**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**

AGRADECIMIENTOS

Queremos expresar nuestro más sincero agradecimiento a todas las personas que nos han apoyado y guiado a lo largo de este Máster en Data Analytics. En primer lugar, queremos agradecer a nuestros responsables del Trabajo Final de Máster, **Pedro Nieto Peláez** y **Sara Adam Goig**, por su apoyo constante y por ofrecernos su valiosa experiencia y orientación tanto en los aspectos académicos como en los prácticos de nuestro proyecto. Vuestra implicación y compromiso han sido fundamentales para nuestro crecimiento profesional y personal durante este tiempo.

Nos gustaría también agradecer a todos los profesores del máster, quienes nos han proporcionado las herramientas, el conocimiento y la inspiración para superar los retos que hemos encontrado en este camino. Gracias a vuestras enseñanzas y consejos, hemos podido desarrollar las habilidades necesarias para afrontar con éxito los desafíos del mundo del análisis de datos y la inteligencia artificial.

A lo largo de este curso, hemos sido afortunados de contar con un equipo docente de gran calidad, cuya pasión por el mundo del dato ha sido una fuente de motivación para nosotros. Cada uno de vosotros ha dejado una huella importante en nuestra formación, y estamos profundamente agradecidos por ello.

Por último, queremos agradecer a la institución académica **EDEM** y **Roquette**, la empresa que nos brindó la oportunidad de aplicar nuestros conocimientos en un contexto real, y a nuestros compañeros de equipo, por su colaboración y esfuerzo. Juntos hemos aprendido, crecido y enfrentado este reto con éxito.

ÍNDICE DE CONTENIDO

[1. Introducción. 6](#_Toc176869761)

[1.1 Objetivos generales. 6](#_Toc176869762)

[2. Metodología de trabajo. 7](#_Toc176869763)

[3. Arquitectura. 8](#_Toc176869764)

[4. Exploración y Preprocesamiento de los Datos. 9](#_Toc176869765)

[4.1 Origen de datos. 10](#_Toc176869766)

[4.2 Tratamiento de valores ausentes. 10](#_Toc176869767)

[4.3 Codificación de variables. 10](#_Toc176869768)

[5. Creación de Modelos 11](#_Toc176869769)

[5.1 Regresión Logística 12](#_Toc176869770)

[5.3 SVM 16](#_Toc176869771)

[5.4 Gradient Boosting 18](#_Toc176869772)

[5.5 Random Forest 20](#_Toc176869773)

[6. Análisis del Modelo Ganador 20](#_Toc176869774)

[7. Conclusiones 25](#_Toc176869775)

ÍNDICE DE FIGURAS

[Figura 1: Trello con las tareas asignadas del 4o Sprint del trabajo 7](#_Toc176869782)

[Figura 2: Diagrama de la arquitectura presentada 9](#_Toc176869783)

[Figura 3: Curva ROC Regresión Logística 12](#_Toc176869784)

[Figura 4: Curva Precision-Recall Regresión Logística 13](#_Toc176869785)

[Figura 5: Curva ROC Regresión Lineal 15](#_Toc176869786)

[Figura 6: Curva Precision-Recall Regresión Lineal 15](#_Toc176869787)

[Figura 7: Curva ROC SVM 17](#_Toc176869788)

[Figura 8: Curva Precision-Recall SVM 17](#_Toc176869789)

[Figura 9: Curva ROC Gradient Boosting 19](#_Toc176869790)

[Figura 10: Curva Precision-Recall SVM 19](#_Toc176869791)

[Figura 11: Curva ROC Random Forest 22](#_Toc176869792)

[Figura 12: Curva Precision-Recall Random Forest 23](#_Toc176869793)

[Figura 13: Curva Number of Trees vs OOB Error 23](#_Toc176869794)

# 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.

Una vez preprocesados los datos y tratados los valores nulos, hemos creado una nueva variable denominada **flag**, que se ha convertido en nuestra variable a predecir. Esta variable se definió en función de un criterio basado en la cantidad de contaminación presente. Si el valor de polución superaba los 2500 pp, asignamos un valor de 1 (indicando presencia de contaminación). En cambio, si el valor era inferior a 2500 pp, asignamos un valor de 0 (indicando ausencia de contaminación). Esta transformación nos permite abordar el problema como una tarea de clasificación binaria, facilitando el desarrollo y evaluación de los modelos de machine learning.

A continuación, nos centramos en la multicolinealidad entre las variables.

Calculamos la matriz de correlación y observamos que algunas máquinas presentaban una correlación muy alta entre sí, con coeficientes entre 0.8 y 1. Una correlación tan elevada indica que estas variables aportan información redundante al modelo, lo que puede afectar negativamente su desempeño y conducir a interpretaciones erróneas.

Para abordar este problema, decidimos eliminar las máquinas (variables) con una correlación superior a 0.8 y menor a 1 con respecto a otras variables. Al hacerlo, reducimos la dimensionalidad del conjunto de datos y eliminamos la redundancia, lo que contribuye a mejorar la eficiencia y la precisión del modelo. Este paso fue fundamental para garantizar que el modelo aprendiera de datos relevantes y no se viera influenciado por información duplicada.

Además de la transformación y selección de variables, procedimos a normalizar los datos utilizando la técnica de *StandardScaler*. Esta normalización fue necesaria debido a que algunos modelos de machine learning pueden ver su proceso de aprendizaje comprometido si las variables tienen escalas muy diferentes. *StandardScaler* ajusta los valores de cada variable para que tengan una media de 0 y una desviación estándar de 1, lo que garantiza que las diferencias en magnitud entre las variables no influyan de manera desproporcionada en el entrenamiento del modelo.

