Universitat Oberta de Catalunya

uoc.edu

Estudio de la presencia del fenómeno Concept Drift en el tratamiento en tiempo real del dataset 3W de Petrobras

Máster Universitario en Ciencia de Datos (Data Science)
TRABAJO FINAL DE MÁSTER Área: 2

Autor: Baltasar Boix Rita
Tutor: Jesús López Lobo
Profesor: Jordi Casas Roma

Índice

1. Introducción. 1.1 Descripción del problema.

Dataset 3W

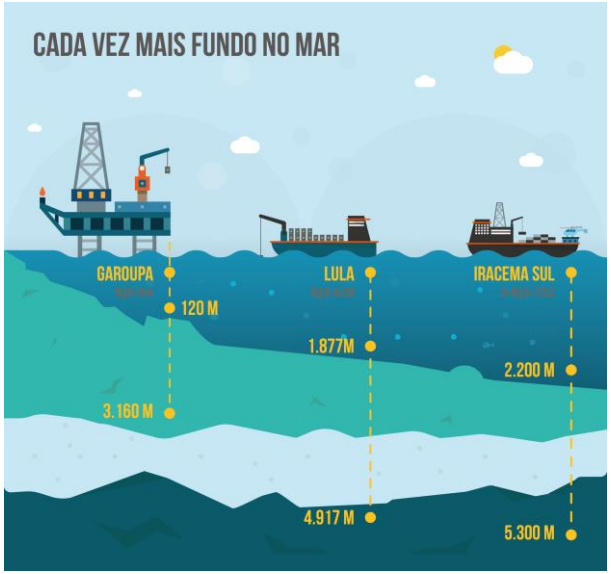
- Colección de series temporales de datos de operación del proceso de extracción de crudo offshore.
- Describen una operación normal y ocho anómalas (etiquetas).
- Petrobras anima a investigar la aplicación de algoritmos de ML sobre el dataset.

1. Introducción. 1.1 Descripción del problema.

Petrobras FPSO (Floating, Production, Storage and Offloading) P-71, instalado en el campo de Itapu, en el área presalina de la Cuenca de Santos, a 200 km de la costa de Río de Janeiro.

https://www.worldenergytrade.com/oil-gas/produccion/brasil-el-nuevo-fpso-de-petrobras-entra-en-produccion

https://es.wikipedia.org/wiki/Cuenca_de_Santos



https://www.conselhoparlamentar.org.br/tecnologia-em-aguas-profundas-coloca-petrobras-no-topo-do-mundo/

1. Introducción. 1.1 Descripción del problema.

Las estadísticas de Seguridad de la Industria de Proceso muestran que los grandes accidentes son infrecuentes pero cuando ocurren pueden transformar a todo un sector (BP Deepwater Horizon Oil Spill).

Control regulatorio. El uso de ordenadores para el control de procesos complejos ha permitido automatizar acciones de bajo nivel como la apertura y cerrado de válvulas.

La respuesta ante situaciones anormales del proceso sigue estando en buena parte en manos de los operadores. Esto implica la detección temprana de las situaciones anormales, diagnosticando sus causas y tomando las acciones de control apropiadas para llevar el proceso de nuevo a un estado normal y seguro.

https://www.britannica.com/event/Deepwater-Horizon-oil-spill

1. Introducción. 1.1 Descripción del problema.

Abnormal Event Management. Se busca la automatización de la detección y diagnóstico de fallos en el proceso industrial.

Se utilizan técnicas de aprendizaje automático (ML) para entrenar modelos que detectan y diagnostican situaciones anómalas a partir de datos en tiempo real. Podemos distinguir dos formas de entrenar los modelos:

- Batch/Offline learning. Se selecciona un dataset de datos y se aplican técnicas de aprendizaje supervisado. Periódicamente se actualiza el dataset y se reentrena el modelo.

- Stream/Online learning. Los datos se adquieren secuencialmente en línea y se utilizan para actualizar el modelo. Utilizan el aprendizaje incremental.

Concept Drift. Consiste en el cambio que se puede producir en la distribución de los datos. Si esto ocurre a menudo los modelos entrenados quedan obsoletos muy rápidamente. Se hace necesario detectar estos cambios en tiempo real y reentrenar los modelos.

1. Introducción. 1.2 Objetivos del TFM.

Objetivo principal.

- En este trabajo se pretende demostrar que es posible llevar a cabo una predicción en tiempo real de los eventos anómalos producidos en la extracción de crudo en el dataset 3W de Petrobras.

Objetivo secundario.

- Considerar de forma adecuada la presencia del fenómeno concept drift durante el proceso de aprendizaje.

1. Introducción. 1.3 Metodología.

En este TFM se van a utilizar métodos de stream learning aplicados a las series temporales del dataset 3W de Petrobras.

El uso de stream learning en este caso presenta varios retos:

- El etiquetado lo realizan expertos en batch observando ventanas de tiempo de 5 minutos a 12 horas según la situación analizada. La observación de un solo registro no contiene suficiente información para clasificarlo y será necesario analizar tendencias.

- El origen de los datos de cada serie temporal es diverso. Hay que pretratarlos para poder utilizar el conjunto.

- Los creadores del dataset advierten que algunos de los tipos de eventos anómalos pueden sufrir de concept drift, cambios en el comportamiento de los datos que ocurren con el tiempo y deterioran el rendimiento de los modelos ajustados. Este va a ser el principal reto del TFM.

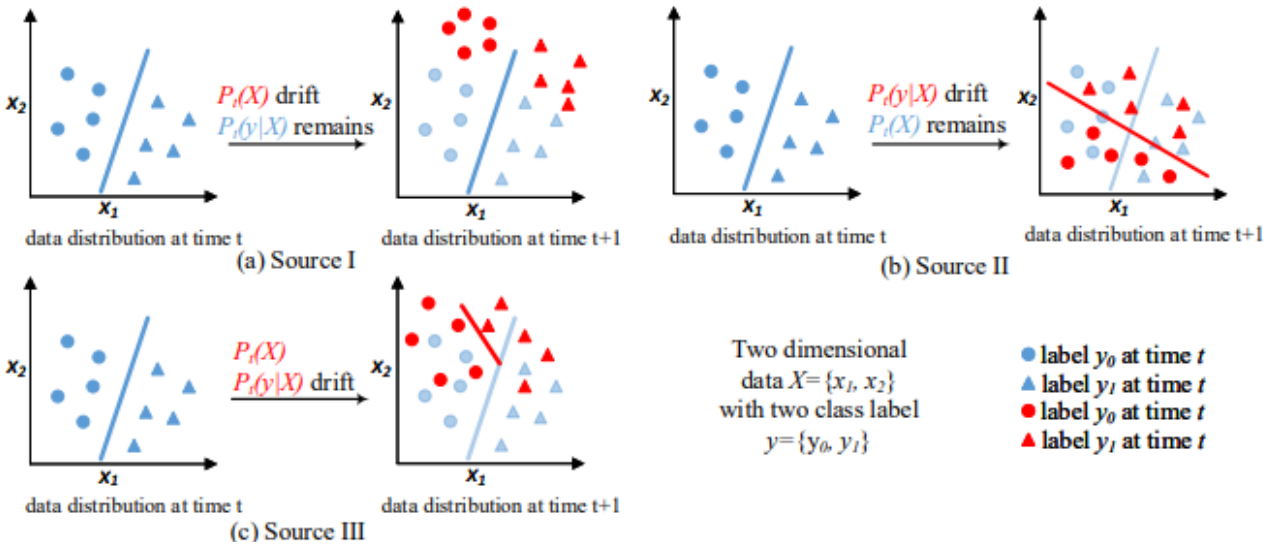
2. Estado del arte. 2.1 Online Learning.

