Diagnóstico de falla de rodamientos con Machine Learning

Cesar Vicuña H., Johan Callomamani B., Adolfo RamonP., Paul Cusi H.,Jairo Pinedo T.
Universidad Nacional de Ingeniería
(Maestría deInteligencia Artificial, Machine Learning)

***Resumen*.**

El objetivo principal fue la implementación de modelos supervisados de machine learning para poder clasificar y predecir los tipos de fallas que pueden tener un rodamiento (falla pista interna, pista externa, elemento rodante), también evaluación del mejor tratamiento de datos para optimizar resultados para muestreos de vibración en el tiempo, para ello se monitoreó vibraciones en un rodamiento el cual presenta un inicio normal y en el tiempo es llega a la falla, con un equipo acelerómetros ICP de cuarzo de alta sensibilidad
PCB 353B33.
Se aplicaron 04 modelos supervisados que fueron Decision tree
Logistic regression, Random forest y Extreme gradient boosting (XGBoost) y se han utilizado 3 datasets distintos:
Dataset dominio en el tiempo (20,480 variables + 1 variables objetivo, 6706 registros)
Dataset dominio en frecuencias (10,240 variables + 1 variables objetivo, 6706 registros)
Dataset características estadísticas de señales (9 variables + 1 variables objetivo, 6706 registros)

Keywords— Machine Learning, rodamientos, falla de rodamientos, vibración, modelos supervisados. (palabras claves)

I. INTRODUCCIÓN

El diagnóstico de fallas en rodamientos es un aspecto crítico en el mantenimiento predictivo de sistemas rotativos, ya que estas fallas pueden llevar a costosas paradas de maquinaria y a fallos catastróficos si no se detectan a tiempo. Los rodamientos son componentes esenciales en diversas aplicaciones industriales y su desempeño confiable es vital para la operación eficiente de equipos como motores, generadores y sistemas de transmisión [1]. Con el avance de las tecnologías de machine learning, se han desarrollado diversos modelos para mejorar la precisión y eficiencia en la detección de estas fallas. Entre estos modelos, los supervisados han demostrado un desempeño destacado debido a su capacidad para aprender patrones complejos a partir de datos etiquetados.

En este trabajo, se investiga el uso de cuatro modelos supervisados para el diagnóstico de fallas en rodamientos: Decision Tree, Logistic Regression, Random Forest y Extreme Gradient Boosting (XGBoost). Cada uno de estos modelos ofrece ventajas únicas en términos de interpretabilidad, capacidad de manejo de datos y precisión predictiva.

El modelo Decision Tree es conocido por su simplicidad e interpretabilidad, facilitando la comprensión de las decisiones del modelo y la identificación de las características más relevantes para la predicción [2]. Este modelo divide el espacio de datos en regiones homogéneas utilizando reglas de decisión jerárquicas, lo que permite una visualización clara de cómo se toman las decisiones. Además, su capacidad para manejar tanto datos numéricos como categóricos lo hace versátil en diversas aplicaciones.

La regresión logística, por otro lado, es una técnica estadística que permite la clasificación binaria y ha sido ampliamente utilizada en diagnósticos médicos y otras aplicaciones [3]. Su fundamento matemático se basa en la modelación de la probabilidad de un evento mediante una función logística, lo que permite una interpretación probabilística de los resultados. Aunque tradicionalmente se ha utilizado para problemas de clasificación binaria, sus extensiones permiten abordar problemas multicategóricos, ampliando su aplicabilidad.

Random Forest, una extensión del Decision Tree, mejora la precisión del modelo al reducir el riesgo de sobreajuste mediante la creación de múltiples árboles de decisión y la combinación de sus resultados [4]. Este enfoque de ensemble learning no solo aumenta la robustez del modelo, sino que también mejora la generalización a datos no vistos. Cada árbol en el bosque se entrena con un subconjunto diferente de los datos y, al final, se realiza un voto mayoritario para la clasificación final, lo que reduce el sesgo y la varianza del modelo.

Finalmente, XGBoost es un algoritmo de boosting altamente eficiente y preciso que ha ganado popularidad en competiciones de machine learning y aplicaciones industriales debido a su capacidad para manejar grandes volúmenes de datos y su excelente rendimiento predictivo [5]. Este modelo se basa en la construcción secuencial de árboles de decisión donde cada nuevo árbol intenta corregir los errores cometidos por los árboles anteriores. La implementación eficiente de XGBoost permite una rápida convergencia y una excelente capacidad de manejo de características ausentes y desbalanceo de clases.

El uso de estos modelos en el diagnóstico de fallas en rodamientos ha sido investigado en varios estudios previos. Por ejemplo, Li et al. [6] demostraron que el uso de Random Forest en la detección de fallas mejora significativamente la precisión comparado con modelos más simples. Zhang et al. [7] exploraron el uso de XGBoost en diagnósticos industriales y encontraron que su capacidad para manejar datos desbalanceados y su eficiencia computacional lo hacen ideal para aplicaciones en tiempo real. Además, el estudio de Wuest et al. [8] resaltó la importancia de la selección adecuada de características y su impacto en la precisión de los modelos de machine learning aplicados a la detección de fallas. Otros estudios han explorado la combinación de estos modelos con técnicas de preprocesamiento de señales y extracción de características para mejorar aún más la precisión del diagnóstico [9], [10].

En las siguientes secciones, se describen en detalle la metodología de implementación de estos modelos, los conjuntos de datos utilizados para el entrenamiento y prueba, y se comparan los resultados obtenidos para determinar el modelo más adecuado para el diagnóstico de fallas en rodamientos.

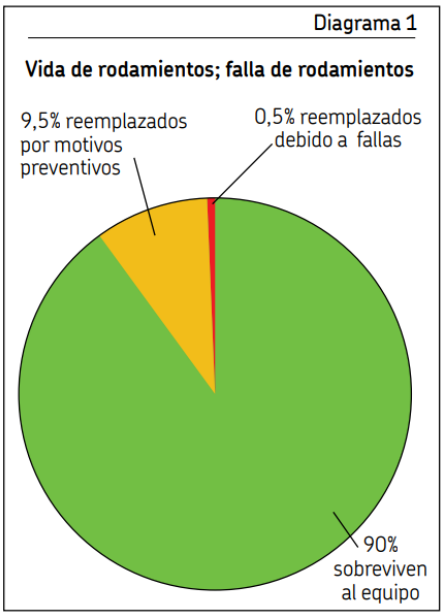


Figura 1. Referencias Daños de rodamientos [SKF]

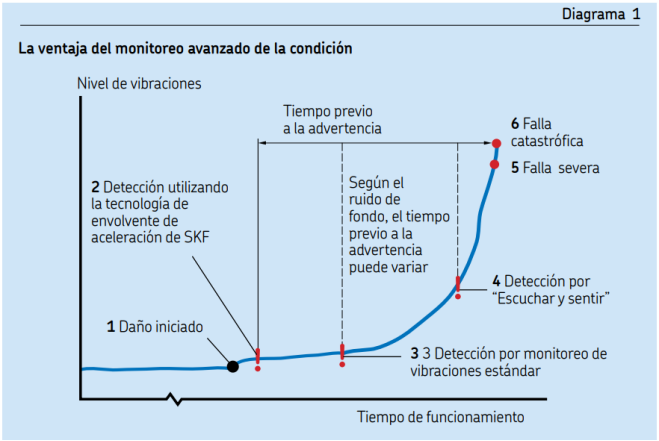


Figura 2. Referencias Daños de rodamientos [SKF]

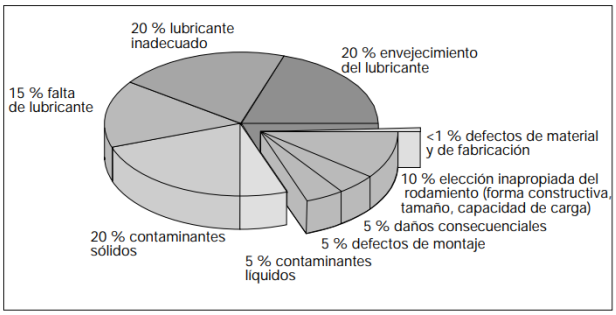


Figura 3. Referencia de Averías de los rodamientos [FAG]

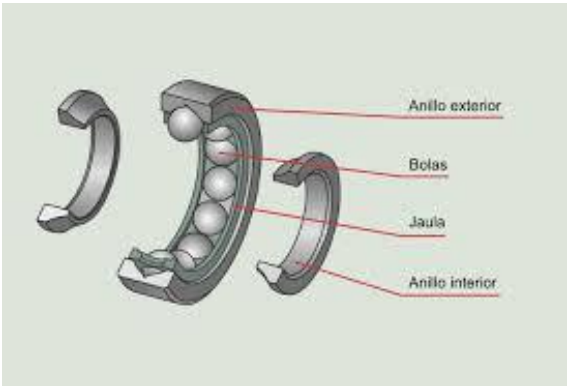


Figura 4. Elemento rodante.