# 5. Creación de Modelos

En este apartado, procedimos a la creación de modelos de machine learning utilizando diferentes algoritmos y evaluamos el rendimiento de cada uno de ellos. Nuestro objetivo fue encontrar el modelo que ofreciera la mayor precisión posible, manteniendo un equilibrio adecuado entre la exactitud predictiva y la eficiencia computacional. Para lograrlo, implementamos una variedad de algoritmos, tanto de clasificación como de regresión, y evaluamos su desempeño utilizando métricas clave como la precisión, recall, F1-score y la curva ROC-AUC, con el fin de obtener una visión completa de su comportamiento.

Además, realizamos un ajuste de hiperparámetros a través de la técnica *Grid Search* para optimizar el rendimiento de cada modelo. Este proceso nos permitió identificar el modelo más adecuado para nuestras necesidades, no solo en términos de precisión, sino también considerando su capacidad para generalizar y evitar problemas de sobreajuste.

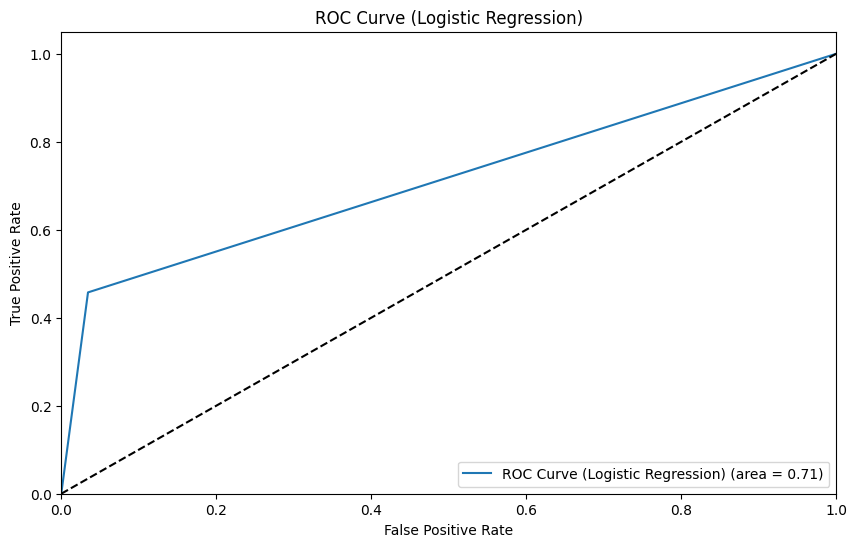
## 5.1 Regresión Logística

En el estudio realizado, el modelo de Regresión Logística implementado alcanzó una accuracy global de 0.89, lo que indicó una tasa de acierto del 89% en sus predicciones. Sin embargo, un análisis más detallado de las métricas de rendimiento reveló disparidades significativas en el desempeño del modelo entre las clases predichas.

La evaluación exhaustiva del modelo mostró un rendimiento asimétrico entre las clases. Para la clase 0 (*Sin polución*), el modelo demostró eficacia en la identificación y clasificación de instancias. La alta precisión de 0.91 indicó una baja tasa de falsos positivos, mientras que el elevado recall de 0.97 sugirió que el modelo capturó con éxito la gran mayoría de los casos negativos reales. El F1-score resultante de 0.94 confirmó un rendimiento robusto y equilibrado para esta clase.

En contraste, el modelo exhibió limitaciones significativas en la clasificación de instancias de la clase 1 (*Polución*). Aunque la precisión de 0.70 fue aceptable, indicando que el 70% de las predicciones positivas fueron correctas, el recall de 0.46 reveló una deficiencia crítica: el modelo solo identificó correctamente el 46% de los casos positivos reales. Esta discrepancia se reflejó en un F1-score moderado de 0.55, evidenciando un rendimiento subóptimo para la clase positiva.

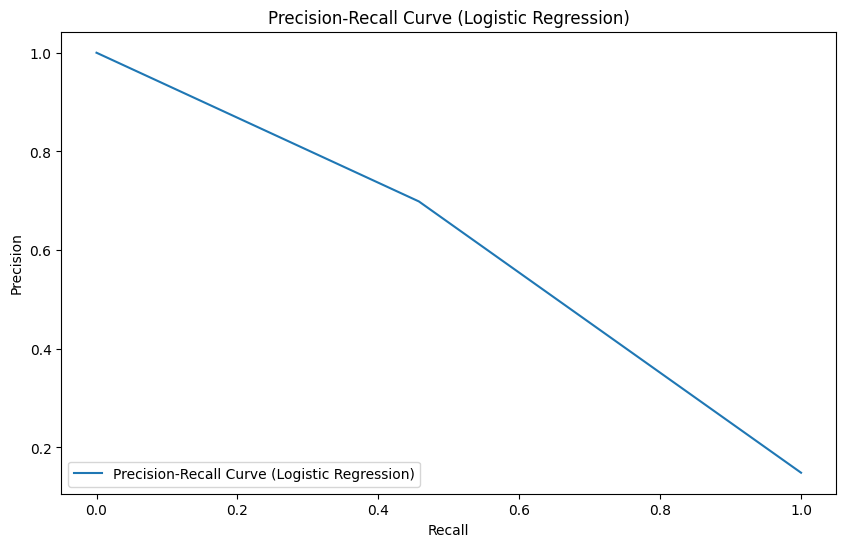
Figura 3: Curva ROC Regresión Logística



En el análisis del rendimiento del modelo, se empleó también la Curva ROC (Receiver Operating Characteristic) como herramienta de evaluación. Esta curva proporcionó una representación gráfica de la capacidad discriminativa del modelo a través de diversos umbrales de clasificación, ilustrando la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos.

