# Exploratory Data Analysis para PiSA Farmacéutica

Jose Angel Govea Garcia[1], Daniel Alberto Sánchez Fortiz[1] and Diego Vértiz Padilla[1]

[1] *Tecnológico de Monterrey, Escuela de Ingeniería y Ciencias, Campus Guadalajara, México*

Reception date of the manuscript: 09/03/2025
Acceptance date of the manuscript: 16/03/2025
Publication date: 16/03/2025

**Abstract**—El presente estudio aborda la problemática del deterioro prematuro de una máquina sopladora de botellas utilizada en la producción de suero Electrolit en PiSA Farmacéutica. A través de un análisis exploratorio de datos (EDA) sobre los registros de más de 100 sensores, se identificaron fluctuaciones anómalas en variables clave de temperatura. Se implementó un modelo de detección de anomalías basado en K-Nearest Neighbors (KNN), que permitió detectar eventos críticos y coincidencias de anomalías en múltiples variables. Para validar estos hallazgos, se desarrolló un modelo de pronóstico temporal SARIMA, el cual sirvió como referencia del comportamiento esperado de las variables analizadas. La comparación entre los datos reales y las predicciones del modelo permitió confirmar la hipótesis de que el estrés térmico contribuye al desgaste acelerado de la máquina. Los resultados obtenidos refuerzan la necesidad de estrategias de mantenimiento predictivo que permitan anticipar fallos y extender la vida útil del equipo.

**Keywords**—Análisis exploratorio de datos, detección de anomalías, KNN, SARIMA, series de tiempo, industria farmacéutica, mantenimiento predictivo, modelos estadísticos, validacion de modelo

## I. Introducción

PiSA Farmacéutica es una empresa mexicana establecida internacionalmente. La fábrica cuenta con más de 1,500 marcas de medicamentos integradas en 17 líneas de especialidad y 14 plantas de producción en México. Su producto más conocido a nivel mundial es el suero Electrolit, el cual se envasa en botellas fabricadas mediante una máquina sopladora. Esta máquina es crítica para el proceso, ya que asegura que las botellas sean infladas de manera adecuada.

## II. Problema

La máquina sopladora encargada de la fabricación de botellas para el suero Electrolit ha reducido significativamente su vida útil operativa. Mientras que se espera que esta maquinaria funcione durante al menos un año antes de requerir reemplazo o mantenimiento mayor, actualmente solo alcanza un máximo de dos a tres meses de operación continua antes de presentar fallos críticos. Esta situación genera paros en la producción, incrementa los costos de mantenimiento y afecta directamente la eficiencia operativa y la rentabilidad de la planta.

Aunque se cuenta con un registro de datos provenientes de más de 100 sensores instalados en la máquina, aún no se ha realizado un análisis profundo que permita entender el comportamiento de las variables monitoreadas y encontrar patrones que expliquen o predigan estos fallos. Frente a esta situación, surge la necesidad de realizar un análisis exploratorio de datos (EDA), aplicar técnicas de detección de anomalías y hacer predicciones que permitan anticipar los problemas y mejorar la confiabilidad de la maquina.

Contact data: Daniel Alberto Sánchez Fortiz, a01255138@tec.mx

## III. Hipótesis

El estrés térmico generado durante el proceso de soplado de las botellas de Electrolit está provocando un deterioro acelerado en la vida útil de la máquina sopladora. Este fenómeno se ve reflejado en las fluctuaciones anómalas de las variables de temperatura registradas por los sensores, indicando posibles fallos en el control térmico del sistema.

## IV. Metodología

### *a. Datos*

Se analizaron dos archivos proporcionados por PiSA, cada uno con un tamaño aproximado de 1.8 GB, correspondientes a los meses de noviembre de 2024 y enero de 2025. Cada dataset contiene alrededor de 45 millones de registros en formato Parquet.

### *b. Procesamiento de los Datos*

Los datos brutos provenientes del sistema de monitoreo de la máquina sopladora fueron procesados utilizando la librería pandas en Python, apoyada por otros pyarrow y numpy. Los archivos de origen estaban en formato Parquet, que es un formato de almacenamiento columnar eficiente, utilizado para manejar grandes volúmenes de datos. Dado el tamaño de los archivos originales, se implementó un esquema de proce-

samiento por lotes (*chunks*) de 100,000 registros. Esto permitió realizar una lectura controlada de la información, reduciendo el consumo de memoria y facilitando el filtrado en ventanas de tiempo específicas.

