UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Estimación de las propiedades PVT del aceite con el uso de algoritmos de aprendizaje automático

TESIS

Que para obtener el título de

Ingeniero Petrolero

P R E S E N T A

José Enrique Leal Castillo

DIRECTOR DE TESIS

Dr. Rodolfo Gabriel Camacho Velázquez

Ciudad Universitaria, Cd. Mx., 2021

Este trabajo fue realizado en colaboración con Ilse Abril Vázquez Sánchez, egresada de Ingeniería en Computación

Agradecimientos

A Dios, por darme la oportunidad de tener la vida que tengo y guiarme en el camino siempre con el bien de su luz.

A mis padres, Alejandro y Silvana, por todo el apoyo que me han dado a lo largo de la carrera y en general de mi vida. Por ser mi ejemplo de vida y por inculcarme los valores fundamentales.

A mi hermana, por llenar mi vida de alegría y ser la mejor compañía a lo largo de mi infancia.

A Abril, por ser complemento de mi vida, mi motivación y darme ese amor incondicional desde que la conocí.

Al Dr. Camacho, por creer en mí y darme la oportunidad de desarrollar el trabajo de investigación con él. Por darme la oportunidad de adquirir nuevos conocimientos siempre con su supervisión y atención a dudas.

A mis sinodales por el tiempo brindando con el objetivo de mejorar el trabajo escrito elaborado y por la dedicación firme en sus labores como docentes de la carrera.

Al Dr. Samaniego, por siempre ser un ejemplo para mi de lo que quiero lograr ser, un hombre de bien y muy dedicado. Por incluirme en el proyecto Petrobowl y por haberme brindado siempre a mi y a los competidores su apoyo, hecho que se vio reflejado en nuestro título mundial.

A mis profesores por inculcarme los conocimientos que hoy tengo de la carrera. Por ser en su mayoría experiencias positivas en mi desarrollo.

A mis amigos por haberme acompañado, por haberme extendido la mano en momentos de necesidad dentro y fuera de la universidad.

Índice general

Introducción

Conocer el comportamiento de los fluidos petroleros es de gran importancia para la industria petrolera, específicamente para Ingeniería de Yacimientos e Ingeniería de Producción. Cálculos de balance de materia, análisis de pruebas de presión, estimados de reservas, simulación numérica de yacimientos y diseño de sistemas de producción superficiales entre otros son dependientes de una correcta y precisa estimación de las propiedades PVT de dichos fluidos.

Las propiedades PVT de los fluidos petroleros pueden ser determinadas de diferentes maneras, siendo las más comunes las correlaciones, constantes de equilibrio y por su puesto pruebas de laboratorio, las cuales son económicamente costosas y ofrecen en ocasiones resultados dudosos.

Un campo relativamente nuevo de la computación llamado aprendizaje automático o "Machine Learning" se ha comenzado a utilizar en diversas ramas de la ingeniería para resolver problemas de forma más eficiente El aprendizaje automático es una rama de la inteligencia artificial que permite que las máquinas "aprendan" sin ser específicamente programadas para ello. Una cualidad indispensable para hacer que los modelos creados sean capaces de identificar patrones entre los datos para hacer predicciones y estimaciones.

En este trabajo se propone el uso de Técnicas de Aprendizaje Automático (TAA) o "Machine Learning", para clasificar los fluidos de yacimiento en aceites negros o volátiles haciendo uso de la información a condiciones de yacimiento conocida, así como de propiedades físicas elementales en superficie. Además, se propone el uso de regresiones lineales, redes neuronales y máquinas de soporte de vectores para obtener una estimación confiable de las diferentes propiedades PVT de aceites en yacimientos de México.

Para ello, se hará uso de reportes de pruebas PVT para entrenar y validar los modelos. Los datos de entrada a los modelos serán parámetros de los fluidos que no requieren de alguna prueba especializada para poder conocerse como lo son la gravedad API, RGA, viscosidad del fluido a presión y temperatura atmosférica, así como la temperatura, presión y profundidad del yacimiento petrolero.

El objetivo es obtener con los modelos creados y con la mayor precisión posible, las curvas que describen las diferentes propiedades PVT del aceite conforme varía la presión a condiciones de temperatura de yacimiento: presión de saturación, factor de volumen del aceite, densidad y viscosidad.

Capítulo 1

Conceptos Fundamentales de las Propiedades PVT

La abreviatura PVT tiene como significado presión, volumen y temperatura. Es utilizada en la industria petrolera para denotar los cambios que tienen las propiedades de los fluidos cuando cambian las condiciones físicas del medio en el que se encuentran.

La variación de las propiedades de un fluido con el cambio de las condiciones físicas del medio depende principalmente de la composición del mismo. Una estimación precisa de las propiedades físicas de los fluidos es de gran importancia en los diversos campos de la Ingeniería Petrolera.

1.1. Definición de las Propiedades PVT del Aceite

1.1.1. Densidad del Aceite ρ_o

Ahmed (2006) La densidad del petróleo crudo se define como la masa que tiene una unidad de volumen a condiciones de presión y temperatura designadas. Generalmente se expresa con la unidad de kilogramo sobre metro cubico. $[kg/m^3]$.

La densidad del aceite puede ser reportada a condiciones atmosféricas (ρ_{osc}), de tanque de almacenamiento (ρ_{oSTB}), de separador (ρ_{osep}), de presión de burbuja (ρ_{ob}) y/o de yacimiento (ρ_{oy})

Si el valor de ρ_{ob} es conocido, se puede estimar el valor de ρ_{oy} en un yacimiento bajo saturado (presión por arriba del punto de burbuja) con la siguiente expresión:

$$\rho_{oy} = \rho_{ob} \ exp \ [C_o \ (p_y - p_b)] \tag{1.1}$$

Donde:
p_y = presión de yacimiento $[psi]$
p_b = presión de burbuja $[psi]$
C_o = compresibilidad isotérmica del aceite $[psi^{-1}]$

1.1.2. Densidad Relativa del Aceite ρ_{ro}

Schlumberger (2020). También llamada gravedad específica del aceite, se define como la densidad del aceite dividida entre la del agua, ambas generalmente medidas a condiciones estándar (60°/60°).

$$\rho_{ro} = \frac{\rho_o}{\rho_w} \tag{1.2}$$

Donde:
ρ_{ro} = densidad relativa del aceite [1]
ρ_o = densidad del aceite $[lb/ft^3]$
ρ_w = densidad del agua $[lb/ft^3]$

$$\rho_{ro} = \frac{\rho_o}{62{,}4} \ , 60°/60° \tag{1.3}$$

La **figura 1.1** muestra la curva típica del comportamiento de la densidad relativa del aceite en función de la presión a temperatura constante (Ej. temperatura de yacimiento). Iniciando a una presión por arriba de la presión de burbuja, el aceite se ubica en la región bajo saturada por lo que solo existe una sola fase en el sistema (aceite + gas disuelto). A medida que la presión se reduce, el volumen de aceite aumenta y por su concepto, la densidad del aceite se reduce. En la p_b, el aceite llega a su valor mínimo de densidad (ρ_{rob}). Mientras se sigue reduciendo la presión por debajo de la P_b (región saturada), la densidad del aceite aumenta a medida que se libera el gas en solución y se pierden los elementos más ligeros. Cuando se llega a valores cercanos de la presión atmosférica, el valor de (ρ_{ro}) se encuentra en su valor máximo debido a que han liberado sus elementos más ligeros. (gas en solución).

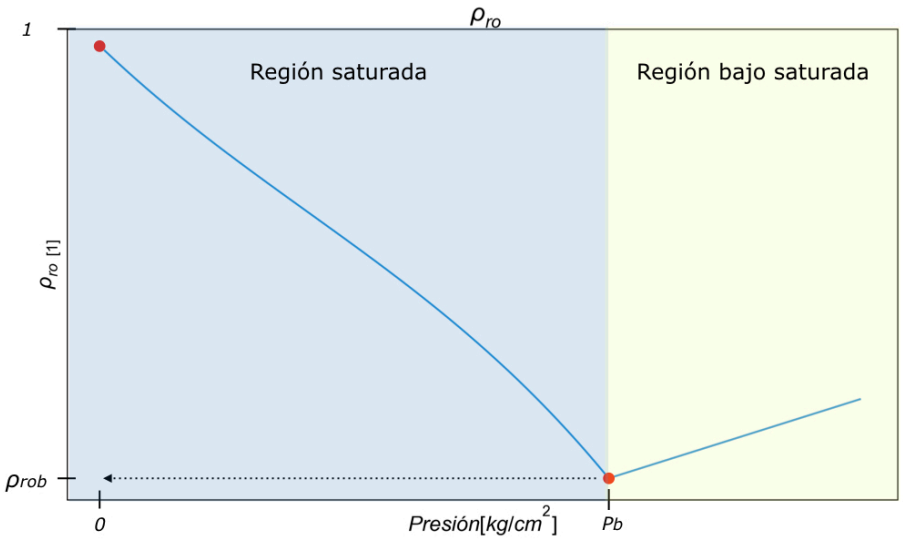

Figura 1.1: *Ahmed (2006)* Comportamiento de la densidad relativa del aceite (ρ_{ro}) vs Presión.

1.1.3. Gravedad API

Ahmed (2006) Es la escala preferida por la industria petrolera para indicar la densidad de los fluidos producidos y fue desarrollada de forma arbitraria por el Instituto Estadounidense del Petróleo (American Petroleum Institute, API). Se relaciona con la densidad relativa mediante la siguiente expresión:

$$^{\circ}API = \frac{141{,}5}{\rho_{ro}} - 131{,}5 \tag{1.4}$$

La gravedad API de los crudos suele oscilar entre los 47° API para los crudos más ligeros y los 10° API o menos para los crudos asfálticos más pesados.

1.1.4. Relación de Solubilidad Gas – Aceite R_s

Se define como la relación que existe entre el volumen de gas disuelto y el volumen de petróleo crudo a presión y temperatura estándar. Las unidades de campo más comunes para esta propiedad son $[scf/STB]$ y $[m^3/m^3]$.

$$R_s = \frac{V_{gd}}{(V_o)_{sc}} \tag{1.5}$$

Donde:
R_s = relación de solubilidad gas aceite $[scf/STB]$
V_{gd} = volumen de gas disuelto $[scf]$
$(V_o)_{sc}$ = volumen de aceite a condiciones estándar $[STB]$

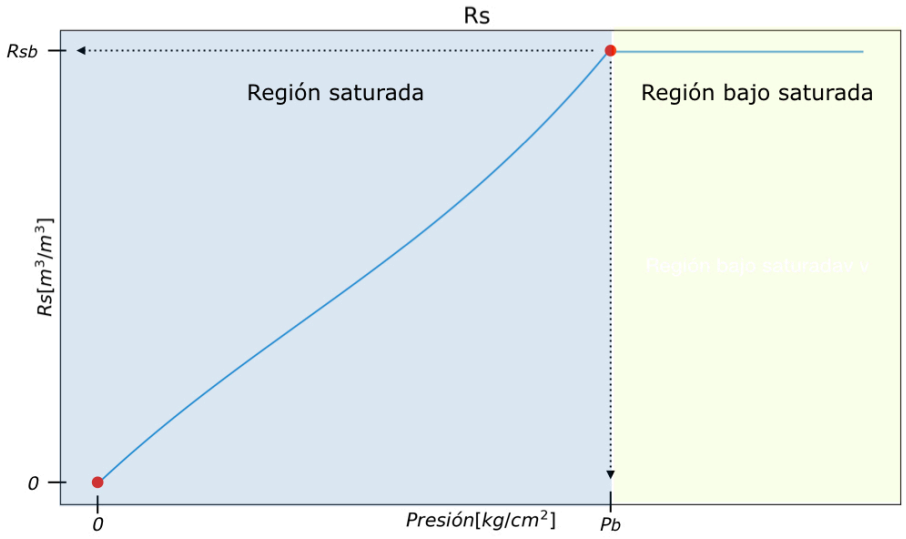

Figura 1.2: *Ahmed (2006)* Comportamiento de la relación gas - aceite (R_s) vs Presión.

La **figura 1.2** describe la curva típica del comportamiento de la relación de solubilidad gas – aceite con el cambio de presión en un sistema a temperatura constante. A una presión mayor que la presión de burbuja, el aceite se localiza en la región bajo saturada y el gas se mantiene completamente disuelto en el aceite, causando un valor de R_s máximo y constante hasta alcanzar la P_b. La P_b es el punto de inicio para el descenso de la cantidad de gas disuelto debido a que se libera la primera burbuja de gas en solución. Al seguir reduciendo la presión por debajo de la P_b (región saturada), la R_s decrece a medida que se libera el gas en solución. Cuando se llega a valores cercanos de la presión atmosférica, el valor de R_s tiende a cero debido a que la mayoría del gas en solución se ha liberado.

1.1.5. Presión de Burbuja P_b

También llamada presión de saturación, en un sistema de hidrocarburos se define como la presión más alta a la que la primera burbuja de gas se libera del petróleo a condiciones de temperatura de yacimiento. El punto de burbuja se puede medir experimentalmente realizando una prueba PVT de expansión a composición constante.

Es de gran importancia conocer el valor de otras propiedades PVT evaluadas a la presión de burbuja ya que representa un punto de inflexión en las curvas que describen sus comportamientos (ej. R_{sb}, B_{ob}, μ_{ob}).

1.1.6. Factor de Volumen del Aceite B_o

Se define como la relación existente entre el volumen de aceite (más el gas en solución) a ciertas condiciones de pozo y el volumen de aceite a condiciones estándar. Por definición general, el factor de volumen del aceite es siempre mayor o igual que uno. Es reportado comúnmente a condiciones presión de burbuja y de yacimiento (B_{ob}, B_{oy}).

El factor de volumen del aceite se expresa matemáticamente como:

$$B_o = \frac{(V_o)_{p,T}}{(V_o)_{sc}} \tag{1.6}$$

Donde:
B_o = factor de volumen del aceite $[bbl/STB]$
$(V_o)_{p,T}$ = volumen de aceite a la condición de presión y temperatura seleccionada. $[bbl]$
$(V_o)_{sc}$ = volumen de aceite a condiciones estándar $[STB]$

Si se conoce el valor de B_{ob}, y se desea estimar el valor de B_{oy} en un yacimiento bajo saturado (presión por arriba del punto de burbuja), se puede hacer uso de la siguiente expresión:

$$B_{oy} = B_{ob}\ exp\ [-C_o\ (p_y - p_b)] \tag{1.7}$$

Donde:
p_y = presión de yacimiento $[psi]$
p_b = presión de burbuja $[psi]$

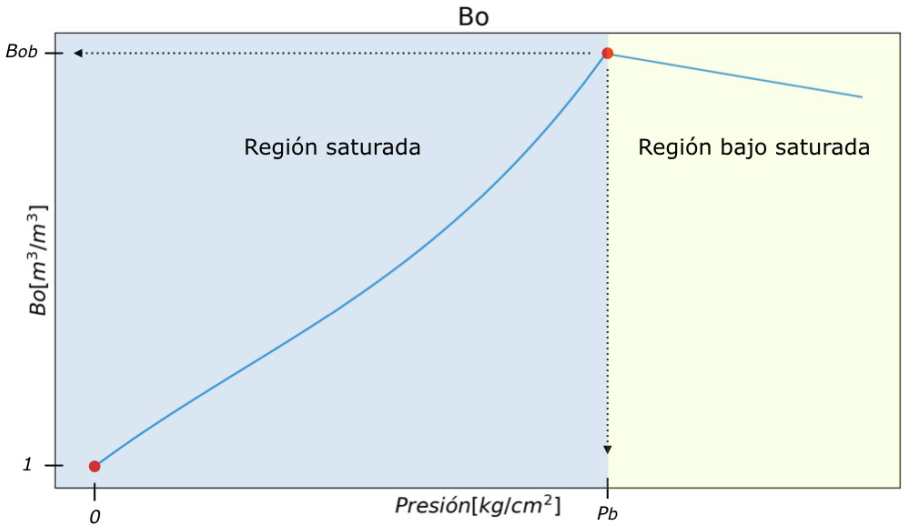

Figura 1.3: *Ahmed (2006)* Comportamiento del factor de volumen del aceite (B_o) vs Presión.

La **figura 1.3** muestra la curva típica del comportamiento del factor de volumen del aceite en función de la presión a temperatura constante. Comenzando a una presión por arriba de la presión de burbuja, el aceite se encuentra en la región bajo saturada por lo que solo existe una sola fase en el sistema. A medida que la presión se reduce, el volumen de aceite aumenta debido a la expansión del aceite, dando como resultado un aumento en el factor de volumen del aceite que continuará hasta que se alcance la presión de burbuja. En la P_b, el aceite llega a su valor máximo de expansión (B_{ob}). A medida que la presión se sigue reduciendo por debajo de la P_b (región saturada), el volumen de aceite disminuye a medida que se libera el gas en solución. Cuando se llega a valores cercanos de la presión atmosférica, el valor de el B_o tiende a uno.

1.1.7. Compresibilidad Isotérmica del Aceite C_o

Los coeficientes de compresibilidad isotérmica son necesarios para resolver muchos problemas de ingeniería de yacimientos, incluidos los problemas de flujo y también son necesarios para determinar las propiedades físicas del petróleo crudo bajo saturado.

Para un sistema de petróleo crudo bajo saturado (presiones por encima del punto de burbuja), el coeficiente de compresibilidad isotérmica del aceite se define mediante una de las siguientes expresiones equivalentes:

$$C_o = \left[\frac{-1}{V}\ \frac{\partial V}{\partial p}\right]_T \tag{1.8}$$

$$C_o = \left[\frac{-1}{B_o}\ \frac{\partial B_o}{\partial p}\right]_T \tag{1.9}$$

$$C_o = \left[\frac{-1}{\rho_o}\ \frac{\partial \rho_o}{\partial p}.\right]_T \tag{1.10}$$

Donde:
C_o = compresibilidad isotérmica del aceite $[psi^{-1}]$
ρ_o = densidad del aceite $[lb/ft^3]$
B_o = factor de volumen del aceite $[bbl/STB]$

Para un sistema de crudo saturado (presiones por debajo del punto de burbuja), la compresibilidad del aceite se define como:

$$C_o = \frac{-1}{B_o}\ \frac{\partial B_o}{\partial p}\ +\ \frac{B_g}{B_o}\ \frac{\partial R_s}{\partial p} \tag{1.11}$$

Donde:
B_g = factor de volumen del gas $[bbl/scf]$

1.1.8. Factor de Volumen Total B_t

Se define como la relación existente entre el volumen total de la mezcla de hidrocarburos (Petróleo y gas presente) evaluado a la condición de interés por cada unidad de volumen de aceite en el tanque de almacenamiento.

Esta propiedad engloba el volumen total de un sistema independientemente del número de fases presentes. Matemáticamente se define como:

$$B_t = \frac{(V_o)_{p,T} + (V_g)_{p,T}}{(V_o)_{sc}} \tag{1.12}$$

Donde:

$(V_o)_{p,T}$ = volumen de aceite a la condición de presión y temperatura seleccionada. $[bbl]$

$(V_g)_{p,T}$ = volumen de gas libre a la condición de presión y temperatura seleccionada. $[bbl]$

$(V_o)_{sc}$ = volumen de aceite a condiciones estándar $[STB]$

1.1.9. Viscosidad del Aceite μ_o

La viscosidad del petróleo crudo es una propiedad física importante que controla su flujo través de medios porosos y tuberías. La viscosidad, en general, se define como la resistencia interna del fluido a fluir.

Según las condiciones presión y temperatura, la viscosidad de los crudos se puede clasificar en tres categorías:

- **Viscosidad del aceite muerto** (μ_{od}) Se define como la viscosidad del petróleo crudo observada a presión atmosférica (sin gas en solución) y la temperatura de separación.
- **Viscosidad del aceite saturado** (μ_{ob}) Se define como la viscosidad del petróleo crudo registrada debajo de la presión de burbuja y temperatura del yacimiento.
- **Viscosidad del aceite bajo saturado** (μ_{ou}) Se define como la viscosidad del petróleo crudo medida a una presión por arriba del punto de burbuja (ej. presión de yacimiento bajo saturado) y la temperatura de yacimiento.

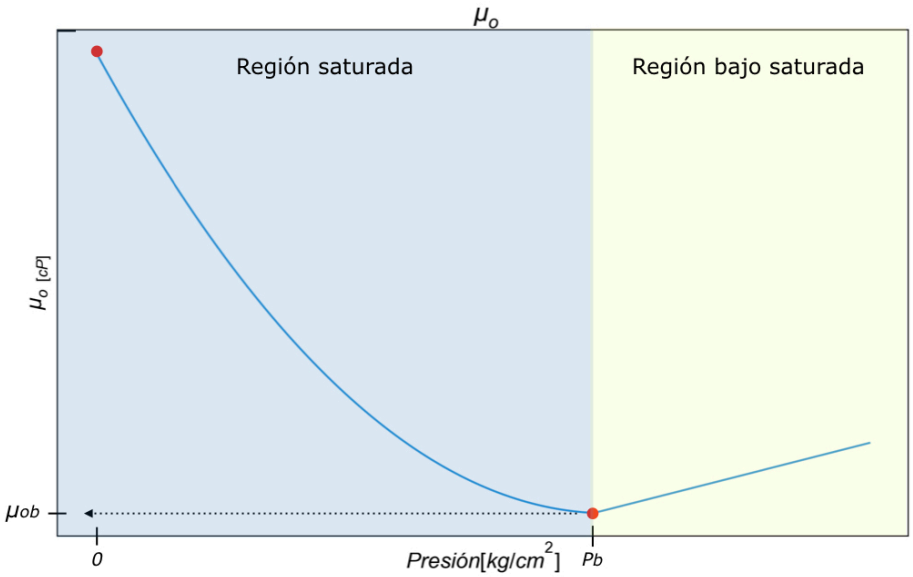

Figura 1.4: *Ahmed (2006)* Comportamiento de la viscosidad del aceite (μ_o) vs Presión.

La **figura 1.4** describe la curva típica del comportamiento de la viscosidad del aceite en función de la presión a temperatura constante. Iniciando a una presión por arriba de la presión de burbuja, el aceite se ubica en la región bajo saturada. Al irse reduciendo la presión, la viscosidad se reduce debido al efecto de expansión del aceite. En la pb, el aceite llega a su valor mínimo de Viscosidad(μ_{ob}). Continuando con la reducción de la presión por debajo de la pb (región saturada), la viscosidad del aceite aumenta a medida que se libera el gas en solución y se pierden los elementos más ligeros (bajos en viscosidad). Cuando se llega a valores cercanos de la presión atmosférica, el valor de viscosidad se encuentra en su valor máximo debido a que han liberado los elementos menos viscosos.

1.2. Definición de las Propiedades PVT del Gas

1.2.1. Peso Molecular Aparente M_a

Es la suma de los pesos moleculares de los componentes de una mezcla de gas multiplicados por su fracción molar. Esta importante propiedad del gas se define matemáticamente por la siguiente ecuación:

$$M_a = \sum_{i=1}^{n} y_i M_i \tag{1.13}$$

Donde:
M_a = peso molecular aparente de una mezcla de gases $[g/mol]$
M_i = peso molecular del componente i $[g/mol]$
y_i = fracción molar del componente i en la mezcla [1]

1.2.2. Densidad del Gas ρ_g

Para cualquier condición de presión y temperatura, la densidad de una mezcla de gases se calcula con la siguiente expresión:

$$\rho_g = \frac{pM_a}{RTz} \tag{1.14}$$

Donde:
ρ_g = densidad del gas $[lb/ft^3]$
M_a = peso molecular aparente de una mezcla de gases $[g/mol]$
p = presión del sistema $[psi]$
R = constante universal de los gases reales para estas unidades$[\frac{psi\ \ ft^3}{lb-mol\ \ R^\circ}]$
T = temperatura del sistema $[R^\circ]$
z = factor compresibilidad del gas [1]

1.2.3. Volumen Específico v

El volumen específico de un gas se define como el volumen que ocupa una unidad de su masa. Es igual al inverso de su densidad y se expresa como:

$$v = \frac{1}{\rho_g} \tag{1.15}$$

Donde:
v = volumen específico de una mezcla de gases $[ft^3/lb]$.
ρ_g = densidad del gas $[lb/ft^3]$

1.2.4. Densidad Relativa del Gas ρ_{rg}

También llamada gravedad específica del gas, se define como la relación existente entre la densidad del gas y la del aire. Ambas densidades se miden a las mismas condiciones de presión y temperatura. Comúnmente, la presión y temperatura estándar (60°/60°) se usan para definir la ρ_{rg}.

$$\rho_{rg} = \frac{\rho_g}{\rho_{air}} \tag{1.16}$$

Donde:
ρ_{rg} = densidad relativa del gas [1]
ρ_g = densidad de la mezcla de gas $[lb/ft^3]$
ρ_{air} = densidad del aire $[lb/ft^3]$

o

$$\rho_{rg} = \frac{M_a}{M_{air}} = \frac{M_a}{28{,}96} \ ,60°/60° \tag{1.17}$$

Donde:

M_a = peso molecular aparente de la mezcla de gases $[g/mol]$
M_{air} = peso molecular aparente del aire $[g/mol]$

1.2.5. Factor de Compresibilidad del gas z

También llamado factor de desviación del gas o simplemente factor z, es una cantidad adimensional que se define como la relación del volumen de n-moles de la mezcla de gas con el volumen del mismo número de moles de un gas ideal a las mismas condiciones de presión y temperatura.

$$z = \frac{V_{actual}}{V_{ideal}} = \frac{V_{actual}}{(nRT)/p} \tag{1.18}$$

Donde:
z = factor de compresibilidad del gas [1]
V_{actual} = volumen de la mezcla de gas $[ft^3]$
V_{ideal} = volumen de un gas ideal$[ft^3]$
n = número de moles de gas $[lb-mol]$

1.2.6. Compresibilidad del gas C_g

Por definición, la compresibilidad isotérmica del gas es el cambio en el volumen por unidad de volumen del mismo con la variación en la presión. Para un gas ideal (z=1), C_g se puede expresar como:

$$C_g = \frac{1}{p} \tag{1.19}$$

Donde:
C_g = compresibilidad del gas $[psi^{-1}]$
p = presión del sistema $[psi]$

1.2.7. Factor de Volumen del Gas B_g

Esta propiedad del gas se define como el volumen real ocupado por una cierta cantidad de gas a una presión y temperatura especifica, dividido por el volumen ocupado por la misma cantidad de gas a condiciones de presión y temperatura estándar.

$$B_g = \frac{(V_g)_{p,T}}{(V_g)_{sc}} \tag{1.20}$$

Donde:
B_g = factor de volumen del gas $[ft^3/scf]$
$(V_g)_{p,T}$ = volumen del gas a la condición de presión y temperatura seleccionada. $[ft^3]$
$(V_g)_{sc}$ = volumen del gas a condiciones estándar $[scf]$

1.3. Clasificación de los Tipos de Yacimiento de Acuerdo a sus Propiedades PVT

Con frecuencia, los ingenieros petroleros tienen la tarea de estudiar el comportamiento y las características de los yacimientos petroleros y determinan el desarrollo de la producción futura con el objetivo de maximizar las ganancias.

Los hidrocarburos que se encuentran de forma natural en los yacimientos de petróleo son mezclas de compuestos orgánicos que exhiben un comportamiento multifásico cambiante en amplios intervalos de presiones y temperaturas. Los sistemas de hidrocarburos pueden existir en estados gaseosos, líquidos y/o sólidos.

Las diferencias en el comportamiento de las fases dan como resultado la existencia de diversos tipos de yacimientos de hidrocarburos. Los yacimientos se clasifican comúnmente por los fluidos que contienen (Ej. aceite o gas). Estas clasificaciones amplias se subdividen aún más según:

- La composición de los fluidos del yacimiento
- Las condiciones de presión y temperatura de yacimiento
- Las propiedades físicas (PVT) de los fluidos contenidos en el yacimiento.

1.3.1. Diagrama de Fases

McKinnel y col. (2020) El diagrama de fases representa de forma gráfica los estados físicos de una sustancia o mezcla de compuesto bajo diferentes condiciones de presión y temperatura. Se compone de la presión a la que se encuentra el sistema en el eje y la temperatura en el eje x

Ahmed (2006) El diagrama de fases es la herramienta más útil para determinar el tipo de yacimiento de acuerdo a los fluidos que contiene. El mecanismo de interpretación de estos diagramas es simple, cuando se cruzan las líneas o curvas, se origina un cambio de fase.

Para comprender de manera correcta los diagramas de fase, es necesario identificar y definir los siguientes elementos clave :

- **Cricondenterma(T_{ct})** Se define como la temperatura máxima por encima de la cual no se puede formar líquido independientemente de las condiciones de presión.
- **Cricondenbara (p_{cb})** Es la presión máxima por encima de la cual no se puede formar gas independientemente de la temperatura .
- **Punto crítico (C)** Para una mezcla multicomponente, se define como el punto de presión y temperatura en el cual todas las propiedades intensivas de las fases liquida y gaseosa son iguales. En este punto, los valores de presión y temperatura correspondientes son denominados presión y temperatura crítica. (p_c , T_c).
- **Envolvente de fase** Se define como la región delineada por la curvas de punto de burbuja y de punto de rocío. Dentro de esta envolvente, el gas y el líquido coexisten en equilibrio.

- **Líneas de calidad** Son las líneas punteadas ubicadas dentro del diagrama que describen las condiciones de presión y temperatura para obtener volúmenes iguales de líquidos. Las líneas de calidad convergen en el punto crítico.
- **Curva de puntos de burbuja** Se define como la línea o curva que divide la región de la fase líquida con la región de dos fases.
- **Curva de punto de rocío** Es la curva o línea que divide la región de la fase de vapor (gas) con la región de dos fases.

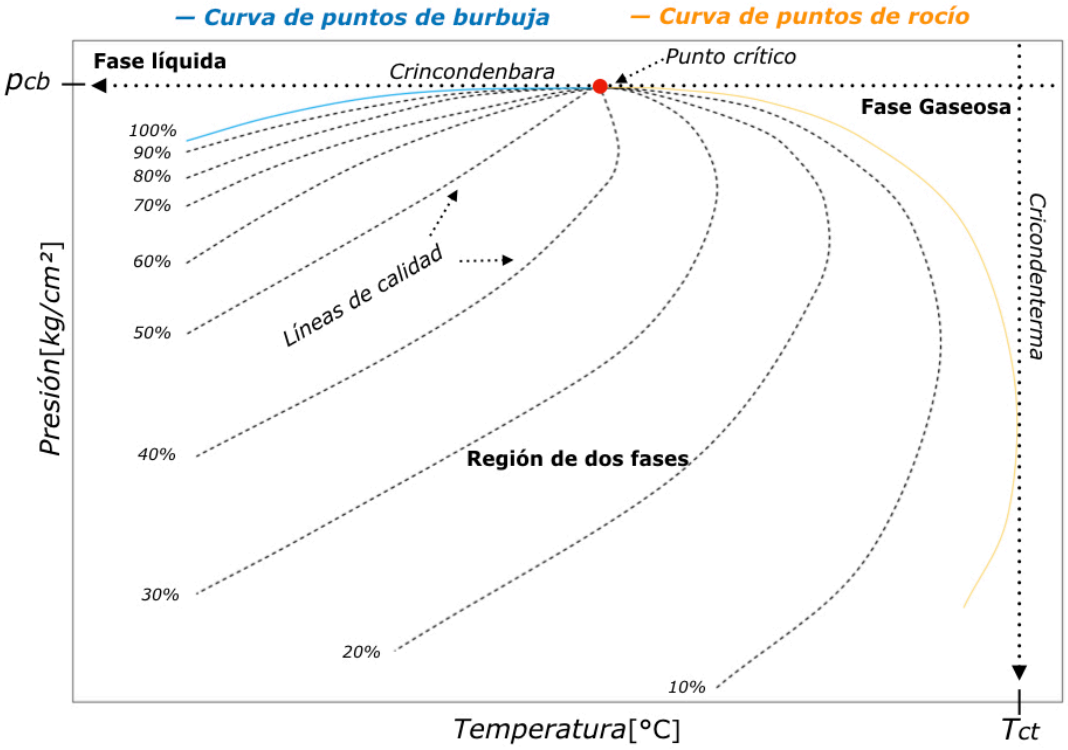

Figura 1.5: *Ahmed (2006)* Diagrama de fases generado para una mezcla de hidrocarburos.

Los yacimientos se clasifican en función de la ubicación del punto que representa su presión y temperatura con respecto al diagrama de fase generado con el fluido que contiene. *McCain (1990)* Por lo tanto, los yacimientos pueden clasificarse de forma general en los siguientes tipos:

- **Aceite negro** También es llamado de aceite crudo de bajo encogimiento. Consiste en una gran variedad de compuestos químicos que incluyen cadenas pesadas y largas de moléculas no volátiles. El gas asociado normalmente es del tipo seco
- **Aceite volátil** También es llamado aceite crudo de alto encogimiento. Comparado con los aceites negros, contiene menos moléculas pesadas y más cadenas intermedias (etanos a hexanos) de hidrocarburos . El gas asociado normalmente es del tipo húmedo.
- **Gas y condensado** Si la temperatura del yacimiento se encuentra entre la temperatura crítica y la cricondenterma del fluido, se clasifica como un yacimiento de gas retrógrado. El hidrocarburo existe en forma de gas en el yacimiento, aunque puede existir la presencia de

líquidos livianos (condesados) precipitados en la vecindad del pozo. A medida que la presión a la que es sometida el fluido decrece a lo largo del trayecto rumbo a superficie, se precipitan cantidades considerables de líquidos livianos llamados condensados. Los condensados se componen principalmente de propano, butano, pentano y otras fracciones más pesadas de hidrocarburos. En esta peculiar categoría de yacimientos de hidrocarburos, el comportamiento termodinámico del fluido es el factor que controla el desarrollo y el proceso de recuperación.

