



Centro Universitario de la Defensa en la Escuela Naval Militar

TRABAJO FIN DE GRADO

*Predicción de fallos en máquinas mediante regresión logística
basada en variables de sensores*

Grado en Ingeniería Mecánica

ALUMNO: Pedro Márquez Franco

DIRECTORES: Débora Cores Carrera
Alicia Vázquez Carpentier

CURSO ACADÉMICO: 2025-2026

Universida_{de}Vigo



Centro Universitario de la Defensa en la Escuela Naval Militar

TRABAJO FIN DE GRADO

*Predicción de fallos en máquinas mediante regresión logística
basada en variables de sensores*

Grado en Ingeniería Mecánica
Intensificación en Tecnología Naval
Cuerpo General / Infantería de Marina

UniversidadeVigo

RESUMEN

En este Trabajo Fin de Grado se analiza la aplicación de modelos de regresión logística para la predicción del riesgo de fallo en maquinaria industrial a partir de variables operativas. El objetivo principal es evaluar la capacidad de estos modelos para identificar patrones asociados al fallo y estudiar la influencia de la calidad de los datos y la selección de variables en su rendimiento.

Para ello, se han empleado distintos conjuntos de datos, tanto iniciales como sintéticos diseñados específicamente para el análisis, que incluyen variables como temperatura, vibración, potencia, humedad, presión y edad. Los modelos han sido implementados en MATLAB, estimando la probabilidad de fallo en función de dichas variables.

Como aportación relevante, se ha incorporado el análisis de la matriz de correlación y de los p-valores, lo que ha permitido identificar relaciones entre variables y justificar la eliminación de variables redundantes o poco significativas. Esta etapa ha resultado clave para mejorar la interpretabilidad del modelo sin comprometer su capacidad predictiva.

El rendimiento se ha evaluado mediante curvas ROC y el área bajo la curva (AUC), observándose que los mejores resultados se obtienen en aquellos conjuntos donde las variables presentan una relación clara con el fallo. En conjunto, el trabajo demuestra que la regresión logística es una herramienta eficaz, siempre que se realice un adecuado tratamiento y selección de los datos.

PALABRAS CLAVE

Regresión logística, predicción de fallos, mantenimiento predictivo, fallo de maquinaria, selección de variables, curva ROC.

AGRADECIMIENTOS

En primer lugar, quiero expresar mi más sincero agradecimiento a mi familia, especialmente a mis padres y a mis hermanas, por su apoyo incondicional a lo largo de toda esta etapa. Gracias por estar siempre al pie del cañón, por vuestra confianza y por acompañarme en todo momento, incluso en los más complicados.

También quiero dar las gracias a mis compañeros y amigos, en especial a Carlos de la Puente, Álvaro Tortosa y Marcos Muño, por su paciencia y por haber estado siempre dispuestos a ayudarme, soportando todas mis preguntas en la “camareta” y haciendo este proceso mucho más llevadero.

A mi amiga Chiqui, por aguantar cada día mis interminables conversaciones sobre el TFG y por sus constantes palabras de ánimo y motivación, que han sido fundamentales para seguir adelante.

Por último, quiero agradecer especialmente a mis tutoras, Débora y Alicia, cuya guía, apoyo y dedicación han sido imprescindibles para la realización de este trabajo. Sin su ayuda, este proyecto no habría sido posible.

CONTENIDO

Contenido	1
Índice de Figuras	3
Índice de Tablas.....	4
1 Introducción y objetivos.....	5
1.1 Motivación	5
1.2 Antecedentes	6
1.3 Objetivo.....	6
1.4 Estructura del trabajo	6
2 Estado del arte	9
2.1 Problemas de clasificación binaria en estadística	9
2.1.1 Que es clasificar.....	9
2.1.2 Variable respuesta binaria.....	10
2.1.3 Ejemplos típicos (fallo/no fallo)	10
2.2 Regresión lineal y sus limitaciones	11
2.2.1 Modelo lineal clásico	11
2.2.2 Justificación de buscar otro modelo	12
2.3 Modelo de regresión logística	12
2.3.1 Definición del modelo	12
2.3.2 Función logística.....	13
2.3.3 Interpretación probabilística	14
2.3.4 Significado de los parámetros.....	14
2.4 Estimación de parámetros mediante máxima verosimilitud	15
2.4.1 Idea de verosimilitud	15
2.4.2 Función de verosimilitud	15
2.4.3 Log- verosimilitud	16
2.4.4 Modelo de optimización logística.....	17
2.5 Uso de herramientas computacionales.....	17
2.5.1 Papel de Matlab en el ajuste del modelo	17
2.5.2 Importancia de la simulación y el análisis numérico.....	18
3 Desarrollo del TFG.....	19
3.1 Planteamiento del problema.....	19
3.2 Modelo de regresión logística	19
3.3 Método del gradiente	20

3.4 Método de Newton.....	20
3.5 Descripción de los datos	21
3.5.1 Origen y estructura de los conjuntos de datos	21
3.6 Preparación de los datos.....	22
3.6.1 Selección de variables explicativas	22
3.7 Implementación del modelo en Matlab.....	22
3.7.1 Estimación de parámetros mediante “glmfit”	23
3.7.2 Cálculo de probabilidades mediante “glmval”	23
3.8 Metodología de evaluación	23
3.8.1 Análisis de correlación entre variables	23
3.8.2 Construcción de la curva ROC mediante “perfcurve”	24
3.8.3 Cálculo del área bajo la curva (AUC).....	25
3.8.4 Interpretación de los parámetros estadísticos del modelo	25
4 Resultados obtenidos.....	29
4.1 Resultados del conjunto de datos CD1	29
4.1.1 CD11 Máquina tipo Mill.....	29
4.1.2 CD12 Máquina tipo Lathe.....	31
4.1.3 CD13 Máquina tipo Drill.....	32
4.1.4 Tabla resumen CD1.....	34
4.2 Resultados del conjunto de datos CD2	35
4.3 Resultados del conjunto de datos de CDS1	36
4.4 Resultados del conjunto de datos de CDS2	39
4.5 Resultados del conjunto de datos de CDS3	42
4.6 Resultados del conjunto de datos de CDS4	44
4.7 Comparación de resultados	46
5 Conclusiones y líneas futuras	49
5.1 Conclusiones	49
5.2 Líneas futuras	50
6 Bibliografía.....	51
Anexo I: Implicaciones Sociales, y/o Económicas, y/o Ambientales	52
Anexo II: Reflexiones Éticas y Sociales	53
Anexo III: Código Matlab	54
Anexo IV: Uso de la IA.....	57

ÍNDICE DE FIGURAS

Figura 2-1 Representación de una función lineal con su error asociado (tomada de [11]).....	12
Figura 4-1 Curva ROC- <i>CD11</i> (figura tomada de Matlab)	31
Figura 4-2 curva ROC- <i>CD12</i> (figura tomada de Matlab)	32
Figura 4-3 Curva ROC- <i>CD13</i> (figura tomada de Matlab)	34
Figura 4-4 Curva ROC- <i>CD2</i> (figura tomada de Matlab)	36
Figura 4-5 Matriz correlación <i>CDS1</i> (figura tomada de Matlab).....	37
Figura 4-6 Curva ROC- <i>CDS1</i> (figura tomada de Matlab)	38
Figura 4-7 Matriz correlación- <i>CDS2</i> (figura tomada de Matlab)	40
Figura 4-8 Curva comparación modelos- <i>CDS2</i> (figura tomada de Matlab)	41
Figura 4-9 Matriz correlación- <i>CDS3</i> (figura tomada de Matlab)	42
Figura 4-10 Curva ROC- <i>CDS3</i> (figura tomada de Matlab)	43
Figura 4-11 Matriz correlación- <i>CDS4</i> (figura tomada de Matlab)	44
Figura 4-12 Curva ROC- <i>CDS4</i> (figura tomada de Matlab)	45

ÍNDICE DE TABLAS

Tabla 3-1 Características de las bases de datos	21
Tabla 4-1 Resultados máquina tipo Mill	30
Tabla 4-2 Indicadores de máquina tipo Mill	30
Tabla 4-3 Resultados máquina tipo Lathe	31
Tabla 4-4 Indicadores de máquina tipo Lathe	31
Tabla 4-5 Resultados máquina tipo Drill	33
Tabla 4-6 Indicadores de máquina tipo Drill.....	33
Tabla 4-7 Resumen de resultados del CD1.	34
Tabla 4-8 Resultados del CD2.....	35
Tabla 4-9 Indicadores del modelo para CD2.....	35
Tabla 4-10 Resultados del CDS1.	37
Tabla 4-11 Indicadores del modelo para CDS1.	37
Tabla 4-12 Resultados del CDS1 reducido.	39
Tabla 4-13 Comparativa de modelos CDS1.....	39
Tabla 4-14 Resultados del conjunto de datos CDS2.	40
Tabla 4-15 Indicadores del modelo para CDS2.	40
Tabla 4-16 Resultados del CDS2 reducido.	41
Tabla 4-17 Indicadores del modelo para CDS2 reducido.	41
Tabla 4-18 Resultados del CDS3.	43
Tabla 4-19 Indicadores del modelo para CDS3.	43
Tabla 4-20 Resultados del CDS4.	45
Tabla 4-21 Indicadores del modelo para CDS4.	45
Tabla 4-22 Comparación de la capacidad predictiva de los modelos.	46
Tabla 4-23. Ranking de modelos.....	47

1 INTRODUCCIÓN Y OBJETIVOS

1.1 Motivación

La fiabilidad y disponibilidad de la maquinaria constituyen factores clave en la operación de sistemas industriales complejos. La aparición de fallos inesperados en equipos mecánicos y electromecánicos puede provocar interrupciones en la operación, pérdidas económicas significativas y, en determinados contextos, comprometer la seguridad de las personas y de las instalaciones [1]. Estas consecuencias adquieren una especial relevancia en entornos donde la continuidad del servicio resulta crítica y las condiciones de operación pueden ser exigentes.

En este contexto, resulta necesario disponer de herramientas que permitan anticipar la aparición de fallos y actuar de manera preventiva antes de que estos se produzcan. La evolución de los sistemas de sonorización y la creciente disponibilidad de datos operativos han impulsado el desarrollo de técnicas de mantenimiento predictivo basadas en el análisis de variables de condición de los equipos [2]. La motivación de este Trabajo de Fin de Grado surge, por tanto, de la necesidad de desarrollar métodos que permitan estimar de manera cuantitativa la probabilidad de fallo de una máquina en función de sus condiciones de funcionamiento, contribuyendo así a una gestión más eficiente del mantenimiento.

Este enfoque resulta especialmente relevante en el ámbito naval y, en particular, en los buques de la Armada, donde la fiabilidad de los sistemas de a bordo constituye un factor crítico para garantizar la operatividad del buque y la seguridad de la dotación. Los buques de guerra integran una elevada complejidad técnica y operan durante largos periodos en condiciones exigentes, lo que incrementa la probabilidad de desgaste y fallo de equipos mecánicos y electromecánicos. Sistemas como la propulsión, la generación y distribución eléctrica, la climatización, los sistemas auxiliares o la maquinaria asociada a servicios esenciales del buque deben mantener elevados niveles de disponibilidad para asegurar el cumplimiento de la misión asignada.

La aparición de una avería imprevista durante la navegación puede limitar de forma significativa la capacidad operativa del buque, obligar a modificar la misión o incluso comprometer la seguridad de la unidad. La aplicación de técnicas de mantenimiento predictivo basadas en datos obtenidos mediante sensores permitiría anticipar el riesgo de fallo de estos sistemas, optimizar la planificación de las tareas de mantenimiento y reducir la dependencia de intervenciones correctivas no planificadas. La posibilidad de estimar de manera cuantitativa la probabilidad de fallo de una máquina en función de sus condiciones de funcionamiento constituye, por tanto, una herramienta de gran interés para apoyar la toma de decisiones en el ámbito del mantenimiento naval y mejorar la disponibilidad operativa de los buques de la Armada.

1.2 Antecedentes

Tradicionalmente, el mantenimiento de la maquinaria se ha basado en estrategias correctivas y preventivas. El mantenimiento correctivo actúa una vez que el fallo ya se ha producido, lo que suele implicar tiempos de parada elevados y costes no planificados. Por su parte, el mantenimiento preventivo establece intervenciones periódicas basadas en el tiempo de operación o en el número de horas de servicio, sin tener en cuenta el estado real del equipo. Aunque este enfoque reduce la frecuencia de fallos, puede dar lugar a intervenciones innecesarias o no evitar determinadas averías [1].

El desarrollo del mantenimiento predictivo ha supuesto un avance significativo frente a estas estrategias tradicionales. Este enfoque se basa en la monitorización continua de la maquinaria mediante sensores capaces de medir variables relevantes, como la temperatura, la vibración o la humedad, permitiendo detectar patrones asociados al deterioro de los equipos [2]. En este contexto, las técnicas estadísticas y de análisis de datos desempeñan un papel fundamental para transformar la información procedente de los sensores en indicadores útiles para la toma de decisiones.

Entre los distintos métodos empleados para la predicción de fallos, los modelos de clasificación probabilística destacan por su capacidad para estimar la probabilidad de ocurrencia de un evento. En particular, la regresión logística ha sido ampliamente utilizada en problemas de ingeniería debido a su simplicidad, robustez e interpretabilidad, permitiendo analizar la influencia de distintas variables sobre el riesgo de fallo [3].

1.3 Objetivo

El objetivo general de este Trabajo de Fin de Grado es desarrollar un modelo de regresión logística que permita predecir la probabilidad de fallo de una máquina a partir de datos procedentes de sensores operativos.

Para alcanzar este objetivo general, se plantean los siguientes objetivos específicos:

- Analizar un conjunto de datos que incluya variables operativas y de condición de la máquina.
- Estudiar la relación entre las variables medidas por los sensores y la aparición de fallos.
- Construir un modelo de regresión logística para la predicción de fallos.
- Evaluar el rendimiento del modelo mediante métricas adecuadas.
- Analizar la influencia de cada variable sobre la probabilidad de fallo.
- Valorar la aplicabilidad del modelo como herramienta de apoyo al mantenimiento predictivo, especialmente en sistemas críticos.

1.4 Estructura del trabajo

El presente documento se estructura en cinco capítulos diferenciados, con el objetivo de organizar de forma clara y ordenada el desarrollo del Trabajo de Fin de Grado, así como los resultados obtenidos. A continuación, se describe brevemente el contenido de cada capítulo:

- Capítulo 1: Introducción.
En este capítulo se contextualiza el trabajo, exponiendo la problemática asociada a los fallos en maquinaria y la necesidad de disponer de herramientas que permitan anticipar su aparición. Se presentan la motivación del estudio, los antecedentes en el ámbito del mantenimiento predictivo y los objetivos que se persiguen. Asimismo, se pone de manifiesto la relevancia de este tipo de técnicas en entornos críticos, como la maquinaria de los buques de la Armada.
- Capítulo 2: Estado del arte.
Se abordan los conceptos fundamentales relacionados con el mantenimiento industrial, haciendo hincapié en el mantenimiento predictivo basado en datos de sensores. Además, se introducen los fundamentos teóricos de la regresión logística, explicando su formulación, funcionamiento e

interpretación, y justificando su idoneidad para la estimación de la probabilidad de fallo de las máquinas.

- **Capítulo 3: Metodología y desarrollo del modelo**
En este capítulo se describe la metodología seguida para el desarrollo del trabajo. Se detallan las variables consideradas, el conjunto de datos empleado y el tratamiento previo de la información. Asimismo, se explica el proceso de construcción del modelo de regresión logística utilizado para la predicción de fallos.
- **Capítulo 4: Resultados y análisis.**
Se presentan y analizan los resultados obtenidos a partir del modelo desarrollado. Se evalúa su capacidad para estimar la probabilidad de fallo de la maquinaria bajo diferentes condiciones operativas y se estudia la influencia de las distintas variables de entrada en el comportamiento del modelo.
- **Capítulo 5: Conclusiones y líneas futuras.**
En este último capítulo se recogen las conclusiones derivadas del trabajo realizado, valorando el grado de cumplimiento de los objetivos planteados. Finalmente, se proponen posibles líneas de trabajo futuro orientadas a mejorar el modelo y a ampliar su aplicación en el ámbito del mantenimiento predictivo de maquinaria, especialmente en entornos navales.
- **Bibliografía y anexos.**
Se incluyen las referencias bibliográficas utilizadas a lo largo del trabajo, así como material complementario relevante, como resultados adicionales o documentación de apoyo.