Figura 5. Cojinete con rodamiento

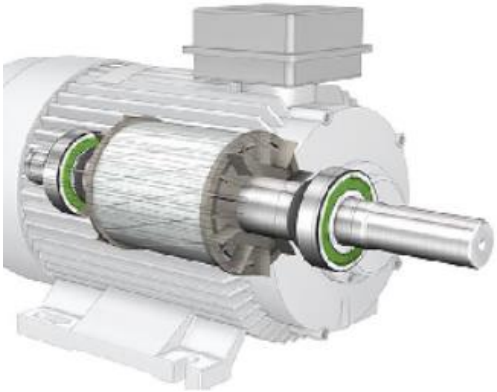


Figura 6. Uso del elemento rodante.

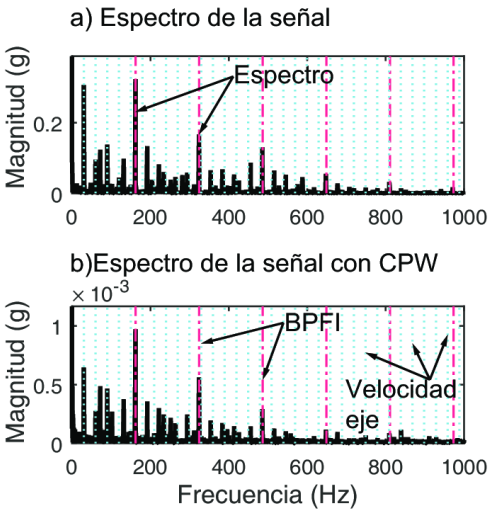


Figura 7. Fallas de rodamientos.

II. MARCO TEÓRICO

a. Diagnóstico de Fallas en Rodamientos

El diagnóstico de fallas en rodamientos es esencial en el mantenimiento predictivo de equipos industriales. Los rodamientos son componentes críticos en diversas máquinas rotativas, y su fallo puede causar paradas imprevistas y costosas reparaciones. Tradicionalmente, se han utilizado técnicas basadas en el análisis de vibraciones y señales acústicas para detectar fallas en rodamientos [11]. Sin embargo, la precisión y eficiencia de estas técnicas pueden ser limitadas por factores como el ruido ambiental y la complejidad de las señales.

b. Machine Learning en el Diagnóstico de Fallas

La aplicación de machine learning en el diagnóstico de fallas ha ganado considerable atención debido a su capacidad para manejar grandes volúmenes de datos y descubrir patrones complejos. Los algoritmos de machine learning pueden ser clasificados en supervisados y no supervisados. En el contexto de diagnóstico de fallas, los modelos supervisados son ampliamente utilizados debido a su capacidad para aprender a partir de datos etiquetados [12]. Estos modelos incluyen técnicas como árboles de decisión, regresión logística, bosques aleatorios y XGBoost.

c. Árboles de Decisión

Los árboles de decisión son algoritmos de clasificación que dividen iterativamente el espacio de características mediante reglas de decisión basadas en los valores de los atributos. Este enfoque permite una visualización clara y una fácil interpretación de los resultados [13]. Los árboles de decisión son eficaces para manejar tanto datos numéricos como categóricos, y son particularmente útiles en aplicaciones donde la interpretabilidad es crucial.
La función de decisión en un árbol de decisión puede ser representada como:

$$f(x) = \sum_{i-1}^{N} w_i \cdot I(x \in R_i)$$

donde *wi* es el peso asignado a la región *Ri, y I(·)* es una función indicadora que es igual a 1 si x pertenece a *Ri* y 0 en caso contrario.

d. Regresión Logística

La regresión logística es una técnica estadística utilizada para la clasificación binaria. Modela la probabilidad de un evento en función de una o más variables independientes utilizando la función logística [14]. La regresión logística es apreciada por su simplicidad y por proporcionar una salida probabilística, lo que permite una interpretación directa de las predicciones.
La función de probabilidad en la regresión logística se define como:

$$P(y = 1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$$

donde β0,β1,…,βn son los coeficientes del modelo que se estiman a partir de los datos.

e. Bosques Aleatorios

El algoritmo de bosques aleatorios es una extensión del árbol de decisión que mejora la precisión al reducir el riesgo de sobreajuste. Este modelo de ensemble learning crea múltiples árboles de decisión utilizando diferentes subconjuntos de datos y características, y combina sus resultados mediante un voto mayoritario [15]. Los bosques aleatorios son robustos y eficaces en la reducción de la varianza del modelo, lo que mejora su capacidad de generalización.
El proceso de combinación de resultados en bosques aleatorios puede ser expresado como:

$$\hat{y} = \text{majority\_vote}\{h_1(x), h_2(x), \ldots, h_k(x)\}$$

donde *hi(x)* representa el resultado del i-ésimo árbol de decisión.

f. XGBoost

XGBoost es un algoritmo de boosting altamente eficiente y preciso que ha demostrado un rendimiento superior en varias competiciones de machine learning. Este algoritmo construye secuencialmente árboles de decisión, donde cada nuevo árbol intenta corregir los errores de los árboles anteriores [16]. XGBoost es conocido por su capacidad para manejar grandes volúmenes de datos y por su eficiencia computacional.
El objetivo en XGBoost se define como la minimización de una función de pérdida regularizada:

$$\mathcal{L}(t) = \sum_{i-1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k-1}^{t} \Omega(f_k)$$

donde *l* es la función de pérdida, $\hat{y}_i^{(t)}$ es la predicción del modelo en la iteración t, y Ω es el término de regularización que penaliza la complejidad del modelo.

g. Extracción de Características

La precisión de los modelos de machine learning depende en gran medida de la calidad de las características extraídas de los datos. En el contexto del diagnóstico de fallas en rodamientos, las características comúnmente extraídas incluyen medidas estadísticas de señales de vibración, transformadas de Fourier, y análisis de tiempo-frecuencia [17]. Técnicas avanzadas como la Transformada Wavelet y la Descomposición de Modos Empíricos también se utilizan para mejorar la calidad de las características [18].

h. Preprocesamiento de Datos

El preprocesamiento de datos es un paso crucial en la implementación de modelos de machine learning. Incluye la normalización de datos, la eliminación de ruido, y el manejo de datos faltantes. La normalización asegura que las características tengan escalas comparables, mientras que la eliminación de ruido mejora la relación señal-ruido [19]. El manejo de datos faltantes puede realizarse mediante imputación o eliminación, dependiendo de la cantidad de datos faltantes y su importancia.

i. Evaluación de Modelos

La evaluación de modelos de machine learning se realiza utilizando métricas como precisión, recall, F1-score y área bajo la curva ROC (AUC-ROC). Estas métricas proporcionan una

visión completa del rendimiento del modelo en términos de su capacidad para predecir correctamente las clases positivas y negativas [20]. La validación cruzada es una técnica comúnmente utilizada para evaluar la robustez del modelo y evitar el sobreajuste.

j. **Aplicaciones Industriales**

Los modelos de machine learning para el diagnóstico de fallas en rodamientos han sido implementados en diversas aplicaciones industriales, incluyendo la manufactura, la energía eólica, y el transporte. Estos modelos han demostrado mejorar significativamente la eficiencia del mantenimiento predictivo y reducir los costos operativos [21]. La integración de estos modelos con sistemas de monitoreo en tiempo real permite la detección temprana de fallas y la planificación de intervenciones de mantenimiento más efectivas [22].