- **Gas húmedo** Contiene más elementos pesados que el gas seco (contenido de la fracción $C_1 < 90\,\%$) . En el yacimiento existe solo en forma de gas, sin embargo al transportarse rumbo al separador, en algún punto cruza levemente la envolvente de fases lo que causa la condensación de líquidos livianos en el proceso (menor cantidad que en los yacimientos de gas y condensado). Esto es causado por una disminución suficiente en la energía cinética de las moléculas pesadas con la caída de temperatura y su posterior cambio a líquido a través de las fuerzas de atracción entre las moléculas.
- **Gas seco** Se compone primordialmente de elementos ligeros (contenido de la fracción $C_1 > 90\,\%$), se mantiene a lo largo de todo su trayecto hacia el separador fuera de la envolvente de fases, por lo que la formación de líquidos tiende a ser nula. Esto es debido a que la energía cinética de la mezcla de gases es tan alta y la atracción entre las moléculas es tan pequeña que ninguna de ellas se une a un líquido en las condiciones de temperatura y presión del tanque de almacenamiento.

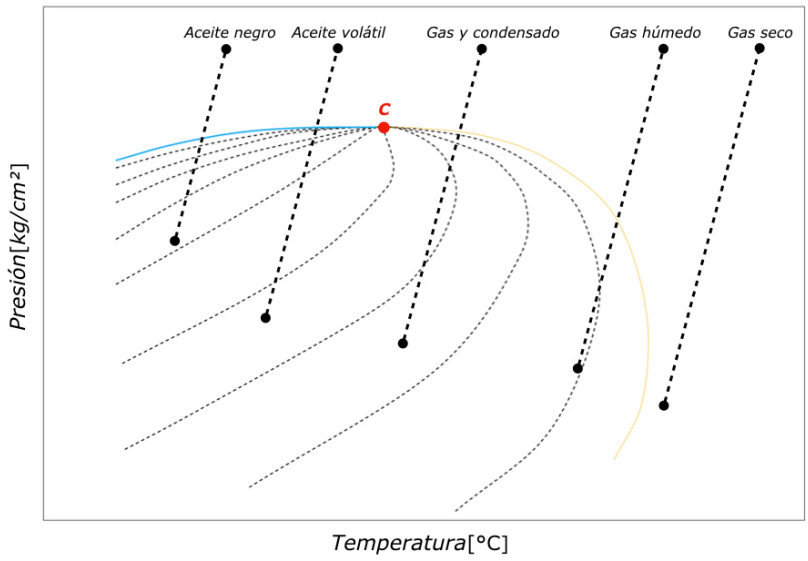

Figura 1.6: *McCain (1990)* Clasificación general del tipo de yacimiento de acuerdo a la ubicación de sus condiciones en el diagrama de fases generado para una muestra de fluidos del mismo.

La **figura 1.6** muestra cómo se ubican típicamente los tipos de yacimientos en un diagrama de fases. El punto superior representa las condiciones de yacimiento. El punto inferior indica las condiciones medidas en el separador a nivel de superficie.

Los yacimientos de aceite negro se encuentran alejados a la izquierda del punto crítico, teniendo así su fase estable en forma líquida dentro del yacimiento, con forme se reducen la temperatura y presión, el sistema pasa a entrar en la región de saturación y liberar gas, lo que genera un leve efecto de encogimiento.

Los yacimientos de aceite volátil por su parte, se encuentran también a la izquierda del punto crítico (fase liquida a nivel yacimiento) en el diagrama de fases, sin embargo a una mucha menor distancia. Lo anterior hace que el fluido entre de manera más agresiva a la región de saturación conforme se transporta a la batería de separación, liberando así mayor cantidad de gas en el proceso que un aceite negro y por ende genera así un efecto de alto encogimiento.

Las condiciones de los yacimiento de gas y condensado se encuentran entre el punto crítico y la cricondenterma, obteniendo así originalmente una fase gaseosa en el yacimiento, sin embargo mientras se va reduciendo la presión y temperatura se entra en la envolvente de fases por su parte superior (curvas de roció) y empieza el fenómeno de condensación retrograda, donde se empieza a condensar parte del líquido debido a que la atracción entre las moléculas del componente pesado se vuelve más efectiva, efecto que dura hasta que se alcanza un punto máximo de condensación más o menos a la mitad de la envolvente. Una vez alcanzado este punto, una mayor reducción de la presión permite que las moléculas pesadas comiencen el proceso normal de vaporización. El proceso de vaporización continúa hasta que la presión de reserva alcanza la presión más baja del punto de rocío. Esto significa que todo el líquido que se formó debe vaporizarse.

Los yacimientos de gas húmedo se ubican a la derecha de la cricondenterma en el diagrama de fases, por lo que a nivel yacimiento solo se encuentra la fase gaseosa, al disminuir las condiciones a las que es sometido el fluido por el proceso de transporte hacia la superficie, se cruza parcialmente la envolvente de fases por lo que se observa también la condensación de líquidos (condensados), sin embargo esto ocurre en menor cantidad que los yacimientos de gas y condensado debido a la ubicación el diagrama de fases

Por último, los yacimientos de gas seco se ubican también a la derecha de la cricondenterma pero están considerablemente más alejados que los yacimientos de gas húmedo, por lo que durante toda su vida productiva el fluido se mantiene en estado gaseoso al no poder entrar en la región de dos fases y el único contenido de líquido que se observa es el agua producida.

1.3.2. Propiedades PVT del Aceite Negro

Los yacimientos de aceite negro contienen cadenas de hidrocarburos más complejas y pesadas que el resto yacimientos, por lo que muestran un rango de características físicas y composicionales definidas (según el autor).

Propiedad	*Mendez*	*McCain*	*Moses*	*Santos **
Factor de volumen del aceite a Pb (B_{ob}) [m^3/m^3]	*< 2.0*	*< 2.0*	*< 2.0*	*< 2.0*
Relación gas - aceite (RGA) [m^3/m^3]	*< 200*	*< 356*	*< 356*	*< 200*
Densidad relativa del aceite (ρ_{ro}) [1]	*>0.85*	*> 0.80*	*> 0.80*	*> 0.834*
Gravedad API del aceite [°API]	*< 34.9*	*< 45*	*< 45*	*< 38*
Contenido de la fracción C7+ [%]	*> 20*	*> 20*	*-*	*> 25*
** Incluye aceite negro y ligero*				

Tabla 1.1: *Santos (2013)* Rango de propiedades PVT reportadas para yacimientos de aceite negro

En la **tabla 1.1** se muestran los rangos de propiedades PVT observadas en el aceite negro por varios autores. Estos rangos son útiles para clasificar debidamente el tipo de fluido encontrado en yacimientos nuevos cuando solo se cuenta con datos básicos del mismo.

1.3.3. Propiedades PVT del Aceite Volátil

Debido al mayor contenido de elementos intermedios, el aceite volátil o de alto encogimiento, muestra características físicas diferentes al aceite negro, en la siguiente tabla se definen por autor los rangos para el aceite volátil.

Propiedad	*Mendez*	*McCain*	*Moses**	*Santos*
Factor de volumen del aceite a Pb (B_{ob}) [m³/m³]	*> 2.0*	*> 2.0*	*< 2.0*	*> 2.0*
Relación gas - aceite (RGA) [m³/m³]	*200 - 1000*	*356 - 587*	*< 356 - 534*	*200 -550*
Densidad relativa del aceite (ρ_{ro}) [1]	*0.75 - 0.85*	*< 0.83*	*< 0.8*	*< 0.834*
Gravedad API del aceite [°API]	*35 - 49.9*	*> 40*	*> 40*	*> 38*
Contenido de la fracción C7+ [%]	*12.5 - 25*	*12.5 - 20*	*12.5 - 22*	*12.7 - 25*
* Yacimientos cercanos al punto critico				

Tabla 1.2: *Santos (2013)* Rango de propiedades PVT reportadas para yacimientos de aceite volátil

La **tabla 1.2** muestra en comparación con el aceite negro, mayores valores de B_{ob} y R_{sb} así como menores de ρ_{ro} y contenido de elementos C_7+ para el aceite volátil.

1.3.4. Propiedades PVT del Gas y Condensado

Los condensados por ser ricos en elementos ligeros y de alto poder calorífico generalmente poseen un valor de mercado mayor que el aceite, pero a su vez requieren un diseño de instalaciones más especializadas para poder aprovecharse al máximo, por lo que es importante la correcta identificación de estos yacimientos . En la siguiente tabla se expresan los rangos de propiedades físicas para los fluidos de yacimientos de gas y condensado

Propiedad	*Mendez*	*McCain*	*Moses**	*Santos*
Factor de volumen del aceite a Pb (B_{ob}) [m³/m³]	-	-	-	-
Relación gas - aceite (RGA) [m³/m³]	*500 - 15,000*	*587 - 8,905*	*534 - 26,716*	*550 - 10,000*
Densidad relativa del aceite (ρ_{ro}) [1]	*0.75 - 0.80*	*0.74 - 0.83*	*0.74 - 0.83*	*0.731 -0.815*
Gravedad API del aceite [°API]	*45 - 57.1*	*40 - 60*	*40 - 60*	*42 - 62*
Contenido de la fracción C7+ [%]	*3- 12.5*	*1 - 12.5*	*< 12.5*	*1 - 12.7*

Tabla 1.3: *Santos (2013)* Rango de propiedades PVT reportadas para yacimientos de gas y condensado

En la **tabla 1.3** se muestran los rangos de propiedades PVT que ocurren en los yacimientos de gas y condensado. Se observa en ocasiones una similitud en densidad con el aceite volátil, sin embargo se distinguen por las altas cantidades de gas y el contenido bajo de la fracción C_7+.

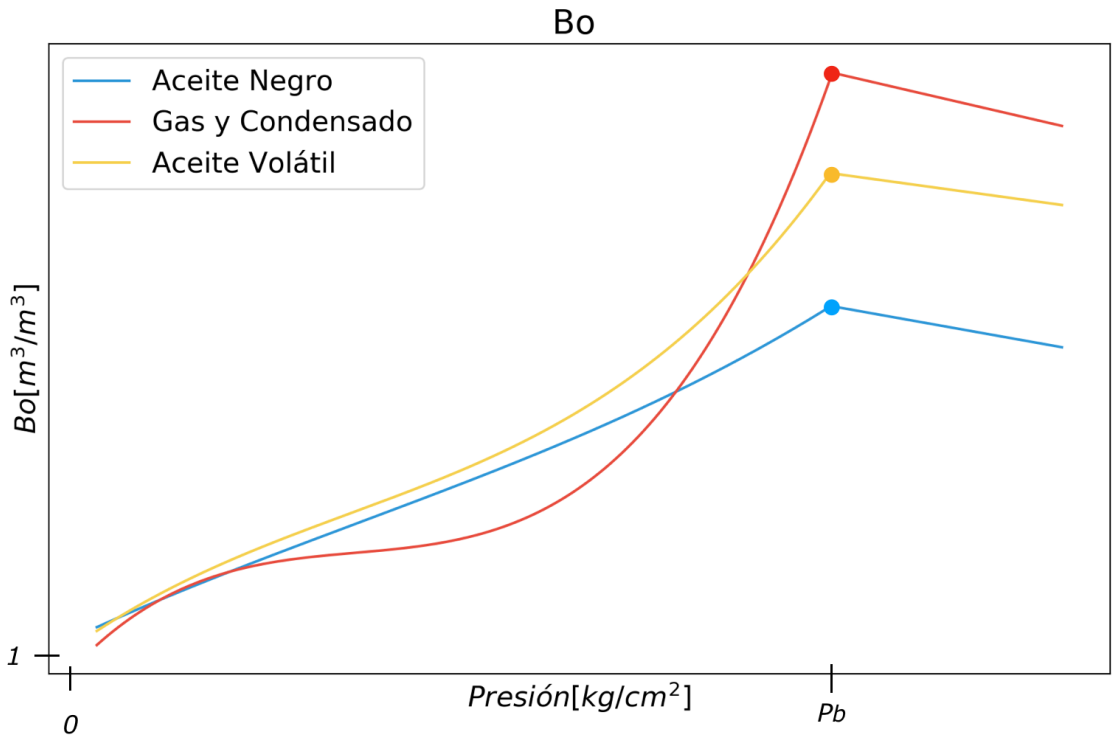

Figura 1.7: Ahmed (2006) Ejemplo del comportamiento del factor de volumen del aceite por tipo de yacimiento con la variación de presión del sistema.
* En el caso de gas y condensado, se gráfica el factor de volumen de la fase líquida.

La **figura 1.7** ejemplifica el comportamiento típico del factor de volumen del aceite cuando existe variación de presión a temperatura constante por tipo de yacimiento. *Carlson (2006)* El parámetro que principalmente controla la forma y magnitud del factor de volumen del aceite es la cantidad de gas en solución existente debido a que el volumen ocupado por el aceite está en función de la cantidad de gas liberado.

Podemos observar que el comportamiento de la curva del B_o para el aceite negro se mantiene cercana a un comportamiento lineal y con un valor relativamente bajo de B_{ob} debido a que la R_s es baja y el gas se libera de manera cercana a constante con la reducción de presión.

El comportamiento de la curva para el aceite volátil se aprecia con un cambio de pendiente más pronunciado y un valor de B_{ob} mayor a los aceites negros debido a su mayor valor de la R_s.

Por último el comportamiento de la curva para los yacimientos de gas y condensado es más brusco debido al alto R_s y por lo tanto la gran cantidad de gas liberado durante parte del proceso.

La **figura 1.8** representa la magnitud típica de la presión de saturación a temperatura constante observada por tipo de yacimiento. El valor de la presión de burbuja de un fluido se puede ver reflejada en un diagrama de fases y es función de la temperatura de yacimiento y la composición del fluido.

Por su posición en el diagrama de fases, los yacimientos de aceite negro exhiben una magnitud de presión de burbuja baja.

Dandekar (2013) Los yacimientos de aceite volátil y de gas y condensado regularmente exhiben una presión de burbuja alta comparada con el aceite negro por su ubicación en el diagrama de fases.

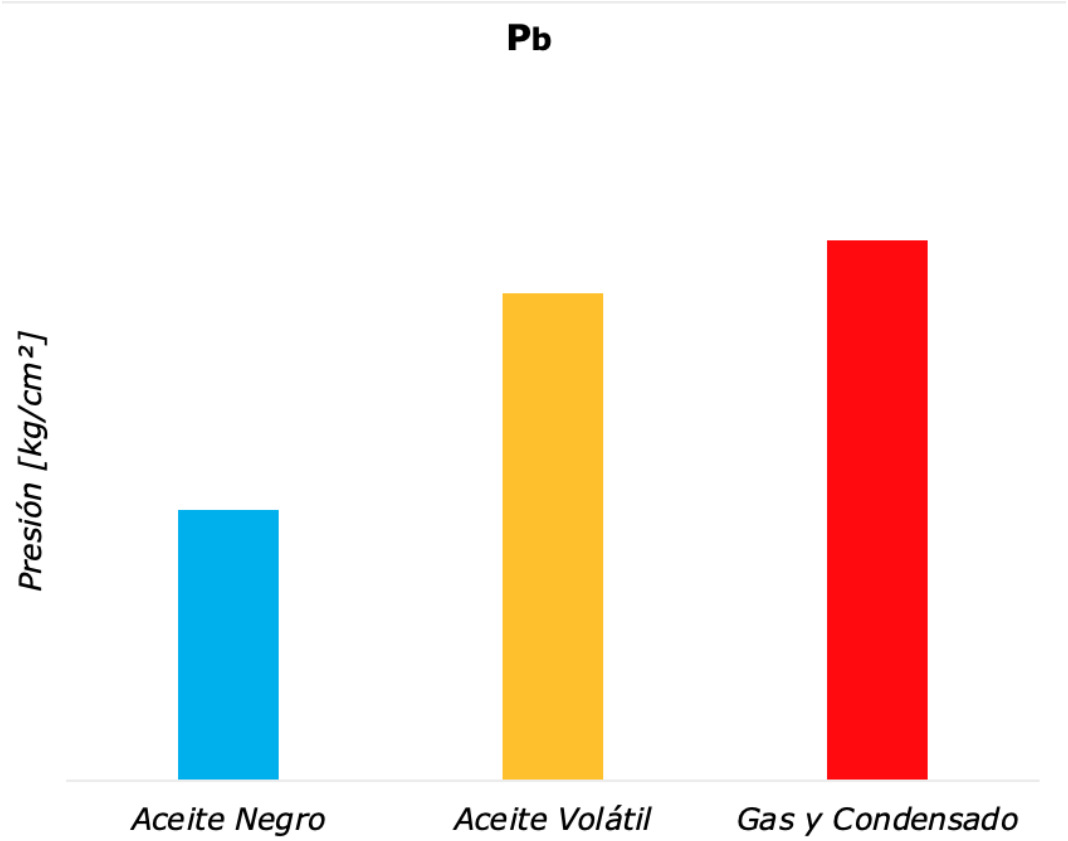

Figura 1.8: *Dandekar (2013)* Valores típicos de presión de burbuja por tipo de yacimiento.

1.4. Impacto de la correcta caracterización de los fluidos petroleros

El ingeniero petrolero debe determinar el tipo de fluido y sus propiedades físicas desde una etapa temprana en la vida del yacimiento ya que son un factor clave en la toma de decisiones para maximizar ganancias, como lo son: la selección del método de muestreo, el diseño de instalaciones en superficie, el cálculo de volúmenes de hidrocarburos, elaboración del plan de explotación y la selección de métodos de recuperación mejorada.

El siguiente caso, es un ejemplo del impacto que genera la incertidumbre de los valores de las propiedades PVT de los fluidos del yacimiento.

El campo Big Butte es un yacimiento de empuje combinado. La presión actual de la formación se estima en 4000 *psi*. Los datos de producción del yacimiento y la información PVT se dan a continuación:

Propiedad	*Condiciones iniciales de yacimiento*	*Condiciones actuales de yacimiento*
$\dot{p}$ [psi]	4000	3000
B_o [bbl/STB]	1.4	1.33
R_s [scf/STB]	600	500
N_p [MMSTB]	0	8
G_p [MMMscf]	0	6
B_w [bbl/STB]	1	1
W_e [MMbbl]	0	3
W_p [MMbbl]	0	0.2
B_g [bbl/scf]	0.0011	0.0015
c_f, c_w	-	-
Volumen bruto de la zona de aceite [ac-ft]	100,000	-
Volumen bruto de la zona de gas [ac-ft]	20,000	-

Tabla 1.4: *Ahmed (2006)* Datos disponibles del yacimiento Big Butte

Como parte de la elaboración del plan de explotación, se requiere calcular volumen original de aceite en sitio utilizando la ecuación de balance de materia.

Solución

Paso 1: Asumiendo la misma porosidad y saturación de agua connata para las zonas de aceite y gas, calcular el parámetro adimensional m.

$$m = \frac{volumen\ de\ la\ zona\ de\ gas}{volumen\ de\ la\ zona\ de\ aceite} \tag{1.21}$$

Donde:
m = factor adimensional de casquete de gas [1]

$$m = \frac{20,000}{100,000} = 0{,}2 \tag{1.22}$$

Paso 2: Calcular la relación gas aceite acumulada R_p

$$R_p = \frac{G_p}{N_p} \tag{1.23}$$

Donde:
R_p = relación gas aceite acumulada $[scf/STB]$
G_p = producción de gas acumulada $[scf]$
N_p = producción de aceite acumulada $[STB]$

$$R_p = \frac{6 \times 10^9}{8 \times 10^6} = 750\ [scf/STB]$$

Paso 3: Calcular volumen original de aceite en sitio (N) utilizando la ecuación de balance de materia (Sin gastos de inyección ni efectos de compresibilidad).

$$N = \frac{N_p[B_o + (R_p - R_s)B_g] - (W_e - W_p B_w)}{B_o - B_{oi} + (R_{si} - R_s)B_g + mB_{oi}[\frac{B_g}{B_{gi}} - 1]} \tag{1.24}$$

Donde:

N = volumen original de aceite en sitio $[STB]$

N_p = producción de aceite acumulada $[STB]$

B_o = factor de volumen del aceite actual $[bbl/STB]$

B_{oi} = factor de volumen del aceite inicial $[bbl/STB]$

B_g = factor de volumen del gas actual $[bbl/scf]$

B_{gi} = factor de volumen del gas inicial $[bbl/scf]$

R_p = relación gas aceite acumulada $[scf/STB]$

R_s = relación de solubilidad gas aceite actual $[scf/STB]$

R_{si} = relación de solubilidad gas aceite inicial $[scf/STB]$

G_p = producción de gas acumulada $[scf]$

m = factor adimensional de casquete de gas [1]

W_e = entrada de agua $[bbl]$

W_p = agua producida $[bbl]$

B_w = factor de volumen del agua $[bbl/STB]$

$$N = \frac{8 \times 10^6[1{,}33 + (750 - 500)0{,}0015] - (3 \times 10^6 - 0.2 \times 10^6)}{1{,}33 - 1{,}40 + (600 - 500)0{,}0015 + (0{,}2 \ * \ 1{,}4)[\frac{0{,}0015}{0{,}0011} - 1]C}$$

$$N = 33.684 \times 10^6 \ [STB]$$

Paso 4: Calcular el volumen recuperable y el ingreso que representa este volumen. Asumiendo que el factor de recuperación (F_{r1P}) esperado para la reserva 1_P es de 10 % y que se tiene un precio de venta promedio fijo del barril a 40$ $[USD]$

$$1_P = F_{r1P} * N \tag{1.25}$$

Donde:

1_P = volumen de la reserva 1_P $[STB]$

F_{r1P} = factor de recuperación esperado para la reserva 1_P [1]

$$1_P = 0{,}1 * 33.684 \times 10^6 = 3.684 \times 10^6 \ [STB]$$

$$Ingresos \ 1_P = 1_P * \$ \ bbl$$

Donde:

$Ingresos$ = ingresos esperados para la reserva 1_P $[USD]$

$\$ \ bbl$ = precio del barril $[USD]$

$$Ingresos \ 1_P = 3.684 \times 10^6 * 40\$ = 134{,}764 \ [MMUSD]$$

Paso 5: Con el objetivo de observar cómo afecta a un proyecto la correcta caracterización de los fluidos petroleros, se hace un análisis utilizando la metodología anterior para diferentes casos propuestos en los que se tiene un error del 1 % y del 5 % en una de las propiedades PVT medidas. La propiedad con error en la medición supuesta es el factor de volumen del aceite a la presión actual. Los resultados de este análisis se expresan en la siguiente tabla:

Caso	*Valor de Bo actual estimado [m³/m³]*	*Reserva 1P estimada [MMbbl]*	*Ingreso estimado [MMUSD]*	*Diferencia de ingreso estimado vs el real [MMUSD]*	*Porcentaje del error relativo absoluto en el ingreso estimado [%]*
Valor del Bo actual sin error (Real)	*1.33*	*3.368*	*134.734*	*-*	*-*
***1%** de error en el valor de Bo (superior)*	*1.343*	*3.548*	*141.922*	*7.188*	*5.33*
***1%** de error en el valor de Bo (inferior)*	*1.317*	*3.203*	*128.117*	*-6.617*	*4.91*
***5%** de error en el valor de Bo (superior)*	*1.397*	*4.454*	*178.162*	*43.428*	*32.23*
***5%** de error en el valor de Bo (inferior)*	*1.264*	*2.655*	*106.181*	*-28.553*	*21.19*

Tabla 1.5: Resultados del análisis realizado para observar el impacto de la variación de una propiedad PVT sobre el proyecto

La **tabla 1.5** expresa los resultados del análisis, se puede observar que a pesar de que solo se alteró el valor de una sola propiedad (factor de volumen del aceite a condiciones actuales), el impacto en los ingresos esperados es muy significativo. Cuando se evaluó el proyecto con el error del B_o en el orden del 1 %, se tuvo un impacto en el ingreso del 5 % aproximadamente. Cuando el error aumentó al 5 %, el impacto creció de manera acelerada y significativa hasta el 32 % para el caso de la estimación alta (superior) del valor del B_o.

El ejercicio anterior es un ejemplo que demuestra la importancia económica y técnica de la reducción en el error de la estimación del valor de las propiedades PVT de los fluidos producidos.

Capítulo 2

Obtención de las Propiedades PVT del Aceite

2.1. Muestreo

Dake (1998) Las muestras de fluido del yacimiento se recolectan normalmente en la etapa temprana de la vida productiva del mismo. Básicamente hay dos formas de recolectar muestras, ya sea por muestreo directo en el fondo del pozo o por recombinación superficial de las fases de petróleo y gas.

Cualquiera que sea la técnica utilizada, existe el mismo problema básico el cual es garantizar que la proporción de gas y petróleo en la muestra compuesta sea la misma que la existente originalmente en el yacimiento.

La estimación o cálculo de las propiedades PVT del aceite se puede realizar de distintas maneras si se cuenta con una muestra de fluido representativa del pozo.

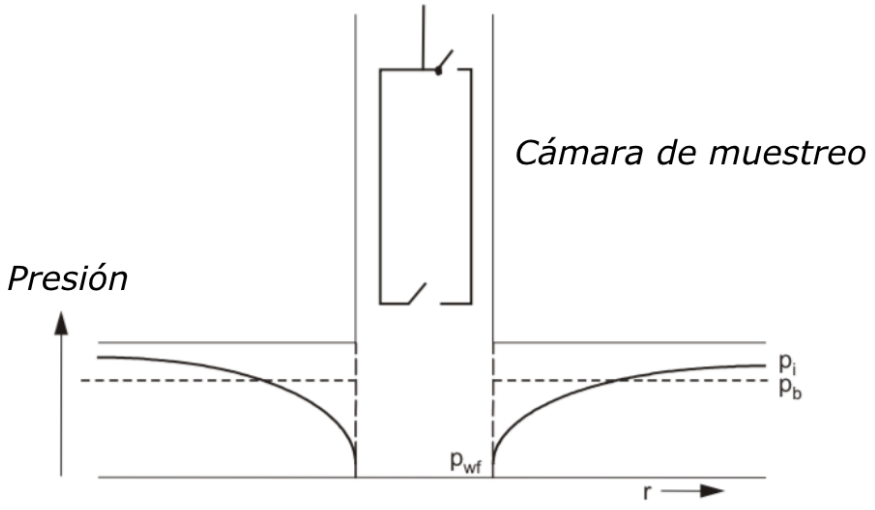

Figura 2.1: *Dake (1998)* Muestreo de fluidos en el fondo del pozo.

La **figura 2.1** muestra el proceso de recolección de una muestra de fondo. Para poder realizar este tipo de muestreo se corre una sonda de muestreo especial en el pozo con línea eléctrica hasta la profundidad del yacimiento. La sonda permite recolectar la muestra del flujo del pozo a condiciones de presión de fondo.

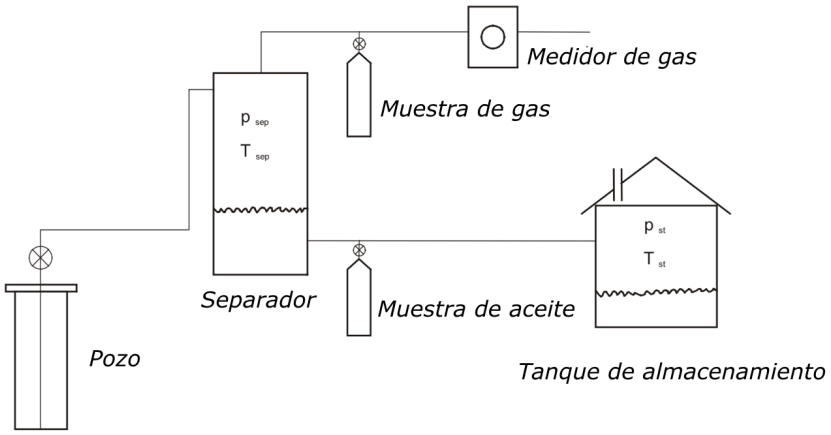

Figura 2.2: *Dake (1998)* Muestreo de fluidos en superficie.

La **figura 2.2** ilustra el proceso de muestreo con recombinación en superficie. Cuando se recolectan muestras de fluido en la superficie, se toman volúmenes separados de aceite y gas a condiciones de separador. El pozo se produce a un gasto constante durante un período de varias horas para conocer la relación gas aceite que es medida en metros cúbicos de gas de separador por barril de aceite en el tanque de almacenamiento. Una vez conocida dicha relación, las fases se recombinan para generar una muestra de fluido compuesta representativa.

2.2. Pruebas de Laboratorio

Ahmed (2006) Una de las formas para conocer las propiedades PVT de los fluidos son las pruebas de laboratorio, si bien son el método más fiable cuando se ejecutan de forma correcta, son el método más costoso económicamente hablando, pues se requiere de equipo de laboratorio especializado.

Es necesario que los estudios de laboratorio para conocer el comportamiento PVT y de equilibrio de fase de los fluidos del yacimiento sean precisos. Dichos estudios son necesarios para caracterizar los fluidos y evaluar su desempeño volumétrico a varios niveles de presión. Hay muchos análisis de laboratorio que se pueden hacer en una muestra. La cantidad de datos deseados determina el número de pruebas realizadas en el laboratorio. En general, existen tres tipos de pruebas de laboratorio utilizadas para medir muestras de yacimientos de hidrocarburos:

- **Pruebas primarias:** Son pruebas de campo simples que involucran las mediciones de la gravedad específica y la relación gas aceite de los hidrocarburos producidos.
- **Pruebas de laboratorio rutinarias:** Son varias pruebas de laboratorio que se realizan de manera rutinaria para caracterizar los hidrocarburos del yacimiento. Incluyen:
  - Análisis composicional

  - Expansión a composición constante o separación flash
  - Liberación diferencial
  - Pruebas de separador
  - Agotamiento a volumen constante

- **Pruebas especiales de laboratorio PVT:** Este tipo de pruebas se realizan para aplicaciones muy específicas. Si un yacimiento se va a explotar con una inyección de gas miscible o cíclica de gas, se pueden realizar las siguientes pruebas:
  - Prueba *Slim Tube*
  - Prueba de hinchamiento

El aparato utilizado para realizar la mayoría de los experimentos anteriores es una celda PVT, como se muestra en la **figura 2.3**.

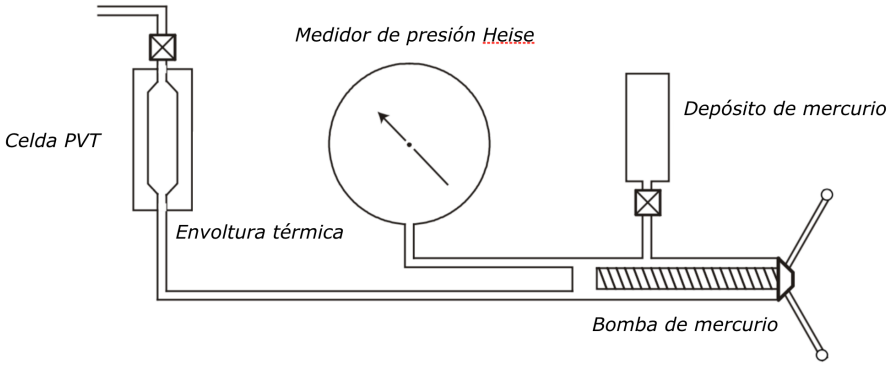

Figura 2.3: *Dake (1998)* Arquitectura de una celda PVT .

Después de recombinar el aceite y el gas en las proporciones correctas se carga el fluido en la celda PVT que se mantiene a temperatura constante la cual es la temperatura del yacimiento medida. La presión de la celda se controla mediante una bomba de mercurio de desplazamiento positivo y se registra en un medidor de presión preciso. El movimiento del émbolo se calibra en términos de volumen de mercurio inyectado o extraído de la celda PVT para que los cambios de volumen en la celda puedan medirse directamente.

No es la intención de este trabajo profundizar en cómo se realiza específicamente cada prueba de rutina, si no especificar cuáles propiedades PVT de los aceites se pueden obtener con ellas.

Experimento	*Propiedades obtenidas*
Análisis composicional	*- Fracciones mol: C_n, H_2S, CO_2, etc.* *- Densidad de la fase liquida y gaseosa*
Expansión a composición constante	*- Volumen relativo* *- P_b, C_o, B_o , ρ_{ro} (región bajo saturada)*
Liberación diferencial	*P_b, R_s, B_o, B_g, B_t, ρ_{ro}, ρ_{rg}, factor z (región saturada)*
Prueba de separador	*-Temperatura y presión optima de separación* *- RGA, ρ_{ro}, ρ_{rg} (condiciones de separación)*
Agotamiento a volumen constante (Aceites volátiles y condensados)	*- Saturación de liquido retrogrado* *- Factor z de las dos fases*

Tabla 2.1:*Ahmed (2006)* Experimentos PVT disponibles

2.3. Ecuaciones de Estado

Un método utilizado para estimar el comportamiento PVT de los fluidos del yacimiento es el uso de ecuaciones de estado modificadas debido a que el comportamiento del petróleo crudo está dictado por la presión, temperatura y composición del sistema.

Wu y Rosenegger (1997) Las ecuaciones de estado (EOS) son funciones semi empíricas que permiten describir el estado de agregación de la materia como una relación matemática entre la temperatura, la presión, el volumen, la densidad, la energía interna e incluso otros parámetros.

Yee (2011) Normalmente el uso de ecuaciones de estado para estimar propiedades PVT tales como P_b, R_{sb} y B_{ob} requiere de un análisis composicional complejo previo, por lo que es una solución condicionada a la existencia de dicho análisis.

Además es necesaria una muy buena comprensión de los parámetros usados para ajustar correctamente las EOS, las cuales se pueden modelar a través de simuladores de composición modernos, generalmente en forma cúbica o polinomial. En la industria petrolera las dos ecuaciones de estado más utilizadas desde hace décadas son: Peng-Robinson (PR) y Soave-Redlich-Kwong (SRK)

2.4. Correlaciones

Banzer (1996) Las correlaciones son un método empírico simple para estimar propiedades PVT que son desarrolladas a partir de datos de laboratorio y/o de campo utilizando variantes de regresiones lineales y métodos gráficos.

Fath y col. (2018) Durante las últimas siete décadas, los investigadores han presentado diversas correlaciones para la estimación de las propiedades PVT de los aceites crudos y se han introducido para una o más ubicaciones geográficas específicas con una composición química dada y un rango de condiciones a nivel yacimiento

La mayoría de estas correlaciones se desarrollan asumiendo que las propiedades PVT son función de la relación gas - aceite en solución, la temperatura del yacimiento, la densidad del petróleo y la densidad del gas. $[PVT = f\ (R_s, T_y, \rho_{ro}, \rho_{rg})]$

Hassn y Sadiq. (2009) La mayoría de estas correlaciones producen resultados razonablemente precisos cuando se aplican a la presión del punto de burbuja. Pero, para presiones por debajo del punto de burbuja, el factor de volumen de formación de aceite calculado puede producir un error considerable.