2 ESTADO DEL ARTE

En este capítulo se presentan los fundamentos teóricos necesarios para el desarrollo del trabajo. Se introduce el modelo de regresión logística, así como los principales conceptos asociados a su interpretación y evaluación, incluyendo la función logística, los coeficientes del modelo y las métricas de rendimiento utilizadas.

2.1 Problemas de clasificación binaria en estadística

2.1.1 *Que es clasificar*

En el ámbito de la estadística aplicada es habitual encontrarse con problemas en los que la variable de interés no toma valores continuos, sino que representa la ocurrencia o no de un determinado evento. Este tipo de problemas se conocen como problemas de clasificación, y cuando la variable respuesta solo puede adoptar dos posibles valores se denominan problemas de clasificación binaria [3].

En un problema de clasificación binaria, la variable respuesta y se define habitualmente como una variable que solo puede tomar dos valores posibles, tomando el valor 1 cuando el evento de interés ocurre y 0 en caso contrario. Ejemplos habituales de este tipo de problemas son la detección de fallos en sistemas industriales, el diagnóstico de enfermedades, la concesión o no de un crédito, o la predicción de éxito o fracaso de un determinado proceso [4]. En todos estos casos, el objetivo no es predecir un valor numérico concreto, sino determinar a qué clase pertenece cada observación.

Desde un punto de vista estadístico, estos problemas se abordan modelando la probabilidad de ocurrencia del evento, condicionada a un conjunto de variables explicativas. Es decir, se busca estimar la probabilidad de que $y = 1$ dadas unas determinadas características observadas, lo que permite no solo clasificar una observación, sino también cuantificar el grado de incertidumbre asociado a dicha clasificación [5].

El uso de modelos probabilísticos en problemas de clasificación binaria resulta especialmente adecuado cuando se requiere interpretar el resultado del modelo y analizar la influencia de las distintas variables explicativas sobre la variable respuesta. En este contexto, los modelos estadísticos permiten establecer una relación clara entre las variables de entrada y la probabilidad del evento, constituyendo una herramienta fundamental en aplicaciones de ingeniería y análisis de datos [3].

La formulación de estos problemas pone de manifiesto la necesidad de emplear modelos específicos que sean capaces de manejar variables respuesta binarias y proporcionar estimaciones probabilísticas coherentes. Esta necesidad conduce al estudio de modelos alternativos a la regresión lineal clásica, los cuales se analizan en los siguientes apartados.

2.1.2 Variable respuesta binaria

En los problemas de clasificación binaria, la variable respuesta representa la ocurrencia o no de un determinado evento y solo puede adoptar dos valores posibles. En el contexto del presente trabajo, dicha variable indica la presencia o ausencia de un fallo en la máquina analizada, tomando el valor 1 cuando se produce un fallo y 0 en caso contrario [3].

Desde un punto de vista estadístico, este tipo de variable se modela habitualmente como una variable aleatoria de tipo Bernoulli. Es decir, para cada observación i , la variable respuesta Y_i puede tomar el valor 1 con una determinada probabilidad p_i , y el valor 0 con probabilidad $1 - p_i$. Esta formulación permite interpretar directamente p_i como la probabilidad de que ocurra un fallo dadas las condiciones operativas asociadas a dicha observación.

La naturaleza discreta de la variable respuesta tiene implicaciones importantes en el modelado. A diferencia de las variables continuas, no resulta adecuado suponer que los errores siguen una distribución normal ni que la relación entre la variable respuesta y las variables explicativas sea lineal en el sentido clásico [6]. En su lugar, el objetivo consiste en modelar la probabilidad de ocurrencia del evento en función de las variables medidas por los sensores, tales como temperatura, vibración, humedad o edad de la máquina.

De este modo, el problema estadístico se reformula como la estimación de la probabilidad condicional

$$P(Y_i = 1 \mid X_i),$$

donde X_i representa el vector de variables explicativas asociado a la observación i . Esta perspectiva probabilística permite no solo clasificar una observación como fallo o no fallo, sino también cuantificar el nivel de riesgo asociado a cada situación operativa, lo cual resulta especialmente relevante en aplicaciones de mantenimiento predictivo [4].

La consideración explícita de la variable respuesta como una variable binaria constituye el punto de partida para el desarrollo de modelos específicos que permitan estimar de forma coherente dichas probabilidades, entre los cuales destaca la regresión logística, que se analiza en los apartados siguientes.

2.1.3 Ejemplos típicos (fallo/no fallo)

Los problemas de clasificación binaria aparecen en numerosos ámbitos de aplicación, tanto en ingeniería como en ciencias de la salud, economía o ciencias sociales. En todos estos casos, la variable de interés representa la ocurrencia o no de un evento, lo que permite formular el problema en términos probabilísticos y aplicar modelos específicos para su análisis [3].

En el ámbito industrial, uno de los ejemplos más representativos es la detección de fallos en maquinaria. A partir de variables medidas por sensores como temperatura, vibración, presión o tiempo de operación se busca determinar si una máquina se encuentra en condiciones normales de funcionamiento o si existe riesgo de fallo. Este tipo de aplicaciones es especialmente relevante en estrategias de mantenimiento predictivo, donde el objetivo consiste en anticipar fallos antes de que se produzcan y minimizar así costes y tiempos de inactividad [4].

En el campo médico, la clasificación binaria se emplea, por ejemplo, para determinar la presencia o ausencia de una enfermedad a partir de variables clínicas o resultados de pruebas diagnósticas. De forma similar, en el ámbito financiero se utiliza para evaluar la probabilidad de impago de un crédito, clasificando a los clientes como solventes o no solventes en función de sus características económicas [6].

En todos estos contextos, el interés no reside únicamente en asignar una etiqueta categórica (0 o 1), sino en estimar la probabilidad asociada a cada caso particular. Esta estimación probabilística permite establecer umbrales de decisión adaptados al nivel de riesgo que se esté dispuesto a asumir, lo que resulta

especialmente útil en entornos donde las decisiones tienen implicaciones económicas o de seguridad significativas [3].

En el presente trabajo, el problema de clasificación binaria se centra en la predicción del riesgo de fallo en máquinas a partir de datos de sensores operativos. La formulación probabilística del problema permite no solo identificar situaciones de mayor riesgo, sino también cuantificar dicho riesgo de forma objetiva, proporcionando una herramienta útil para la toma de decisiones en la gestión y mantenimiento de sistemas industriales.

2.2 Regresión lineal y sus limitaciones

2.2.1 Modelo lineal clásico

El modelo de regresión lineal establece una relación entre una variable respuesta Y y un conjunto de variables explicativas $x_1, x_2, \dots, x_p$ mediante una combinación lineal de estas últimas. En su forma general, el modelo puede expresarse como:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_j x_{ij} + \beta_k x_{ik} + \varepsilon_i.$$

En esta expresión:

- β_0 es el término independiente o intercepto. ($x_{i0} = 1$ para todo i)
- β_j representa el efecto de la variable x_j sobre la variable respuesta.
- ε_i es el término de error asociado a la observación i .

Este modelo resulta adecuado cuando la variable respuesta es continua, ya que permite estimar el valor esperado de Y como una función lineal de las variables explicativas.

Si se intenta aplicar el modelo lineal anterior a una variable respuesta binaria, por ejemplo:

$$Y_i = \begin{cases} 1 & \text{si hay fallo} \\ 0 & \text{si no hay fallo} \end{cases}$$

el modelo adoptaría la forma:

$$E(Y_i | X_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}.$$

Dado que, en el caso binario, la esperanza condicional cumple:

$$E(Y_i | X_i) = P(Y_i = 1 | X_i),$$

la combinación lineal anterior estaría representando directamente la probabilidad de fallo.

Sin embargo, este planteamiento presenta un problema fundamental: la expresión lineal

$$\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ik}$$

puede tomar cualquier valor real, mientras que una probabilidad debe estar comprendida entre 0 y 1. Por tanto, el modelo lineal puede generar predicciones negativas o superiores a uno como se muestra en la Figura 2-1, lo que carece de interpretación probabilística.

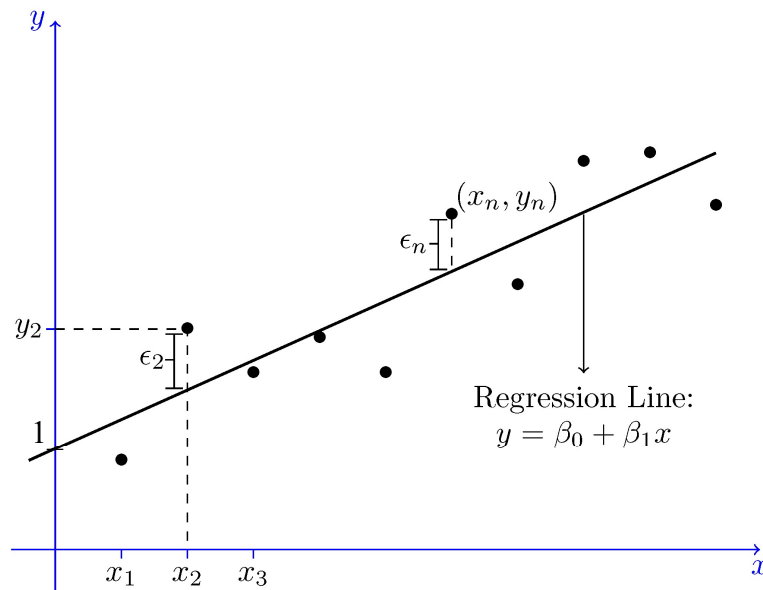


Figura 2-1 Representación de una función lineal con su error asociado (tomada de [11])

2.2.2 Justificación de buscar otro modelo

Además del problema de acotación de las probabilidades, el modelo lineal presenta otras limitaciones cuando se aplica a variables binarias:

1. No linealidad real del fenómeno:

En muchos problemas prácticos, la relación entre las variables explicativas y la probabilidad de ocurrencia del evento no es lineal. La probabilidad suele crecer de forma progresiva y estabilizarse en los extremos, comportamiento que no puede representarse adecuadamente mediante una función lineal.

2. Supuestos del modelo lineal:

El modelo clásico supone que el término de error sigue una distribución normal con varianza constante. Sin embargo, en el caso de una variable binaria, la varianza depende de la probabilidad p_i , ya que:

$$\text{Var}(Y_i | X_i) = p_i(1 - p_i),$$

lo que implica que no es constante. Estas limitaciones justifican la necesidad de introducir un modelo alternativo que permita:

- Mantener la interpretación probabilística.
- Garantizar que las predicciones estén entre 0 y 1,
- Adaptarse mejor a la naturaleza discreta de la variable respuesta.

Este modelo alternativo es la regresión logística, que se introduce en el apartado siguiente.

2.3 Modelo de regresión logística

2.3.1 Definición del modelo

Una vez puesto de manifiesto que la regresión lineal clásica no resulta adecuada para modelar variables respuesta que solo pueden tomar dos valores, es necesario emplear un modelo específicamente diseñado para este tipo de situaciones. En este contexto se introduce la regresión logística, ampliamente utilizada en estadística para problemas de clasificación binaria y estimación de probabilidades.

El objetivo del modelo no es predecir directamente un valor numérico, sino estimar la probabilidad de que ocurra un determinado evento (en este trabajo, el fallo de una máquina) en función de un conjunto de variables explicativas medidas por sensores.

El modelo parte de una combinación lineal de dichas variables:

$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

donde:

- $x_1, x_2, \dots, x_k$ representan las variables observadas del sistema (temperatura, vibración, humedad, edad de la máquina, etc.).
- $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ es el vector de parámetros desconocidos que cuantifican la influencia de cada variable sobre el riesgo de fallo.

Esta expresión lineal no se interpreta directamente como probabilidad, sino como una magnitud intermedia que posteriormente se transforma mediante la función logística, obteniendo:

$$p = P(Y = 1 | X) = \frac{1}{1 + e^{-Z}}$$

donde p representa la probabilidad de que se produzca el fallo de la máquina bajo las condiciones observadas.

De este modo, el modelo permite relacionar variables físicas medibles con la probabilidad de fallo mediante una formulación matemática continua, interpretable y adecuada para su aplicación en estrategias de mantenimiento predictivo y análisis de fiabilidad de sistemas técnicos [6].

2.3.2 Función logística

El modelo logístico introduce la función sigmoide con el objetivo de transformar la combinación lineal de variables en una cantidad que pueda interpretarse como probabilidad.

Como en la regresión lineal, se parte de una expresión de tipo lineal:

$$Z_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} \quad (1)$$

Esta magnitud Z_i puede tomar cualquier valor real, desde $-\infty$ hasta $+\infty$. Sin embargo, una probabilidad debe estar necesariamente comprendida entre 0 y 1, por lo que es necesario aplicar una transformación que convierta esos valores en un rango válido.

Para ello se utiliza la función logística o sigmoide, definida como:

$$p_i = \frac{1}{1 + e^{-Z_i}} \quad (2)$$

donde p_i representa la probabilidad de que ocurra el evento $Y_i = 1$ en la observación i .

La elección de esta función no es arbitraria, sino que responde a su comportamiento matemático:

- Si Z_i toma valores muy grandes y positivos, entonces e^{-Z_i} tiende a 0, y, por tanto:

$$p_i \approx 1$$

lo que indica que la probabilidad del evento (fallo) es muy alta.

- Si Z_i toma valores muy negativos, entonces e^{-Z_i} crece mucho, y:

$$p_i \approx 0$$

lo que indica que la probabilidad de fallo es muy baja.

- Para valores intermedios de Z_i , la función crece de forma suave, describiendo una transición progresiva entre ambos estados.

Este comportamiento en forma de “S” permite modelar situaciones reales en las que el sistema pasa gradualmente de una zona segura a una zona de alto riesgo, en lugar de cambiar de manera brusca.

Se denota por p_i la probabilidad de que ocurra el suceso $Y_i = 1$, siendo la variable respuesta:

$$Y_i = \begin{cases} 1 & \text{si ocurre el fallo} \\ 0 & \text{si no ocurre} \end{cases} \quad (3)$$

El modelo debe comportarse de forma coherente con los datos observados:

- Cuando en una observación se tiene $Y_i = 1$ (ha habido fallo), interesa que la probabilidad estimada p_i sea grande, es decir, cercana a 1.
- Cuando $Y_i = 0$ (no hay fallo), interesa que p_i sea pequeña, es decir, cercana a 0.

Por tanto, el ajuste del modelo consistirá en encontrar los valores del vector de parámetros β que hagan que las probabilidades p_i calculadas mediante la función logística se parezcan lo máximo posible a los resultados reales observados. Esta idea será la base del método de estimación por máxima verosimilitud que se desarrolla posteriormente, para más detalles ver [8].

2.3.3 Interpretación probabilística

Una vez definida la función logística, el modelo puede interpretarse directamente en términos probabilísticos. Para cada observación i , el modelo calcula:

$$p_i = P(Y_i = 1 \mid X_i)$$

es decir, la probabilidad de que ocurra el evento de interés (fallo de la máquina) está condicionada a los valores observados de las variables explicativas x_i .

Dado que la variable respuesta Y_i solo puede tomar dos valores, la distribución de Y_i se modela mediante una variable aleatoria de tipo Bernoulli cuya probabilidad de éxito es precisamente p_i . Esto significa que:

$$P(Y_i = 1) = p_i \text{ y } P(Y_i = 0) = 1 - p_i \quad (4)$$

De este modo, el modelo logístico no proporciona una predicción determinista, sino una medida del riesgo asociado a cada conjunto de condiciones de operación. En lugar de afirmar que una máquina fallará o no fallará, el modelo cuantifica qué probabilidad existe de que falle, lo cual resulta más adecuado en sistemas reales donde intervienen incertidumbres y variabilidad en las mediciones [8] [7].

Esta interpretación probabilística permite utilizar el modelo como herramienta de apoyo a la toma de decisiones. Por ejemplo, si la probabilidad estimada supera un determinado umbral, puede activarse una acción de mantenimiento preventivo antes de que se produzca el fallo, integrando así el análisis estadístico dentro de estrategias de mantenimiento predictivo.

2.3.4 Significado de los parámetros

En el modelo de regresión logística, los coeficientes $\beta_0, \beta_1, \dots, \beta_k$ son parámetros desconocidos que deben estimarse a partir de los datos y que determinan cómo influyen las variables medidas en la probabilidad de fallo.