III. IDENTIFICACIÓN DEL PROBLEMA

La falla de rodamientos es una de las principales causas de fallos en maquinaria industrial. Un rodamiento defectuoso puede causar paradas imprevistas, lo que resulta en costosos tiempos de inactividad y posibles daños adicionales a la maquinaria.

Se requiere un sistema eficiente para monitorear y controlar la evolución de las fallas en los rodamientos. Esto permitirá programar el reemplazo de los rodamientos de manera planificada, minimizando el impacto en la producción y evitando paradas imprevistas.

Para la presente investigación se instalaron acelerómetros ICP de cuarzo de alta sensibilidad PCB 353B33 en la carcasa del cojinete (dos acelerómetros para cada cojinete [ejes x e y] para el conjunto de datos.

Figura 8. Sensor de vibración.

IV. SOLUCIÓN

Para determinar el tipo de falla de un rodamiento vamos a realizar un modelo de clasificación, con lo cual vamos a poder determinar el tipo de falla que presenta el rodamiento, los modelos que se van utilizar son:

1) Árboles de decisión.
2) Regresión logística
3) Random forest
4) Clasificador XGBoost

V. DESCRIPCIÓN DEL CONJUNTO DE DATOS

La información utilizada en el presente estudio abarca datos obtenidos de la toma de vibración de 3 datasets distintos:

- Dataset dominio en el tiempo (20,480 variables + 1 variables objetivo, 6706 registros)
- Dataset dominio en frecuencias (10,240 variables + 1 variables objetivo, 6706 registros)
- Dataset características estadísticas de señales (9 variables + 1 variables objetivo, 6706 registros)

Para la presente investigación se utilizaron Dataset en dominio del tiempo (20,480 variables + 1 variables objetivo, 6706 registros)

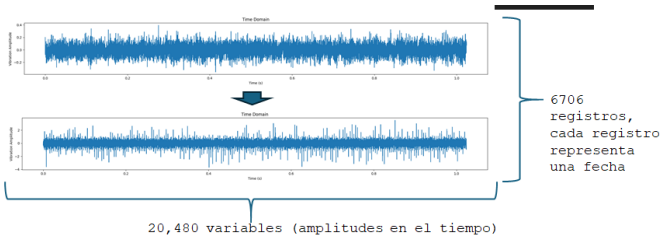


Figura 8. Distribución de datos de vibración en el dominio del tiempo

Dataset dominio en el tiempo (20,480 variables + 1 variables objetivo, 6706 registros.

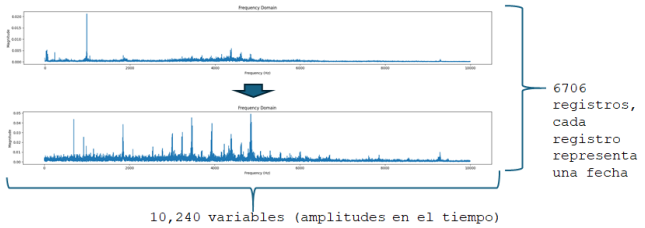
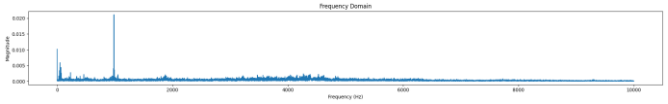


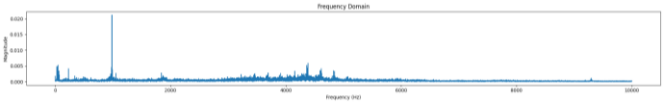
Figura 9. Distribución de datos de vibración en el dominio de frecuencias.

Variable objetivo se asigna estudiando el dominio de frecuencias, se caracteriza la falla y se agrega la columna clase al dataset.

Rodamiento Normal:



Rodamiento Inicio de falla.



Rodamiento con falla catastrófica

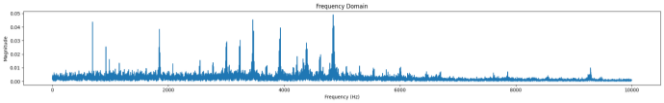


Figura 10. Identificación de la falla del rodamiento, en el dominio de frecuencias.

Dataset características estadísticas de señales (9 variables + 1 variables objetivo, 6706 registros)

Caracteristicas de las variables

Para el tratamiento y extracción de las características estadisticas de las señales se realiza un calculo por cada registro transformando un data set de 20480 variables independientes y 6706 registros en 9 variables independientes y 6706 registros.

. Máximos
Se considera el máximo valor de vibración por registro.

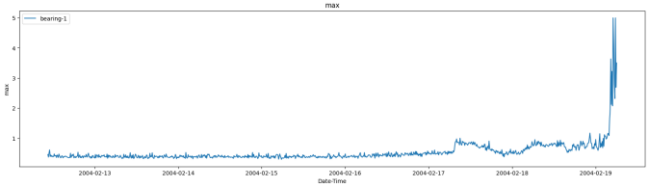


Figura 11. Valores máximos de vibración de un rodamiento

. Desviación estándar
Para la desviación estándar se considera:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

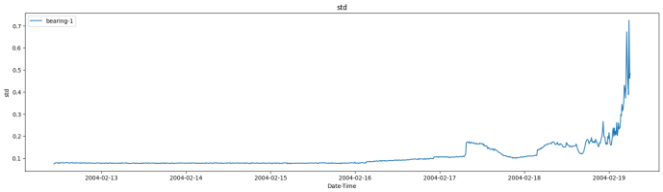


Figura 12. Valores desviación estándar de vibración de un rodamiento

. Kurtosis
Para el calculo de kurtosis se considera:

$$\mathrm{K} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \overline{x})^4}{\sigma^4}$$

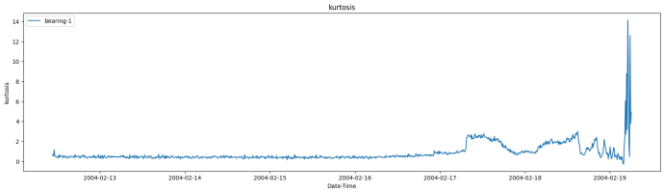


Figura 13. Valores de kurtosis de vibración de un rodamiento

. Mínimo
Se considera el mínimo valor de vibración en un registro

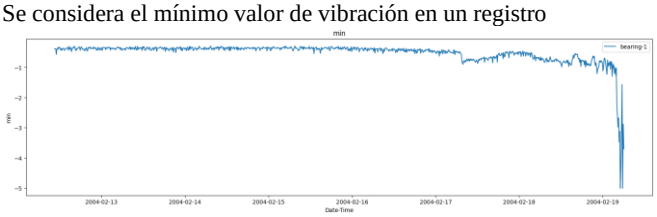


Figura 14. Valores mínimos de vibración de un rodamiento

. Media cuadrática
Para el cálculo de rms se considera:

$$x_{rms} = \sqrt{(\frac{1}{N})\sum_{i=1}^{N}(x)^2}$$

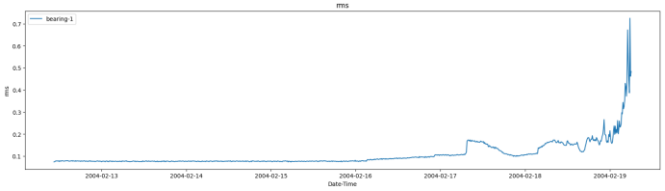


Figura 15. Valores de rms de vibración de un rodamiento

. Factor de cresta
Para el calculo de factor de cresta se considera:

$$x_{crest} = \frac{\text{max value}}{\mathrm{x_{rms}}}$$

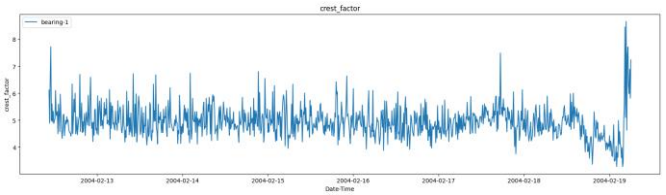


Figura 16. Valores factor de cresta de vibración de un rodamiento

. Media
Para el cálculo de la media absoluta se considera:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N}|x_i|$$