Los investigadores han descubierto que la densidad del gas es un factor de correlación fuerte con la mayoría de la propiedades PVT pero desafortunadamente, esta es a menudo una de las variables medidas con el menor grado de consistencia. La gravedad específica del gas depende de la presión y la temperatura de los separadores, que pueden no estar disponibles.

Correlación	P_b	B_o	R_s	μ_{od}	μ_{ob}	μ_{ou}
Standing	✔	✔	✔	✘	✘	✘
Al-Marhoun	✔	✔	✘	✘	✘	✘
Glaso	✔	✔	✔	✔	✘	✘
Dokla & Osman	✔	✔	✘	✘	✘	✘
Lasater	✔	✘	✔	✘	✘	✘
Vasquez & Beggs	✔	✔	✔	✘	✘	✔
Kartoatmodjo & Schmidt	✔	✔	✔	✔	✔	✔
Beggs & Robinson	✘	✘	✘	✔	✔	✘

Tabla 2.2:*Banzer (1996)* Correlaciones PVT más conocidas

La **tabla 2.2** describe de manera resumida, las propiedades PVT del aceite que pueden ser estimadas con las correlaciones más destacadas disponibles en la literatura.

Correlación	Zona geográfica	Rango de gravedad API [1]	Rango de R_s [Scf/STB]	Rango de temperatura [°C]	Numero de muestras
Standing	California	16.5 - 63.8	4 - 259	38 - 126	22
Al-Marhoun	Medio Oriente	19.4 - 44.6	5 - 285	23 - 116	69
Glaso	Mar del Norte	22.3 - 48.1	16 - 470	27 - 138	45
Dokla & Osman	Emiratos Árabes	28.2 - 40.3	14 - 404	88 - 135	51
Lasater	Canadá, Estados Unidos y Sudamerica	17.9 - 51.1	1 - 517	28 - 133	137
Vasquez & Beggs	Mundo (desconocido)	5.3 - 59.5	0 - 392	Sin especificar	600
Kartoatmodjo & Schmidt	Indonesia, Medio Oriente y América	14.4 - 58.9	0 - 515	24 - 160	740
Beggs & Robinson	Mundo (desconocido)	16-58	4 - 369	21 - 146	600

Tabla 2.3:*Banzer (1996)* Descripción de las correlaciones PVT más conocidas

La **tabla 2.3** muestra información y los rangos en los que las correlaciones de la **tabla 2.2** son aplicables.

2.5. Técnicas de Inteligencia Artificial

En recientes décadas, se ha introducido el uso de técnicas de *Machine Learning* o aprendizaje automático para la estimación de propiedades PVT.

APD. (2019) El *Machine Learning* es una rama de la inteligencia artificial que permite hacer automáticas una serie de operaciones con el fin de reducir la necesidad de que intervengan los seres humanos.

Lo que se denomina como "aprendizaje" consiste en la capacidad de un sistema para identificar una serie de patrones complejos determinados por parámetros de entrada. En el caso de estimación de propiedades PVT, se busca estimar las propiedades objetivo (ej. $B_o, P_b, etc.$) a partir de los parámetros de entrada (ej. $\rho_{ro}, Ty, etc.$).

Se han utilizado por diversos autores, técnicas tales como redes neuronales artificiales, máquinas de soporte de vectores y lógica difusa. Mostrando resultados positivos y superiores a los obtenidos con el uso de correlaciones PVT clásicas.

Estado del Arte

La estimación de las propiedades PVT de los fluidos petroleros por medio de técnicas de aprendizaje automático no es un tema que apenas se ha explorado vagamente. Desde hace más de dos décadas se han implementado diversos algoritmos para estimaciones por diferentes métodos como redes neuronales artificiales (ANN: *Artificial Neural Networks*), máquinas de vectores de soporte (SVM: *Support Vector Machines*), lógica difusa, algoritmos de caja transparente abierta (TOB: *Transparent Open Box*) e incluso sistemas híbridos.

Gharbi, Elsharkawy y col. (1999) propusieron un modelo novedoso con el objetivo de desarrollar una red neuronal artificial universal para estimar algunas propiedades PVT, como la presión de burbuja (P_b) y el factor de volumen del aceite a condiciones de presión de burbuja (B_{ob}), de varios sistemas de petróleo crudo en el mundo. El modelo estimó para el conjunto de datos de prueba, la Pb con un porcentaje absoluto de error relativo promedio (E_a) de 6.48 % y B_{ob} con un E_a de 1.97 %. Utilizando como variables de entrada al modelo la relación gas – aceite disuelto (R_s), la gravedad específica del gas (ρ_{rg}), la gravedad específica del aceite (ρ_{ro}) y la temperatura del yacimiento (T).

Años después, *Osman, Al-Mahourn y col. (2001)* realizaron un estudio para crear un modelo de redes neuronales artificiales que predijera el factor de volumen del aceite en la presión del punto de burbuja (B_{ob}) a partir de la temperatura del yacimiento, la relación gas-aceite disuelto (R_s), la gravedad específica del gas (ρ_{rg}) y la gravedad API del aceite. Para ello, hicieron uso de una arquitectura de red 4-5-1 entrenada a partir del algoritmo de retropropagación con 803 registros obtenidos de pozos de todo el mundo, presentando un E_a de 1.79 % y un coeficiente de correlación de 0.988.

Los autores anteriores *Osman, Al-Mahourn. (2002)* presentaron un modelo de red neuronal artificial para estimar la presión de burbuja y el factor de volumen del aceite en la presión de burbuja. Dicho modelo fue entrenado con 283 registros de datos obtenidos de aceites crudos de Arabia Saudita y presentó un porcentaje absoluto de error relativo promedio de 0.52 % con un coeficiente de correlación de 0.999, mejorando notablemente los resultados obtenidos anteriormente debido a una excelente correlación entre los datos de entrada y salida del conjunto de datos utilizado.

Nagy, Ahmed y col. (2009) exploraron otras técnicas de aprendizaje automático y publicaron un trabajo en el que investigaron la capacidad de las máquinas de vectores de soporte (SVM) para modelar las propiedades PVT de los sistemas de petróleo crudo debido a los inconvenientes de las redes neuronales artificiales, las cuales, mencionan en su trabajo, no operan con precisión al funcionar únicamente para un cierto rango de características del fluido del yacimiento y área geo-

gráfica con composiciones de fluidos similares. En dicho trabajo, los autores llegaron a la conclusión de que las máquinas de vectores de soporte tienen un rendimiento mejor, eficiente y confiable en comparación con las redes neuronales artificiales al obtener un coeficiente de correlación de 0.997 para el B_{ob} y del 0.977 para P_b.

Posteriormente, *Selamat, Raheem y col. (2012)* propusieron el uso de lógica difusa y, compararon su desempeño con redes neuronales artificiales entrenadas a partir del método de aprendizaje lineal basado en sensibilidad. El E_a registrado por la lógica difusa fue del 1.49 % para el B_{ob} y 20.65 % para P_b, a comparación de las redes neuronales que obtuvieron un E_a del 1.2 % para B_{ob} y 35.54 % para P_b.

Baarimah y col. (2015), trabajaron con técnicas de lógica difusa y redes neuronales para estimar una mayor cantidad de propiedades PVT, entre ellas, el factor de volumen del aceite en el punto de burbuja (B_{ob}), presión de saturación (P_b), la relación gas-aceite disuelto (R_s), la gravedad API del aceite y la gravedad específica del gas (ρ_{rg}). Al igual que en el trabajo de Selamat y Raheem, llegaron a la conclusión de que la lógica difusa superaba a las redes neuronales en cuanto a precisión en la resolución de este problema usando el mismo conjunto de datos para alimentar dichos modelos.

Oloso y col. (2015), propusieron el uso de máquinas de vectores de soporte para la estimación de propiedades que no habían sido consideradas antes en ningún trabajo de este tipo: la viscosidad de aceite muerto, la viscosidad del aceite saturado y la viscosidad del aceite bajo saturado. Los resultados obtenidos en este trabajo fueron satisfactorios al obtener un E_a del 10.32 % para la viscosidad del aceite muerto, de 7.04 % para la viscosidad del aceite en el punto de burbuja y de 1.19 % para la viscosidad del aceite a condiciones del yacimiento.

Ramírez y col. (2017), retomaron el uso de redes neuronales artificiales para realizar regresiones no lineales, combinándolas con una decorrelación lineal de los datos adquiridos a través del uso de análisis de componentes principales (PCA) con el objetivo de estimar con mayor precisión la P_b y el B_{ob} obteniendo así un E_a de 14.73 % y 1.47 % respectivamente.

Por último, *Wood y Choubineh (2019)* utilizaron algoritmos de caja transparente abierta (TOB) para estimar Bob a partir de los registros de un set de datos. Posteriormente se compararon las predicciones de Bob del algoritmo TOB para tres conjuntos de datos con un modelo de red neuronal artificial de perceptrón multicapa estándar. Se Llego a la conclusión de que la precisión de la estimación del modelo TOB es buena y además muy parecida a la obtenida con redes neuronales artificiales (mejor que la de correlaciones empíricas).

Cabe recalcar que el uso de técnicas de aprendizaje automático para la estimación de propiedades PVT en aceites ha comenzado a ser objeto de estudio en la Universidad Nacional Autónoma de México en años recientes.

Camargo (2016), realizó la estimación de propiedades PVT para yacimientos de aceite negro y volátil de la Región Sur de México mediante el uso de redes neuronal artificiales. Se obtuvo con resultados satisfactorios la presión de burbuja (P_b), relación de solubilidad gas – aceite (R_{sb}), com-

presibilidad del aceite (C_{ob}) y factor de volumen del aceite a la presión de saturación (B_{ob}). A partir de información básica de pozos de la región.

Hernández (2017), desarrolló modelos de redes neuronales para determinar las propiedades de los fluidos petroleros de manera indirecta y con mayor exactitud, junto con una metodología de validación del cumplimiento de las leyes físicas del yacimiento de cada modelo de red desarrollado. Se logro reconstruir de forma satisfactoria la curva del B_o vs presión para varias muestras de aceite, además se estimó con un E_a de 5.48 % la P_b de dichas muestras.

Capítulo 3

Aprendizaje Automático

El aprendizaje automático es una rama de la Inteligencia Artificial que busca diseñar algoritmos que permitan que una computadora aprenda. La palabra aprendizaje, en este contexto, no implica necesariamente la conciencia, sino el encontrar regularidades estadísticas y otros patrones en los datos. *Ayodele. (2010)*

Para ello, se hace uso de una variedad de algoritmos que detectan de forma automatizada patrones significativos en los datos para describir y predecir resultados. A medida que estos algoritmos se alimentan de los datos de entrenamiento, es posible producir modelos más precisos basados en dicha información.

En otras palabras, un sistema de aprendizaje automático se encarga de leer un conjunto de datos y optimizar un modelo para resolver un determinado problema.

Tanto el campo de la inteligencia artificial como el de aprendizaje automático no son nuevos. Los inicios de la inteligencia artificial se remontan al año 1950, cuando Arthur Lee Samuel, un investigador de IBM, desarrolló uno de los primeros programas de aprendizaje automático para jugar a las damas. El modelo mejoraba su juego cuanto más practicaba pues esto le permitía estudiar cuales movimientos constituían las estrategias ganadoras para utilizarlas en partidas futuras.*Samuel. (1959)* Explicó el enfoque de aprendizaje automático utilizado en un artículo publicado en 1959 en el IBM Journal of Research and Development.

Desde entonces y a lo largo del tiempo muchas áreas tales como la medicina y las finanzas han utilizado técnicas de inteligencia artificial para optimizar sus sistemas obteniendo una mayor eficiencia en sus procesos.

3.1. Sistemas de aprendizaje automático

En el campo del aprendizaje automático, un sistema es un algoritmo encargado de realizar iteraciones para optimizar un conjunto de datos, inicializar el modelo y alimentar a dicho modelo con los datos para aprender de ellos. La **figura 3.1** muestra los pasos en el funcionamiento de un

sistema de aprendizaje automático

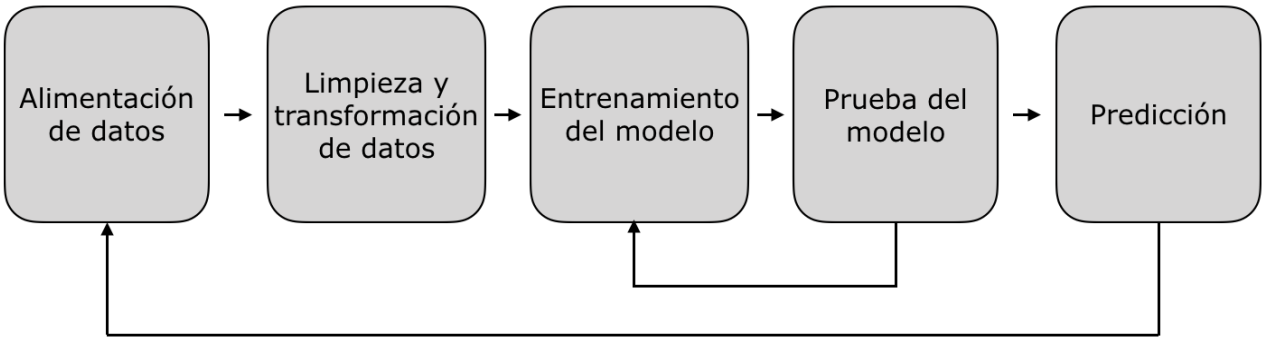

Figura 3.1: Diagrama de bloques de un sistema de aprendizaje automático.

Ayodele. (2010) Los sistemas de aprendizaje automático pueden clasificarse dependiendo de su enfoque de aprendizaje en *supervisados, no supervisados y por refuerzo*. A grandes rasgos, el *aprendizaje supervisado* implica intervención por parte de un humano para indicar cuándo una predicción es correcta o incorrecta a través de etiquetas, mientras que en el *aprendizaje no supervisado* el algoritmo simplemente categoriza los datos en función de su estructura oculta. La **figura 3.2** nos muestra de forma gráfica ambos conceptos

Por su parte, el *aprendizaje por refuerzo* es un poco diferente a los anteriores en el sentido de que el aprendizaje se lleva a cabo mediante recompensas. En este caso, el sistema es llamado *agente* e intenta aprender mientras maximiza las recompensas.

Dekel. (2010) También, otra forma de clasificar los sistemas de aprendizaje automático es dependiendo de si el modelo obtenido puede aprender de forma incremental sobre la marcha. En los sistemas de *aprendizaje en línea (online learning)* los modelos se entrenan de forma incremental al alimentar instancias secuencialmente, ya sea individualmente o en grupos pequeños llamados mini lotes. La **figura 3.3** muestra el proceso del aprendizaje en linea. Por otro lado, en los sistemas de *aprendizaje por lotes (batch learning)*,los modelos son incapaces de aprender de forma incremental. Primero, el modelo se entrena con todos los datos y luego es lanzado a producción. La **figura 3.4** muestra el proceso del aprendizaje por lotes.

3.2. El ciclo del aprendizaje automático

El proceso de creación de una aplicación de aprendizaje automático es iterativo debido a que se obtiene nueva información cada día. Debido a ello, se debe mantener el modelo actualizado una

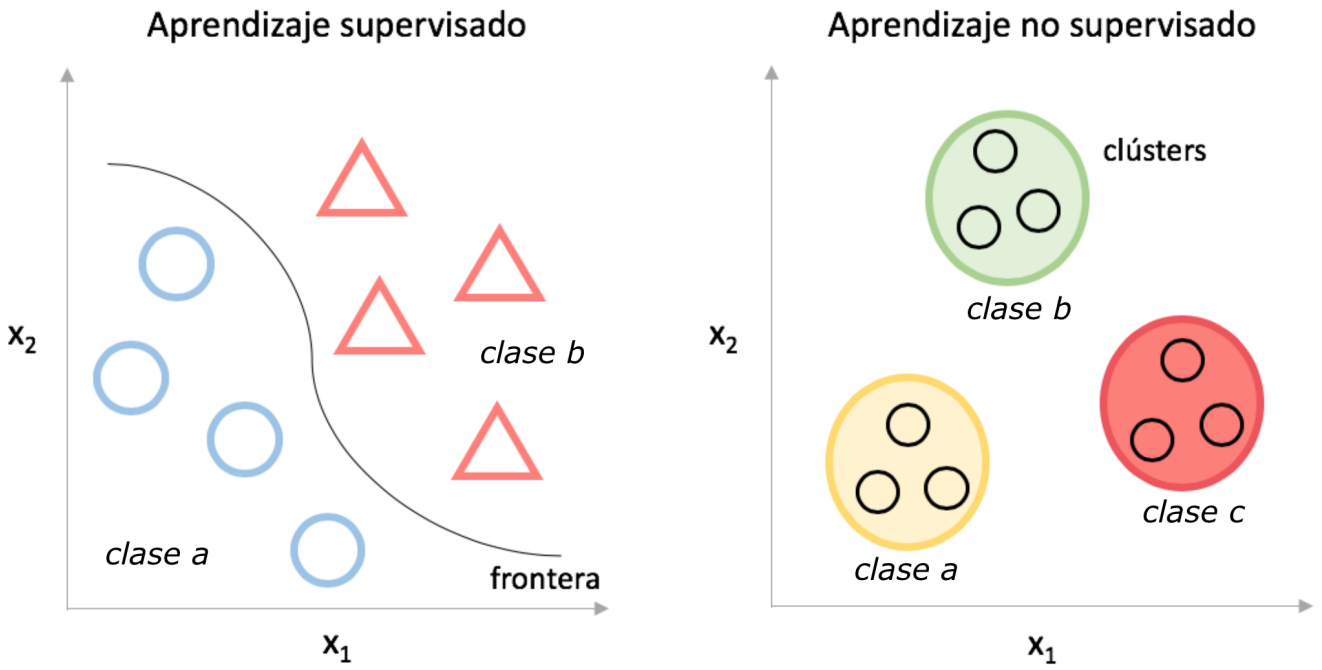

Figura 3.2: Aprendizaje supervisado y no supervisado.

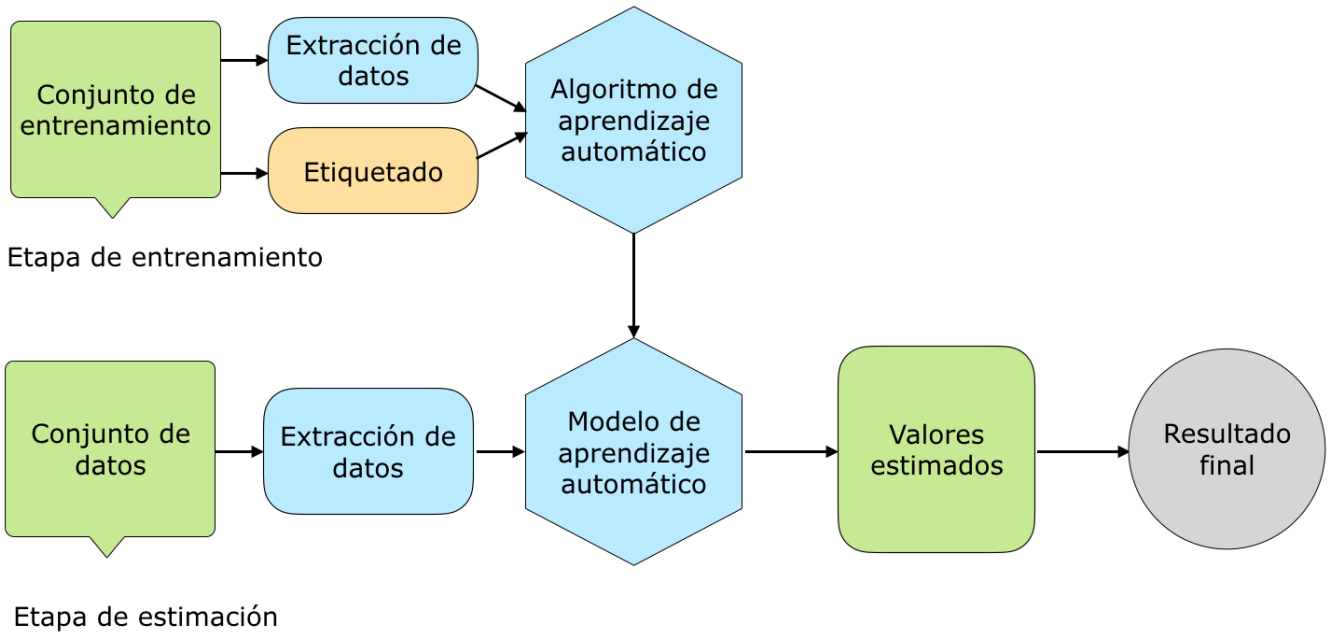

Figura 3.3: Aprendizaje en linea.

vez que entre a producción.

Hurwitz y Kirsch (2018) Los pasos a seguir en el ciclo de aprendizaje automático son los siguientes:

1. Identificación de las fuentes de datos: este paso incluye la identificación de fuentes de datos

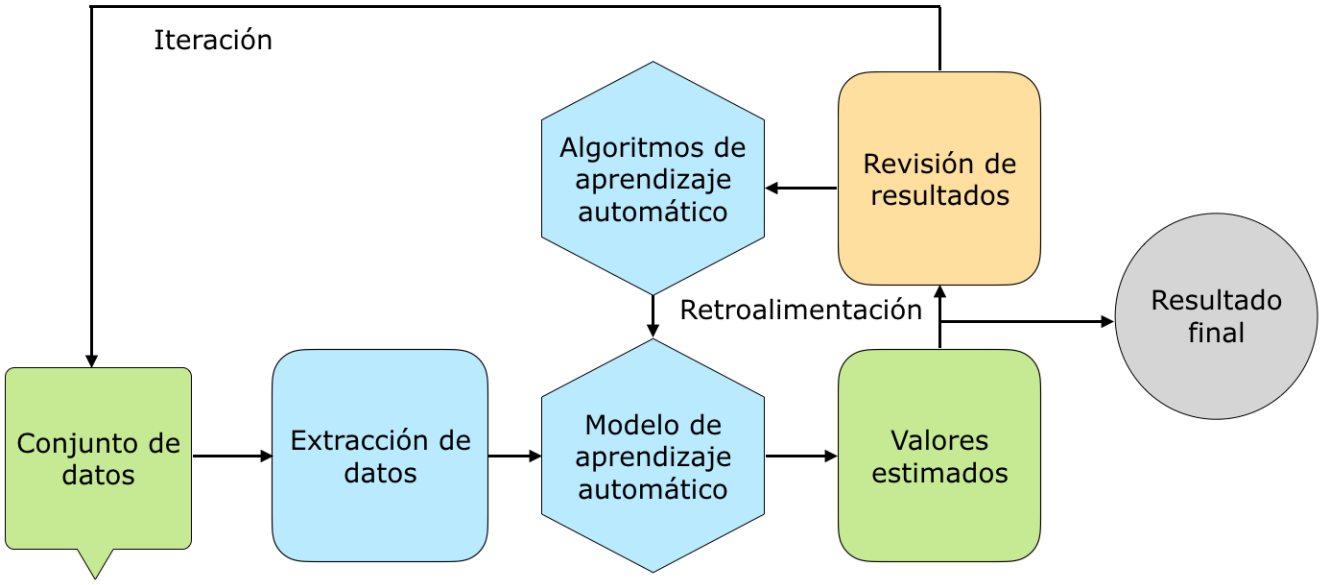

Figura 3.4: Aprendizaje por lotes.

relevantes para resolver el problema. Además, se debe considerar expandir los datos a futuro para mejorar el modelo.

2. Preparación de los datos: los datos deben estar limpios y seguros. La aplicación de aprendizaje automático fallará si se construye basada en datos inexactos.
3. Selección del algoritmo de aprendizaje automático: es posible elegir el algoritmo a utilizar a partir de los datos reunidos y del desafío que se enfrentará.
4. Entrenamiento: consiste en la creación del modelo. Dependiendo del tipo de datos y el algoritmo, el proceso de aprendizaje puede ser supervisado, no supervisado o por refuerzo.
5. Evaluación: en este paso se evalúan los modelos obtenidos en el paso anterior para encontrar aquel que tiene un mejor rendimiento.
6. Implementación: los modelos obtenidos pueden implementarse en aplicaciones en la nube y locales.
7. Predicción: una vez implementado el modelo, se pueden comenzar a hacer predicciones a partir de nuevos datos de entrada.
8. Evaluación de predicciones: la información recopilada del análisis de la validez de las nuevas predicciones es de utilidad para retro-alimentar el ciclo de aprendizaje automático para mejorar la precisión en predicciones futuras.

3.3. Conjuntos de datos

Los algoritmos de aprendizaje automático obtienen a menudo la mayor parte de la atención cuando se habla de este tema. Sin embargo, el éxito depende, en mayor medida, de la selección de buen

conjunto de datos.

Es necesario comprender los datos con los que se trabajarán pues, si se crea un modelo basado en información defectuosa, las predicciones serán inexactas. Además, es necesario identificar la relevancia del conjunto de datos para reducir su complejidad, de manera que el modelo pueda aprender fácilmente de ellos.

Además, es útil conocer el tipo de datos a utilizar para caracterizar el problema de aprendizaje que se desea resolver ya que puede ser de gran ayuda cuando se enfrenta un nuevo desafío: a menudo, los problemas con tipos de datos similares se pueden resolver con técnicas parecidas.

Un conjunto de datos es una colección de datos, es decir, hace referencia al contenido de una tabla de una base de datos o a una única matriz de datos estadísticos, en donde cada columna de la tabla representa una variable particular y, cada fila corresponde a un miembro dado del conjunto de datos en cuestión.

Una vez que se cuenta con el conjunto de datos, se debe realizar pre-procesamiento que incluyen siguientes acciones:

1. Formato: Si los datos están distribuidos en diferentes archivos, deben reunirse en uno solo para formar el conjunto de datos.
2. Limpieza de datos: en este paso deben eliminarse los miembros del conjunto de datos con valores faltantes y eliminar los caracteres no deseados en dichos valores.
3. Extracción de características: este paso se basa en el análisis y optimización de la cantidad de características. Debe averiguarse qué características son importantes para la predicción y seleccionarlas para cálculos más rápido y con bajo consumo de memoria.

Una vez hechos los pasos anteriores, se debe hacer una selección de datos para conformar los dos siguientes subconjuntos de datos necesarios:

- Conjunto de datos de entrenamiento. Estos datos son utilizados para entrenar al algoritmo, de manera que aprenda de los datos de entrada para producir los resultados indicados en las salidas esperadas. Generalmente, este conjunto de datos constituyen al rededor del 80 % de los datos totales.
- Conjunto de datos de prueba. Estos datos son utilizados para evaluar qué tan bien fue entrenado el algoritmo a partir del conjunto de datos de entrenamiento. Este conjunto representa el 20 % de los datos totales y, es importante recalcar que, no es recomendable utilizar datos del conjunto de entrenamiento en él ya que el modelo sabrá de antemano el resultado esperado, por lo que la evaluación no podrá llevarse de forma correcta.

Es posible que al construir los conjuntos de datos, se encuentren algunos problemas que pueden afectar el proceso de aprendizaje. El primer problema que se puede enfrentar es la insuficiencia de datos para entrenar el modelo. Por ejemplo, *Banko y Bill (2001)* mostraron que los algoritmos de aprendizaje automático muy diferentes, incluidos los más simples, funcionaban de manera idéntica

ante un problema complejo una vez que se les proporcionaban datos suficientes. Con ello, demostraron la importancia de tener suficiente cantidad de datos de entrenamiento.

Otro factor que puede afectar el proceso de aprendizaje del modelo es la baja calidad de la información, es decir, que los conjuntos de datos de entrenamiento tengan errores no intencionados, ruido y valores atípicos que dificulten al modelo detectar patrones.

Alpaydin (2014) También, se puede obtener un sobreajuste (overfitting) cuando el modelo es demasiado complejo en relación con la cantidad de datos y su ruido. El problema de sobreajuste significa que el modelo funciona para los datos de entrenamiento, pero no se generaliza bien. En otras palabras, al ocurrir un sobreajuste, el modelo estudia tan bien los datos de entrenamiento que los "memoriza", por lo que, al recibir datos nuevos, su desempeño es pobre. Por otro lado, cuando el modelo es demasiado simple para entender los datos puede ocurrir un subajuste (underfitting).

3.4. Tipos de aprendizaje

Los algoritmos de aprendizaje automático son organizados en una taxonomía basada en el resultado deseado del algoritmo. Las categorías que incluya esta taxonomía son:

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje semi-supervisado
- Aprendizaje por refuerzo

Dependiendo de la naturaleza del problema que se está abordando, se debe elegir entre los diferentes enfoques mencionados anteriormente según el tipo y volumen de los datos. A continuación se explicarán a detalle cada uno de estos conceptos.

3.4.1. Aprendizaje supervisado

En el aprendizaje supervisado generalmente se desea clasificar un conjunto de datos establecido encontrando patrones en los datos que puedan aplicarse a un proceso analítico. Además, es importante mencionar que los datos utilizados en este tipo de aprendizaje tienen características etiquetadas que definen su significado.

Dentro de la categoría de aprendizaje supervisado entran los problemas de clasificación, regresión y pronóstico.

- Clasificación: en las tareas de clasificación, el modelo de aprendizaje automático debe sacar una conclusión de los valores observados y determinar a qué categoría pertenecen las nuevas observaciones.
- Regresión: el modelo de aprendizaje automático debe estimar y comprender las relaciones entre las variables. En análisis de regresión se centra en una variable dependiente y una o varias variables cambiantes.
- Pronóstico: es el proceso de hacer predicciones sobre el futuro en función de los datos pasados y presentes y, comúnmente, es utilizado para analizar tendencias.

Generalmente, cuando los valores de las etiquetas son continuos, se trata de una regresión y, cuando son valores finitos, se trata de un problema de clasificación. En el caso de la regresión, el aprendizaje supervisado ayuda a entender la correlación existente entre las diferentes variables de entrada. Por otro lado, en un problema de clasificación, el aprendizaje automático mapea un vector de entrada en una de varias clases después de estudiar varios ejemplos de entradas-salidas.

El aprendizaje supervisado es la técnica más común para entrenar redes neuronales y árboles de decisión. En ambas técnicas, el éxito obtenido depende en gran medida de la información proporcionada por las clasificaciones predeterminadas.

En el caso de las redes neuronales, la clasificación predeterminada es utilizada para determinar el error de la red y luego ajustarla para minimizar dicho error. Por su parte, en los árboles de decisión las clasificaciones son utilizadas para determinar los atributos que aportan más información y, por lo tanto, pueden utilizarse para resolver el problema de clasificación.

3.4.2. Aprendizaje no supervisado

El aprendizaje no supervisado es más adecuado cuando el problema a resolver requiere utilizar una gran cantidad de datos que no tienen etiqueta.

Ayodele. (2010) Este tipo de aprendizaje tiene dos enfoques. En el primer enfoque se le enseña al modelo dándole algún tipo de sistema de recompensa para indicar el nivel de éxito. Este tipo de aprendizaje generalmente encaja en el marco de un problema de decisión porque el objetivo no es producir una clasificación sino tomar decisiones que maximicen las recompensas. Además, puede ser utilizada una forma de aprendizaje por refuerzo para que el modelo base sus acciones en las recompensas y castigos. De esta manera, el agente sabe qué hacer sin ningún procesamiento ya que sabe la recompensa exacta que espera lograr por cada acción que pueda tomar. Este enfoque puede ser beneficioso en casos en los que el cálculo de cada posibilidad consuma mucho tiempo.

El segundo enfoque de aprendizaje no supervisado se llama clustering (agrupación). En este tipo de aprendizaje, el objetivo no es maximizar una función de utilidad, sino simplemente encontrar similitudes en los datos de entrenamiento. A menudo, los grupo descubiertos coincidirán razonablemente bien con una clasificación intuitiva. Aunque el algoritmo no podrá asignar nombres a los grupos, puede producirlos y luego utilizarlos para asignar nuevas entradas en uno u otro de los

grupos encontrados.

Ghahramani (2004) Los algoritmos de aprendizaje no supervisados están diseñados para extraer la estructura de las muestras de datos. La calidad de una estructura se mide con una función de costo que generalmente se minimiza para inferir parámetros óptimos que caracterizan la estructura oculta en los datos.

3.4.3. Aprendizaje semi-supervisado

El aprendizaje semi-supervisado fue introducido para resolver los problemas encontrados en los dos enfoques de aprendizaje mencionados anteriormente. El principal inconveniente de cualquier algoritmo de aprendizaje supervisado es que el conjunto de datos debe ser etiquetado previamente a mano. Este proceso puede llegar a ser muy costoso, especialmente cuando se trabaja con un volumen grande de datos. Por su parte, la desventaja del aprendizaje no supervisado es que su espectro de aplicación es limitado.

Chapelle y col. (2006) El aprendizaje semi-supervisado se encuentra entre el aprendizaje supervisado y no supervisado: emplea pocos datos etiquetados y muchos datos no etiquetados dentro del conjunto de datos de entrenamiento. Esto permite que el algoritmo deduzca patrones e identifique las relaciones entre su variable a predecir y el resto del conjunto de datos en función de la información que ya tiene.

Los algoritmos que hacen uso del aprendizaje semi-supervisado tratan de explorar la información estructural que contienen los datos no etiquetados con el objetivo de generar modelos predictivos que funcionen mejor que los que sólo utilizan datos etiquetados.

Chapelle y col. (2006) El procedimiento básico a seguir para hacer uso de aprendizaje supervisado consiste en, primero, agrupar datos similares haciendo uso de un algoritmo de aprendizaje no supervisado y luego, usar los datos etiquetados existentes para etiquetar el resto de los datos no etiquetados. Los casos de uso típicos de este tipo de algoritmo tienen una propiedad común entre ellos: la adquisición de datos sin etiquetas es relativamente barata, mientras que etiquetar dichos datos es muy costoso.

3.4.4. Aprendizaje por refuerzo

Hurwitz y Kirsch (2018) El aprendizaje por refuerzo es un modelo de aprendizaje conductual en el que se capacitan a los modelos (agentes) de aprendizaje automático para tomar una secuencia de decisiones.

