MANTENIMIENTO PREDICTIVO BASADO EN ANÁLISIS QUÍMICO DE ACEITES DE MOTORES

Aprendizaje automático y sus aplicaciones en la industria 4.0

Federico Krell, TECHINT E&C, fkrell@techint.com

Palabras claves

Mantenimiento predictivo, análisis químico, motores, industria 4.0, técnicas estadísticas, aprendizaje automático.

Resumen

Este estudio se centra en el desarrollo de técnicas de mantenimiento predictivo basadas en el análisis químico de los aceites de los motores. Se utilizan métodos estadísticos y de aprendizaje automático para identificar el estado de los motores en una clasificación binaria. Se recolectaron y clasificaron muestras químicas de los aceites de motores en la industria minera, obtenidas de una base de datos pública de la University of Western Australia. El objetivo principal fue evaluar la eficacia de diferentes modelos en la toma de decisiones informadas y la optimización de los planes de mantenimiento.

Los resultados obtenidos destacan la utilidad del enfoque de mantenimiento predictivo basado en el análisis químico de los aceites de los motores. Se observó que los modelos estadísticos y de aprendizaje automático pueden proporcionar información valiosa para la gestión del mantenimiento. Se encontró una correlación significativa entre ciertos atributos químicos y el estado de los motores, lo que puede servir como una señal de alerta temprana de posibles problemas.

Se destacó la importancia de definir métricas de evaluación adecuadas para los modelos, considerando la capacidad de ahorro de recursos y la optimización de los planes de mantenimiento. Aunque se encontraron diferencias en la relevancia de los atributos entre los métodos estadísticos y el modelo de aprendizaje automático, se concluyó que ambos enfoques pueden ser útiles en la detección de problemas y la toma de decisiones informadas.

Este estudio contribuye al campo del mantenimiento predictivo al explorar la aplicación del análisis químico de los aceites de los motores. Los resultados obtenidos pueden ayudar a las empresas a optimizar sus estrategias de mantenimiento, reducir costos y mejorar la eficiencia en la gestión de los equipos. Además, este enfoque tiene un amplio potencial de aplicación en el desarrollo de nuevas industrias, ya que proporciona una base sólida para la implementación de técnicas de mantenimiento predictivo en diversos sectores.

En resumen, este estudio demuestra el potencial de utilizar técnicas estadísticas y de aprendizaje automático en el mantenimiento predictivo basado en el análisis químico de los aceites de los motores. Los resultados obtenidos ofrecen perspectivas útiles para la toma de decisiones informadas en la gestión del mantenimiento. Se sugiere continuar investigando y refinando estos enfoques para mejorar aún más la eficacia del mantenimiento predictivo en el desarrollo de nuevas industrias.

Abstract

This study focuses on the development of predictive maintenance techniques based on the chemical analysis of engine oils. Statistical methods and machine learning are used to identify the condition of engines in a binary classification. Chemical samples of engine oils were collected and classified in

the mining industry, obtained from a public database at the University of Western Australia. The main objective was to evaluate the effectiveness of different models in making informed decisions and optimizing maintenance plans.

The results highlight the usefulness of the predictive maintenance approach based on the chemical analysis of engine oils. It was observed that statistical and machine learning models can provide valuable information for maintenance management. A significant correlation was found between certain chemical attributes and the engine condition, which can serve as an early warning signal for potential issues.

The importance of defining appropriate evaluation metrics for the models was emphasized, considering the resource-saving capacity and optimization of maintenance plans. Although differences in attribute relevance were found between the statistical methods and the machine learning model, it was concluded that both approaches can be useful in problem detection and informed decision-making.

This study contributes to the field of predictive maintenance by exploring the application of chemical analysis of engine oils. The obtained results can help companies optimize their maintenance strategies, reduce costs, and improve equipment management efficiency. Furthermore, this approach has broad potential application in the development of new industries, as it provides a solid foundation for implementing predictive maintenance techniques across various sectors.

In summary, this study demonstrates the potential of using statistical and machine learning techniques in predictive maintenance based on the chemical analysis of engine oils. The obtained results offer valuable insights for informed decision-making in maintenance management. Further research and refinement of these approaches are suggested to further enhance the effectiveness of predictive maintenance in the development of new industries.

Introducción

El mantenimiento predictivo basado en el análisis químico de los aceites de los motores ha surgido como una estrategia prometedora en la gestión del mantenimiento de los equipos industriales. Como parte de la actualización a Industrias 4.0, los métodos basados en estadística y ciencia de datos pueden ser de utilidad para incorporarse a los procesos ya existentes de mantenimiento preventivo como un primer filtro que permita bajar costos optimizando la cantidad de equipos que deben ser llevados a mantenimiento.
El objetivo de este trabajo es utilizar diferentes técnicas estadísticas y de aprendizaje automático para poder identificar en una clase binaria el estado de los motores. Finalmente, se busca explorar las capacidades de los diferentes modelos para poder tomar decisiones informadas.
Como base de datos se utilizaron muestras químicas recolectadas y clasificadas de los aceites de motores en la industria minera. Estos muestreos se realizan habitualmente para evaluar el requerimiento, o no, de algún tipo de mantenimiento. Los datos se obtuvieron de una base pública de la University of Western Australia, y fueron utilizados con anterioridad en el desarrollo del trabajo "Classifying machinery condition using oil samples and binary logistic regression" (Philips et al., 2015)
La estructura de este trabajo comprende, en principio, una sección en la que se presenta el marco teórico que se utiliza, antecedentes de investigación y el estado actual de este tipo de desarrollos de software, para brindar la base conceptual necesaria para el alcance de este trabajo. Luego se procede a describir la metodología, el origen de los datos y un análisis exploratorio inicial. Por último, se presentan los resultados del trabajo, el análisis que conllevan los mismos, y las conclusiones obtenidas.

