HACKATÓN TALENTO TECH – PLANTILLA OFICIAL DEL RETO

1. **NOMBRE DEL EQUIPO**
   Starlight

2. **NOMBRE DEL PROYECTO**
   Modelo de predicción de fallas en la maquinaria

3. **DESCRIPCIÓN GENERAL DEL PROYECTO (MÁX. 100 PALABRAS)**
   ¿Qué problema empresarial aborda su solución?
   ¿Qué variable están intentando predecir?
   ¿Qué tipo de empresa o sector imaginaron?

   Nuestra solución aborda el problema de ineficiencia en producción, prediciendo anticipadamente si una máquina fallará usando datos operativos como temperatura, vibración, consumo y productividad. La variable objetivo es fallo_detectado (fallo sí o no). Pensamos en implementarla principalmente en el sector manufacturero o industrial, donde cualquier parada no planificada impacta directamente la producción y calidad. Este enfoque permite reducir pérdidas operativas, optimizar mantenimiento y mejorar la continuidad de la línea productiva.

4. **DATASET UTILIZADO**

De dónde se obtuvo?

El data set fue proporcionado por la universidad y proviene de registros operativos de sensores de maquinaria industrial: temperatura, vibración, humedad, eficiencia operativa, consumo energético, tiempos de ciclo, entre otros indicadores.

2. ¿Qué variables contiene?

- **Timestamp**: sello de tiempo de cada lectura.
- **Variables operativas**: temperatura, vibración, humedad, consumo energía, eficiencia_porcentual, tiempo_ciclo, paradas.
- **Variables de control/identificación**: operador_id, maquina_id, producto_id, turno, cantidad producida, unidades defectuosas.
- **Variables de falla**: fallo_detectado (binaria), tipo_fallo (categórica).

3. ¿Cuántos registros y columnas tiene?

El tamaño de la data sets es de 6000 **registros** con 18 columnas

4. ¿Qué limpieza o preprocesamiento se aplicó?

1. EDA descriptivo para conocer la calidad de datos.
2. Auditoría y limpieza: detección de duplicados y eliminación de columnas con muchos nulos.
3. Llenado e imputación de datos faltantes según tipo de columna.
4. Tipificación de los datos.
5. Outliers detectados y tratados.
6. Escalado de numéricas sensible a modelo.
7. Codificación de variables categóricas.

8. Analisis predictivo.

9. Estadísticas generales con gráficos.

5. **MODELO PREDICTIVO IMPLEMENTADO**

1. ¿Qué algoritmo utilizaron?

Se evaluó tres modelos predictivos: **regresión logística**, **árbol de decisión** y finalmente **Random Forest**. Tomamos la decisión de usar Random Forest después de analizar la **matriz de correlaciones (heatmap)** entre las variables numéricas y los tipos de fallo. Vimos que muchas variables tenían correlaciones lineales débiles o moderadas, lo que indica que un modelo lineal simple (como regresión logística) podría no capturar patrones complejos en los datos. El Random Forest resulta ideal porque puede detectar **relaciones no lineales e interacciones entre variables**, reduce el sobreajuste gracias al uso de múltiples árboles y ofrece un ranking de importancia de variables consistente con lo identificado en el heatmap, validando su elección empíricamente.

¿Qué variable predijeron?

La variable objetivo fue **fallo_detectado**, binaria (1 = se detectó fallo, 0 = sin fallo). El modelo aprende a prever cuándo ocurre un fallo basándose en las variables operativas registradas (temperatura, vibración, humedad, consumo de energía, tiempo de ciclo, etc.).

Rendimiento del modelo

- El modelo predictivo implementado —un Random Forest Classifier— alcanzó un 90 % de precisión, lo que significa que 9 de cada 10 casos fueron clasificados

correctamente. Esto corresponde a un margen de error del 10 %, es decir, aproximadamente una de cada diez predicciones fue incorrecta (error de clasificación).

Supuestos realizados

- Se asumió que cada instancia con fallo_detectado = 1 corresponde a un evento de fallo real que el modelo debe aprender a reconocer.
- Se considera que los datos fueron registrados correctamente y sincronizados: sin desfasajes temporales entre sensores y etiquetado de fallos.
- Se asume que las variables operativas reflejan el estado real de la máquina en tiempo de registro, sin errores de captura.
- El modelo también presupone que la distribución de fallos en los datos históricos es representativa del entorno futuro (predicción en condiciones similares).

6. **VISUALIZACIONES**

¿Qué tipo de gráficos generaron?
¿Qué muestran las gráficas?
¿Cómo se comparan las predicciones con los datos reales?
¿Qué herramienta usaron? (Matplotlib, Seaborn, Power BI, etc.)