Este tipo de aprendizaje difiere de los anteriores porque el sistema no está capacitado con el conjunto de datos de muestra, sino, el sistema aprende a través de prueba y error. Debido a ello, una

secuencia de decisiones exitosas dará como resultado que el proceso sea reforzado"porque resuelve mejor el problema en cuestión.

Para que el agente haga lo que el programador desea, se le ofrecen recompensas o penalizaciones por las acciones que realiza pero no se le dan pistas ni sugerencias de cómo resolver el problema. Depende totalmente del modelo descubrir cómo se realiza la tarea siguiendo el objetivo de maximizar la recompensa total.

En este caso, la participación humana se limita a cambiar el entorno de aprendizaje y a ajustar el sistema de recompensas y sanciones. A medida que el agente maximiza la recompensa, es propensa a buscar formas inesperadas de hacerlo. Por ello, se requiere intervención humana para motivar al sistema a realizar la tarea de la forma esperada.

Este enfoque de aprendizaje es de gran utilidad cuando no hay una forma adecuada para realizar una tarea, pero hay reglas que el modelo debe seguir para realizar dicha tarea correctamente.

3.5. Algoritmos de aprendizaje supervisado

En la sección anterior se mencionaron diferentes tipos de aprendizaje. Para el caso específico de este trabajo, el cual está centrado en un problema de clasificación y un problema de regresión, se hizo uso de algoritmos de aprendizaje supervisado.

En esta sección, serán explicados los algoritmos utilizados para resolver dichos problemas.

3.5.1. Regresión lineal

Alpaydin (2014) Aunque este método puede parecer simple a comparación de otros enfoques de aprendizaje automático, la regresión lineal es un método de aprendizaje útil y ampliamente utilizado.

La regresión lineal simple es un enfoque muy sencillo para predecir una respuesta cuantitativa Y sobre la base de una variable de predicción simple X, suponiendo que hay una relación lineal entre X e Y que matemáticamente puede describirse como $Y \approx \beta_0 + \beta_1 X$. De esta ecuación, β_0 y β_1 son dos constantes desconocidas que representan la intersección y la pendiente del modelo lineal, respectivamente. Juntas, β_0 y β_1 se conocen como coeficientes o parámetros del modelo.

Una vez que los datos de entrenamiento son utilizados para producir estimaciones de los valores de β_0 y β_1, es posible estimar nuevos valores haciendo uso del modelo obtenido mediante la expresión $\overline{y} \approx \widehat{\beta}_0 + \widehat{\beta}_1 x$. De la expresión anterior, $\overline{y}$ indica una predicción de Y sobre la base de X = x. En este caso, se utiliza el símbolo de sombrero para denotar el valor estimado de un parámetro o

coeficiente desconocido, o para denotar el valor predicho de la respuesta. El objetivo es encontrar las mejores estimaciones de los coeficientes para minimizar los errores en la predicción de $\widehat{y}$ a partir de x.

El enfoque más común para determinar los valores de estos parámetros es el criterio de mínimos cuadrados. Sea $\widehat{y}_i \approx \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ la predicción para Y basada en el i-ésimo valor de X, entonces $e_i = y_i - \widehat{y}_i$ representa el i-ésimo residual, que es la diferencia entre el i-ésimo valor de respuesta observado y el i-ésimo valor de respuesta que estima el modelo lineal. La suma residual de cuadrados (RSS) es definida como:

$$[RSS = e_1^2 + e_2^2 + ... + e_n^2] \tag{3.1}$$

El enfoque de mínimos cuadrados elige valores de β_0 y β_1 para minimizar el RSS. Los minimizadores son:

$$[\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}] \tag{3.2}$$

$$[\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}] \tag{3.3}$$

Las dos expresiones anteriores definen los coeficientes de mínimos cuadrados para la regresión lineal simple.

Si bien, la regresión lineal simple es un enfoque útil para predecir una respuesta sobre la base de una sola variable predictiva, en la práctica, a menudo se tiene más de un predictor. Debido a esto, es posible extender el modelo de regresión lineal simple descrito anteriormente para acomodar directamente múltiples predictores. Para ello, es posible dar a cada variable predictora un coeficiente de pendiente separado para un solo modelo.

Suponiendo que se tienen n predictores distintos, el modelo de regresión lineal múltiple toma la forma $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$.

De igual manera que en la regresión lineal simple, los coeficientes de regresión β_0, β_1, ..., β_n son desconocidos y deben estimarse utilizando el mismo enfoque de mínimos cuadrados, es decir, deben elegirse valores de β_0, β_1, ..., β_n que minimicen la suma de los residuos al cuadrado.

Los valores β_0, β_1, ..., β_n que minimizan la expresión mostrada anteriormente son las estimaciones de coeficientes de regresión de mínimos cuadrados múltiples. A diferencia de las estimaciones de regresión lineal simples, las estimaciones de coeficientes de regresión múltiple son representadas más fácilmente usando álgebra matricial.

3.5.2. Regresión logística

Este método de aprendizaje modela la probabilidad de que Y pertenezca a una categoría particular, es decir, es utilizado para resolver problemas de clasificación binaria.

Para ello, utiliza la misma estructura que la regresión lineal, pero transformando la variable respuesta en una probabilidad. *Smola y Vishwanathan (2008)* En este caso, se debe modelar p(X) usando una función que proporcione salidas entre 0 y 1 para todos los valores de X. Muchas funciones cumplen con esta descripción, sin embargo, en la regresión logística se utiliza la función logística, la cual es descrita con la siguiente expresión:

$$[p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}] \tag{3.4}$$

Al igual que en la regresión lineal, los coeficientes β_0 y β_1 son desconocidos y deben estimarse con base en los datos de entrenamiento disponibles. En este caso, se puede utilizar el método de máxima verosimilitud. La intuición básica detrás del uso de este método para ajustar un modelo de regresión logística es la siguiente: se buscan estimaciones para β_0 y β_1 de modo que la probabilidad pronosticada $\widehat{p}(x_i)$ de incumplimiento para cada muestra, utilizando la función logística, corresponda lo más cerca posible al estado de la muestra observada.

En otras palabras, se intenta encontrar β_0 y β_1 de modo que al conectar estas estimaciones en el modelo para p(X), se obtenga un número cercano a uno para todas las muestras que incumplieron, y un número cercano a cero para todas las muestras que no lo hicieron. Esta intuición se formaliza utilizando una ecuación matemática llamada función de verosimilitud:

$$[\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_i))] \tag{3.5}$$

Las estimaciones β_0 y β_1 son elegidas para maximizar esta función de probabilidad. La función de máxima verosimilitud es un enfoque muy general que se utiliza para adaptarse a varios modelos no lineales de aprendizaje. En la configuración de regresión lineal, el enfoque de mínimos cuadrados es, de hecho, un caso especial de máxima verosimilitud.

Las salidas obtenidas a partir del modelo de regresión logística se interpretan de la siguiente manera:

- Si la probabilidad obtenida es igual o superior a 0.5, se le asigna la clase 1.
- Si la probabilidad es menor a 0.5, se le asigna la clase 0.

Hasta este punto, se consideró la regresión logística haciendo uso de una sola variable predictora, sin embargo, al igual que en la regresión lineal, es posible extender este método a una regresión logística múltiple que utilice dos o más variables predictoras.

Por analogía con la extensión de regresión lineal simple a múltiple, es posible generalizar de la siguiente manera:

$$[p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}}] \tag{3.6}$$

De esta expresión X = (X_1, ..., X_n) son las n variables predictoras y, los parámetros $\beta_1, ..., \beta_n$ son desconocidos y deben estimarse mediante el método de máxima verosimilitud.

3.5.3. Árboles de decisión

Alpaydin (2014) Un árbol de decisión es una estructura de datos jerárquica de aprendizaje supervisado que implementa la estrategia de divide y vencerás. Es un método no paramétrico eficiente, que puede usarse tanto para clasificación como para regresión.

Un árbol de decisión está compuesto por nodos de decisión internos y hojas terminales. Cada nodo de decisión m implementa una función de prueba $f_m(x)$ con resultados discretos que etiquetan las ramas. Esta función de prueba es implementada al plantear una serie de preguntas sobre las características asociadas con los elementos.

Dada una entrada, en cada nodo, se aplica la función de prueba y se toma una de las ramas posibles que redirigen a un nodo secundario dependiendo del resultado obtenido. Este proceso comienza en la raíz y se repite recursivamente hasta que se llega a un nodo hoja, en cuyo punto el valor descrito en este nodo constituye la salida.

Generalmente, los nodos hoja tienen asociada una clase que le es asignada al elemento de entrada. En algunas variaciones, cada hoja contiene una distribución de probabilidad sobre las clases que estima la probabilidad condicional de que un elemento que llega a la hoja pertenezca a una clase dada.

Los árboles de decisión son comúnmente más fáciles de interpretar que otros clasificadores, como las redes neuronales y las máquinas de vectores de soporte, porque combinan preguntas simples sobre los datos de una manera comprensible. Desafortunadamente, pequeños cambios en los datos de entrada a veces pueden conducir a grandes cambios en el árbol construido.

Cabe recalcar que, los árboles de decisión son lo suficientemente flexibles para manejar elementos con una combinación de características con valores reales y categóricos, así como elementos con algunas características faltantes. Además, son lo suficientemente expresivos como para modelar muchas particiones de los datos que no se logran tan fácilmente con clasificadores que se basan en un límite de decisión único, como la regresión logística o las máquinas de vectores de soporte.

Otra ventaja de los árboles de decisión es que soportan naturalmente problemas de clasificación con más de dos clases y pueden modificarse para manejar problemas de regresión. También, una vez construidos, clasifican nuevos elementos rápidamente.

Los árboles de decisión se construyen agregando nodos de preguntas de forma incremental, utilizando ejemplos de entrenamiento etiquetados para guiar la elección de las preguntas. Idealmente, una sola pregunta simple dividiría perfectamente los ejemplos de entrenamiento en sus clases. Si no existe una pregunta que proporcione una separación tan perfecta, se elige una pregunta que separe los ejemplos de la manera más limpia posible.

Shai y Shai (2014) Una buena pregunta debe dividir un conjunto de elementos con etiquetas heterogéneas en subconjuntos con etiquetas casi homogéneas, estratificando los datos para que haya poca variación en cada estrato. Para evaluar el grado de impureza (falta de homogeneidad), la medida más común para los árboles de decisión es la entropía.

Si se desean clasificar los elementos en m clases usando un conjunto de elementos de entrenamiento E. Sea $p_i(i = 1, ..., m)$ la fracción de elementos de E que pertenecen a la clase i. La entropía de la distribución de probabilidad $(p_i)_{i=1}^m$ da una medida razonable de la impureza del conjunto E. La entropía, $\sum_{i=1}^m -p_i log_2(p_i)$, es más baja cuando un solo p_i es igual a 1 y todos los demás son 0, mientras que se maximiza cuando todos los p_i son iguales.

Dada una medida de impureza I, se elige aquella pregunta que minimiza el promedio ponderado de la impureza de los nodos hijos resultantes. Si I es la función de entropía, la diferencia entre la entropía de la distribución de las clases en el nodo principal y el promedio ponderado de la entropía de los hijos se denomina ganancia de información. Es decir, la ganancia de información es expresada como $I(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} I(S_v)$.

Se continúan seleccionando preguntas recursivamente para dividir los elementos de entrenamiento en subconjuntos cada vez más pequeños, lo que resulta en un árbol. Un aspecto crucial para aplicar los árboles de decisión es limitar su complejidad para que no se sobre-ajusten a los ejemplos de entrenamiento. Una técnica es detener la división cuando ninguna pregunta reduce la entropía de los subconjuntos más que una pequeña cantidad. O bien, alternativamente, se puede elegir construir el árbol completamente hasta que no se pueda subdividir más la hoja y, posteriormente, "podar" el árbol para evitar el sobre-ajuste.

3.5.4. Máquinas de vectores de soporte

Shai y Shai (2014) Una máquina de vectores de soporte es un modelo lineal para problemas de clasificación y regresión. Puede resolver tanto problemas lineales como no lineales y funciona bien para muchos problemas prácticos.

La idea de una máquina de vectores de soporte es simple: dados los datos de entrenamiento etiquetados, el algoritmo genera una línea o un hiperplano óptimo para separar los datos en clases y clasificar nuevas entradas.

En matemáticas, un hiperplano H es un subespacio lineal de un espacio vectorial V, de modo que la base de H tiene una cardinalidad menos que la cardinalidad de la base para V. En otras

palabras, si V es un espacio vectorial n-dimensional, H es un subespacio (n-1)-dimensional.

Para el caso de la clasificación, las máquinas de vectores de soporte buscan el hiperplano que obtenga la superficie óptima que delimite cada una de las clases involucradas en el problema. Mientras que, en un problema de regresión, obtienen una curva que modele la tendencia de los datos para, a partir de ella, estimar cualquier otro dato a futuro.

Es bien sabido que el rendimiento de una máquina de vectores de soporte depende de una buena configuración de los parámetros C, γ y la función kernel.

El parámetro C le indica a la máquina de vectores de soporte cuánto desea evitar clasificar erróneamente cada ejemplo de entrenamiento. Para valores grandes de C, la optimización elegirá un hiperplano de menor margen si ese hiperplano hace un mejor trabajo al clasificar correctamente todos los puntos de entrenamiento.

Por otro lado, el parámetro γ define hasta donde llega la influencia de un solo ejemplo del conjunto de datos de entrenamiento. Los valores bajos de gamma indican que los puntos alejados de la línea de separación plausible se consideran en el cálculo de la línea de separación. Por su parte, un valor alto en gama significa que los puntos cercanos a la línea plausible se consideran dentro del cálculo.

Finalmente, la función kernel es utilizada cuando los problemas no son lineales. Este tipo de funciones se encargan de trasladar los problemas no lineales a un hiperplano en donde la solución es lineal y, por lo tanto, más sencilla de obtener. Una vez resuelto el problema, la solución se transforma de nuevo al espacio original. *Alpaydin (2014)* Entre los kernels más populares para usar con máquinas de vectores de soporte se encuentran:

- **Kernel lineal**. Cuantifica la similitud de un par de observaciones usando la correlación de Pearson. Se expresa mediante la ecuación

$$[K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x^{\smile} i' j] \tag{3.7}$$

- **Kernel polinómico**. Un kernel polinómico de grado d (siendo d >1) permite un límite de decisión mucho más flexible. Su caracteriza por la expresión

$$[K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x^{\smile} i' j)^d] \tag{3.8}$$

- **Kernel radial**. Tiene un comportamiento muy local, en el sentido de que sólo las observaciones de entrenamiento cercanas a una observación de prueba tendrán efecto sobre su clasificación. La expresión incluye un parámetro γ que es una constante positiva que, cuanto mayor sea, mayor flexibilidad le proporcionará a la máquina de vectores de soporte. Sin embargo, es importante tener en cuenta que una mayor flexibilidad puede provocar un problema

de sobre-ajuste a los datos de entrenamiento.

$$[K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2)] \tag{3.9}$$

Una de las ventajas de las máquinas de vectores de soporte, tanto en problemas de clasificación como de regresión, es que pueden utilizarse para evitar las dificultades de usar funciones lineales en espacios con características de alta dimensión.

3.5.5. Redes neuronales artificiales

Shai y Shai (2014) Una red neuronal artificial es un sistema de aprendizaje supervisado construido con un conjunto de elementos simples, llamados perceptrones o neuronas, que se encuentran organizados en capas interconectadas. Cada perceptrón es capaz de tomar decisiones simples con las cuales alimenta a otros perceptrones.

Alpaydin (2014) Un perceptrón es un algoritmo de clasificación binaria cuyo modelado fue basado en el funcionamiento de una neurona del cerebro humano. Este algoritmo, a pesar de tener una estructura simple, es capaz de aprender y resolver problemas complejos.

Básicamente, el proceso de aprendizaje de un perceptrón es el siguiente:

1. El perceptrón se alimenta de las variables de entrada, las multiplica por sus respectivos pesos y suma los resultados obtenidos.
2. Suma el número uno multiplicado por un peso de sesgo.
3. Ingresa el resultado de la suma a la función de activación. En un perceptrón simple, la función de activación, generalmente, es una función escalonada.
4. Genera los resultados. El resultado de la función de activación es la salida del algoritmo.

En conjunto, una red neuronal artificial es capaz de emular prácticamente cualquier función y resolver cualquier problema, siempre y cuando se tengan muestras suficientes de entrenamiento y la potencia computacional adecuada.

Existen dos tipos de redes neuronales artificiales: superficiales y profundas. Las redes neuronales superficiales tienen únicamente tres capas de neuronas organizadas de la siguiente manera:

1. Una capa de entrada que recibe las variables o entradas independientes del modelo.
2. Una capa oculta encargada de procesar las entradas.
3. Una capa de salida encargada de entregar los resultados después de procesar los datos de entrada.

Goodfellow y col. (2016) Por otro lado, una red neuronal profunda tiene una estructura similar, sin embargo, ésta se caracteriza por tener dos o más capas ocultas de neuronas. En 2016, Goodfellow, Bengio y Courville demostraron que si bien, las redes neuronales superficiales son capaces de abordar problemas complejos, las redes profundas pueden llegar a ser más precisas a medida que se agregan más capas de neuronas. Además, descubrieron que las capas adicionales son útiles hasta un límite de 9 a 10, después de lo cual su poder predictivo comienza a disminuir.

Después de que se define la estructura de una red neuronal es necesario asignarle pesos iniciales, con los cuales se genera una predicción inicial. El resultado es evaluado mediante una función de error que es utilizada para definir que tan lejos está el modelo de la predicción verdadera.

El objetivo es encontrar los pesos óptimos para cada perceptrón, de modo que los resultados obtenidos sean más precisos y minimicen la función de error. Existen muchos algoritmos posibles para realizar esto; por ejemplo, podría utilizarse una búsqueda por fuerza bruta para encontrar los pesos que generen el error más pequeño. Sin embargo, mientras más grande es una red neuronal, es necesario utilizar un algoritmo que sea eficiente computacionalmente.

El algoritmo de retro-propagación es el más utilizado debido a que es capaz de descubrir los pesos óptimos con una rapidez relativa, incluso para una red con millones de pesos. *Shai y Shai (2014)* Los pasos que realiza el algoritmo de retro-propagación, a grandes rasgos, son los siguientes:

1. Los pesos se inicializan y las entradas del conjunto de datos de entrenamiento son introducidos en la red. Los datos son procesados y el modelo genera una predicción inicial.
2. A partir de la predicción inicial, se calcula el resultado de la función de error para verificar que tan lejos está el valor predicho del valor conocido.
3. Primero, el algoritmo calcula las derivadas parciales del error con respecto a los pesos que unen la última capa oculta con la capa de salida. Luego, el algoritmo calcula las derivadas parciales del error con respecto a los pesos que unen la capa de entrada con la capa oculta. El resultado de la retro-propagación es un conjunto de pesos que minimizan la función de error.
4. Los pesos se actualizan.

Por lo general, el algoritmo de retro-propagación es ejecutado luego de procesar un lote de muestras del conjunto de datos de entrenamiento. El tamaño de dicho lote y el número de lotes utilizados son dos hiperparámetros importantes que deben ajustarse para obtener mejores resultados.

Finalmente, es importante mencionar otro elemento necesario para las redes neuronales artificiales: la función de activación. Una función de activación es una ecuación matemática que determina la salida de cada elemento en la red neuronal. Esta función toma la entrada de cada neurona y la transforma en una salida cuyo valor, generalmente, se encuentra dentro del rango de 0 y 1 o de -1 a 1.

En una red neuronal, los valores de entrada se introducen en las neuronas de la red. Cada neurona tiene un peso y las entradas se multiplican por dicho peso para, posteriormente, alimentar a la

función de activación.

La salida de cada neurona es la entrada de las neuronas de la siguiente capa de la red, por lo que las entradas caen en cascada a través de múltiples funciones de activación hasta que. finalmente, la capa de salida genera una estimación. La derivada de la función de activación ayuda a la red a aprender mediante el algoritmo de retro-propagación explicado anteriormente.

La selección de una función de activación es crítica para la construcción y entrenamiento de la red. *Sharma y col. (2020)* Las funciones de activación más comúnmente utilizadas son las siguientes:

- Función sigmoide: tiene un gradiente suave y genera valores entre cero y uno. Para valores muy altos o bajos de los parámetros de entrada, la red puede ser muy lenta para alcanzar una predicción.
- Función TanH: se encuentra centrada en cero, lo que facilita el modelado de entradas que son fuertemente negativas, muy positivas o neutrales.
- Función ReLu: computacionalmente es muy eficiente pero no puede procesar entradas que se acercan a cer o son negativas.
- Función Leaky ReLu: tiene una pequeña pendiente positiva en su área negativa, lo que le permite procesar valores iguales a cero o negativos.
- Función Softmax: es una función de activación especial que se utiliza para las neuronas de salida. Normaliza las salidas para cada clase entre cero y uno y, devuelve la probabilidad de que la entrada pertenezca a una clase específica.

Capítulo 4

Clasificación de los Fluidos Petroleros Utilizando TAA

En este trabajo se plantea el uso algoritmos de aprendizaje supervisado para lograr una clasificación binaria de los tipos de aceites (1: negro, 0: volátil) de manera sencilla y práctica cuando solamente se cuenta con datos de fácil obtención.

Para armar el conjunto de datos se utilizaron reportes de resultados de experimentos PVT, de donde se obtuvieron para las entradas de los algoritmos los valores de la densidad relativa del gas, la densidad relativa del aceite, la temperatura del yacimiento, la profundidad del yacimiento, la presión del pozo, viscosidad del aceite a 20º [C] y el Rs de 106 pozos de una región de México. Por su parte, para las salidas, es decir, los tipos de aceite, se utilizaron los criterios de clasificación de *Méndez y col. (1979)*, mostrados en la **tabla 4.1**, para determinar a qué tipo de yacimiento pertenecía cada muestra de los pozos elegidos.

Propiedad	*Aceite negro*	*Aceite volátil*
ρ_{ro} *[1]*	*>0.85*	*0.75 - 0.85*
Rs [m³/m³]	*<200*	*200 - 1000*
Bob [m³/m³]	*<2.0*	*≤2.0*

Tabla 4.1: Clasificación de Méndez para los tipos de aceites.

Como se puede observar en la **tabla 4.2**, en total se utilizaron 106 pozos de una región de la República Mexicana para entrenar y validar los modelos de clasificación. Además, en la **tabla 4.3** pueden verse los valores utilizados para normalizar las variables de entrada del conjunto de datos.

Posteriormente, a partir de los resultados obtenidos mediante el mapa de correlación mostrado en la **figura 4.1** se hizo una reducción de las variables de entrada para elegir únicamente aquellas que tuvieran una correlación alta con la salida esperada de los modelos, es decir, aquellas cuya

Tipo	Cantidad [1]
Aceite negros	53
Aceite volátil	53
Total	106

Tabla 4.2: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de clasificación.

Rango de datos	Min	Max	Promedio
ρro [1]	0.798	0.933	0.854
Rs [m³/m³]	11.1	783.1	221.8
Ty [°C]	42.6	162.8	116.4

Tabla 4.3: Rango de valores de las propiedades del conjunto de datos.

correlación se encuentra en el rango de [0.66, 1] y [-1, -0.66]. A partir de esto, se seleccionaron únicamente la temperatura del yacimiento, la densidad relativa del aceite y el R_s.

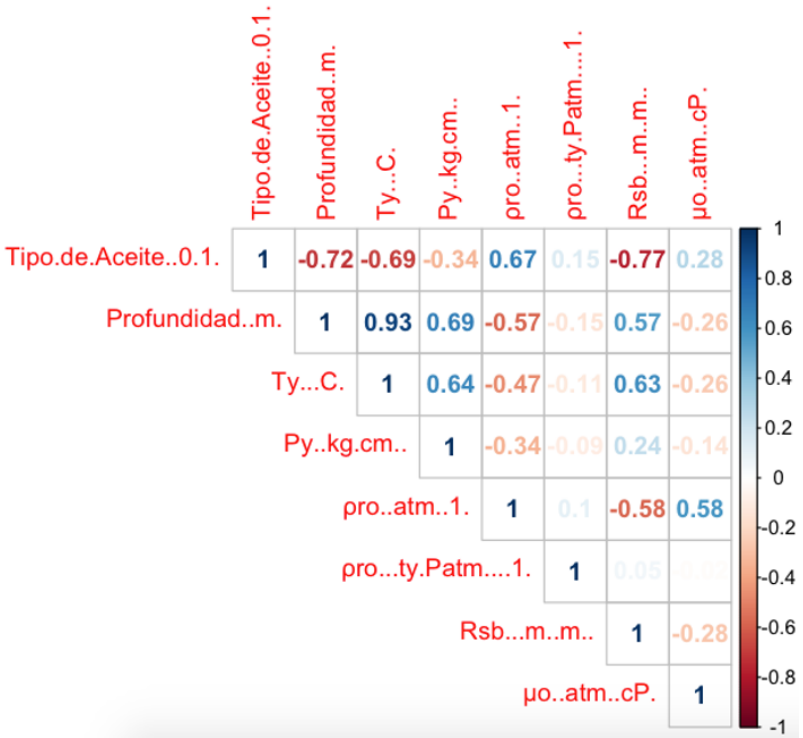

Figura 4.1: Mapa de correlación para elegir las variables de clasificación.

En el caso de la profundidad y la temperatura del yacimiento, se puede observar que ambas tienen un valor de correlación alto con el tipo de aceite y entre ellas. Debido a esto, se decidió eliminar la profundidad como variable de entrada ya que puede no ser viable utilizarla en algunos casos debido a que por eventos geológicos, pueden existir excepciones (fallamientos, erosiones, entre otros) que provoquen ruido durante el entrenamiento de los modelos (*Shizhen y col. (2016)*).

Luego, antes de entrenar los modelos de aprendizaje automático, se llevó a cabo una normalización de datos mediante el método min-max haciendo uso de la expresión 4.1.

$$v' = \frac{v - min_A}{max_A - min_A}, \tag{4.1}$$

de donde v' es el valor normalizado, v es el valor a normalizar, min_A es el valor mínimo del conjunto de datos y max_A es el valor máximo del conjunto de datos.

La normalización es una técnica necesaria utilizada como parte de la preparación de datos para el aprendizaje automático cuyo objetivo es cambiar los valores numéricos del conjunto de datos para usar una escala común, sin distorsionar las diferencias en los rangos de valores ni perder información, de esta manera, los algoritmos modelan los datos correctamente.

Finalmente, se separaron los datos en dos conjuntos: un conjunto de entrenamiento con el 80 % de los datos y otro conjunto de validación con el 20 % restante.

Para crear los modelos de clasificación se utilizaron herramientas de código abierto con las cuales se realizaron múltiples pruebas con la finalidad de optimizar los parámetros de entrenamiento para obtener los mejores resultados sin caer en el efecto de sobreajuste.

4.1. Regresión logística

Para el modelo de regresión logística fue necesario eliminar la variable R_s ya que, debido a la alta correlación de esta variable con la temperatura del yacimiento y la densidad relativa, el algoritmo presentaba problemas para converger.

A partir de los valores de los coeficientes obtenidos mediante este algoritmo, mostrados en la **tabla 4.4**, se construyó la expresión 4.2, a partir de la cual es posible realizar la clasificación de los aceites.

$$Tipo = -7{,}196 - 7{,}165(Temperatura) + 29{,}531(Densidadrelativa). \tag{4.2}$$

A partir de la expresión 4.2, se realizó la clasificación de 20 aceites de yacimientos petroleros pertenecientes al conjunto de datos de prueba, a partir de los cuales se obtuvieron los resultados

Variable	*Coeficiente*
Intersección	*-7.196*
Temperatura	*-7.165*
Densidad Relativa	*29.531*

Tabla 4.4:Coeficientes de la regresión logística.

Modelo Logístico		
	Volátil	Negro
Volátil	10	0
Negro	3	7

Cuadro 4.1: Matriz de confusión de la clasificación mediante regresión logística.

mostrados en la matriz de confusión mostrados en el **cuadro 4.1**.

De la matriz de confusión mostrada en el **Cuadro 4.1**, se puede observar que el modelo clasificó correctamente 10 aceites volátiles y 7 aceites negros. Y, por otro lado, clasificó incorrectamente 3 aceites negros como volátiles. Con estos resultados, se obtuvieron una exactitud del 85 % y una tasa de error del 0.15 %.

4.2. Árboles de decisión

Para el caso de los árboles de decisión, las reglas de inducción a las que se llegaron mediante este algoritmo son mostradas en la **figura 4.2**:

1. Si $R_s \geq 0{,}24$ y $\rho_{ro} < 0{,}46$ entonces $Tipo = volátil$.
2. Si $R_s \geq 0{,}24$ y $\rho_{ro} \geq 0{,}46$ entonces $Tipo = negro$.
3. Si $R_s < 0{,}24$ y $\rho_{ro} < 0{,}15$ entonces $Tipo = volátil$.
4. Si $R_s < 0{,}24$ y $\rho_{ro} \geq 0{,}15$ entonces $Tipo = negro$

Puede observarse que el mismo algoritmo "poda"las variables no significativas para evitar un sobreajuste. En este caso, la variable eliminada es la temperatura del yacimiento.

Con el uso de las reglas de inducción obtenidas, se realizó la clasificación de 20 aceites de yacimientos petroleros pertenecientes al conjunto de datos de prueba, con los cuales se obtuvo la matriz de confusión mostrada en el **cuadro 4.2** .

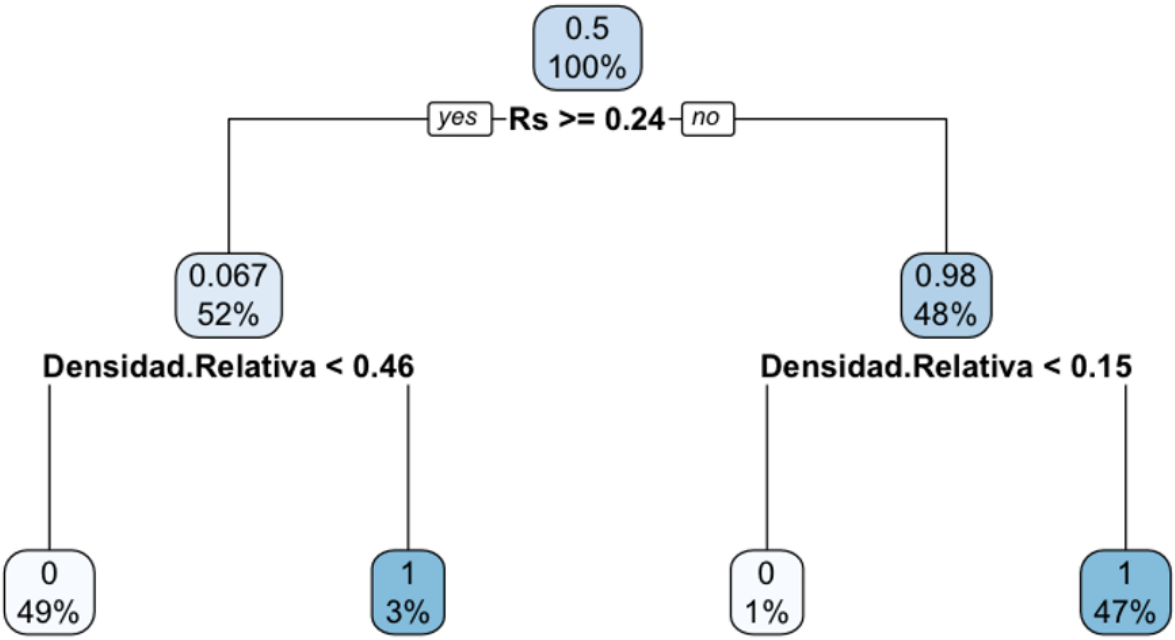

Figura 4.2: Árbol de decisión generado para clasificar aceites petroleros.

Árboles de decisión		
	Volátil	Negro
Volátil	10	0
Negro	0	10

Cuadro 4.2: Matriz de confusión de la clasificación mediante árboles de decisión.

El **Cuadro 4.2**, muestra la matriz generada con los resultados, se puede observar que el modelo clasificó correctamente 10 aceites volátiles y 10 aceites negros. Con estos resultados, se obtuvieron una exactitud del 100 % y una tasa de error del 0 %.

4.3. Redes neuronales artificiales

Finalmente, se resolvió el problema de clasificación mediante redes neuronales artificiales. En este caso, se entrenó una red neuronal mediante el algoritmo de retro-propagación con la estructura mostrada en la siguiente estructura.

1. Una capa de entrada con tres neuronas correspondientes a cada una de las variables seleccionadas.
2. Dos capas ocultas de dos neuronas y una neurona, respectivamente.
3. Una capa de salida de una neurona perteneciente al tipo de aceite.

A partir de la red neuronal artificial generada, se clasificaron los 20 aceites pertenecientes al conjunto de datos de prueba, obteniendo la matriz de confusión mostrada en el **Cuadro 4.3**

Como se puede observar, la red neuronal artificial logró clasificar correctamente 10 aceites volá-

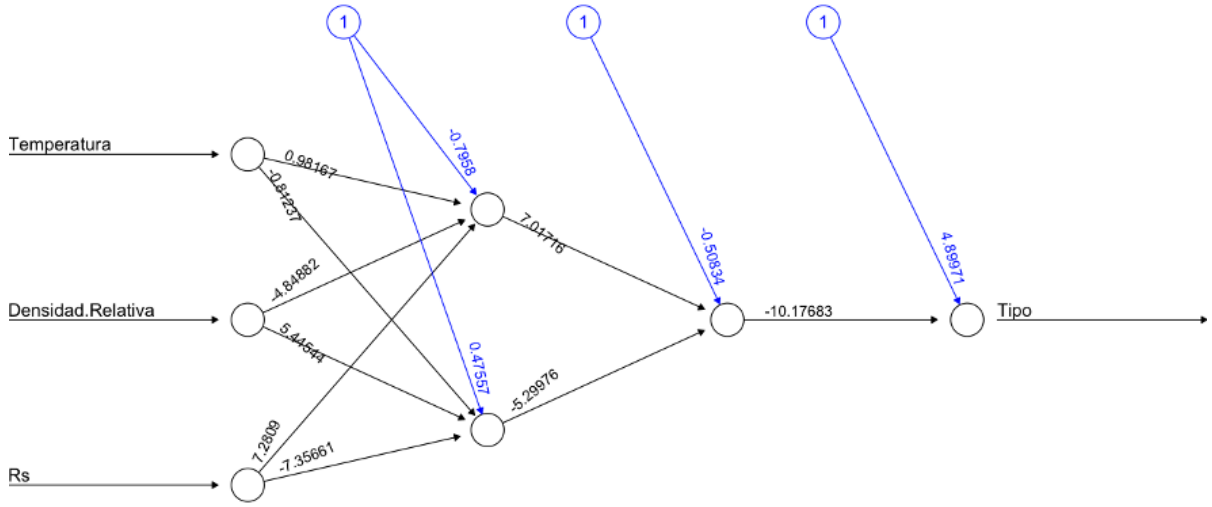

Figura 4.3: Red neuronal obtenida para clasificar aceites petroleros.