Marco Teórico

El mantenimiento predictivo es una etapa de la gestión de equipos industriales que está en desarrollo e investigación actualmente. La implementación de este tipo de herramienta podría eliminar el factor de riesgo y el costo de mantenimiento, permitiendo así optimizar los planes de mantenimiento, eliminando los procedimientos innecesarios y generando nuevos recursos en la toma de decisiones de los sectores de confiabilidad de las industrias (Sircar et al., 2021).
Varios de los métodos a utilizar requieren grandes volúmenes de datos. Para justificar esta inversión, es necesario realizar un estudio previo sobre las capacidades que tendrían los métodos de inteligencia artificial para ahorrar costos.
Por otro lado, la necesidad de generar métricas adecuadas para evaluar los modelos es crítica para poder realizar un entrenamiento adecuado.
Parte de esto se discute en el artículo "Classifying machinery condition using oil samples and binary logistic regression " (Philips et al., 2015), en el cual se menciona la importancia de los falsos negativos al implementar un modelo de clasificación automática. La no clasificación de un equipo en mal estado puede llevar a fallas críticas, mientras que la no clasificación de un equipo en buen estado solo llevaría a un mantenimiento evitable. Es importante tener en cuenta esta diferencia al evaluar el modelo.
El mantenimiento predictivo es una opción que se está desarrollando en lo que se conoce como la Industria 4.0, la cual incorpora capacidades tecnológicas y la ciencia de datos en la toma de decisiones. Existen ejemplos de investigaciones en el área automotriz, como el artículo "Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry" (Theissler, 2020), y ejemplos de integración en sistemas existentes, como el artículo "Integrating machine learning techniques into optimal maintenance scheduling" (Yeardley, 2018). Estos estudios resaltan la importancia de los regímenes de mantenimiento adecuados para minimizar los tiempos muertos y contribuir a la minimización de defectos y accidentes. Es crucial aplicar los nuevos métodos digitales de la Industria 4.0 para obtener estrategias de mantenimiento inteligentes y eficientes.
En este contexto, utilizar la base de datos de aceites de motores es una forma de ejemplificar las capacidades de diferentes métodos para realizar un mantenimiento predictivo y la posible dependencia de los costos en la optimización adecuada.

Metodología

En esta sección de describen los datos utilizados, su origen y un análisis exploratorio, así como las técnicas a utilizar y de qué forma se utilizan.

Datos Utilizados

La base de datos utilizada en este trabajo se obtuvo del artículo de investigación titulado "Classifying machinery condition using oil samples and binary logistic regression" (Philips et al., 2015) de la University of Western Australia, el cual es de acceso público. Los datos utilizados en este trabajo fueron obtenidos de la siguiente biblioteca en línea:
https://prognosticsdl.systemhealthlab.com/dataset/oil-analysis-on-diesel-engines.
Sin embargo, para acceder a los datos originales, se utilizó el repositorio público en GitHub titulado "Maintenance4Ind4.0" de la autora original del paper M.R. Hodkiewicz, disponible en la siguiente dirección:
https://github.com/mhodki/Maintenance4Ind4.0/tree/main/data.
Este repositorio contiene datos de aceites de motores de equipos utilizados en la industria minera y fue descargado en febrero del 2023. Los datos incluyen características químicas de los aceites de los motores, así como información sobre si los motores fueron enviados a mantenimiento o no, lo cual se utilizará como variable objetivo para la clasificación binaria en este trabajo de especialista. El

respaldo del artículo de investigación y el uso de datos de un repositorio público garantizan la calidad y representatividad del dataset utilizado en este estudio.
El análisis exploratorio inicial muestra un dataset que tiene una llave o id para cada muestra, un unitno que muestra diferentes unidades analizadas, un modelid que identifica entre 3 modelos y un evalcode que muestra con diferentes valores si el motor es o no apto. La columna evalcode se simplifica en la columna asample con un valor binario 1 y 0 siendo 1 el motor apto evaluado como A y 0 el motor no apto evaluado distinto de A.
La base de datos no tiene datos faltantes ni nulos, por ser una base de datos utilizada en un ámbito académico puede que haya sido previamente curada para su utilización.
Las demás columnas del conjunto de datos son numéricas y contienen información sobre cómo fue evaluado previamente, si fue evaluado A en la última revisión o no, si se cambió el aceite desde la última revisión o no, la cantidad de horas que lleva ese aceite en el motor y varios análisis químicos.
Para este trabajo se utilizan todas las columnas numéricas y las generadas binarias como numéricas.

Análisis exploratorio inicial

La cantidad de motores en buen estado es de 211 y de motores que deben ser mandados a mantenimiento es de 1121. Debe tenerse en cuenta que hay diferentes grados de falla dentro de los motores no aptos y no toda falla es fatal. Esta observación es importante en el momento de identificar de forma binaria ya que de no identificar un motor en mal estado puede que el mismo no tenga una falla grave y su no identificación no tenga la misma gravedad que la de otros motores en peor estado.
Hay 611 cuyo código es B, 429 con código C y 81 con código X. Al binarizar para este trabajo incluyendo al B, C y X en la misma categoría se pierde la diferencia de gravedad. En trabajos futuros puede ser de utilidad no hacer la binarización.
En un primer análisis se analiza la correlación de las variables con asample obteniendo valores de correlación absolutos máximos de 0.4. En general no se tiene una correlación estadística significativa.
Se estudia la media para algunas de las variables que en el paper original se marcaron como importantes agrupadas por código evaluado.