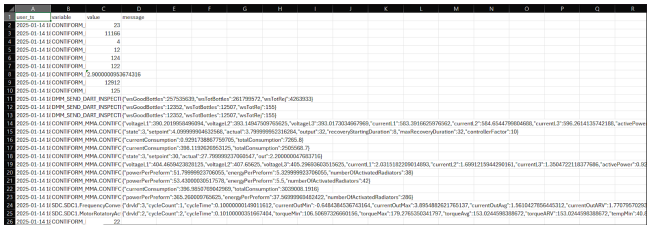

**Fig. 1:** Dataset Original.

Durante el recorrido de cada chunk, los registros fueron filtrados para mantener aquellos cuya marca de tiempo (`user ts`) se encontrara dentro del rango establecido. Los resultados se consolidaron en un DataFrame único, que representaba todos los datos disponibles en la ventana seleccionada.

El conjunto de datos crudo contenía cuatro columnas clave:user_ts, variable, valor y message. En particular, la columna message almacenaba información adicional en formato JSON, la cual fue convertida a un diccionario para su posterior expansión en columnas adicionales.

Se definió una lista de variables de interés, seleccionadas por su relevancia para el monitoreo del comportamiento térmico de la máquina. Entre estas se incluyeron temperaturas de los cuellos de los preformas, temperaturas de los circuitos de enfriamiento, y lecturas de capas de control de calefacción, entre otras.

Posteriormente, se realizó un proceso de pivoteo, en el que se transformó la estructura de los datos de formato largo a formato ancho. Esto permitió agrupar las mediciones por timestamp (user ts) y distribuir cada variable en una columna independiente.

Al observar que muchos de los datos eran nulos, realizamos en cada variable una agrupación por segundo, juntando los registros por segundo en lugar de milésimas de segundo. Esto nos permitió tener un mejor manejo de los datos para su análisis como series de tiempo.

A continuación, se realizó una limpieza adicional de los nombres de las variables para facilitar su interpretación, eliminando prefijos redundantes como `CONTIFORM MMA.CONTIFORM MMA1.`

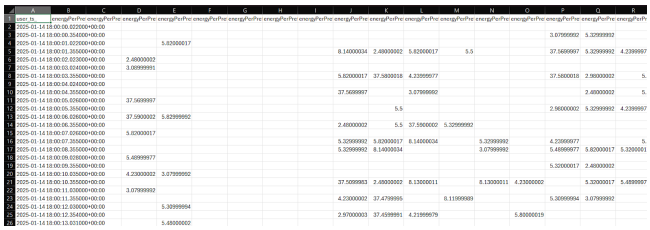

**Fig. 2:** Dataset pivoteado.

## *c. Selección de Variables*

Para reducir el ruido y asegurar la calidad del análisis, se seleccionaron únicamente aquellas variables de temperatura que contaran con suficiente cantidad de datos válidos. Se aplicó un percentil 80 sobre el conteo de valores disponibles por variable, de modo que solo aquellas con mayor cobertura fueran incluidas en el análisis posterior.

El DataFrame resultante incluía los registros filtrados y listos para el análisis de series de tiempo, en el cual se aseguró que:

- Los timestamps fueran únicos.
- Se ajustara la zona horaria a UTC.
- Se agruparan las mediciones por timestamp tomando el valor máximo en caso de duplicados.

Algunas de las variables que utilizamos:

1. `energyPerPreform_CurrentPreform NeckFinishTemperature.0`
2. `energyPerPreform_CurrentPreform TemperatureOvenInfeed.0`
3. `numberOfActivatedRadiators_CurrentPreform NeckFinishTemperature.0`
4. `powerPerPreform_CurrentPreform NeckFinishTemperature.0`
5. `value_ActualTemperatureCoolingCircuit2.0`
6. `value_CurrentPreformTemperature OvenInfeed.0`

## *d. Detección de Anomalías*

Para la detección de anomalías en las series de tiempo analizadas, se empleó el algoritmo K-Nearest Neighbors (KNN), utilizando la implementación proporcionada por la librería PyOD (Python Outlier Detection). Este método se basa en la premisa de que las observaciones anómalas presentan un comportamiento significativamente distinto al del resto de los datos, lo que se refleja en una mayor distancia respecto a sus vecinos más cercanos dentro del espacio multidimensional.

