Predicción de formación de hidratos en pozos de petróleo

Trabajo Final de Especialización

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Facultad de Ciencias Exactas y Naturales

Facultad de Ingeniería

Universidad de Buenos Aires

Celeste Savone

Octubre 2025

Contenido

Resumen: En las operaciones de producción de hidrocarburos ocurren eventos no deseados, entre ellos la formación de hidratos, con consecuencias sobre las instalaciones, las personas y el medio ambiente. El objetivo de este trabajo es predecir en tiempo real la formación de hidratos en las líneas de producción de pozos de petróleo, a partir de mediciones de presión y temperatura, mediante la aplicación de algoritmos de aprendizaje supervisado. Con datos reales de pozos offshore, la aplicación de técnicas de tratamiento de datos, ingeniería de atributos y utilizando *K-Nearest Neighbor, Random Forest*, y *redes neuronales* como algoritmos clasificadores, se diseñaron diferentes modelos predictivos multiclase para identificar la etapa previa, transitoria y estacionaria de formación de hidratos.

El ajuste de los modelos se realizó priorizando la detección temprana del evento asociado a la etapa transitoria y penalizando las falsas alarmas. Bajo esta premisa la red neuronal combinada con la técnica de *undersampling* fue el modelo con mejor rendimiento, pudiendo distinguir con alto desempeño la etapa normal de la etapa transitoria del evento. Ninguno de los modelos logró predecir correctamente la etapa estacionaria de formación de hidratos.

1. INTRODUCCIÓN

1.1. Marco teórico

Dentro de los numerosos eventos no deseados que pueden ocurrir en operaciones de extracción y transporte de hidrocarburos, la formación de hidratos despierta interés por sus consecuencias sobre la integridad de las instalaciones y la seguridad de las operaciones. Los hidratos de gas son sólidos cristalinos formados como resultado de la inclusión de una molécula en la red cristalina de otra. Su presencia genera obstrucciones, altera la reología del fluido, y bajo ciertas condiciones puede ocurrir la liberación incontrolada del gas presente. Para controlar su formación se recurre a métodos de reducción de contenido de agua, calentamiento y adición de inhibidores [1].

Por lo tanto, la predicción de su formación resulta importante para garantizar operaciones de extracción y transporte de hidrocarburos rentables y seguras.

El abordaje tradicional de esta temática se basa en el uso de correlaciones empíricas, diagramas de fase [2] o simuladores termodinámicos que permiten determinar las condiciones de formación de hidratos. Las variables de entrada varían según la complejidad del método, pero en general estas herramientas utilizan la presión, temperatura y gravedad específica o composición del gas para evaluar si las condiciones operativas del sistema se encuentran dentro de la zona de formación de hidratos.

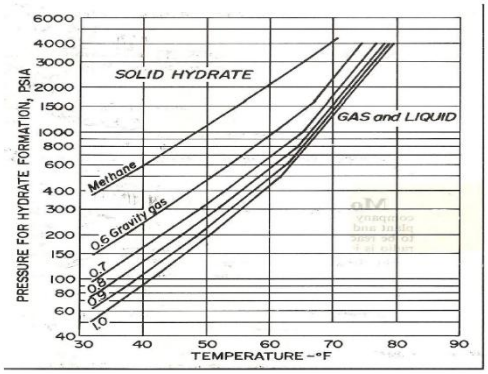


Fig. 1. Gráfico de curvas de formación de hidratos para diferentes composiciones de gas. Por encima de las curvas se encuentra la zona de formación de hidratos [5]

1.2. Objetivos

El objetivo del presente trabajo es predecir la formación de hidratos en la línea de producción de un pozo de petróleo, a partir de datos de presión y temperatura del sistema, clasificando la condición del pozo según la etapa de desarrollo del evento no deseado.

La presión y temperatura son variables habitualmente monitoreadas en pozos de petróleo y gas, por ser críticas para el control operativo. Dichas variables resultan además fundamentales en la predicción de la formación de hidratos [3].

En función de ello, se desarrollarán modelos que permitan la detección en tiempo real de condiciones asociadas a la formación de hidratos, diferenciando entre períodos sin formación (clase "Normal"), etapas iniciales de formación (clase "transitoria") y fases en las que la formación de hidratos está consolidada ("clase "falla")

El trabajo privilegiará la detección temprana, enfocándose especialmente en predecir la condición inicial de formación de hidratos (clase "transitoria"), ya que en esta fase es posible implementar estrategias de recuperación que eviten el desarrollo completo de la falla y las consecuencias asociadas.

2. METODOLOGÍA

2.1. Origen de los datos

Los datos utilizados en el presente trabajo fueron obtenidos de la base de datos 3W [4], de acceso público, la cual contiene información sobre eventos no deseados en operaciones de producción de petróleo y gas. Esta base proporciona registros provenientes de sensores ubicados en el pozo, en líneas submarinas y en la plataforma, así como como datos simulados o generados manualmente para representar eventos muy poco frecuentes [5].

Cada archivo o dataset (denominado *instancia* en el contexto de la base de datos 3W) corresponde a una serie de tiempo multivariable que abarca el período continuo de monitoreo. Para este trabajo, se analizaron únicamente aquellos datasets en los que se registró formación de hidratos y cuya fuente de información proviene directamente de sensores, con el objetivo de trabajar exclusivamente con datos reales.

Cada observación dentro de cada dataset está etiquetada para diferenciar la etapa de desarrollo del evento no deseado. Dicha etiqueta indica si la observación corresponde a una condición de operación normal, es decir sin formación de hidratos, a una condición inicial de formación de hidratos, o una condición estacionaria de formación de hidratos donde la falla se encuentra consolidada.

Para el presente trabajo se seleccionaron tres datasets que corresponden a diferentes períodos de monitoreo de un mismo pozo de la base de datos 3W, para los que se registró formación de hidratos en la línea de producción. Los datasets que contienen observaciones en los años 2012 y 2014 fueron usados para el entrenamiento de los modelos. En el dataset con observaciones del año 2015 es donde se realiza la predicción para la evaluación de la performance. Para reducir el tiempo de procesamiento se disminuyó la frecuencia de muestreo original de 1 segundo, a un registro cada 10 segundos.

Sobre esta base se desarrollaron modelos de aprendizaje supervisado para clasificación de la condición del pozo respecto a la formación de hidratos, a partir de mediciones de presión y temperatura en el pozo, en línea de producción y en la plataforma, diferenciándose tres

clases: condición normal del pozo sin presencia de anomalía, transitorio del evento de formación de hidratos que corresponde a la etapa inicial del evento no deseado, estado estacionario de formación de hidratos, donde la anomalía es consolidada.

Los modelos de clasificación multiclase fueron optimizados privilegiando la detección del estado transitorio, asumiendo que la detección temprana del evento permitirá aplicar medidas correctivas que eviten la consolidación de este o minimicen los costos de reacondicionamiento.

Para cumplir con el objetivo se realizó el preprocesamiento de datos, análisis exploratorio, ingeniería de atributos, y se desarrollaron experimentos con distintas estrategias de modelado y distintos clasificadores, para finalmente evaluar la performance de los modelos.

Todas las etapas fueron desarrolladas en código Python con uso de librerías soporte.