En el análisis de datos, se utilizan diversos tipos de gráficos que permiten una mejor comprensión de la información. A continuación, se describen algunos de ellos:

Boxplot

Permite visualizar los cuartiles de los datos.
Facilita el análisis de los outliers (valores atípicos).

Gráfica de Barras Agrupadas

Muestra la distribución de las variables numéricas.
Ayuda a comparar diferentes categorías de manera efectiva.

Diagrama de Comparación de Métricas por Modelo

Se utilizó para comparar diferentes métricas como Accuracy, Precision, Recall y F1-score entre varios modelos.

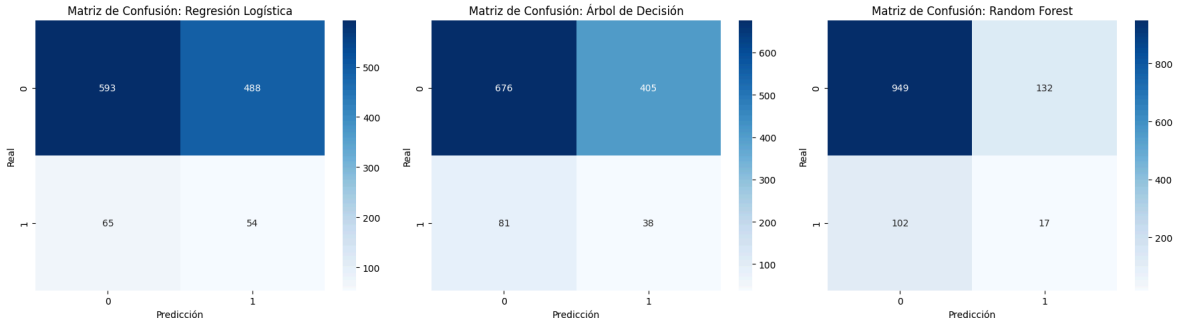
En la gráfica se puede observar claramente el rendimiento de modelos como la Regresión Logística, el Árbol de Decisión y Random Forest.

Esta gráfica proporciona un análisis visual que permite entender mejor las diferencias en el rendimiento de los modelos evaluados.

Sin embargo las siguientes son las que consideramos mas importantes:

Matriz de confusión (*Confusion Matrix*)

- **Qué muestra**: cuántas instancias fueron clasificadas correctamente o incorrectamente: verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.
- **Comparativa**: permite visualizar directamente cómo las predicciones se alinean con los valores reales, revelando patrones de error.
- **Herramienta**: generada con scikit-learn + matplotlib o seaborn a través de librerías como *scikit-plot*.



Evaluación de métricas de los modelos

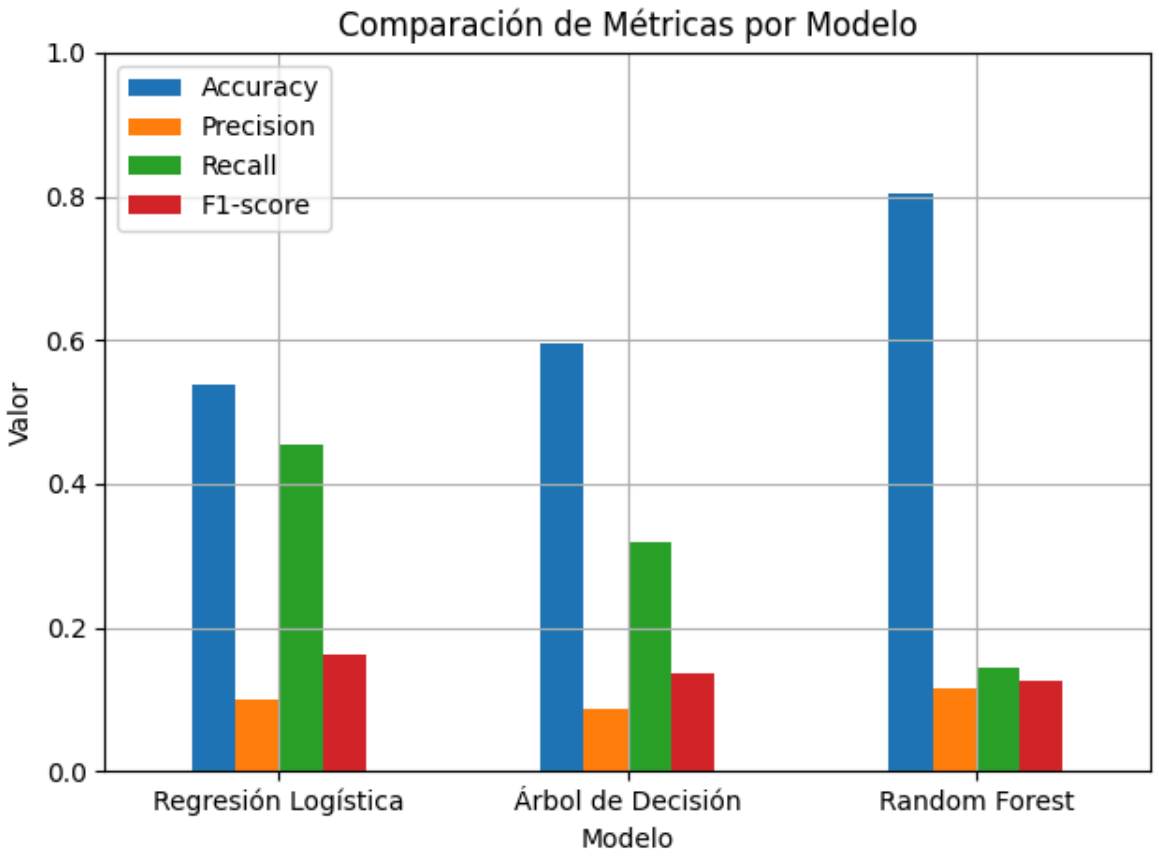
- **Métricas Representadas:**
- Accuracy
- Precision
- Recall
- F1-score
- **Modelos Comparados:**
- Regresión Logística
- Árbol de Decisión
- Random Forest
- **Comparación de Predicciones con Datos Reales**