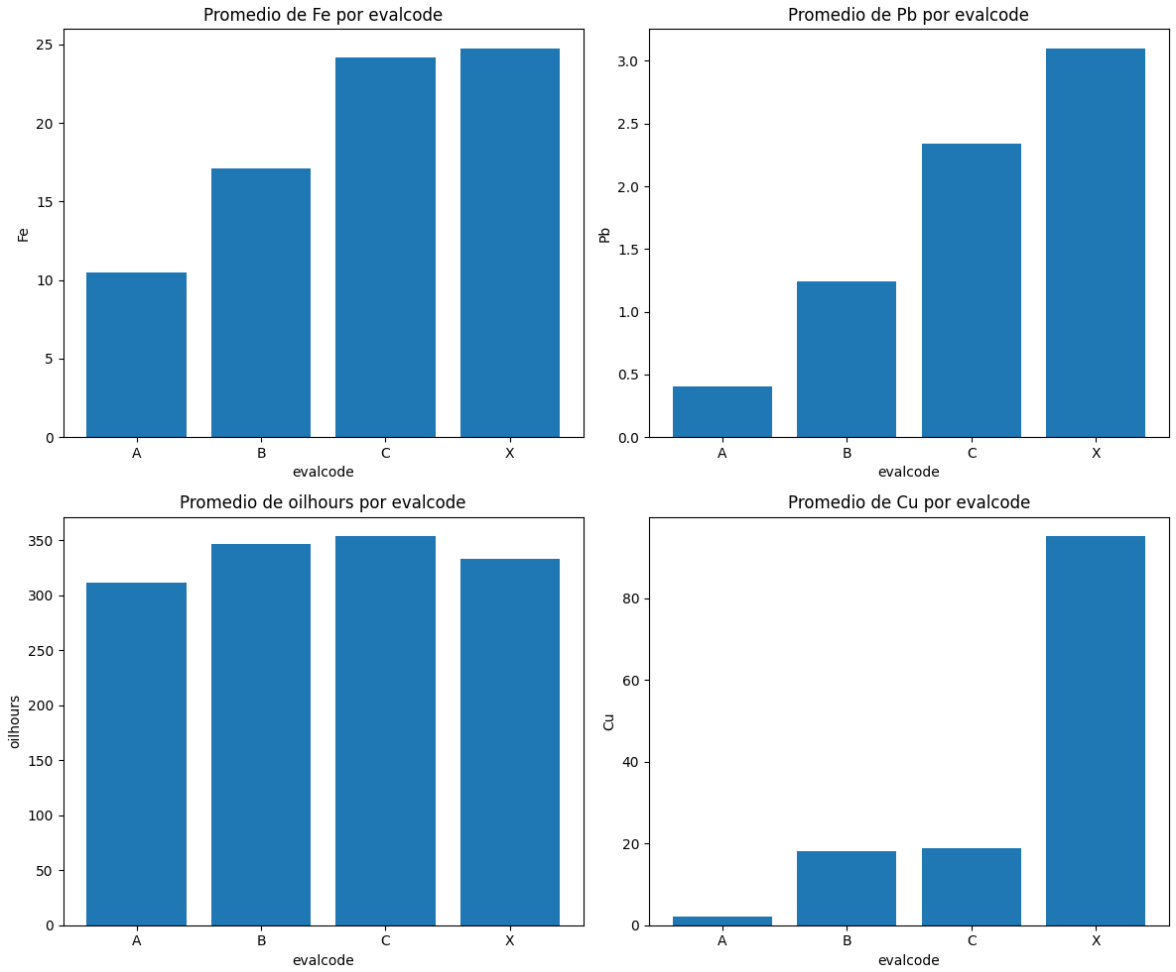

Figura 2 Valores medios de Hierro, Plomo, cantidad de horas y cobre por cada tipo de clasificación

Las cuatro variables son las que el paper identifica con menores p-valor. Puede verse que hay un salto en general entre A y no A en todas menos OilHours. En el desarrollo de este trabajo se tendrá en cuenta la capacidad de los modelos de determinar las importancias de variable y podrá compararse con los resultados del paper de la universidad de Australia.

Por último, se toma en cuenta la idea del paper y la columna de resultados previos se transforma en 4 columnas binarias prev"n" llamadas preva, prevb, prevc y prevx.

Metodología especifica:

Métricas
Se definen dos métricas para comparación.
Accuracy: Es la métrica mas intuitiva en este tipo de análisis, le asigna un punto a cada valor identificado correctmente y no toma en cuenta los falsos negativos ni falsos positivos.
Metrica personalizada: Esta metrica toma una relación 2 a 1 en cuanto a la no identificación de la variable objetivo. De esta forma el score asigna una unidad arbitraria a identificar correctamente un motor en buen estado y descontara uno al no identificarlo. Por otro lado, se asignan dos unidades a la

identificación de motor en mal estado y se restan dos a no identificar. Dando una penalización extra a la no identificación de un motor en mal estado.
Esta métrica se crea para dar valor especial a la identificación de los motores en mal estado y ser coherente con el costo asociado a la no identificación de un motor en mal estado. La relación 2 a 1 es arbitraria ya que cada problema de mantenimiento predictivo puede tener una relación distinta.
Como ejemplo ilustrativo la matriz de confusión de la regresión lineal es:

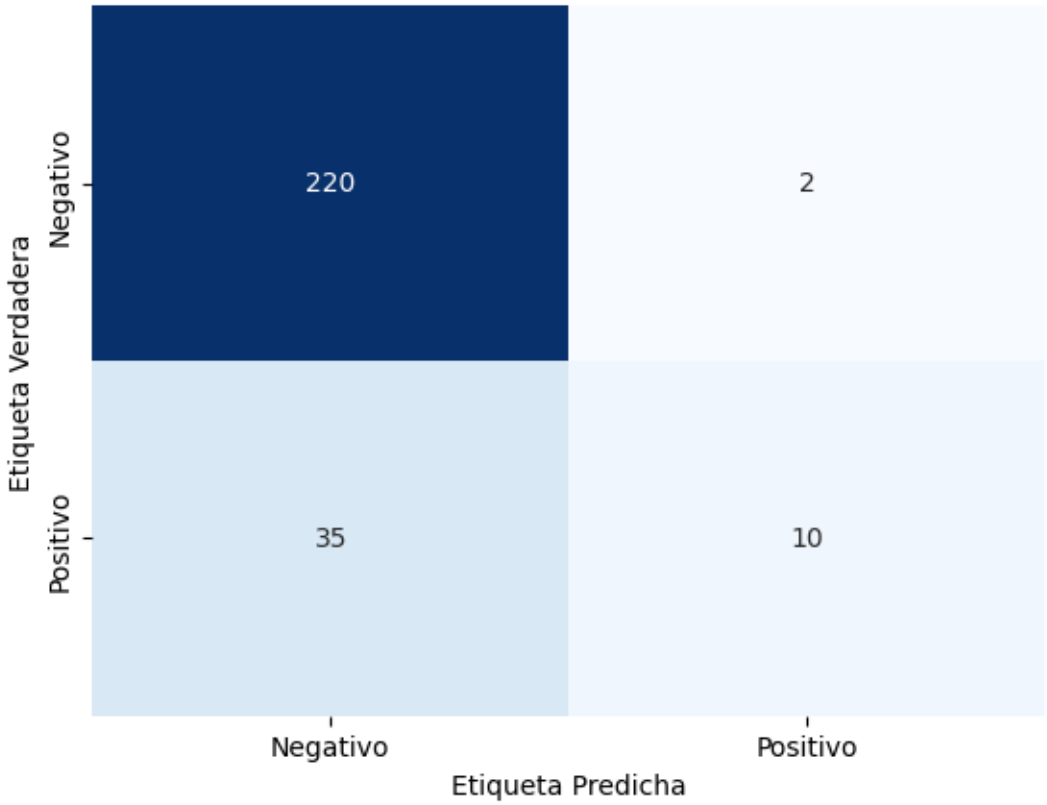

Figura 6 Matriz de confusión de regresión Lineal

Y el puntaje obtenido es 411.
El objetivo es tener una forma de comparar los métodos estadísticos con lightgbm. Siendo lightgbm optimizado para las dos métricas obteniendo dos modelos distintos y comprobando la importancia de definir los valores asignados a cada espacio de la matriz de confusión.

División de entrenamiento y validación
En nuestro enfoque de modelado en el campo del mantenimiento predictivo, es común dividir nuestros datos en conjuntos de entrenamiento y validación. Esta separación permite evaluar y ajustar nuestros modelos de manera objetiva. Por lo general, utilizamos una proporción de 80/20, donde el 80% de los datos se asigna al conjunto de entrenamiento y el 20% restante se reserva para la validación.
La razón detrás de esta separación radica en garantizar la eficacia y la generalización de nuestros modelos. Al utilizar el conjunto de entrenamiento para entrenar el modelo, podemos ajustar sus parámetros y encontrar la mejor configuración para lograr un rendimiento óptimo. Luego, al evaluar el modelo en el conjunto de validación, podemos verificar su capacidad para generalizar y hacer predicciones precisas en datos no vistos.
La elección de la proporción 80/20 se basa en un equilibrio entre tener suficientes datos para entrenar el modelo de manera efectiva y mantener una muestra significativa para evaluar su rendimiento. Al utilizar el 80% de los datos para el entrenamiento, se aprovecha una gran cantidad de información y

patrones en los datos. El 20% restante se utiliza como conjunto de validación independiente para obtener una evaluación imparcial y confiable del rendimiento del modelo en situaciones del mundo real. Esta metodología nos permite desarrollar modelos sólidos y confiables en el campo del mantenimiento predictivo.