El entorno para Online Learning impone unas restricciones computacionales con respecto a la configuración tradicional del Batch/Offline Learning:

- Cada muestra se procesa solo una vez a su llegada, y los modelos deben poder procesar las muestras secuencialmente tan pronto como se reciben.
- El procesamiento de cada muestra debe realizarse dentro del intervalo de tiempo entre muestras.
- El algoritmo debe usar solo una cantidad de memoria preasignada y finita.
- Tras cada exploración de datos, el modelo resultante debe ser válido para sus uso.
- El modelo producido debe ser equivalente al que se construiría por Batch Learning.

En este TFM se utilizan métodos de stream learning del módulo de Python River

2. Estado del arte. 2.2 Concept drift.



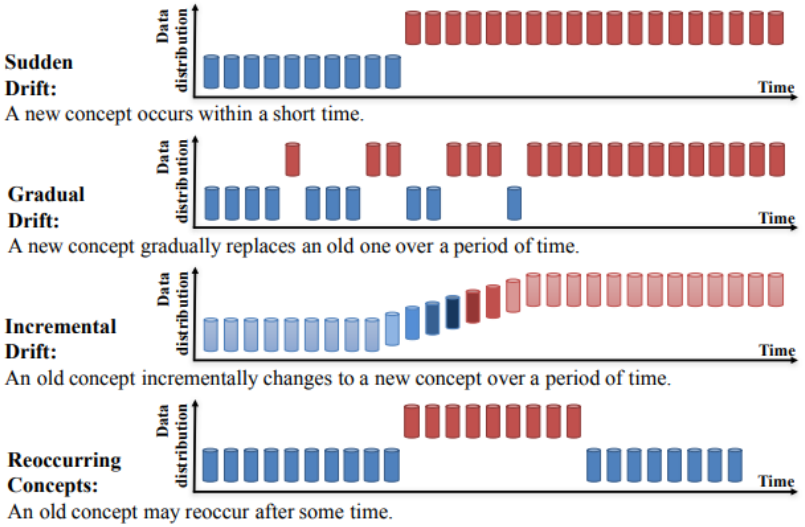
Concept Drift. Cambio en la probabilidad conjunta de X e y en el tiempo t.

$$P_t(X, y) = P_t(X)\ P_t(y|X)$$

Origen I: Virtual Drift

Origen II: Real Drift

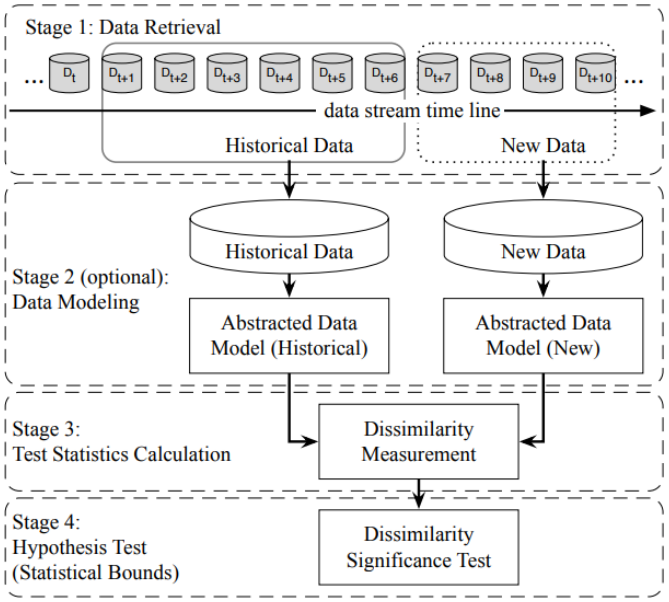
Origen III: Mezcla de I y II



Tipo de Drift

Uno de los principales retos de los métodos para detectar Concept Drift es distinguirlo del ruido en el flujo de datos

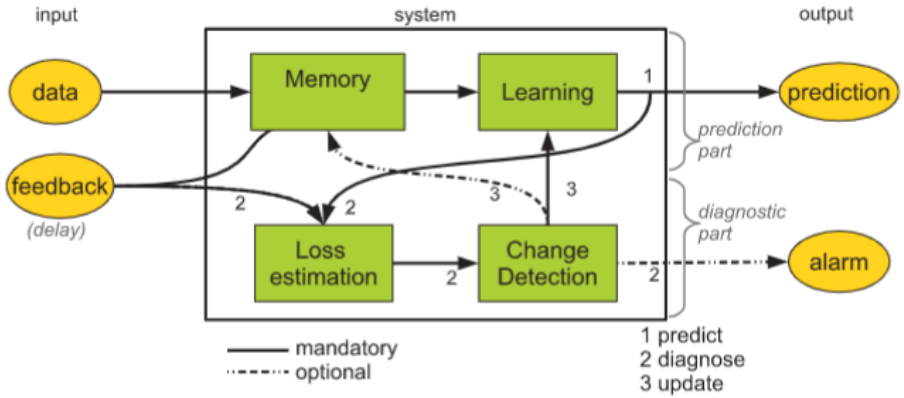
2. Estado del arte. 2.2 Concept drift.



Esquema de un detector de Concept Drift

La detección de Concept Drift se ha abordado tradicionalmente como una tarea supervisada, con datos etiquetados que se utilizan constantemente para validar el modelo aprendido. Aunque efectivas, estas técnicas no siempre son prácticas.

Las técnicas de detección de cambios no supervisadas son poco fiables. La ineficacia de las técnicas no supervisadas radica en la exclusión de las características del clasificador del proceso de detección.



Esquema genérico para un online adaptive learning algorithm

En este TFM se aplica este esquema con datos etiquetados.

2. Estado del arte. 2.3 Trabajos sobre el 3W dataset

- Publicación del dataset en 2019

Se han publicado diversos artículos que han aplicado distintas técnicas de ML para la clasificación de los eventos anormales. Marins *et al* utilizan un esquema de cálculo que comparten casi todos los artículos:

- Cada observación se forma a partir de un ventana de tiempo deslizante (*rolling time window*) de la que se calculan una colección de estadísticos (media, desviación estándar, máximo, etc.).

- Se aplica una reducción del número de variables (mediante técnicas como PCA).

- Los mejores resultados se obtienen utilizando Random Forest (RF) consiguiendo precisiones por encima del 90%.

