

Análisis de defectos en sistemas industriales, combinando la visualización y detección de patrones, con el procesamiento de lenguaje natural

Ismael Espinoza Arias¹, Samuel González-López¹,
Jesús Raúl Cruz-Rentería¹, Jesús Miguel García-Gorrostieta²,
Aurelio López-López³

¹ Tecnológico Nacional de México/Tecnológico de Nogales,
México

² Universidad de la Sierra,
México

³ Instituto Nacional de Astrofísica, Óptica y Electrónica,
Puebla, México

{M24340931,samuel.gl,jesus.cr}@nogales.tecnm.mx,
jgarcia@unisierra.edu.mx, allopez@inaoep.mx

Resumen. Este trabajo se centra en la supervisión y el análisis de las pruebas eléctricas realizadas en transformadores, inductores y productos relacionados. Los componentes eléctricos como transformadores e inductores desempeñan un papel fundamental en este campo debido a sus funcionalidades, que van desde un cargador doméstico convencional hasta dispositivos más sofisticados que regulan la corriente suministrada a toda una ciudad. Nuestro trabajo busca identificar relaciones significativas entre diversos parámetros operativos. Estas correlaciones se utilizan para detectar tendencias, permitiendo una comprensión más profunda de las interacciones entre diferentes variables. Se presenta una metodología de tres etapas. La primera busca identificar las coocurrencias de las fallas en diferentes tipos de errores, utilizando las técnicas Apriori y FPGrowth. En la segunda etapa se desarrolla un análisis cualitativo, a través de entrevistas abiertas. La última etapa es el entrenamiento de un modelo computacional para mejorar una de las fuentes identificadas generadora de ciertos fallos. Se logró obtener un modelo predictivo para proveer posibles soluciones a problemas identificadas en máquinas de embobinado, alcanzando un 84.11 % de F-measure.

Palabras clave: Análisis predictivo, toma de decisiones, procesamiento de lenguaje natural, aprendizaje automático y sistemas industriales.

Approach for the Analysis of Defects in Industrial Systems by Combining Pattern Visualization and Detection with Natural Language Processing

Abstract. This work focuses on the monitoring and analysis of electrical tests performed on transformers, inductors, and related products. Electrical components such as transformers and inductors play a fundamental role in this field due to their wide range of functionalities—from a conventional household charger to sophisticated devices that regulate the current supplied to an entire city. Our aim is to identify meaningful relationships between various operational parameters. These correlations are used to detect trends, enabling a deeper understanding of the interactions between different variables. A three-stage methodology is presented. The first stage identifies fault co-occurrences across different error types using Apriori and FP-Growth techniques. In the second stage, a qualitative analysis is conducted through open-ended interviews. The final stage involves training a computational model to improve one of the identified sources that contributes to certain failures. A predictive model was successfully developed to propose possible solutions for issues detected in winding machines, achieving an F-measure of 84.11 %.

Keywords: Predictive analysis, decision-making, natural language processing, machine learning, and industrial systems.

1. Introducción

La industria eléctrica tiene una gran importancia en la actualidad, ya que interviene de un modo u otro en muchos aspectos de la vida cotidiana. En concreto, los componentes eléctricos como transformadores e inductores desempeñan un papel fundamental en este campo debido a sus funcionalidades, que van desde un cargador doméstico convencional hasta dispositivos más sofisticados que regulan la corriente suministrada a toda una ciudad. En el ámbito de las pruebas eléctricas a las que se someten estos componentes, es crucial garantizar que cumplen la funcionalidad, la estética y las dimensiones requeridas para desempeñar eficazmente el papel que se les ha asignado [1]. El análisis de los tiempos de prueba puede revelar variaciones significativas que muestran los defectos más comunes en ciertos probadores, lo que permite identificar ineficiencias en los procesos de prueba. Encontrar estas correlaciones es crucial, ya que ayuda a detectar problemas recurrentes, optimizar los procedimientos y mejorar la fiabilidad de los productos. Esto no solo reduce costos y mejora la eficiencia, sino que también previene fallos futuros, garantiza estándares de calidad más altos y mejora la satisfacción del cliente. En resumen, permite a las empresas ser más competitivas y ofrecer productos de mejor calidad [9,10].