Recordando la expresión dada por (1) y (2), los parámetros β actúan como factores de ponderación que cuantifican la contribución de cada variable explicativa al riesgo de que ocurra el evento $Y_i = 1$. En particular:

- β_0 (término independiente): representa el valor base del modelo cuando todas las variables explicativas toman valor cero. puede interpretarse como el nivel inicial de riesgo del sistema en ausencia de efectos adicionales.

- β_j (coeficientes asociados a las variables): indican cómo cambia la probabilidad de fallo cuando la variable x_j aumenta, manteniendo las demás constantes.
 - Si $\beta_j > 0$: un aumento de x_j incrementa el riesgo de fallo.
 - Si $\beta_j < 0$: un aumento de x_j reduce la probabilidad de fallo.
 - Si $\beta_j \approx 0$: la variable tiene poca influencia en el comportamiento del sistema.

En el contexto del mantenimiento de maquinaria, estos parámetros permiten identificar qué variables son más determinantes en la aparición de fallos.

Por ejemplo:

- Un coeficiente positivo elevado asociado a la vibración indicaría que incrementos en vibración aumentan notablemente la probabilidad de avería.
- Un coeficiente pequeño en la humedad sugeriría que esta variable apenas afecta al funcionamiento del equipo.
- Un coeficiente positivo en la edad de la máquina reflejaría el efecto del desgaste acumulado.

De esta manera, el modelo no solo predice probabilidades, sino que también proporciona información sobre los factores que más influyen en la fiabilidad del sistema, lo cual resulta de gran utilidad para priorizar acciones de mantenimiento y supervisión técnica [7].

2.4 Estimación de parámetros mediante máxima verosimilitud

2.4.1 Idea de verosimilitud

Una vez formulado el modelo de regresión logística, es necesario estimar el vector de parámetros β que permiten describir la relación entre las variables explicativas y la probabilidad de fallo. Para ello se emplea el método de máxima verosimilitud, ampliamente utilizado en la estimación de modelos estadísticos [7].

Este método consiste en determinar los valores de los parámetros que maximizan la probabilidad de observar los datos disponibles. En el problema considerado, se dispone de un conjunto de observaciones en las que se conocen tanto las variables medidas por los sensores como el resultado real del sistema (fallo o no fallo). El modelo logístico asigna a cada observación una probabilidad estimada de ocurrencia del fallo:

$$p_i = P(Y_i = 1 \mid X_i)$$

La estimación por máxima verosimilitud selecciona los parámetros del modelo β , de forma que las probabilidades calculadas sean lo más coherentes posible con los resultados observados, obteniendo así el ajuste que mejor representa el comportamiento real del sistema analizado.

Este procedimiento proporciona un criterio objetivo y fundamentado matemáticamente para el ajuste del modelo, permitiendo su aplicación en problemas de clasificación binaria y, en particular, en el análisis predictivo de fallos en maquinaria.

2.4.2 Función de verosimilitud

Dado un conjunto de n observaciones independientes correspondientes al funcionamiento de una máquina. Para cada observación i se dispone de:

- Un vector de variables explicativas

$$X_i = (x_{i0}, x_{i1}, x_{i2}, \dots, x_{ik}) \in \mathbb{R}^{k+1}$$

que recoge las mediciones de los sensores u otros dispositivos.

- Una variable respuesta

$$y_i \in \{0,1\}$$

El modelo de regresión logística establece que la probabilidad de fallo en la observación i viene dada por:

$$p_i = P(Y_i = 1 \mid X_i) = \frac{1}{1 + e^{-z_i}}$$

donde

$$Z_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}.$$

Dado que Y_i solo puede tomar los valores 0 ó 1, cada observación sigue una distribución de Bernoulli con parámetro p_i . Por tanto:

$$P(Y_i = y_i) = p_i^{y_i} (1 - p_i)^{1-y_i}$$

Esta expresión recoge simultáneamente los dos casos posibles:

- Si $y_i = 1 \Rightarrow P(Y_i = 1) = p_i$
- Si $y_i = 0 \Rightarrow P(Y_i = 0) = 1 - p_i$

Considerando las n observaciones independientes, la función de verosimilitud del modelo se define como el producto de las probabilidades individuales:

$$L(\beta) = (\beta_0, \beta_1, \dots, \beta_k) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}.$$

Esta función mide la probabilidad de observar el conjunto de datos reales en función de los parámetros del modelo. En consecuencia, el problema de estimación consiste en encontrar el vector de parámetro β que hacen máxima dicha verosimilitud, es decir, aquellos que hacen que los datos observados sean los más probables posible; de tal forma que los casos que han ocurrido tengan alta probabilidad y los que no han ocurrido tengan baja probabilidad [7].

2.4.3 Log- verosimilitud

Trabajar directamente con el producto en la función de verosimilitud resulta poco manejable, especialmente cuando el número de observaciones es grande. Por ello, se considera su logaritmo natural, definiendo la denominada función de log-verosimilitud.

$$\ell(\beta_0, \beta_1, \dots, \beta_k) = \log L(\beta).$$

Aplicando propiedades del logaritmo, el producto se transforma en una suma:

$$\ell(\beta) = \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (6)$$

El uso del logaritmo no modifica el problema de estimación, ya que la función logaritmo es estrictamente creciente. Esto significa que si se tienen dos valores a y b tales que:

$$a > b \Rightarrow \log(a) > \log(b),$$

por lo que el punto donde la función original alcanza su máximo coincide exactamente con el punto donde lo alcanza su logaritmo. En consecuencia:

$$\begin{aligned} \arg \max L(\beta) &= \arg \max \log L(\beta). \\ \beta &\in \mathbb{R}^{k+1} \quad \beta \in \mathbb{R}^{k+1} \end{aligned}$$

Es decir, maximizar la verosimilitud o maximizar la log-verosimilitud conduce a los mismos valores estimados de los parámetros.

Las ventajas que nos proporciona esta transformación son las siguientes:

- Convierte productos de muchos términos en sumas, lo que simplifica el tratamiento analítico.
- Evita problemas numéricos derivados de multiplicar probabilidades muy pequeñas.
- Facilita la aplicación de métodos de optimización iterativos para estimar los parámetros del modelo.

Por tanto, en la práctica, el ajuste del modelo logístico se realiza maximizando la log-verosimilitud en lugar de la verosimilitud directa, sin que ello altere el resultado final de la estimación [7].

2.4.4 Modelo de optimización logística

Una vez definida la función de log-verosimilitud, la función (6) de los parámetros del modelo consiste en determinar los valores $\beta_0, \beta_1, \dots, \beta_k$ que la maximizan. Este proceso se conoce como el criterio de optimización asociado al método de máxima verosimilitud [7].

Matemáticamente, el problema puede expresarse como la búsqueda de los parámetros que proporcionan el mayor valor de la función:

$$\max_{\beta \in \mathbb{R}^{k+1}} \ell(\beta) = \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (7)$$

A diferencia de otros modelos más simples, esta función no permite obtener una solución analítica directa. Es decir, no existe una expresión cerrada que proporcione los coeficientes del modelo de forma explícita. Por ello, la estimación debe realizarse mediante métodos numéricos iterativos, que van ajustando progresivamente los valores de los parámetros hasta alcanzar el máximo de la función.

Estos algoritmos parten de una estimación inicial de los coeficientes y, en cada iteración, evalúan cómo varía la log-verosimilitud al modificar dichos valores, desplazándose en la dirección que produce una mejora del ajuste. El procedimiento finaliza cuando los cambios entre iteraciones sucesivas son suficientemente pequeños, lo que indica que se ha alcanzado el óptimo.

Este enfoque permite ajustar el modelo logístico incluso en presencia de múltiples variables explicativas y grandes volúmenes de datos, siendo la base de los métodos implementados en herramientas de cálculo científico y software estadístico empleados en aplicaciones reales de clasificación y mantenimiento predictivo.

2.5 Uso de herramientas computacionales

2.5.1 Papel de Matlab en el ajuste del modelo

La estimación de los parámetros del modelo de regresión logística requiere la aplicación de algoritmos numéricos iterativos que permitan resolver el modelo (7). Debido a la complejidad de estos cálculos, resulta necesario emplear herramientas de programación y cálculo científico que automaticen dicho proceso.

En el presente trabajo se utiliza MATLAB como entorno de desarrollo para la implementación del modelo. Este software permite:

- Importar y gestionar los datos procedentes de los sensores de operación.
- Implementar el ajuste del modelo logístico mediante algoritmos de optimización numérica.
- Calcular las probabilidades estimadas de fallo para cada observación.
- Analizar la influencia de las variables explicativas a través de los coeficientes obtenidos.

El uso de MATLAB facilita, por tanto, la aplicación práctica del modelo teórico desarrollado, permitiendo realizar el entrenamiento del modelo de forma eficiente y reproducible, así como modificar o ampliar el análisis cuando se disponga de nuevos datos [7].

2.5.2 Importancia de la simulación y el análisis numérico

El análisis de datos mediante simulación constituye una herramienta esencial en problemas de mantenimiento predictivo, donde el comportamiento del sistema depende de múltiples variables y presenta un componente de incertidumbre. A través del tratamiento numérico de los datos es posible evaluar distintos escenarios de funcionamiento y estimar el riesgo de fallo bajo diversas condiciones operativas [7].

La simulación permite:

- Validar el modelo estadístico utilizando conjuntos de datos reales o sintéticos.
- Analizar la sensibilidad del sistema ante variaciones en las variables medidas.
- Identificar patrones que incrementan la probabilidad de fallo.
- Apoyar la toma de decisiones en estrategias de mantenimiento preventivo.

De este modo, el uso combinado de modelización estadística y herramientas computacionales transforma los datos recogidos por los sistemas de monitorización en información útil para la gestión de la maquinaria, constituyendo la base del enfoque de mantenimiento basado en condición.

3 DESARROLLO DEL TFG

Este capítulo describe la metodología seguida en el desarrollo del trabajo. En primer lugar, se presentan los conjuntos de datos utilizados, así como las variables consideradas en cada caso. A continuación, se explica el proceso de modelización mediante regresión logística, incluyendo el ajuste del modelo en MATLAB y el cálculo de los principales indicadores estadísticos. Finalmente, se introduce el análisis de correlación entre variables, utilizado para mejorar la selección de variables y la interpretación de los resultados.

3.1 Planteamiento del problema

El objetivo del estudio es predecir el riesgo de fallo de una máquina industrial a partir de datos obtenidos mediante sensores u otros dispositivos instalados en el sistema. Estos dispositivos registran diferentes variables relacionadas con el funcionamiento de la máquina, tales como temperatura, presión, vibración, humedad o edad de la máquina.

A partir de estas mediciones se pretende construir un modelo matemático capaz de estimar la probabilidad de que se produzca un fallo en el sistema. Este tipo de análisis resulta de gran interés en el ámbito del mantenimiento predictivo, ya que permite anticipar posibles fallos y planificar intervenciones antes de que se produzcan averías que puedan interrumpir el proceso productivo.

Para abordar este problema se utilizará el modelo de regresión logística (7), que permite modelizar la probabilidad de ocurrencia de un evento binario (fallo o no fallo) en función de un conjunto de variables explicativas.

3.2 Modelo de regresión logística

El modelo matemático que usaremos para predecir la probabilidad de fallo de una maquinaria es el descrito en el estado del arte, el cual viene dado en (7), donde n es el número de observaciones (o datos), las variables explicativas $x_{i1}, x_{i2}, \dots, x_{ik}$ correspondientes a las mediciones de los dispositivos, β es el vector de parámetros a determinar (incógnitas del problema), donde la probabilidad de fallo se obtiene mediante la ecuación (2) que depende del vector de parámetros β , según la ecuación (1).

En términos generales, los métodos iterativos se utilizan para resolver problemas generales de la forma:

$$\min_{X \in \mathbb{R}^n} f(X)$$

donde $f: \mathbb{R}^n \rightarrow \mathbb{R}$ es una función no lineal, continua y dos veces diferenciable. Si X^* es un mínimo de f , deben cumplirse ciertas condiciones de optimalidad.

Condición necesaria si X^* es un mínimo de f entonces:

$$\nabla f(X^*) = 0$$

Condición suficiente: si la matriz Hessiana $\nabla^2 f(X^*)$ es definida positiva y $\nabla f(X^*) = 0$, entonces X^* corresponde a un mínimo local de f .

3.3 Método del gradiente

El método del gradiente es un algoritmo iterativo basado en la dirección de máximo descenso de la función objetivo de f .

A partir de un punto inicial X^0 , el método genera una sucesión de aproximaciones mediante la expresión:

$$X^{H+1} = X^H + \alpha_H d^H$$

donde:

$d^H = -\nabla f(X^H)$ es la dirección de descenso

α_H es el tamaño de paso de la iteración, que se escoge de forma que garantice convergencia a un mínimo de f (condición Armijo).

El proceso iterativo continúa hasta que se cumple un criterio de parada basado, por ejemplo, en la magnitud del gradiente. Para más detalles ver [9]

3.4 Método de Newton

El método de Newton es un algoritmo de optimización que utiliza información de primer y segundo orden de la función objetivo.

En cada iteración se resuelve el sistema:

$$\nabla^2 f(X^H) \Delta X^H = -\nabla f(X^H)$$

y se actualiza la solución mediante:

$$X^{H+1} = X^H + \alpha_H \Delta X^H$$

donde α_H se escoge para garantizar convergencia a un mínimo de f (condición Armijo). [9]

Este método suele presentar una velocidad de convergencia superior al método del gradiente, ya que incorpora información sobre la curvatura de la función. Para más detalles ver [9].

En el caso de la regresión logística, el método de Newton se utiliza para maximizar la función de log-verosimilitud del modelo (7).

El gradiente de la función puede escribirse como:

$$\nabla l(\beta) = X^T (y - p)$$

donde:

X es la matriz de variables explicativas, que se describe como.

$$X = \begin{pmatrix} x_{10} & x_{11} & x_{12} & \dots & x_{1k} \\ x_{20} & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n0} & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times (k+1)}$$

$y = (y_1, y_2, \dots, y_n)$ es el vector de observaciones

$p = (p_1, p_2, \dots, p_n)$ es el vector de probabilidades estimadas.

La matriz Hessiana viene dada por:

$$\nabla^2 l(\beta) = -X^T W X$$

donde W es una matriz diagonal cuyos elementos son:

$$W = \begin{pmatrix} p_1(1-p_1) & 0 & 0 & \dots & 0 \\ 0 & p_2(1-p_2) & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & p_n(1-p_n) \end{pmatrix} \in \mathbb{R}^{n \times n}$$

El método de Newton genera iterativamente nuevas estimaciones de los parámetros hasta alcanzar la convergencia del algoritmo.

3.5 Descripción de los datos

3.5.1 Origen y estructura de los conjuntos de datos

Para el desarrollo del modelo de regresión logística se dispone de seis conjuntos de datos independientes obtenidos a partir de registros del estado operativo de una máquina monitorizada mediante sensores y otros dispositivos.

Con el objetivo de caracterizar estructuralmente los conjuntos de datos disponibles, se realizó un análisis exploratorio inicial que permitió determinar el número de observaciones, el número de variables, la proporción de fallos, y darle un nombre de referencia a cada archivo. La Tabla 3-1 muestra las características de los conjuntos de datos a utilizar en este trabajo.

Referencia	Nº Observaciones	Variables	Proporción de fallos
CD1 (Mill)	1000	Temperatura, humedad, vibración, potencia (4)	0.30
CD1 (Lathe)	1000	Temperatura, humedad, vibración, potencia (4)	0.28
CD1 (Drill)	1000	Temperatura, humedad, vibración, potencia (4)	0.26
CD2	3000	Temperatura, vibración, potencia, presión, edad (5)	0.0993
CDS1	700	Temperatura, humedad, vibración, potencia, presión, edad (6)	0.5143
CDS2	700	Temperatura, humedad, vibración, potencia, presión, edad (6)	0.3629
CDS3	12000	Temperatura, humedad, vibración, potencia, presión, edad, carga (7)	0.3345
CDS4	500	Temperatura, vibración, potencia, presión, edad (5)	0.1600

Tabla 3-1 Características de las bases de datos

El conjunto de datos utilizado en este trabajo se basa en el *dataset* público de predicción de fallos de maquinaria disponible en la plataforma *Kaggle* [10]. A partir de este *dataset*, se han generado distintas

versiones y subconjuntos de datos, con el objetivo de analizar el comportamiento del modelo bajo diferentes configuraciones y estructuras de variables.