El área bajo la curva ROC (AUC) resultante fue de 0.71, lo cual indicó un desempeño moderado del clasificador. Este valor sugirió que el modelo poseía una capacidad razonable para diferenciar entre las clases, aunque no alcanzó un nivel óptimo. En el contexto de la interpretación del AUC, donde un valor de 1.0 habría representado un clasificador perfecto y 0.5 habría equivalido a una clasificación aleatoria, el resultado de 0.71 posicionó al modelo en un rango de efectividad intermedia. La utilización de la Curva ROC no solo permitió una evaluación más completa del rendimiento global del modelo, sino que también facilitó la comparación potencial con otros modelos y la identificación del umbral de clasificación óptimo para equilibrar la precisión y el recall de manera efectiva.

Figura 4: Curva Precision-Recall Regresión Logística



El análisis del rendimiento del modelo se complementó con la evaluación de la Curva Precision-Recall, la cual ilustró la relación entre estos dos importantes indicadores a través de diversos umbrales de clasificación. La curva reveló un patrón característico de los modelos de clasificación binaria, donde se observó una disminución en la precisión a medida que se incrementaba el recall. Este comportamiento puso de manifiesto el compromiso inherente entre la capacidad del modelo para detectar correctamente la clase positiva y su habilidad para mantener una alta precisión en sus predicciones.

En el caso específico de este modelo, la curva demostró que para alcanzar niveles elevados de recall, es decir, para clasificar correctamente la mayoría de los casos positivos, el modelo sacrificó una cantidad considerable de precisión. Esta situación resultó en un aumento en la tasa de falsos positivos. Tal observación sugirió que, si bien existía la posibilidad de ajustar el modelo para mejorar su recall, dicha mejora vendría acompañada de una disminución en la precisión, un escenario que no necesariamente resultaría deseable.

5.2 Regresión Lineal

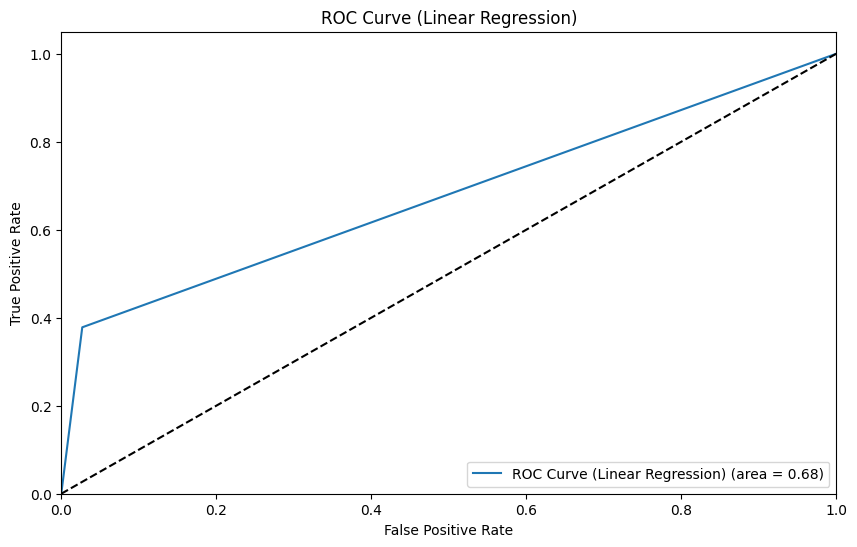
El modelo de Regresión Lineal obtuvo una accuracy de 0.884, lo que significó que el 88.4% de las predicciones fueron correctas. Sin embargo, fue importante observar otras métricas para comprender mejor el rendimiento del modelo. La precisión para la clase negativa resultó ser del 90%, indicando que una gran mayoría de las predicciones de esta clase fueron acertadas. En contraste, para la clase positiva, la precisión fue del 71%, lo que reflejó un desempeño moderado al clasificar correctamente las observaciones de esta categoría.

En cuanto al recall, se observó una marcada diferencia entre ambas clases. Para la clase positiva, solo el 38% de los casos reales fueron detectados correctamente, mientras que, para la clase negativa, el recall alcanzó un 97%. Esto evidenció una eficacia notablemente superior del modelo para identificar la clase negativa.

El F1-score, métrica que combina precisión y recall, arrojó resultados dispares: 0.49 para la clase positiva y 0.93 para la negativa. Esta disparidad reflejó un rendimiento significativamente más robusto en la clasificación de la clase negativa, contrastando con debilidades considerables en la detección de la clase positiva.