2. Trabajos relacionados

La Industria 4.0 ha impulsado la aplicación de la inteligencia artificial (IA) en la fabricación, con una atención destacada a la detección de defectos. El aprendizaje profundo se ha consolidado como una herramienta eficaz para identificar anomalías complejas en productos y procesos [4]. Paralelamente, la visualización de datos emerge como un componente crucial para monitorizar y comprender anomalías en entornos industriales [5]. Además, el procesamiento del lenguaje natural (PLN) emerge como una técnica valiosa para el diagnóstico automatizado de fallos, que analiza los informes y registros de mantenimiento [6]. El reconocimiento de patrones, a través del aprendizaje automático, también contribuye significativamente a la detección de defectos mediante la identificación de características específicas en datos visuales y de sensores [7]. Por último, la integración de la detección de anomalías basada en IA con la visualización de datos en tiempo real optimiza la supervisión de procesos, permitiendo respuestas rápidas y decisiones informadas [8]. Juntas, estas tecnologías impulsan la evolución hacia sistemas de fabricación más inteligentes y autónomos. En contraste nuestro trabajo implementa una solución que va de lo general a lo específico, es decir, primero se buscan los fallos más frecuentes para posteriormente tratar de llegar a las posibles fuentes de los fallos.

3. Metodología

Para el desarrollo del análisis de las correlaciones entre errores, se realizó combinando técnicas de análisis exploratorio y técnicas de procesamiento de lenguaje natural (PLN) Ver Figura 1 .

3.1. Recolección y preparación de Datos

El conjunto de datos se compuso de registros de errores en a partir de una base de datos de un sistema industrial, organizados en un archivo Excel de manera mensual por todo un año. Cada fila contenía hasta tres errores registrados en conjunto, clasificados en tres categorías:

- **Errores tipo A** (A1, A2, ..., A10). Son errores de pruebas de Funcionalidad del producto.
- **Errores tipo B** (B1, B2, B3). Corresponde a errores de pruebas de Dimensión del producto.
- **Errores tipo C** (C1, C2, ..., C5). Son errores de pruebas de Estética del producto.

3.2. Análisis exploratorio de datos

Se implementó un enfoque exploratorio para identificar patrones en los datos:

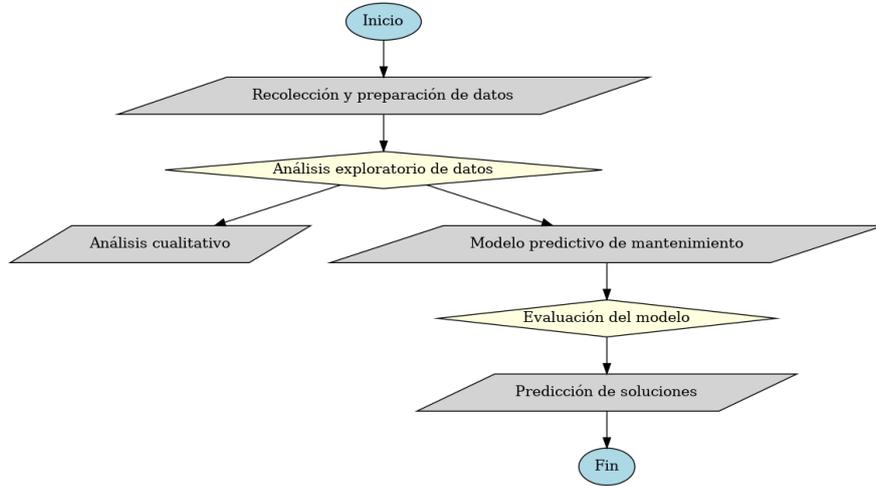


Fig. 1. Diagrama de flujo de la metodología implementada.

- **Heatmap de correlaciones:** Se utilizó la biblioteca `seaborn` para visualizar la frecuencia con la que ciertos pares de errores aparecen juntos. Esto permitió identificar combinaciones recurrentes de errores.