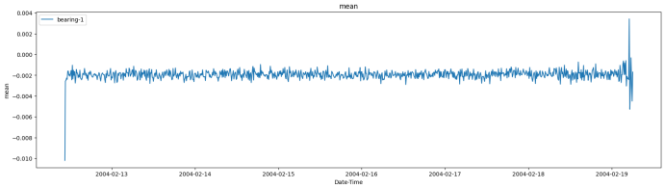


Figura 17. Valores de media de vibración de un rodamiento

. Sknewness
Para el calculo de sknewness se considera:

$$\mathrm{Sk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \overline{x})^3}{\sigma^3}$$

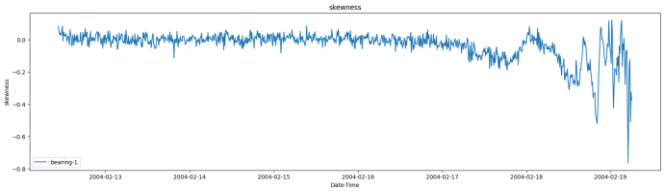


Figura 18. Valores de sknewness de vibración de un rodamiento

. Factor de forma
Para el cálculo de factor de forma se considera:

$$\frac{x_{rms}}{\overline{x}}$$

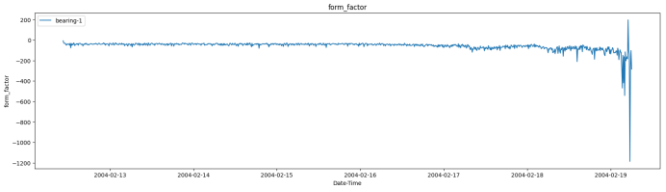


Figura 19. Valores de factor de forma de vibración de un rodamiento

VI. PREPARACIÓN DE LA INFORMACION

Se realizó con el programa Python

1. Transformación a dominio de frecuencias
2. Búsqueda del inicio de falla
3. Asignación de la variable objetivo
4. Transformación características de señales

VII. DESCRIPCIÓN DE TÉCNICAS UTILIZADAS

Algoritmos de machine learning:
Para los 3 datasets utilizamos los siguientes modelos de machine learning supervisados, decision tree, logistic regression, random forest, extreme gradient boosting (XGBoost).
No aplicamos escalamiento, para las dataset del dominio del tiempo y frecuencia, si para el dataset de caracteristicas.
Dividimos la data de entrenamiento y de prueba en 80% y 20% respectivamente.
Hacemos uso de la matriz de confusión para analizar el resultado de los modelos.

VIII. RESULTADOS

- Decision tree:

1. Dataset de vibración en el dominio de tiempo

	precision	recall	f1-score	support
0	0.00	0.00	0.00	11
1	0.85	0.85	0.85	1030
2	0.45	0.44	0.44	127
3	0.34	0.35	0.35	174
accuracy			0.74	1342
macro avg	0.41	0.41	0.41	1342
weighted avg	0.74	0.74	0.74	1342

AUC ROC : 0.6220294055911942

Figura 20. Resultado del modelo decision tree al dataset de vibración del dominio en el tiempo.

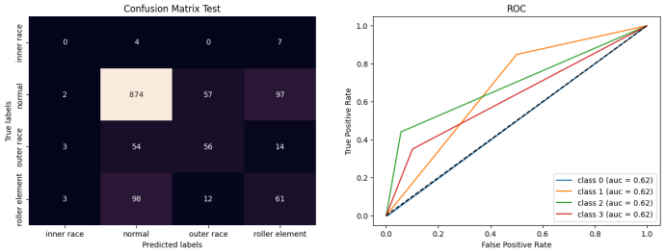


Figura 21. Matriz de confusión del modelo decision tree al dataset de vibración del dominio en el tiempo.

2. Dataset de vibración en el dominio de frecuencia

	precision	recall	f1-score	support
0	0.90	0.82	0.86	11
1	1.00	0.99	0.99	1030
2	0.96	0.98	0.97	127
3	0.97	0.98	0.97	174
accuracy			0.99	1342
macro avg	0.96	0.94	0.95	1342
weighted avg	0.99	0.99	0.99	1342

AUC ROC : 0.967450268075493

Figura 22. Resultado del modelo decision tree al dataset de vibración del dominio de frecuencias.

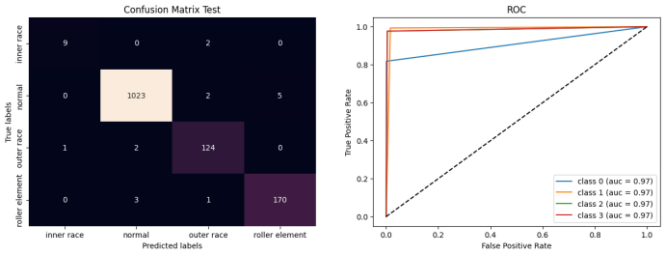


Figura 23. Matriz de confusión del modelo decision tree al dataset de vibración de frecuencias.

3. Dataset de características

	precision	recall	f1-score	support
0	0.92	1.00	0.96	11
1	0.99	0.99	0.99	1030
2	1.00	0.98	0.99	127
3	0.92	0.95	0.94	174
accuracy			0.98	1342
macro avg	0.96	0.98	0.97	1342
weighted avg	0.98	0.98	0.98	1342

AUC ROC : 0.9853873692478821

Figura 24. Resultado del modelo decision tree al dataset de características de vibración.

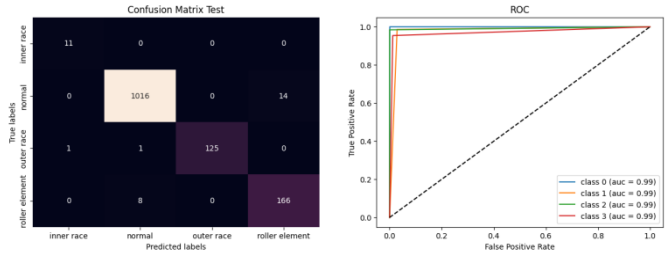


Figura 25. Matriz de confusión del modelo decision tree al dataset de características de vibración.

- Logistic regression:

1. Dataset dominio en el tiempo

0	0.00	0.00	0.00	11
1	0.78	0.99	0.87	1030
2	0.94	0.13	0.22	127
3	0.08	0.01	0.01	174
accuracy			0.77	1342
macro avg	0.45	0.28	0.28	1342
weighted avg	0.70	0.77	0.69	1342

AUC ROC : 0.521584902868997

Figura 26. Resultado del modelo logistic regression al dataset de vibración en el domino del tiempo.

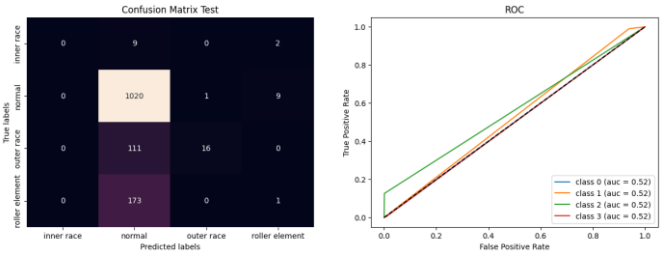


Figura 27. Matriz de confusión del modelo logistic regression al dataset de vibración del dominio del tiempo.