Métodos estadísticos utilizados

Regresión lineal:
Regresión lineal es una técnica estadística fundamental que permite construir un modelo simple sin parámetros de ajuste complejos. Esta técnica es ampliamente utilizada debido a su capacidad para calcularse rápidamente y proporcionar una primera aproximación en el análisis de datos.
En este estudio de mantenimiento predictivo, se aplica la regresión lineal para abordar la predicción de una variable binaria. Sin embargo, para adaptar este modelo a un problema de clasificación, es necesario realizar una binarización de las predicciones, de manera de poder asignar etiquetas de clase a cada instancia.
La regresión lineal también proporciona información sobre la correlación entre las variables predictoras y la variable objetivo. Esto se puede obtener mediante el análisis de los p-valores asociados a cada variable. Los p-valores nos indican la probabilidad de obtener un resultado tan extremo como el observado, asumiendo que no hay una relación significativa entre la variable predictor y la variable objetivo. De esta manera, los p-valores nos permiten evaluar la relevancia de cada característica en la predicción de la variable binaria.
Al aplicar la regresión lineal, se obtienen no solo los coeficientes que representan las ponderaciones de cada variable en el modelo, sino también los p-valores asociados a cada uno. Estos p-valores brindan una medida de la significancia estadística de cada característica, lo que ayuda a comprender su influencia en la predicción de la variable binaria.
En resumen, la regresión lineal es una técnica básica y rápida que puede servir como punto de partida en el análisis de datos para la predicción de una variable binaria en el mantenimiento predictivo. Al binarizar las predicciones y analizar los p-valores, se obtiene información sobre la correlación y la relevancia de las características utilizadas en el modelo.

Elastic net
Elastic Net es una técnica que combina los conceptos de regresión lineal, regresión de Ridge y regresión de Lasso, lo que resulta en un enfoque más robusto para la selección de características y la regularización del modelo. Esta técnica se utiliza en este análisis de modelos de mantenimiento predictivo para predecir una variable binaria. En este estudio, se busca explorar la eficacia de un modelo más complejo y menos común en problemas binarios, como es el caso del Elastic Net. Para ello, se aplica un tratamiento de binarización a los datos y ajustamos los valores de los parámetros l1 y alpha, que controlan la cantidad de regularización L1 y L2 aplicada en el modelo, respectivamente.
Al variar los valores de alpha y l1_ratio, se generan múltiples modelos de regresión Elastic Net y se evalúa su desempeño utilizando matrices de confusión y puntajes personalizados. Las matrices de confusión permiten visualizar la precisión del modelo en la clasificación de las etiquetas verdaderas y predichas, mientras que los puntajes personalizados brindan una medida específica de la efectividad del modelo en relación con nuestras necesidades.
Además de la evaluación del desempeño, el Elastic Net también proporciona información valiosa sobre la importancia relativa de cada característica utilizada en el modelo. Los coeficientes lambda obtenidos del método pueden considerarse como una medida análoga a la importancia de cada característica, lo que ayuda a comprender mejor la contribución de cada una en la predicción de la variable binaria.

En resumen, al emplear el método Elastic Net en nuestro análisis, se busca obtener un modelo más robusto y explorar su eficacia en la predicción de la variable binaria en el contexto del mantenimiento

predictivo. A través de la optimización de los parámetros y la evaluación exhaustiva del desempeño, podemos obtener una comprensión más profunda de las características relevantes y mejorar la capacidad predictiva de nuestro modelo.

Regresión logística

La regresión logística es una técnica ampliamente utilizada en problemas de clasificación binaria, como el que abordamos en nuestro estudio de mantenimiento predictivo. Esta técnica se basa en el paper "Classifying machinery condition using oil samples and binary logistic regression", cuyo dataset utilizamos en nuestro trabajo de análisis.

En este análisis, se aplica la regresión logística regularizada para predecir si los motores deben ser enviados a mantenimiento o no. Se utilizan diferentes valores de alpha, que controlan la complejidad del modelo, para comparar los resultados de las matrices de confusión.

Después de ajustar el modelo de regresión logística regularizada, se realizan predicciones en el conjunto de validación. Estas predicciones se convierten en clasificaciones binarias utilizando un umbral de 0.5. Luego, se calculan y muestran las matrices de confusión para evaluar el desempeño del modelo.

Las matrices de confusión permiten visualizar la precisión del modelo en la clasificación de las etiquetas verdaderas y predichas.

Además de las matrices de confusión, se calcula una puntuación personalizada para evaluar el desempeño del modelo en relación con nuestros criterios específicos. Esta puntuación personalizada se basa en una función definida por nosotros y nos permite cuantificar la efectividad del modelo en la clasificación de las instancias.

En resumen, la regresión logística regularizada es una técnica fundamental en nuestro análisis de mantenimiento predictivo. Se utilizan diferentes valores de alpha para ajustar el modelo y comparar las matrices de confusión. A través de esta técnica, se busca predecir si los motores deben ser enviados a mantenimiento o no, y evaluar el desempeño del modelo mediante medidas personalizadas.

LightGBM

LightGBM es una técnica de gradient boosting altamente adaptable que permite mejorar la precisión en el análisis de mantenimiento predictivo. En este enfoque, se utiliza la optimización bayesiana para ajustar los hiperparámetros del modelo LightGBM según dos tipos de puntajes distintos: la exactitud (accuracy) y un puntaje personalizado. Esta capacidad de adaptación permite encontrar el equilibrio óptimo entre los parámetros del modelo y los objetivos específicos del análisis.

Para ajustar el modelo de manera precisa y eficiente, se utiliza un conjunto exploratorio de hiperparámetros. Este conjunto incluye variables como:

- n_estimators: Rango de valores de 1 a 1000.
- num_leaves: Rango de valores de 2 a 1000.
- learning_rate: Rango de valores de 0.001 a 0.3, utilizando una distribución logarítmica.
- min_data_in_leaf: Rango de valores de 10 a 10000.
- feature_fraction: Rango de valores de 1e-3 a 1.0, utilizando una distribución logarítmica.

Estas variables representan diferentes aspectos del modelo LightGBM que pueden influir en su rendimiento y capacidad predictiva.