Para complementar el análisis exploratorio, se aplicaron los algoritmos de minería de datos:

- **Apriori:** Utilizado para generar reglas de asociación basadas en la probabilidad de que ciertos errores ocurran juntos [2].
- **FP-Growth:** Implementado para identificar patrones frecuentes en los datos de manera más eficiente en comparación con Apriori, aprovechando la estructura FP-tree [3].

Ambos algoritmos permitieron obtener reglas de la forma:

$$\text{Si ocurre } A3 \rightarrow \text{Probabilidad del } 80\% \text{ de que ocurra } B3. \quad (1)$$

Estas reglas fueron comparadas con las correlaciones observadas en el heatmap, permitiendo validar los patrones previstos y descubrir nuevas asociaciones no previstas.

3.3. Análisis cualitativo

Se realizaron entrevistas semiestructuradas en la planta industrial para estimar los costos de retrabajos y fabricación de los productos manufacturados. El diseño de las entrevistas fue de tipo abierto, permitiendo que los ingenieros, supervisores y operarios proporcionaran detalles adicionales que consideraran

relevantes. Se formularon preguntas sobre tiempos de retrabajo, uso de materiales adicionales, intervención de técnicos y paros de línea.

Posteriormente, se realizó un análisis temático de las entrevistas. Se agruparon las respuestas en categorías comunes, como “desgaste de maquinaria”, “ajustes de parámetros”, “fallos de sensores” y “problemas de calibración”. Este análisis permitió identificar factores críticos que impactaban en los costos y la eficiencia operativa.

De los datos recolectados, se encontró que las máquinas embobinadoras eran una de las principales fuentes de fallos y desperdicios en producción, debido a su antigüedad y falta de mantenimiento preventivo. Esto proporcionó la base para el desarrollo de un modelo predictivo orientado al mantenimiento de estos equipos.

3.4. Modelo predictivo

Derivado del análisis exploratorio y cualitativo, se identificó como una fuente significativa de fallos a las máquinas embobinadoras. En las entrevistas abiertas, los técnicos señalaron que estas máquinas generaban un alto porcentaje de desperdicio debido a fallas recurrentes.

Para abordar esta problemática, se construyó un modelo de predicción de mantenimiento basado en técnicas de aprendizaje automático. El objetivo fue anticipar el tipo de intervención requerida, minimizando los tiempos de solución y optimizando los recursos técnicos.

Carga y limpieza de datos:

- Se importó un archivo CSV con registros históricos de mantenimiento de un año.
- Se eliminaron espacios en blanco, valores nulos y registros inconsistentes.
- Se seleccionaron como variables de interés el “problema reportado por el operador” y la “solución realizada por el técnico”.

Categorización de problemas comunes:

- Se definieron dos categorías principales de soluciones: **Ajustes** (modificaciones de parámetros o configuraciones) y **Sensores** (calibración, limpieza o reemplazo de sensores).
- Las categorías se asignaron manualmente a cada registro de solución, validando con especialistas de la planta.

Transformación y modelado:

- Se utilizó **TF-IDF Vectorizer** para transformar los textos de los reportes en representaciones numéricas, capturando la importancia relativa de las palabras.
- Los datos se dividieron en conjuntos de entrenamiento (70 %) y prueba (30 %).

- Se entrenaron tres modelos de clasificación supervisada:
 - Regresión Logística
 - Random Forest
 - Máquinas de Soporte Vectorial (SVM)
- Los modelos se evaluaron utilizando métricas de Precisión, Exactitud, Recall y F1-Score.

Generación de recomendaciones:

- El modelo de mejor desempeño (Regresión Logística) se utilizó para generar **tres recomendaciones** para cada fallo reportado.
- Las recomendaciones se obtuvieron considerando las soluciones históricas más frecuentes asociadas a cada tipo de problema, ponderadas según la similitud de la descripción del fallo.

Visualización y validación:

- Se generaron mapas de calor de matrices de confusión para comparar los modelos.
- Se construyó una tabla con las principales recomendaciones agrupadas por tipo de problema.

Carga y limpieza de datos

- Se importó un archivo CSV con registros de mantenimiento.
- Se eliminaron espacios en blanco, valores nulos y se filtraron los datos para centrarnos en las columnas del problema reportado por el operador y la solución que realizó el técnico.