- Se observa que el modelo de Random Forest tiene las mejores métricas en accuracy, precision, y recall, seguido por el Árbol de Decisión y finalmente la Regresión Logística.

- El modelo de Random Forest supera significativamente a los otros dos, especialmente en F1-score, lo que indica un mejor balance entre precision y recall.

- **Herramienta Usada**

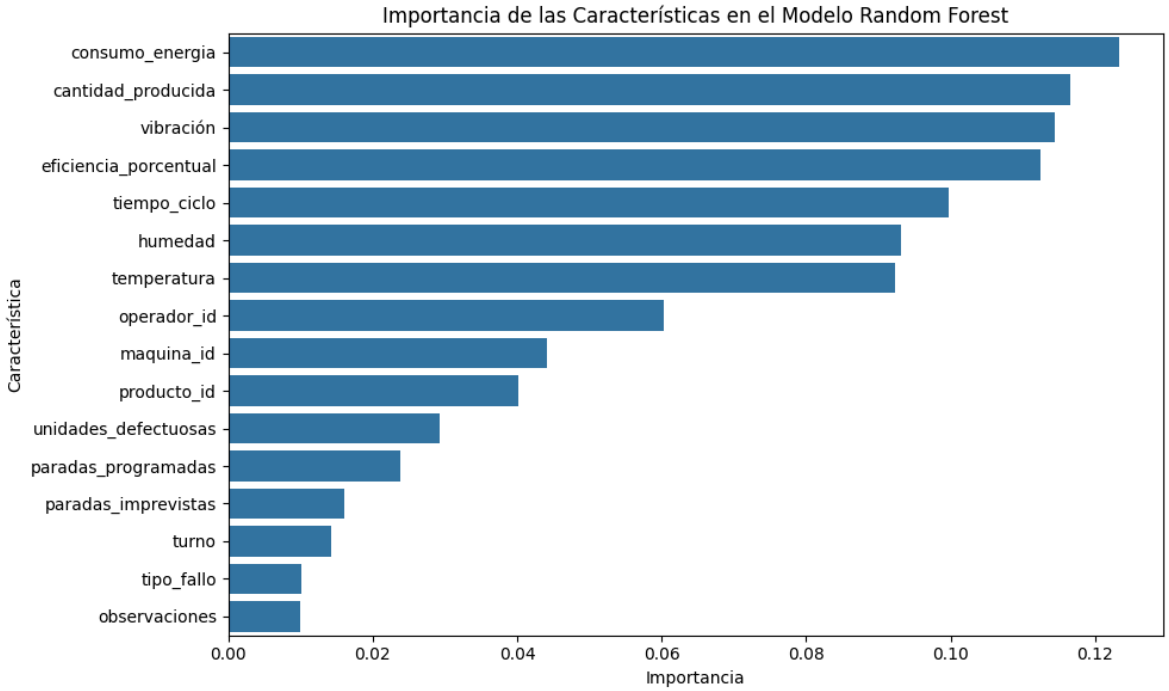
- La gráfica parece estar creada con Matplotlib, dada la estética y el formato de visualización.