Se eligió KNN debido a su simplicidad y efectividad para la detección de *outliers* en *datasets* con características no lineales y sin supuestos de distribución específicos. Además, su capacidad de adaptarse a distintas escalas de datos lo convierte en una opción adecuada para la problemática planteada.

Previamente a la aplicación del modelo, las series temporales fueron suavizadas mediante una media móvil de 10 observaciones, con el objetivo de reducir el impacto del ruido extremo y evitar falsos positivos en la detección.

El algoritmo KNN identifica anomalías calculando la distancia media desde cada punto hacia sus k vecinos más cercanos. En este analisis, se determinaron los hiperparámetros de la siguiente manera:

- n neighbors = 20
- contamination = 0.0025

Esto significa que el modelo asume que el 0.25% de los datos analizados son anómalos. El ajuste de estos hiperparámetros se realizó con el objetivo de lograr un equilibrio adecuado

entre la sensibilidad del modelo (capacidad de detectar verdaderas anomalías) y su precisión (minimización de falsos positivos).

Tras la detección, se realizó un análisis de coincidencias temporales, identificando eventos donde múltiples variables presentaban anomalías de manera simultánea, lo que refuerza la hipótesis de la existencia de fallas sistémicas.

El resultado del análisis permite identificar observaciones cuyo comportamiento difiere de manera significativa respecto a la tendencia general de la serie de tiempo. Dichas observaciones fueron clasificadas como anómalas y se destacan visualmente en las gráficas de resultados.

### *e. SARIMA*

Tras desarrollar nuestro detector de anomalías, buscamos fortalecer el análisis para aproximarnos a una refutación sólida de nuestra hipótesis. Con este objetivo, decidimos realizar un pronóstico del comportamiento esperado de las variables de temperatura bajo condiciones normales. Esto nos permitió comparar el comportamiento real de los datos en los momentos de anomalía con el comportamiento proyectado, identificando así desviaciones significativas.

Para este pronóstico implementamos un modelo SARIMA (Seasonal AutoRegressive Integrated Moving Average), seleccionado por su capacidad para modelar series temporales que presentan tanto tendencias como estacionalidades, características observadas en las variables analizadas.

El proceso que llevamos a cabo para implementar el modelo SARIMA fue el siguiente:

1. **Selección de la variable objetivo**: El primer paso fue seleccionar las variables a modelar. Se decidió enfocar el análisis en tres variables, las cuales presentaron hallazgos significativos durante la detección de anomalías. Estas variables fueron:

    - `energyPerPreform_CurrentPreformNeckFinishTemperature.0`
    - `powerPerPreform_CurrentPreformNeckFinishTemperature.0`
    - `numberOfActivatedRadiators_CurrentPreformNFinishTemperature.0`

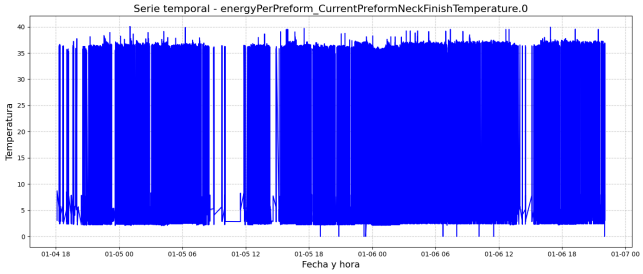


**Fig. 3:** Serie temporal de la variable energyPerPreform$_{C}urrentPreformNeckFinishTemperature$.0.

2. **Pruebas preliminares**: Posteriormente, se analizaron las series temporales de las variables seleccionadas para comprender su comportamiento. Se realizaron pruebas de Dickey-Fuller aumentada (ADF) para verificar la estacionariedad y se examinaron las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) para validar la viabilidad de un modelo SARIMA.

3. **Muestreo temporal**: Debido a las limitaciones computacionales, no fue posible trabajar con la totalidad de los datos en su resolución original (por segundo). Se optó por realizar un muestreo a intervalos de 10 segundos y, posteriormente, se aplicó un proceso de interpolación para aproximar la cantidad real de datos durante la fase de predicción. Esto permitió conservar la mayor cantidad de información relevante, reduciendo la carga computacional.