2. Dataset dominio en la frecuencia

	precision	recall	f1-score	support
0	0.00	0.00	0.00	11
1	0.89	1.00	0.94	1030
2	1.00	0.43	0.60	127
3	0.91	0.67	0.77	174
accuracy			0.90	1342
macro avg	0.70	0.53	0.58	1342
weighted avg	0.90	0.90	0.88	1342

AUC ROC : 0.7103256640705581

Figura 28. Resultado del modelo logistic regression al dataset de vibración en el domino de frecuencias.

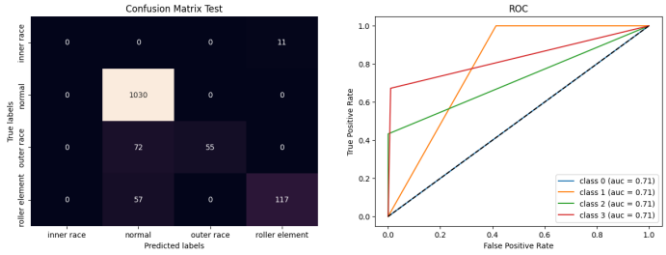


Figura 29. Matriz de confusión del modelo logistic regression al dataset de vibración del dominio de frecuencias.

3. Dataset de caracteristicas

0	0.82	0.82	0.82	11
1	0.84	0.94	0.89	1030
2	0.93	0.60	0.73	127
3	0.38	0.21	0.27	174
accuracy			0.81	1342
macro avg	0.74	0.64	0.67	1342
weighted avg	0.79	0.81	0.79	1342

AUC ROC : 0.7385467384797968

Figura 30. Resultado del modelo logistic regression al dataset de características de vibración.

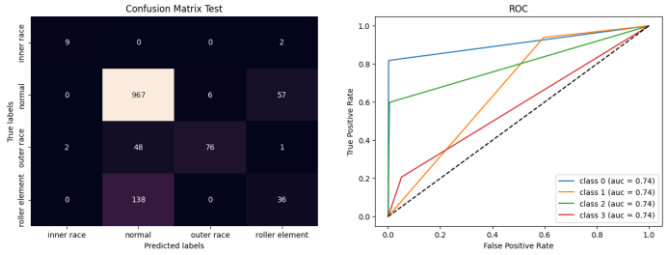


Figura 31. Matriz de confusión del modelo logistic regression al dataset de características de vibración.

- Random forest:

1. Dataset dominio en el tiempo

	precision	recall	f1-score	support
0	0.00	0.00	0.00	11
1	0.83	1.00	0.90	1030
2	1.00	0.60	0.75	127
3	0.52	0.06	0.11	174
accuracy			0.83	1342
macro avg	0.59	0.42	0.44	1342
weighted avg	0.80	0.83	0.78	1342

AUC ROC : 0.6199754222156594

Figura 26. Resultado del modelo random forest al dataset de vibración en el domino del tiempo.

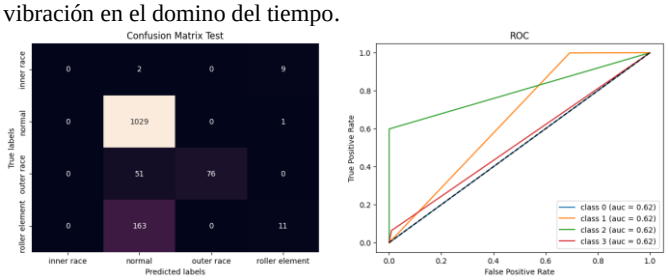


Figura 27. Matriz de confusión del modelo random forest al dataset de vibración del dominio del tiempo.

2. Dataset dominio en la frecuencia

	precision	recall	f1-score	support
0	0.80	0.73	0.76	11
1	0.99	1.00	1.00	1030
2	1.00	0.95	0.98	127
3	0.99	0.98	0.99	174
accuracy			0.99	1342
macro avg	0.95	0.92	0.93	1342
weighted avg	0.99	0.99	0.99	1342

AUC ROC : 0.9539477882053344

Figura 28. Resultado del modelo random forest al dataset de vibración en el domino de frecuencias.

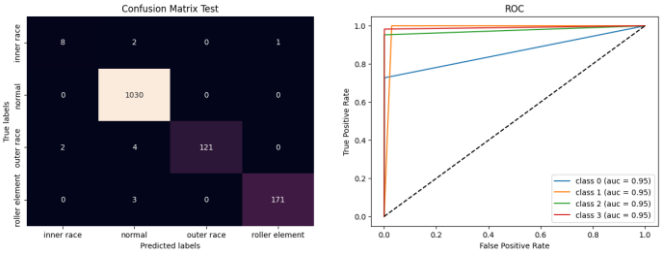


Figura 29. Matriz de confusión del modelo random forest al dataset de vibración del dominio de frecuencias.

3. Dataset de características

	precision	recall	f1-score	support
0	0.92	1.00	0.96	11
1	0.99	0.99	0.99	1030
2	1.00	0.99	1.00	127
3	0.95	0.96	0.96	174
accuracy			0.99	1342
macro avg	0.97	0.99	0.98	1342
weighted avg	0.99	0.99	0.99	1342

AUC ROC : 0.9892615726997428

Figura 30. Resultado del modelo random forest al dataset de características de vibración.

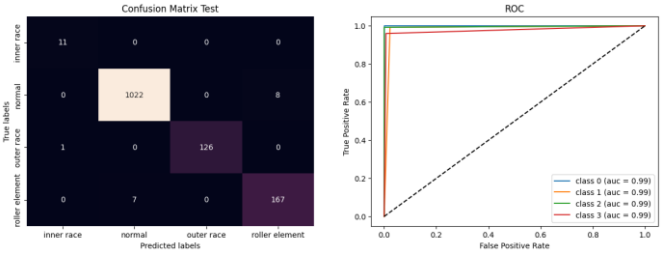


Figura 31. Matriz de confusión del modelo random forest al dataset de características de vibración.

- Extreme gradient boosting (XGBoost):

1. Dataset dominio en el tiempo

	precision	recall	f1-score	support
0	0.00	0.00	0.00	11
1	0.81	1.00	0.90	1030
2	1.00	0.47	0.64	127
3	0.44	0.04	0.07	174
accuracy			0.82	1342
macro avg	0.56	0.38	0.40	1342
weighted avg	0.77	0.82	0.76	1342

AUC ROC : 0.5925253862617472

Figura 32. Resultado del modelo XGBoost al dataset de vibración en el domino del tiempo.

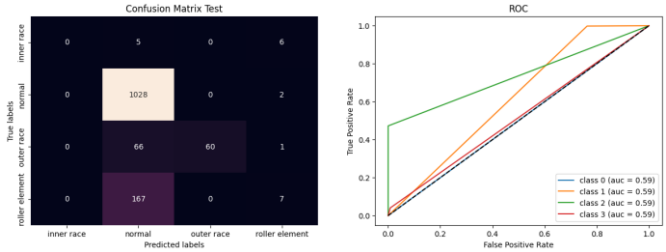


Figura 33. Matriz de confusión del modelo XGBoost al dataset de vibración del dominio del tiempo.

4. Dataset dominio en la frecuencia

	precision	recall	f1-score	support
0	1.00	1.00	1.00	11
1	1.00	1.00	1.00	1030
2	1.00	0.98	0.99	127
3	0.99	1.00	1.00	174
accuracy			1.00	1342
macro avg	1.00	1.00	1.00	1342
weighted avg	1.00	1.00	1.00	1342

AUC ROC : 0.997523834489406

Figura 34. Resultado del modelo XGBoost al dataset de vibración en el domino de frecuencias.

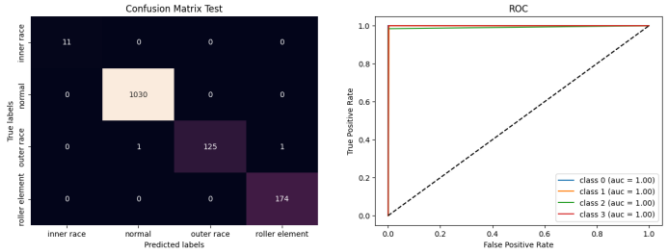


Figura 35. Matriz de confusión del modelo XGBoost al dataset de vibración del dominio de frecuencias.