2.2. Preprocesamiento

2.2.1. Estructura inicial del dataset

Los tres datasets iniciales extraídos de la base de datos 3W (df_2012, df_2014 y df_2015), contienen variables obtenidas desde el sistema de instrumentación y etiquetas de clase, agregadas para cada observación en la etapa de etiquetado. A continuación, se muestra el tamaño original de cada dataset:

Dataset	Observaciones	Período de datos	Variables
df_2012	243137	2 days 19:32:16	29
df_2014	314604	3 days 15:23:23	29
df_2015	167599	1 days 22:33:18	29

Tabla 1. Estructura de datasets iniciales.
El "periodo de datos" indica la duración en tiempo del monitoreo continuo registrado en cada dataset

Las variables son de tipo *numéricas continuas* para representar mediciones de proceso como presión, temperatura y caudal, *numéricas discretas* para representar la posición de válvulas, y las etiquetas de clase de cada observación. Estas últimas son *variables categóricas representadas numéricamente*. Solo una de las etiquetas de clase será la variable objetivo y corresponde a aquella que indica la etapa de desarrollo de la formación de hidratos. La segunda etiqueta diferencia las maniobras operativas y será descartada del análisis.

Estas variables no presentan tratamiento previo por lo que existen datos faltantes, variables perdidas y variables con valores únicos (frozen variables).

2.2.2. Remuestreo

Se realizó el remuestreo de los datasets iniciales, para disminuir la frecuencia de muestreo a una observación cada 10 segundos, agrupando 10 observaciones consecutivas mediante el cálculo de la media para todas las variables, excepto para las etiquetas de clase, en las que se mantuvo el valor de la última observación del intervalo.

2.2.3. Limpieza de datos

a. Filtrado

Variables predictoras: teniendo en cuenta el objetivo del trabajo se realizó el filtrado de variables, seleccionando aquellas que corresponden a mediciones de presión y temperatura, obteniéndose 11 variables predictoras.

Variables de clase: cada observación presenta dos etiquetas de clase, una etiqueta asociada a la condición del pozo respecto a la ocurrencia de eventos anormales ('class') y una etiqueta relacionada con el estado operativo del pozo según las maniobras operativas ('state'). Esta segunda etiqueta fue descartada del análisis.

b. Datos faltantes

Se encontraron 4 variables con todos sus datos faltantes, probablemente sensores no activos, por lo que no se realizó imputación y fueron eliminadas del análisis.

También se registraron 3600 observaciones con etiqueta de clase faltante, correspondientes a la primera hora de cada dataset. Estas observaciones fueron eliminadas del análisis.

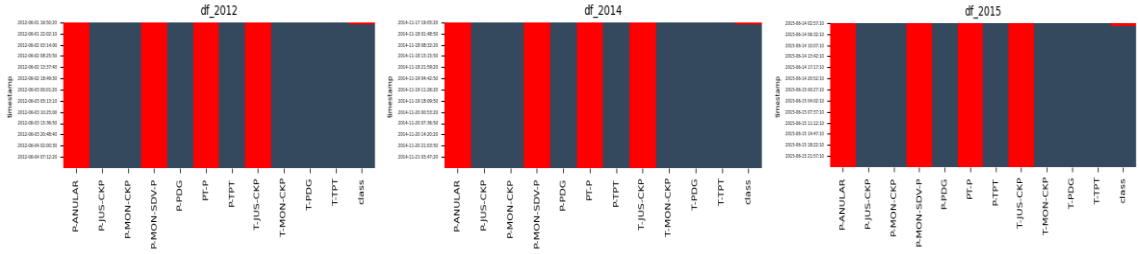


Figura 2. Visualización de datos faltantes en datasets

c. Variables con valores únicos

No se encontraron variables con valores únicos (frozen variables) en los datasets filtrados.

d. Conversión

Para mejorar la interpretación de los resultados se recodificó la variable de clase numérica 'class', por una variable categórica como se indica a continuación:

- Condición normal: "0" en dataset original, "Normal" en dataset preprocesado.
- Transitorio de anomalía: "108" en dataset original, "Transitorio" en dataset preprocesado.
- Anomalía consolidada: "8" en dataset original, "Falla" en dataset preprocesado.

También se realizó conversión de unidades de medias en las variables de presión (de Pa. a bar).

2.2.4. Datasets preprocesados

Luego del preprocesamiento se obtuvieron los datasets correspondientes a los años 2012, 2014, y 2015 con la estructura que se indica en la tabla 2.

Dataset preprocesado	**Observaciones**	**Período de datos**	**Variables**
df_2012	23954	2 days 18:32:10 (Normal: 1 days 04:33:00 Transitorio: 1 days 08:16:10 Falla: 0 days 05:42:40)	8
df_2014	31101	3 days 14:23:20 (Normal: 0 days 11:22:10 Transitorio: 3 days 00:43:40 Falla: 0 days 02:17:10)	8
df_2015	16401	1 days 21:33:20	8
Variable	**Descripción**		
P-TPT	Pressure at the TPT (temperature and pressure transducer) [Pa] (well sensor)		

T-TPT	Temperature at the TPT (temperature and pressure transducer) [°C] (well sensor)
P-PDG	Pressure at the PDG (permanent downhole gauge) [Pa] (well sensor)
T-PDG	Temperature at the PDG (permanent downhole gauge) [°C] (well sensor)
P-MON-CKP	Upstream pressure of the PCK (production choke) [Pa] (platform sensor)
T-MON-CKP	Upstream temperature of the PCK (production choke) [°C] (platform sensor)
P-JUS-CKP	Upstream pressure of the PCK (production choke) [Pa] (platform sensor)
class	Label of the observation

Tabla 2. Estructura de datasets preprocesados

2.2.5. Visualización de variables

En las figuras 2 y 3 se muestran las series temporales de las variables de presión y temperatura en los datasets preprocesados, diferenciando los períodos de operación normal, transitorio y falla en aquellos datasets que fueron utilizados como entrenamiento.

