CIENCIA DE DATOS APLICADA

Entrega 1

Modelo de clasificación de fallas de sistemas de levantamiento artificial

Juan David Ayala Nariño
Brayan Steven Garcia Cardenas
Alberto Jose Mendoza Peñaloza
Carlos Fernando Montaña Herrera

Universidad de los Andes

Departamento de Ingeniería de Sistemas y Computación

Bogotá D.C., Septiembre 17 de 2022

1. **Definición de la problemática y entendimiento del negocio**

- **Organización**: HOCOL S.A

- **Contexto**:

Hocol es una filial del Grupo Ecopetrol. Las actividades de producción y exploración se concentran especialmente en las cuencas del norte de Colombia (Guajira, Sinú San Jacinto, Valle Inferior del Magdalena y Cesar Ranchería), los Llanos (norte del Meta y sur del Casanare) y el Valle Superior del Magdalena (Huila y Tolima).

En Hocol se implementan procesos de aplicación de nanotecnología para el mejoramiento de crudos y de su movilización, y se cuenta con avanzadas facilidades de procesamiento de fluidos. Estas competencias se complementan, además, con importantes reducciones en los costos de levantamiento y desarrollo y en la disminución de pérdidas de producción.

- **Estrategia Corporativa (Producción)**:

Asegurar una producción bruta de 38,9 Barriles de producción promedio diaria de crudo para el año 2024, habilitada en la continuidad operacional y la operación limpia de barriles.

- **Oportunidad**:

Modelo de clasificación de fallas de sistemas de levantamiento artificial

Las pérdidas por fallas en los sistemas de levantamiento artificial de crudo pueden ser significativas en la organización, ya que pueden provocar interrupciones en la producción y altos costos en tiempos de inactividad, cuando un sistema de levantamiento artificial falla, la producción de crudo se detiene o se reduce significativamente, lo que resulta en una pérdida directa de ingresos para HOCOL entre un 3% y 5% de la producción diaria del pozo.

La reparación de sistemas de levantamiento artificial puede ser costosa, especialmente si se requiere la sustitución de componentes clave, estos costos incluyen piezas de repuesto, mano de obra y el tiempo necesario para llevar a cabo las reparaciones. Para evitar fallas en el sistema de levantamiento artificial, es necesario llevar a cabo un mantenimiento preventivo regular y controlado, lo que conlleva a una disminución de los costos de mantenimiento correctivo en un 30% en promedio.

El alcance de este proyecto está enfocado en disminuir los costos de paradas de producción en el campo, por eventos relacionados con mantenimientos correctivos y fallas en elementos críticos del pozo, los cuales pueden detectarse de manera proactiva y ser mitigados controladamente, asegurando la continuidad operacional del campo y la mínima afectación económica en la producción de crudo del bloque.

- **Objetivo del Proyecto**

Diseñar e implementar un modelo de clasificación de fallas de los sistemas de levantamiento artificial para el campo Ocelote, que permita disminuir las pérdidas de producción de crudo y las paradas no programadas de pozo.

- **Métricas de negocio**

    - **# de horas hombre disminuidas en mantenimientos correctivos por detección temprana de fallas** – Este indicador se refiere a la cantidad de horas hombre que se dejan de ejecutar por la detección temprana de fallos, en todos los frentes relacionados con las áreas de mantenimiento y operación (Inventarios, bodega, Operación, etc).

    - **# de barriles producidos** – Este indicador se asocia a la cantidad de barriles que no se ven afectados por una parada programada de pozo, frente a los barriles que se dejan de producir por los mayores tiempos que conllevan las paradas no programadas.

    - **% de disminución de tiempos en paradas no programas de pozos –** Este indicador se asocia al % de disminución de tiempo que el pozo deja de producir crudo por paradas no programadas. Uptime del pozo.

2. **Ideación**

- Usuarios del Producto: **Vicepresidencia de Desarrollo y Producción**

- Procesos relacionados con la problemática

En la actualidad, HOCOL no cuenta con un proceso automatizado para la detección temprana de fallos en los sistemas de levantamiento, los trabajos y procesos relacionados con esta función, están centrados en información histórica catalogada por la jefatura de tecnología de producción, y con la cual, se realizan las programaciones de los mantenimientos preventivos en los pozos del bloque Guarrojo.