5. Dataset de caracteristicas

	precision	recall	f1-score	support
0	0.92	1.00	0.96	11
1	1.00	0.99	0.99	1030
2	1.00	0.99	1.00	127
3	0.95	0.99	0.97	174
accuracy			0.99	1342
macro avg	0.96	0.99	0.98	1342
weighted avg	0.99	0.99	0.99	1342

AUC ROC : 0.9943999723084441

Figura 36. Resultado del modelo XGBoost al dataset de características de vibración.

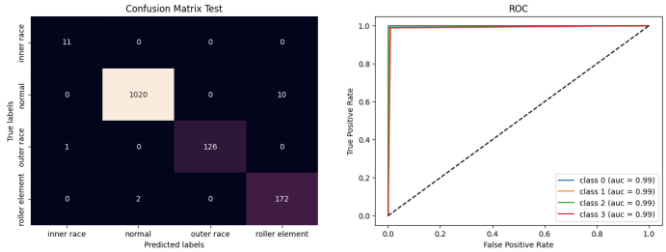
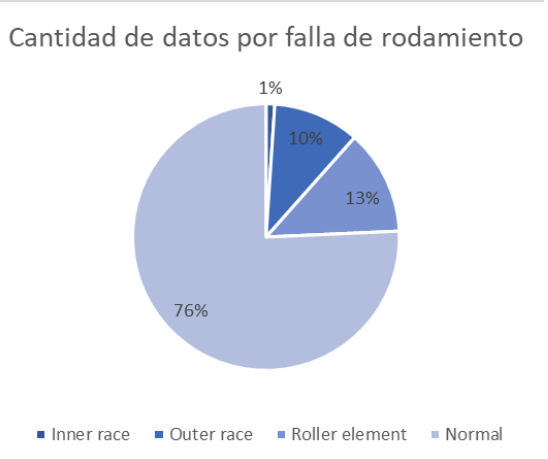


Figura 37. Matriz de confusión del modelo XGBoost rest al dataset de características de vibración.

Re-muestreo de datos de falla de rodamientos.
Realizamos un análisis de los datos que tenemos en base a los tipos de fallas, evidenciando que se cuenta con una cantidad desigual por tipo de falla de rodamiento.

Tipos de falla rodamientos	Cant. Datos
Inner race	70
Outer race	705
Roller element	854
Normal	5077

Tabla 01. Cantidad de datos por tipos de fallas de rodamientos

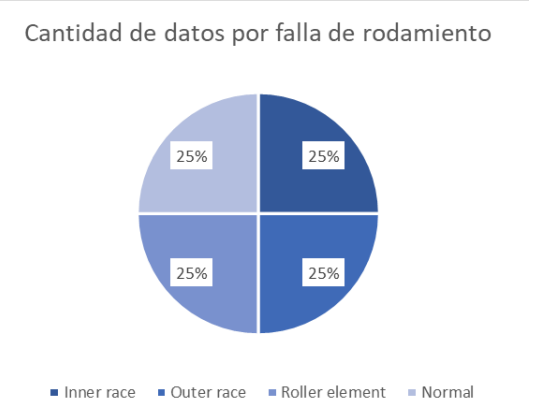


Grafica 01. Pie de distribución de datos por falla de rodamientos

Debido a que tenemos una distribución desigual, siendo el 76% como condición normal de los rodamientos y el 24% de las fallas, con un valor mínimo para la falla de pista interna (inner race) de la cantidad de datos por categorías de fallas vamos a realizar re-muestreo para ello elegimos categoría de mayor tamaño la categoría 'normal' y los demás como minoritarios.

Tipos de falla rodamientos	Cant. Datos
Inner race	5077
Outer race	5077
Roller element	5077
Normal	5077

Tabla 01. Cantidad de datos por tipos de fallas de rodamientos despues del remuestreo.



Grafica 01. Pie de distribución de datos por falla de rodamientos despues del re-muestreo.

Con el re-muestreo se procede aplicar los modelos decisión tree, logistic regression, random forest y XGBoost.

- Decision tree:

6. Dataset dominio en el tiempo

	precision	recall	f1-score	support
0	0.99	1.00	0.99	993
1	0.98	0.76	0.86	1056
2	0.89	0.99	0.94	995
3	0.89	0.99	0.94	1018
accuracy			0.93	4062
macro avg	0.94	0.94	0.93	4062
weighted avg	0.94	0.93	0.93	4062

AUC ROC : 0.9573907504948114

Figura 38. Resultado del modelo decision tree al dataset de vibración en el domino del tiempo.

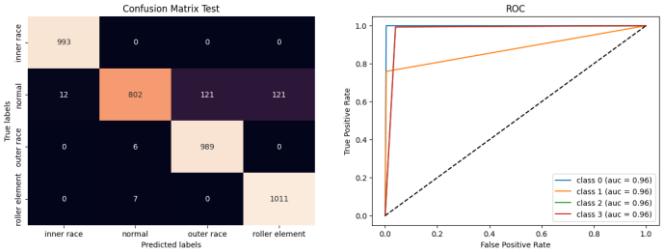


Figura 39. Matriz de confusión del modelo decision tree al dataset de vibración del dominio del tiempo.

7. Dataset dominio en la frecuencia

	precision	recall	f1-score	support
0	1.00	1.00	1.00	993
1	1.00	0.99	0.99	1056
2	0.99	1.00	1.00	995
3	1.00	1.00	1.00	1018
accuracy			1.00	4062
macro avg	1.00	1.00	1.00	4062
weighted avg	1.00	1.00	1.00	4062

AUC ROC : 0.9980895443320028

Figura 40. Resultado del modelo decision tree al dataset de vibración en el domino de frecuencias.

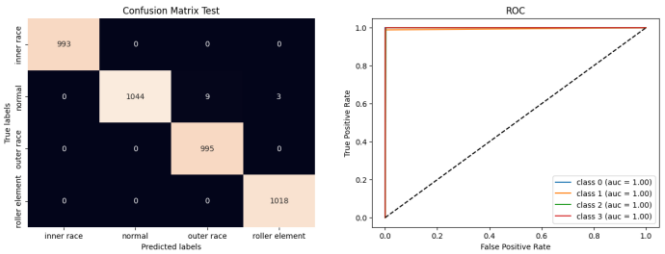


Figura 41. Matriz de confusión del modelo decisión tree al dataset de vibración del dominio de frecuencias.

8. Dataset de caracteristicas

	precision	recall	f1-score	support
0	1.00	1.00	1.00	993
1	1.00	0.99	0.99	1056
2	1.00	1.00	1.00	995
3	0.99	1.00	0.99	1018
accuracy			1.00	4062
macro avg	1.00	1.00	1.00	4062
weighted avg	1.00	1.00	1.00	4062

AUC ROC : 0.9977533964868583

Figura 42. Resultado del modelo decisión tree al dataset de características de vibración.

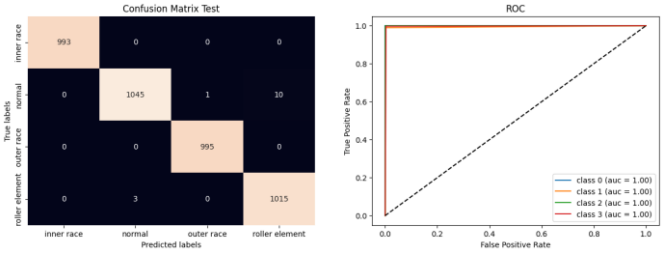


Figura 43. Matriz de confusión del modelo decisión tree al dataset de características de vibración.