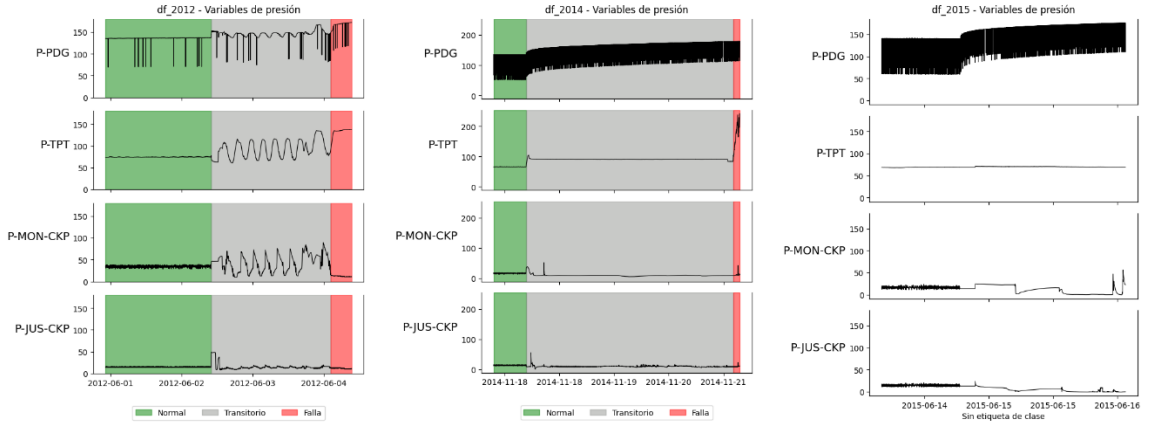


Figura 3. Variables de presión en datasets preprocesados

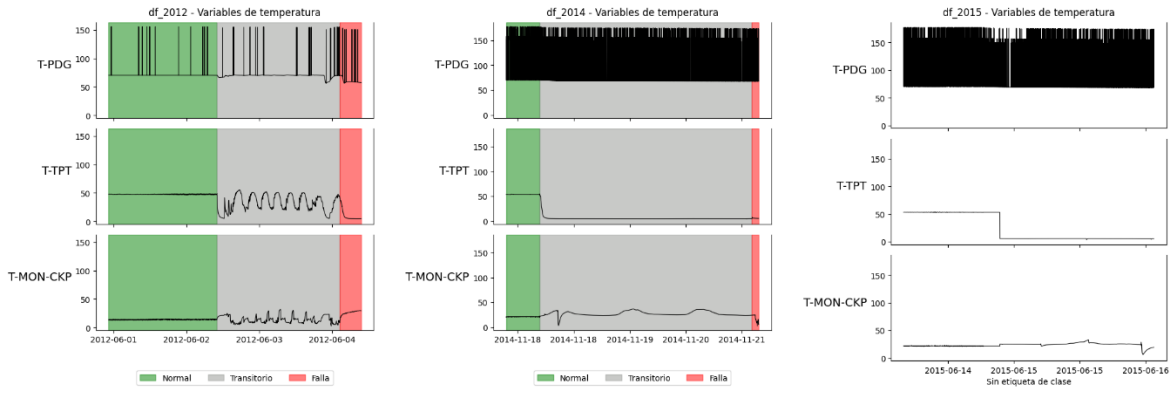


Figura 4. Variables de temperatura en datasets preprocesados

Las variables P-PDG y T-PDG presentaron una característica oscilatoria para los datasets de los años 2014 y 2015.

2.3. Análisis de Datos

2.3.1. Datos de entrenamiento y testeo

Para el desarrollo de los modelos predictivos se trabajó con un conjunto de datos de entrenamiento (df_train) obtenido a partir de las observaciones de los años 2012 y 2014 (df_2012, df_2014), y un conjunto de datos de prueba (df_test) obtenido a partir de las observaciones del año 2015 (df_2015).

Datasets preprocesado	Observaciones	Período de datos
df_train (df_2012 + df_2014)	55055	6 days 8:55:30
df_test (df_2015)	16401	1 day 21:33:20

Tabla 3. Estructura de conjunto de datos de entrenamiento y prueba

2.3.2. Análisis exploratorio de datos

El conjunto de datos de entrenamiento es desbalanceado, como se observa en la figura 4, donde la clase minoritaria es la condición de "Falla", y la clase mayoritaria es el estado "Transitorio".

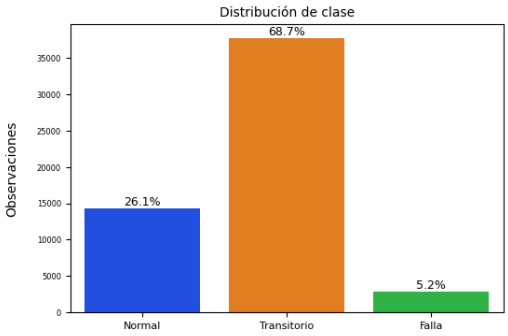


Figura 5. Distribución de clases en dataset de entrenamiento

En las figuras 5 y 6 se muestra la distribución de las variables de presión y temperatura agrupadas por clase para el conjunto de entrenamiento, y los diagramas de caja correspondientes.

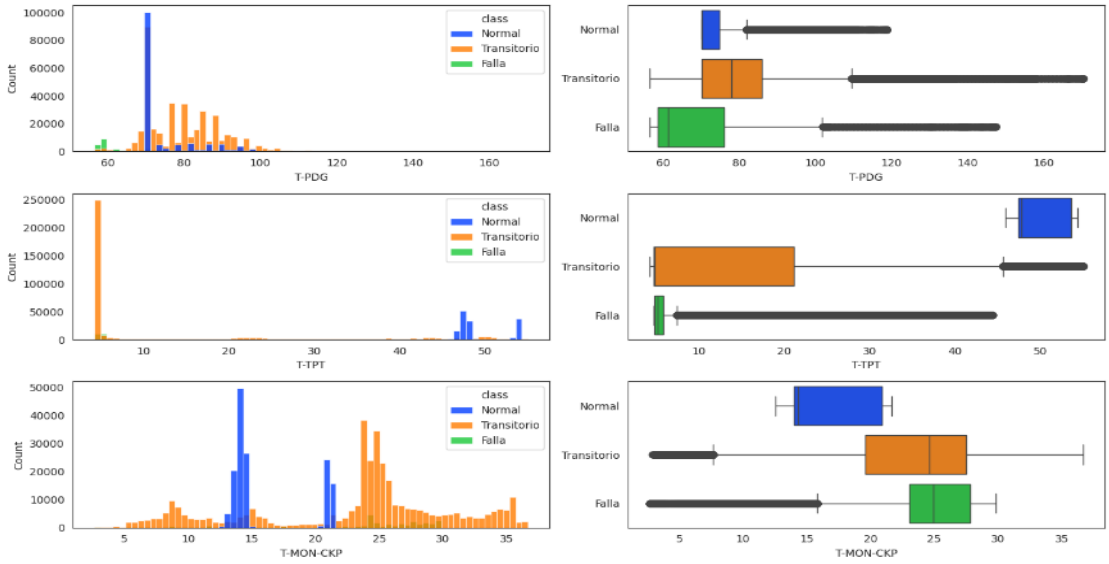


Figura 6. Distribución y diagramas de caja de variables de temperatura dataset de entrenamiento agrupadas por clase

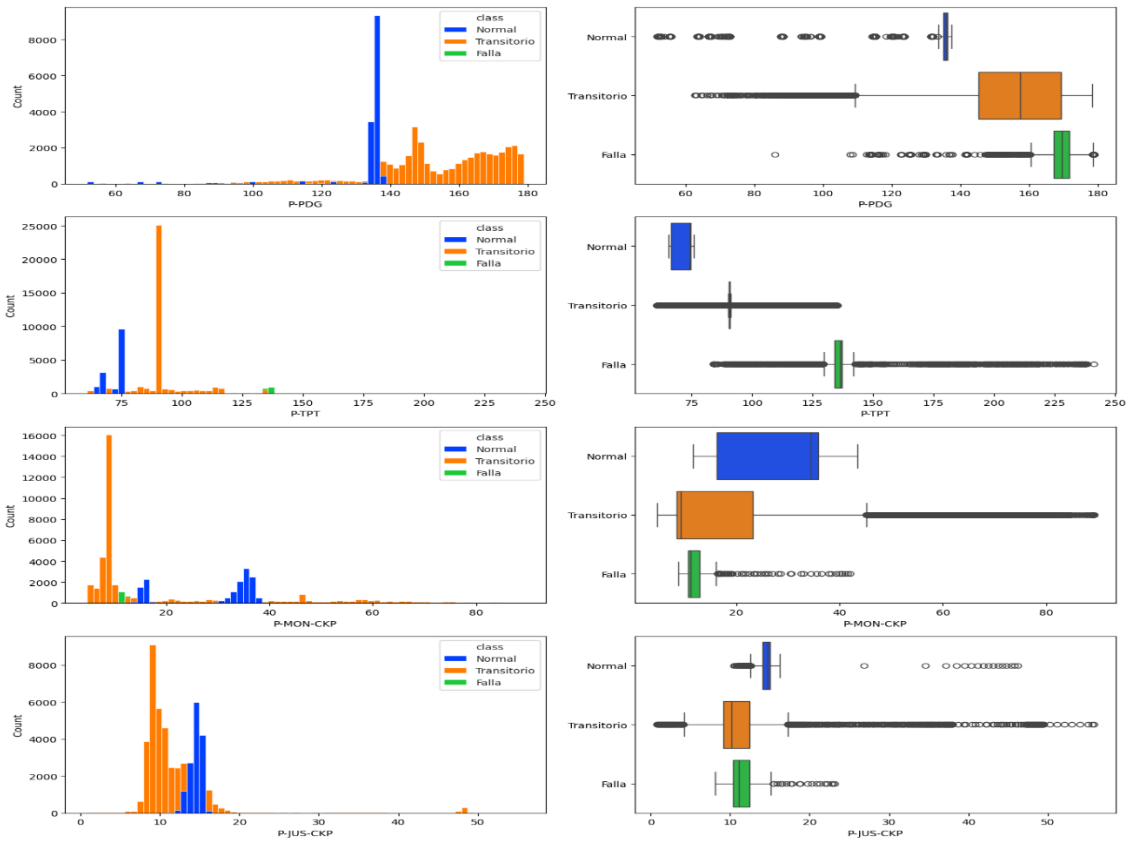


Figura 7. Distribución y diagramas de caja de variables de presión en dataset de entrenamiento agrupadas por clase

Pueden verse distribuciones en general sesgadas, con presencia de outliers.