Redes neuronales artificiales		
	Volátil	Negro
Volátil	10	0
Negro	0	10

Cuadro 4.3: Matriz de confusión de la clasificación mediante redes neuronales artificiales.

tiles y 10 aceites negros, obteniendo los mismos resultados que los árboles de decisión. Con estos resultados, se obtuvieron una exactitud del 100 % y una tasa de error del 0 %.

Los resultados obtenidos en este capítulo pueden resumirse en la tabla mostrada a continuación.

Se observa una exactitud de clasificación perfecta en el conjunto de datos de prueba para los métodos de árboles de decisión y redes neuronales mientras que el modelo de regresión logística, al ser un algoritmo más simple, presenta resultados inferiores en cuanto a la exactitud obtenido al clasificar.

Apegándose al conocimiento generado previamente por investigadores de la Ingeniería Petrolera, puede concluirse que las técnicas de aprendizaje automático representan una alternativa eficaz y fiables para realizar la tarea de clasificación de tipos de aceite de acuerdo a sus características físicas fundamentales, siempre y cuando se haga una selección correcta de las variables de entrada para asegurar el correcto entrenamiento de los modelos.

#Pozo	Ty [°C]	ρro [1]	Rs [m³/m³]	Tipo (Valor real)	Modelo Logístico	Árboles de Decisión	Redes Neuronales
1	0.544	0.700	0.187	1	1	1	1
2	0.502	0.388	0.198	1	1	1	1
3	0.802	0.201	0.377	0	0	0	0
4	0.669	0.347	0.453	0	0	0	0
5	0.968	0.283	0.435	0	0	0	0
6	0.386	0.539	0.125	1	1	1	1
7	0.037	0.455	0.090	1	1	1	1
8	0.200	0.327	0.138	1	1	1	1
9	0.881	0.140	0.358	0	0	0	0
10	0.801	0.242	0.361	0	0	0	0
11	0.819	0.384	0.394	0	0	0	0
12	0.236	0.264	0.225	1	0	1	1
13	0.935	0.319	0.134	1	0	1	1
14	0.943	0.347	0.287	0	0	0	0
15	0.394	0.619	0.117	1	1	1	1
16	0.561	0.613	0.087	1	1	1	1
17	0.727	0.320	0.240	0	0	0	0
18	0.769	0.334	0.389	0	0	0	0
19	0.968	0.230	0.248	0	0	0	0
20	0.478	0.341	0.224	1	0	1	1

Tabla 4.5: Evaluación de los modelos de clasificación con el conjunto de datos de prueba (no conocido previamente por la red). Entradas normalizadas con el método de min / max.

Algoritmo	Exactitud	Tasa de error
ρro [1]	85 %	15 %
Rs [m³/m³]	100 %	0 %
Ty [°C]	100 %	0 %

Tabla 4.6: Comparación de los resultados obtenidos con los modelos de clasificación entrenados.

Capítulo 5

Estimación de las Propiedades PVT Utilizando TAA

En este capítulo, se plantea el uso de algoritmos de aprendizaje automático para estimar las propiedades PVT de aceites negros y volátiles a partir de otras propiedades de fácil obtención. Las propiedades PVT calculadas se presentan a diferentes condiciones de presión para generar la curva que describe el comportamiento del fluido desde superficie hasta el yacimiento, tal como se hace en las pruebas de laboratorio PVT.

Las propiedades PVT que se busca estimar de los aceites a condiciones de temperatura de yacimiento conforme varía la presión son las siguientes:

- P_b
- Curva de B_o
- Curva de ρ_o
- Curva de R_s
- Curva de μ_o

Para armar el conjunto de datos utilizado para entrenar, probar y validar los algoritmos se utilizaron reportes de resultados de experimentos PVT (expansión a composición constante y liberación diferencial) completos, de donde se obtuvieron para posibles entradas de los algoritmos los valores de la densidad relativa del gas, la densidad relativa del aceite a condiciones de tanque de almacenamiento (aceite muerto), la temperatura del yacimiento, la profundidad del yacimiento, la presión del pozo, viscosidad del aceite a 20º $[C]$ y la R_s de pozos de una región de México.

Las propiedades de entrada a utilizar fueron elegidas a partir de su valor de correlación con las propiedades a calcular. Dichos valores de correlación pueden observarse en la **figura 5.1**.

Por su parte, para las salidas, se obtuvieron los valores de las propiedades PVT objetivo reportadas a cinco diferentes presiones ($\frac{1}{5}P_b,\ \frac{2}{5}P_b,\ \frac{3}{5}P_b,\ \frac{4}{5}P_b\ y\ P_b$).

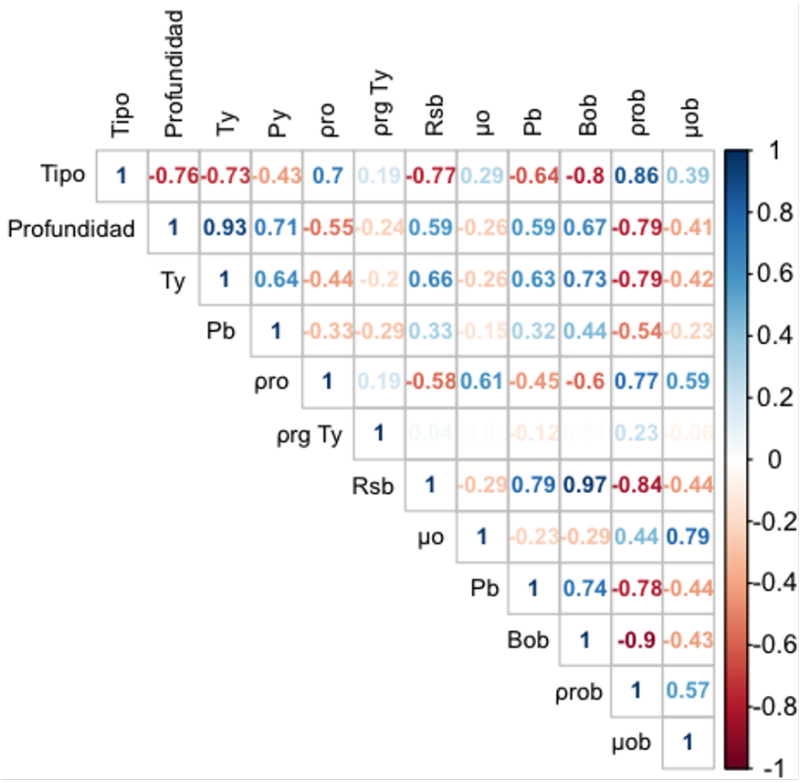

Figura 5.1: Mapa de Correlación entre las propiedades PVT del aceite.

Es importante recalcar que, como el comportamiento de las propiedades en región saturada es diferente al de la región bajo saturada, se decidió generar dos modelos diferentes para cada región que, posteriormente, pueden ser combinados para generar la curva completa del comportamiento del fluido con la variación de presión.

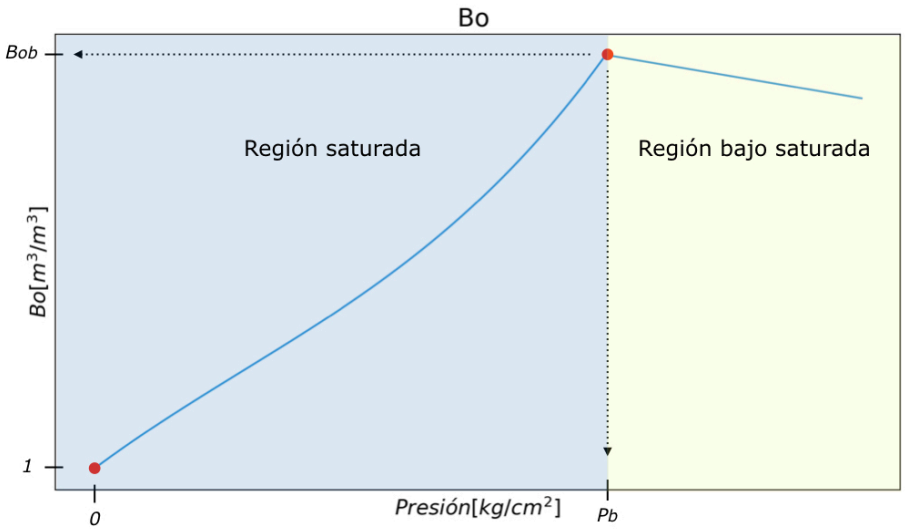

Figura 5.2: Ejemplo de la gráfica objetivo a generar para cada propiedad con los modelos propuestos.

5.1. Preparación de salidas de la Región Saturada

Para llevar las salidas a las mismas condiciones se propuso encontrar cinco puntos a valores arbitrarios de presión, definiendo estos valores como: $\frac{1}{5}P_b$, $\frac{2}{5}P_b$, $\frac{3}{5}P_b$, $\frac{4}{5}P_b$ *y* P_b.

Para encontrar los valores de las propiedades a dichas presiones es necesario encontrar, en primer lugar, la Pb registrada en los reportes PVT para, posteriormente, encontrar la función matemática del grado adecuado que describen el comportamiento de las propiedades en la región saturada (desde $\frac{1}{5}P_b$ hasta P_b), mediante la técnica de regresión polinomial por mínimos cuadrados. Se decidió usar un polinomio de tercer grado para representar el valor de las distintas propiedades PVT a lo largo de la región saturada debido a que la geometría caprichosa mostrada es demasiado compleja para ajustarse a un polinomio de segundo grado de forma aceptable.

$$f(x) = A_1X^3 + A_2X^2 + A_3X + A_4 \tag{5.1}$$

Por ejemplo, para un modelo de Bo en la región saturada, cuya curva es:

$$B_{osat}(X) = (3{,}16 * 10^{-8})X^3 - (1{,}49 * 10^{-5})X^2 + (3{,}4 * 10^{-3})X + 1{,}057 \tag{5.2}$$

En la cual, la Pb reportada es 293 kg/cm2 , de esta expresión, se obtuvieron los puntos mostrados en la **tabla 5.1** , con los cuales es posible entrenar el modelo de la región saturada.

Propiedad	*(1/5) Pb*	*(2/5) Pb*	*(3/5) Pb*	*(4/5) Pb*	*Pb*
Presión [kg/cm²]	*58.6*	*117.2*	*175.8*	*234.4*	*293*
Bo [m³/m³]	*1.214*	*1.306*	*1.372*	*1.450*	*1.579*

Tabla 5.1: Ejemplo de salidas recolectadas para la región saturada.

5.2. Preparación de salidas de la Región Bajo Saturada

Para simplificar el conjunto de salidas de la región bajo saturada a un solo punto y obtener mejores resultados con los modelos, se decidió calcular la función de grado 1 que describe la recta entre los valores registrados en la P_b y los valores registrados en la P_y.

$$f(x) = A_1X + A_2 \tag{5.3}$$

Posteriormente, se calcula la diferencia entre el valor de la propiedad objetivo registrado en la P_b y el valor de la propiedad objetivo evaluada a $P_b + 200[\frac{kg}{cm^2}]$ en la función lineal encontrada para la región bajo saturada (valor de presión arbitrario para normalizar).

El proceso para encontrar el valor de las propiedades a P_y después de estimar esta diferencia con los modelos es sencillo: se divide dicha diferencia entre 200 y se multiplica por el valor de la P_y menos el valor de la P_b para, finalmente, sumarse o restarse dependiendo del valor de Pb estimado con el modelo de la región saturada.

Por ejemplo, para un pozo cuya curva en la región bajo saturada es:

$$B_{bsat} = -3{,}29 * 10^{-4}X + 1{,}668 \tag{5.4}$$

donde X toma un valor de presión dado y tenemos que

$$Pb = 293kg/cm^2$$
$$Py = 309{,}92kg/cm^2$$

Los valores calculados para entrenar al modelo de regresión son los mostrados en la **tabla 5.2**.

Propiedad	*Pb*	*Pb + 200*	△
Presión [kg/cm²]	*293*	*493*	*200*
Bo [m³/m³]	*1.572*	*1.506*	*0.066*

Tabla 5.2: Ejemplo de salida recolectada para la región bajo saturada.

De la anterior operación encontramos que:

$$(B_o@Pb + 200[kg/cm^2] - B_o@Pb) = 0{,}066[1] \tag{5.5}$$

De forma inversa,se muestra el proceso para encontrar B_o en P_y a partir de conocer $(B_o@Pb + 200[kg/cm^{]-B_o@Pb)}$:
$B_o\ en\ P_y[m^3/m^3] = B_o@pb - [((B_o@pb - B_o@p_b + 200)/200) * (P_y - P_b)]$
$B_o\ en\ P_y[m^3/m^3] = 1{,}572 - [((1{,}572 - 1{,}506)/200) * (309{,}92 - 293)]$
$B_o\ en\ P_y[m^3/m^3] = 1{,}566$

Una vez obtenidos estos puntos, se utiliza de nuevo el método de mínimos cuadrados para encontrar la recta que describe a la propiedad en la región bajo saturada.

Cabe recalcar que para crear los conjuntos de datos utilizados para elaborar los modelos de estimación, se utilizó la siguiente configuración haciendo una selección de datos de forma aleatoria:

Conjunto	*Porcentaje del total de los datos [%]*
Conjunto de prueba	*20*
Conjunto de entrenamiento	*80*

Tabla 5.3: Configuración utilizada para construir los conjuntos de datos.

Tal y como se recomienda en el Capítulo 3 de este trabajo, se utilizó el 80 % de los datos para formar el conjunto de datos de entrenamiento y el 20 % para conformar el conjunto de datos de prueba

5.3. Estimación de P_b

5.3.1. Preparación de entradas

A partir del mapa de correlación mostrado en la **figura 5.1**, se seleccionaron como entradas para los algoritmos de aprendizaje automático las propiedades con valores de correlación más alto, es decir, la temperatura del pozo (T_y), la relación gas-aceite disuelto (R_s) y la densidad relativa del aceite (ρ_{ro}). En el caso de la profundidad del pozo (P_y), a pesar de presentar una alta correlación con la presión de burbuja, fue descartada debido a su alta dependencia con la profundidad del pozo.

El conjunto de datos utilizado para el entrenamiento de los modelos de regresión para calcular la P_b se compone de muestras de 105 pozos con las características mostradas en la **tabla 5.4**.

Tipo	*Cantidad [1]*
Aceite negros	*53*
Aceite volatil	*52*
Total	*105*

Tabla 5.4: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de Pb.

Rango de datos	*Min*	*Max*	*Promedio*
ρro [1]	*0.798*	*0.933*	*0.854*
Rs [m³/m³]	*11.1*	*783.1*	*222.0*
Ty [°C]	*42.6*	*162.8*	*116.0*
Pb [kg/cm²]	*31.6*	*408.3*	*234*

Tabla 5.5: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de regresión de la P_b y normalizar las variables de entrada durante la etapa de procesamiento de datos.

A partir de los datos de la **tabla 5.5**, se normalizaron todos los datos de entrada a partir del método mín-máx. Es decir, para un pozo cuya temperatura es de 102 C, este dato se normalizó, a partir de las características identificadas en la tabla anterior, con la expresión

$$Ty' = \frac{102 - 42{,}6}{162{,}8 - 42{,}6} = 0{,}4941, \tag{5.6}$$

de donde, el valor $Ty' = 0.4941$ corresponde al valor normalizado de $Ty = 102$C.

5.3.2. Regresión lineal

Mediante este algoritmo es posible estimar el valor normalizado del punto de burbuja a partir de la temperatura del yacimiento y la relación gas-aceite disuelto a partir de la expresión obtenida con el conjunto de datos de entrenamiento.

$$P_b' = 0{,}1003T_y + 0{,}9478R_s + 0{,}2105 \tag{5.7}$$

Una vez obtenido este valor de P_b es necesario desnormalizar para, posteriormente, poder utilizarlo para graficar las demás propiedades del aceite.

$$P_b = P_b'(max - min) + min \tag{5.8}$$

Para evaluar el desempeño de los algoritmos se utilizaron dos diferentes métricas las cuales son el error absoluto promedio E_a y el coeficiente de determinación R^2, los cuales están definidos de forma precisa en el **Anexo A** al final de este trabajo

Las métricas de error obtenidas para este algoritmo en el conjunto de datos de prueba son mostradas en la **tabla 5.6**.

Indicador	*Valor [%]*
E_a	*17.39*
R^2	*83.39*

Tabla 5.6: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba.

5.3.3. Regresión con máquina de vectores de soporte SVR

El modelo de *Support Vector Regression* o regresión con máquina de vectores de soporte (SVR) fue entrenado con la función Kernel radial y, haciendo uso del mismo conjunto de datos que el algoritmo anterior. A partir de esto, se obtuvieron los siguientes parámetros con el conjunto de prueba con este modelo

Indicador	*Valor [%]*
E_a	*14.19*
R^2	*93.5*

Tabla 5.7: Métricas de error obtenidas para el algoritmo SVR en el conjunto de datos de prueba.

5.3.4. Redes Neuronales Artificiales

El modelo de redes neuronales artificiales fue entrenado con el mismo conjunto de datos que los modelos anteriores y a partir del algoritmo de retropropagación. En este caso, se propuso la siguiente estructura de red debido que dio los mejores resultados con base en prueba y error:

- **Capa de Entrada:** tres neuronas correspondientes a las propiedades de entrada para calcular Pb.
- **Primera Capa Oculta:** tres neuronas
- **Segunda Capa Oculta:** dos neuronas
- **Capa de Salida:** una capa de salida con una neurona equivalente al valor de Pb calculado.
- **Función de activación:** sigmoidal

En la **figura 5.3** se muestra la estructura de la red neuronal artificial propuesta:

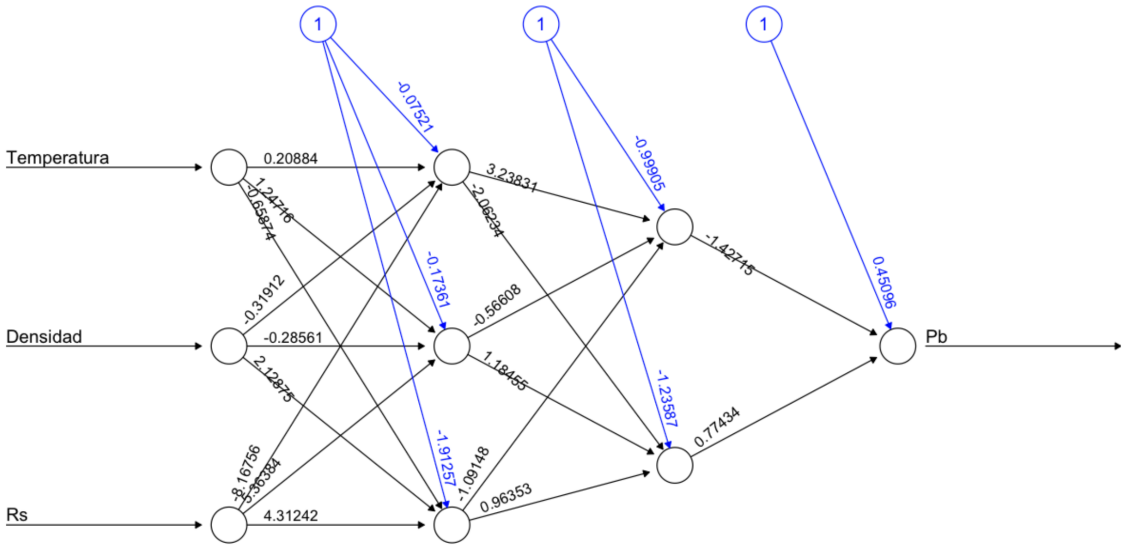

Figura 5.3: Estructura del modelo RNA para la estimación de la P_b

A continuación se muestran las métricas obtenidas con el modelo RNA creado para estimar la P_b

Indicador	*Valor [%]*
E_a	*9.53*
R^2	*92.63*

Tabla 5.8: Métricas de error obtenidas para el algoritmo RNA en el conjunto de datos de prueba.

5.3.5. Comparación de los 3 modelos

La **tabla 5.9** muestra una comparación de los tres modelos propuestos para la estimación de la presión de burbuja a partir de los datos de entrada propuestos.

Modelo generado	*Ea [%]*
Regresión lineal	*17.39*
SVR	*14.19*
RNA	*9.53*

Tabla 5.9: Métricas de error obtenidas con los modelos propuestos evaluados en conjunto de datos de prueba.

Como se mencionó anteriormente, el conjunto de datos es dividido en tres partes: uno de entrenamiento otra de prueba y uno más de de validación, este último es construido con el fin de observar gráficamente los resultados de los modelos, se hizo uso del conjunto de datos de validación, mostrado en la **tabla 5.10**.

Los valores de Pb calculados con los tres algoritmos mencionados anteriormente, así como el valor de P_b registrado en los reportes PVT, pueden apreciarse en la **tabla 5.11**.

La **figura 5.4** permite ver de forma gráfica los resultados conseguidos por los modelos de aprendizaje automático utilizados en este trabajo. Mientras más cercanos a la línea de P_b *Real* se encuentren los resultados obtenidos mediante las técnicas de aprendizaje automático, puede decirse que, mejor rendimiento tiene el modelo.
En este caso, puede observarse que, para estimar la presión de burbuja, el modelo de aprendizaje automático que menos error consigue con el conjunto de datos de prueba son las redes neuronales

Pozo	Ty [°C]	ρro [1]	Rs [m³/m³]	μo@20°C [cP]	Tipo de aceite	Pb [kg/cm²]
A	127	0.8995	96.397	18.57	Negro	184
B	102	0.8567	175.7	9.34097322	Negro	248
C	47	0.8596	80.6	10.9581313	Negro	126.6
D	126	0.8514	401.1	7.04652138	Volatil	320.7
E	147	0.8338	233.466	5.54659092	Volatil	275
F	124	0.8497	373.7	4.84738061	Volatil	318.5

Tabla 5.10: Descripción del conjunto de datos de validación utilizado

Pozo	Pb real [kg/cm²]	Pb Regresión lineal [kg/cm²]	Pb SVR [kg/cm²]	Pb RNA [kg/cm²]
A	184	176.874	204.861	175.364
B	248	205.698	252.004	233.785
C	126.6	144.427	98.841	124.907
D	320.7	317.494	320.329	330.582
E	275	246.558	244.205	256.02
F	318.5	304.192	318.274	323.542

Tabla 5.11: Estimaciones de la P_b con los modelos de aprendizaje automático propuestos.

artificiales, con un valor de E_a del 9.53 %, por lo que, para obtener esta propiedad, puede considerarse como el modelo más adecuado.

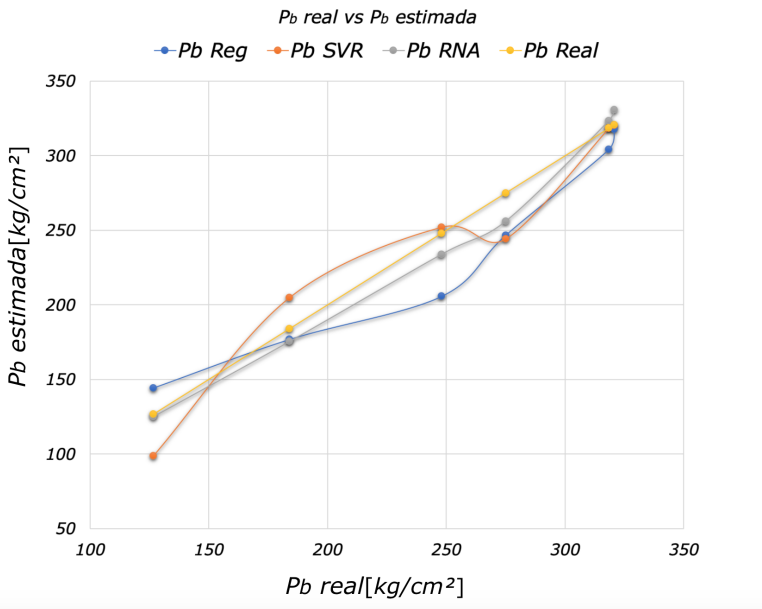

Figura 5.4: Gráfica de las estimaciones de la P_b obtenida con los modelos de aprendizaje automático propuestos vs el valor real medido de la P_b

5.4. Estimación del B_o en su región saturada

5.4.1. Preparación de entradas

A partir del mapa de correlación mostrado en la **figura 5.1** , se seleccionaron como entradas para los algoritmos de aprendizaje automático las propiedades con valores de correlación más alto, que coinciden con las usadas para los modelos de estimación de P_b. En el caso de la profundidad del pozo también fue descartada debido a su alta dependencia con la profundidad del pozo.

El conjunto de datos utilizado para el entrenamiento de los modelos de regresión para calcular el Bo en su región saturada se compone de muestras de 101 pozos con las características mostradas en la **tabla 5.12** y en la **tabla 5.13** . A partir de esta información, se normalizaron los datos de todo el conjunto de datos utilizado a partir del método mín-máx.

Tipo	*Cantidad [1]*
Aceite negros	*51*
Aceite volatil	*50*
Total	*101*

Tabla 5.12: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación del B_o en su región saturada.

Rango de datos	*Min*	*Max*	*Promedio*
ρro [1]	0.798	0.933	0.855
Rs [m³/m³]	11.1	783.1	221.8
Ty [°C]	43.6	162.8	116.5
Bob [m³/m³]	1.10	3.90	1.82

Tabla 5.13: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de regresión del B_o en su región saturada y normalizar las variables de entrada durante la etapa de procesamiento de datos.

5.4.2. Regresión lineal

Con este algoritmo es posible estimar los valores normalizados de B_o en 5 puntos arbitrarios de presión, haciendo uso de las expresiones obtenidas a partir de este algoritmo, el cual utiliza los valores de las entradas mencionadas anteriormente.

Punto 1

$$B_{o@\frac{1}{5}P_b} = 0{,}16209T_y - 0{,}09263\rho_o + 0{,}73161R_s + 0{,}10266 \tag{5.9}$$

La **tabla 5.14** muestra las métricas de error obtenidas con el modelo para la estimación de B_o en el primer punto de presión propuesto.

Rango de datos	*Min*	*Max*	*Promedio*
ρro [1]	0.798	0.933	0.855
Rs [m³/m³]	11.1	783.1	221.8
Ty [°C]	43.6	162.8	116.5
Bob [m³/m³]	1.10	3.90	1.82

Tabla 5.14: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba para el primer punto de presión.

Punto 2

$$B_{o@\frac{2}{5}P_b} = 0{,}1866T_y - 0{,}1486\rho_o + 0{,}6073R_s + 0{,}1435 \tag{5.10}$$

La **tabla 5.15** muestra las métricas de error obtenidas con el modelo para la estimación de B_o en el segundo punto de presión propuesto.

Indicador	*Valor [%]*
E_a	*3.63*
R^2	*91.07*

Tabla 5.15: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba para el segundo punto de presión.

Punto 3

$$B_{o@\frac{3}{5}P_b} = 0{,}2115T_y - 0{,}1591\rho_o + 0{,}6571R_s + 0{,}1398 \tag{5.11}$$

La **tabla 5.16** muestra las métricas de error obtenidas con el modelo para la estimación de B_o en el tercer punto de presión propuesto.

Indicador	*Valor [%]*
E_a	*3.79*
R^2	*94.53*

Tabla 5.16: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba para el tercer punto de presión.

Punto 4

$$B_{o@\frac{4}{5}P_b} = 0{,}18073T_y - 0{,}10739\rho_o + 0{,}86433R_s + 0{,}05474 \tag{5.12}$$

La **tabla 5.17** muestra las métricas de error obtenidas con el modelo para la estimación de B_o en el cuarto punto de presión propuesto.

Indicador	*Valor [%]*
E_a	*2.68*
R^2	*97.77*

Tabla 5.17: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba para el cuarto punto de presión.

Punto 5

$$B_{o@P_b} = 0{,}10124T_y - 0{,}03649\rho_o + 0{,}92621R_s - 0{,}03973 \tag{5.13}$$

La **tabla 5.18** muestra las métricas de error obtenidas con el modelo para la estimación de B_o en el último punto de presión propuesto.

5.4.3. SVR

El modelo de SVR fue con la función Kernel radial y, al igual que el algoritmo anterior, fue entrenado con el mismo conjunto de datos de entrenamiento que el algoritmo anterior.

Indicador	*Valor [%]*
E_a	*3.6*
R^2	*99.0*

Tabla 5.18: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba para el último punto de presión.

Se obtuvieron para cada punto de presión propuesto, los siguientes parámetros de error en el conjunto de datos de prueba:

Indicador	*Valor [%]*
E_a	*2.84*
R^2	*89.87*

Tabla 5.19: Métricas de error obtenidas para el algoritmo de SVR en el conjunto de datos de prueba para el primer punto de presión propuesto.

Indicador	*Valor [%]*
E_a	*3.21*
R^2	*93.47*

Tabla 5.20: Métricas de error obtenidas para el algoritmo de SVR en el conjunto de datos de prueba para el segundo de presión propuesto.

Indicador	*Valor [%]*
E_a	*3.25*
R^2	*96.08*

Tabla 5.21: Métricas de error obtenidas para el algoritmo de SVR en el conjunto de datos de prueba para el tercer punto de presión propuesto.

Indicador	*Valor [%]*
E_a	*3.21*
R^2	*97.75*

Tabla 5.22: Métricas de error obtenidas para el algoritmo de SVR en el conjunto de datos de prueba para el cuarto punto de presión propuesto.

Indicador	*Valor [%]*
E_a	*3.36*
R^2	*98.07*

Tabla 5.23: Métricas de error obtenidas para el algoritmo de SVR en el conjunto de datos de prueba para el último punto de presión propuesto.

5.4.4. Redes Neuronales Artificiales

El modelo de redes neuronales artificiales fue entrenado con el mismo conjunto de datos que los dos modelos anteriores. En este caso, la estructura de red utilizada para estimar el B_o en su región saturada en los 5 puntos de presión propuestos dio los mejores resultados con la siguiente arquitectura:

- **Capa de Entrada:** tres neuronas correspondientes a las propiedades de entrada para calcular B_o.
- **Primera Capa Oculta:** cuatro neuronas
- **Segunda Capa Oculta:** tres neuronas
- **Capa de Salida:** una capa de salida con cinco neuronas equivalente al valor de B_o calculado.
- **Función de activación:** sigmoidal

En la **figura 5.5** se muestra la estructura de la red neuronal artificial propuesta:

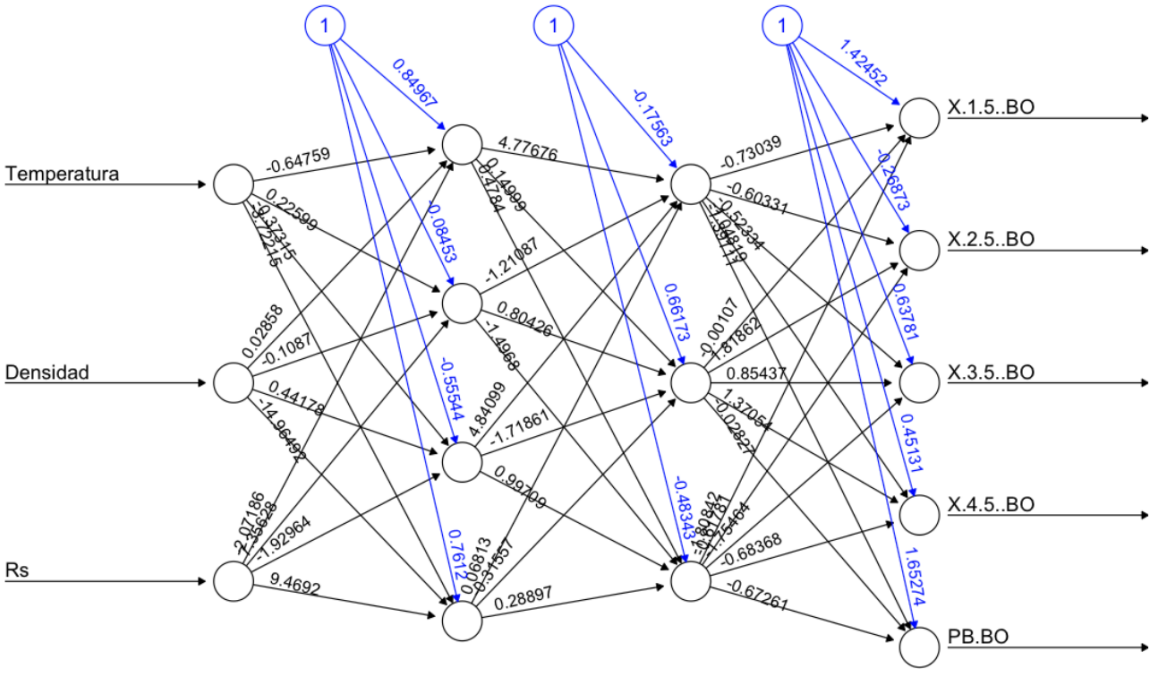

Figura 5.5: Estructura del modelo RNA para la estimación de los puntos de B_o en su región saturada

A continuación, en la **tabla 5.24**, se muestran las métricas obtenidas con el modelo redes neuronales creado para estimar $B_{o\ sat}$ en los cinco puntos de presión propuestos.

Punto	Indicador	Valor [%]
1/5 Pb	E_a	2.6
	R^2	88.82
2/5 Pb	E_a	2.77
	R^2	93.91
3/5 Pb	E_a	2.75
	R^2	96.6
4/5 Pb	E_a	2.21
	R^2	98.42
Pb	E_a	2.16
	R^2	99.05

Tabla 5.24: Métricas de error obtenidas para el algoritmo de RNA en el conjunto de datos de prueba para los cinco puntos de presión propuesto.