Del CD1 se pueden extraer 3 conjuntos de datos diferentes correspondientes a tres tipos de máquinas y los denotaremos como $CDIi$ con $i=1,2,3$.

Del análisis presentado se observa:

- Una variabilidad considerable en el tamaño muestral de los conjuntos.
- Diferencias en la proporción de fallos, comprendidas entre aproximadamente el 10% y el 30%.
- Diferencias estructurales en el número de variables disponibles.

Estas características justifican la necesidad de un proceso de homogeneización previo al ajuste del modelo logístico, como hemos mostrado en la Tabla 3-1.

3.6 Preparación de los datos

3.6.1 Selección de variables explicativas

Tal y como se ha observado en el análisis exploratorio preliminar, los conjuntos de datos disponibles no presentan una estructura completamente homogénea. En particular, se identificaron diferencias en el número y tipo de variables incluidas en cada archivo.

El modelo se construye a partir de un vector de variables explicativas:

$$X = (x_1, x_2, x_3, x_4, x_5, x_6, x_7,) \\ = (\text{Temperatura, humedad, vibración, potencia, presión, edad, carga})$$

Entonces dependiendo de los datos de estudio se tienen diferentes vectores:

$$X = (x_1, x_2, x_3, x_4) = (\text{Temperatura, humedad, vibración, potencia})(\text{CD1})$$

$$X = (x_1, x_3, x_4, x_5, x_6) = (\text{Temperatura, vibración, potencia, presión, edad})(\text{CD2})$$

$$X = (x_1, x_2, x_3, x_4, x_5, x_6) = \\ (\text{Temperatura, humedad, vibración, potencia, presión, edad})(\text{CDS1})$$

$$X = (x_1, x_2, x_3, x_4, x_5, x_6) = \\ (\text{Temperatura, humedad, vibración, potencia, presión, edad})(\text{CDS2})$$

$$X = (x_1, x_2, x_3, x_4, x_5, x_6, x_7) = \\ (\text{Temperatura, humedad, vibración, potencia, presión, edad, carga})(\text{CDS3})$$

$$X = (x_1, x_3, x_4, x_5, x_6) = (\text{Temperatura, vibración, potencia, presión, edad})(\text{CDS4})$$

3.7 Implementación del modelo en Matlab

La implementación del modelo de regresión logística se realizó en el entorno MATLAB, utilizando funciones específicas para modelos lineales generalizados. El código desarrollado permite:

- Cargar distintos conjuntos de datos.
- Seleccionar subconjuntos de observaciones y variables.
- Analizar la correlación entre variables.
- Ajustar el modelo mediante máxima verosimilitud.
- Calcular probabilidades estimadas.
- Evaluar la capacidad discriminante mediante la curva ROC.

A continuación, se describe la correspondencia entre cada bloque del código y la formulación matemática presentada en el Capítulo 2.

3.7.1 Estimación de parámetros mediante “glmfit”

El ajuste del modelo se realiza mediante la función:

`glmfit(X, y, 'binomial', 'link', 'logit')`

Que ajusta modelos lineales generalizados mediante máxima verosimilitud (7).

Por tanto, la función `glmfit` estima el vector de parámetros:

$$\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_k) \text{ con } k = 4, 5, 6$$

mediante la maximización numérica de la log-verosimilitud descrita en el apartado 2.4.

Internamente, MATLAB emplea un algoritmo iterativo basado en métodos de tipo Newton-Raphson o *Iteratively Reweighted Least Squares* (IRLS), resolviendo el problema de optimización hasta alcanzar convergencia.

El resultado devuelve:

- Los coeficientes estimados $\hat{\beta}$.
- El valor de la desviación.
- Estadísticos asociados (errores estándar, `stats.t`, `stats.se`, etc.).

De este modo, la implementación computacional materializa directamente la formulación teórica desarrollada previamente.

3.7.2 Cálculo de probabilidades mediante “glmval”

Una vez estimados los parámetros, se procede al cálculo de las probabilidades estimadas de fallo utilizando:

`p=glmval(y, X, 'logit')`

Desde el punto de vista matemático, esta instrucción calcula:

$$p_i = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_4 x_{ik})}}$$

Es decir, aplica la función logística a la combinación lineal estimada.

Es importante destacar que el modelo no produce directamente una clasificación binaria, sino una probabilidad continua de fallo asociada a cada observación. La decisión final depende del umbral que se establezca, aspecto que se analiza mediante la curva ROC.

3.8 Metodología de evaluación

3.8.1 Análisis de correlación entre variables

Con el objetivo de analizar la relación existente entre las variables explicativas, se realizó un estudio previo de correlación utilizando MATLAB. Para ello, se empleó la función *corrcoef*, que permite calcular tanto la matriz de correlación lineal de Pearson (R) como la matriz de p-valores asociada (P) a partir de los datos.

La matriz R contiene coeficientes de correlación entre -1 y 1, donde valores próximos a ± 1 indican una relación lineal fuerte entre variables, mientras que valores cercanos a 0 indican ausencia de correlación. En este trabajo, se consideró que existe una correlación relevante cuando el valor absoluto del coeficiente supera 0.7.

Por otro lado, la matriz de p-valores P permite evaluar la significancia estadística de dichas relaciones. Se consideró que una correlación es significativa cuando el valor p es inferior a 0.05.

Adicionalmente, mediante el uso de la función *heatmap* de MATLAB, se representó gráficamente la matriz de correlación, lo que facilita la identificación visual de posibles patrones de dependencia entre variables.

A partir de este análisis, en algunos conjuntos de datos se procedió a la eliminación de variables redundantes o poco relevantes, con el objetivo de simplificar el modelo y mejorar su estabilidad e interpretabilidad.

3.8.2 Construcción de la curva ROC mediante “*perfcurve*”

Para evaluar la capacidad discriminante del modelo se utiliza la función:

$$[X_f, y_f, T, AUC] = \text{perfcurve}(y, p, 1)$$

Esta instrucción construye la curva ROC variando sistemáticamente el umbral de clasificación y calculando:

- Tasa de verdaderos positivos (sensibilidad).
- Tasa de falsos positivos.

Matemáticamente, este procedimiento evalúa la capacidad del modelo para ordenar correctamente las observaciones según su probabilidad estimada de fallo.

La curva ROC (*Receiver Operating Characteristic*) se construye a partir de dos magnitudes fundamentales: la tasa de falsos positivos y la tasa de verdaderos positivos. Estas cantidades permiten evaluar la capacidad de un modelo para distinguir entre observaciones positivas y negativas.

La tasa de falsos positivos se define como:

$$X_f = \frac{FP}{FP + TN}$$

donde:

- *FP* representa el número de falsos positivos, es decir, observaciones negativas que el modelo clasifica incorrectamente como positivas.
- *TN* representa el número de verdaderos negativos, es decir, observaciones negativas que el modelo clasifica correctamente.

Por tanto, esta magnitud mide la proporción de observaciones negativas que se clasifican erróneamente como positivas.

Por otra parte, la tasa de verdaderos positivos se define como

$$Y_f = \frac{TP}{TP + FN}$$

donde:

- *TP* representa el número de verdaderos positivos, es decir, observaciones positivas correctamente clasificadas.
- *FN* representa el número de falsos negativos, es decir, observaciones positivas que el modelo clasifica incorrectamente como negativas.

Esta cantidad mide la proporción de observaciones positivas que el modelo es capaz de identificar correctamente.

La clasificación final depende del valor de un umbral de decisión *T* aplicado a las probabilidades estimadas por el modelo.

En este trabajo se considera clase positiva la ocurrencia de un fallo en la máquina, mientras que la clase negativa corresponde al funcionamiento normal del sistema. El modelo logístico proporciona

probabilidades estimadas de fallo, que posteriormente se transforman en una clasificación binaria mediante la aplicación de un umbral de decisión T .

Cuando el valor del umbral T es grande, únicamente las observaciones con probabilidades muy altas se clasifican como positivas, por lo que casi ninguna observación se clasifica como positiva.

Por el contrario, cuando el umbral T es pequeño, muchas observaciones superan dicho umbral, lo que provoca que un número elevado de observaciones se clasifique como positivo.

Al variar el valor del umbral T se obtienen diferentes valores de X_f e Y_f , lo que permite construir la curva ROC, que representa gráficamente la relación entre la tasa de falsos positivos y la tasa de verdaderos positivos para distintos umbrales de clasificación.

3.8.3 Cálculo del área bajo la curva (AUC)

Una vez construida la curva ROC, es habitual resumir su información mediante una medida escalar conocida como Área Bajo la Curva (*Area Under the Curve*, AUC).

Esta cantidad se define como el área comprendida bajo la curva ROC y puede expresarse analíticamente como

$$AUC = \int_0^1 Y_f(X_f) d(X_f)$$

Desde el punto de vista interpretativo, el AUC mide la capacidad del modelo para discriminar entre observaciones positivas y negativas.

En términos probabilísticos, el AUC puede interpretarse como la probabilidad de que el modelo asigne una probabilidad de fallo mayor a una observación que realmente presenta fallo que a una observación que no lo presenta.

El valor del AUC toma valores entre 0 y 1:

- $AUC = 0.5 \rightarrow$ el modelo no tiene capacidad discriminante (equivalente a una clasificación aleatoria).
- $0.5 < AUC < 0.7 \rightarrow$ capacidad discriminante baja.
- $0.7 < AUC < 0.8 \rightarrow$ capacidad discriminante aceptable.
- $0.8 < AUC < 0.9 \rightarrow$ buen rendimiento del modelo.
- $AUC > 0.9 \rightarrow$ capacidad discriminante excelente.

Por tanto, el análisis del AUC permite comparar distintos modelos y evaluar su capacidad predictiva de forma independiente del umbral de clasificación elegido.

3.8.4 Interpretación de los parámetros estadísticos del modelo

Una vez estimado el modelo de regresión logística mediante la función `glmfit` de MATLAB, el procedimiento proporciona un conjunto de estadísticas que permiten evaluar la significación y fiabilidad de los parámetros estimados. Estas estadísticas se encuentran almacenadas en la estructura `stats`.

- El vector `stats.beta` contiene los coeficientes estimados del modelo, es decir, los valores de los parámetros β_i asociados a cada variable explicativa.

Estos coeficientes determinan la influencia de cada variable sobre la probabilidad de que ocurra el evento de interés (en este caso, el fallo de la máquina).

- El vector `stats.se` representa el error estándar asociado a cada coeficiente estimado β_i ,

El error estándar mide la incertidumbre de la estimación de cada parámetro. Cuanto menor es el error estándar, mayor es la precisión con la que se ha estimado el coeficiente correspondiente.

Este valor se utiliza posteriormente para calcular el estadístico t.

- El estadístico t se calcula como

$$t_i = \frac{\beta_i}{se_i}$$

donde:

β_i es el coeficiente estimado del modelo.

se_i es el error estándar asociado a dicho coeficiente.

Este estadístico mide cuántas desviaciones estándar se encuentra el coeficiente estimado respecto al valor cero.

Valores elevados del estadístico t_i en valor absoluto indican que el coeficiente está significativamente alejado de cero, lo que sugiere que la variable asociada tiene una influencia relevante en el modelo.

- El valor *stats.p* representa la probabilidad asociada al estadístico t bajo la hipótesis nula

$$H_0: \beta_i = 0$$

Esta hipótesis plantea que la variable explicativa correspondiente no tiene efecto sobre la variable respuesta.

La interpretación habitual es la siguiente:

- Si $p_i < 0.05$, el coeficiente se considera estadísticamente significativo.
- Si $p_i > 0.05$, el coeficiente se considera no significativo.

Por tanto, el análisis de los valores p permite identificar qué variables explicativas contribuyen realmente a la predicción del fallo.

- El parámetro *stats.sfit* representa la desviación estándar de los residuos del modelo.

Los residuos se definen como la diferencia entre el valor observado y el valor estimado por el modelo:

$$residuo_i = y_i - \hat{p}_i$$

donde:

y_i es el valor observado.

$\hat{p}_i$ es la probabilidad estimada por el modelo para la observación i.

La desviación estándar de los residuos puede expresarse como

$$s = \sqrt{\frac{1}{(n - k)} \sum_{i=1}^n (y_i - \hat{p}_i)^2}$$

donde:

- n es el número total de observaciones,
- k es el número de parámetros del modelo,
- $\hat{p}_i$ representa la probabilidad estimada para la observación i.

Esta medida cuantifica la desviación media entre los valores observados y las predicciones del modelo. En general, cuanto menor es este valor, mejor es el ajuste del modelo a los datos.

4 RESULTADOS OBTENIDOS

Este capítulo presenta los resultados al problema de regresión logística (7) para el conjunto de datos descrito en la Tabla 3-1.

En las tablas de resultados se incluyen, además de los coeficientes estimados, los errores estándar, el estadístico t y los valores p asociados a cada parámetro. El estadístico t se obtiene como el cociente entre el coeficiente estimado y su error estándar, y se utiliza para evaluar la significancia estadística de las variables mediante el correspondiente valor p .

En el análisis posterior se pone especial atención en los coeficientes estimados y en los valores p , ya que estos permiten interpretar de forma directa la influencia de cada variable en la probabilidad de fallo y determinar si dicha influencia resulta estadísticamente significativa. Por este motivo, aunque el error estándar y el estadístico t se incluyen en las tablas para completar la información estadística del modelo, su interpretación se refleja indirectamente a través del valor p , por lo que no se analizan de forma individual en cada caso. De manera similar, la desviación estándar del ajuste se presenta como indicador complementario del modelo, mientras que la capacidad predictiva se evalúa principalmente mediante el área bajo la curva ROC (AUC), que resulta más adecuada para analizar el rendimiento del modelo de clasificación. También evaluaremos en algunos casos la correlación para ver qué datos podemos eliminar y así mejorar el conjunto de datos obteniendo unos mejores análisis.

4.1 Resultados del conjunto de datos CD1

Con el objetivo de analizar la relación entre las variables operativas de la máquina y la probabilidad de fallo, se ajustó un modelo de regresión logística utilizando las observaciones correspondientes a las máquinas del conjunto de datos CD1.

El modelo considera como variables explicativas el vector $X = (x_1, x_2, x_3, x_4)$, descritas en el anterior capítulo.

4.1.1 CD1₁ Máquina tipo Mill

En las siguientes tablas se muestran los datos obtenidos y a continuación un análisis de estos:

Variable	Descripción	Coeficiente β	Error estándar	Estadístico t	p-valor
β_0	Intercepto	-2.7620	1.8622	-1.4832	0.1380
$\beta_1 (x_1)$	Temperatura	0.0281	0.0124	2.2686	0.0233

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_2 (x_2)	Vibración	0.0083	0.0247	0.3378	0.7355
β_3 (x_3)	Potencia	-0.0390	0.0615	-0.6341	0.5260
β_4 (x_4)	Humedad	-0.0072	0.0255	-0.2825	0.7775

Tabla 4-1 Resultados máquina tipo Mill

Métrica	Valor
Área bajo la curva ROC (AUC)	0.588
Desviación estándar del ajuste (sfit)	1.0069
Número de observaciones	1000
Número de variables explicativas	4

Tabla 4-2 Indicadores de máquina tipo Mill

Como podemos ver en la Tabla 4-1 y la Tabla 4-2, el coeficiente asociado a la variable x_1 (temperatura) presenta un valor positivo ($\beta_1 = 0.0281$) y un valor p igual a 0.0233, inferior al nivel de significancia habitual de 0.05. Este resultado indica que la temperatura tiene una influencia estadísticamente significativa en la probabilidad de fallo de la máquina. En términos prácticos, este coeficiente positivo sugiere que, manteniendo constantes las demás variables del modelo, incrementos en la temperatura se asocian con un aumento en el riesgo de que se produzca un fallo en el sistema.

Por el contrario, las variables x_2 (vibración), x_3 (consumo de potencia) y x_4 (humedad) presentan valores p superiores al nivel de significancia considerado. Esto indica que, dentro del conjunto de datos analizado, no existe evidencia estadística suficiente para afirmar que dichas variables tengan una influencia significativa en la probabilidad de fallo de la máquina. No obstante, esto no implica necesariamente que estas variables no estén relacionadas con el fenómeno físico del fallo, sino que su efecto no puede ser identificado con claridad a partir de los datos disponibles o puede estar condicionado por la presencia de otros factores no incluidos en el modelo.