La comparación de medianas como medida de tendencia central, muestra que existen diferencias de este parámetro en las variables según la clase. Para las variables “P-TPT” y ”P-PDG”, se observa un incremento en los valores de presión luego de iniciada la formación de hidratos (clase “Transitorio” y “Falla”). Estos sensores están ubicados aguas arriba (upstream) de la línea de producción por lo que el taponamiento por formación de hidratos podría explicar este comportamiento.

La comparación de los rangos intercuartílicos refleja diferencias en la dispersión de los datos de una misma variable según la clase.

Pueden observarse también distribuciones bimodales, como ocurre en la clase “Normal” en las variables “T-MON-CKP” y “P-MON-CKP”, pudiendo atribuirse este hecho a las diferencias de los valores centrales de estas variables en los datasets usados para la generación del conjunto de entrenamiento (df_2012 y df_2014).

Respecto de los valores atípicos, su presencia es marcada en todas las variables, especialmente para las clases “Transitorio” y “Normal”, con outliers moderados y severos.

En el conjunto de testeo, las figuras 7 y 8 muestran asimetría de las distribuciones de algunas variables, especialmente las variables de temperatura, con concentración de valores en más de una posición, lo que puede indicar agrupamiento. Se evidencia presencia de outliers en algunas variables.

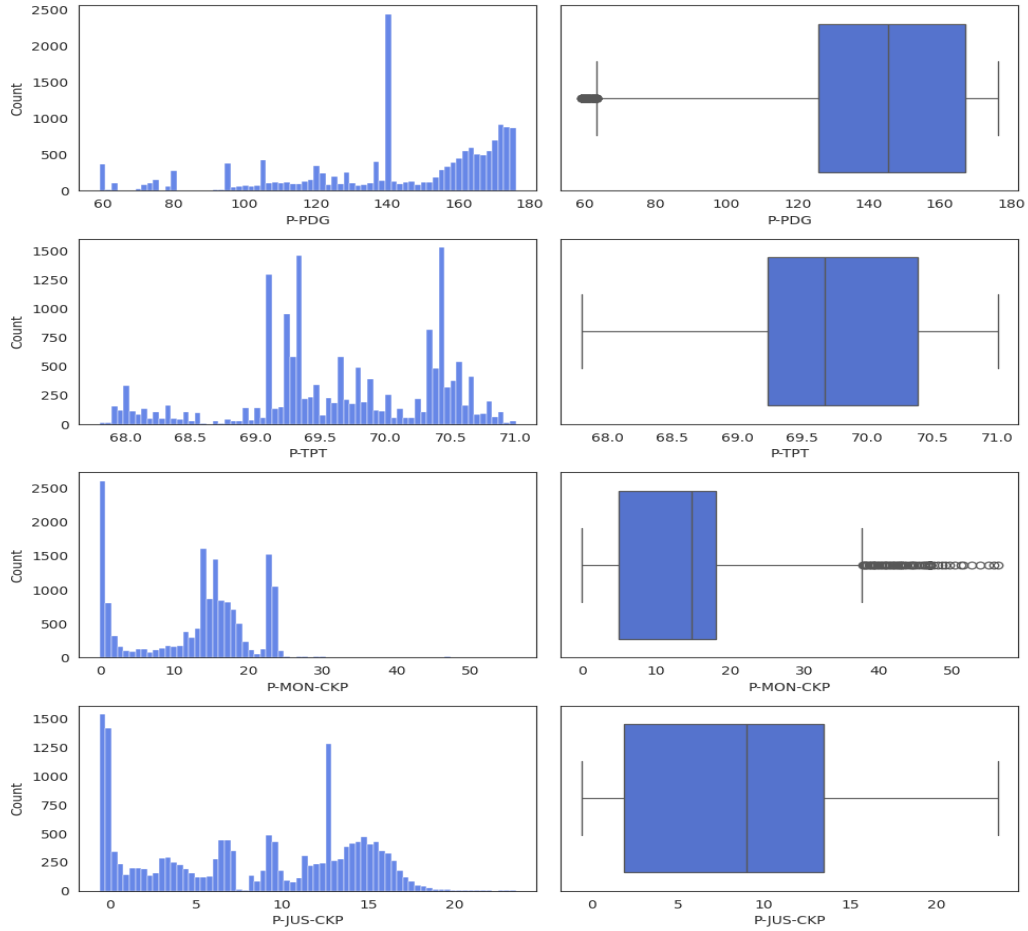


Figura 8. Distribución y diagramas de caja de variables de presión en dataset de prueba

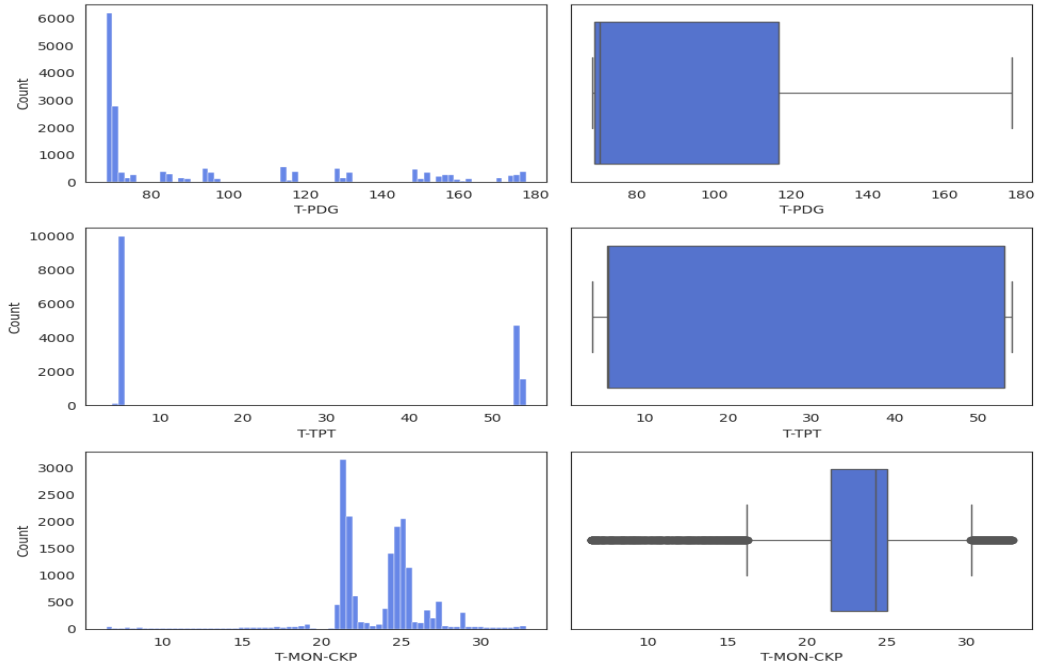


Figura 9. Distribución y diagramas de caja de variables de temperatura en dataset de prueba

2.3.3. Correlación de variables

Se evaluaron los coeficientes de correlación entre las variables numéricas del conjunto de entrenamiento, calculado con el método de Spearman.