5.4.5. Comparación de los 3 métodos

La **tabla 5.25** muestra una comparación de los tres modelos propuestos para la estimación del factor de volumen del aceite en la región saturada a partir de los datos de entrada propuestos. Se utiliza el promedio de los E_a en los 5 puntos propuestos para comparar los modelos.

Modelo generado	Ea de Bo saturado observado en los 5 puntos de presión en el conjunto de prueba [%]
Regresión lineal	3.31
SVR	3.17
RNA	2.5

Tabla 5.25: Métricas de error promedio obtenidas para los algoritmos propuestos en el conjunto de datos de prueba para los cinco puntos de presión propuestos.

Se puede observar que, para estimar esta propiedad, el modelo de aprendizaje automático que menos error obtuvo con el conjunto de datos de prueba fue el de RNA con un E_a igual a 2.5 %, por lo que, puede concluirse que para obtener el factor de volumen del aceite en su región saturada, éste es el modelo más competitivo.

5.5. Estimación del B_o en su región Bajo saturada

5.5.1. Preparación de entradas

A partir del mapa de correlación mostrado en la **figura 5.1**, se seleccionaron como entradas para los algoritmos de aprendizaje automático las propiedades que mostraron los valores de correlación más alto, que coinciden nuevamente con las usadas para los modelos de estimación de Pb y Bo en la región saturada.

El conjunto de datos utilizado para el entrenamiento de los modelos de regresión para calcular $B_{o\ bajosat}$ se compone de muestras de 98 pozos con las características mostradas en la **tabla 5.26**. A partir de esta información, se normalizaron los datos de todo el conjunto de datos utilizado a partir del método mín-máx.

Tipo	*Cantidad [1]*
Aceite negros	*50*
Aceite volatil	*48*
Total	*98*

Tabla 5.26: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de B_o en su región bajo saturada.

5.5.2. Regresión lineal

Mediante este algoritmo es posible estimar el valor normalizado de $(B_o@Pb - B_o@Pb + 200[kg/cm])m/m]$a partir de la temperatura del yacimiento y la relación gas-aceite disuelto con la expresión

$$(B_o@Pb - B_o@Pb + 200[kg/cm]) = 0{,}05105T_y + 0{,}80606R_s - 0{,}07659 \tag{5.14}$$

En **tabla 5.28** se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $B_{o\ bajosat}$.c

Rango de datos	Min	Max	Promedio
ρro [1]	0.798	0.933	0.855
Rs [m^3/m^3]	11.1	783.1	217.7
Ty [°C]	42.6	162.8	114.9
$\triangle$(Bo@Pb - Bo@Pb+200[kg/cm^2]) [m^3/m^3]	0.0142	0.5808	0.1102

Tabla 5.27: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de regresión del B_o en su región bajo saturada y normalizar las variables de entrada durante la etapa de procesamiento de datos.

Indicador	Valor [%]
E_a	29.26
R^2	95.8

Tabla 5.28: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba.

5.5.3. SVR

El modelo de SVR fue entrenado con función Kernel radial y, haciendo uso del mismo conjunto de datos. A partir de esto, estas fueron las métricas de error obtenidas en el conjunto de prueba con este modelo.

Indicador	Valor [%]
E_a	26.36
R^2	95.22

Tabla 5.29: Métricas de error obtenidas para el algoritmo SVR en el conjunto de datos de prueba.

5.5.4. Redes Neuronales Artificiales

Este modelo fue entrenado con el mismo conjunto de datos que los modelos anteriores. La estructura de red propuesta para la estimación obtuvo los mejores resultados con la siguiente arquitectura $(B_o@Pb - B_o@Pb + 200[kg/cm])m/m]$

- **Capa de Entrada:** tres neuronas correspondientes a las propiedades de entrada para calcular Pb.
- **Primera Capa Oculta:** seis neuronas
- **Segunda Capa de Oculta:** seis neuronas
- **Capa de Salida:** una capa de salida con una neurona equivalente al valor de $(B_o@Pb - B_o@Pb + 200[kg/cm])m/m]$ calculado.
- **Función de activación:** sigmoidal

En la **figura 5.6** se muestra la estructura de la red neuronal artificial propuesta:

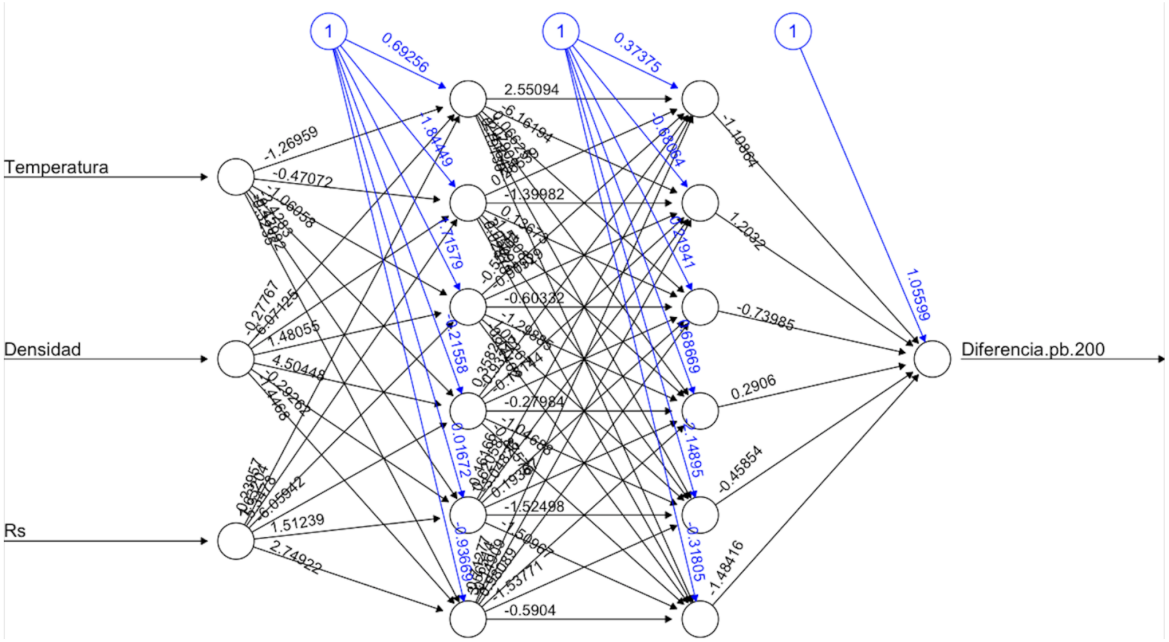

Figura 5.6: Estructura del modelo RNA para la estimación de $(B_o@Pb - B_o@Pb + 200[kg/cm])$

En la **tabla 5.30** se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $(B_o@Pb - B_o@Pb + 200[kg/cm])m/m]$.

Indicador	Valor [%]
E_a	11.99
R^2	94.23

Tabla 5.30: Métricas de error obtenidas para el algoritmo RNA en el conjunto de datos de prueba.

5.6. Comparación de los tres métodos

Hay que recordar que, una vez obtenido el valor de $(B_o@Pb - B_o@Pb + 200[kg/cm])m/m]$, es necesario normalizar dicho resultado para, posteriormente, llevarlo a condiciones de presión de yacimiento (B_{oy}). En la siguiente tabla, se muestra una comparación del porcentaje de error absoluto medio obtenido con los valores de B_o normalizado a condiciones de presión de yacimiento.

Modelo generado	Ea del Boy para el conjunto de validación [%]
Regresión lineal	3.13
SVR	2.22
RNA	1.91

Tabla 5.31: Métricas de error obtenidas para los algoritmos en el conjunto de datos de prueba.

En la **tabla 5.31** , se puede observar que para estimar el B_{oy}, el modelo de aprendizaje automático que menos error obtiene con el conjunto de pruebas es las red neuronal artificial con un porcentaje de error absoluto medio del 1.91 %, por lo que, para obtener esta propiedad, este es el modelo más competitivo.

Los valores de B_{oy} calculados con los tres algoritmos mencionados anteriormente en el conjunto de validación, así como el valor de B_{oy} registrado en los reportes PVT, pueden apreciarse en la **tabla 5.32**

Pozo	Boy Real [m³/m³]	Boy Regresión lineal [m³/m³]	Boy SVR [m³/m³]	Boy RNA [m³/m³]
A	1.340	1.379	1.321	1.365
B	1.496	1.584	1.544	1.554
C	1.243	1.184	1.277	1.238
D	2.309	2.299	2.254	2.303
E	1.848	1.870	1.869	1.883
F	2.152	2.230	2.205	2.221

Tabla 5.32: Resultados obtenidos para todos los algoritmos en el conjunto de datos de validación.

5.7. Generación de las curvas completas de B_o

A continuación, se muestran las gráficas generadas a partir de la unión de los resultados obtenidos con los modelos de aprendizaje automático para estimar las curvas de las regiones saturada y bajo saturada de B_o.

Las **figuras 5.7, 5.8, 5.9, 5.10, 5.11 y 5.12** nos permiten ver de manera gráfica las curvas generadas por la unión de resultados obtenidos con los modelos de B_o en la región saturada y bajo saturada elaborados en este trabajo. De esta manera, podemos confirmar visualmente que para estimar el B_o, el modelo de RNA muestra mejores resultados en este caso.

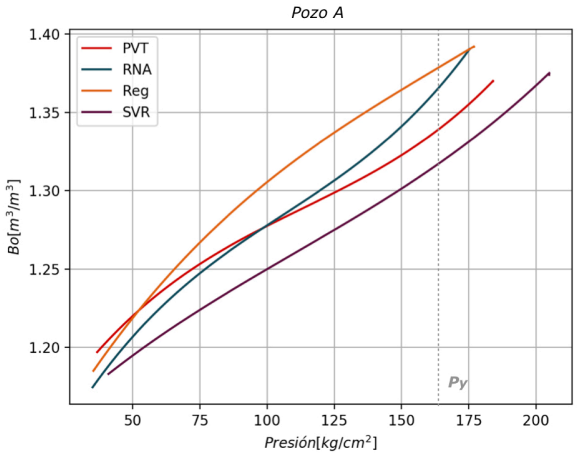

Figura 5.7: Curvas generadas para el Pozo A con los modelos de estimación de B_o propuesto

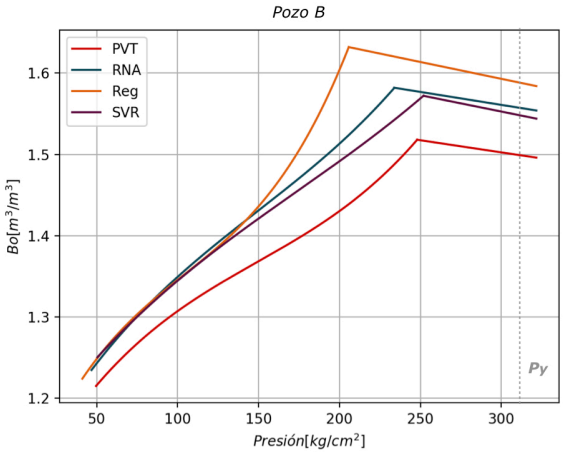

Figura 5.8: Curvas generadas para el Pozo B con los modelos de estimación de B_o propuesto

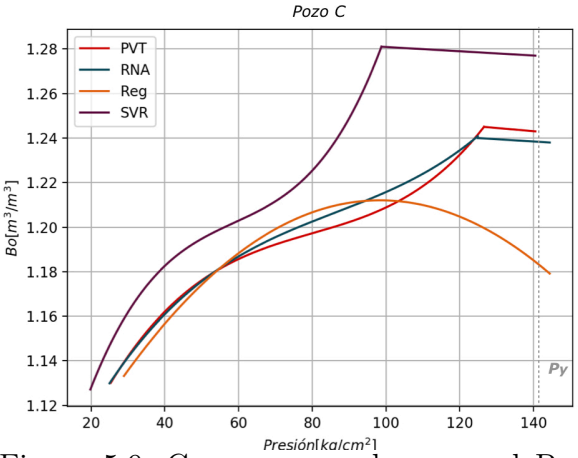

Figura 5.9: Curvas generadas para el Pozo C con los modelos de estimación de B_o propuesto

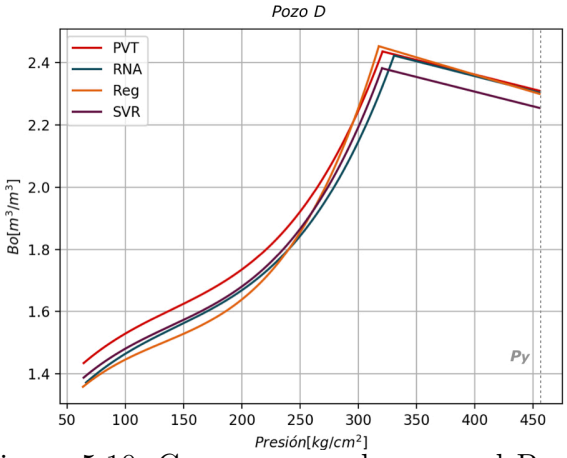

Figura 5.10: Curvas generadas para el Pozo D con los modelos de estimación de B_o propuesto

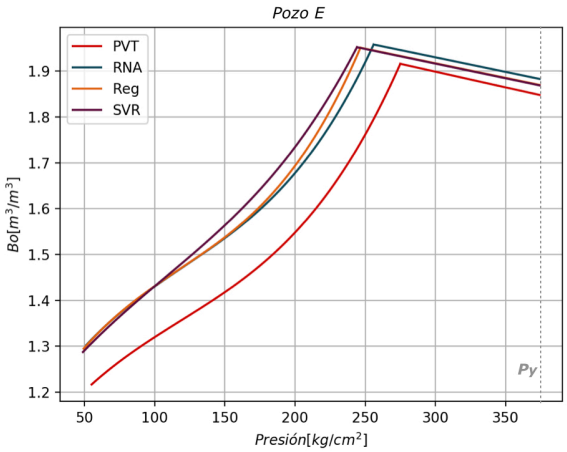

Figura 5.11: Curvas generadas para el Pozo E con los modelos de estimación de B_o propuesto

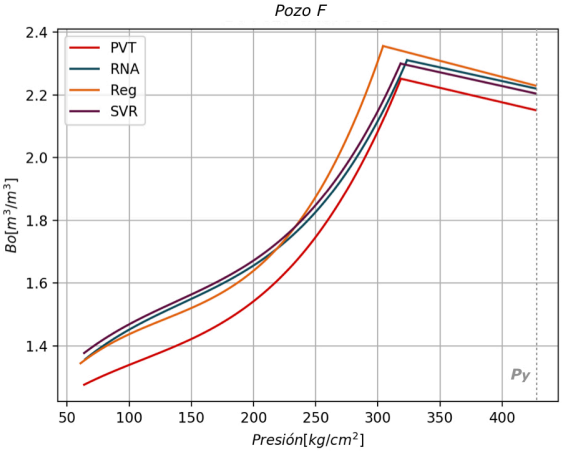

Figura 5.12: Curvas generadas para el Pozo F con los modelos de estimación de B_o propuesto

5.8. Estimación de ρ_{ro} en su región saturada

5.8.1. Preparación de entradas

Partiendo del mapa de correlación generado **figura 5.1**, se seleccionaron como entradas la temperatura del pozo (T_y), la relación gas-aceite disuelto (R_s), la densidad relativa del aceite (ρ_{ro}) a condiciones de superficie y la viscosidad del aceite a 20 °C ($\mu_{o\ @20^oC}$) debido a su valor alto de correlación.

El conjunto de datos utilizado para el entrenamiento de los modelos de regresión para calcular $\rho_{o\ sat}$ está conformado por muestras de 103 pozos con las características mostradas en las **tablas 5.32 y 5.33** A partir de esta información, se normalizaron los datos de todo el conjunto de datos utilizado con el método mín-máx.

Tipo	Cantidad [1]
Aceite negros	52
Aceite volatil	51
Total	103

Tabla 5.32: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de ρ_{ro} en su región saturada.

Rango de datos	Min	Max	Promedio
ρro [1]	0.798	0.933	0.855
Rs [m³/m³]	11.1	783.1	220.6
Ty [°C]	43.6	162.8	116.6
μo@20°C [cP]	2.84	302.74	18.91
ρrob [1]	0.396	0.874	0.629

Tabla 5.33: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de regresión del ρ_{ro} en su región saturada y normalizar las variables de entrada durante la etapa de procesamiento de datos.

5.8.2. Regresión lineal

Con este algoritmo se estimaron los valores normalizados de la ρ_{ro} en 5 puntos arbitrarios de presión, usando las 4 entradas mencionadas anteriormente y generando las siguiente expresiones

Punto 1

$$\rho_{ro@\frac{1}{5}P_b} = -0{,}20232T_y + 0{,}35113\rho_{ro} - 0{,}16774R_s + 0{,}04018\mu_{o@20^oC} + 0{,}63691 \tag{5.15}$$

Punto 2

$$\rho_{ro@\frac{2}{5}P_b} = -0{,}242T_y + 0{,}3846\rho_{ro} - 0{,}1946R_s + 0{,}0388\mu_{o@20^oC} + 0{,}6086 \tag{5.16}$$

Punto 3

$$\rho_{ro@\frac{3}{5}P_b} = -0{,}26338T_y + 0{,}38907\rho_{ro} - 0{,}24570R_s + 0{,}04755\mu_{o@20^oC} + 0{,}60618 \tag{5.17}$$

Punto 4

$$\rho_{ro@\frac{4}{5}P_b} = -0{,}27755T_y + 0{,}39801\rho_{ro} - 0{,}35589R_s + 0{,}04574\mu_{o@20^oC} + 0{,}61174 \tag{5.18}$$

Punto 5

$$\rho_{ro@P_b} = -0{,}28408T_y + 0{,}42414\rho_{ro} - 0{,}54034R_s + 0{,}02232\mu_{o@20^oC} + 0{,}62248 \tag{5.19}$$

La **tabla 5.34** muestra las métricas de error obtenidas con el modelo para la estimación de la ρ_{ro} en los puntos de presión propuestos.

Punto	*Indicador*	*Valor [%]*
1/5 Pb	E_a	*2.04*
	R^2	*92.42*
2/5 Pb	E_a	*2.82*
	R^2	*91.37*
3/5 Pb	E_a	*2.98*
	R^2	*93.19*
4/5 Pb	E_a	*3.15*
	R^2	*95.03*
Pb	E_a	*4.32*
	R^2	*95.49*

Tabla 5.34: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba para los puntos de presión.

5.8.3. SVR

El modelo de SVM fue entrenado con el mismo conjunto de datos de entrenamiento. Se utilizó la función Kernel radial, ya que dio mejores resultados. Se obtuvieron para cada punto de presión propuesto, los siguientes valores de error en el conjunto de datos de prueba:

Se obtuvieron para los puntos de presión propuestos, los siguientes parámetros de error en el conjunto de datos de prueba:

Punto	Indicador	Valor [%]
1/5 Pb	E_a	2.44
	R^2	89.95
2/5 Pb	E_a	2.97
	R^2	90.37
3/5 Pb	E_a	2.81
	R^2	93.06
4/5 Pb	E_a	2.66
	R^2	95.69
Pb	E_a	3.77
	R^2	96.33

Tabla 5.35: Métricas de error obtenidas para el algoritmo de SVR en el conjunto de datos de prueba para los puntos de presión propuestos.

5.8.4. Redes Neuronales Artificiales

El modelo de redes neuronales artificiales fue entrenado con el mismo conjunto de datos que los dos modelos anteriores. En este caso, la estructura de red utilizada para estimar ρ_{ro} en su región saturada en los 5 puntos de presión propuestos dio los mejores resultados con la siguiente arquitectura:

- **Capa de Entrada:** cuatro neuronas correspondientes a las propiedades de entrada para calcular ρ_{ro}.
- **Primera Capa Oculta:** cuatro neuronas
- **Segunda Capa Oculta:** tres neuronas
- **Capa de Salida:** una capa de salida con cinco neuronas equivalente al valor de ρ_{ro} calculado en cada punto de presión.
- **Función de activación:** sigmoidal

En la **figura 5.13** se muestra la estructura de la red neuronal artificial propuesta:

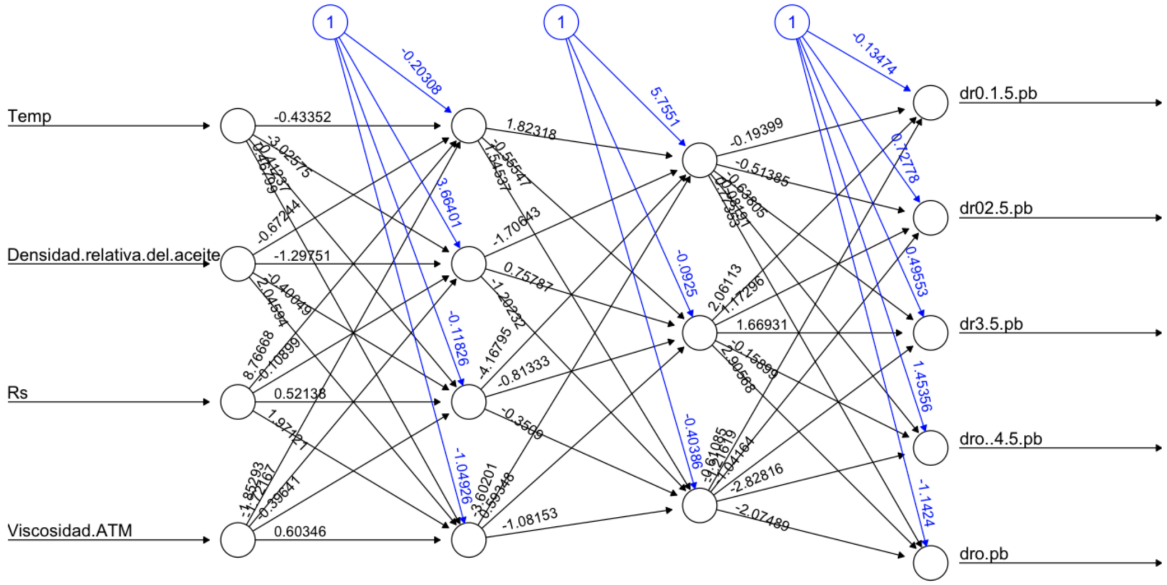

Figura 5.13: Estructura del modelo RNA para la estimación de los puntos de ρ_{ro} en su región saturada

En la **tabla 5.36** se muestran las métricas de error del modelo redes neuronales creado para estimar el ρ_{ro} en los 5 puntos de presión propuestos de la región saturada.

Punto	Indicador	Valor [%]
1/5 Pb	E_a	2.08
	R^2	91.66
2/5 Pb	E_a	2.68
	R^2	91.19
3/5 Pb	E_a	2.82
	R^2	93.49
4/5 Pb	E_a	2.95
	R^2	95.33
Pb	E_a	4.42
	R^2	95.42

Tabla 5.36: Métricas de error obtenidas para el algoritmo de RNA en el conjunto de datos de prueba para los cinco puntos de presión propuestos.

5.8.5. Comparación de los 3 métodos

La **tabla 5.37** muestra una comparación de los tres modelos propuestos para la estimación de la densidad del aceite en la región saturada a partir de los datos de entrada propuestos.

Modelo generado	Ea de pro observado en los 5 puntos de presión en el conjunto de prueba [%]
Regresión lineal	3.06
SVR	2.93
RNA	2.99

Tabla 5.37: Métricas de error promedio obtenidas para los algoritmos propuestos en el conjunto de datos de prueba para los cinco puntos de presión propuestos.

Se puede observar que para estimar esta propiedad, el modelo de aprendizaje automático que menos error consigue en el conjunto de pruebas es el de SVM con un E_a de 2.93 %, por lo que es el modelo más competitivo

5.9. Estimación de la ρ_{ro} en su región bajo saturada

5.9.1. Preparación de entradas

A partir del mapa de correlación mostrado en la **figura 5.1**, se seleccionaron como entradas de alimentación a los algoritmos de aprendizaje automático la temperatura del pozo (Ty), la relación gas-aceite disuelto (Rs) y la densidad relativa del aceite.

El conjunto de datos utilizado para el entrenamiento de los modelos de regresión para calcular o en su región bajo saturada se compone de muestras de 88 pozos con las características mostradas en las **tablas 5.38 y 5.39**. A partir de esta información, se normalizaron los datos de todo el conjunto de datos utilizado a partir del método mín-máx.

Tipo	*Cantidad [1]*
Aceite negros	*42*
Aceite volatil	*46*
Total	*88*

Tabla 5.38: Número de muestras del conjunto de datos utilizado para entrenar y probar los modelos de ρ_{ro} en su región bajo saturada

Rango de datos	*Min*	*Max*	*Promedio*
ρro [1]	*0.798*	*0.9332*	*0.853*
Rs [m³/m³]	*11.1*	*783.1*	*227*
Ty [°C]	*42.6*	*162.8*	*117*
△(ρro@Pb+200[kg/cm²] - ρro@Pb) [1]	*0.0030*	*0.0767*	*0.0330*

Tabla 5.39: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar y validar los modelos de regresión de la ρ_{ro} en su región bajo saturada y normalizar las variables de entrada durante la etapa de procesamiento de datos.

5.9.2. Regresión lineal

Mediante este algoritmo es posible estimar el valor normalizado de $(\rho_{ro}@Pb - \rho_{ro}@Pb + 200[kg/cm])1]$ a partir de la temperatura del yacimiento y la relación gas-aceite disuelto con el uso de la expresión:

$$(\rho_{ro}@Pb - \rho_{ro}@Pb + 200[kg/cm]) = 0{,}08028T_y + 0{,}63891R_s + 0{,}16392 \tag{5.20}$$

Las métricas de error obtenidas para este algoritmo en el conjunto de datos de prueba se observan en la **tabla 5.40** .

Indicador	*Valor [%]*
E_a	*12.38*
R^2	*95.09*

Tabla 5.40: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba.

5.9.3. SVR

El modelo de SVR fue entrenado con función Kernel radial y, haciendo uso del mismo conjunto de datos. Estas fueron las métricas de error obtenidas en el conjunto de prueba con este modelo.

Indicador	*Valor [%]*
E_a	*12.30*
R^2	*96.22*

Tabla 5.41: Métricas de error obtenidas para el algoritmo SVR en el conjunto de datos de prueba.

5.9.4. Redes Neuronales Artificiales

Estos son los parámetros de aprendizaje utilizados en el entrenamiento del modelo de redes neuronales para la estimación de $(\rho_{ro}@Pb - \rho_{ro}@Pb + 200[kg/cm])$

- **Capa de Entrada:** tres neuronas correspondientes a las propiedades de entrada para calcular $(\rho_{ro}@Pb - \rho_{ro}@Pb + 200[kg/cm])$.
- **Primera Capa Oculta:** seis neuronas
- **Segunda Capa de Oculta:** seis neuronas
- **Capa de Salida:** una capa de salida con una neurona equivalente al valor de $(\rho_{ro}@Pb - \rho_{ro}@Pb + 200[kg/cm])$ calculado.
- **Función de activación:** sigmoidal

En la **figura 5.14** se muestra la estructura de la red neuronal artificial propuesta:

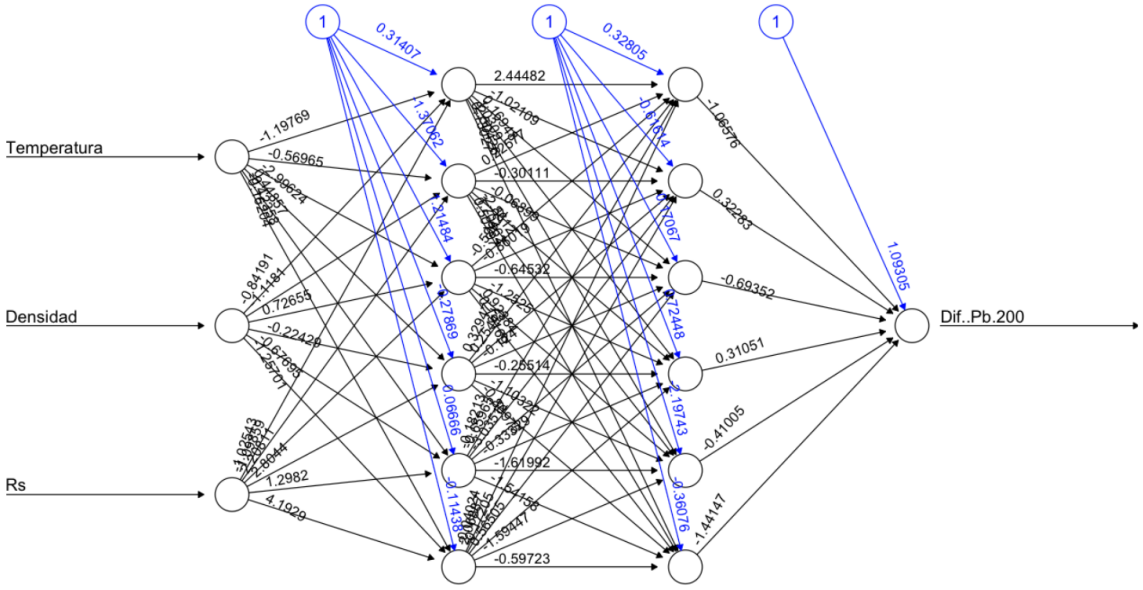

Figura 5.14: Estructura del modelo RNA para la estimación de $(\rho_{ro}@Pb - \rho_{ro}@Pb + 200[kg/cm])$

En la **figura 5.42** se muestran las métricas de error del modelo redes neuronales creado para estimar $(\rho_{ro}@Pb - \rho_{ro}@Pb + 200[kg/cm])$ con los cuales se obtuvieron los mejores resultados:

Indicador	Valor [%]
E_a	14.21
R^2	92.53

Tabla 5.42: Métricas de error obtenidas para el algoritmo RNA en el conjunto de datos de prueba.

5.10. Comparación de los tres métodos

El cálculo de $(\rho_{ro}@Pb - \rho_{ro}@Pb + 200[kg/cm])[m/m]$ tiende a suponer errores mayores que el error real, por lo que se debe normalizar dicho resultado llevándolo a condiciones de presión de yacimiento (ρ_{roy}).

La **tabla 5.43** muestra una comparación del Ea de los tres modelos propuestos para la estimación de la ρ_{ro} ya normalizada a condiciones de presión de yacimiento evaluado en el set de validación.

Modelo generado	Ea del Boy para el conjunto de validación [%]
Regresión lineal	2.32
SVR	2.47
RNA	2.13

Tabla 5.43: Métricas de error obtenidas para los algoritmos en el conjunto de datos de prueba.

En la **tabla 5.43** se puede observar que para estimar la ρ_{roy}, el modelo de aprendizaje automático que menos error consigue en el conjunto de pruebas es las redes neuronales artificiales con un error porcentual absoluto medio de 2.13 %, por lo que, para obtener esta propiedad, este es el modelo más competitivo.

Los valores de ρ_{roy} calculados con los tres algoritmos mencionados anteriormente en el conjunto de validación, así como el valor de ρ_{roy} registrado en los reportes PVT, pueden apreciarse en la **5.44**.

5.11. Generación de las curvas completas de ρ_{ro}

A continuación, se muestran las gráficas generadas a partir de la unión de los resultados obtenidos con los modelos de aprendizaje automático para estimar las curvas de las regiones saturada y bajo saturada de ρ_{ro}.

Con las **figuras 5.15, 5.16, 5.17, 5.18, 5.19 y 5.20** es posible observar de manera gráfica que,

Pozo	ρ_{roy} Real [1]	ρ_{roy} Regresión lineal [1]	ρ_{roy} SVR [1]	ρ_{roy} RNA [1]
A	0.726	0.730	0.735	0.744
B	0.708	0.677	0.672	0.671
C	0.754	0.762	0.761	0.747
D	0.560	0.579	0.570	0.573
E	0.558	0.578	0.580	0.562
F	0.579	0.584	0.569	0.573

Tabla 5.44: Resultados obtenidos para todos los algoritmos en el conjunto de datos de validación.

para estimar la ρ_{ro}, los modelos de aprendizaje automático que tienen una mejor aproximación son la red neuronal artificial y las máquinas de vectores de soporte de regresión.

Con las **figuras 5.15, 5.16, 5.17, 5.18, 5.19 y 5.20** es posible observar de manera gráfica que, para estimar la ρ_{ro}, los modelos de aprendizaje automático que tienen una mejor aproximación son la red neuronal artificial y las máquinas de vectores de soporte de regresión.