4. **Definición de hiperparámetros SARIMA**: Se identificaron los mejores parámetros para el componente ARIMA mediante un proceso iterativo de validación empírica. Una vez definidos, se determinaron los parámetros de la parte estacional del modelo. La separación de este proceso se debió a las restricciones computacionales mencionadas.

5. **Entrenamiento del modelo SARIMA**: Con los hiperparámetros establecidos, se entrenó el modelo utilizando una ventana deslizante (*sliding window*) de 7 días para entrenamiento y 2 días para predicción, avanzando 2 días por iteración hasta cubrir el período de predicción requerido. Nuevamente, se trabajó con el muestreo a 10 segundos para mantener la eficiencia computacional.

6. **Análisis de los residuos**: Tras el entrenamiento, se realizó un análisis exhaustivo de los residuos. Se confirmó que se comportan en su mayoría como ruido blanco, sin autocorrelación significativa. Asimismo, se observó homocedasticidad, evidenciada en una varianza prácticamente constante a lo largo del tiempo. También se evaluó la distribución de los residuos, obteniendo resultados próximos a la normalidad.

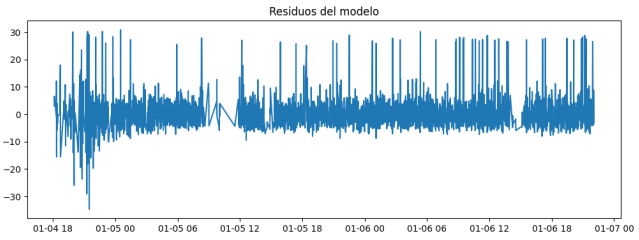


**Fig. 4:** Comportamiento de los residuos del modelo SARIMA

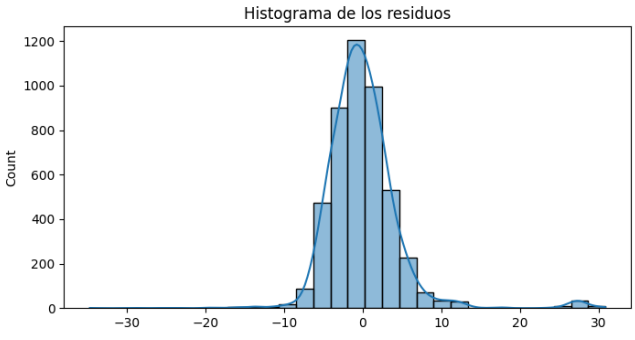


**Fig. 5:** Histograma de los residuos del modelo SARIMA.

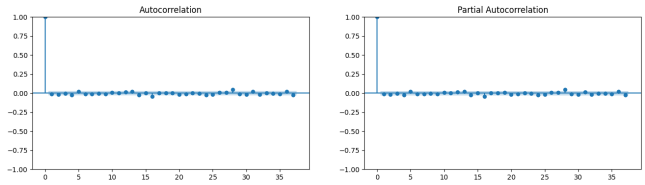


**Fig. 6:** Funciones de Autocorrelación (ACF) y Autocorrelación Parcial (PACF) de los residuos del modelo SARIMA.

7. **Evaluación del modelo**: Se calcularon las métricas de desempeño para cada predicción, utilizando el MAE (*Mean Absolute Error*) y el RMSE (*Root Mean Squared Error*). Los valores obtenidos indicaron un nivel aceptable de error en las predicciones, reflejando la capacidad del modelo para capturar el comportamiento esperado de la serie temporal bajo condiciones normales.

El pronóstico obtenido mediante SARIMA sirvió como modelo de referencia para contrastar los datos reales observados. En aquellos casos donde el comportamiento real se desvió significativamente del pronóstico, se evidenció la ocurrencia de anomalías detectadas previamente por el algoritmo KNN, lo que refuerza la hipótesis de fallas o comportamientos anómalos en el sistema.

El código completo de implementación del modelo SARIMA, así como el procesamiento de los datos y la detección de anomalías mediante KNN, se encuentra disponible en el siguiente repositorio:

## V. Resultados

Los resultados se dividen en dos fases principales: la detección de anomalías utilizando el algoritmo K-Nearest Neighbors (KNN) y la posterior validación mediante pronóstico SARIMA.