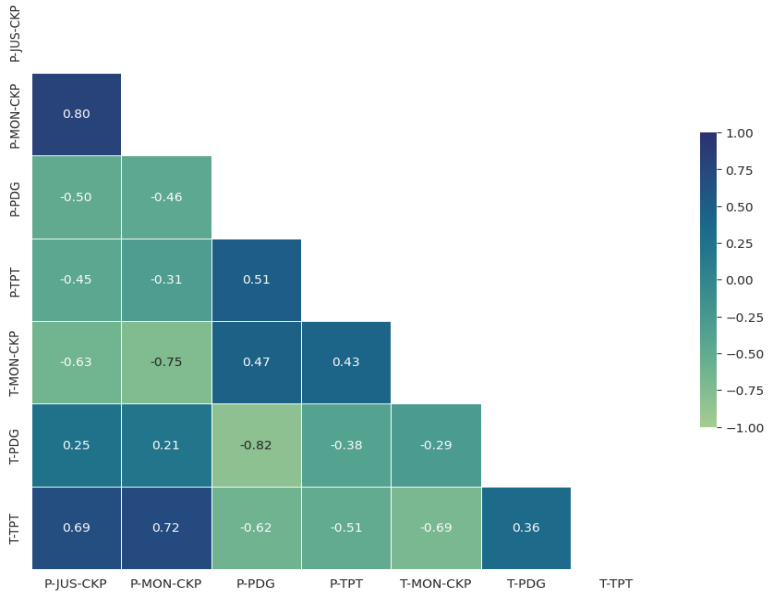


Figura 10. Diagrama de calor con coeficientes de correlación

Puede observarse en la figura 9 un índice de correlación alto positivo entre las variables de presión 'P-MON-CKP' y 'P-JUS-CKP', que corresponden a sensores de presión ubicados aguas arriba y abajo de la válvula de control de producción, y un coeficiente de correlación negativa elevado para las variables 'T-PDG' y 'P-PDG', o 'T-MON-CKP' y 'P-MON-CKP', que corresponden a medidas de presión y temperatura en un mismo punto del proceso.

2.4. Tratamiento de datos e ingeniería de atributos

Se describen a continuación las técnicas de tratamiento de datos e ingeniería de atributos utilizadas.

2.4.1. Suavizado (smoothing)

Para suavizar las oscilaciones de las variables 'P-PDG' y 'T-PDG' se utilizó la técnica de media móvil, con un tamaño de ventana de 1 minuto y 5 minutos con los efectos mostrados en la figura 10 y 11.

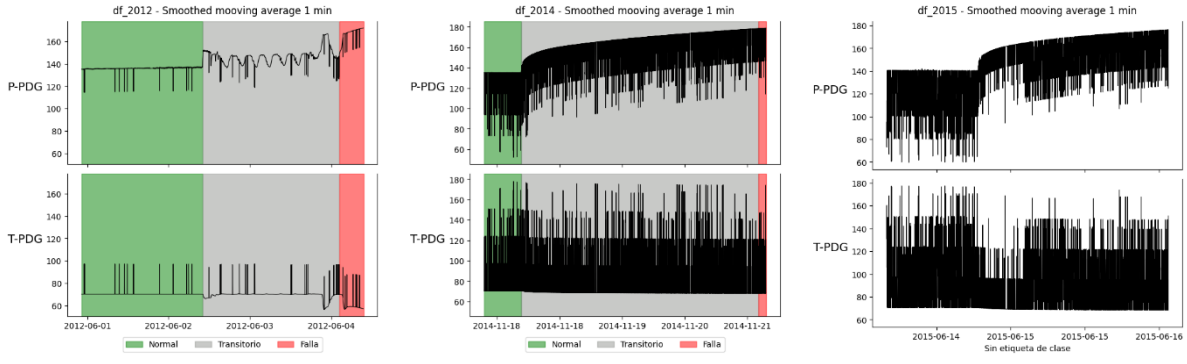


Figura 11. Efecto de suavizado por media móvil con ventana de 1 minuto

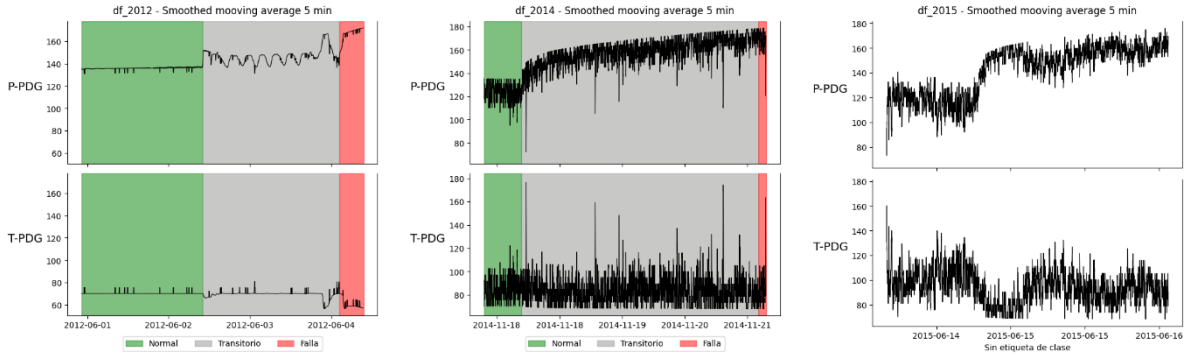


Figura 12. Efecto de suavizado por media móvil con ventana de 5 minutos

2.4.2. Variables históricas

Se generaron dos nuevas variables por cada variable predictora mediante la técnica de ventana deslizante, calculando el rango y el cociente entre valor final e inicial, en una ventana de 1 minuto. De esta forma se obtuvieron 14 nuevas variables predictoras con información histórica, que se agregaron a las 7 variables predictoras de los datasets preprocesados.

2.4.3. Undersampling

Teniendo en cuenta el desbalanceo de clases en el conjunto de datos de entrenamiento, se implementó la técnica de undersampling, sobre la clase mayoritaria con el objeto de evaluar su impacto en la clasificación de las clases minoritarias. Se resampleó la clase mayoritaria (clase “Transitorio”) con un número de observaciones igual al de la segunda minoría del dataset (clase “Normal”).

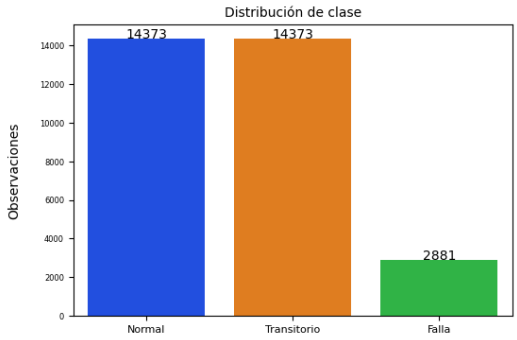


Figura 13. Distribución de clases con undersampling en dataset de entrenamiento

2.5. Experimentos y modelado

Se desarrollaron 12 experimentos obtenidos por combinación de 3 algoritmos de clasificación con 4 estrategias de modelado. Los experimentos fueron agrupados según el clasificador utilizado.