- Utilizan métodos de Batch/Offline Learning

3. Materiales y métodos. 3.1 Descripción del dataset 3W

- El dataset 3W está formado por un conjunto de series temporales que describen la operación de un pozo de extracción.
- Las series tienen como origen:

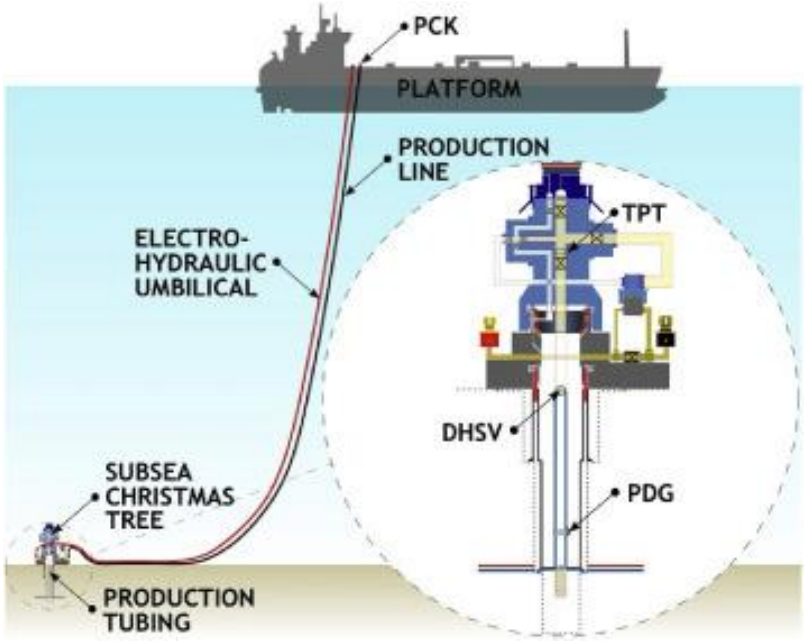
  Datos reales de pozos. Identifican el pozo con un número.

  Datos generados por simulación.

  Datos manuales generadas por expertos.

- Cada serie temporal se almacena en un archivo CSV.
- Cada observación es una línea en el archivo CSV (una por segundo).
- En cada observación: 8 variables del proceso + etiqueta

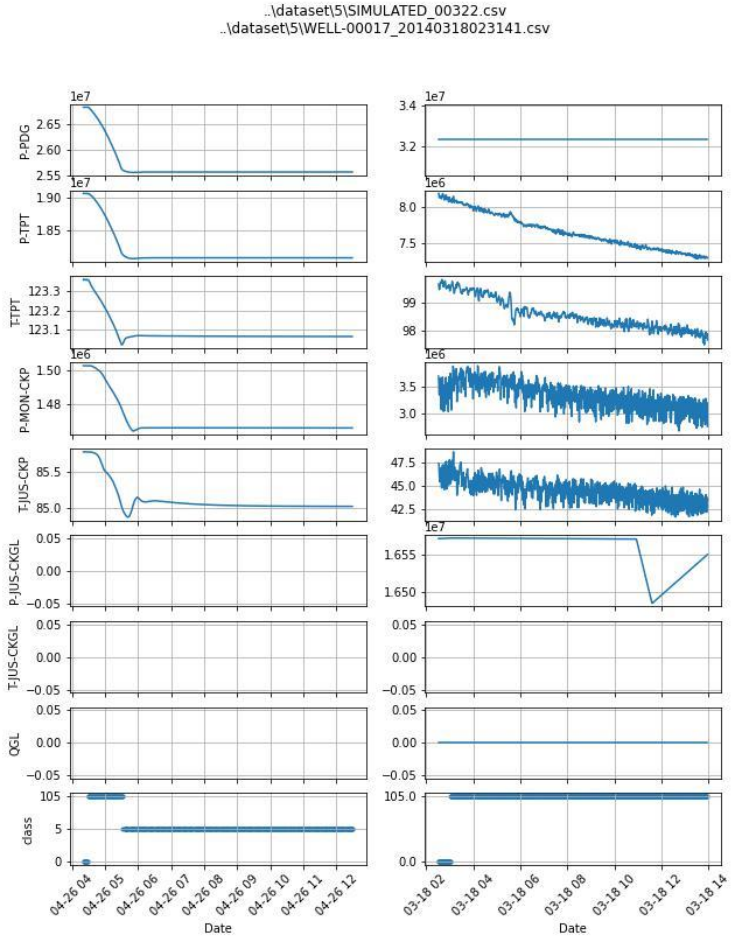
3. Materiales y métodos. 3.1 Descripción del dataset 3W



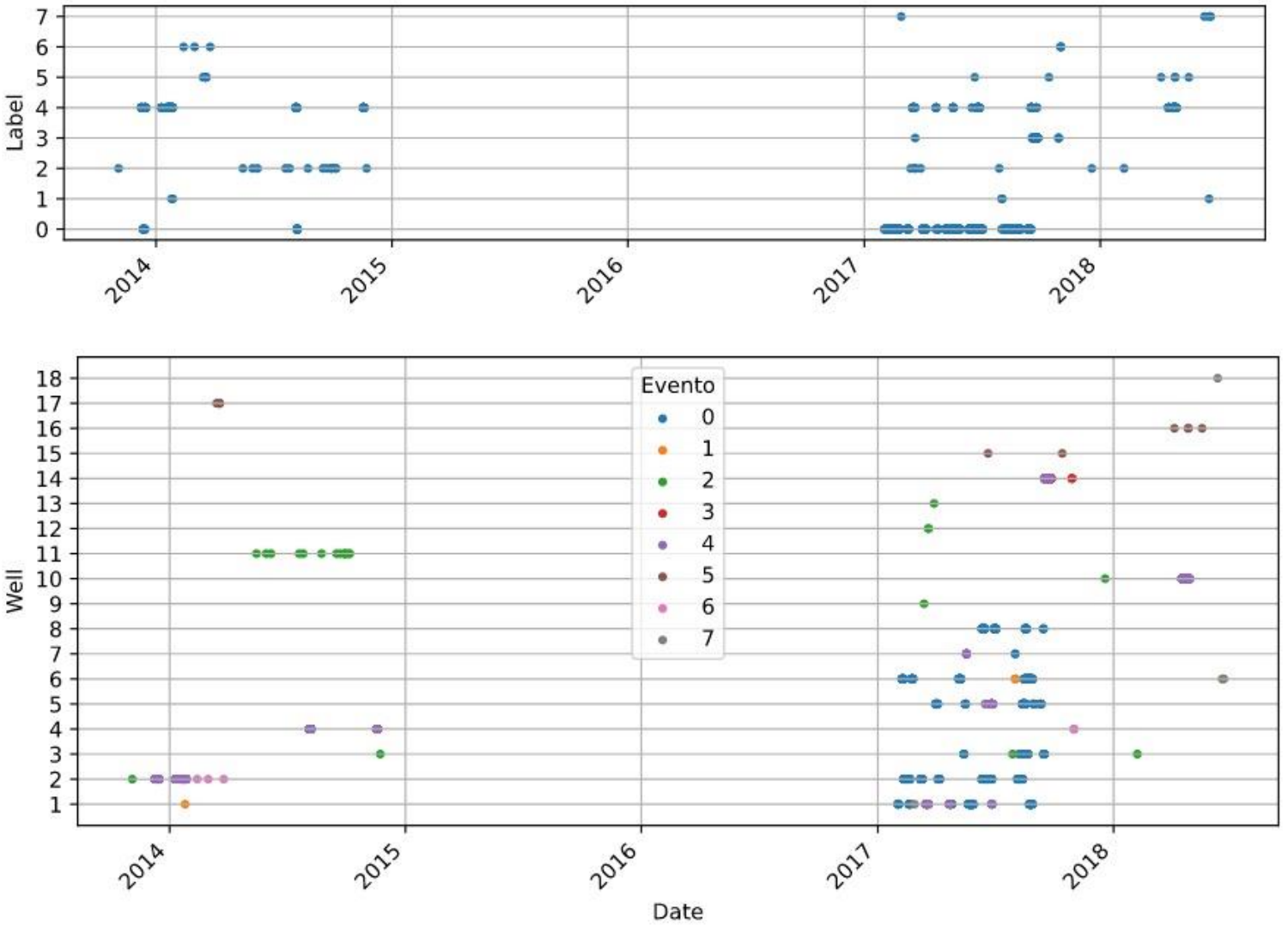
P-PDG: pressure variable at the Permanent Downhole Gauge (PDG);
P-TPT: pressure variable at the Temperature and Pressure Transducer (TPT);
T-TPT: temperature variable at the Temperature and Pressure Transducer (TPT);
P-MON-CKP: pressure variable upstream of the production choke (CKP);
T-JUS-CKP: temperature variable downstream of the production choke (CKP);
P-JUS-CKGL: pressure variable upstream of the gas lift choke (CKGL);
T-JUS-CKGL: temperature variable upstream of the gas lift choke (CKGL);
QGL: gas lift flow rate;