En conjunto, el análisis de los coeficientes sugiere que la temperatura aparece como la variable operativa más relevante para explicar el riesgo de fallo en las máquinas tipo Mill dentro del conjunto de datos CD1, mientras que las demás variables consideradas presentan una contribución más limitada en el comportamiento del modelo.

Con el fin de evaluar la capacidad del modelo para distinguir entre situaciones de fallo y funcionamiento normal, se analizó la curva ROC correspondiente al modelo ajustado, que se muestra en la Figura 4-1.

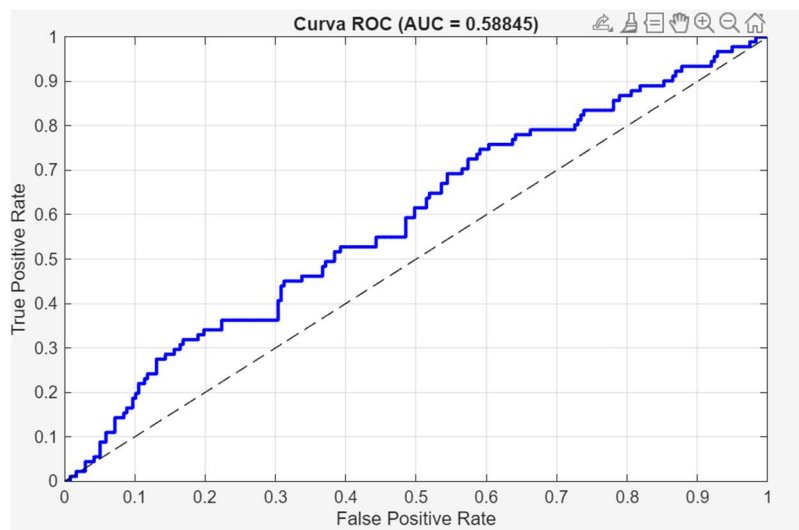


Figura 4-1 Curva ROC- $CD1_1$ (figura tomada de Matlab)

La figura 4-1 representa la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para diferentes valores del umbral de clasificación aplicado a las probabilidades estimadas por el modelo.

El área bajo la curva ROC obtenida fue $AUC = 0.588$, lo que indica que el modelo presenta una capacidad discriminante limitada, aunque ligeramente superior a la de un clasificador aleatorio. En términos probabilísticos, este resultado puede interpretarse como que existe aproximadamente un 58.8 % de probabilidad de que el modelo asigne una mayor probabilidad de fallo a una observación que realmente presenta fallo que a una observación que no lo presenta.

4.1.2 $CD1_2$ Máquina tipo Lathe

A continuación, se presentan los resultados obtenidos del análisis realizado para este conjunto de datos:

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_0	Intercepto	1.4523	1.7616	0.8244	0.4097
$\beta_1 (x_1)$	Temperatura	-0.0092	0.0121	-0.7594	0.4476
$\beta_2 (x_2)$	Vibración	-0.0292	0.0239	-1.2206	0.2222
$\beta_3 (x_3)$	Potencia	0.0170	0.0604	0.2818	0.7781
$\beta_4 (x_4)$	Humedad	-0.0099	0.0214	-0.4640	0.6426

Tabla 4-3 Resultados máquina tipo Lathe

Métrica	Valor
Área bajo la curva ROC (AUC)	0.553
Desviación estándar del ajuste (sfit)	1.0067
Número de observaciones	1000
Número de variables explicativas	4

Tabla 4-4 Indicadores de máquina tipo Lathe

Como se observa en la Tabla 4-3 y la Tabla 4-4, ninguno de los coeficientes asociados a las variables explicativas presenta valores p inferiores al nivel de significancia habitual. Esto indica que, para el conjunto de datos analizado, no se puede afirmar que las variables consideradas tengan una influencia estadísticamente significativa en la probabilidad de fallo de las máquinas tipo Lathe.

En particular, la variable x_1 (temperatura) presenta un coeficiente negativo ($\beta_1 = -0.0092$) y un valor p igual a 0.4476, lo que indica que no existe evidencia estadística suficiente para considerar que esta variable tenga una relación significativa con el riesgo de fallo en este conjunto de datos. De manera similar, la variable x_2 (vibración) presenta un coeficiente ($\beta_2 = -0.0292$) con un valor p igual a 0.2222, mientras que las variables x_3 (consumo de potencia) y x_4 (humedad) presentan coeficientes ($\beta_3 = 0.0170$ y $\beta_4 = -0.0099$) con valores p de 0.7781 y 0.6426 respectivamente. En todos los casos, los valores p indican que no se puede rechazar la hipótesis nula de que los coeficientes sean iguales a cero.

En conjunto, el análisis de los coeficientes sugiere que ninguna de las variables operativas consideradas presenta una relación estadísticamente significativa con el riesgo de fallo en las máquinas tipo Lathe dentro del conjunto de datos CD1. Este resultado puede deberse a que el fenómeno de fallo en este tipo de máquina esté influenciado por otros factores no incluidos en el modelo o a que la relación entre las variables analizadas y el fallo sea más compleja.

Con el fin de evaluar la capacidad del modelo para distinguir entre situaciones de fallo y funcionamiento normal, se analizó la curva ROC correspondiente al modelo ajustado, que se muestra en la Figura 4-2.

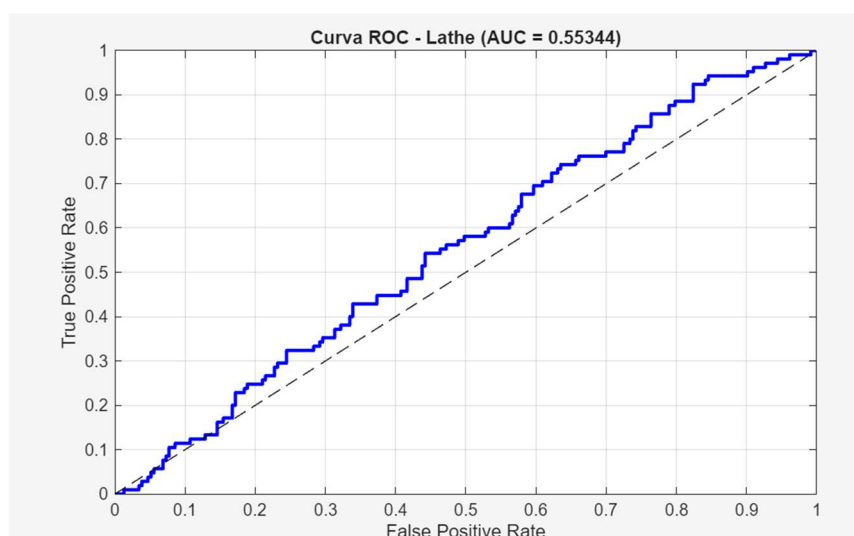


Figura 4-2 curva ROC- $CD1_2$ (figura tomada de Matlab)

El área bajo la curva ROC obtenida fue $AUC = 0.553$, lo que indica que el modelo presenta una capacidad discriminante baja, muy cercana a la de un clasificador aleatorio. En términos probabilísticos, este resultado puede interpretarse como que existe aproximadamente un 55.3 % de probabilidad de que el modelo asigne una mayor probabilidad de fallo a una observación que realmente presenta fallo que a una observación que no lo presenta.

4.1.3 $CD1_3$ Máquina tipo Drill

A continuación, se muestran los resultados del modelo de regresión logística ajustado para este conjunto de datos:

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_0	Intercepto	-2.1648	1.8702	-1.1575	0.2471

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
$\beta_1 (x_1)$	Temperatura	-0.0007	0.0128	-0.0527	0.9580
$\beta_2 (x_2)$	Vibración	0.0144	0.0247	0.5844	0.5589
$\beta_3 (x_3)$	Potencia	0.0858	0.0631	1.3613	0.1734
$\beta_4 (x_4)$	Humedad	-0.0059	0.0244	-0.2398	0.8105

Tabla 4-5 Resultados máquina tipo Drill

Métrica	Valor
Área bajo la curva ROC (AUC)	0.557
Desviación estándar del ajuste (sfit)	1.0082
Número de observaciones	1000
Número de variables explicativas	4

Tabla 4-6 Indicadores de máquina tipo Drill

Como se observa en la Tabla 4-5 y la Tabla 4-6, ninguno de los coeficientes asociados a las variables explicativas presenta valores p inferiores al nivel de significancia habitual, al igual que en el CD_{12} , teniendo las mismas consecuencias.

En particular, la variable x_1 (temperatura) presenta un coeficiente muy próximo a cero ($\beta_1 = -0.0007$) y un valor p igual a 0.9580, lo que indica que no existe evidencia estadística suficiente para considerar que esta variable tenga una relación significativa con el riesgo de fallo en este conjunto de datos. De manera similar, la variable x_2 (vibración) presenta un coeficiente positivo ($\beta_2 = 0.0144$) con un valor p igual a 0.5589, mientras que las variables x_3 (consumo de potencia) y x_4 (humedad) presentan coeficientes ($\beta_3 = 0.0858$ y $\beta_4 = -0.0059$) con valores p de 0.1734 y 0.8105 respectivamente.

En todos los casos, los valores p indican que no se puede rechazar la hipótesis nula de que los coeficientes sean iguales a cero.

En conjunto, el análisis de los coeficientes sugiere que ninguna de las variables operativas consideradas presenta una relación estadísticamente significativa con el riesgo de fallo en las máquinas tipo Drill dentro del conjunto de datos CD1.

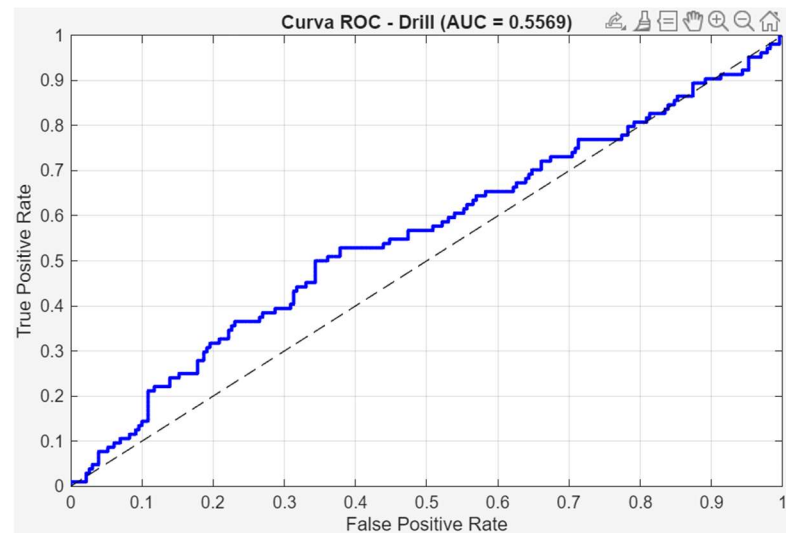


Figura 4-3 Curva ROC- $CD1_3$ (figura tomada de Matlab)

El área bajo la curva ROC obtenida fue $AUC = 0.557$ como se ve en la Figura 4-3, lo que indica que el modelo presenta una capacidad discriminante baja, muy cercana a la de un clasificador aleatorio. En términos probabilísticos, este resultado puede interpretarse como que existe aproximadamente un 55.7 % de probabilidad de que el modelo asigne una mayor probabilidad de fallo a una observación que realmente presenta fallo que a una observación que no lo presenta.

4.1.4 Tabla resumen $CD1$

A continuación, se presenta una tabla resumen que recoge los principales resultados obtenidos para los tres tipos de máquina considerados dentro del conjunto de datos $CD1$.

Tipo de máquina	AUC	Variables significativas
Mill	0.588	Temperatura
Lathe	0.553	Ninguna
Drill	0.557	Ninguna

Tabla 4-7 Resumen de resultados del $CD1$.

Como se observa en la Tabla 4-7, el modelo correspondiente a las máquinas tipo Mill presenta el valor más alto del área bajo la curva ROC, con un $AUC = 0.588$, lo que indica una capacidad discriminante ligeramente superior a la obtenida en los modelos ajustados para los otros tipos de máquina. Además, en este caso la variable x_1 (temperatura) resulta estadísticamente significativa, lo que sugiere que esta variable tiene una influencia relevante en la probabilidad de fallo de este tipo de máquina.

Por el contrario, en los modelos correspondientes a las máquinas tipo Lathe y Drill no se identifican variables estadísticamente significativas dentro del conjunto de variables consideradas. Asimismo, los valores de AUC obtenidos para estos modelos, 0.553 y 0.557 respectivamente, se sitúan muy próximos al valor 0.5, lo que indica que la capacidad del modelo para discriminar entre situaciones de fallo y funcionamiento normal es limitada.

En términos generales, estos resultados sugieren que las variables operativas analizadas presentan una mayor capacidad explicativa del riesgo de fallo en las máquinas tipo Mill, mientras que en los otros tipos de máquina el fenómeno de fallo podría estar influenciado por otros factores no incluidos en el modelo o por relaciones más complejas entre las variables.

En consecuencia, el análisis del conjunto de datos CD1 pone de manifiesto que el modelo de regresión logística presenta un comportamiento diferente según el tipo de máquina considerado, siendo más informativo en el caso de las máquinas tipo Mill que en los casos de Lathe y Drill.

4.2 Resultados del conjunto de datos CD2

El modelo de regresión logística ajustado para el conjunto de datos CD2 considera como variables explicativas el vector:

$$X = (x_1, x_2, x_3, x_4, x_5, x_6)$$

correspondiente a temperatura, humedad, vibración, edad de la máquina, presión y consumo de potencia.

A continuación, se muestran los resultados obtenidos tras el ajuste del modelo.

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_0	Intercepto	-40.1798	86.5268	-0.4644	0.6424
$\beta_1 (x_1)$	Temperatura	0.2072	0.7957	0.2604	0.7945
$\beta_2 (x_2)$	Humedad	0.1397	0.5423	0.2575	0.7968
$\beta_3 (x_3)$	Vibración	-0.5435	0.8166	-0.6656	0.5057
$\beta_4 (x_4)$	Potencia	-0.5953	1.2688	-0.4692	0.6389
$\beta_5 (x_5)$	Presión	0.3206	1.0552	0.3038	0.7613
$\beta_6 (x_6)$	Edad	0.0016	0.0097	0.1612	0.8720

Tabla 4-8 Resultados del CD2.

Métrica	Valor
Área bajo la curva ROC (AUC)	0.682
Desviación estándar del ajuste (sfit)	0.9939
Número de observaciones	3000
Número de variables explicativas	6

Tabla 4-9 Indicadores del modelo para CD2.

A partir de los resultados mostrados en la Tabla 4-8 y Tablas 4-9 se observa que ninguna de las variables presenta un valor p inferior al nivel de significancia, por lo que al igual que en los anteriores casos, no existe evidencia estadística suficiente para afirmar que alguna de las variables consideradas tenga una influencia significativa sobre la probabilidad de fallo de la máquina cuando se analizan de forma individual dentro del modelo.

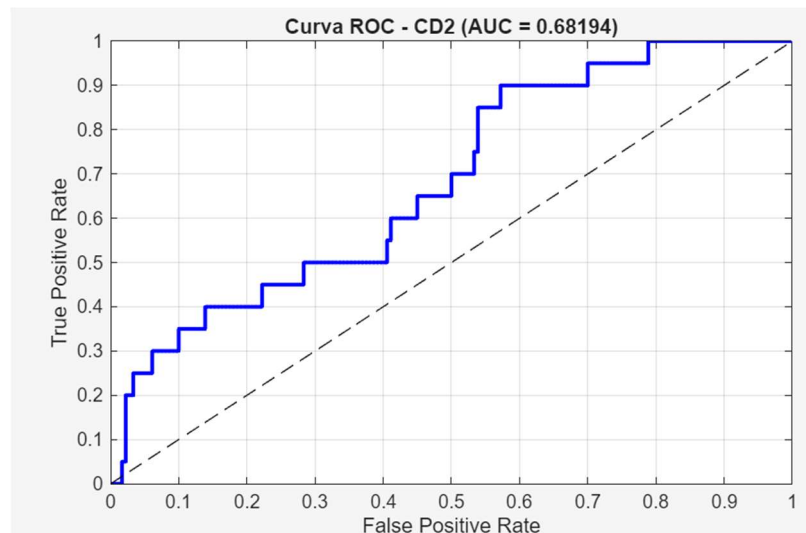


Figura 4-4 Curva ROC-CD2 (figura tomada de Matlab)

Como se observa en la Figura 4-9, el área bajo la curva obtenida es $AUC = 0.682$, lo que indica una capacidad discriminante moderada. Este valor es superior al obtenido en varios de los conjuntos de datos analizados previamente, lo que sugiere que el modelo presenta una mayor capacidad para distinguir entre observaciones con fallo y sin fallo en este conjunto de datos.