En todos los casos el dataset de entrenamiento corresponde a las observaciones del año 2012 y 2014. El dataset de testeo corresponde a las observaciones del año 2015.

2.5.1. Estrategias de modelado

Las estrategias de modelado describen el tratamiento de datos aplicado, definiendo el conjunto de datos sobre el que se aplicaron los algoritmos clasificadores.

- Base: se aplicaron los algoritmos clasificadores sobre los datasets preprocesados, indicados en el ítem 2.4, sin ninguna estrategia adicional.
- Smoothing: se aplicaron los algoritmos clasificadores sobre los datasets obtenidos luego de aplicar un filtro de media móvil para una ventana de 1 minuto, sobre las variables P-PDG y T-PDG, como se indicó en el ítem 4.1. El resto de las variables se mantuvo sin cambios.
- Variables históricas: se aplicaron los algoritmos clasificadores sobre los dataset obtenidos luego de agregar las 14 nuevas variables de rango y cociente en una ventana móvil de 1 minuto, como se indicó en el ítem 4.2.
- Undersampling: se aplicaron los algoritmos clasificadores sobre los dataset obtenidos luego de resamplear la clase mayoritaria en el dataset de entrenamiento, como se indicó en el ítem 4.3.

2.5.2. Métricas de performance

Para evaluar el rendimiento de los modelos se utilizó una función ganancia que privilegia el acierto de la clase 'Transitorio' y penaliza los falsos positivos para esta clase. Dicha función es una estimación del ahorro en días de producción que implicaría para una empresa productora, la detección temprana de la formación de hidratos ('Transitorio') asumiendo que en esta etapa puede revertirse tal condición, y estima también las pérdidas que implican tomar acciones correctivas ante "falsas alarmas". También se reportarán la Tasa de Falsos Positivos, la Recuperación o Recall, y el F1-score para la clase "Transitorio", complementando la evaluación.

- Ganancia = 15 * TP – 2* FP
- TFP _T (tasa de falsos positivos para la clase "Transitorio") = $\frac{FP}{FP+TN}$
- Recall_T (tasa de verdaderos positivos para la clase "Transitorio") = $\frac{TP}{TP+FN}$
- F1-score_T = $\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$

Siendo TP, TN, FP, y FN los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos para la clase "Transitorio", respectivamente.

2.5.3. Algoritmos de clasificación

Se emplearon 3 algoritmos de clasificación:

- **K-Nearest Neighbor (KNN) [6]:** para este algoritmo clasificador se optimizó el número de vecinos cercanos (parámetro K o n_neighbors) en cada experimento, previo escalado de los datos, utilizando validación cruzada de 5-folds para evaluar la ganancia del modelo. El espacio de búsqueda del parámetro se indica en la tabla 4.

KNN			
Parámetro	**Inferior**	**Superior**	**Iteraciones**
k (n_neighbors)	3	40	10

Tabla 4. Espacio de búsqueda de parámetros en KNN

- **Random Forest (RF) [7]:** en cada experimento de este clasificador se optimizó el número de árboles de decisión individuales ('n_estimators'), la máxima profundidad del árbol ('max_depth'), y el número mínimo de muestras necesarias para dividir un nodo interno ('min_samples_split'). Se aplicó una búsqueda Bayesiana de hiperparámetros con 30 iteraciones en el espacio de búsqueda que se muestra en la tabla 5, con validación cruzada de 5-folds para evaluar la ganancia del modelo.

Random Forest			
Parámetro	**Inferior**	**Superior**	**Iteraciones**
n_estimators	10	150	
max_depth	5	50	30
min_samples_split	2	100	

Tabla 5. Espacio de búsqueda de hiperparámetros en RF

- **Red neuronal artificial (ANN) [8]:** se modeló una red neuronal con una capa oculta, previo escalado y codificación de la clase (one-hot encoding). Se optimizó el número de neuronas ('units') y la función de activación, además del tamaño del lote ('batch_size'), mediante la evaluación de 10 combinaciones de parámetros en el espacio de búsqueda que se indica en la tabla 6. Del conjunto de entrenamiento se separó un 20% de los datos, en forma estratificada, para la validación del modelo y se usaron 10 ciclos completos de entrenamiento ('epoch').

ANN		
Parámetro	**Valores**	**Iteraciones**
units	[20,30, 40, 50]	
activation	['relu', 'sigmoid']	10
batch_size	[16,32,64]	

Tabla 6. Espacio de búsqueda de parámetros en ANN

3. RESULTADOS

En cada experimento los parámetros óptimos son los que maximizan la función ganancia indicada en el punto 5 y las métricas que se reportan son las evaluadas en el conjunto de datos de prueba (df_test). En la figura 13 se muestra una línea de tiempo con la clasificación real del conjunto de datos de prueba, para mejor interpretación de los resultados.

Figura 14. Línea de tiempo para la clase verdadera en el conjunto de prueba (df_test)

Los experimentos se compararon también con las métricas obtenidas por un modelo que predice la clase mayoritaria para todas las observaciones ('Lazy model'), es decir como "Transitorio".

	Lazy_model
Ganancia	153297
TFP_T	1,000
Recall_T	1,000
F1-score_T	0,801

Tabla 7. Métricas de performance en modelo que predice todas las observaciones como "Transitorio"

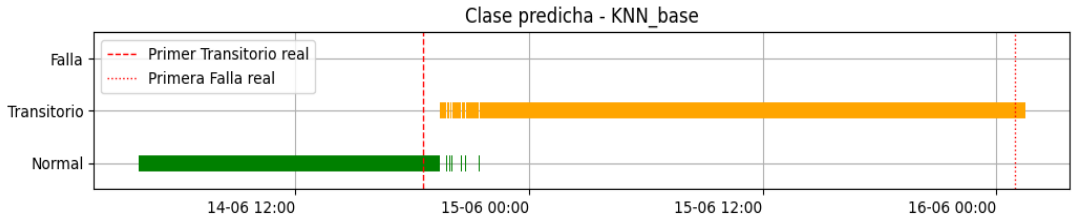
3.1. Experimentos KNN

Dentro de los experimentos KNN la estrategia de undersampling es la que logró la mejor performance, mientras que el agregado de variables históricas arrojó las peores métricas, con una ganancia inferior a la del Lazy model.

	KNN_base	KNN_smoothing	KNN_variables_históricas	KNN_undersampling
K_óptimo	35	33	15	40
Ganancia	156834,0	158004,0	146261	**158724**
TFP_T	0,036	0,036	0,039	**0,036**
Recall_T	0,958	0,965	0,893	**0,969**
F1-score_T	0,969	0,973	0,934	0,975

Tabla 8. Parámetros óptimos y métricas de performance en experimentos KNN

Si bien todos los experimentos tienen una alta recuperación de los "Transitorios" (Recall superior al 89%), confunden dicha clase con la "Normal", principalmente al inicio de la etapa transitoria, lo que tiende a retrasar la detección de los transitorios. Quien menos confusión presenta es KNN_undersampling.
Ninguno de los modelos logra predecir la clase "Falla", confundiéndola en todos los casos con la clase "Transitorio" (excepto KNN_variables_históricas que logró algún acierto), lo que impacta en la tasa de falsos positivos.
La clase "Normal" es perfectamente predicha por los experimentos, con alguna confusión únicamente en KNN_variables_históricas.