3. Materiales y métodos. 3.2 Análisis Exploratorio

SOURCE	REAL	SIMULATED	HAND-DRAWN	TOTAL
INSTANCE LABEL				
0 - Normal Operation	597	0	0	597
1 - Abrupt Increase of BSW	5	114	10	129
2 - Spurious Closure of DHSV	22	16	0	38
3 - Severe Slugging	32	74	0	106
4 - Flow Instability	344	0	0	344
5 - Rapid Productivity Loss	12	439	0	451
6 - Quick Restriction in PCK	6	215	0	221
7 - Scaling in PCK	4	0	10	14
8 - Hydrate in Production Line	0	81	0	81
TOTAL	1022	939	20	1981

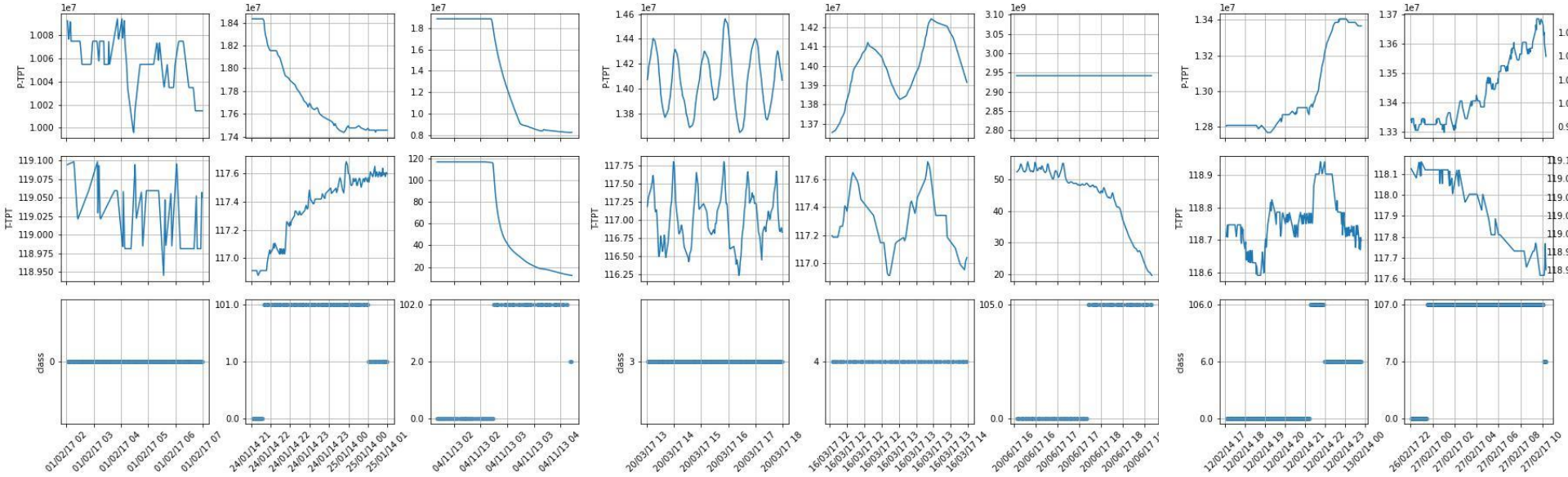


3. Materiales y métodos. 3.2 Análisis Exploratorio



Cronograma de las series reales del dataset 3W

3. Materiales y métodos. 3.2 Análisis Exploratorio



TYPE OF UNDESIRABLE EVENT	WINDOW SIZE
1 - ABRUPT INCREASE OF BSW	12h
2 – SPURIOUS CLOSURE OF DHSV	5min–20min
3 – SEVERE SLUGGING	5h
4 – FLOW INSTABILITY	15min
5 – RAPID PRODUCTIVITY LOSS	12h
6 – QUICK RESTRICTION IN PCK	15min
7 – SCALING IN PCK	72h
8 – HYDRATE IN PRODUCTION LINE	30min–5h

3. Materiales y métodos. 3.3 Preprocesamiento de los datos

- Se ha decidido reducir el tamaño del dataset utilizando el promedio por minuto en vez de por segundo de las variables. Para la etiqueta class se usa el valor más frecuente en el minuto. La dinámica del proceso no requiere una frecuencia de muestreo tan alta.

- Las series temporales tienen como origen datos reales de 18 pozos diferentes, series simuladas y otras rellenadas manualmente. Para poder utilizar el dataset en conjunto las variables se han normalizado (StandardScaler).

- Los datos no disponibles (Nan) se han estimado con el método forward fill cuando en una serie hay menos del 20 %; si hay más, toda la serie se sustituye por cero.

- Para gestionar los ficheros de las series temporales se ha programado la python class d3w que permite seleccionar las series según su origen, evento o pozo y generar conjuntos para entrenamiento, test y validación.

- Las series seleccionadas con d3w se alimenta a una python class que genera los datos de forma continua concatenando los registros de cada serie temporal para cada tipo de modelo.

3. Materiales y métodos. 3.4 Métodos online

Se utiliza el dataset 3W para emular un entorno en tiempo real en el que los datos se procesan registro a registro.

Es un ejercicio teórico en el que se hacen algunas aproximaciones que nos alejan del problema real:

- Se considera que se dispone de la etiqueta del registro con solo un minuto de retraso mientras que en el caso real el etiquetado lo han realizado expertos a posteriori.

- Se concatenan las series temporales como si fuera una sola cuando pueden distar entre ellas años.

- Las series que tienen como origen una simulación o están creadas manualmente no las podemos situar en el tiempo.