Categorización de problemas comunes Los problemas se clasificaron en dos categorías:

- **Ajustes:** Soluciones relacionadas con cambios de parámetros y configuraciones.
- **Sensores:** Soluciones que involucran calibración, limpieza o reemplazo de sensores.

Preparación del modelo

- Se utilizó **TF-IDF Vectorizer** para transformar los reportes en representaciones numéricas.
- Se dividieron los datos en **conjuntos de entrenamiento y prueba**.

Características utilizadas:

- Como características (features) de entrada para el modelo se utilizaron los reportes de fallos escritos por los operadores, los cuales fueron procesados mediante la técnica de **TF-IDF Vectorization**.

- Cada reporte se transformó en un vector numérico que representa la relevancia de las palabras clave en el contexto de fallos y soluciones históricas.
 - El modelo clasifica cada reporte en una de las dos categorías: “**Ajustes**” o “**Sensores**”, basándose en el contenido textual del problema reportado.
- Se probaron varios modelos de Machine Learning:
- **Regresión Logística**
 - **Random Forest**
 - **SVM (Máquinas de Soporte Vectorial)**
- Se evaluaron los modelos en términos de **precisión, exactitud y matriz de confusión**.

Predicción de soluciones El modelo entrenado genera soluciones basadas en problemas reportados, con tres respuestas posibles por falla. Se analizaron los siguientes casos: cortos e insuficiencias, sensor con falso contacto, falta de flux, falló la máquina y sensor quemado.

Creación de tabla comparativa Se organizó una tabla con tres soluciones por problema. Cuando no había suficientes datos, se buscaron respuestas similares en la misma categoría.

Visualización y análisis Se generaron gráficos de calor para analizar la matriz de confusión y evaluar el rendimiento del modelo, identificando patrones en las fallas más comunes.

4. Resultados

4.1. Análisis exploratorio

Evaluando la coherencia del comportamiento de las correlaciones encontradas en este sistema industrial, se validaron las que se tenía una noción anteriormente. Además, se identificaron nuevas relaciones que podrían ser útiles para el desarrollo de estrategias preventivas y correctivas. Como se puede ver en la siguiente figura. Esta matriz fue generada con el histórico mensual de los fallos más recurrentes. Por ejemplo, se observa A3 y B3 con un valor de 214, esto significa que un error de tipo B (de Dimensionalidad) afecta directamente a la prueba de Funcionalidad (A3) realizada al producto. Podemos decir que B3 influye en A3. Dada la magnitud alta de productos fabricados, la visualización en una matriz permite identificar rápidamente los fallos.

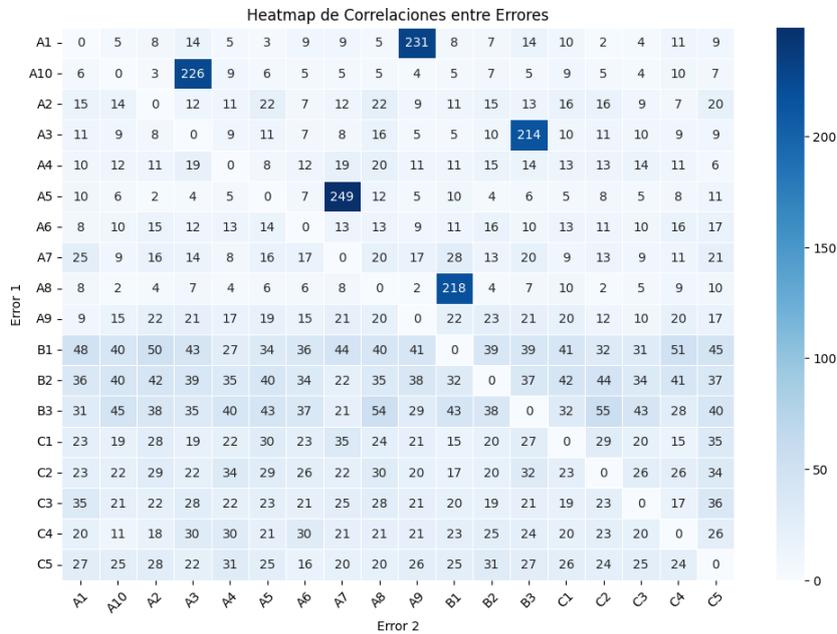


Fig. 2. Mapa de calor anual perteneciente al año 2024.