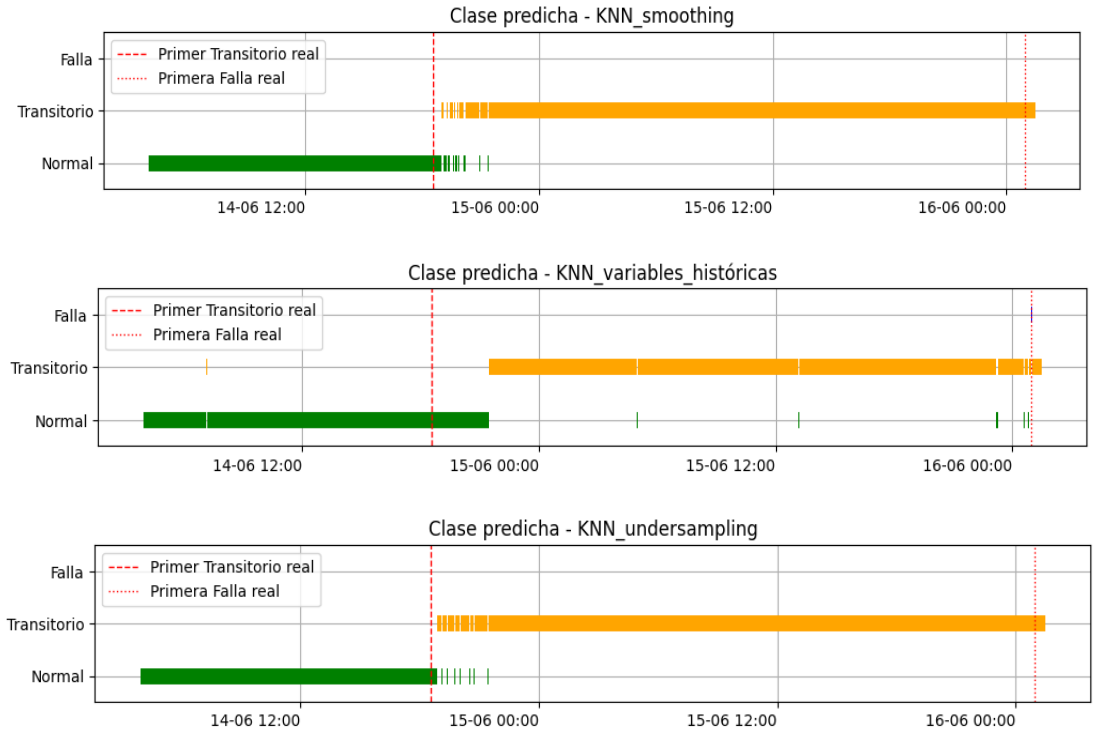


Figura 15. Líneas de tiempo para la clase predicha por experimentos KNN

3.2. Experimentos RF

EL experimento RF_base no supera la ganancia del Lazy model. Las estrategias para RF_smoothing, RF_variables_históricas y RF_undersampling logran mejorar las ganancias respecto al RF_base, y superar la ganancia del Lazy Model. La estrategia de undersampling es la que mostró mejor perfomance del clasificador.

Todos los experimentos RF presentan una elevada tasa de falsos positivos (entre el 48% y 74%) siendo el experimento RF_base el que mayor cantidad de falsas alarmas presentó. Puede verse en la figura 15 la elevada confusión entre clase “Normal” y “Transitorio” en la etapa normal, con falsas alarmas tempranas teniendo en cuenta la ocurrencia del primer transitorio. En este sentido su performance se asemeja a la del ‘Lazy model’, inclinando la predicción hacia la clase mayoritaria. Ninguno de los experimentos RF logra captar la clase “Falla”, prediciéndola como “Transitorio”.

	RF_base	**RF_smoothing**	**RF_variables_históricas**	**RF_undersampling**
n_estimators_óptimo	39	32	38	76
max_depth_óptimo	47	43	5	39
min_samples_split_óptimo	24	53	100	98
Ganancia	152939	154522	155640	**155673**
TFP_T	0,739	0,697	0,487	**0,482**
Recall_T	0,980	0,987	0,980	**0,980**
F1-score_T	0,835	0,846	0,883	**0,883**

Tabla 9. Parámetros óptimos y métricas de performance en experimentos RF

Figura 16. Líneas de tiempo para la clase predicha por experimentos RF

En particular para los experimentos RF se obtuvo la importancia de variables en la clasificación. Puede verse en la Tabla 10 que las variables que miden presión en el pozo ('P-PDG' y 'P-TPT') son las dos variables más importantes en todos los experimentos. Las variables agregadas en RF_variables_históricas tienen menor importancia en la clasificación que las variables originales y no se muestran en la tabla.

Importancia de variables en RF			
RF_base	RF_smoothing	RF_variables_históricas	RF_undersampling
P-PDG	P-PDG	P-TPT	P-PDG
P-TPT	P-TPT	P-PDG	P-TPT
T-TPT	T-TPT	T-TPT	T-PDG
P-JUS-CKP	P-JUS-CKP	T-MON-CKP	T-TPT
T-PDG	T-PDG	P-JUS-CKP	T-MON-CKP
P-MON-CKP	P-MON-CKP	T-PDG	P-MON-CKP
T-MON-CKP	T-MON-CKP	P-MON-CKP	P-JUS-CKP

Tabla 10. Variables ordenadas por orden de importancia en experimentos RF

3.3. Experimentos ANN

Todos los experimentos ANN superaron la performance del Lazy_model, obteniéndose la mayor ganancia con ANN_undersampling. La estrategia en el experimento ANN_variables_históricas no mejora la ganancia del experimento ANN_base, aunque es el experimento que menor tasa de falsos positivos presenta, lo que minimiza las falsas alarmas.

	ANN_base	ANN_smoothing	ANN_variables_históricas	ANN_undersampling
units_óptimo	30	30	20	50
activation_óptimo	relu	relu	relu	relu
batch_size_óptimo	16	16	16	32
Ganancia	159714,0	160834,0	156755	**162177**
TFP_T	0,039	0,049	**0,037**	0,076
Recall_T	0,975	0,983	0,957	0,993
F1-score_T	0,978	0,979	0,969	0,978

Tabla 11. Parámetros óptimos y métricas de performance en experimentos ANN

Los experimentos ANN presentaron confusión entre la clase “Normal” y “Transitorio”, al inicio de la etapa transitoria. En el experimento ANN_undersampling y ANN_smoothing la confusión entre dichas clases también se produce durante la etapa normal, generando falsas alarmas tempranas.
Ninguno de los experimentos predice correctamente la clase “Falla”, confundiéndola con la clase “Transitorio”.