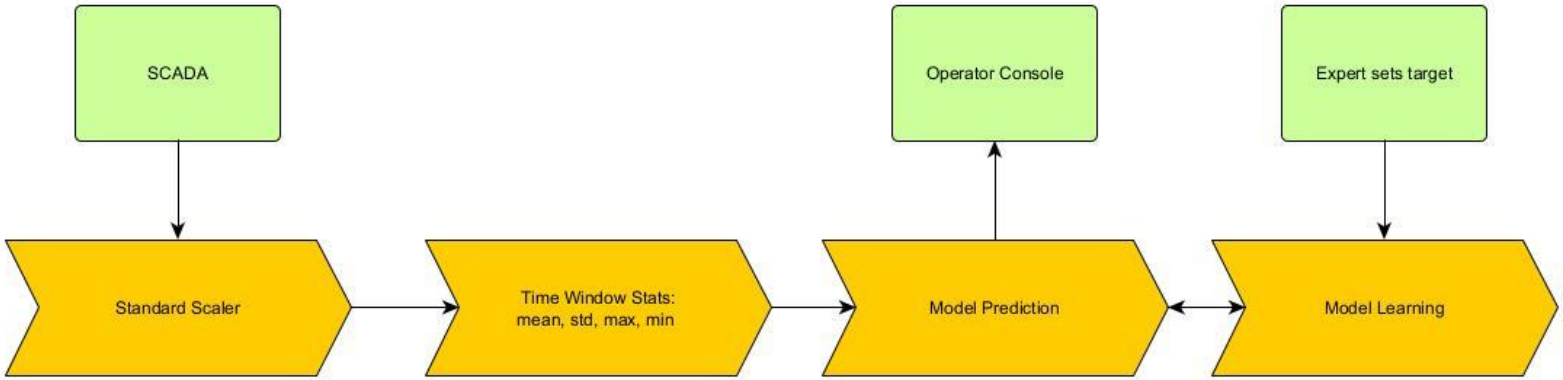
3. Materiales y métodos. 3.4 Métodos online

Modelo Base

Se hace un uso extensivo de las rutinas del módulo de Python River:

- Normalización de las variables (StandarScaler).
- Cálculo de las variables estadísticas de una ventana de tiempo deslizante.
- Predicción de la etiqueta del evento anómalo.
- Actualización del modelo.

Se reinicializa la normalización y el cálculo de variables con el cambio de serie temporal



3. Materiales y métodos. 3.4 Métodos online

Clasificador

- El modelo de aprendizaje incremental de River escogido ha sido un Hoeffding Tree Classifier. Un algoritmo clasificador multiclase incremental de inducción de un árbol de decisión.

- Aunque la ejecución del modelo se ha realizado para cada evento anómalo independientemente se requiere un clasificador multiclase ya que algunos distinguen tres situaciones (0: Operación Normal, 10x: transición, x: Evento anómalo).

- Se han probado los modelos linear_model.LogisticRegression y linear_model.ALMAClassifier (con un wrapper multiclase). Se ha comparado la precisión obtenida para cada modelo ajustado en modo solo predicción. Se comportan peor que el Hoeffding Tree.

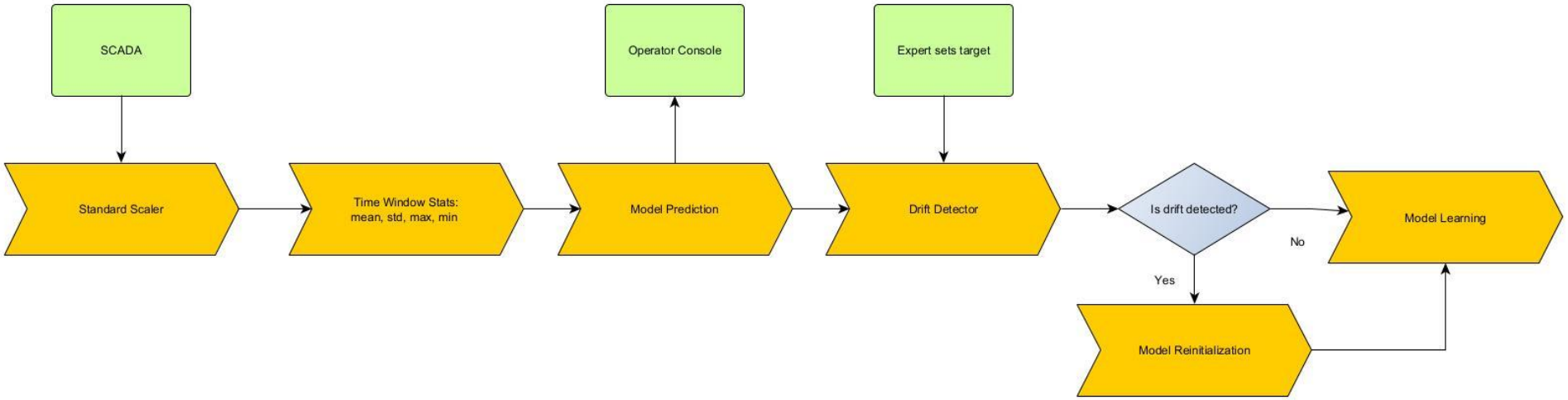
3. Materiales y métodos. 3.4 Métodos online

Detección de drift

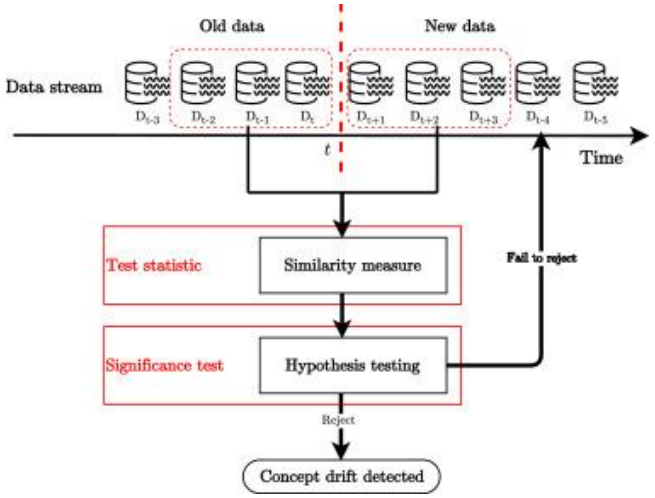
Se introduce un detector de drift en el esquema del modelo Base.

Cuando se detecta drift se reinicia el aprendizaje del modelo siempre que el modelo haya visto un mínimo de registros.

Se busca observar si se obtiene una mejora de precisión global.