4.2. Entrevistas

Se realizaron entrevistas a 10 personas de la planta, entre ellas personal de ingeniería, gerencia y producción para identificar las causas de fallo en el proceso, se utilizó un método de entrevista del tipo de respuesta abierta para obtener un tipo de respuesta más viable y relajante. Se determinó que el problema podría originarse por el desgaste y la falta de ajuste de las máquinas de embobinado, debido a su antigüedad. Por ello, se decidió implementar un modelo predictivo para abordar esta problemática.

4.3. Modelo predictivo

Se implementó un modelo de predicción de mantenimiento capaz de sugerir soluciones técnicas automatizadas basadas en datos históricos, mejorando la eficiencia en la resolución de fallas.

El gráfico de matrices de confusión compara los modelos Naive Bayes, Random Forest y Regresión Logística. Los tonos oscuros en la diagonal principal indican predicciones correctas, mientras que los valores fuera de la diagonal representan errores de clasificación. La Regresión Logística destaca por tener menos errores y mayor precisión en comparación con los otros modelos.

Tabla 1. Comparación de métricas de modelos de clasificación.

Modelo	Precisión	Exactitud	Recall	F1-score
Naive Bayes	81.95 %	82.33 %	81.95 %	81.97 %
Random Forest	79.70 %	80.17 %	79.70 %	79.71 %
Regresión Logística	84.21 %	84.49 %	84.21 %	84.11 %

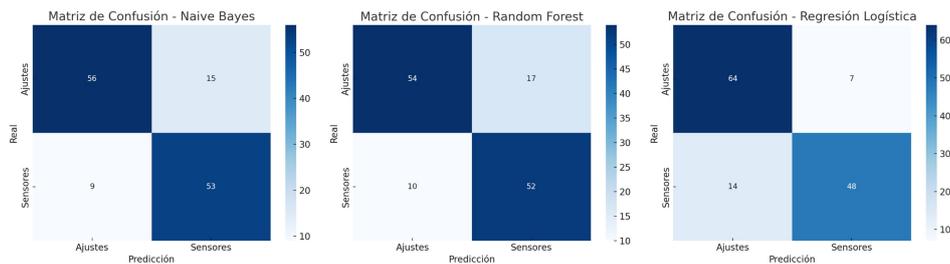


Fig. 3. Mapa de calor perteneciente a los modelos utilizados.

5. Discusión

Se puede apreciar la utilidad que nos proporciona cada una de estas herramientas en este campo específico, como lo son:

- **Heatmaps**, que facilitaron la observación directa de las combinaciones más frecuentes.
- **Grafos de correlaciones**, que destacaron los errores con mayor conexión.
- **Tablas de reglas de asociación**, que revelaron patrones emergentes y permitieron identificar errores clave en el sistema.
- **Algoritmos de clasificación**, que de una forma semiautomática permite al operador dar una solución rápida al fallo encontrado en las máquinas.

5.1. Discusión de Costo-Beneficio

El modelo predictivo propuesto aportaría beneficios económicos significativos para la planta industrial. A partir de los datos obtenidos en las entrevistas, se estimó que el costo promedio por retrabajo de un producto defectuoso oscila entre \$25 y \$40 USD, considerando materiales, tiempo de intervención y paros de línea. Realizamos el análisis bajo un escenario de 120 fallos mensuales relacionados con las máquinas embobinadoras lo cual representa un gasto estimado de entre \$3,000 y \$4,800 USD mensuales en retrabajos.

Con la ayuda del modelo predictivo se podría tener los siguientes beneficios :

- Un 25 % de fallos atendidos de forma reactiva, al anticipar ajustes preventivos en las máquinas.

Tabla 2. Soluciones recomendadas para problemas del operador.