### *a. Detección de anomalias - KNN*

El análisis reveló la presencia de múltiples anomalías a lo largo de la serie temporal. Sin embargo, la mayoría de estas correspondieron a eventos singulares, es decir, anomalías que se presentaban de manera aislada en una sola variable, dificultando la interpretación y la identificación de un patrón significativo.

El hallazgo más relevante surgió al identificar que tres variables específicas presentaban anomalías de manera simultánea, coincidiendo exactamente en el mismo segundo en múltiples ocasiones. Estas variables fueron:

- `energyPerPreform_CurrentPreformNeckFinish Temperature.0`
- `powerPerPreform_CurrentPreformNeckFinish Temperature.0`
- `numberOfActivatedRadiators_CurrentPreformNeck FinishTemperature.0`

La ocurrencia conjunta de anomalías en estas tres variables sugiere la existencia de eventos anómalos relacionados con un mal funcionamiento de la máquina o condiciones operativas fuera de los parámetros establecidos. A partir de este hallazgo, se decidió enfocar el análisis únicamente en estas tres variables.

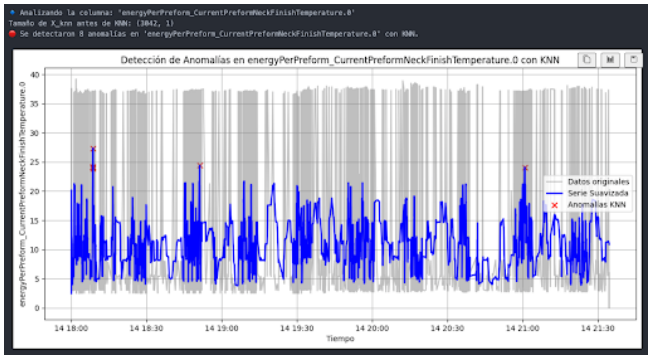


**Fig. 7:** Detección de Anomalías de la variable energyPerPreform_CurrentPreformNeckFinish Temperature.0

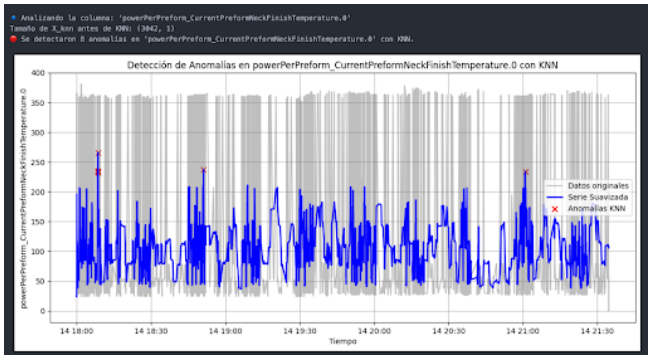


**Fig. 8:** Detección de Anomalías de la variable powerPerPreform_CurrentPreformNeckFinish Temperature.0

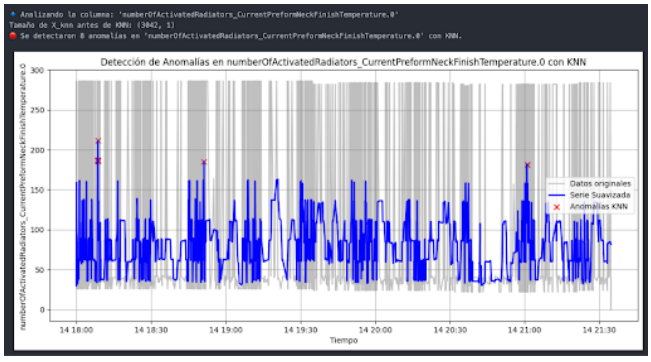


**Fig. 9:** Detección de Anomalías de la variable numberOfActivatedRadiators_CurrentPreformNeck FinishTemperature.0

### *b. SARIMA*

Los resultados obtenidos mediante el modelo SARIMA reflejan un desempeño sólido en la predicción de las tres variables seleccionadas. El modelo logró capturar de manera efectiva las tendencias y estacionalidades presentes en las series temporales, lo que se evidencia en la coherencia del pronóstico general y en las métricas de error obtenidas (MAE y RMSE), que se mantuvieron en rangos aceptables durante las diferentes iteraciones de la ventana deslizante. Esto indica que, bajo condiciones normales de operación, el modelo es capaz de predecir el comportamiento esperado de las variables analizadas con un nivel de precisión adecuado.