En términos probabilísticos, un valor de $AUC = 0.682$ implica que existe aproximadamente un 68.2 % de probabilidad de que el modelo asigne una mayor probabilidad de fallo a una observación que realmente presenta fallo que a una observación que no lo presenta. Por tanto, aunque el modelo no identifica variables individuales estadísticamente significativas, sí muestra una capacidad global razonable para diferenciar entre ambos estados del sistema.

4.3 Resultados del conjunto de datos de CDS1

El modelo de regresión logística ajustado para el conjunto de datos CDS1 considera como variables explicativas el vector:

$$X = (x_1, x_2, x_3, x_4, x_5, x_6)$$

Este conjunto de datos sintético ha sido diseñado para analizar la influencia de las distintas variables operativas de la máquina en la probabilidad de fallo y estudiar la posible existencia de correlaciones entre ellas antes de ajustar el modelo de regresión logística.

Con el objetivo de identificar posibles relaciones lineales entre las variables explicativas, se calculó la matriz de correlación de Pearson como se muestra en la Figura 4-5. Este análisis permite detectar situaciones de multicolinealidad, que pueden afectar a la estabilidad del modelo de regresión logística cuando varias variables contienen información redundante.

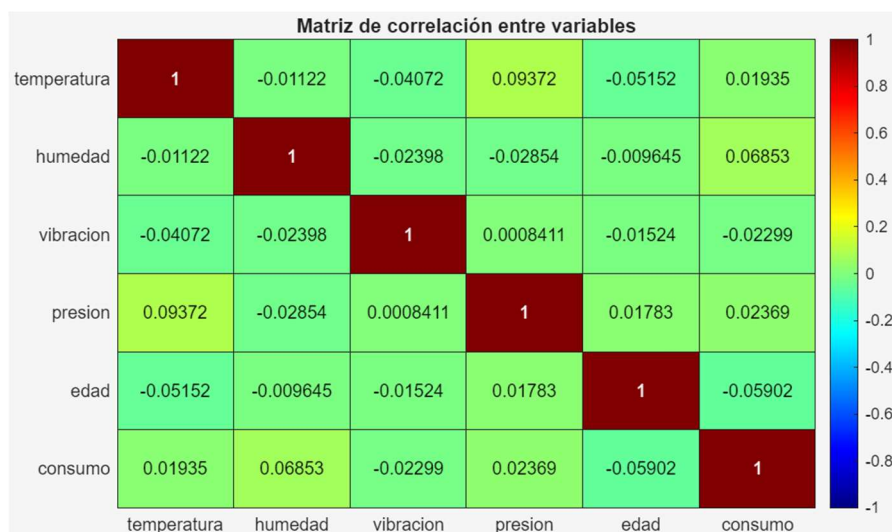


Figura 4-5 Matriz correlación CDS1 (figura tomada de Matlab)

Los resultados obtenidos muestran que los coeficientes de correlación entre las distintas variables son en general muy bajos, con valores cercanos a cero. En particular, no se observan valores de correlación superiores al umbral habitual de $|R| > 0.7$, lo que indica que no existe una correlación lineal fuerte entre las variables consideradas.

En la matriz de p -valores únicamente se observa una relación estadísticamente significativa entre las variables temperatura y presión, con un valor p inferior a 0.05. Sin embargo, el valor del coeficiente de correlación asociado es relativamente bajo ($R \approx 0.094$).

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_0	Intercepto	-7.2170	1.1703	-6.1670	0.0000
$\beta_1 (x_1)$	Temperatura	0.0225	0.0104	2.1735	0.0297
$\beta_2 (x_2)$	Humedad	-0.0024	0.0081	-0.2991	0.7649
$\beta_3 (x_3)$	Vibración	0.1907	0.0683	2.7913	0.0052
$\beta_4 (x_4)$	Potencia	0.0203	0.0034	6.0348	0.0000
$\beta_5 (x_5)$	Presión	0.1873	0.0564	3.3220	0.0009
$\beta_6 (x_6)$	Edad	0.1079	0.0262	4.1129	0.0000

Tabla 4-10 Resultados del CDS1.

Métrica	Valor
Área bajo la curva ROC (AUC)	0.688
Desviación estándar del ajuste (sfit)	1.0082
Número de observaciones	700
Número de variables explicativas	6

Tabla 4-11 Indicadores del modelo para CDS1.

Como puede observarse en la Tabla 4-10 y Tabla 4-11, varias de las variables consideradas presentan valores p inferiores al nivel de significancia habitual de 0.05. En particular, las variables temperatura, vibración, presión, edad de la máquina y consumo de potencia muestran una relación estadísticamente significativa con la probabilidad de fallo del sistema.

Por el contrario, la variable humedad presenta un valor p elevado, lo que indica que su contribución al modelo es limitada dentro de este conjunto de datos.

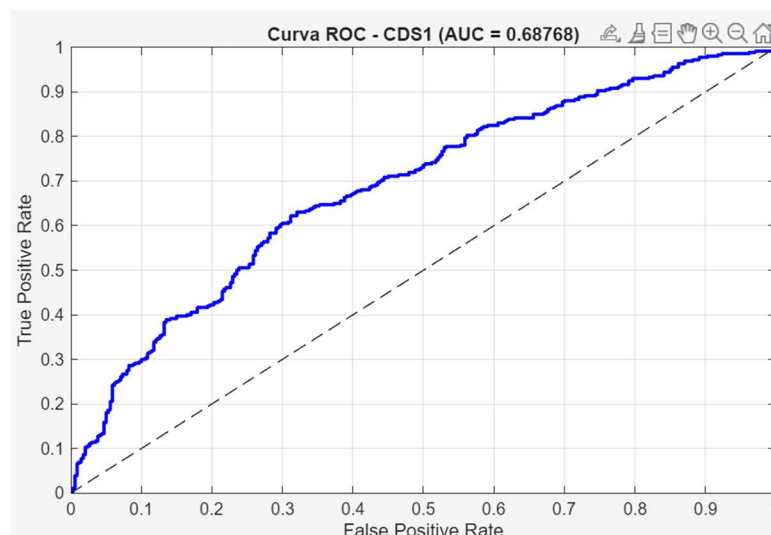


Figura 4-6 Curva ROC-CDS1 (figura tomada de Matlab)

La figura 4-6 presenta un valor $AUC = 0.688$ como se observa en la tabla 4-11, lo que indica una capacidad discriminante moderada para distinguir entre situaciones de fallo y funcionamiento normal. Este resultado sugiere que las variables consideradas contienen información relevante para explicar el comportamiento del sistema, aunque el modelo todavía presenta cierto margen de mejora en términos de precisión predictiva.

A partir del análisis del modelo logístico inicial se observa que la variable humedad presenta un valor p elevado ($p = 0.7649$), lo que indica que su contribución al modelo es limitada dentro de este conjunto de datos.

Por este motivo se procede a ajustar un segundo modelo eliminando dicha variable, con el objetivo de analizar si la reducción del número de variables permite mejorar o mantener el rendimiento del modelo.

Tras eliminar la variable humedad (x_2) del modelo inicial, el modelo reducido queda definido por el vector de variables explicativas $X = (x_1, x_3, x_4, x_5, x_6)$, correspondientes a temperatura, vibración, potencia, presión y edad de la máquina.

La Tabla 4-12 muestra cómo queda este nuevo conjunto de datos:

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_0	Intercepto	-7.3504	1.0830	-6.7870	0.0000
$\beta_1 (x_1)$	Temperatura	0.0225	0.0104	2.1747	0.0297
$\beta_3 (x_3)$	Vibración	0.1911	0.0683	2.7978	0.0051
$\beta_4 (x_4)$	Potencia	0.0203	0.0034	6.0310	0.0000
$\beta_5 (x_5)$	Presión	0.1877	0.0564	3.3305	0.0009

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_6 (x_6)	Edad	0.1079	0.0262	4.1133	0.0000

Tabla 4-12 Resultados del CDS1 reducido.

Modelo	Variables	AUC
Modelo completo	Temperatura, Humedad, Vibración, Potencia, Presión, Edad	0.688
Modelo reducido	Temperatura, Vibración, Potencia, Presión, Edad	0.687

Tabla 4-13 Comparativa de modelos CDS1.

Con el objetivo de simplificar el modelo y mejorar su interpretabilidad, se ajustó un segundo modelo de regresión logística eliminando la variable humedad, ya que en el modelo inicial presentaba un valor p elevado ($p = 0.7649$), lo que indicaba una contribución poco relevante a la predicción del fallo.

Los resultados del modelo reducido se muestran en la Tabla 4-12 y la comparación en la 4-13. Puede observarse que todas las variables restantes presentan valores p inferiores a 0.05, lo que indica que contribuyen de forma significativa a la explicación de la probabilidad de fallo del sistema.

Al comparar ambos modelos, se observa que el valor del área bajo la curva ROC pasa de $AUC = 0.688$ en el modelo completo a $AUC = 0.687$ en el modelo reducido. Esta diferencia es prácticamente despreciable, lo que indica que la eliminación de la variable humedad no afecta de forma significativa a la capacidad discriminante del modelo.

Por tanto, el modelo reducido presenta un comportamiento predictivo muy similar al modelo original, pero con un menor número de variables, lo que permite obtener un modelo más sencillo e interpretable sin pérdida apreciable de rendimiento.

4.4 Resultados del conjunto de datos de CDS2

El modelo de regresión logística ajustado para el conjunto de datos CDS2 considera como variables explicativas el vector:

$$X = (x_1, x_2, x_3, x_4, x_5, x_6)$$

En este conjunto de datos sintéticos se incluyen variables que no presentan una relación significativa con el riesgo de fallo del sistema. El objetivo de este análisis es estudiar cómo se comporta el modelo de regresión logística cuando algunas variables explicativas no aportan información relevante para la predicción del fallo. Para ello se analizan los coeficientes estimados y los valores p asociados a cada variable.

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_0	Intercepto	-6.4998	1.2833	-5.0648	0.0000
β_1 (x_1)	Temperatura	0.0191	0.0121	1.5738	0.1155
β_2 (x_2)	Humedad	0.0005	0.0070	0.0668	0.9467
β_3 (x_3)	Vibración	0.4198	0.0755	5.5581	0.0000
β_4 (x_4)	Potencia	0.0146	0.0039	3.7654	0.0002
β_5 (x_5)	Presión	0.0068	0.0633	0.1078	0.9142

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_6 (x_6)	Edad	0.0632	0.0280	2.2594	0.0239

Tabla 4-14 Resultados del conjunto de datos CDS2.

Métrica	Valor
Área bajo la curva ROC (AUC)	0.663
Desviación estándar del ajuste (sfit)	0.9995
Número de observaciones	700
Número de variables explicativas	6

Tabla 4-15 Indicadores del modelo para CDS2.

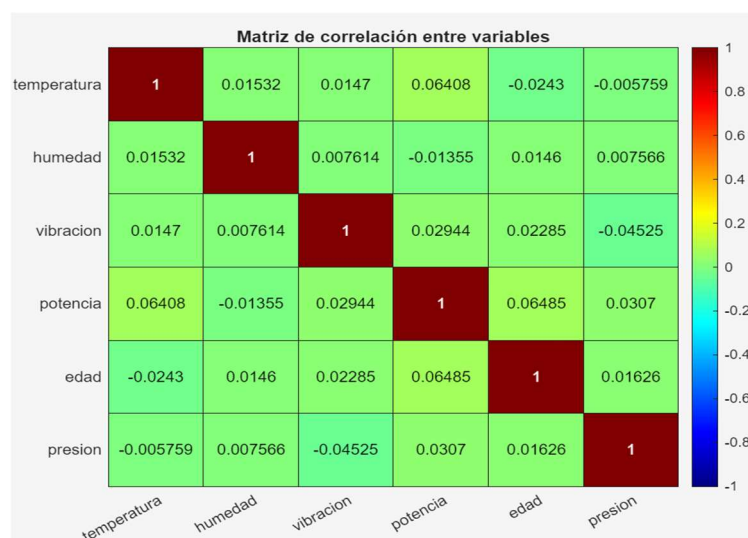


Figura 4-7 Matriz correlación-CDS2 (figura tomada de Matlab)

El análisis mostrado en la Tabla 4-14 y 4-15 y la Figura 4-7, vemos que la matriz de correlación muestra que los coeficientes de correlación entre las variables explicativas presentan valores muy reducidos, todos ellos claramente por debajo del umbral habitual de $|R| > 0.7$ utilizado para identificar relaciones lineales fuertes entre variables. Asimismo, en la matriz de p-valores no se observan relaciones estadísticamente significativas entre pares de variables.

Estos resultados indican que no existe un problema de multicolinealidad entre las variables del modelo, por lo que todas ellas pueden incluirse inicialmente en el ajuste de la regresión logística.

Posteriormente, a partir del análisis de los coeficientes del modelo logístico, se observa que las variables temperatura, humedad y presión presentan valores p elevados, lo que indica que su contribución a la predicción del fallo es limitada. Por este motivo se decide eliminarlas del modelo, obteniendo un modelo reducido formado por las variables vibración, potencia y edad.

En consecuencia, se va a construir un modelo reducido eliminando variables no significativas.

El modelo reducido queda definido por:

$$X = (x_3, x_4, x_6)$$

Quedando la tabla de la siguiente manera:

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_0	Intercepto	-5.0324	0.7117	-7.0715	0.0000
β_3 (x_3)	Vibración	0.4201	0.0754	5.5738	0.0000
β_4 (x_4)	Potencia	0.0145	0.0039	3.7552	0.0002
β_6 (x_6)	Edad	0.0626	0.0279	2.2433	0.0249

Tabla 4-16 Resultados del CDS2 reducido.

Métrica	Valor
Área bajo la curva ROC (AUC)	0.658
Desviación estándar del ajuste (sfit)	0.9983
Número de observaciones	700
Número de variables explicativas	3

Tabla 4-17 Indicadores del modelo para CDS2 reducido.

A partir del análisis del modelo inicial se observó que varias variables presentaban valores p superiores al nivel de significancia habitual de 0.05. En particular, las variables temperatura, humedad y presión no mostraban una contribución estadísticamente significativa en la predicción del fallo del sistema.

Por este motivo se ajustó un modelo reducido eliminando dichas variables y conservando únicamente vibración, potencia y edad, que sí presentaban valores p inferiores a 0.05.

Los resultados del modelo reducido se muestran en la Tabla 4-16 y Tabla 4-17. Puede observarse que todas las variables restantes resultan estadísticamente significativas, lo que indica que contribuyen de forma relevante a explicar la probabilidad de fallo.

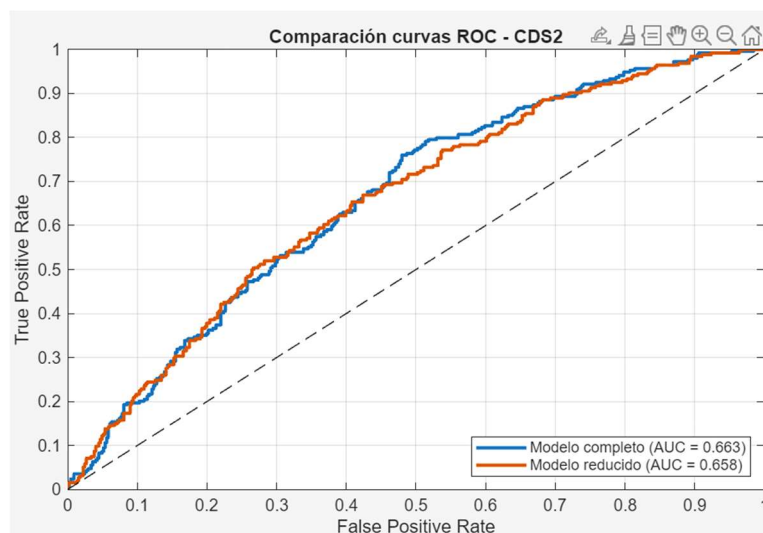


Figura 4-8 Curva comparación modelos-CDS2 (figura tomada de Matlab)

Al comparar ambos modelos como se ve en la Figura 4-8, se observa que el valor del área bajo la curva ROC pasa de $AUC = 0.663$ en el modelo completo a $AUC = 0.658$ en el modelo reducido. Esta diferencia es pequeña, lo que indica que la eliminación de las variables no significativas permite obtener un modelo más simple y fácilmente interpretable sin una pérdida apreciable de capacidad predictiva.

