Métodos, modelos y tecnologías por materia

- Curso SI7009/SI6002 Aprendizaje Automático
 - Modelos predictivos para estimar el RUL (XGBoost y GNN)
 - Ajuste de hiperparámetros y métricas de desempeño.
 - Feature importance con XGBoost y Lasso
 - Filtrado con Autoencoder
- Curso SI7006/SI6003 Almacenamiento y Procesamiento de Grandes Datos
 - Gestión de datos en entornos distribuidos.
 - Uso de pipelines para procesamiento de series temporales multivariadas.
- Curso SI7007/SI6004 Visualización de Datos
 - Construcción de dashboards para monitoreo del RUL.
 - Visualización de trayectorias de motores y predicciones.
 - Integración de analítica visual en el prototipo final.

Cronograma general (octubre–noviembre)

Semana	Actividad principal
W1	Definición del problema, entendimiento del negocio
W2	Análisis exploratorio de datos (EDA)
W3	Preparación y preprocesamiento de datos
W4	Entrenamiento inicial de modelos
W5	Ajuste y optimización de modelos
W6	Validación

W7	Pruebas comparativas
W8	Integración de visualización, conclusiones y entrega

Restricciones del proyecto

En datos:

- Cobertura variable por subset (FD001/FD003 con 1 condición; FD002/FD004 con 6).
- **Ruido y simulación:** señales con ruido; dataset sintético, no idéntico a motores reales.
- **Etiquetas de RUL:** en test el fallo no ocurre; el RUL verdadero se entrega aparte.
- **Variabilidad entre motores:** desgaste inicial y diferencias de fabricación desconocidas.

En divulgación:

- **Confusión de alcance:** El dataset que se usa es uno simulado; **no promete** aplicabilidad directa a motores reales.

En código:

- Riesgo de sobreajuste de modelo.
- Falta de generalización a motores reales.
- Dependencia de poder computacional.

Operativas del equipo:

- **Tiempo/alcance:** Cubrir todos los subsets puede exceder la fecha límite de entrega.

Contribución o aporte al trabajo original

Como contribución principal, este proyecto propone la implementación de un **modelo no convencional basado en Graph Neural Networks (GNN)** para la predicción del ciclo de vida útil.

restante (RUL) en motores turbofán, utilizando el conjunto de datos C-MAPSS. El modelo propuesto se comparará con un modelo clásico de aprendizaje automático, con el objetivo de analizar las diferencias en desempeño, generalización y capacidad predictiva. Además, se incluye un análisis de importancia de variables mediante XGBoost y Lasso Regression, junto con un filtrado de características utilizando Autoencoders para la reducción de ruido y dimensionalidad, buscando mejorar la interpretabilidad y precisión de los resultados obtenidos.

Referencias bibliográficas

- Babu, G. S., Zhao, P., & Li, X. (2016). Deep convolutional neural network based regression approach for estimation of remaining useful life. International Conference on Database Systems for Advanced Applications, 214–228. Springer.
https://doi.org/10.1007/978-3-319-32025-0_17
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Li, X., Ding, Q., & Sun, J. Q. (2019). Remaining useful life estimation in prognostics using deep convolution neural networks. Reliability Engineering & System Safety, 172, 1–11.
<https://doi.org/10.1016/j.ress.2017.11.021>
- NASA Ames Research Center. (2008–2024). C-MAPSS Jet Engine Simulated Data. NASA Open Data. Disponible en: <https://data.nasa.gov/dataset/cmapss-jet-engine-simulated-data>
- Saxena, A., & Goebel, K. (2008). Turbofan engine degradation simulation data set. NASA Ames Prognostics Data Repository. NASA Ames Research Center. Disponible en: <https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/>

- Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017). Long short-term memory network for remaining useful life estimation. 2017 IEEE International Conference on Prognostics and Health Management (ICPHM), 88–95. IEEE.
<https://doi.org/10.1109/ICPHM.2017.7998311>