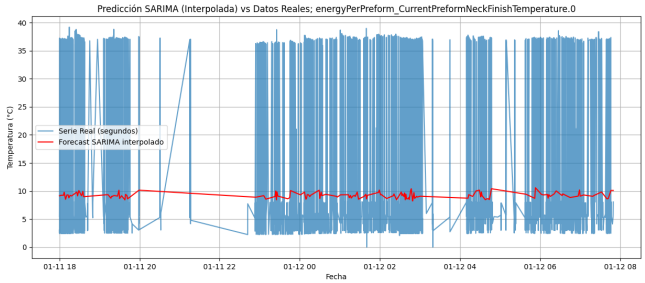


**Fig. 10:** Predicción del modelo SARIMA respecto a los valores reales de la variable energyPerPreform_CurrentPreformNeckFinishTemperature.0.

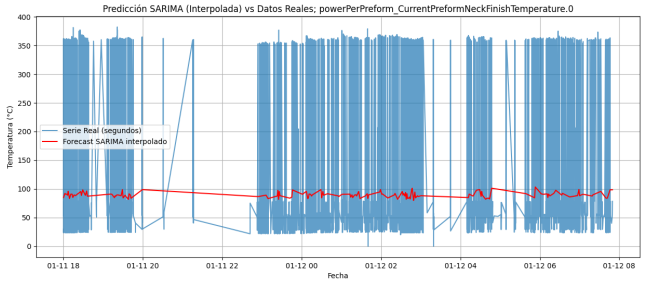


**Fig. 11:** Predicción del modelo SARIMA respecto a los valores reales de la variable powerPerPreform_CurrentPreformNeckFinishTemperature.0.

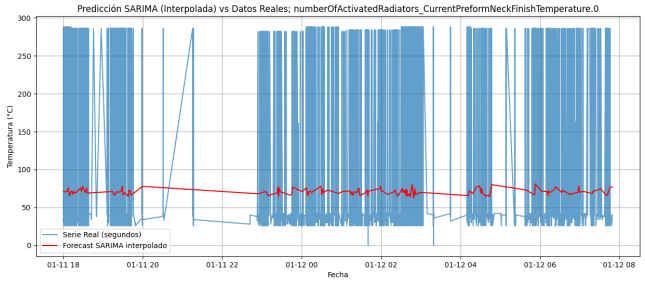


**Fig. 12:** Predicción del modelo SARIMA respecto a los valores reales de la variable numberOfActivatedRadiators_CurrentPreformNeckFinishTemperature.0.

Sin embargo, es importante destacar que debido al proceso de muestreo utilizado para reducir la granularidad de los datos (de registros por milisegundo a registros cada 10 segundos), el modelo genera un pronóstico suavizado. Esta suavización puede limitar la sensibilidad del modelo para capturar eventos abruptos o fluctuaciones rápidas en la señal, lo que dificulta confirmar con absoluta certeza la magnitud y el momento exacto de ciertas anomalías detectadas previamente por el algoritmo KNN.

A pesar de estas limitaciones, el modelo SARIMA cumplió con su propósito principal: proporcionar una referencia confiable del comportamiento normal de las variables, permitiendo contrastar los datos reales y evidenciar desviaciones significativas en los periodos donde se detectaron anomalías grupales. Esto refuerza la hipótesis planteada sobre el deterioro acelerado de la máquina y sienta las bases para el desarrollo de un sistema de mantenimiento predictivo más robusto en el futuro.

## VI. Conclusiones

Estos hallazgos permiten confirmar la hipótesis planteada: las fluctuaciones anómalas en las temperaturas registradas por los sensores generan estrés térmico adicional en la máquina sopladora, contribuyendo al deterioro acelerado de sus componentes. La combinación de métodos de detección de anomalías y modelos de pronóstico no solo valida esta hipótesis, sino que también constituye una base sólida para implementar estrategias de mantenimiento predictivo orientadas a extender la vida útil del equipo y reducir los tiempos de inactividad.