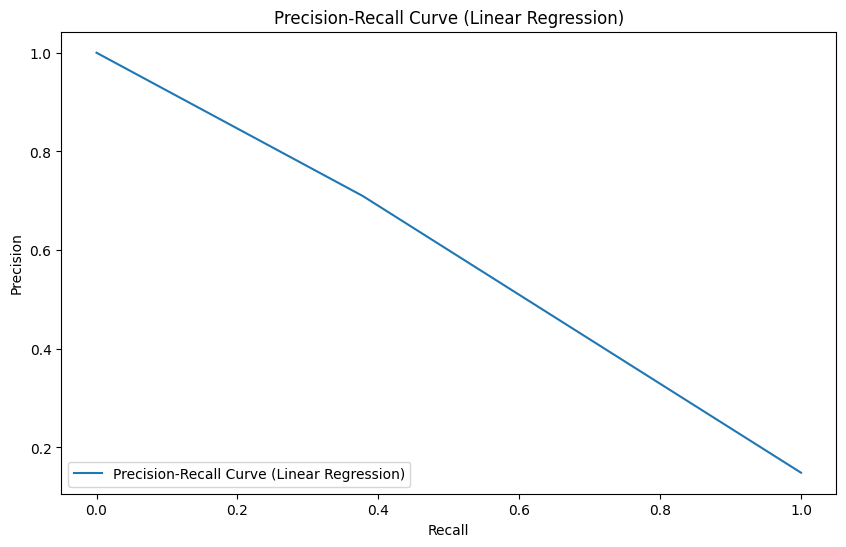
El análisis del informe de clasificación y la matriz de confusión sugirió que el modelo enfrentó dificultades importantes para detectar correctamente los ejemplos positivos, mostrando un sesgo hacia la clase negativa.

Figura 5: Curva ROC Regresión Lineal



La Curva ROC para la regresión lineal mostró un área bajo la curva (AUC) de 0.68, lo cual indicó un rendimiento que superó el azar, pero no alcanzó a ser particularmente robusto. Este resultado reflejó un desempeño moderado. Con 0.68, el modelo evidenció una capacidad limitada para diferenciar efectivamente entre las clases positiva y negativa, lo que puso de manifiesto sus restricciones en cuanto a precisión y eficacia en la clasificación.

Figura 6: Curva Precision-Recall Regresión Lineal



La Curva Precision-Recall reveló el delicado equilibrio entre precisión y recall en diversos umbrales de clasificación. El gráfico mostró una tendencia clara: conforme el recall aumentaba, la precisión experimentaba una disminución considerable. Este comportamiento evidenció que, si bien el modelo logró mejorar su capacidad para identificar casos positivos (aumentando el recall), lo hizo a costa de generar un número significativo de falsos positivos, lo cual impactó negativamente en la precisión.

## 5.3 SVM

El algoritmo SVM demostró un rendimiento notable, logrando una accuracy del 91.19%. Esta cifra indicó que más de nueve de cada diez predicciones realizadas por el modelo fueron acertadas, lo cual representó una mejora significativa respecto a los modelos de regresión previamente analizados.

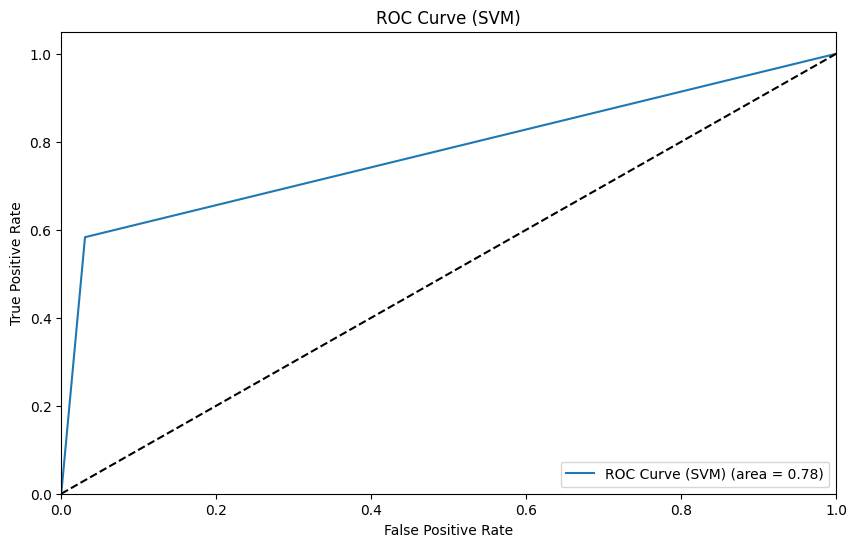
Un examen detallado del informe de clasificación y la matriz de confusión arrojó luz sobre el comportamiento del SVM en distintas facetas. En cuanto a la precisión, el modelo exhibió un desempeño particularmente sólido en la clase negativa, alcanzando un 93% de aciertos en sus predicciones. Para la clase positiva, aunque menos favorable, la precisión del 77% seguía siendo considerable, indicando que más de tres cuartas partes de las observaciones clasificadas como positivas fueron efectivamente correctas.

El análisis del recall reveló una disparidad interesante entre las clases. Para la clase negativa, el modelo demostró que era capaz de identificar correctamente el 97% de los casos. Sin embargo, en la clase positiva, el recall del 58% sugirió que el SVM enfrentó desafíos al intentar detectar todos los casos positivos reales, dejando espacio para mejoras en este aspecto.

El F1-score, métrica que sintetiza precisión y recall, reflejó esta asimetría en el rendimiento del modelo. Con un valor de 0.95 para la clase negativa, el SVM demostró un equilibrio casi óptimo entre precisión y exhaustividad para esta categoría. En contraste, el F1-score de 0.66 para la clase positiva subrayó las dificultades del modelo en mantener un balance igualmente favorable para ambas clases.

En síntesis, el informe de clasificación puso de manifiesto que el SVM logró un equilibrio entre precisión y recall para la clase negativa, mientras que evidenció ciertas limitaciones en su capacidad para detectar y clasificar correctamente los casos de la clase positiva.