Una vez obtenidos los mejores parámetros de la optimización bayesiana, se realizan entrenamientos y evaluaciones de modelos con 50 semillas diferentes utilizando los parámetros optimizados. Cada modelo se entrena con una semilla diferente para garantizar la variabilidad en las configuraciones del modelo. A continuación, se obtiene la salida del vector de probabilidades de cada modelo para el conjunto de validación.

La combinación de las 50 semillas se realiza antes de la predicción final utilizando el vector de probabilidades obtenido de cada modelo. En lugar de realizar el corte y luego promediar, se promedian los 50 vectores de probabilidades antes de aplicar el umbral de corte. Esto permite obtener un vector de probabilidades final con menos incertidumbre al hacer el corte que el que se hubiera obtenido utilizando el enfoque tradicional de promediar después del corte.
Finalmente, se aplica el umbral de corte al vector de probabilidades promediado para obtener las predicciones del nuevo modelo. Estas predicciones se utilizan para evaluar el rendimiento del modelo promedio en el conjunto de validación. Además, se puede visualizar el desempeño del modelo promedio utilizando una matriz de confusión.
En resumen, LightGBM demuestra su capacidad de adaptación a través de la optimización bayesiana y el uso de múltiples semillas. La combinación de las 50 semillas se realiza previo a la predicción final, promediando los vectores de probabilidades de cada modelo. Esto permite obtener un modelo final más robusto y confiable, con menos incertidumbre en el proceso de corte. Con esta técnica, se mejora la precisión y la calidad de las predicciones en el análisis de mantenimiento predictivo.

Resultados

En esta sección se analizarán los resultados de cada método utilizado. Se tiene en cuenta el resultado de su matriz de confusión, así como las diferentes formas de evaluar intrínsecas al método. La matriz de confusión se calcula en base a la predicción de un conjunto del 20% previamente separado para poder ser identificado sin ser utilizado en el entrenamiento.

Regresión lineal

Se utiliza LinearRegression de scikit learn y OLS de statsmodels para la búsqueda de p-valores.
La binarización se hace en el límite de 0.5 y se obtiene en el set de entrenamiento un score de 411 con la matriz de confusión siendo:

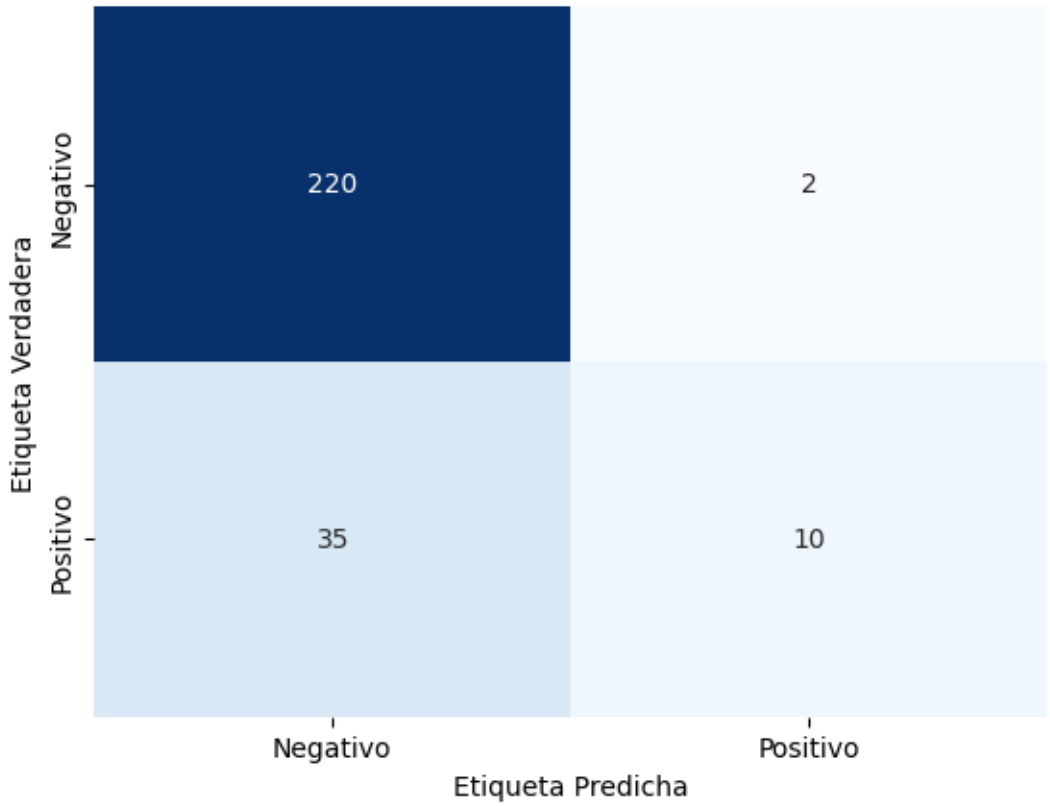

Figura 7 Matriz de confusión de regresión Lineal

El r2 es de 0.19 lo cual es bajo en general, pero la capacidad de identificar es relativamente buena teniendo solo 2 motores en mal estado no identificados y 10 que no requieren mantenimiento identificados.

Los p-valores obtenidos menores a 0.05 son 3, Na, Mo, y ST sin embargo los valores con mayor influencia en la regresión son los preva, prevc, prevx y prevb. Estos resultados son consistentes con la regresión logística del trabajo original.

Regresión elastic net

Se utiliza el mismo límite de 0.5 para las predicciones, en este caso se hacen combinaciones de las variables alpha y l1_ratio para poder optimizar la respuesta y elegir el modelo con mejor puntaje.

Los parámetros utilizados fueron:

alphas = [0.1, 0.5, 1.0]

l1_ratios = [0.1, 0.3, 0.5]

Los mejores resultados se obtienen con alpha 0.1 y ratios de 0.1 y 0.3. Obteniendo la misma matriz de confusión:

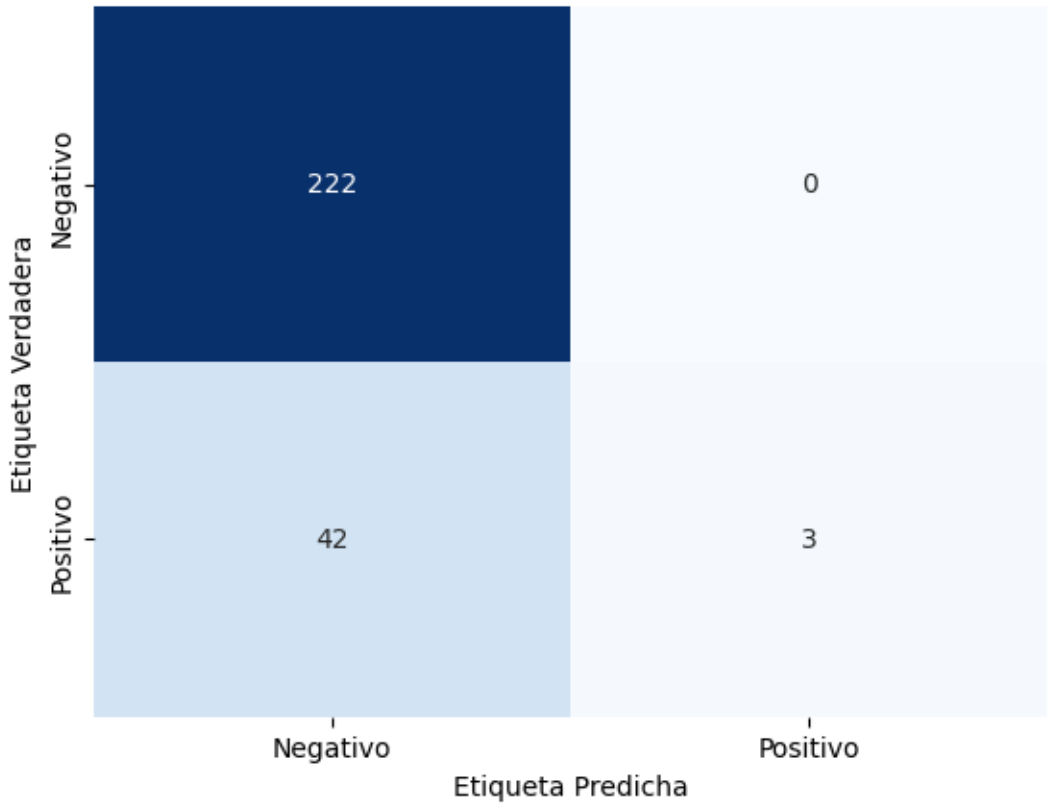

Figura 8 Matriz de confusión de regresión Elastic Net

Con un puntaje de 405, vale aclarar que es menor que el obtenido con la regresión lineal simple. Por otro lado, en este caso no elige ningún motor en mal estado y se ahorran 3 mantenimientos por lo que el costo asociado depende fuertemente del costo asociado.

En este caso las variables identificadas con mayor magnitud en la regresión son Fe, P y ST. Siendo el hierro la de mayor magnitud.

Regresión logística.

En la regresión logística se utiliza el parámetro alpha para optimizar el puntaje. Los valores 10, 9, 8, 5, 4, 2 y 1 fueron seleccionados y tanto el valor de 10 como el valor de 5 obtienen un puntaje de 403.

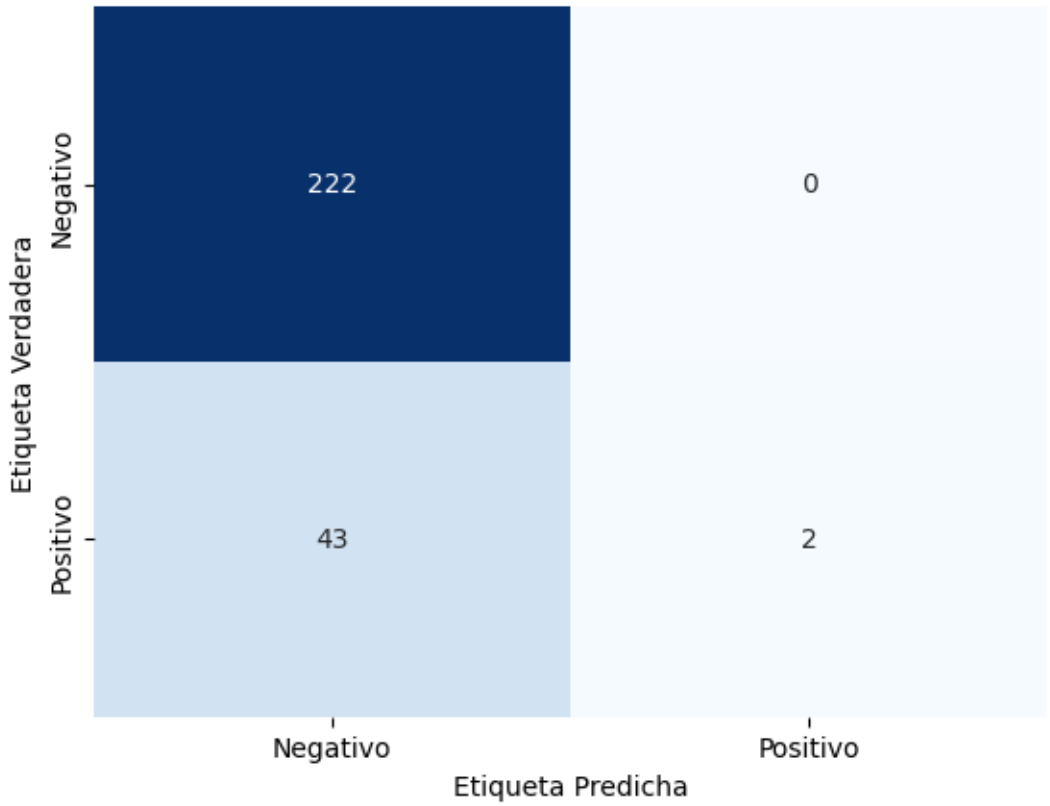

Figura 9 Matriz de confusión de regresión logistica

El puntaje obtenido es menor que el de la regresión lineal. Nuevamente no se identifica ningún motor en buen estado como motor en mal estado y se ahorran dos manteamientos por lo que dependerá de los costos asociados.

Las variables Fe, oilhours y preva tienen los p-valores menores a 0.05 y las variables preva, Fe y Si tiene los componentes de mayor magnitud para la regresión.

LightGBM

Este método puede ser optimizado directamente con el método de puntaje. Es por esto que se compara la respuesta al puntaje clásico de accuracy con el puntaje de este trabajo.