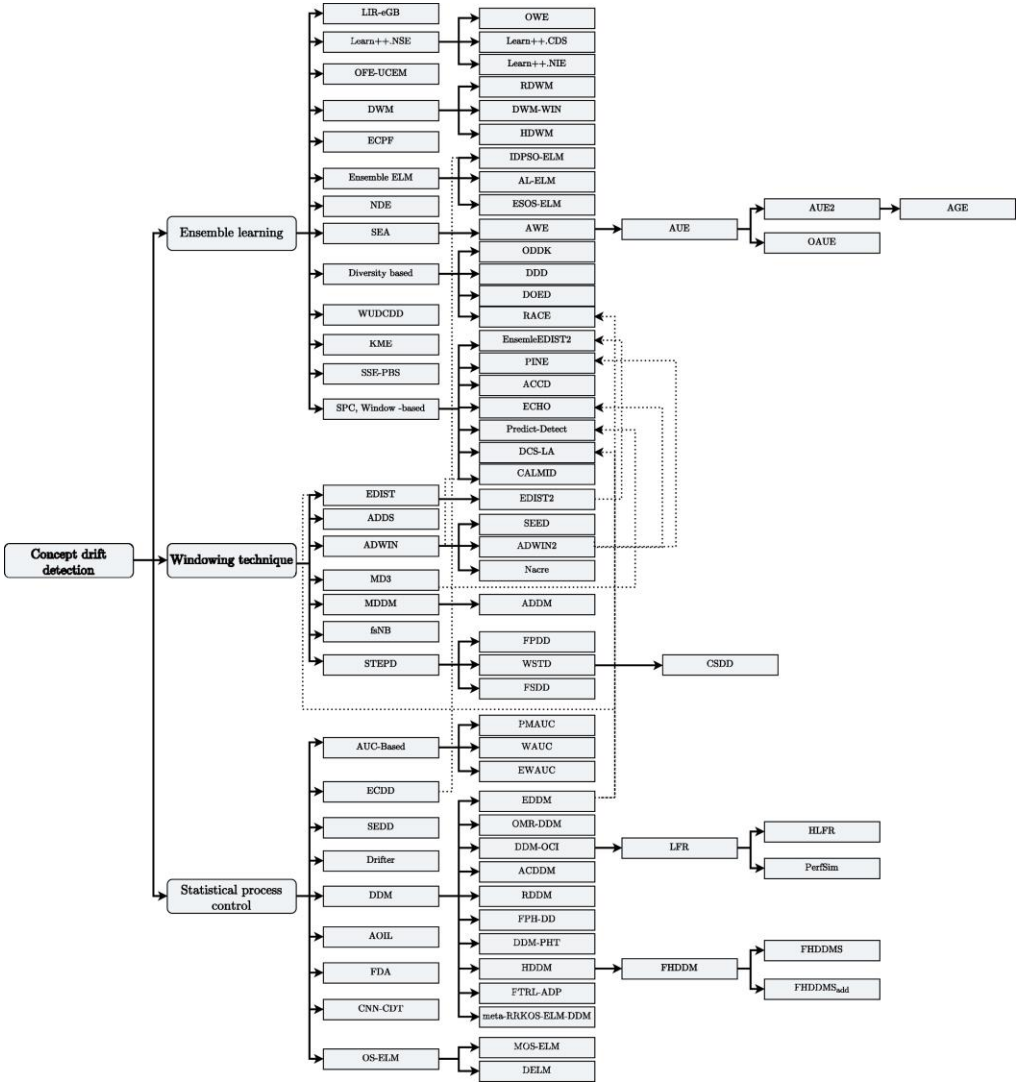
3. Materiales y métodos. 3.4 Métodos online



Se han ensayado los detectores de drift de River:

- ADWIN (ADaptive WINdowing) compara los promedios entre dos ventanas de tiempo recientes.
- DDM (Drift Detection Method) detecta un incremento en la tasa de error.
- EDDM (Early Drift Detection Method) detecta un incremento en la distancia media entre dos errores.
- HDDM_A (Hoeffding's bound Drift Detection Method) basado en la media móvil.
- HDDM_W(Hoeffding's bound Drift Detection Method) basado en EWMA (media movil ponderada exponencialmente)

3. Materiales y métodos. 3.4 Métodos online



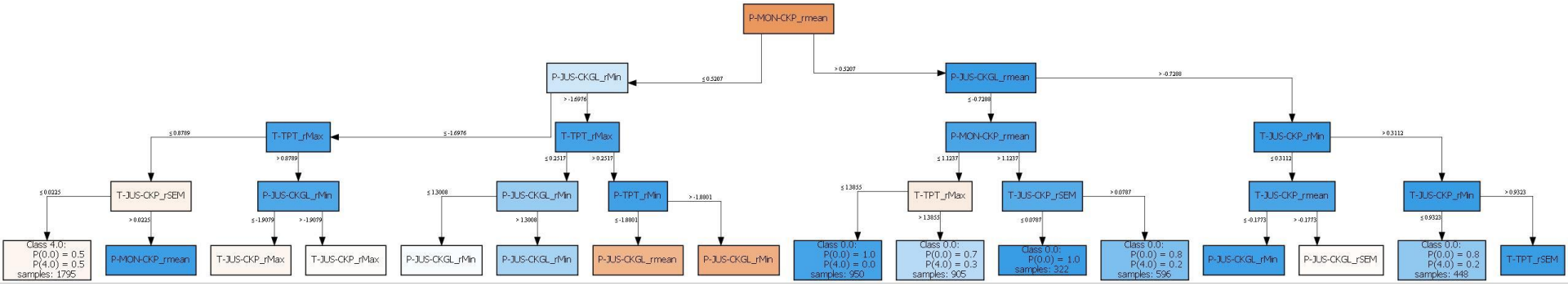
4. Experimentación y Resultados

Modelo

Representación gráfica de los primeros 4 de 12 niveles del Hoeffding Tree resultante del procesamiento de las series de datos reales ordenados por fecha con etiquetas:

- 0 (Operación Normal)
- 4 (Flujo inestable).

n_nodes	n_branches	n_leaves	n_active_leaves	n_inactive_leaves	height
183.0	91.0	92.0	92.0	0.0	12.0



4. Experimentación y Resultados

Modelo

Informe de clasificación del Hoeffding Tree resultante del procesamiento de las series de datos reales con etiquetas durante el entrenamiento y solo predicción:

- 0 (Operación Normal)
- 4 (Flujo inestable).

Se igualan los resultados si se entrenan unas 10 épocas en vez de una sola pasada.

BalancedAccuracy: 93.75%

	Precision	Recall	F1	Support
0.0	98.35%	95.07%	96.68%	149201
4.0	79.82%	92.43%	85.66%	31501
Macro	89.08%	93.75%	91.17%	
Micro	94.61%	94.61%	94.61%	
Weighted	95.12%	94.61%	94.76%	

Entrenamiento

BalancedAccuracy: 88.17%

	Precision	Recall	F1	Support
0.0	96.45%	92.44%	94.40%	149201
4.0	70.07%	83.90%	76.37%	31501
Macro	83.26%	88.17%	85.38%	
Micro	90.95%	90.95%	90.95%	
Weighted	91.85%	90.95%	91.26%	

