**EDEM**

**MÁSTER DATA ANALYTICS**

**CURSO 2023/2024**

**Trabajo de Fin de Máster**

|  |
| --- |
| **TÍTULO: Modelo de detección de polución en planta** |

|  |
| --- |
| **Alumnos:**  Jorge Domínguez Lozano  Ignacio Reyes Vázquez  Borja Cabo Huélamo  Francisco Tudela Muñoz |

|  |
| --- |
| **Tutor EDEM:** Pedro Nieto Peláez  **Tutor Empresa:** Sara Adam Goig  **Empresa:** Roquette |

Septiembre de 2024



**Resumen**

La contaminación industrial es un problema que ha estado siempre presente en la sociedad moderna y que afecta negativamente tanto al medio ambiente como a la salud pública. Identificar rápidamente las fuentes de las que procede la contaminación es esencial para poder mitigar sus efectos y mejorar la sostenibilidad de los procesos industriales. En nuestro estudio, hemos llevado a cabo una investigación en el uso de técnicas de machine learning para analizar los datos obtenidos de sensores industriales. Mediante el desarrollo de modelos predictivos, hemos logrado identificar patrones y relaciones en los datos que permiten detectar el origen de la contaminación de manera más precisa y oportuna. Este enfoque no solo ayuda a identificar las fuentes de contaminación, sino que también proporciona una herramienta valiosa para la toma de decisiones en tiempo real y la optimización de procesos industriales, con el fin de reducir el impacto ambiental.

**Palabras clave: Contaminación, Machine Learning, Sensores, Impacto Ambiental**

**Abstract**

Industrial pollution is an ever-present problem in modern society, adversely affecting both the environment and public health. Quickly identifying the sources of pollution is essential to mitigate its effects and improve the sustainability of industrial processes. In our study, we have conducted research on the use of machine learning techniques to analyze data obtained from industrial sensors. By developing predictive models, we have been able to identify patterns and relationships in the data that allow us to detect the source of pollution in a more accurate and timely manner. This approach not only helps to identify the sources of pollution, but also provides a valuable tool for real-time decision making and optimization of industrial processes to reduce environmental impact.

**Keywords: Pollution, Machine Learning, Sensors, Environmental Impact**

ÍNDICE DE CONTENIDO

[1. Introducción. 4](#_Toc176281717)

[1.1 Objetivos generales. 4](#_Toc176281718)

[2. Metodología de trabajo. 4](#_Toc176281719)

[3. Arquitectura. 5](#_Toc176281720)

[4. Exploración y Preprocesamiento de los Datos. 6](#_Toc176281721)

[4.1 Origen de datos. 6](#_Toc176281722)

[4.2 Tratamiento de valores ausentes. 7](#_Toc176281723)

[4.3 Codificación de variables. 7](#_Toc176281724)

[5. Creación de Modelos 7](#_Toc176281725)

[5.1 Selección de Variables 7](#_Toc176281726)

[5.2 Regresión Logística 7](#_Toc176281727)

[5.3 Random Forest 7](#_Toc176281728)

[6.Análisis del Modelo Ganador 7](#_Toc176281729)

[7. Conclusiones 7](#_Toc176281730)

[8.Referencias y Bibliografía 7](#_Toc176281731)

[Anexos 7](#_Toc176281732)

ÍNDICE DE FIGURAS

[Figura 1: Trello con las tareas asignadas del 4o Sprint del trabajo 6](#_Toc176285175)

[Figura 2: Diagrama de la Arquitectura presentada 7](#_Toc176285176)

# 1. Introducción.

Roquette es una empresa multinacional francesa especializada en la producción de ingredientes de origen vegetal para los sectores de alimentación, nutrición y salud. Fundada en 1933 por los hermanos Dominique y Germain Roquette, la compañía se ha convertido en uno de los líderes mundiales en la transformación de materias primas vegetales como maíz, trigo, patatas y guisantes.

En nuestro caso, en Valencia se opera la producción de maíz.  
Sus principales productos incluyen almidones, proteínas vegetales, fibras dietéticas y polialcoholes.