4.5 Resultados del conjunto de datos de CDS3

El modelo de regresión logística ajustado para el conjunto de datos CDS3 considera como variables explicativas el vector:

$$X = (x_1, x_2, x_3, x_4, x_5, x_6, x_7)$$

Para este conjunto se añade x_7 que es la carga de trabajo de la máquina, variable adicional incorporada en este conjunto de datos.

Este conjunto de datos ha sido generado de forma sintética siguiendo una relación logística conocida entre las variables explicativas y el riesgo de fallo. Esto permite analizar el comportamiento del modelo de regresión logística en un escenario controlado en el que la relación entre variables y probabilidad de fallo está claramente definida.

A continuación, se muestran los datos obtenidos al ser analizado a través de Matlab:

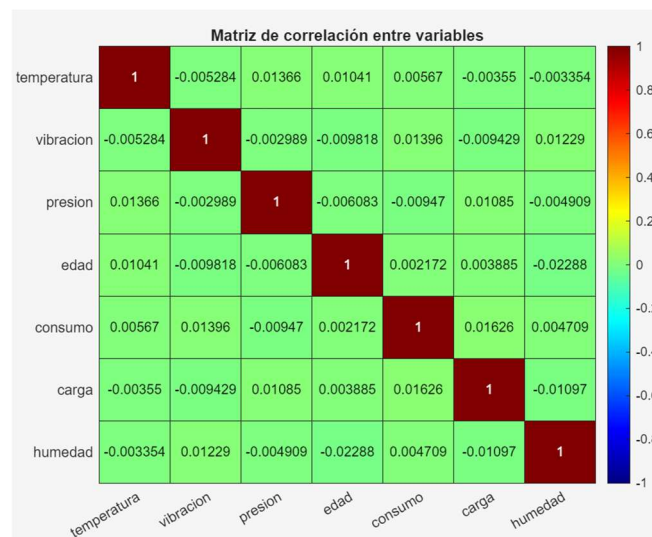


Figura 4-9 Matriz correlación-CDS3 (figura tomada de Matlab)

La matriz de correlación de la Figura 4-9 muestra que no existen correlaciones lineales fuertes entre las variables, ya que todos los coeficientes de correlación presentan valores muy próximos a cero.

Esto indica que las variables incluidas en el modelo aportan información diferente y no existe multicolinealidad significativa entre ellas.

En la matriz de p -valores únicamente se observa una relación estadísticamente significativa entre la edad de la máquina (x_4) y humedad (x_7) con un valor $p = 0.012$, inferior al nivel de significación habitual de 0.05.

Sin embargo, el coeficiente de correlación asociado es muy bajo ($|R| = 0.0229$), por lo que la relación es estadísticamente significativa pero muy débil desde el punto de vista práctico. Por este motivo se decidió mantener ambas variables en el modelo.

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
β_0	Intercepto	-18.0041	0.4025	-44.7274	< 0.001
$\beta_1 (x_1)$	Temperatura	0.0774	0.0028	27.9444	< 0.001
$\beta_2 (x_2)$	Humedad	-0.0131	0.0023	-5.6938	1.24×10^{-8}
$\beta_3 (x_3)$	Vibración	0.6357	0.0187	33.9787	< 0.001

Variable	Descripción	Coefficiente β	Error estándar	Estadístico t	p-valor
$\beta_4 (x_4)$	Potencia	0.0275	0.0011	24.6186	< 0.001
$\beta_5 (x_5)$	Presión	0.4251	0.0215	19.7978	< 0.001
$\beta_6 (x_6)$	Edad	0.1392	0.0068	20.5392	< 0.001
$\beta_7 (x_7)$	Carga	0.0312	0.0016	19.5428	< 0.001

Tabla 4-18 Resultados del CDS3.

Métrica	Valor
Área bajo la curva ROC (AUC)	0.819
Desviación estándar del ajuste (sfit)	1.0085
Número de observaciones	12000
Número de variables explicativas	7

Tabla 4-19 Indicadores del modelo para CDS3.

Los resultados del ajuste del modelo de regresión logística se presentan en la Tabla 4-18 y la Tabla 4-19. En este caso se observa que todas las variables incluidas presentan valores p extremadamente pequeños, lo que indica que contribuyen de forma significativa a explicar la probabilidad de fallo en el sistema.

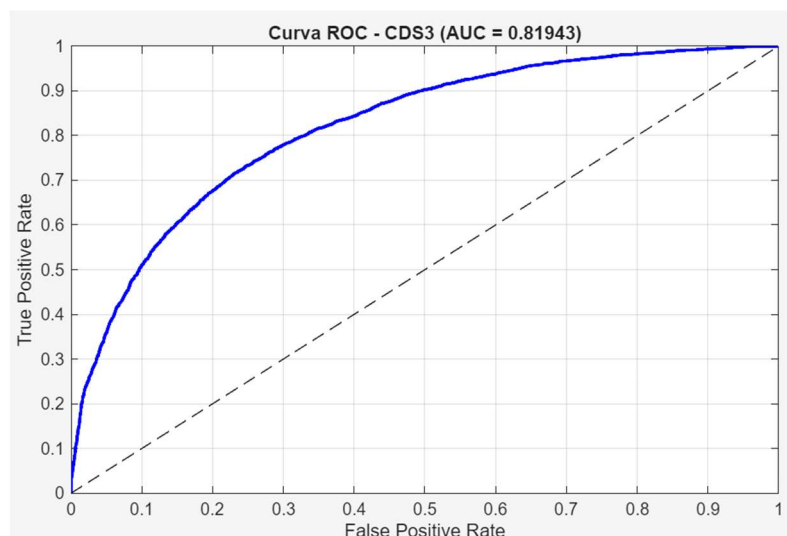


Figura 4-10 Curva ROC-CDS3 (figura tomada de Matlab)

Por otra parte, como se observa en la Figura 4-10 el modelo alcanza un área bajo la curva ROC de $AUC = 0.819$, lo que indica una capacidad de discriminación elevada. Este resultado implica que el modelo es capaz de distinguir correctamente entre situaciones de fallo y funcionamiento normal en aproximadamente el 81.9 % de los casos, mostrando un comportamiento claramente superior al observado en varios de los conjuntos de datos analizados anteriormente.

En conclusión, el conjunto de datos CDS3 presenta un comportamiento notablemente mejor que los anteriores ya que no existe multicolinealidad entre variables, la mayoría de las variables son

estadísticamente significativas y el modelo alcanza una AUC superior a 0.8, lo que indica una buena capacidad predictiva.

Esto sugiere que el conjunto de datos ha sido generado de forma que las variables operativas influyen de manera clara en la probabilidad de fallo, permitiendo a la regresión logística capturar correctamente la relación entre las variables y el evento de fallo.

Adicionalmente, se ha realizado un análisis complementario sobre el conjunto de datos CDS3, que presenta un gran rendimiento y que, además, es el que dispone de un mayor número de observaciones (12000). Con el objetivo de evaluar la influencia del tamaño muestral en el comportamiento del modelo, se ha repetido el ajuste de la regresión logística empleando únicamente un subconjunto reducido de 1000 observaciones seleccionado de dicho *dataset*.

Los resultados obtenidos muestran que los valores de AUC y las métricas de rendimiento del modelo apenas experimentan variaciones significativas respecto al caso original. Esto indica que, en este problema, el incremento en el número de observaciones no supone una mejora sustancial del modelo cuando las variables ya contienen información relevante.

Por tanto, puede concluirse que la calidad y relevancia de las variables explicativas tienen un impacto mucho mayor en el rendimiento del modelo que el tamaño del conjunto de datos, reforzando así la idea de que un conjunto de datos bien definido es más determinante que un gran volumen de datos.

4.6 Resultados del conjunto de datos de CDS4

El modelo de regresión logística ajustado para el conjunto de datos CDS4 considera como variables explicativas el vector

$$X = (x_1, x_3, x_4, x_5, x_6)$$

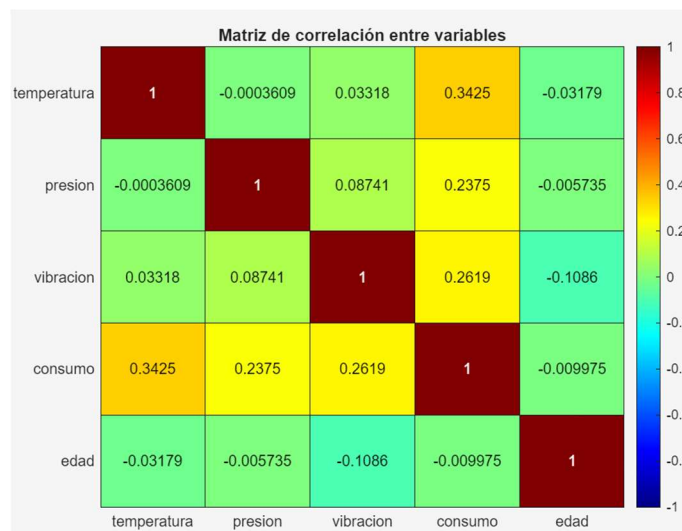


Figura 4-11 Matriz correlación-CDS4 (figura tomada de Matlab)

Antes de ajustar el modelo de regresión logística se analizó la matriz de correlación entre las variables explicativas como se puede ver en la Figura 4-11. Los coeficientes de correlación obtenidos presentan valores relativamente bajos, siendo el mayor de ellos el correspondiente a la relación entre temperatura y consumo energético ($R = 0.342$). Este valor se encuentra claramente por debajo del umbral habitual de 0.7 utilizado para detectar multicolinealidad, por lo que puede considerarse que no existe una correlación lineal fuerte entre las variables del modelo.

La matriz de p -valores muestra algunas relaciones estadísticamente significativas, especialmente entre el consumo energético y varias de las variables operativas, así como entre la vibración y la edad

de la máquina. No obstante, los coeficientes de correlación asociados a estas relaciones presentan magnitudes moderadas o bajas, por lo que se decidió mantener todas las variables en el modelo.

Variable	Descripción	Coeficiente β	Error estándar	Estadístico t	p-valor
β_0	Intercepto	-12.1560	1.7872	-6.8017	< 0.001
$\beta_1 (x_1)$	Temperatura	0.0373	0.0167	2.2356	0.025
$\beta_3 (x_3)$	Vibración	0.3222	0.1562	2.0627	0.039
$\beta_4 (x_4)$	Potencia	0.0225	0.0073	3.1008	0.0019
$\beta_5 (x_5)$	Presión	0.1908	0.0967	1.9734	0.048
$\beta_6 (x_6)$	Edad	0.0639	0.0300	2.1285	0.033

Tabla 4-20 Resultados del CDS4.

Métrica	Valor
Área bajo la curva ROC (AUC)	0.712
Desviación estándar del ajuste (sfit)	1.0445
Número de observaciones	500
Número de variables explicativas	5

Tabla 4-21 Indicadores del modelo para CDS4.

Los resultados del ajuste del modelo se presentan en la Tabla 4-20 y la Tabla 4-21. En este caso todas las variables explicativas presentan valores p inferiores a 0.05, lo que indica que contribuyen de forma significativa a explicar la probabilidad de fallo en el sistema.

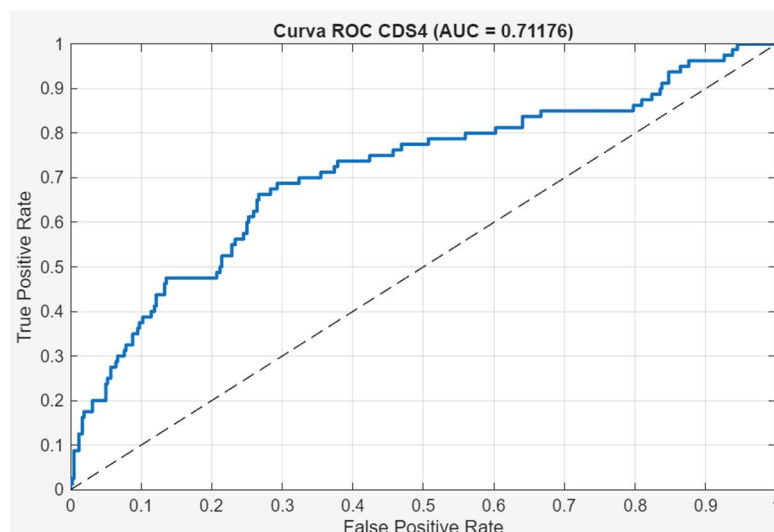


Figura 4-12 Curva ROC-CDS4 (figura tomada de Matlab)

Como se muestra en la Figura 4-12, el modelo alcanza un área bajo la curva ROC de $AUC = 0.712$, lo que indica una capacidad de discriminación moderada. Este resultado sugiere que el modelo es capaz de distinguir entre situaciones de fallo y funcionamiento normal en aproximadamente el 71.2 % de los casos, mostrando un rendimiento superior al de algunos conjuntos de datos analizados anteriormente.

El modelo alcanza un área bajo la curva ROC de $AUC = 0.712$, lo que indica una capacidad de discriminación moderada. Este resultado sugiere que el modelo es capaz de distinguir entre situaciones de fallo y funcionamiento normal en aproximadamente el 71.2 % de los casos, mostrando un rendimiento superior al de algunos conjuntos de datos analizados anteriormente, aunque inferior al observado en el conjunto CDS3.

4.7 Comparación de resultados

Con el objetivo de analizar el comportamiento de los distintos modelos de regresión logística obtenidos a lo largo del trabajo, se realizó una comparación global de los resultados obtenidos para cada uno de los conjuntos de datos analizados. En el caso del conjunto CD1, el análisis se realizó de forma independiente para cada tipo de máquina (Mill, Lathe y Drill), mientras que en los conjuntos sintéticos posteriores se ajustaron modelos únicos sobre cada *dataset*. En la Tabla 4-22 se resumen las principales métricas de rendimiento de todos los modelos obtenidos.

Conjunto datos	Tipo	Nº variables	Variables utilizadas	AUC
CD1	Mill	4	Temperatura, Vibración, Potencia, Humedad	0.588
CD1	Lathe	4	Temperatura, Vibración, Potencia, Humedad	0.553
CD1	Drill	4	Temperatura, Vibración, Potencia, Humedad	0.557
CD2	Completo	6	Temperatura, Humedad, Vibración, Potencia, Presión, Edad	0.682
CDS1	Completo	6	Temperatura, Humedad, Vibración, Potencia, Presión, Edad	0.688
CDS1	Reducido	5	Temperatura, Vibración, Potencia, Presión, Edad	0.687
CDS2	Completo	6	Temperatura, Humedad, Vibración, Potencia, Presión, Edad	0.663
CDS2	Reducido	3	Vibración, Presión, Edad	0.658
CDS3	Completo	7	Temperatura, Humedad, Vibración, Potencia, Presión, Edad, Carga	0.819
CDS4	Completo	5	Temperatura, Vibración, Potencia, Presión, Edad	0.712

Tabla 4-22 Comparación de la capacidad predictiva de los modelos.

A partir de los resultados mostrados en la Tabla 4-22 puede observarse que los modelos obtenidos presentan comportamientos diferentes en función del conjunto de datos utilizado. Los modelos correspondientes al conjunto CD1, ajustados de forma independiente para cada tipo de máquina, muestran en general los valores más bajos de AUC, lo que indica una capacidad limitada para discriminar entre situaciones de fallo y funcionamiento normal.

El conjunto CD2, correspondiente al *dataset* original completo, presenta un rendimiento intermedio, con un valor de AUC de 0.682, situándose por encima de los modelos de CD1 y próximo a los obtenidos en algunos conjuntos sintéticos iniciales.

Por su parte, los conjuntos sintéticos CDS1, CDS2, CDS3 y CDS4 muestran en general valores de AUC superiores, destacando especialmente CDS3, que alcanza el mejor resultado global. Asimismo, en algunos casos la eliminación de variables permitió simplificar el modelo sin producir variaciones relevantes en su capacidad predictiva.