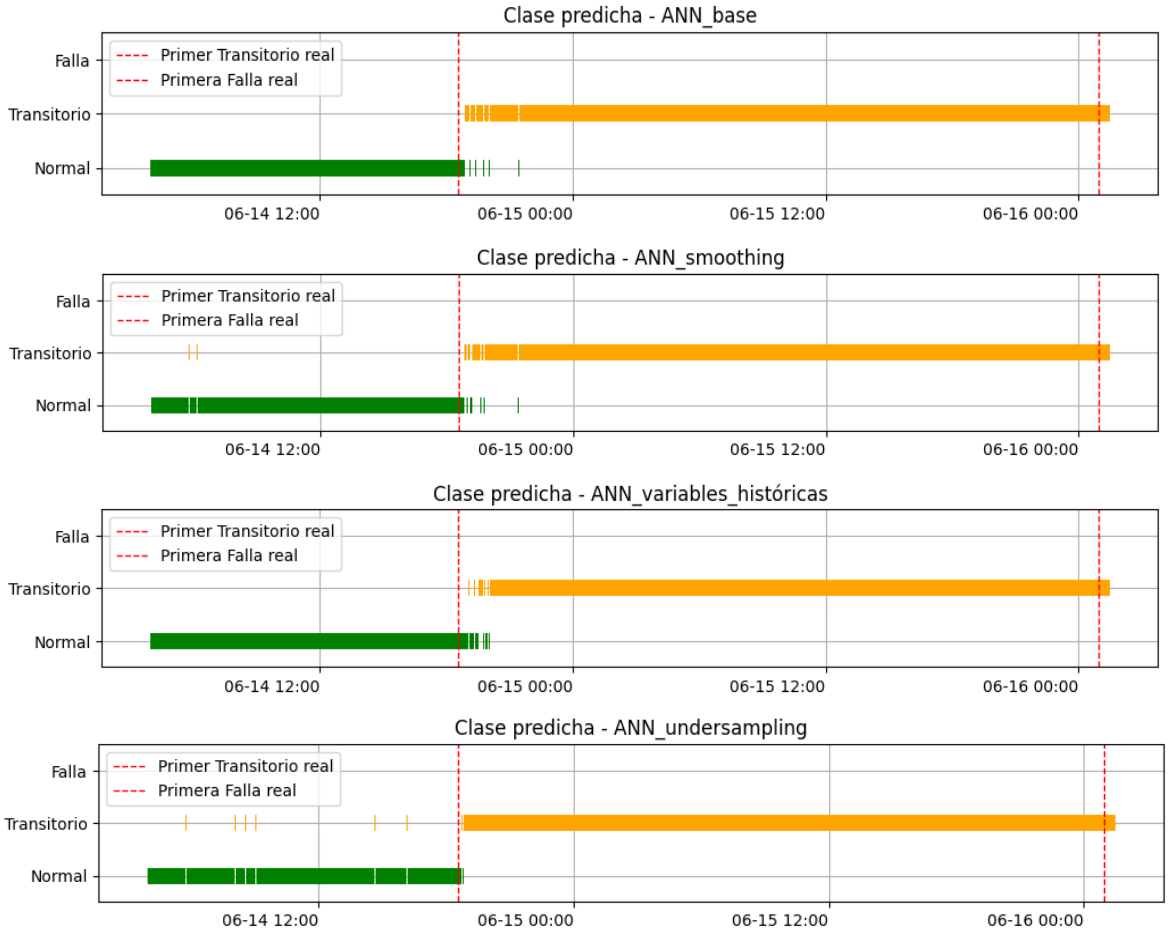


Figura 17. Líneas de tiempo para la clase predicha por experimentos ANN

4. CONCLUSIONES

Los experimentos presentan diferente performance según el clasificador y la estrategia de modelado. Ninguno de los modelos logra predecir correctamente la etapa consolidada de la formación de hidratos ("Falla"), confundiéndola en todos los casos con la etapa transitoria, por lo que la diferencia en la performance de los modelos se debe a las recuperaciones de la clase "Normal" y "Transitorio" en cada uno.

Los experimentos con K-Nearest Neighbor como clasificador presentaron las menores tasas de falsos positivos para la clase "Transitorio", con excelente desempeño para captar la clase "Normal", a costa de una menor recuperación de la clase "Transitorio".

Los experimentos con Random Forest como clasificador obtuvieron las mejores tasas de recuperación de la clase "Transitorio" a costa de una mayor tasa de falsos positivos por confusión entre las clases "Normal" y "Transitorio".

Los experimentos con redes neuronales lograron un mejor balance entre la recuperación de la clase "Transitorio", y la tasa de falsos positivos, arrojando las mejores ganancias para cada estrategia de modelado.

Respecto del impacto de las estrategias de tratamiento de datos sobre la performance de los modelos dentro de un grupo de experimentos, la estrategia de smoothing y undersampling lograron mejorar el desempeño de los experimentos respecto del experimento base, siendo undersampling la que mejores resultados arrojó. El efecto de agregar variables históricas no fue el mismo para todos los clasificadores, mejorando la performance solo para Random Forest respecto al experimento base.

La mejor performance para detectar los "Transitorios" sin incurrir en falsas alarmas, la arrojó, en consecuencia, el modelo que combinó la estrategia de undersampling con una red neuronal como clasificador, con una tasa de falsos positivos de 0,08, una recuperación del 0,99 y F1-score de 0,98, métricas calculadas para la clase "Transitorio".

De esta forma puede concluirse que, con el clasificador y la estrategia de tratamiento de datos adecuados, puede predecirse la formación de hidratos en pozos de petróleo de forma online, pudiendo distinguirse con alto desempeño la etapa normal de la etapa transitoria del evento. Los límites de decisión entre las etapas con formación de hidratos ('Transitorio' y 'Falla') no lograron ser captados por ningún modelo.

Debe tenerse presente que todas las predicciones están basadas en datos reales crudos por lo que su tratamiento es un paso determinante en la performance de los modelos. Queda para futuras revisiones analizar el impacto de aplicar otras estrategias como tratamiento de outliers, eliminación de variables correlacionadas o técnicas más agresivas para las variables oscilatorias o ruidosas.

Adicionalmente, teniendo en cuenta las particularidades de cada clase y sus límites de decisión, es interesante evaluar a futuro la performance de un experimento de ensamblado de modelos o incluso plantear otras estrategias como clasificación binaria o one-vs-rest.

5. REFERENCIAS

[1] Makwashi, N., & Ahmed, T. (2021). Gas hydrate formation: Impact on oil and gas production and prevention strategies. *Nigerian Research Journal of Engineering and Environmental Sciences, 6*(1), 61–75. https://doi.org/10.5281/zenodo.5047631

[2] Janna, F. (2015). *Correlation for predicting hydrate formation* (B.Eng. thesis). Universiti Teknologi PETRONAS. https://utpedia.utp.edu.my/Dissertation_14689

[3] Sun, W., Wei, N., Zhao, J., Zhou, S., Zhang, L., Li, Q., Jiang, L., Zhang, Y., Li, H., Xu, H., Li, C., Shen, X., & Xiong, C. (2021). Wellbore temperature and pressure field in deep-water drilling and the applications in prediction of hydrate formation region. *Frontiers in Energy Research, 9*, Article 696392. https://doi.org/10.3389/fenrg.2021.696392

[4] Petrobras. (2023). *3W* [GitHub repository]. https://github.com/petrobras/3W

[5] Vargas, R. E. V., Munaro, C. J., Ciarelli, P. M., Medeiros, A. G., do Amaral, B. G., Barrionuevo, D. C., de Araújo, J. C. D., Ribeiro, J. L., & Magalhães, L. P. (2019). A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering, 181*, 106223. https://doi.org/10.1016/j.petrol.2019.106223

[6] Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13 (1), 21-27,* https://doi.org/10.1109/TIT.1967.105396

[7] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:101093340432

[8] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

[9] Tariq, Z., Aljawad, M. S., Hasan, A., Murtaza, M., Mohammed, E., El-Husseiny, A., Alarifi, S. A., Mahmoud, M., & Abdulraheem, A. (2021). A systematic review of data science and machine learning applications to the oil and gas industry. *Journal of Petroleum Exploration and Production Technology, 11*, 4339–4374. https://doi.org/10.1007/s13202-021-01302-2

[10] Turan, E. M., & Jäschke, J. (2021). Classification of undesirable events in oil well operation. In *Proceedings of the 23rd International Conference on Process Control (PC)* (pp. 157–162). IEEE. https://doi.org/10.1109/PC52310.2021.9447527

[11] Villamil, R. H. (2024, November 10–14). Assessment of deep learning techniques for anomaly detection in offshore oil wells. In *Proceedings of the 20th Brazilian Congress of Thermal Sciences and Engineering (ENCIT 2024)*. https://doi.org/10.26678/ABCM.ENCIT2024.CIT24-0603