- Requerimientos del producto de datos

Información acerca de las variables reportadas por el dispositivo IoT, su nivel de importancia para el negocio y unidades de medida. Además de información detallada de los pozos que permita determinar su condición en un momento determinado de tiempo, extraída de los sistemas Scada Experion y almacenadas temporalmente en las bases de CosmoDB. Esta información es reportada cada 10 minutos a través de dispositivos IoT en cada yacimiento, por lo que existe un gran volumen de datos. Por esto, se estima que con menos de un mes de registros es posible implementar un modelo de ML que atienda el objetivo de negocio.

- Componentes tecnológicos

Entre los requerimientos tecnológicos se encuentra el lenguaje Python y la librería sckit-learn para desarrollar los modelos de ML. Una infraestructura para desplegar el dashboard a presentar en negocio como streamlit o github. Además, será necesario un host gratuito para desplegar el servicio API REST, desarrollado en frameworks de python tales como Django. Este servicio brindará información de las predicciones derivadas del modelo, permitiendo filtrar por pozo y fecha.

- Mockup

3. **Responsible**

- **Ética**
  - **Responsabilidad**: Asumir la responsabilidad por el funcionamiento y los posibles resultados erróneos o daños causados por fallos en la detección.

- **Privacidad**
  - **Recopilación de datos**: Asegurarse que la recopilación de datos cumple con las leyes de privacidad y que se obtenga el consentimiento adecuado de las partes involucradas.
  - **Acceso restringido**: Limitar el acceso a los datos a personas autorizadas y garantizar que se utilicen solo para los fines previstos.

- **Confidencialidad**:
  - **Seguridad de datos:** Implementar medidas de seguridad sólidas para proteger los datos utilizados en la detección de fallos, especialmente si incluyen información confidencial o sensible.
  - **Contratos y acuerdos de confidencialidad**: Establecer acuerdos claros de confidencialidad con terceros que puedan tener acceso a los datos o algoritmos.
  - **Auditoría**: Permitir la auditoría de los sistemas para verificar su funcionamiento y la calidad de las decisiones.

- **Aspectos regulatorios**:

- **Cumplimiento normativo**: Cumplir con todas las leyes y regulaciones aplicables relacionadas con la recopilación y el uso de datos, así como con el funcionamiento de sistemas de IA
- **Normas de la industria**: Seguir las mejores prácticas y estándares de la industria para garantizar la calidad y la seguridad de los sistemas.

4. **Enfoque Analítico**

La hipótesis de negocio es que existen un conjunto de variables relevantes reportadas por el dispositivo IoT en los yacimientos, con las cuáles se puede definir cuándo los sistemas de levantamiento mecánico se encuentran en buen o mal estado. Estas variables están basadas en carga de levantamiento, corriente de motor, frecuencia de salida, presión, temperatura, entre otros. El ingeniero de producción de la empresa nos proporcionará unos rangos normales para estas variables, con las cuáles se podrá adelantar un proceso de etiquetado binario para implementar un modelo de clasificación. Entre las propuestas de modelos para desarrollar se encuentran árboles de decisión (random forest), Naive Bayes y regresión logística. Las métricas para evaluar la calidad del modelo serán la precisión y el recall. Especialmente es importante el recall negativo, para saber qué porcentaje de los sistemas de levantamiento en mal estado no será capaz de identificar el modelo y generarán pérdidas al negocio. La técnica de visualización elegida en este caso, es generar un listado diario de los pozos que se encuentran en buen y mal estado con el propósito de poder monitorearlos constantemente. Además, mostrar el porcentaje de yacimientos del campo ocelote que se encuentran en mal estado y una gráfica histórica de cómo ha evolucionado el estado de los pozos en el tiempo.