Figura 7: Curva ROC SVM

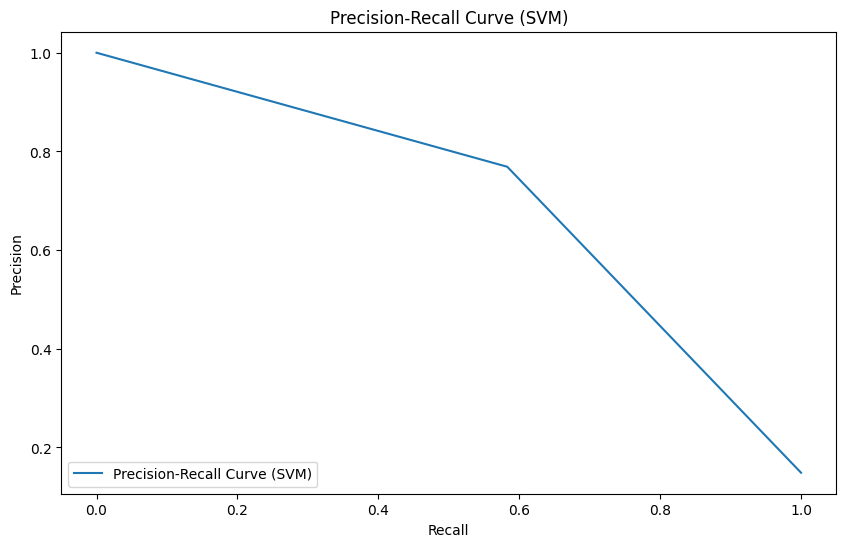


La evaluación de la Curva ROC para el modelo SVM reveló un área bajo la curva (AUC) de 0.78. Este valor demostró que el modelo poseía una capacidad discriminativa considerable entre las clases, superando significativamente el umbral de predicción aleatoria.

El valor de 0.78 posicionó al SVM en un rango de rendimiento bastante favorable. Este resultado sugirió que el modelo fue capaz de capturar patrones relevantes en los datos, permitiéndole distinguir entre las clases con una precisión notable.

Sin embargo, la distancia entre 0.78 y el ideal de 1.0 indicó que aún existía un margen sustancial para mejorar. El SVM demostró un desempeño sólido, pero no alcanzó el nivel de un clasificador óptimo.

Figura 8: Curva Precision-Recall SVM



El análisis de la Curva Precision-Recall para el modelo SVM proporcionó insights valiosos sobre su comportamiento en diferentes umbrales de decisión. Esta curva reveló una relación inversa característica entre la precisión y el recall, ilustrando el delicado equilibrio que el modelo debió manejar.

En los segmentos iniciales de la curva, donde el recall se mantuvo bajo, el modelo exhibió una precisión notable. Esto indicó que, cuando el SVM fue más selectivo en sus predicciones positivas, logró un alto grado de acierto. Sin embargo, a medida que se intentó aumentar el recall para capturar una mayor proporción de casos positivos reales, se observó una caída pronunciada en la precisión.

Esta tendencia descendente en la precisión conforme aumentaba el recall reflejó un desafío fundamental en la tarea de clasificación. El modelo se enfrentó a la disyuntiva de mantener una alta precisión a costa de identificar menos casos positivos, o bien, incrementar la detección de positivos a expensas de generar más falsos positivos.

## 5.4 Gradient Boosting

El algoritmo de Gradient Boosting, logro una precisión global del 90.20%. Este resultado posicionó al modelo como una opción competitiva en el conjunto de técnicas evaluadas, demostrando su capacidad para realizar predicciones acertadas en más de nueve de cada diez casos.

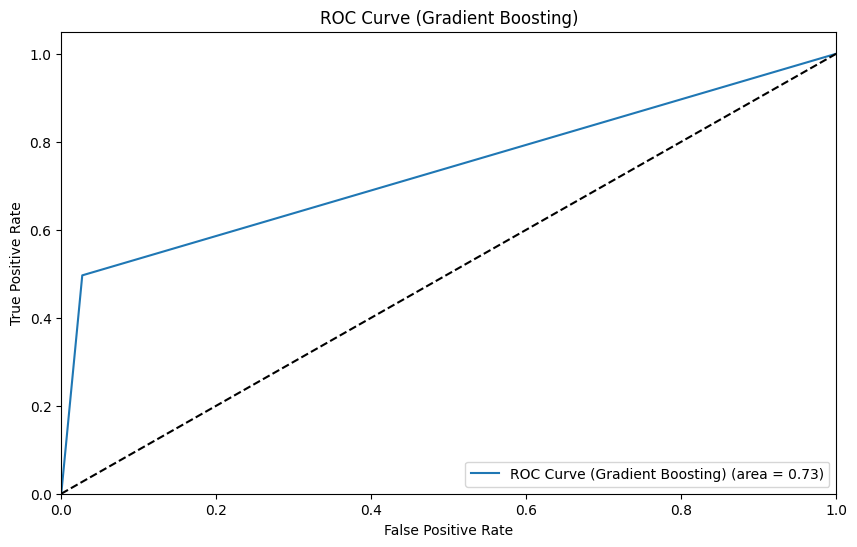
Un análisis minucioso de las métricas de rendimiento reveló matices interesantes en el comportamiento del modelo. La precisión para la clase negativa alcanzó un 92%, evidenciando una identificación correctamente los casos de esta categoría. Por otro lado, la clase positiva, aunque con un desempeño menor, mantuvo una precisión del 76%.

Sin embargo, el examen del recall expuso una marcada asimetría entre las clases. Mientras que para la clase negativa el modelo logró una detección casi perfecta del 97%, la clase positiva presentó un recall del 50%, señalando una deficiencia significativa en la identificación exhaustiva de casos positivos reales.