Con el objetivo de comparar de forma global el rendimiento de todos los modelos ajustados, se elaboró un ranking basado en el valor del área bajo la curva ROC (AUC).

Posición	Conjunto de datos	Tipo de modelo	AUC
1	CDS3	Completo	0.819
2	CDS4	Completo	0.712
3	CDS1	Completo	0.688
4	CDS1	Reducido	0.687
5	CD2	Completo	0.682
6	CDS2	Completo	0.663
7	CDS2	Reducido	0.658
8	CD1	Mill	0.588
9	CD1	Drill	0.557
10	CD1	Lathe	0.553

Tabla 4-23. Ranking de modelos.

Con el objetivo de comparar de forma global el rendimiento de todos los modelos ajustados, se elaboró un ranking basado en el valor del área bajo la curva ROC (AUC) obtenido en cada caso.

Tal como se observa en la Tabla 4-23, el modelo correspondiente al conjunto de datos CDS3 presenta el mejor rendimiento, seguido por CDS4 y los modelos asociados a CDS1 y CD2, que muestran valores de AUC relativamente próximos entre sí.

En las últimas posiciones del ranking se sitúan los modelos obtenidos a partir del conjunto CD1, correspondientes a los distintos tipos de máquina, los cuales presentan los valores más reducidos de AUC dentro del conjunto de modelos analizados.

Además del análisis global del modelo, resulta interesante interpretar su utilidad práctica a partir de la función logística estimada. En el caso del conjunto de datos CDS3, la probabilidad de fallo puede expresarse mediante la función:

$$p_i = \frac{1}{1 + e^{-z}}$$

donde

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

y, utilizando los coeficientes obtenidos para este conjunto de datos, resulta:

$$z = -18.0041 + 0.0774x_1 - 0.0131x_2 + 0.6357x_3 + 0.0275x_4 + 0.4251x_5 + 0.1392x_6 + 0.0312x_7$$

Siendo

$$X = (x_1, x_2, x_3, x_4, x_5, x_6, x_7) \\ = (\text{Temperatura, Humedad, Vibración, Potencia, Presión, Edad, Carga})$$

De este modo, el modelo permite estimar la probabilidad de fallo para cualquier nueva observación que no pertenezca al conjunto de datos original.

Por ejemplo, para un nuevo estado de funcionamiento de la máquina dado por:

$$x = (75, 50, 5, 110, 6, 10, 70)$$

se obtiene:

$$z \approx 1.06$$

y, por tanto,

$$p_i = \frac{1}{1 + e^{-1.06}} \approx 0.743$$

Esto significa que, para ese conjunto de condiciones operativas, el modelo estima una probabilidad de fallo del 74.3 %, lo que puede interpretarse como una situación de riesgo moderado-alto. En términos prácticos, este resultado muestra que el modelo puede utilizarse como herramienta de apoyo para que un operario estime el riesgo de fallo de la máquina a partir de los valores observados en un instante determinado, permitiendo anticipar posibles averías y planificar acciones de mantenimiento preventivo.

5 CONCLUSIONES Y LÍNEAS FUTURAS

En este capítulo se recogen las principales conclusiones derivadas del trabajo realizado, evaluando el grado de cumplimiento de los objetivos planteados al inicio de este. Asimismo, se analizan las limitaciones encontradas y se proponen posibles líneas futuras de investigación orientadas a mejorar y ampliar los resultados obtenidos.

5.1 Conclusiones

El objetivo principal de este trabajo ha sido analizar la capacidad de la regresión logística para predecir fallos en maquinaria industrial a partir de distintas variables operativas, así como estudiar la influencia de la selección de variables y de la estructura de los datos en el rendimiento del modelo.

A la vista de los resultados obtenidos, puede afirmarse que los objetivos planteados se han alcanzado de forma satisfactoria. Se ha comprobado que la regresión logística es capaz de modelizar adecuadamente la probabilidad de fallo cuando las variables explicativas contienen información relevante, alcanzando valores elevados de AUC en los conjuntos mejor estructurados (como CDS3 y CDS4). En estos casos, el modelo presenta una clara capacidad de discriminación, lo que valida su utilidad en problemas de mantenimiento predictivo.

Sin embargo, el trabajo también ha puesto de manifiesto una limitación clave: la regresión logística no es robusta frente a conjuntos de datos mal definidos o con variables poco informativas. En *datasets* como CDS2, donde las variables presentan baja relación con el fallo, el modelo ofrece un rendimiento notablemente inferior. Esto demuestra que el éxito del modelo no depende tanto del algoritmo en sí, sino de la calidad del diseño del conjunto de datos.

Un aspecto especialmente relevante ha sido la incorporación del análisis de correlación mediante la matriz R y la matriz de p -valores P , implementadas en MATLAB. Este paso ha permitido identificar relaciones lineales entre variables y justificar la eliminación de variables redundantes. Se ha observado que la eliminación de variables irrelevantes o redundantes no solo simplifica el modelo, sino que en muchos casos mantiene (o incluso mejora ligeramente) su capacidad predictiva, lo que refuerza la importancia del preprocesamiento de datos.

Asimismo, el análisis de los coeficientes, junto con sus errores estándar, estadísticos t y p -valores, ha permitido interpretar la relevancia de cada variable. En los modelos más robustos, varias variables presentan significación estadística clara (p -valores muy bajos), lo que indica que contribuyen de forma consistente a la predicción del fallo. En cambio, en otros casos se observan coeficientes no significativos, evidenciando la presencia de ruido o falta de relación real con la variable respuesta.

En el contexto del ámbito naval, estos resultados adquieren una relevancia especial, ya que la capacidad de anticipar fallos en sistemas de maquinaria a bordo puede contribuir significativamente a mejorar la seguridad, la eficiencia operativa y la planificación del mantenimiento naval. La aplicación de modelos como el desarrollado en este trabajo permitiría, a partir de variables monitorizadas en tiempo real, estimar el riesgo de fallo de determinados equipos críticos, facilitando la toma de decisiones preventivas y reduciendo tanto los costes asociados a averías como el riesgo de incidentes durante la navegación.

En conjunto, este trabajo demuestra que la regresión logística es una herramienta eficaz, interpretable y adecuada para problemas de clasificación binaria tanto en el ámbito industrial como en el naval, pero su rendimiento está fuertemente condicionado por la calidad del *dataset*, la selección adecuada de variables, y la ausencia de redundancias o correlaciones innecesarias.

5.2 Líneas futuras

Como posibles líneas de mejora y ampliación del trabajo, se proponen las siguientes:

- Aplicación a datos reales industriales, lo que permitiría validar los resultados en un entorno práctico y evaluar la capacidad real del modelo.
- Implementación de métodos automáticos de selección de variables, como LASSO o selección hacia adelante/atrás, para optimizar el modelo sin intervención manual.
- Comparación con modelos más complejos, como *random forest*, máquinas de soporte vectorial o redes neuronales, para analizar posibles mejoras en la capacidad predictiva.
- Estudio del desbalanceo de clases, aplicando técnicas como SMOTE o ponderación de clases en caso de *datasets* con pocos fallos.
- Incorporación de información temporal, considerando la evolución de las variables en el tiempo para construir modelos más realistas de mantenimiento predictivo.

6 BIBLIOGRAFÍA

- [1] R. K. Mobley, *An Introduction to Predictive Maintenance*, Oxford: Elsevier, 2002.
- [2] A. K. S. Jardine, D. Lin y D. Banjevic, «A review on machinery diagnostics and prognostics implementing condition-based maintenance,» *Mechanical Systems and Signal Processing*, p. 1483–1510, 1 julio 2006.
- [3] D. W. Hosmer, S. Lemeshow y R. X. Sturdivant, *Applied Logistic Regression*, Hoboken: Willey, 2013.
- [4] G. James, D. Witten, T. Hastie y R. Tibshirani, *An Introduction to Statistical Learning*, New York: Springer, 2013.
- [5] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York: Springer, 2006.
- [6] A. Agresti, *An Introduction to Categorical Data Analysis*, Hoboken: Wiley, 2018.
- [7] H. Pishro-Nik, *Introduction to Probability, Statistics, and Random Processes*, Amherst: University of Massachusetts Amherst, 2014.
- [8] L. Y. B. U. I. SILVA AYCAGUER, *Regresión logística*, Madrid: La Muralla , 2004.
- [9] J. E. J. & S. R. B. Dennis, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Philadelphia: Classics in Applied Mathematics, Vol. 16. , 1996.
- [10] S. Hussain, «Machine Failure Prediction Dataset,» [En línea]. Available: https://www.kaggle.com/datasets/saquib7hussain/machine-failure-prediction-dataset?utm_source=chatgpt.com&select=machine_failure_data.csv. [Último acceso: 2 Febrero 2026].
- [11] A. Jovell, *Análisis de regresión logística*, Madrid: Centro de Investigaciones Sociológicas, 1995.

ANEXO I: IMPLICACIONES SOCIALES, Y/O ECONÓMICAS, Y/O AMBIENTALES

El presente trabajo, centrado en el análisis de fallos en maquinaria industrial mediante técnicas de regresión logística, presenta implicaciones relevantes en los ámbitos social, económico y ambiental.

Desde el punto de vista social, el desarrollo de modelos de predicción de fallos contribuye a mejorar la seguridad en entornos industriales. La detección anticipada de posibles averías permite reducir el riesgo de accidentes laborales, protegiendo tanto a los operarios como a las instalaciones. Asimismo, favorece la implementación de estrategias de mantenimiento más seguras y planificadas.

En el ámbito económico, este tipo de modelos tiene un impacto directo en la eficiencia operativa de las empresas. La capacidad de anticipar fallos permite reducir costes asociados a paradas no planificadas, minimizar pérdidas de producción y optimizar las tareas de mantenimiento. Además, facilita una mejor gestión de recursos, incrementando la competitividad industrial.

Desde la perspectiva ambiental, la mejora en el mantenimiento de la maquinaria contribuye a un uso más eficiente de los recursos y a la reducción de residuos. Evitar fallos graves implica alargar la vida útil de los equipos, disminuir el consumo innecesario de energía y reducir la generación de desechos derivados de averías o sustituciones prematuras.

En conjunto, el trabajo desarrollado no solo tiene un interés técnico, sino que también aporta beneficios relevantes en términos de seguridad, eficiencia y sostenibilidad, alineándose con los principios de ingeniería responsable.

ANEXO II: REFLEXIONES ÉTICAS Y SOCIALES

El desarrollo del presente trabajo, basado en la predicción de fallos en maquinaria industrial mediante técnicas de regresión logística, plantea diversas consideraciones en el ámbito ético y social que conviene tener en cuenta.

En primer lugar, el uso de modelos predictivos en entornos industriales implica una responsabilidad en la toma de decisiones, ya que las predicciones obtenidas pueden influir en acciones relacionadas con el mantenimiento, la seguridad o la operativa de la maquinaria. En este sentido, es fundamental interpretar los resultados del modelo de forma crítica, evitando una dependencia ciega de las predicciones y considerando siempre el contexto técnico y operativo.

Por otro lado, es importante destacar la transparencia y la interpretabilidad del modelo utilizado. La regresión logística presenta la ventaja de ser un modelo fácilmente interpretable, lo que permite comprender cómo influyen las variables en la probabilidad de fallo. Esta característica resulta especialmente relevante desde el punto de vista ético, ya que facilita la trazabilidad de las decisiones y evita el uso de modelos opacos.

Asimismo, el desarrollo de este tipo de herramientas contribuye a mejorar las condiciones de trabajo, al permitir anticipar fallos y reducir riesgos, lo que tiene un impacto positivo en la seguridad de los trabajadores. No obstante, también plantea la necesidad de una correcta implementación, evitando que estas herramientas sustituyan completamente el criterio humano o se utilicen sin una validación adecuada.

Finalmente, este trabajo pone de manifiesto la importancia de utilizar los datos de forma responsable, garantizando su calidad y evitando interpretaciones erróneas que puedan derivar en decisiones inadecuadas.

En conjunto, la aplicación de modelos predictivos en el ámbito industrial debe abordarse desde una perspectiva ética, combinando el uso de herramientas analíticas con el juicio experto y la responsabilidad profesional.

ANEXO III: CÓDIGO MATLAB

En este anexo se presenta el código base utilizado en MATLAB para el análisis de los distintos conjuntos de datos considerados en el trabajo. Este código ha servido como estructura común para todos los casos analizados, incluyendo la carga de datos, el ajuste del modelo de regresión logística, el cálculo de probabilidades, la obtención de la curva ROC y el análisis estadístico de los coeficientes.

Las diferencias entre los distintos conjuntos de datos (CD1, CD2, CDS1, CDS2, CDS3 y CDS4) se han limitado principalmente a la selección de variables explicativas, en función del número de variables disponibles y de los resultados obtenidos en el análisis de correlación. No obstante, la estructura general del código se ha mantenido constante en todos los casos.

```
%% ANÁLISIS GENERAL - REGRESIÓN LOGÍSTICA
clc; clear; close all;

%% 1. Cargar datos
filename = 'dataset.csv'; % Cambiar según el conjunto de datos
data = readtable(filename);

%% 2. Selección de variables
% Eliminar columnas no necesarias (ejemplo)
% data(:, [1]) = []; % eliminar id si existe
% Separar variables explicativas y variable respuesta
Xtable = data(:, 1:end-1);
y = data{:, end};
X = table2array(Xtable);
varNames = Xtable.Properties.VariableNames;
disp('Primeras 5 filas de X:')
disp(X(1:5, :))

%% 3. Matriz de correlación
[R, P] = corrcoef(X);
corr_table = array2table(R, 'VariableNames', varNames, 'RowNames', varNames);
disp('Tabla de correlación:')
disp(corr_table)
disp('Matriz de correlación R:')
disp(R)
disp('Matriz de p-valores P:')
format long
disp(P)
format short

%% 4. Identificación de correlaciones
```

```
thresholdR = 0.7;
[row,col] = find(abs(R) > thresholdR & abs(R) < 1);
disp('Pares de variables con correlación alta:')
disp([row col])
thresholdP = 0.05;
[row1,col1] = find(P < thresholdP & P > 0);
disp('Pares de variables con p-valores < 0.05:')
disp([row1 col1])

%% 5. Heatmap
figure
h = heatmap(varNames, varNames, R);
h.Title = 'Matriz de correlación';
h.Colormap = jet;
h.ColorLimits = [-1 1];

%% 6. Regresión logística
[b,dev,stats] = glmfit(X,y,'binomial','link','logit');

%% 7. Probabilidades estimadas
p_fallo = glmval(b,X,'logit');
disp('Primeras 10 probabilidades estimadas:')
disp(p_fallo(1:10))

%% 8. Curva ROC
[Xf,Yf,T,AUC] = perfcurve(y,p_fallo,1);
fprintf('AUC del modelo: %.3f\n',AUC);
figure
plot(Xf,Yf,'LineWidth',2)
hold on
plot([0 1],[0 1],'k--')
xlabel('False Positive Rate')
ylabel('True Positive Rate')
title(['Curva ROC (AUC = ', num2str(AUC), ')'])
grid on

%% 9. Estadísticos del modelo
disp('Coeficientes del modelo:')
disp(b)
disp('--- Estadísticos ---')
disp('Coeficientes beta:')
```

```
disp(stats.beta)
disp('Errores estándar:')
disp(stats.se)
disp('Estadístico t:')
disp(stats.t)
disp('p-valores:')
for i = 1:length(stats.p)
    fprintf('%.3e\n', stats.p(i));
end
disp('Desviación estándar del ajuste:')
disp(stats.sfit)
```

ANEXO IV: USO DE LA IA

En el desarrollo del presente Trabajo Fin de Grado se ha hecho uso puntual de herramientas de inteligencia artificial como apoyo en distintas fases del trabajo.

En concreto, estas herramientas se han utilizado para mejorar la redacción de algunos apartados, revisar la claridad de las explicaciones y proponer alternativas en la formulación de textos, con el objetivo de lograr una mayor coherencia en el documento final.

Asimismo, se han empleado como apoyo para la resolución de dudas técnicas relacionadas con el uso del software MATLAB, especialmente en aspectos concretos de implementación y depuración de código, así como para la consulta puntual de información general.

El uso de estas herramientas se ha limitado a funciones de asistencia y apoyo durante el proceso de elaboración del trabajo.