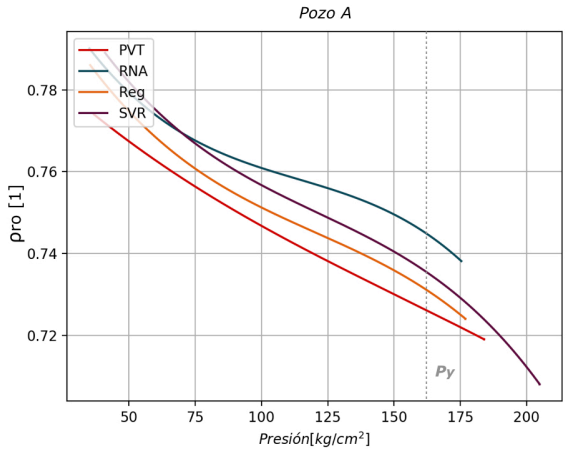

Figura 5.15: Curvas generadas para el Pozo A con los modelos de estimación de ρ_{ro} propuestos

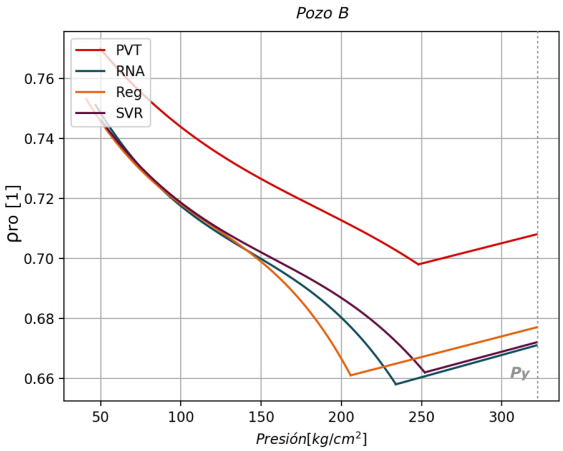

Figura 5.16: Curvas generadas para el Pozo B con los modelos de estimación de ρ_{ro} propuestos

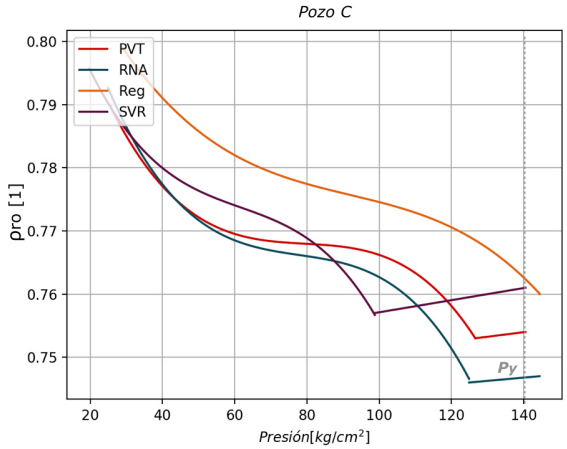

Figura 5.17: Curvas generadas para el Pozo C con los modelos de estimación de ρ_{ro} propuestos

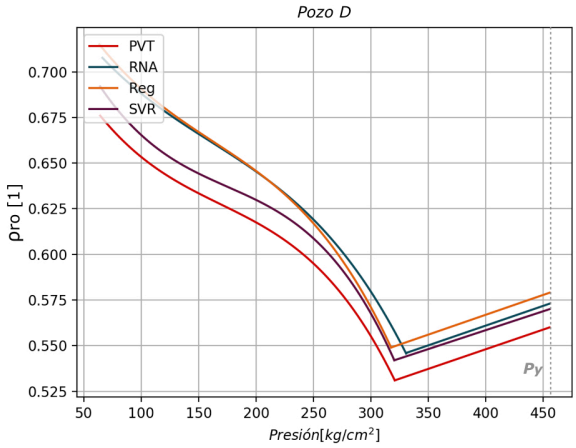

Figura 5.18: Curvas generadas para el Pozo D con los modelos de estimación de ρ_{ro} propuestoso

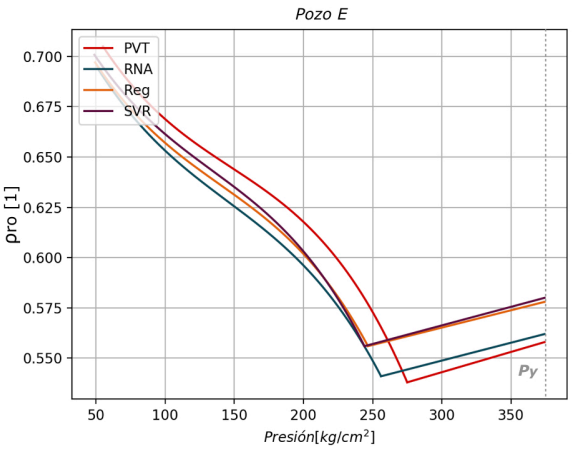

Figura 5.19: Curvas generadas para el Pozo E con los modelos de estimación de ρ_{ro} propuestos

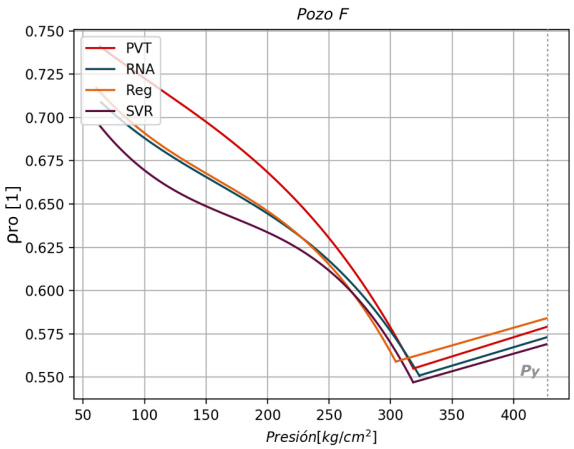

Figura 5.20: Curvas generadas para el Pozo F con los modelos de estimación de B_o propuesto

5.12. Estimación de R_s

5.12.1. Preparación de entradas

Se utilizó el mapa de correlación mostrado en la **figura 5.1** para seleccionar como entradas para los algoritmos de aprendizaje automático las propiedades con valores de correlación más alto. En este caso, la temperatura del pozo (T_y), la relación gas-aceite disuelto a la presión de saturación del aceite (Rsb) y la densidad relativa del aceite (ρ_{ro}).

El conjunto de datos utilizado para el entrenamiento de los modelos de regresión para calcular la curva de R_s se compone de muestras de 101 pozos con las características mostradas en las **tablas 5.45 y 5.46**

Tipo	Cantidad [1]
Aceite negros	51
Aceite volatil	50
Total	101

Tabla 5.45: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de R_s en su región saturada.

Rango de datos	Min	Max	Promedio
ρro [1]	0.798	0.9332	0.855
Rs [m³/m³]	11.1	783.1	221.8
Ty [°C]	43.6	162.8	116.5

Tabla 5.46: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de regresión del R_s en su región saturada y normalizar las variables de entrada durante la etapa de procesamiento de datos.

A partir de los datos de la **tabla 5.46**, se normalizaron todos los datos de entrada con el método mín-máx.

5.12.2. Regresión lineal

Con este algoritmo se estimaron los valores normalizados de R_s en 4 puntos de presión propuestos ($\frac{1}{5}P_b$, $\frac{2}{5}P_b$, $\frac{3}{5}P_b$ y $\frac{4}{5}P_b$), haciendo uso de las entradas mencionadas anteriormente, a partir de las expresiones mostradas a continuación.

Punto 1

$$_{s@\frac{1}{5}P_b} = -0{,}08053T_y - 0{,}09973\rho_{ro} + 1{,}06831R_s + 0{,}27982 \tag{5.21}$$

Punto 2

$$_{s@\frac{2}{5}P_b} = -0{,}01702T_y - 0{,}11366\rho_{ro} + 1{,}10395R_s + 0{,}30148 \tag{5.22}$$

Punto 3

$$_{s@\frac{3}{5}P_b} = 0{,}02671T_y - 0{,}07784\rho_{ro} + 1{,}07759R_s + 0{,}20631 \tag{5.23}$$

Punto 4

$$_{s@\frac{4}{5}P_b} = 0{,}02223T_y - 0{,}02847\rho_{ro} + 1{,}00876R_s + 0{,}07854 \tag{5.24}$$

La **tabla 5.47** muestra las métricas de error obtenidas con el modelo para la estimación de la R_s en los puntos de presión propuestos.

Punto	*Indicador*	*Valor [%]*
1/5 Pb	E_a	*11.21*
	R^2	*91.68*
2/5 Pb	E_a	*10.79*
	R^2	*88.68*
3/5 Pb	E_a	*12.21*
	R^2	*88.01*
4/5 Pb	E_a	*8.31*
	R^2	*95.81*

Tabla 5.47: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba para los puntos de presión.

5.12.3. SVR

El modelo de SVR para estimar la curva de R_s fue entrenado con función Kernel radial haciendo uso del mismo conjunto de datos que el algoritmo anterior. Las métricas de error obtenidas en el conjunto de prueba con este algoritmo para cada punto se muestran a continuación.

Punto	*Indicador*	*Valor [%]*
1/5 Pb	E_a	*9.34*
	R^2	*95.77*
2/5 Pb	E_a	*6.67*
	R^2	*96.66*
3/5 Pb	E_a	*6.48*
	R^2	*96.14*
4/5 Pb	E_a	*5.99*
	R^2	*98.26*

Tabla 5.48: Métricas de error obtenidas para el algoritmo de SVR en el conjunto de datos de prueba para los puntos de presión propuestos.

5.12.4. Redes Neuronales Artificiales

El modelo de redes neuronales artificiales fue entrenado con el mismo conjunto de datos que los dos modelos anteriores. En este caso, la estructura de red utilizada para estimar R_s en su región saturada en los 4 puntos de presión propuestos dio los mejores resultados con la siguiente arquitectura:

- **Capa de Entrada:** tres neuronas correspondientes a las propiedades de entrada para calcular R_s.
- **Primera Capa Oculta:** cuatro neuronas
- **Segunda Capa Oculta:** tres neuronas
- **Capa de Salida:** una capa de salida con cuatro neuronas equivalente al valor de R_s calculado en cada punto de presión.
- **Función de activación:** sigmoidal

En la **figura 5.21** se muestra la estructura de la red neuronal artificial propuesta:

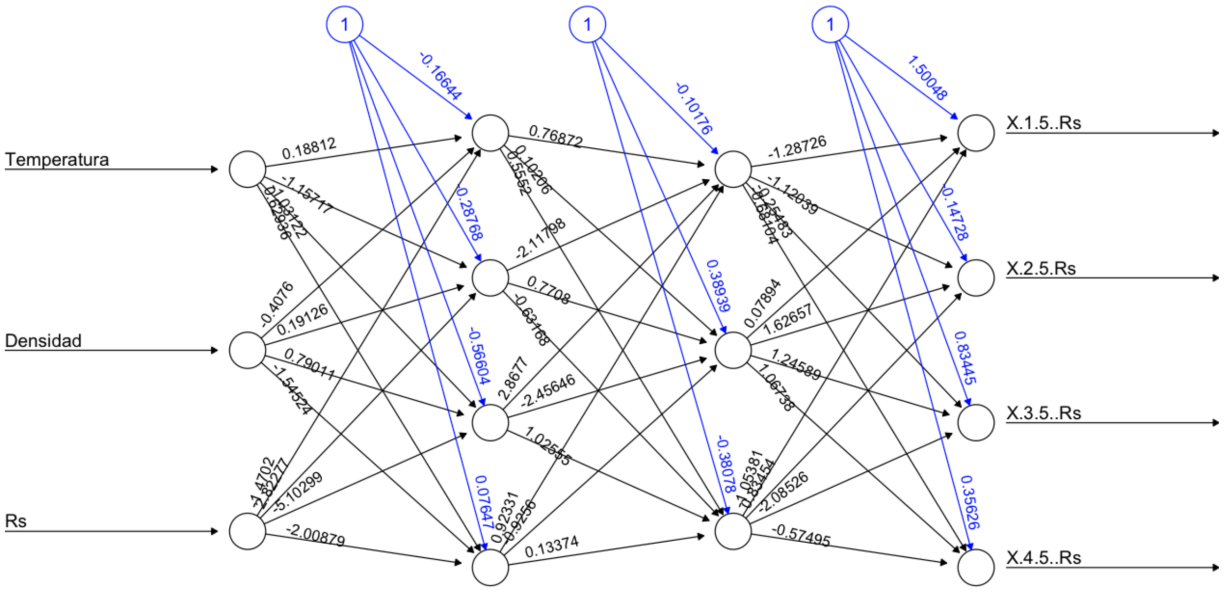

Figura 5.21: Estructura del modelo RNA para la estimación de los puntos de R_s en su región saturada

En la **tabla 5.49** se muestran las métricas de error del modelo redes neuronales creado para estimar la curva de R_s

Punto	*Indicador*	*Valor [%]*
1/5 Pb	E_a	*10.96*
	R^2	*93.11*
2/5 Pb	E_a	*7.22*
	R^2	*95.71*
3/5 Pb	E_a	*5.73*
	R^2	*95.34*
4/5 Pb	E_a	*6.79*
	R^2	*97.83*

Tabla 5.49: Métricas de error obtenidas para el algoritmo de RNA en el conjunto de datos de prueba para los cinco puntos de presión propuestos.

5.12.5. Comparación de los 3 métodos

La **tabla 5.50** muestra una comparación de los tres modelos propuestos para la estimación de R_s a partir de los datos de entrada propuestos. Se utiliza el promedio de los E_a para comparar los

modelos.

Modelo generado	*Ea de Rs observado en los 4 puntos de presión en el conjunto de prueba [%]*
Regresión lineal	*10.63*
SVR	*7.12*
RNA	*7.68*

Tabla 5.50: Métricas de error promedio obtenidas para los algoritmos propuestos en el conjunto de datos de prueba para los cuatro puntos de presión propuestos.

Para esta propiedad, es posible ver que el modelo de aprendizaje automático que menos error obtiene con el conjunto de datos de prueba es la máquina de vectores de soporte de regresión con un E_a igual a 7.12 %, compitiendo contra las redes neuronales que obtienen un E_a igual a 7.68 %.

5.13. Generación de las curvas completas de R_s

A continuación, se muestran las gráficas generadas a partir de los resultados obtenidos con los modelos de aprendizaje automático para estimar las curva de R_s en el set de validación descrito en la **tabla 5.10** .

Con las **figuras 5.22, 5.23, 5.24, 5.25, 5.26 y 5.27** podemos ver que los tres modelos obtienen resultados muy cercanos a los obtenidos en los reportes PVT, sin embargo, se puede confirmar visualmente que, para estimar R_s, el modelo de aprendizaje automático que tiene una mejor aproximación es la máquina de vectores de soporte de regresión.

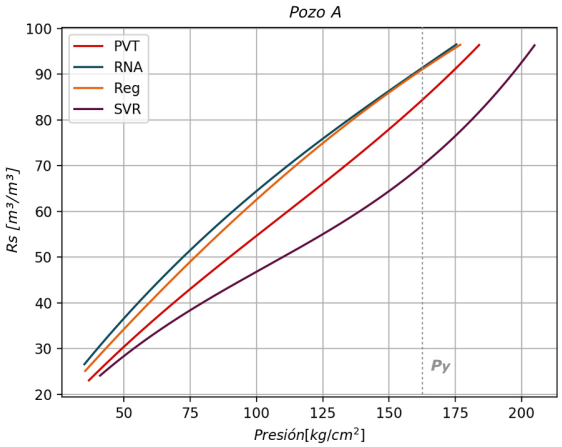

Figura 5.22: Curvas generadas para el Pozo A con los modelos de estimación de R_s propuesto

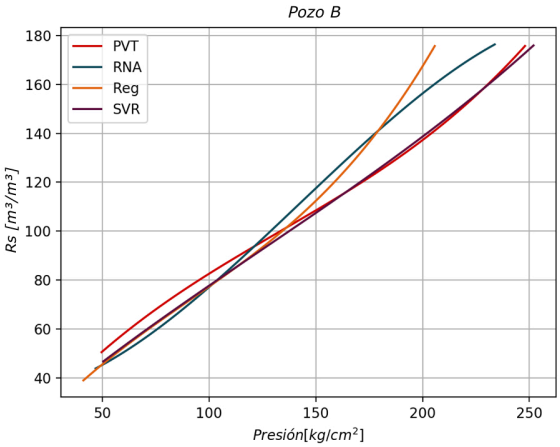

Figura 5.23: Curvas generadas para el Pozo B con los modelos de estimación de R_s propuesto

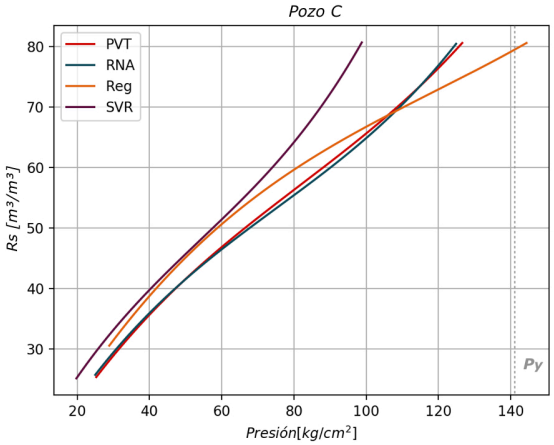

Figura 5.24: Curvas generadas para el Pozo C con los modelos de estimación de R_s propuesto

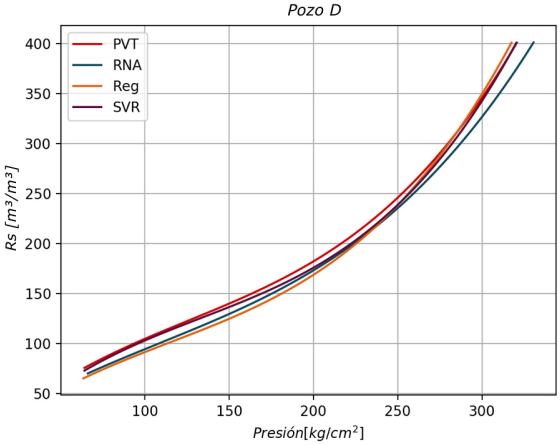

Figura 5.25: Curvas generadas para el Pozo D con los modelos de estimación de R_s propuesto

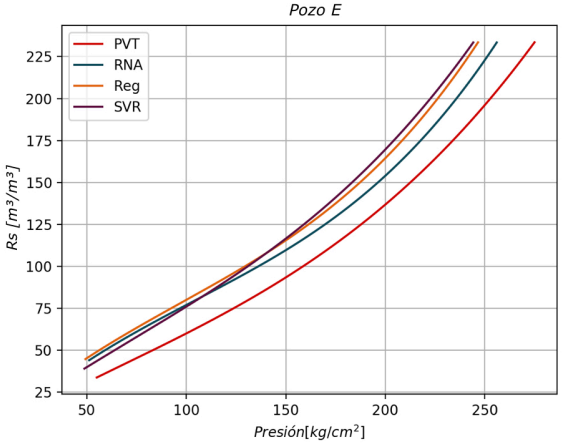

Figura 5.26: Curvas generadas para el Pozo E con los modelos de estimación de R_s propuesto

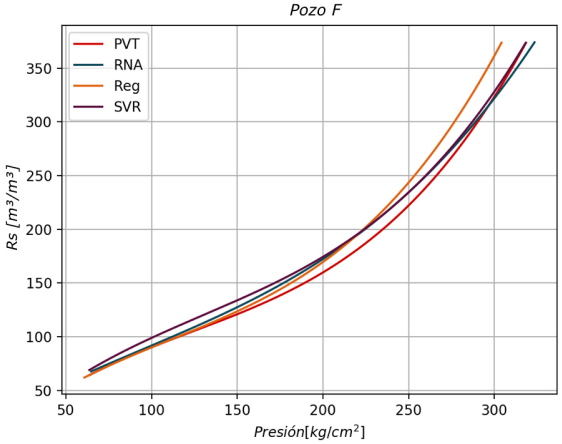

Figura 5.27: Curvas generadas para el Pozo F con los modelos de estimación de R_s propuesto

5.14. Estimación de μ_o en su región saturada

5.14.1. Preparación de entradas

Con base en el mapa de correlación generado (**figura 5.1**), se seleccionaron como entradas T_y, R_s, ρ_{ro} y $\mu_{o@20^oC}$.

El conjunto de datos utilizado para el entrenamiento de los modelos de regresión para calcular la μ_o en su región saturada se compone de muestras de 92 pozos con las características mostradas en las **tablas 5.51 y 5.52**.. A partir de estos datos, se normalizaron todos los datos a partir del método mín-máx.

Tipo	*Cantidad [1]*
Aceite negros	*46*
Aceite volatil	*46*
Total	*92*

Tabla 5.51: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de μ_o en su región saturada.

Rango de datos	*Min*	*Max*	*Promedio*
ρro [1]	*0.798*	*0.933*	*0.854*
Rs [m³/m³]	*11.1*	*503.3*	*220.1*
Ty [°C]	*43.6*	*160*	*116.9*
μo@20°C [cP]	*2.84*	*302.74*	*19.35*
μob [cP]	*0.07*	*14.19*	*0.83*

Tabla 5.52: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar. y validar los modelos de regresión del μ_o en su región saturada y normalizar las variables de entrada durante la etapa de procesamiento de datos.

5.14.2. Regresión lineal

Con este algoritmo se estimaron los valores normalizados de la μ_o en 5 puntos arbitrarios de presión, usando las 4 entradas mencionadas anteriormente y generando las siguiente expresiones

Punto 1

$$\mu_{o@\frac{1}{5}P_b} = -0{,}04988T_y + 0{,}04280\rho_{ro} - -0{,}01361R_s + 0{,}56470\mu_{o@20^oC} + 0{,}03933 \tag{5.25}$$

Punto 2

$$\mu_{o@\frac{2}{5}P_b} = -0{,}053302T_y + +0{,}038406\rho_{ro} - 0{,}02092R_s + 0{,}55183\mu_{o@20^oC} + 0{,}04692 \quad (5.26)$$

Punto 3

$$\mu_{o@\frac{3}{5}P_b} = -0{,}05576T_y + 0{,}04457\rho_{ro} - 0{,}03261R_s + 0{,}55512\mu_{o@20^oC} + 0{,}05360 \quad (5.27)$$

Punto 4

$$\mu_{o@\frac{4}{5}P_b} = -0{,}05837T_y + 0{,}043991\rho_{ro} - 0{,}03887R_s + 0{,}55392\mu_{o@20^oC} + 0{,}05839 \quad (5.28)$$

Punto 5

$$\mu_{o@P_b} = -0{,}06127T_y + 0{,}04330\rho_{ro} - 0{,}04498R_s + 0{,}55394\mu_{o@20^oC} + 0{,}06289 \quad (5.29)$$

La **tabla 5.53** muestra las métricas de error obtenidas con el modelo para la estimación de la μ_o en los puntos de presión propuestos.

Punto	*Indicador*	*Valor [%]*
1/5 Pb	E_a	*62.84*
	R^2	*76.23*
2/5 Pb	E_a	*63.87*
	R^2	*74.90*
3/5 Pb	E_a	*65.22*
	R^2	*75.74*
4/5 Pb	E_a	*69.22*
	R^2	*75.66*
Pb	E_a	*75.66*
	R^2	*77.29*

Tabla 5.53: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba para los puntos de presión.

5.14.3. SVR

El modelo SVR fue entrenado con el mismo conjunto de datos utilizando la función Kernel radial. Se obtuvieron para cada punto de presión propuesto, los siguientes valores de error en el conjunto de datos de prueba.

Se obtuvieron para los puntos de presión propuestos, los siguientes parámetros de error en el conjunto de datos de prueba:

Punto	*Indicador*	*Valor [%]*
1/5 Pb	E_a	*35.92*
	R^2	*92.77*
2/5 Pb	E_a	*43-02*
	R^2	*86.73*
3/5 Pb	E_a	*34.33*
	R^2	*91.14*
4/5 Pb	E_a	*31.86*
	R^2	*93.94*
Pb	E_a	*28.55*
	R^2	*95.87*

Tabla 5.54: Métricas de error obtenidas para el algoritmo de SVR en el conjunto de datos de prueba para los puntos de presión propuestos.

5.14.4. Redes Neuronales Artificiales

El modelo de redes neuronales artificiales para estimar la curva de μ_o en la región saturada también fue entrenado con el conjunto de datos que los modelos anteriores, los parámetros de aprendizaje utilizados en el entrenamiento del modelo fueron los siguientes:

- **Capa de Entrada:** cuatro neuronas correspondientes a las propiedades de entrada para calcular μ_{ro}.
- **Primera Capa Oculta:** cuatro neuronas
- **Segunda Capa Oculta:** tres neuronas
- **Capa de Salida:** una capa de salida con cinco neuronas equivalente al valor de μ_{ro} calculado en cada punto de presión.
- **Función de activación:** sigmoidal

En la **tabla 5.55** se muestran las métricas de error del modelo redes neuronales creado para estimar μ_o en los 5 puntos de presión propuestos de la región saturada.

Punto	Indicador	Valor [%]
1/5 Pb	E_a	45.84
	R^2	75.87
2/5 Pb	E_a	40.91
	R^2	87.87
3/5 Pb	E_a	36.47
	R^2	91.56
4/5 Pb	E_a	27.76
	R^2	95.17
Pb	E_a	24.03
	R^2	96.1

Tabla 5.55: Métricas de error obtenidas para el algoritmo de RNA en el conjunto de datos de prueba para los cinco puntos de presión propuestos.

5.14.5. Comparación de los 3 métodos

La **tabla 5.56** muestra una comparación de los tres modelos propuestos para la estimación de la viscosidad del aceite en la región saturada a partir de los datos de entrada propuestos.

Modelo generado	*Ea de μo observado en los 5 puntos de presión en el conjunto de prueba [%]*
Regresión lineal	*67.40*
SVR	*34.74*
RNA	*37.00*

Tabla 5.56: Métricas de error promedio obtenidas para los algoritmos propuestos en el conjunto de datos de prueba para los cinco puntos de presión propuestos.

Para esta propiedad, podemos ver que el porcentaje de error para los tres modelos utilizados es grande, sin embargo, los modelos que compiten para obtener una mejor aproximación son la máquina de vectores de soporte de regresión con un E_a igual a 34.74 % y las redes neuronales artificiales con un E_a igual a 37 %.

5.15. Estimación de la μ_o en su región bajo saturada

5.15.1. Preparación de entradas

A partir del mapa de correlación mostrado en la **figura 5.1**, se seleccionaron como entradas de alimentación a los algoritmos de aprendizaje automático: T_y, R_s y ρ_{ro}.

El conjunto de datos utilizado para el entrenamiento de los modelos de regresión para calcular la μ_o en su región bajo saturada se compone de muestras de 99 pozos con las características mostradas en las **tablas 5.57 y 5.58** . A partir de esta información, se normalizaron todos los datos de entrada a partir del método mín-máx.

Tipo	*Cantidad [1]*
Aceite negros	*51*
Aceite volatil	*48*
Total	*99*

Tabla 5.57: Número de muestras del conjunto de datos utilizado para entrenar y probar los modelos de μ_o en su región bajo saturada

Rango de datos	Min	Max	Promedio
ρro [1]	0.798	0.933	0.854
Rs [m³/m³]	11.1	783.1	220
Ty [°C]	42.6	162.8	116.1
μo@20°C [cP]	2.84	302.74	19.29
△(μo@Pb+200[kg/cm²] - μo@Pb) [1]	0.001	4.972	0.221

Tabla 5.58: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar y validar los modelos de regresión de la μ_o en su región bajo saturada y normalizar las variables de entrada durante la etapa de procesamiento de datos.

5.15.2. Regresión lineal

Mediante este algoritmo es posible estimar el valor normalizado de $(\mu_o@Pb - \mu_o@Pb + 200[kg/cm])cP]$ a partir de la temperatura del yacimiento y la relación gas-aceite disuelto con el uso de la expresión:

$$\mu_o@Pb - \mu_o@Pb + 200[kg/cm]) = -0{,}07838T_y - 0{,}15239R_s + 0{,}13595 \tag{5.30}$$

Las métricas de error obtenidas para este algoritmo en el conjunto de datos de prueba se observan en la **tabla 5.59** .

Indicador	Valor [%]
E_a	298.93 %
R^2	75.04

Tabla 5.59: Métricas de error obtenidas para el algoritmo de Regresión Lineal en el conjunto de datos de prueba.

5.15.3. SVR

El modelo de SVR fue entrenado con función Kernel radial y, haciendo uso del mismo conjunto de datos. Estas fueron las métricas de error obtenidas en el conjunto de prueba con este modelo.

Indicador	*Valor [%]*
E_a	*532.04 %*
R^2	*78.43*

Tabla 5.60: Métricas de error obtenidas para el algoritmo SVR en el conjunto de datos de prueba.

5.15.4. Redes Neuronales Artificiales

Estos son los parámetros de aprendizaje que mejores resultados obtuvieron para realizar el entrenamiento del modelo de redes neuronales para la estimación de $(\mu_o@Pb - \mu_o@Pb + 200[kg/cm])$

- **Capa de Entrada:** tres neuronas correspondientes a las propiedades de entrada para calcular $(\rho_{ro}@Pb - \rho_{ro}@Pb + 200[kg/cm])$.
- **Primera Capa Oculta:** seis neuronas
- **Segunda Capa Oculta:** seis neuronas
- **Capa de Salida:** una capa de salida con una neurona equivalente al valor de $(\mu_o@Pb - \mu_o@Pb + 200[kg/cm])$ calculado.
- **Función de activación:** sigmoidal

En la **figura 5.28** se muestra la estructura de la red neuronal artificial propuesta

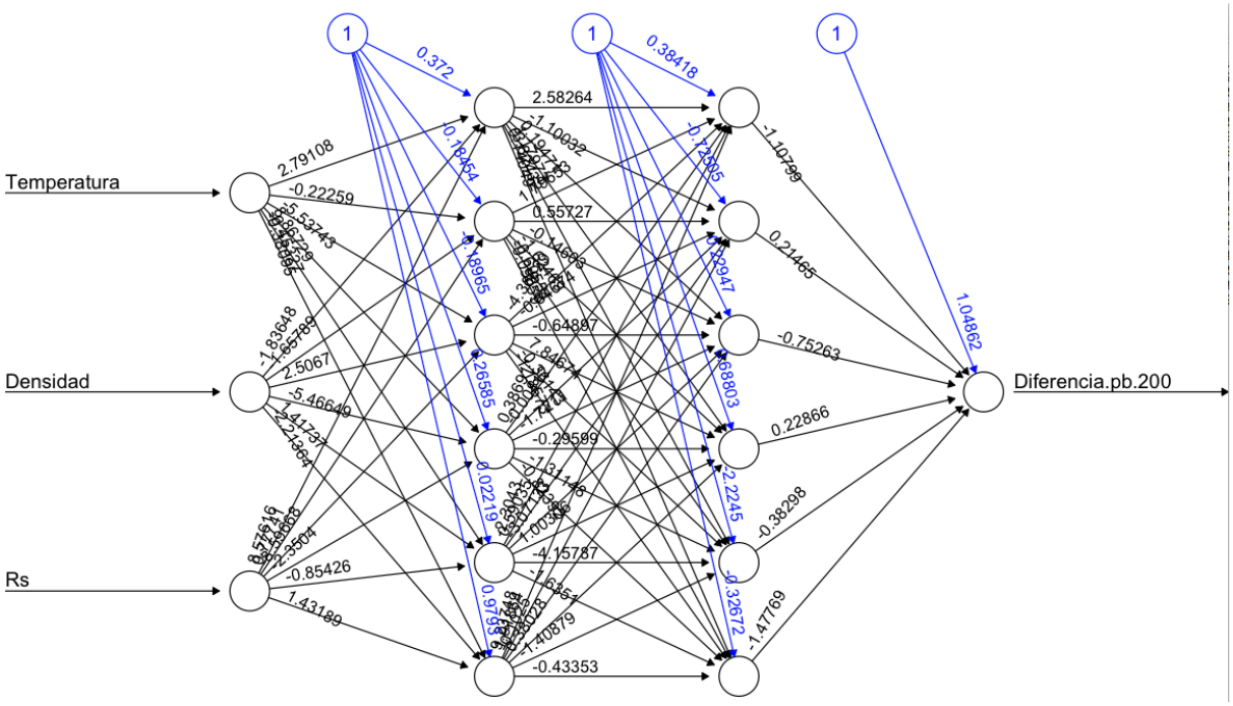

Figura 5.28: Estructura del modelo RNA para la estimación de $(\mu_o@Pb - \mu_o@Pb + 200[kg/cm])$

En la **tabla 5.61** se muestran las métricas de error del modelo redes neuronales creado para estimar $(\mu_o@Pb - \mu_o@Pb + 200[kg/cm])$

Indicador	*Valor [%]*
E_a	*216.16*
R^2	*59.03*

Tabla 5.61: Métricas de error obtenidas para el algoritmo RNA en el conjunto de datos de prueba.

5.16. Comparación de los tres métodos

El cálculo de $(\mu_o@Pb - \mu_o@Pb + 200[kg/cm])$ tiende a suponer errores mayores que el error real, por lo que se debe normalizar dicho resultado llevándolo a condiciones de presión de yacimiento (μ_{oy}).

La**tabla 5.62** muestra una comparación del E_a de los tres modelos propuestos para la estimación de μ_{oy} ya normalizado a condiciones de presión de yacimiento evaluado en el set de validación.

Modelo generado	Ea del μoy para el conjunto de validación [%]
Regresión lineal	43.36
SVR	44.74
RNA	21.17

Tabla 5.62: Métricas de error obtenidas para los algoritmos en el conjunto de datos de validación.

Se puede observar que para estimar la μ_{oy}, el modelo de aprendizaje automático que menos error consigue en el conjunto de pruebas es las redes neuronales artificiales con un error porcentual absoluto medio de 21.17 %, por lo que, para obtener esta propiedad, este es el modelo más competitivo.

Los valores de μ_{oy} calculados con los tres algoritmos mencionados anteriormente en el conjunto de validación, así como el valor de μ_{oy} registrado en los reportes PVT, pueden apreciarse en la **5.63**.

Pozo	μoy Real [1]	μoy Regresión lineal [1]	μoy SVR [1]	μoy RNA [1]
A	0.59	1.15	1.58	0.88
B	0.50	0.94	0.47	0.66
C	1.70	1.35	1.62	1.66
D	0.31	0.22	0.42	0.30
E	0.28	0.22	0.26	0.23
F	0.22	0.21	0.33	0.27

Tabla 5.63: Resultados obtenidos para todos los algoritmos en el conjunto de datos de validación.

5.17. Generación de las curvas completas de μ_o

A continuación se muestran las gráficas generadas a partir de la unión de los modelos de aprendizaje automático elaborados en este trabajo para estimar μ_y en la región saturada y bajo saturada.

Con las **figuras 5.29, 5.30, 5.31, 5.32, 5.33 y 5.34** es posible observar de manera gráfica que, para estimar la μ_o, el de aprendizaje automático que tiene una mejor aproximación es el de neuronal artificial.