Esta disparidad se reflejó claramente en los F1-scores: un robusto 0.94 para la clase negativa contrastó con un modesto 0.60 para la positiva, subrayando el desequilibrio en el rendimiento del modelo entre ambas categorías.

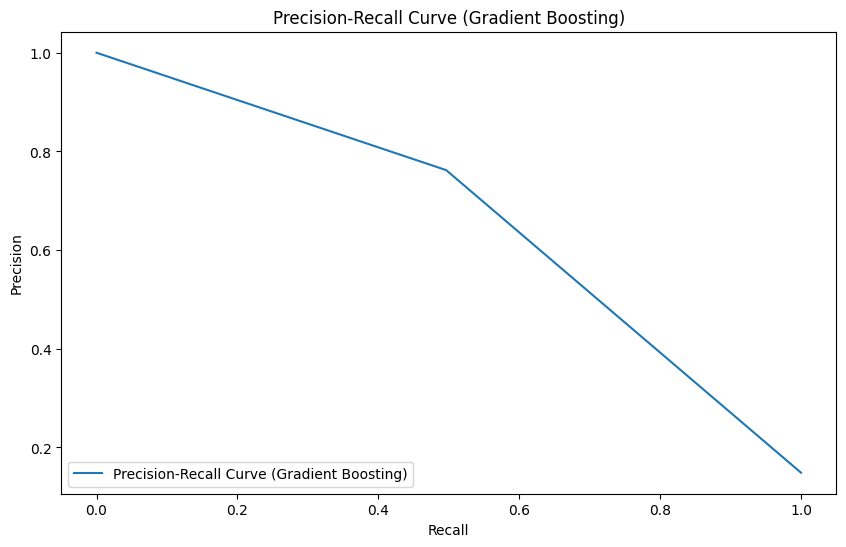
La evaluación integral de estos resultados sugirió un sesgo pronunciado hacia la clase negativa, fenómeno que podría atribuirse tanto a un posible desequilibrio en la distribución de clases del conjunto de datos como a las características intrínsecas del algoritmo de Gradient Boosting, que tiende a favorecer la clase mayoritaria en su proceso de optimización.

Figura 9: Curva ROC Gradient Boosting



La evaluación de la Curva ROC para el modelo de Gradient Boosting reveló un área bajo la curva (AUC) de 0.73. Este valor demostró que el modelo poseía una capacidad discriminativa moderada entre las clases, superando el umbral de predicción aleatoria, pero sin alcanzar niveles óptimos.

Figura 10: Curva Precision-Recall SVM



En las etapas tempranas de la curva, donde el recall se mantuvo en niveles conservadores, el modelo desplegó una precisión impresionante. Esta fase inicial sugirió que el Gradient Boosting, cuando adoptaba un enfoque más selectivo en sus predicciones positivas, lograba un nivel de acierto notable. No obstante, conforme se buscaba expandir la red de detección para abarcar una fracción más amplia de casos positivos reales, se desencadenó un declive acentuado en la precisión.

## 5.5 Random Forest

Tras un exhaustivo proceso de evaluación que abarcó diversos algoritmos de clasificación, emergió un claro vencedor: el Random Forest. Este modelo no solo se destacó por su impresionante precisión, sino que también demostró una notable versatilidad al enfrentar el desafío de identificar tanto las clases predominantes como las menos representadas. La habilidad del Random Forest para navegar la complejidad de nuestros datos, sumada a su destreza en la selección de características relevantes, lo coronó como la elección óptima para nuestra tarea de clasificación.

El Random Forest exhibió un equilibrio envidiable entre precisión y exhaustividad en ambas clases, superando las limitaciones observadas en otros modelos. Su robustez frente a conjuntos de datos complejos y su capacidad para mitigar el sobreajuste lo posicionaron a la vanguardia de nuestras opciones.

En el siguiente apartado, "6. Análisis del Modelo Ganador", nos sumergiremos en una exploración detallada del rendimiento del Random Forest. Desentrañaremos los factores clave de su éxito y examinaremos cómo sus características únicas se alinean con las particularidades de nuestro conjunto de datos.

# 6. Análisis del Modelo Ganador

El modelo de Random Forest demostró un rendimiento excepcional, alcanzando una precisión global del 93.33%. Este resultado sobresaliente posicionó al algoritmo como la opción más efectiva entre todas las técnicas evaluadas, evidenciando su capacidad para realizar predicciones acertadas en más de nueve de cada diez casos.

Un análisis pormenorizado de las métricas de rendimiento reveló matices fascinantes en el comportamiento del modelo:

La precisión para la clase negativa alcanzó un impresionante 97%, demostrando una habilidad casi perfecta para identificar correctamente los casos de esta categoría. Por su parte, la clase positiva, aunque con un desempeño ligeramente menos deslumbrante, mantuvo una precisión respetable del 76%, indicando que más de tres cuartas partes de las predicciones positivas fueron acertadas.