El modelo ajustado con accuracy da la matriz:

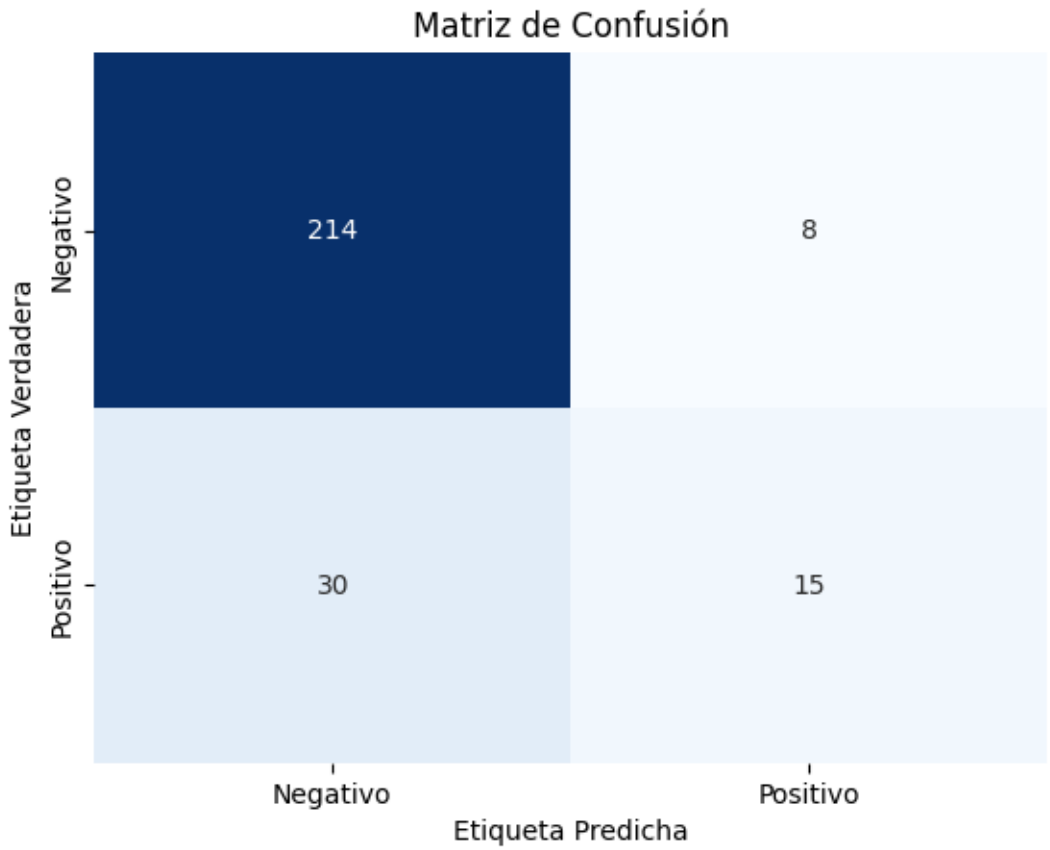

Figura 10 Matriz de confusión de LightGBM Accuracy

Con un puntaje de 397.

El modelo ajustado optimizado para la métrica utilizada da la matriz:

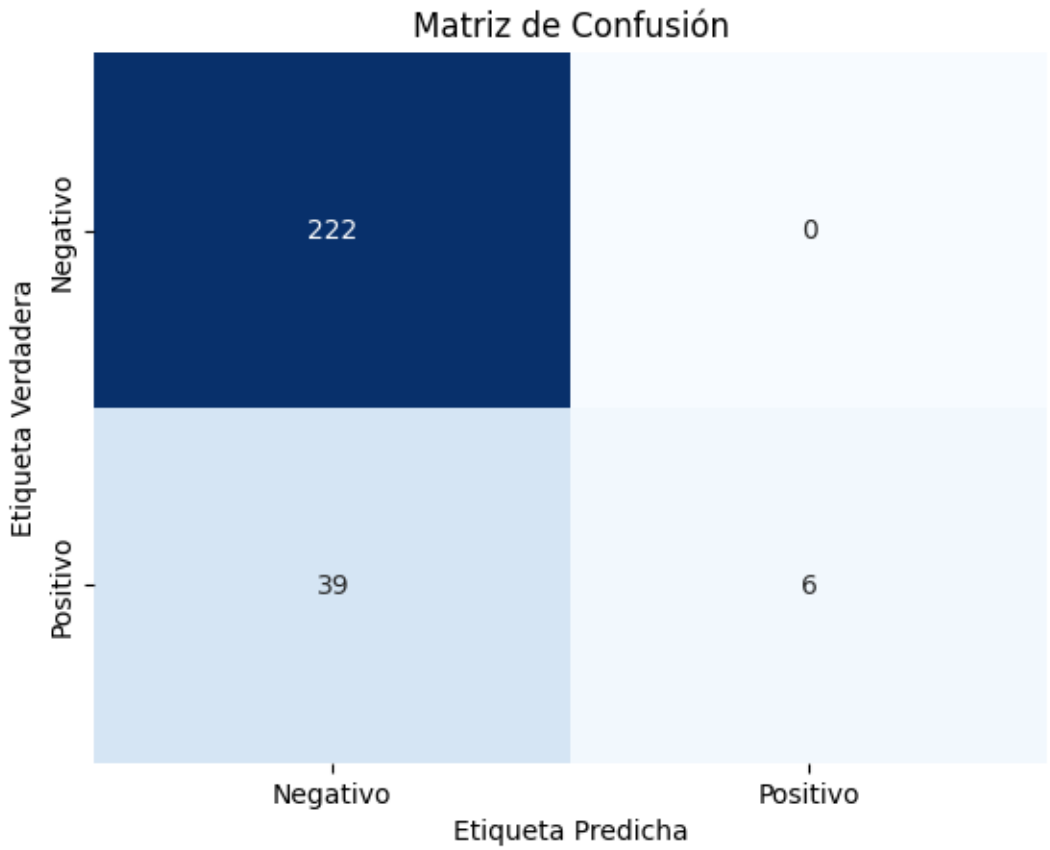

Figura 11 Matriz de confusión de LightGBM

Con un puntaje de 411 igual al de la regresión lineal, en este caso no tiene errores en la identificación de modelos en mal estado.

Como corolario del método, solo 11 de las 50 semillas aleatorias proporcionadas al modelo obtienen un puntaje mayor a 411 y el promedio de los puntajes de todas las semillas es de 409 por lo que el modelo de respuesta promediada de 50 semillas obtiene un puntaje mayor al promedio de las semillas. Este resultado parcial sobre la importancia del uso de varias semillas y la respuesta promediada de la matriz de predicciones reafirma la importancia de esta práctica en la implementación de modelos complejos con gran sensibilidad a la semilla aleatoria del modelo. De esta forma se minimiza la varianza del modelo y se obtiene un resultado mejor que el resultado promedio de las semillas.

Por otro lado, el modelo permite hacer una importancia de atributos el cual da como más importantes al Zn, V40, Ca, P, ST y Na. Estando las variables prev”n” creados al final de la lista.

Es interesante notar que se llega a atributos completamente diferentes a los de los otros modelos. Y que este es el único método que se ajusta a la métrica buscada, por lo que presenta mayor versatilidad.

Análisis FODA y posibles pasos para implementación

La implementación de un método de mantenimiento predictivo requiere de un alto compromiso por parte de las industrias. La inversión que debe hacerse no es solamente en capital humano y tecnológico sino también en tiempo.

A continuación, se describen un posible camino para la implementación de este tipo de sistemas.

1) Definir objetivos claros. En nuestro caso de estudio el objetivo es minimizar la cantidad de equipos cuyo mantenimiento ordinario hubiera sido tardío con los esquemas tradicionales y como segunda instancia optimizar, con la menor cantidad de manteamientos la menor cantidad de fallas, el esquema ordinario de mantenimiento.