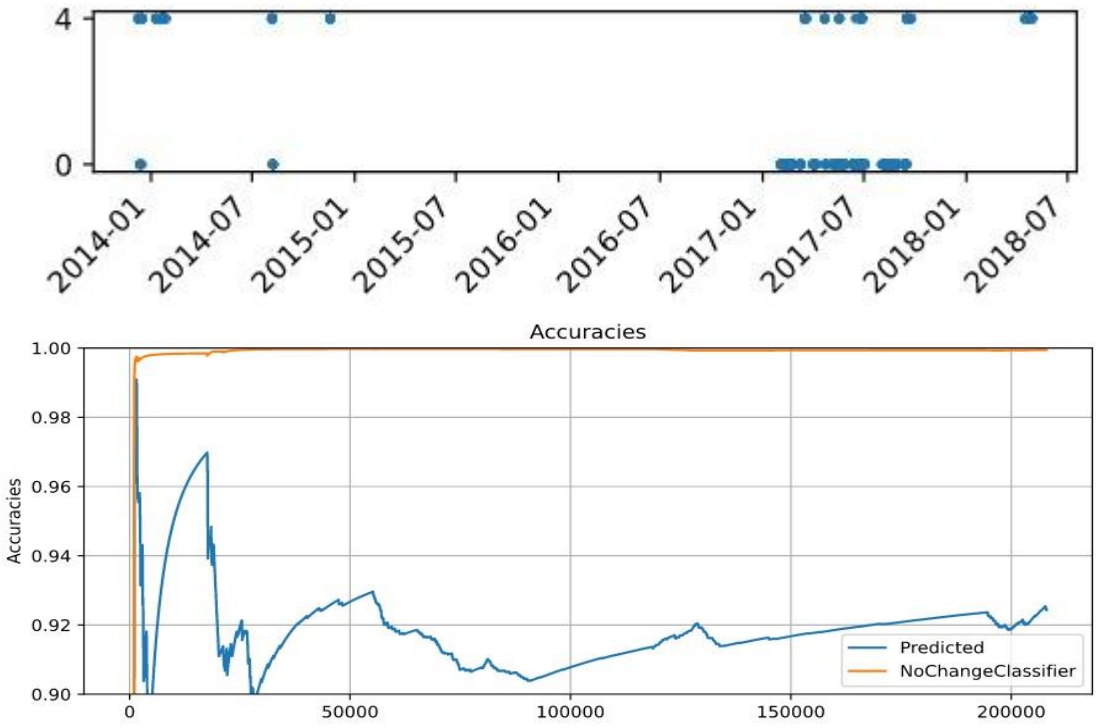
Evaluación

4. Experimentación y Resultados

Las etiquetas muestren una alta autocorrelación que, por otro lado, es muy frecuente en las series temporales.

Se compara la precisión del modelo con respecto al modelo de River dummy.NoChangeClassifier que simplemente asigna la etiqueta anterior al registro actual.

En la parte superior se muestra el cronograma de las series temporales reales utilizadas para el evento anómalo 4 (Flow Instability).

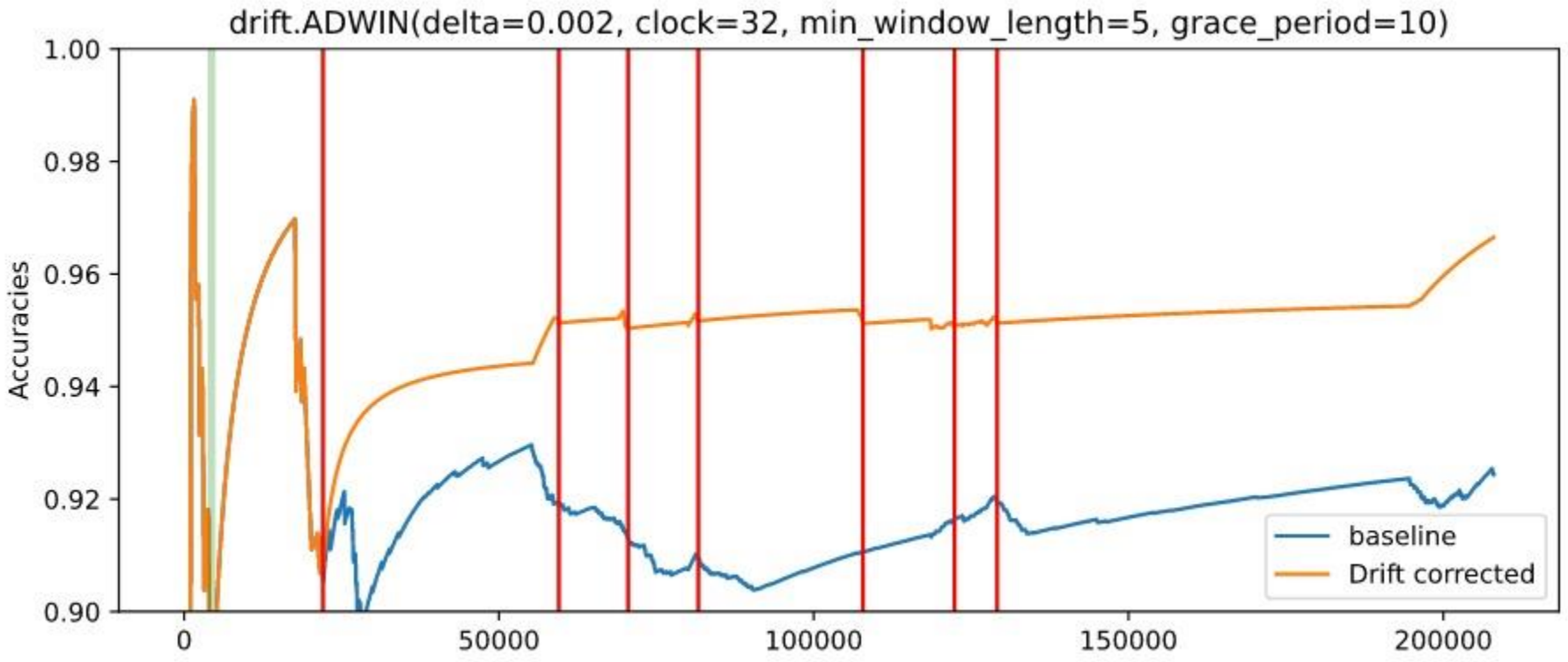


4. Experimentación y Resultados

Detección de drift

Aplicación del métodos de detección de drift ADWIN a las series reales con etiquetas 0 (Operación Normal) y 4 (Flujo inestable) ordenados por fecha.

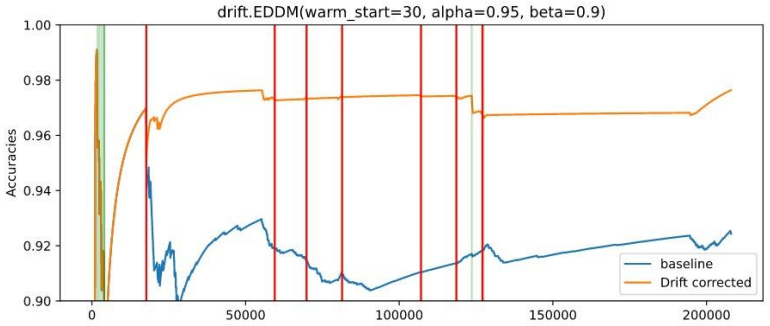
Se reinicializa el aprendizaje del modelo cuando se detecta drift y el modelo al menos ha visto 5000 registros.