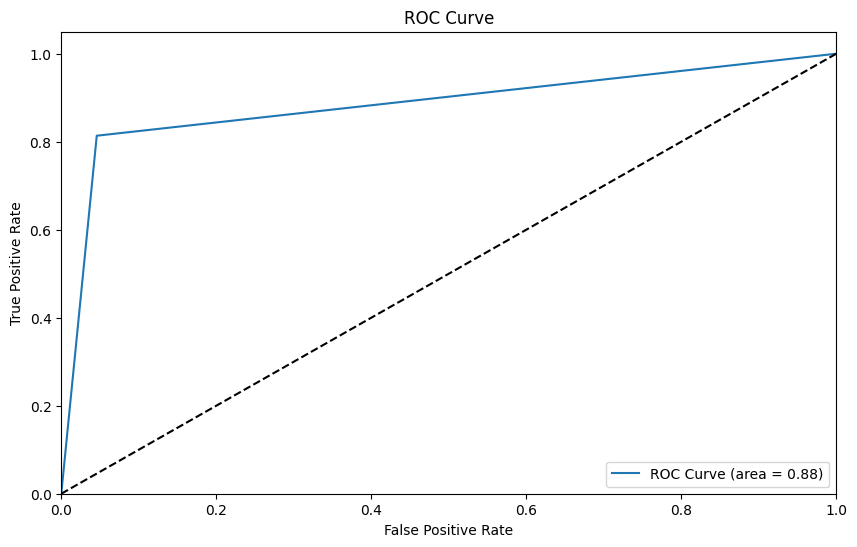
El examen del recall expuso un equilibrio notable entre las clases, un logro significativo en comparación con los modelos anteriores. Para la clase negativa, el modelo logró una detección sobresaliente del 95%, mientras que para la clase positiva alcanzó un robusto 81%. Esta simetría en el recall sugiere que el Random Forest superó con éxito el desafío de identificar exhaustivamente casos de ambas categorías.

Los F1-scores reflejaron claramente este rendimiento balanceado: un excepcional 0.96 para la clase negativa se complementó con un sólido 0.78 para la positiva. Estos valores subrayan la capacidad del modelo para mantener un equilibrio óptimo entre precisión y exhaustividad en ambas clases.

La evaluación integral de estos resultados sugiere que el Random Forest logró mitigar eficazmente el sesgo hacia la clase mayoritaria, un problema común en tareas de clasificación con clases desbalanceadas. Esta habilidad para manejar el desequilibrio de clases, combinada con su alta precisión general, posiciona al Random Forest como una herramienta excepcionalmente adecuada para el problema en cuestión.

Estos hallazgos subrayan la robustez y versatilidad del Random Forest, demostrando su capacidad para capturar patrones complejos en los datos y realizar predicciones precisas en ambas clases.

Figura 11: Curva ROC Random Forest



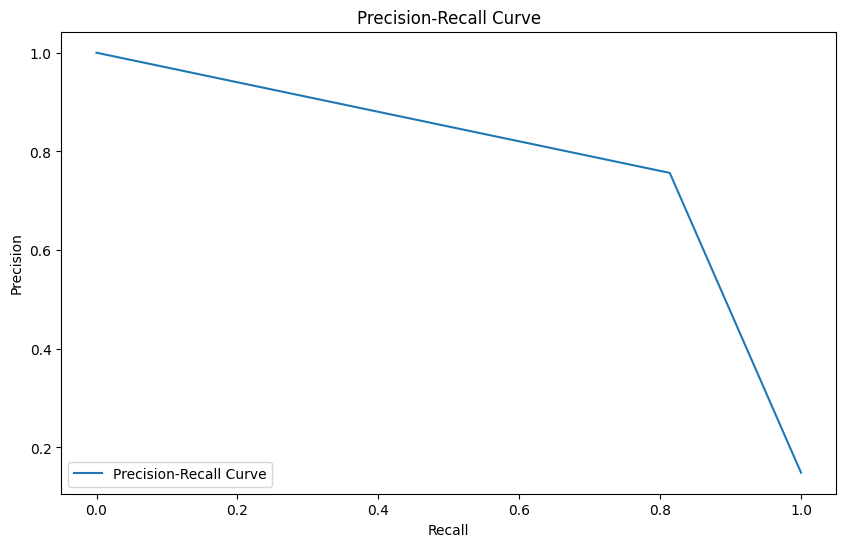
La evaluación de la Curva ROC para el modelo de Random Forest reveló un resultado extraordinario, con un área bajo la curva (AUC) de 0.88. Este valor, que se acerca notablemente al ideal de 1.0, pone de manifiesto la excepcional capacidad discriminativa del modelo entre las clases.

Este resultado no solo supera significativamente el umbral de predicción aleatoria, sino que también eclipsa el desempeño de los modelos previamente analizados.

La proximidad del AUC a 0.9 sugiere que el Random Forest ha logrado capturar patrones sutiles y complejos en los datos, permitiéndole distinguir entre las clases con una precisión remarcable.

El valor de 0.88 no solo refleja una alta precisión general, sino que también indica una robustez considerable del modelo frente a posibles desbalances en las clases. Esta característica es crucial, ya que sugiere que el Random Forest mantiene su eficacia tanto en la identificación de la clase mayoritaria como en la detección de la clase minoritaria, un equilibrio difícil de lograr en muchos problemas de clasificación.