- Logistic regression:

4. Dataset dominio en el tiempo

	precision	recall	f1-score	support
0	1.00	1.00	1.00	993
1	0.98	0.93	0.96	1056
2	0.99	0.99	0.99	995
3	0.95	0.99	0.97	1018
accuracy			0.98	4062
macro avg	0.98	0.98	0.98	4062
weighted avg	0.98	0.98	0.98	4062

AUC ROC : 0.986185894193061

Figura 44. Resultado del modelo logistic regression al dataset de vibración en el domino del tiempo.

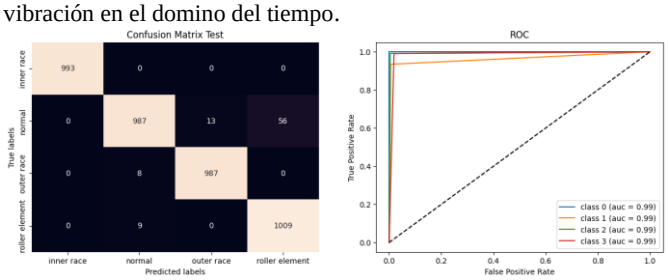


Figura 45. Matriz de confusión del modelo logistic regression al dataset de vibración del dominio del tiempo.

5. Dataset dominio en la frecuencia

	precision	recall	f1-score	support
0	1.00	1.00	1.00	993
1	0.83	0.98	0.90	1056
2	0.99	0.82	0.90	995
3	0.99	0.96	0.97	1018
accuracy			0.94	4062
macro avg	0.95	0.94	0.94	4062
weighted avg	0.95	0.94	0.94	4062

AUC ROC : 0.9601332171002135

Figura 46. Resultado del modelo logistic regression al dataset de vibración en el domino de frecuencias.

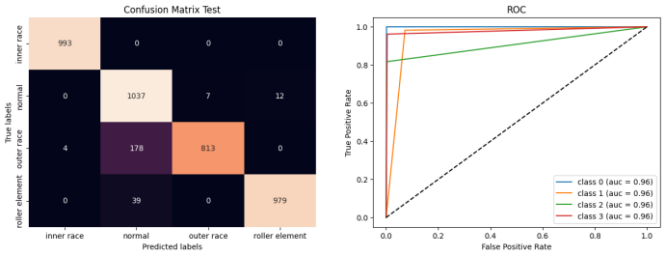


Figura 47. Matriz de confusión del modelo logistic regression al dataset de vibración del dominio de frecuencias.

6. Dataset de caracteristicas

	precision	recall	f1-score	support
0	0.82	0.82	0.82	11
1	0.84	0.94	0.89	1030
2	0.93	0.60	0.73	127
3	0.38	0.21	0.27	174
accuracy			0.81	1342
macro avg	0.74	0.64	0.67	1342
weighted avg	0.79	0.81	0.79	1342

AUC ROC : 0.7385467384797968

Figura 48. Resultado del modelo logistic regression al dataset de características de vibración.

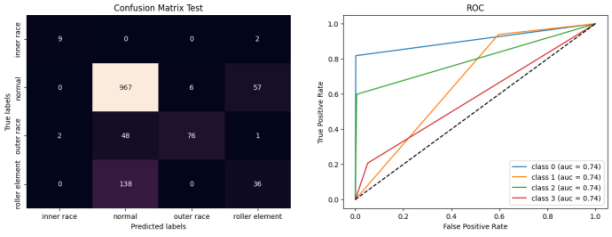


Figura 49. Matriz de confusión del modelo logistic regression al dataset de características de vibración.

- Random forest:

4. Dataset dominio en el tiempo

	precision	recall	f1-score	support
0	1.00	1.00	1.00	993
1	0.99	1.00	0.99	1056
2	1.00	0.99	1.00	995
3	1.00	0.99	1.00	1018
accuracy			1.00	4062
macro avg	1.00	1.00	1.00	4062
weighted avg	1.00	1.00	1.00	4062

AUC ROC : 0.9973529973222017

Figura 50. Resultado del modelo random forest al dataset de vibración en el domino del tiempo.

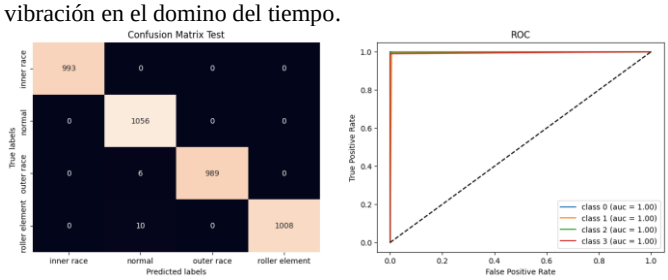


Figura 51. Matriz de confusión del modelo random forest al dataset de vibración del dominio del tiempo.

5. Dataset dominio en la frecuencia

	precision	recall	f1-score	support
0	1.00	1.00	1.00	993
1	1.00	1.00	1.00	1056
2	1.00	1.00	1.00	995
3	1.00	1.00	1.00	1018
accuracy			1.00	4062
macro avg	1.00	1.00	1.00	4062
weighted avg	1.00	1.00	1.00	4062

AUC ROC : 1.0

Figura 52. Resultado del modelo random forest al dataset de vibración en el domino de frecuencias.

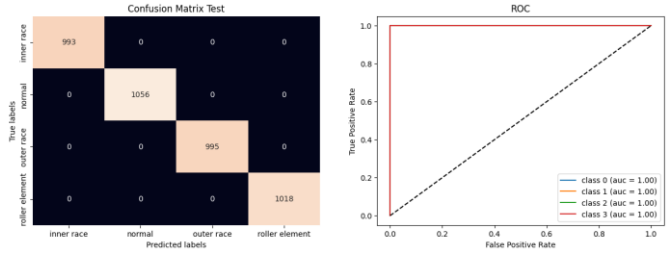


Figura 53. Matriz de confusión del modelo random forest al dataset de vibración del dominio de frecuencias.

6. Dataset de caracteristicas

	precision	recall	f1-score	support
0	1.00	1.00	1.00	993
1	1.00	0.99	1.00	1056
2	1.00	1.00	1.00	995
3	0.99	1.00	1.00	1018
accuracy			1.00	4062
macro avg	1.00	1.00	1.00	4062
weighted avg	1.00	1.00	1.00	4062

AUC ROC : 0.9985653875396965

Figura 54. Resultado del modelo random forest al dataset de características de vibración.

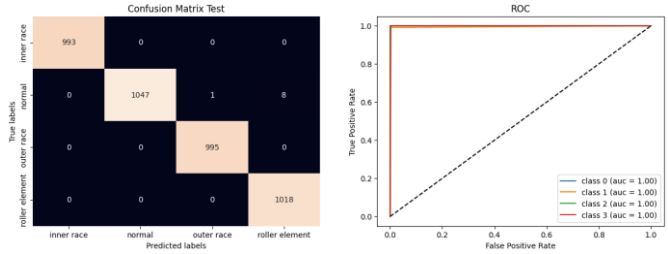


Figura 55. Matriz de confusión del modelo random forest al dataset de características de vibración.

- Extreme gradient boosting (XGBoost):

2. Dataset dominio en el tiempo

	precision	recall	f1-score	support
0	1.00	1.00	1.00	993
1	0.99	0.97	0.98	1056
2	1.00	1.00	1.00	995
3	0.97	0.99	0.98	1018
accuracy			0.99	4062
macro avg	0.99	0.99	0.99	4062
weighted avg	0.99	0.99	0.99	4062

AUC ROC : 0.9937241196625763

Figura 50. Resultado del modelo XGBoost al dataset de vibración en el domino del tiempo.

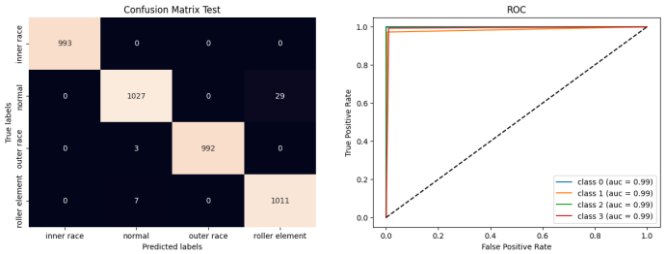


Figura 51. Matriz de confusión del modelo XGBoost al dataset de vibración del dominio del tiempo.