5. **Recolección de datos**

Para el EDA se nos proporcionó la información de un día de las variables de los sistemas de levantamiento mecánico. Estos datos son reportados por un dispositivo IoT en formato json en el storage de azure de la empresa. Esta información será apoyada con un documento de rangos normales de estas variables que será proporcionado por negocio para etiquetar el dataset y proceder a la implementación de los modelos mencionados. Los json contienen la fecha de reporte, el yacimiento al que pertenecen y el siguiente listado de mediciones:

Variable	Unidad	Descripción
I_MOT.PV	Amperios (A)	Corriente del motor
PIP.PV	PSI (lb/in2)	Presión de entrada a la bomba
T_INT.PV	Farenheit (°F)	Temperatura a la entrada de la bomba
THP.PV	PSI (lb/in2)	Presión de Cabeza Tubería
V_OUT.PV	Voltios (V)	Voltaje de Salida del equipo
T_MOT.PV	Farenheit (°F)	Temperatura del motor
V_MOT.PV	Voltios (V)	Voltaje en el motor
FRE_OUT.PV	Hertz (HZ)	Frecuencia de Salida
FRE_OUT_PMM.PV	Hertz (HZ)	Frecuencia de salida motor imanes permanentes
CHP.PV	PSI (lb/in2)	Presión de gas anular

STR.PV	indefinido	Arranque del Aviador
FRE.PV	Hertz (HZ)	Frecuencia del dixie del generador o la red
RPM.PV	RPM	Revoluciones por minuto del motor

6. Entendimiento de los datos

El procedimiento de calidad de datos consistió en la carga del archivo json en la librería pandas, evidenciando un porcentaje de menos de 1% de duplicados y nulos que fueron eliminados. Luego se encontraron dos columnas del dataset que estaban en diccionarios, los cuales fueron normalizados para contar con todos los datos en columnas. Además se encontró una columna que contenía concatenados el yacimiento del cual fueron extraídos los datos y la variable medida, por lo que se procedió a estructurar estas dos variables como columnas del dataset. Por último, se procedió a darle formato a las fechas y eliminar las columnas que contenían ids. Acerca del EDA, se encontraron diferencias entre la cantidad de registros en los yacimientos y las variables reportadas, hallando algunas variables que tenían menos del 1% de datos que debían tener para un reporte cada 10 minutos. Se probó con el chi cuadrado que la distribución de registros por hora no es uniforme. En adición, en las variables se encontraron una gran cantidad de outliers y datos inconsistentes, datos negativos y positivos muy altos en las que no hay correspondencia con el tipo de variable. Estos problemas con los datos, procederán a ser consultados con negocio para aclarar y entender cómo proceder según sus directrices. Este procedimiento fue documentado en un paso a paso detallado en el jupyter notebook de analysis.ipynb.

7. Conclusiones y próximos pasos

El proyecto que se ha estructurado en este documento es de alto impacto para la organización, ya que aporta a que se puedan alcanzar los objetivos de producción de la empresa y que se disminuyan las pérdidas de ingresos por mantenimiento. Sin embargo, aún existen desafíos en la calidad de datos como los outliers y valores inconsistentes que podrían sesgar el modelo, por lo que hay un reto en consultar con negocio y desarrollar un producto de datos con buen rendimiento, según sus directrices. Se reporta una gran pérdida de información en las variables en términos de datos que no siguen la periodicidad de 10 minutos. Teniendo en cuenta esto, puede que sea necesaria solicitar una cantidad de datos considerable para cumplir con los requerimientos de implementar un buen modelo. Es relevante para la evaluación de los modelos la métrica del recall negativo para saber qué porcentaje de los sistemas en mal estado no podrá clasificar el modelo. También sería importante observar cómo coincide el árbol de decisión desarrollado con los rangos de los parámetros entregados por negocio. Los próximos pasos consisten en solucionar las dudas en los datos con el equipo de la empresa, para desarrollar las pruebas de concepto con los primeros modelos. Se requiere además ir gestionando apoyo de tecnologías como github y streamlit para desplegar el dashboard del proyecto, y el cloud hosting de render.com para el API REST a realizar.