Estos ingredientes se utilizan en una amplia gama de aplicaciones, desde alimentos y bebidas hasta productos farmacéuticos y cosméticos operando en más de 100 países. Cuenta con más de 25 plantas industriales en todo el mundo y emplea a alrededor de 8,000 personas. La empresa es conocida por su enfoque en la innovación y la sostenibilidad, invirtiendo constantemente en investigación y desarrollo para crear soluciones que respondan a las cambiantes demandas del mercado y a los desafíos globales en materia de salud y nutrición.

## 1.1 Objetivos generales.

En el contexto de la producción industrial, es importante destacar que las máquinas tienden a contaminar más al iniciar su ciclo de producción. Como ingenieros de datos y de inteligencia artificial, nuestra misión es crucial para mitigar este problema. Mediante el procesamiento de los datos generados por las máquinas y la aplicación de métodos de aprendizaje automático (ML), podemos desarrollar modelos predictivos capaces de anticipar la ocurrencia de contaminación e identificar qué máquinas específicas son propensas a causarla. Esta capacidad de predicción nos permite implementar medidas preventivas, optimizar los procesos de inicio y, en última instancia, reducir significativamente la huella ambiental de la producción industrial.

# 2. Metodología de trabajo.

Nuestro equipo implementó una metodología ágil para la ejecución del proyecto, estructurando el trabajo en sprints quincenales. Este enfoque y con la ayuda de la herramienta Trello, nos permitió:

1. Asignar tareas específicas a cada integrante del equipo.
2. Establecer objetivos claros y alcanzables para cada período de dos semanas.
3. Mantener un ritmo de trabajo constante y focalizado.

Además, para asegurar una comunicación efectiva y un seguimiento preciso del avance, instauramos dos tipos de reuniones clave que se implementaron: reuniones semanales “weeklys”, en las que, cada miembro compartía sus progresos, identificaba obstáculos y proponía soluciones, fomentando la colaboración y el apoyo mutuo y, por otro lado, las revisiones de sprint: Al concluir cada ciclo de dos semanas, evaluábamos los logros, analizábamos áreas de mejora y planificábamos el siguiente sprint.

Nuestro enfoque no se limitó únicamente al aspecto técnico. Reconociendo la importancia de comprender el contexto operativo real de nuestro cliente, decidimos realizar una visita a la planta de producción en Benifaió (Valencia). Esta visita nos permitió entender de primera mano los desafíos que enfrentaban en su proceso productivo y nos proporcionó valiosas perspectivas para desarrollar una solución más efectiva y alineada con sus necesidades reales.

Esta combinación de prácticas ágiles y conocimiento directo del entorno del cliente nos permitió avanzar de manera coherente y eficaz. El resultado fue el desarrollo de una solución que no solo cumplía con los requisitos técnicos, sino que también se ajustaba perfectamente a la realidad operativa del cliente, maximizando así su valor y aplicabilidad.

Figura 1: Trello con las tareas asignadas del 4o Sprint del trabajo

Captura de pantalla de un celular de un mensaje con una foto en el agua

Descripción generada automáticamente con confianza media

# 3. Arquitectura.

La arquitectura para procesar los datos de Roquette y prepararlos para el análisis predictivo comienza con la ingesta de datos donde Roquette sube los archivos CSV a través de una API segura con un endpoint específico, estos archivos se almacenan automáticamente en un bucket de Google Cloud Storage. Para su procesamiento inicial en la capa Bronce se utiliza una Cloud Function de GCP que se activa cuando se carga un nuevo archivo en el bucket, esta función lee el CSV procesa los datos y los carga en una tabla de BigQuery en la capa Bronce, utilizando un modelo incremental para agregar solo los nuevos datos a la tabla existente. Luego se procede a la transformación de datos para crear la capa Gold, donde los datos de la capa Bronce se procesan aplicando transformaciones, como el ajuste de formato de hora para consistencia y la “des-pivotación” de columnas para crear una columna por máquina, mejorando así la estructura para el análisis. los datos transformados se almacenan en una nueva tabla de BigQuery en la capa Gold. esta arquitectura permite un flujo de datos eficiente desde la fuente hasta un formato optimizado para el análisis predictivo de contaminación.