3. Dataset dominio en la frecuencia

	precision	recall	f1-score	support
0	1.00	1.00	1.00	993
1	1.00	1.00	1.00	1056
2	1.00	1.00	1.00	995
3	1.00	1.00	1.00	1018
accuracy			1.00	4062
macro avg	1.00	1.00	1.00	4062
weighted avg	1.00	1.00	1.00	4062

AUC ROC : 1.0

Figura 52. Resultado del modelo XGBoost al dataset de vibración en el domino de frecuencias.

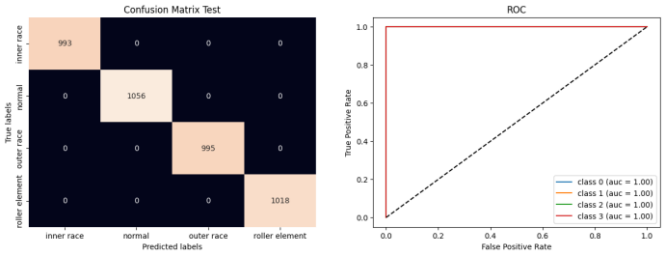


Figura 53. Matriz de confusión del modelo XGBoost al dataset de vibración del dominio de frecuencias.

4. Dataset de Caracteristicas

	precision	recall	f1-score	support
0	1.00	1.00	1.00	993
1	1.00	0.99	1.00	1056
2	1.00	1.00	1.00	995
3	0.99	1.00	1.00	1018
accuracy			1.00	4062
macro avg	1.00	1.00	1.00	4062
weighted avg	1.00	1.00	1.00	4062

AUC ROC : 0.9985653875396965

Figura 54. Resultado del modelo XGBoost al dataset de características de vibración.

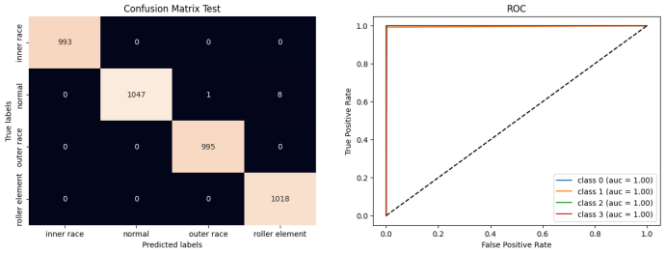


Figura 55. Matriz de confusión del modelo XGBoost al dataset de características de vibración.

Como resumen de los resultados tenemos las siguientes tablas:

	Precisión			
Modelos	Dataset Tiempo	Dataset Frecuenca	Dataset Caracteristicas	Precisión promedio
Decision tree	0.41	0.96	0.96	0.78
Logistic regression	0.52	0.71	0.73	0.65
Random forest	0.59	0.95	0.97	0.84
XGBoost	0.59	0.99	0.99	**0.86**

Tabla 02. Precisión por modelo antes del re-muestreo de datos.

	Precisión			
Modelos	Dataset Tiempo	Dataset Frecuenca	Dataset Caracteristicas	Precisión promedio
Decision tree	0.95	0.99	0.99	0.98
Logistic regression	0.98	0.96	0.92	0.95
Random forest	0.99	1	0.99	**0.99**
XGBoost	0.99	1	0.99	**0.99**

Tabla 03. Precisión por modelo después del re-muestreo de datos.

	Tiempo de ejeciución			
Modelos	Dataset Tiempo	Dataset Frecuenca	Dataset Caracteristicas	Tiempo promedio
Decision tree	5:34	2:14	0:00	2:36
Logistic regression	1:32	1:41	0:00	1:04
Random forest	2:04	3:06	0:02	1:44
XGBoost	17:29	3:06	0:01	**6:52**
o por dataset	6:40	2:32	0:01	

Tabla 04. Tiempo de ejecución de los modelos después del re-muestreo de datos.

Peso del archivo csv		
Dataset Tiempo	Dataset Frecuenca	Dataset Caracteristicas
1,483Mb	880Mb	**1Mb**

Tabla 05. Peso de los dataset para cada tipo.

IX. CONCLUSIONES

1. Se concluye que el mejor dataset para trabajar es el de características con menor peso, mejor tiempo de ejecución y precisión (1Mb peso, 0.1 tiempo de ejecución), el cual tiene ocupe menor espacio comparado con el dataset de tiempo (1,483Mb peso, 6:40 tiempo de ejecución) y frecuencias (880Mb peso, 2:32 tiempo de ejecución).

2. Se concluye que el modelo con mejor desempeño tanto en precisión, tiempo de ejecución, para el dataset de características es Random fores con presión de 0.99 y tiempo de ejecución promedio 0.2 segundos.

3. Se concluye que el remuestreo fue acertado ya que mejora el desempeño en general de todos los modelos de machine learning, en especial logistic regression la cual se ve beneficiada incrementando una precisión promedio de 0.65 a 0.95 para los modelos de machine learning.

X. RECOMENDACIONES

- ✓ Se recomienda categorizar cada falla en normal, alerta, crítico por cada tipo de falla, para poder tener mayor control de las fallas de los rodamientos, y generar mejor análisis de estos.
- ✓ Se recomienda pasar un filtro pasa altas y hallar la envolvente de las frecuencias para que se resalte mejor las fallas de rodamientos.

XI. REFERENCIAS

[1] R. K. Mobley, *An Introduction to Predictive Maintenance*, 2nd ed., Butterworth-Heinemann, 2002.

[2] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.

[3] D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression*, 2nd ed., Wiley-Interscience, 2000.

[4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[6] X. Li, Q. Ding, y J. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," *Reliability Engineering & System Safety*, vol. 172, pp. 1-11, 2018.

[7] J. Zhang, L. Zhang, X. Si, y M. Zhang, "Machine Learning Algorithms in the Application of Aerospace Bearing Fault Diagnosis: A Review," *Sensors*, vol. 21, no. 9, p. 3114, 2021.

[8] T. Wuest, D. Weimer, C. Irgens, y K.-D. Thoben, "Machine learning in manufacturing: advantages, challenges, and applications," *Production & Manufacturing Research*, vol. 4, no. 1, pp. 23-45, 2016.

[9] S. M. Salih, A. J. Abdulrahman, y A. M. Shareef, "Vibration-Based Bearing Fault Diagnosis Using Machine Learning Approaches," *IEEE Access*, vol. 8, pp. 66233-66245, 2020.

[10] M. E. Torres, E. Hernández, y F. A. Fernández, "Feature extraction methods for machine learning in condition monitoring and fault diagnosis of industrial processes," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1541-1551, 2017.

[11] R. K. Mobley, An Introduction to Predictive Maintenance, 2nd ed., Butterworth-Heinemann, 2002.

[12] T. Hastie, R. Tibshirani, y J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.

[13] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.

[14] D. W. Hosmer, S. Lemeshow, y R. X. Sturdivant, Applied Logistic Regression, 3rd ed., Wiley, 2013.

[15] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[16] T. Chen y C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[17] S. Mallat, A Wavelet Tour of Signal Processing: The Sparse Way, 3rd ed., Academic Press, 2008.

[18] N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 454, no. 1971, pp. 903-995, 1998.

[19] I. Guyon y A. Elisseeff, "An introduction to feature extraction," en Feature Extraction: Foundations and Applications, Springer, 2006, pp. 1-25.

[20] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.

[21] S. M. Salih, A. J. Abdulrahman, y A. M. Shareef, "Vibration-Based Bearing Fault Diagnosis Using Machine Learning Approaches," IEEE Access, vol. 8, pp. 66233-66245, 2020.

[22] M. E. Torres, E. Hernández, y F. A. Fernández, "Feature extraction methods for machine learning in condition monitoring and fault diagnosis of industrial processes," IEEE Transactions on Industrial Informatics, vol. 13, no. 4, pp. 1541-1551, 2017.