Problema del operador	Solución 1	Solución 2	Solución 3
Cortos insuficiencias	Ajuste de parámetros	Ajuste de parámetros, TC:10MIN	Se ajustó presión a aflux y se ajustó sensor
Sensor con falso contacto	Se ajustó sensor	Se limpió sensor	Se cambió sensor
Falta de flux	Ajuste a cilindro de flux presión de aire	Ajuste de flux parámetro	Se llenó depósito y se estuvieron haciendo ajustes
Falla máquina	Ajuste de parámetros	Ajuste de sensor	Ajuste de olas
Sensor quemado	Se ajustó sensor	Se limpió sensor	Se cambió sensor

- Esto representa un ahorro estimado de entre \$750 y \$1,200 USD mensuales.

Para poder implementar nuestra propuesta los costos serian:

- El desarrollo y entrenamiento del modelo requiere aproximadamente 60 horas hombre.
- El costo estimado del proyecto sería de alrededor de \$3,000 USD (considerando salarios de especialistas en datos).

Por lo tanto, el retorno de inversión (ROI) podría alcanzarse en aproximadamente **tres a cuatro meses** de operación continua.

Este análisis preliminar sugiere que invertir en soluciones predictivas de mantenimiento no solo es técnicamente viable, sino económicamente rentable en el contexto industrial analizado.

Aunque este marco de solución se realizó para una empresa en particular, es posible llevar este esquema a otras fábricas. También logramos construir una solución que involucra tanto al personal de la empresa como el uso de técnicas computacionales.

6. Conclusiones

El análisis de correlaciones entre defectos en sistemas industriales ha demostrado ser una herramienta valiosa para la optimización de los procesos de ensayo de componentes eléctricos. A través de la implementación de técnicas de minería de datos, como los algoritmos Apriori y FP-Growth, fue posible identificar patrones recurrentes de errores, lo que facilita la toma de decisiones basada en datos para la mejora de la calidad y la eficiencia.

Los mapas de calor y los grafos de correlación permitieron visualizar las relaciones más significativas entre los distintos tipos de errores, lo que contribuye

a un diagnóstico más preciso y a la formulación de estrategias preventivas y correctivas. Estas metodologías no solo validaron correlaciones previamente conocidas, sino que también revelaron nuevas asociaciones que pueden ser clave para reducir fallos en el futuro. Encontramos que para la predicción de soluciones a fallos en máquinas, el algoritmo de regresión logística tuvo un buen desempeño. La integración de enfoques basados en datos en el análisis de defectos industriales, puede mejorar significativamente la fiabilidad de los productos, reducir costos y optimizar los procesos de prueba en una industria.

Referencias

1. Chapman, S. J.: Máquinas eléctricas. McGraw-Hill Education (2016)
2. Radhakrishnan, S., Pillai, V.: Comparative study on Apriori algorithm and FP Growth algorithm with pros and cons. *International Journal of Computer Science Trends and Technology*, 4(4), 161–165 (2016)
3. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, pp. 487–499, Santiago, Chile (1994)
4. Zhang, Y., Wang, X.: Deep Learning-Based Defect Detection in Manufacturing Systems: A Review. *Journal of Intelligent Manufacturing*, 32(3), 567–589 (2021)
5. Brown, P., Liu, J.: Visualization Techniques for Anomaly Detection in Industrial Processes. *Industrial Data Science Journal*, 15(2), 102–118 (2020)
6. Chen, H., Zhao, L.: Natural Language Processing for Automated Fault Diagnosis in Industry 4.0. *IEEE Transactions on Industrial Informatics*, 25(4), 3210–3225 (2020)
7. Patel, R., Kumar, S.: Pattern Recognition in Industrial Defects: A Machine Learning Approach. *Machine Vision Applications*, 18(1), 45–62 (2022)
8. Lee, C., Thompson, B.: Integrating AI-Based Anomaly Detection with Real-Time Data Visualization. *Smart Manufacturing Review*, 10(1), 88–101 (2023)
9. Garvin, D. A.: Competing on the eight dimensions of quality. *Harvard Business Review*, 65(6), 101–109 (1987)
10. Deming, W. E.: *Out of the Crisis*. MIT Press (1987)