Para entrenar el modelo, se crea una API que, al ser llamada, extrae datos de los últimos días de la capa Gold en BigQuery. Esta API utiliza estos datos para entrenar el modelo predefinido que hemos desarrollado y que explicaremos posteriormente, ejecutándose esta API actualmente a demanda del cliente. Una vez entrenado, el modelo se guarda en Google Cloud para su posterior uso. Luego, se implementa otra API que lee la información más reciente de las máquinas. Esta segunda API utiliza el modelo guardado para determinar la probabilidad de que haya polución en el proceso productivo. Además, esta API identifica y proporciona un ranking de las cinco máquinas que más influyen en esa probabilidad de contaminación. Este enfoque permite una actualización continua del modelo y una evaluación casi a tiempo real (con gran posibilidad de pasar a ser de tiempo real usando métodos de streaming de datos como Kafka) de los riesgos de contaminación en la planta de producción de Roquette en Valencia, facilitando así la toma de decisiones preventivas y la optimización de los procesos industriales para minimizar el impacto ambiental.

Diagrama

Descripción generada automáticamenteFigura 2: Diagrama de la arquitectura presentada

# 4. Exploración y Preprocesamiento de los Datos.

Hemos desarrollado un flujo de trabajo integral para la exploración y preprocesamiento de datos, que constituye una parte esencial de la arquitectura de procesamiento de datos diseñada para el proyecto.

Como hemos mencionado en la parte de la arquitectura, la ingesta de datos se realiza a través de la carga de archivos CSV en un bucket de Google Cloud Storage. Este proceso es automatizado mediante el uso de Cloud Functions, que se activan de manera inmediata al detectar la carga de un nuevo archivo en el bucket. Esta función es responsable de leer los archivos CSV y convertirlos en DataFrames utilizando la biblioteca Pandas. Durante esta etapa, se presta especial atención a la estandarización de los datos temporales, mediante la conversión del campo de tiempo a un formato uniforme, haciendo que los formatos en fecha que nos viene por parte del cliente (inutilizables), sean usables. El siguiente paso es su carga en BigQuery.

Tras esto, se realiza una transformación avanzada de los datos para optimizar su estructura y facilitar el análisis. Para ello, aplicamos técnicas de agregación y despivotación. La agregación de valores máximos nos permite identificar los picos dentro de los datos, lo que es particularmente útil para detectar tendencias o anomalías en los mismos. La despivotación, por su parte, reorganiza los datos en un formato más adecuado para el análisis, convirtiendo las diferentes máquinas en columnas separadas. Este formato reformateado es almacenado en la Capa Gold de BigQuery, donde los datos quedan listos para ser utilizados en los modelos de ML.

## 4.1 Origen de datos.

Los datos que utilizamos en este proyecto provienen de múltiples archivos CSV proporcionados por Roquette con datos mensuales. Estos archivos contienen información detallada sobre el funcionamiento de diversas máquinas en sus instalaciones. Cada CSV captura la carga de las diferentes máquinas. Los intervalos de tiempo registrados en estos archivos no son uniformes.

## 4.2 Tratamiento de valores ausentes.

Para garantizar la integridad y consistencia de los datos que analizamos, utilizamos una técnica que rellena los valores faltantes basándose en el último valor disponible en la secuencia de datos. Esto es especialmente útil para mantener la continuidad temporal, ya que se asume que las condiciones no han cambiado significativamente entre los registros si no se dispone de nuevos datos en el tiempo, al ser máquinas constantes.  
  
En situaciones donde no existe un valor previo para completar la información faltante, procedemos a reemplazar los valores ausentes con un valor estándar de 0. Este paso asegura que todos los registros estén completos, evitando así problemas en análisis posteriores. Con esta estrategia combinada, nos aseguramos de que los datos sean consistentes y fiables, lo que es esencial para obtener resultados precisos en los análisis.

## 4.3 Codificación de variables.

# 5. Creación de Modelos

## 5.1 Selección de Variables

## 5.2 Regresión Logística

## 5.3 Random Forest

# 6.Análisis del Modelo Ganador

# 7. Conclusiones

# 8.Referencias y Bibliografía

# Anexos