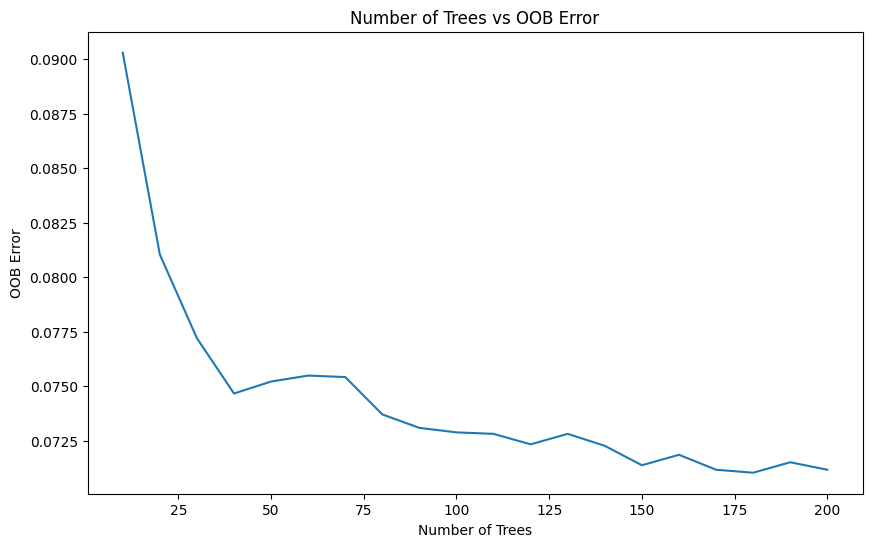
Figura 12: Curva Precision-Recall Random Forest



El análisis de la Curva Precision-Recall del modelo Random Forest mostró una tendencia característica: conforme el recall se incrementaba, se observaba una disminución gradual en la precisión. Este fenómeno, lejos de ser una debilidad, refleja la naturaleza intrínseca de los modelos de clasificación, donde el esfuerzo por identificar una mayor proporción de casos positivos inevitablemente conlleva un aumento en los falsos positivos.

Lo verdaderamente destacable de esta curva fue la capacidad del Random Forest para mantener niveles de precisión notablemente altos incluso ante incrementos sustanciales del recall. Esta resiliencia en el rendimiento subraya la robustez del modelo, demostrando su habilidad para navegar eficazmente el delicado equilibrio entre la identificación exhaustiva de casos positivos y la minimización de falsas alarmas.

Figura 13: Curva Number of Trees vs OOB Error



La evaluación del Out-of-Bag (OOB) Error en función del número de árboles en el modelo Random Forest proporcionó conclusiones cruciales sobre su comportamiento y optimización.

La gráfica reveló una tendencia, conforme se incrementaba el número de árboles, se observaba una disminución pronunciada en el error OOB. Este patrón ilustra vívidamente cómo la diversidad y la redundancia inherentes al ensemble contribuyen a la robustez del modelo. La curva de aprendizaje mostró una inflexión notable alrededor de los 100-150 árboles, punto en el cual el error comenzó a estabilizarse.

Este punto de inflexión es particularmente revelador, ya que sugiere que la configuración de 200 árboles elegida para el modelo final no solo es adecuada, sino que representa un equilibrio óptimo entre rendimiento y eficiencia computacional. La estabilización del error indica que añadir más árboles más allá de este punto ofrece rendimientos decrecientes, evitando así el riesgo de sobreajuste sin sacrificar capacidad predictiva.

El OOB Score de 0.9288, traducido a un error de apenas 7.12% en datos no vistos durante el entrenamiento, subraya, la excepcional capacidad del Random Forest para generalizar sus predicciones a nuevos datos.

Esta métrica no solo valida la elección del Random Forest como el modelo óptimo para esta tarea, sino que también ofrece una garantía sólida de su desempeño futuro. La consistencia entre el OOB Score y la accuracy general del modelo (93.33%) refuerza aún más la confianza en su capacidad predictiva, sugiriendo que su rendimiento se mantendrá estable incluso cuando se enfrente a datos completamente nuevos.

En conjunto, el análisis del OOB Error y Score proporciona una visión profunda de la arquitectura interna del modelo, confirmando su robustez y justificando su selección como la herramienta más adecuada para abordar este caso de uso.

# 7. Conclusiones

El proyecto de desarrollo de un modelo de detección de polución en la planta de Roquette en Valencia ha demostrado la eficacia de las técnicas de machine learning para abordar problemas industriales complejos. Las principales conclusiones son:

1. **Rendimiento del modelo:** El algoritmo Random Forest emergió como la solución más efectiva, logrando una precisión general del 93.33%. Este modelo destacó por su equilibrio entre la detección de casos de polución y la correcta identificación de máquinas no contaminantes.
2. **Comparativa de algoritmos**: Se implementaron y evaluaron diversos algoritmos de clasificación, incluyendo Regresión Logística, SVM y Gradient Boosting. El análisis comparativo fue crucial para seleccionar el modelo más adecuado.
3. **Arquitectura de datos:** La implementación de un flujo de datos automatizado, desde la ingesta de archivos CSV hasta el procesamiento en Google Cloud, demostró ser eficiente y escalable. Las técnicas de preprocesamiento, como la despivotación de columnas y el manejo de datos faltantes, fueron fundamentales para la calidad del dataset.
4. **Metodología ágil:** El enfoque ágil en la gestión del proyecto, junto con la comunicación constante con el cliente y la visita in situ, permitió desarrollar una solución altamente adaptada a las necesidades reales de la planta.
5. **Impacto y aplicabilidad:** El modelo no solo proporciona una herramienta precisa para la detección de polución, sino que también demuestra el potencial del machine learning para optimizar procesos industriales, promoviendo una producción más sostenible y eficiente.
6. **Perspectivas futuras:** Los resultados sugieren que este enfoque podría extenderse a otras plantas y procesos productivos, ofreciendo beneficios significativos en la reducción del impacto ambiental y la optimización operativa en la industria.

Este proyecto ilustra cómo la aplicación del machine learning puede transformar la gestión ambiental en entornos industriales, proporcionando herramientas predictivas precisas que permiten una toma de decisiones más informada y proactiva.