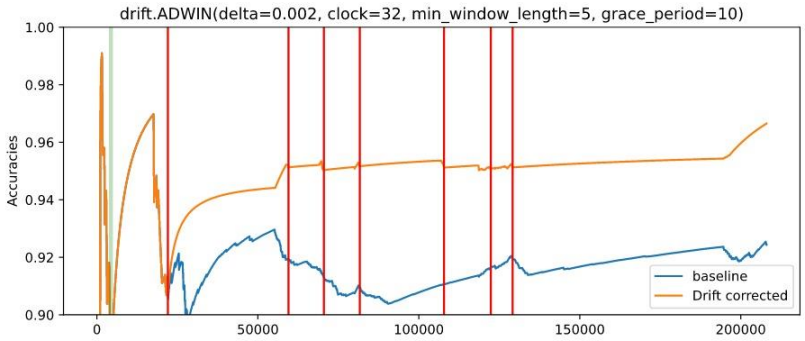


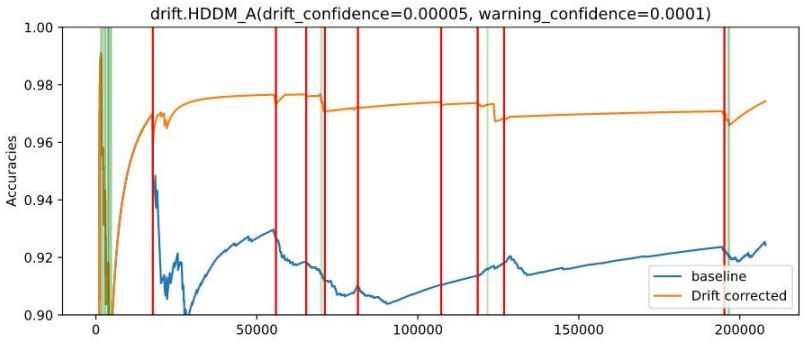
4. Experimentación y Resultados

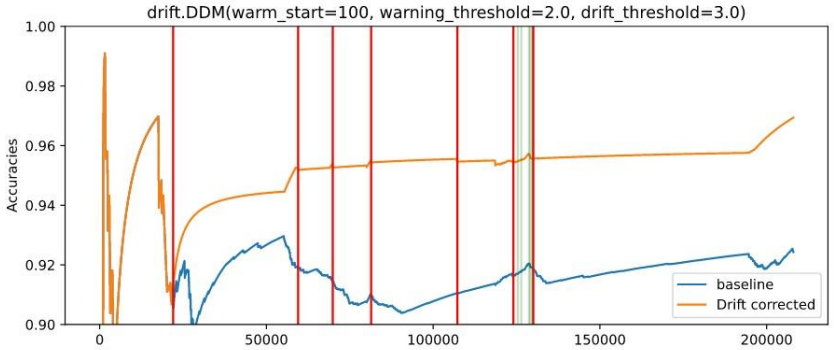
Detección de drift

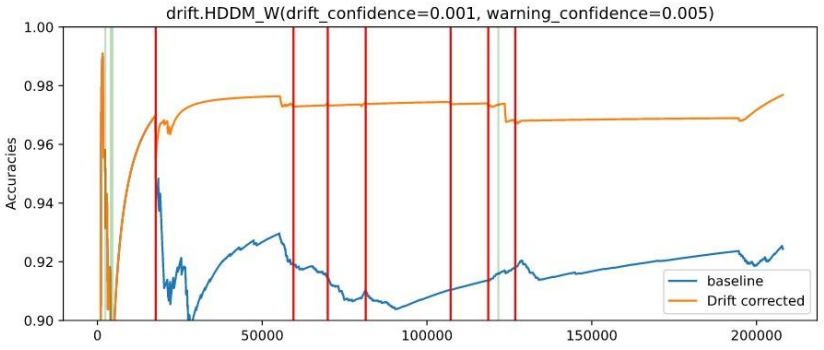
Se observa que la mayor parte de las detecciones ocurren en aproximadamente los mismos puntos y en todos los casos mejora la precisión.







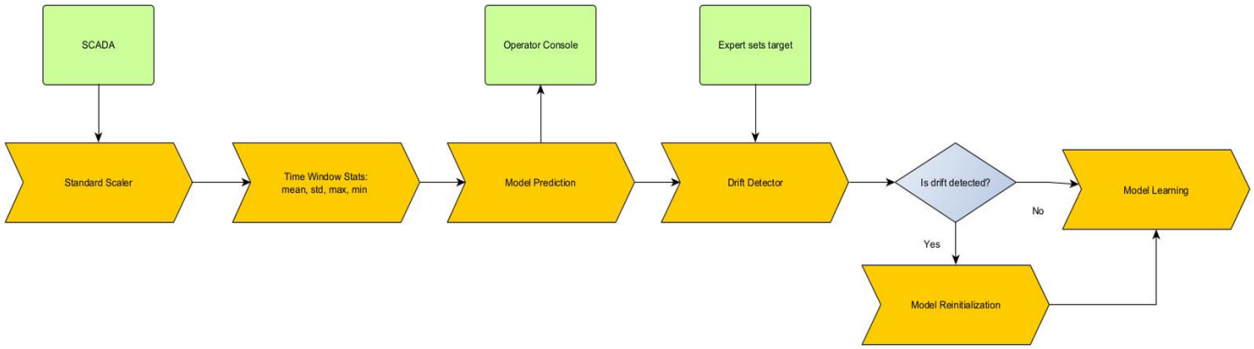




5. Conclusiones del TFM

De los estudios publicados hasta la fecha podemos concluir que es posible desarrollar una aplicación de soporte a la operación de los pozos de petróleo offshore que explote los modelos offline ajustados con el dataset 3W.

El enfoque novedoso de este TFM es la aplicación de técnicas de aprendizaje incremental y de detección de drift.



Se observa una clara mejora cuando se utilizan datos reales de dataset ordenados por fecha. Lamentablemente únicamente en el caso del evento anómalo Flow Instabilility se dispone de suficientes datos.

Una posible explicación del concept drift detectado podría ser que los expertos que han etiquetado las series hayan cambiado con los años o haya cambiado el criterio aplicado.

Aunque los métodos de detección drift han dado resultados similares, ADWIN es el más configurable y permite un ajuste más fino.

5. Conclusiones del TFM

- Se puede concluir que es posible realizar una detección en tiempo real fiable de eventos anómalos durante el proceso de extracción de crudo que era el objetivo principal del TFM.

- En cuanto al objetivo secundario de considerar de forma adecuada la presencia del fenómeno concept drift durante el proceso de aprendizaje, podemos afirmar que se consigue detectar aunque no tenemos evidencia de si es real o de los motivos que lo generan.

- Hay que resaltar que el uso de un proceso de aprendizaje incremental como el que se ha emulado sólo sería posible realmente si dispusiéramos de las etiquetas con un minuto de retraso. En realidad el etiquetado lo realiza un experto offline.

5. Conclusiones del TFM. 5.1 Líneas de trabajo futuras

- Ajustar un modelo offline con datos de los años 2014 y 2015 con un detector de drift aplicado sobre los datos de entrada al modelo que reportara si la predicción deja de ser fiable con datos de los años posteriores.

- El dataset 3W se actualiza periódicamente y, por tanto, se podrían actualizar los modelos o comprobar si siguen siendo válidos los modelos actuales.

- Crear una aplicación que detecte concept drift en los datos del SCADA del proceso de extracción de crudo y aísle series temporales que posteriormente un experto pueda etiquetar.

Universitat Oberta
de Catalunya

UOC.universitat
@UOCuniversidad
UOCuniversitat