Importancia de variables

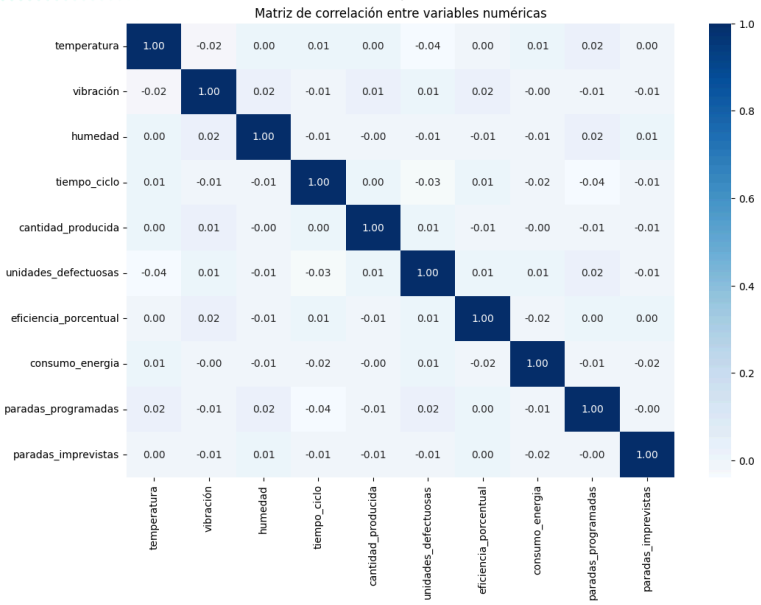
- **Qué muestra**: ranking de las variables más influyentes en las predicciones del modelo, basado en *mean decrease in impurity* o importancia de permutación.
- **Comparativa**: permite confirmar si las variables con mayor importancia coinciden con las que previamente mostraron alta correlación en el heatmap.

- **Herramienta**: atributo feature_importances_ del Random Forest y visualización con matplotlib o seaborn.



Matriz de correlacion

- **Qué muestra**: La matriz de correlación permite identificar la relación lineal entre las variables numéricas del dataset.
- **Comparativa**: ideal para ver el nivel global de coincidencia y dispersión entre predicción y realidad.
- **Herramienta**: matplotlib.



7. **RECOMENDACIONES ESTRATÉGICAS**
   Redacten al menos 2 recomendaciones empresariales derivadas del análisis. Deben estar justificadas con base en las predicciones y gráficas.

Se recomienda implementar un sistema que genere alertas automáticas cuando se detecten patrones predictivos de fallo (por ejemplo, picos de vibración o temperatura fuera de rango). Estas alertas activan acciones de mantenimiento oportuno y evitan paradas inesperadas. Si tuviera información sobre el tipo de fallo, podríamos no sólo predecir el fallo, sino también sugerir medidas específicas según el tipo (como lubricación si el fallo es desgaste o alineación si es desbalance), haciéndolo un enfoque mucho más efectivo y detallado en el tiempo.

Dado que escogimos el modelo de random forest le recomendamos a la empresa tener en cuenta el consumo de energía porque es una variable muy significativa, asi como la cantidad producida, vibración y eficiencia porcentual para predecir futuros fallos.

8. **VALOR DIFERENCIAL DE LA SOLUCIÓN (MÁX. 80 PALABRAS)**
   ¿Qué hace único su enfoque?
   ¿Cómo podría escalarse o mejorarse esta solución en una siguiente etapa?

Se combina de forma única un modelo Random Forest con análisis visual previo (heatmap de correlaciones y gráfico de importancia de variables), lo que permite entender desde el primer día qué factores realmente influyen en los fallos de las máquinas. Esto facilita decisiones claras y predictivas. En una siguiente etapa, podríamos escalarla incorporando sensores con procesamiento local para obtener alertas en tiempo real y crear gemelos digitales que simulen escenarios, permitiendo anticipar fallos con mayor precisión y prescribir acciones óptimas.

9. **ENLACES DE ENTREGA**
   - Repositorio de código (GitHub, GitLab, etc.):
   - Notebook / Dashboard interactivo (si aplica):
   - Video demo (opcional, máx. 3 minutos):

10. **PITCH FINAL**
    - Estructura sugerida para la presentación:
    - Problema y variable a predecir (30s)
    - Análisis de datos y modelo implementado (1 min)
    - Visualizaciones clave y resultados (1 min)
    - Recomendaciones y posible impacto (30s)
    - Cierre creativo o llamado a la acción (30s)

11. **ENTREGABLES ESPERADOS**
    - Dataset limpio y documentado
    - Notebook o script de análisis
    - Mínimo 2 visualizaciones
    - Reporte breve con 2 recomendaciones estratégicas
    - Repositorio en GitHub o ZIP organizado
    - Pitch final en vivo (opcional: video corto)