2) Obtener los datos y crear una base de datos. Construir una base de datos con una cantidad razonable de muestras es el proceso más costoso de este tipo de sistemas y depende de cada industria como debe hacerse. En el caso de motores puede hacerse este tipo de análisis quimico con una determinada frecuencia y luego el seguimiento adecuado para poder identificar el estado del motor y su necesidad de mantenimiento. Siendo un modelo supervisado se debe contar con la variable que se busca predecir en el conjunto de datos de entrenamiento. Es importante aclarar que la frecuencia de muestreo debe definirse también de acuerdo al problema y a la experiencia del equipo de mantenimiento, algunos procesos pueden tener incluso muestreo continuo.
3) En cada industria se debe hacer un análisis exploratorio de los datos recopilados para identificar posibles problemas de muestreo o falta de información relevante.
4) Se debe seleccionar y desarrollar diferentes tipos de modelos. Este trabajo busca ser un ejemplo de una metodología de selección y evaluación de modelos.
5) En varios casos los modelos nos permiten establecer umbrales de alerta, un ejemplo es el uso de feature importance en los modelos de tipo árbol simples y sus umbrales de corte para las ramas troncales.
6) Una vez desarrollados los modelos se debe buscar la integración del mantenimiento predictivo en la metodología de mantenimiento ya existente. Un primer paso es tenerla como alerta temprana, pero sin desencadenar una acción. De esta forma se puede ganar confianza con el método. Este tipo de acciones permite además hacer un primer paso en el AB testing.
7) Una vez que se tenga confianza se puede generar un conjunto chico de maquinaria que se monitorea con acciones de mantenimiento predictivo. Con el modelo indicando si una maquina requiere un mantenimiento antes de lo programado. En una parte se hace el mantenimiento y en otra parte se mantiene el cronograma original. De esta forma puede realizarse un control.
8) Debe capacitarse al personal tanto los alcances del modelo, el plan de monitoreo que se le esté realizando y fundamentalmente en la forma en la que se vaya a realizar el muestreo. Puesto que si el muestreo es consistente con el formato original el método tendrá mejor validez. Si los datos que entran al modelo son recolectados de otra forma el modelo puede fallar con facilidad.
9) Por ultimo debe tenerse un enfoque de mejora continua donde se revise los resultados y se retroalimente con la nueva información al modelo para mantenerlo actualizado.

A partir de esta posible implementación se realiza un análisis de enfoque FODA para este trabajo.

Figura 12 FODA

Conclusiones

De acuerdo a los resultados obtenidos se pueden obtener varios puntos de análisis.

- Una regresión lineal puede ser un buen punto de partida para este tipo de problemas, no solo por haber tenido efectividades similares al resto sino también como base de comparación.
- Los atributos con mayor magnitud en los métodos estadísticos y los atributos con mayor importancia en el método de LightGBM son distintos. Por lo tanto, no sería recomendable tomar menos características del aceite en el laboratorio. Diferentes métodos pueden llegar a utilizar otros atributos para llegar a resultados similares.
- Es importante definir bien la métrica. En la evaluación sobre la utilidad de un método la capacidad que tenga el mismo de ahorrar recursos va a ser más valorada que el nivel de correlación.
- La capacidad que tiene un método como LightGBM de adaptarse a diferentes métricas y tener una respuesta adecuada según las necesidades del problema lo convierten en un método más eficiente en este tipo de problemas.
- Los métodos estadísticos pueden servir para identificar problemas generalizados, siendo capaces de mostrar una correlación con relevancia estadística se pueden obtener señales de alerta temprana como lo es que los motores en mal estado suelen tener un contenido alto de Hierro.

En resumen, este trabajo busca explorar las diferentes opciones dentro del Machine Learning para utilizar la información disponible en los ensayos de laboratorio en plantas o talleres. Se ha demostrado que los métodos predictivos tienen la capacidad de ahorrar dinero en la industria. Aprovechar los métodos estadísticos para correlacionar variables y comprender el problema físico puede ser de gran utilidad. Sin embargo, para obtener predicciones más precisas, se requiere un método capaz de adaptarse a diferentes métricas y optimizar las ganancias de manera adecuada. Es importante

mencionar que este trabajo no incluyó el Feature Engineering ni la búsqueda adicional de información, lo cual podría enriquecer aún más los nuevos algoritmos como LightGBM.

Por último, este trabajo busca mostrar la importancia que tienen este tipo de modelos en el desarrollo de nuevas tecnologías para las industrias. Este tipo de análisis puede hacerse en cualquier tipo de equipo que puedan hacerse análisis rutinarios y se tenga una base de datos actualizada y completa sobre el estado del motor en base a dicho análisis. La industria minera, química, refinerías y cualquier otra relacionada debe en primera instancia avanzar hacia la digitalización y toma de muestras para generar las bases de datos. Una vez creadas este tipo de modelos pueden utilizarse y obtenerse grandes beneficios.

Referencias:

Philips, J., Cripps, E., Lau, J. W., & Hodkiewicz, M. (2015). Classifying machinery condition using oil samples. ELSEVIER, 316-325.
Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). Application of machine learning and artificial intelligence in oil and gas industry. Petroleum Research, 6(4), 379-391.
Theissler, A. K. (2020). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. Procedia CIRP.
Yeardley, A. (2018). Integrating machine learning techniques into optimal maintenance scheduling. Journal of Quality.
https://lightgbm.readthedocs.io/en/latest/index.html Información del paquete lightgbm
https://scikit-learn.org/stable/ Información sobre los paquetes de los modelos utilizados
https://scikit-optimize.github.io/stable/ Información sobre el paquete de optimización bayesiana utilizado.

Anexo

Link al repositorio con script de trabajo para replicabilidad:
https://github.com/FUkrell/TrabajoEspecialista/blob/main/Federico_Krell_Trabajo_Especialista_UBA.ipynb

Curriculum vitae:

Federico Uriel Krell

Título y lugar de estudio:

Ingeniero Químico, Facultad de Ingeniería de la Universidad de Buenos Aires, Argentina, 2021.
Master en Exploración de Datos y Descubrimiento de Conocimiento (Data Science), Exactas y Naturales de la Universidad de Buenos Aires, Argentina, 2022-2023 (En Desarrollo).

Cargo actual:
Ingeniero de Procesos Ssr en TECHINT E&C. (2019-Actualidad)

Trayectoria:
Investigador Ad honorem para laboratorio de Medios Porosos UBA/CONICET y Laboratorio de Procesos Catalíticos UBA/CONICET.
Ayudante segundo Algebra Lineal Facultad de ingeniería UBA.

Tesis de grado:
Aprendizaje automático aplicado al control industrial.