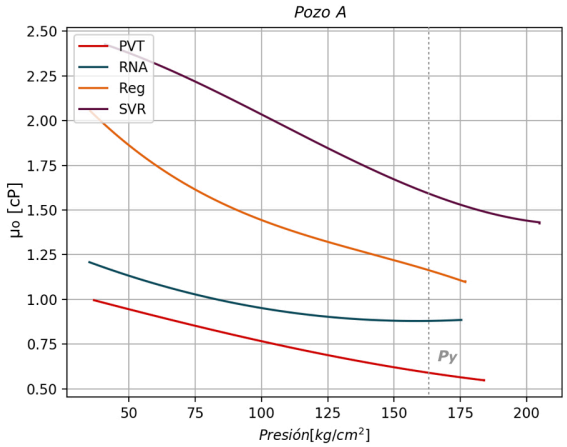

Figura 5.29: Curvas generadas para el Pozo A con los modelos de estimación de μ_{ro} propuesto

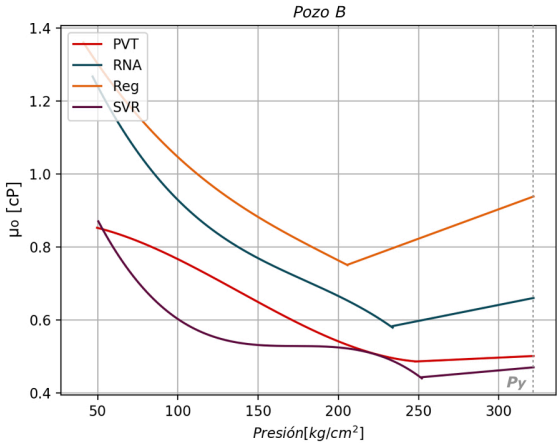

Figura 5.30: Curvas generadas para el Pozo B con los modelos de estimación de μ_{ro} propuesto

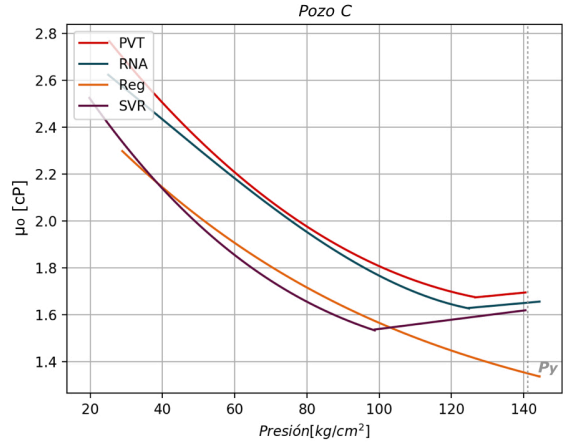

Figura 5.31: Curvas generadas para el Pozo C con los modelos de estimación de μ_{ro} propuesto

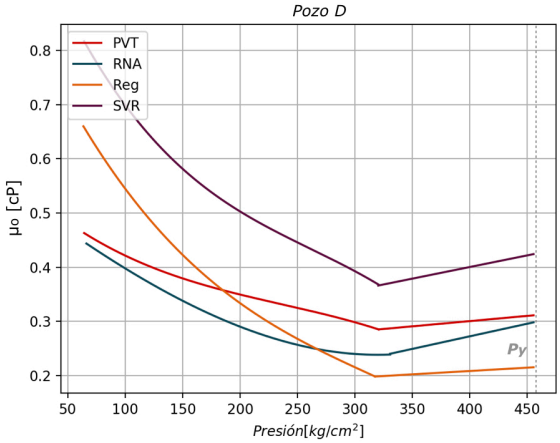

Figura 5.32: Curvas generadas para el Pozo D con los modelos de estimación de μ_{ro} propuesto

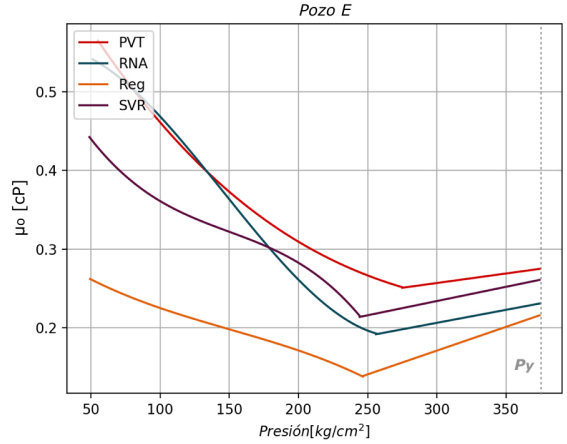

Figura 5.33: Curvas generadas para el Pozo E con los modelos de estimación de μ_{ro} propuesto

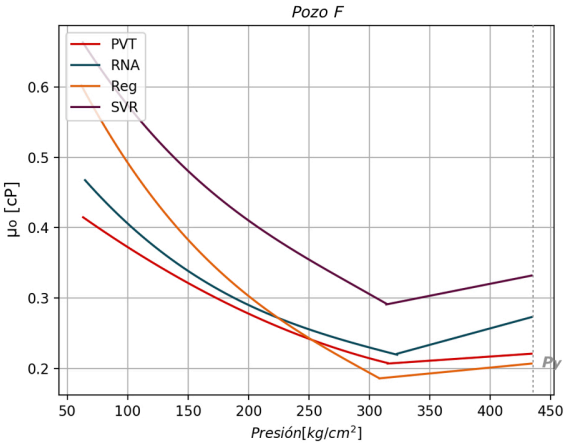

Figura 5.34: Curvas generadas para el Pozo F con los modelos de estimación de μ_{ro} propuesto

Capítulo 6

Comparaciones de resultados con correlaciones

Con el fin de obtener una referencia de los resultados obtenidos en este trabajo, tomaremos el mejor modelo de aprendizaje automático propuesto para la estimacion de cada propiedad y lo compararemos contra los resultados obtenidos implementando las correlaciones de Standing, Vazquez y Kartoatmojdo las cuales son de uso común y están ampliamente disponibles en la literatura.

6.1. P_b

La primera propiedad a comparar es la presión de punto de burbuja. Debido a su buen rendimiento, escogimos los resultados del modelo de redes neuronales artificiales como punto de comparación con las correlaciones implementadas en el conjunto de datos de validación.

Pozo	Pb PVT [kg/cm²]	Pb RNA [kg/cm²]	Pb Standing [kg/cm²]	Pb Vázquez [kg/cm²]	Pb Kartoatmojdo[kg/cm²]
A	184.0	175.4	239.4	224.1	237.3
B	248.0	233.8	252.6	231.5	255.9
C	126.6	124.9	129.8	148.5	141.6
D	320.7	330.6	511.4	428.9	516.0
E	275.0	256.0	322.2	265.7	303.8
F	318.5	323.5	467.4	396.5	475.4

Tabla 6.1: resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos utilizando el conjunto de datos de validación

Se puede observar en las **figuras 6.1 y 6.2**que el modelo de RNA elaborado supera con creces al rendimiento obtenido por las correlaciones de uso tradicional en el conjunto de validación

6.2. B_o

La siguiente propiedad a comparar es el factor de volumen del aceite, específicamente la curva generada conforme varía la presión del sistema y la temperatura se mantiene constante (tempe-

Método	Ea de Pb estimado [%]
RNA	3.89
Standing	26.31
Vázquez	17.89
Kartoatmojdo	27.44

Tabla 6.2: comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar la P_b en el conjunto de validación

ratura del yacimiento). Debido a su buen rendimiento, escogimos los resultados del modelo de redes neuronales artificiales como punto de comparación con las correlaciones implementadas en el conjunto de datos de validación. Es importante recalcar que para hacer estas comparaciones de resultados ponderamos los resultados en la región saturada y bajo saturada de cada muestra.

Método	Ea de Bo observado en los 6 puntos de presión en el conjunto de validación [%]
RNA	2.96
Standing	5.66
Vázquez	4.86
Kartoatmojdo	5.14

Tabla 6.3: comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar la B_o en el conjunto de validación

Las **figuras 6.1, 6.2, 6.3, 6.4, 6.5 y 6.6** muestran los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos utilizando el conjunto de datos de validación y las presiones de saturación estimadas con sus respectivos modelos

Se puede observar que el modelo de RNA elaborado supera con el rendimiento obtenido por las correlaciones de uso tradicional en la mayoría de los pozos de el conjunto de validación

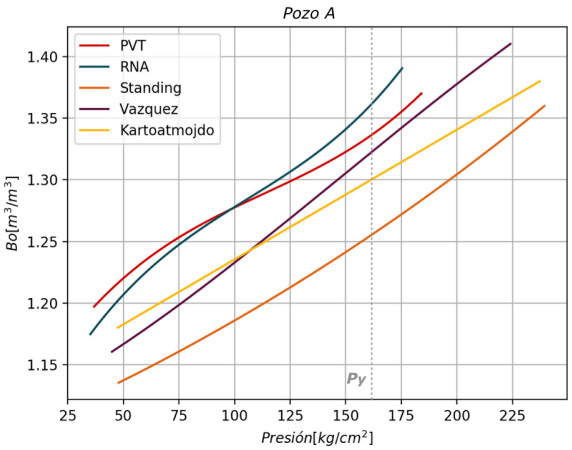

Figura 6.1: Curvas de B_o Pozo A

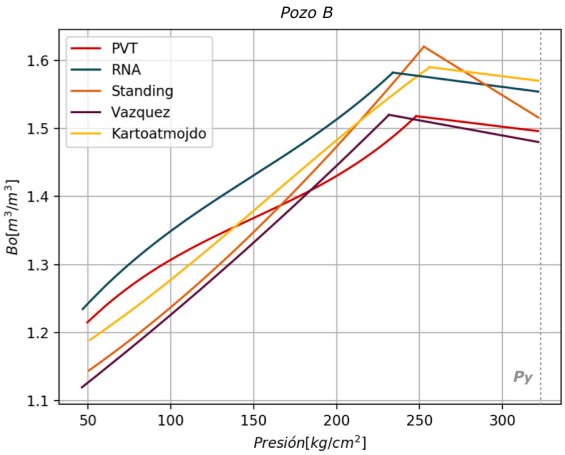

Figura 6.2: Curvas de B_o Pozo B

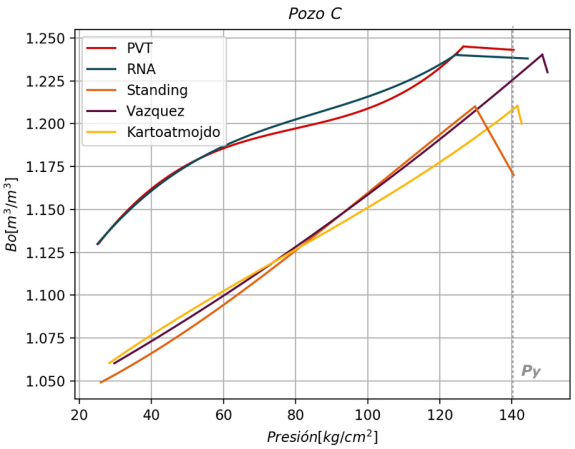

Figura 6.3: Curvas de B_o Pozo C

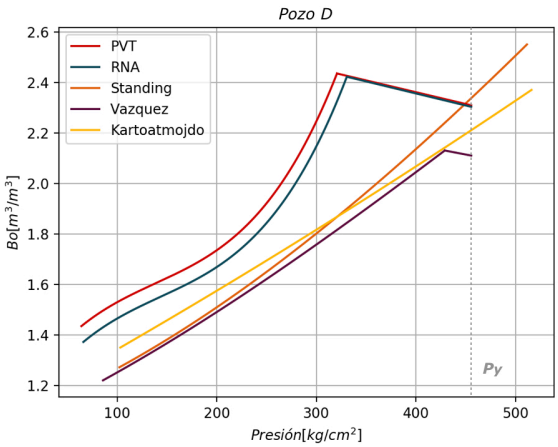

Figura 6.4: Curvas de B_o Pozo D

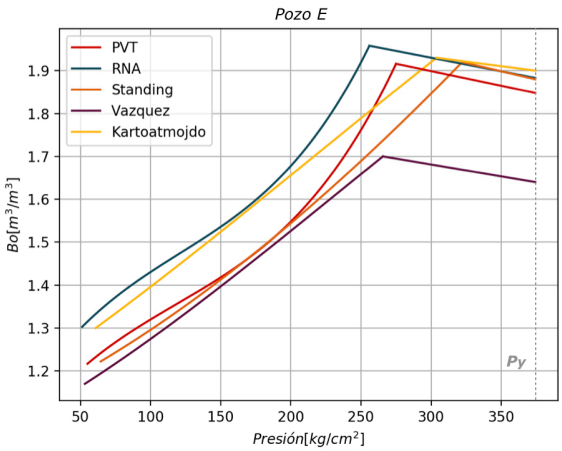

Figura 6.5: Curvas de B_o Pozo E

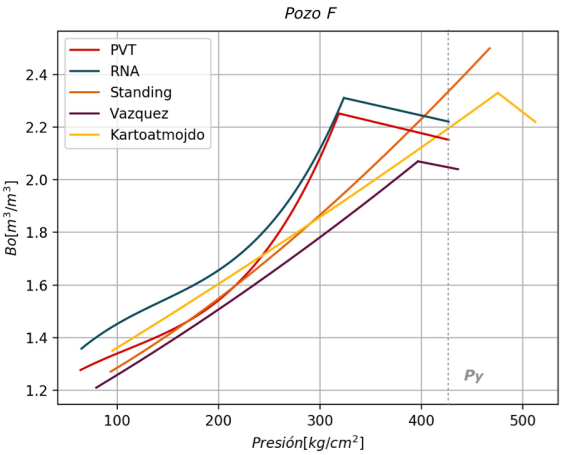

Figura 6.6: Curvas de B_o Pozo F

6.3. ρ_{ro}

La siguiente propiedad a comparar en este trabajo es la densidad relativa del aceite, específicamente la curva generada conforme varía la presión del sistema y la temperatura se mantiene constante (temperatura del yacimiento). Debido a su buen rendimiento, escogimos los resultados del modelo de SVR como punto de comparación con las correlaciones implementadas en el conjunto de datos de validación. Es importante recalcar que para hacer estas comparaciones de resultados ponderamos los resultados en la región saturada y bajo saturada de cada muestra.

Modelo	*Ea de pro observado en los 6 puntos de presión en el conjunto de validación [%]*
SVR	*2.26*
Standing	*5.29*
Vázquez	*3.92*
Kartoatmojdo	*4.55*

Tabla 6.4: comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar la ρ_{ro} en el conjunto de validación

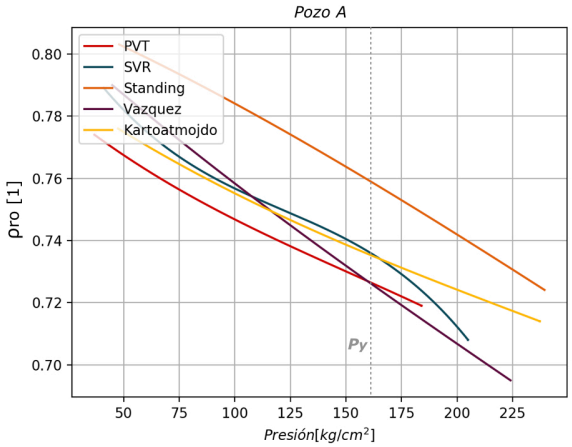

Figura 6.7: Curvas de ρ_{ro} Pozo A

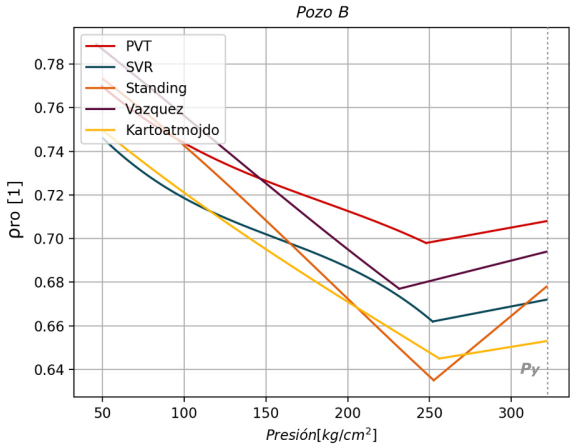

Figura 6.8: Curvas de ρ_{ro} Pozo B

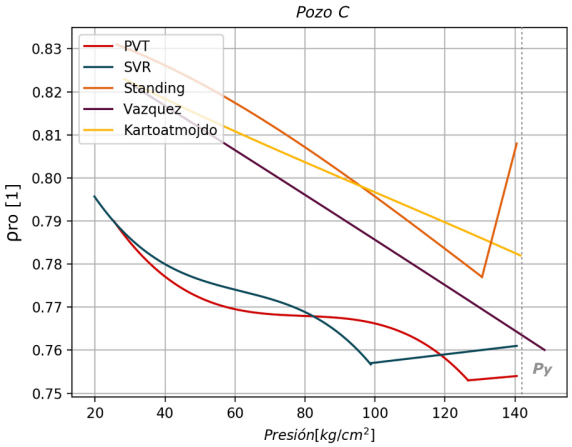

Figura 6.9: Curvas de ρ_{ro} Pozo C

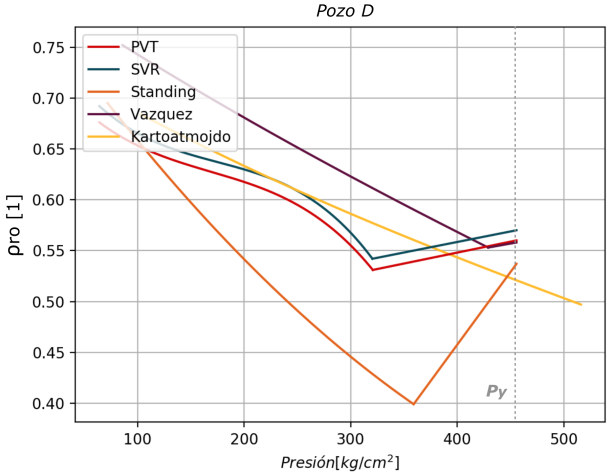

Figura 6.10: Curvas de ρ_{ro} Pozo D

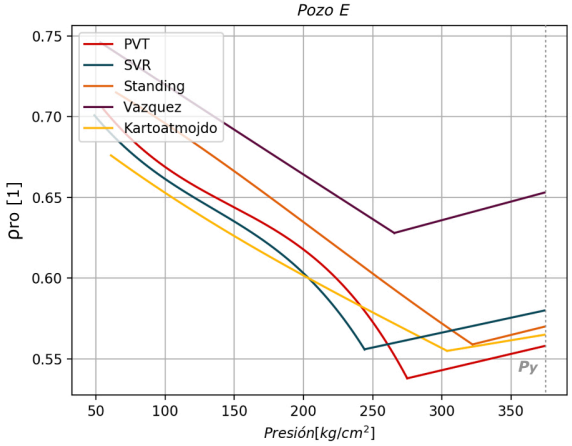

Figura 6.11: Curvas de ρ_{ro} Pozo E

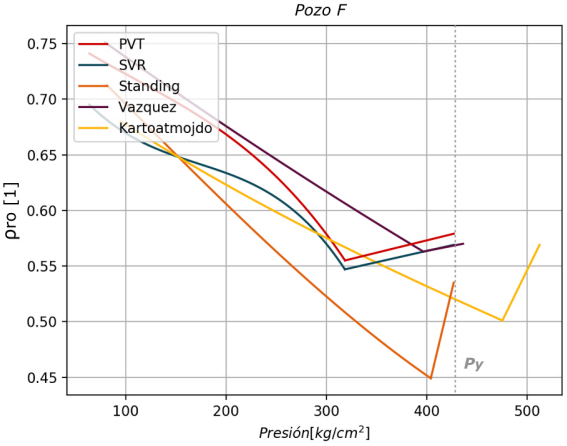

Figura 6.12: Curvas de ρ_{ro} Pozo F

Las **figuras 6.7, 6.8, 6.9, 6.10, 6.11 y 6.12** muestran los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos utilizando el conjunto de datos de validación y las presiones de saturación estimadas con sus respectivos modelos

Se puede observar que el modelo de SVR elaborado supera el rendimiento obtenido por las correlaciones de uso tradicional en la mayoría de los pozos del conjunto de validación, sin embargo el modelo de Vázquez se observa también muy competitivo en algunos pozos.

6.4. R_s

La siguiente propiedad a comparar en este trabajo es la relacion gas aceite, específicamente la curva generada conforme varía la presión del sistema y la temperatura se mantiene constante (temperatura del yacimiento). Debido a su buen rendimiento, escogimos los resultados del modelo de RNA como punto de comparación con las correlaciones implementadas en el conjunto de datos de validación. Es importante recalcar que para hacer estas comparaciones de resultados ponderamos los resultados en la región saturada.

Método	*Ea de Pb estimado [%]*
RNA	*3.89*
Standing	*26.31*
Vázquez	*17.89*
Kartoatmojdo	*27.44*

Tabla 6.5: comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar la R_s en el conjunto de validación

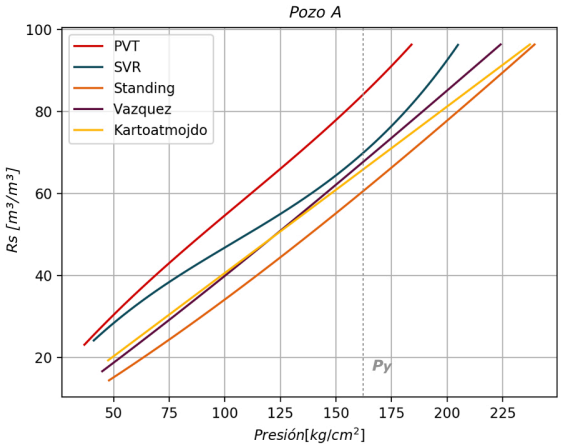

Figura 6.13: Curvas de R_s Pozo A

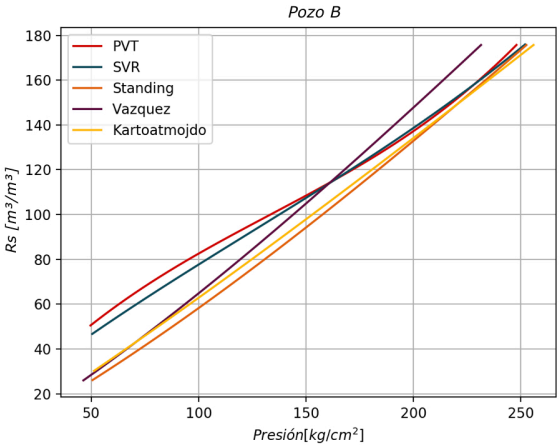

Figura 6.14: Curvas de R_s Pozo B

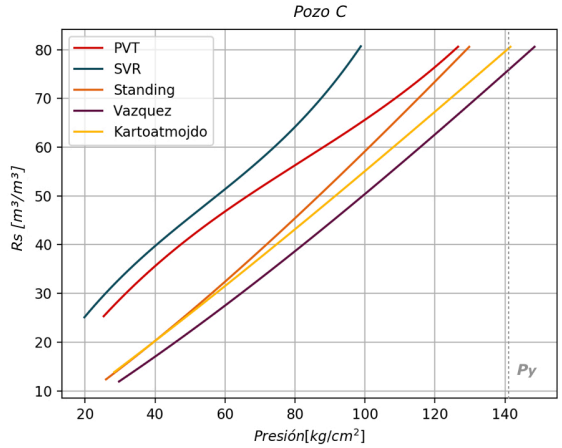

Figura 6.15: Curvas de R_s Pozo C

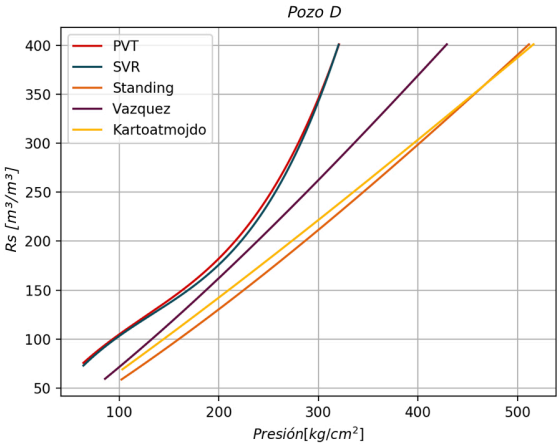

Figura 6.16: Curvas de R_s Pozo D

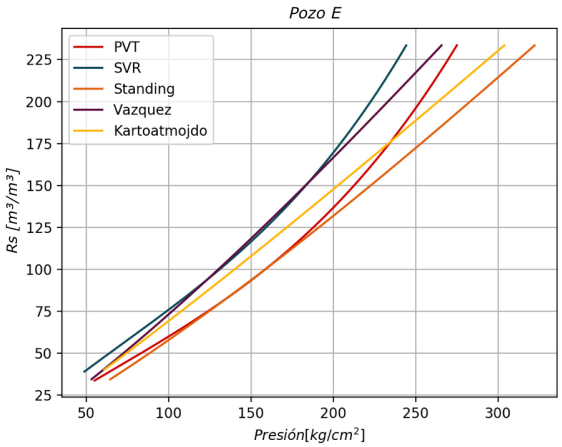

Figura 6.17: Curvas de R_s Pozo E

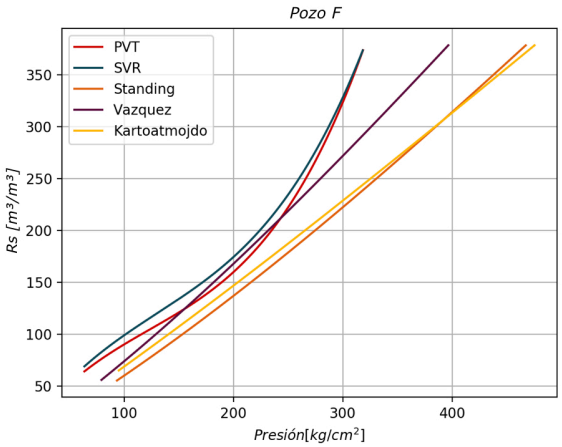

Figura 6.18: Curvas de R_s Pozo F

Las **figuras 6.13, 6.14, 6.15, 6.16, 6.17 y 6.18** muestran los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos utilizando el conjunto de datos de validación y las presiones de saturación estimadas con sus respectivos modelos.

Se puede observar que el modelo de RNA elaborado supera el rendimiento obtenido por las correlaciones de uso tradicional en la mayoría de los pozos del conjunto de validación.

6.5. μ_o

La última propiedad a comparar en este trabajo es la viscosidad del aceite, específicamente la curva generada conforme varía la presión del sistema y la temperatura se mantiene constante (temperatura del yacimiento). Debido a su buen rendimiento, escogimos los resultados del modelo de RNA como punto de comparación con las correlaciones implementadas en el conjunto de datos de validación. Es importante recalcar que para hacer estas comparaciones de resultados ponderamos los resultados en la región saturada y bajo saturada de cada muestra.

Modelo	*Ea de μo observado en los 6 puntos de presión en el conjunto de validación [%]*
RNA	*17.00*
Vázquez	*15.74*
Kartoatmojdo	*38.62*

Tabla 6.6: comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar la μ_o en el conjunto de validación. No se ha incluido a Standing debido a que no cuenta con correlación para el cálculo de la viscosidad del aceite.

Las **figuras 6.19, 6.20, 6.21, 6.22, 6.23 y 6.24** muestran los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos utilizando el conjunto de datos de validación y las presiones de saturación estimadas con sus respectivos modelos

Se puede observar que el modelo de RNA elaborado tiene un rendimiento competitivo en la mayoría de los pozos del conjunto de validación, sin embargo el modelo de Vázquez es más competitivo en cálculo de la viscosidad en la mayoría de los pozos, esto puede ser debido a que las variaciones reológicas presentes en los aceites estudiados son mejor representadas para este conjunto de datos por dicho modelo.

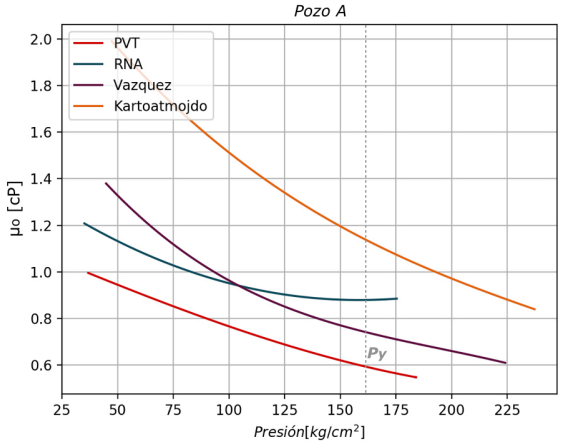

Figura 6.19: Curvas de μ_o Pozo A

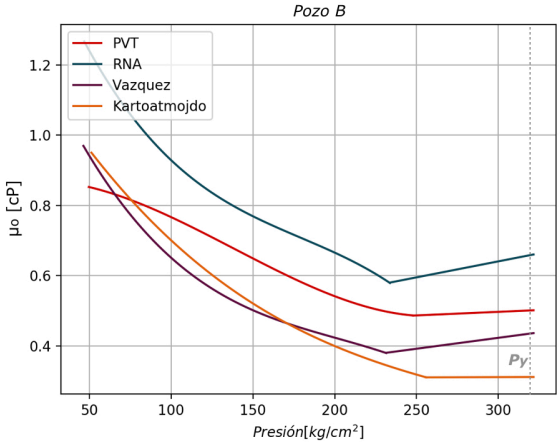

Figura 6.20: Curvas de μ_o Pozo B

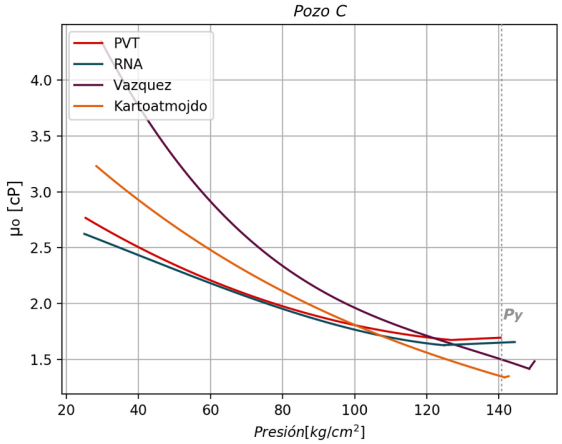

Figura 6.21: Curvas de μ_o Pozo C

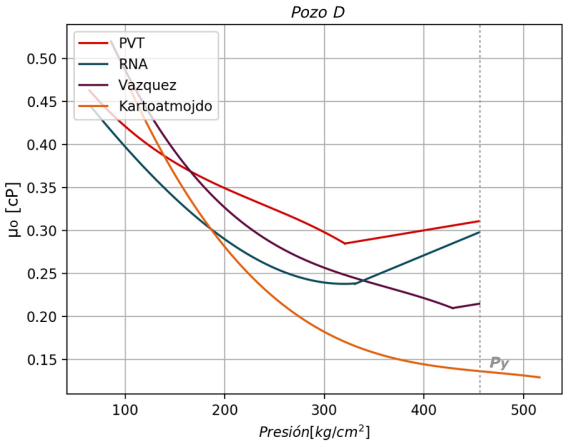

Figura 6.22: Curvas de μ_o Pozo D

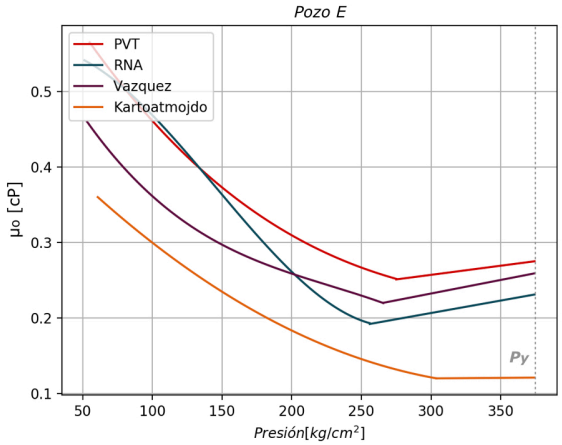

Figura 6.23: Curvas de μ_o Pozo E

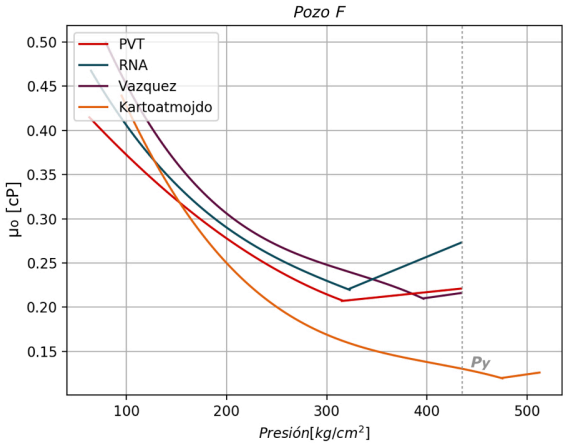

Figura 6.24: Curvas de μ_o Pozo F

Conclusiones

La industria del petróleo se enfrenta a un futuro difícil en donde se requiere explorar procesos más eficientes con el fin de ahorrar costos y hacer el negocio más rentable. Por eso es importante realizar la introducción del aprendizaje automático en diferentes disciplinas de la ingeniería petrolera.

En este trabajo se pudo comprobar que el desempeño del aprendizaje automático, cuando se cuenta con la información suficiente y de buena calidad, supera en algunos casos el rendimiento de los métodos de estimación de propiedades PVT disponibles en la literatura actual.

La correcta implementación de estas técnicas para estimar resultados requiere de la existencia de un banco de datos amplio y lo suficientemente correlacionado con las variables de salida. Por lo tanto es recomendable aplicar esta tecnología en campos maduros en los que se cuenta con suficiente información de pruebas PVT tomadas con anterioridad. Su uso representa un ahorro en el número de las futuras tomas de pruebas PVT y por lo tanto un ahorro en el presupuesto del proyecto.

También debemos dejar claro que el modelo tendrá los mejores resultados en su región de origen, es decir si se evalúa con aceites de otras regiones es probable que muestre resultados más pobres. Por lo que se debe en teoría si se quiere aplicar esta tecnología en proyectos reales, crear un modelo para cada región, campo o yacimiento en específico.

Recomendaciones

Para seguir con el desarrollo del tema se recomienda en futuras entregas las siguientes líneas de investigación:

- La elaboración de un conjunto de datos aún mayor al disponible para este trabajo, con el fin de obtener mejores resultados que los obtenidos. Ya que a mayor número de muestras, mayor la eficiencia de los modelos de aprendizaje automático.
- La preparación de una interfaz gráfica con el fin de facilitar la experiencia de un usuario que desea estimar los resultados utilizando el modelo, ya que actualmente la carga de información se hace de forma manual en archivo con extensión .csv y se corre desde consola.
- La creación de un modelo de aprendizaje automático capaz de calcular R_{sb} a partir de datos superficiales como la RGA, diámetro de estrangulador, condiciones de separación entre otros.
- La realización de un modelo de aprendizaje automático capaz de calcular la viscosidad del aceite a condiciones de superficie cuando no se conoce, a partir de datos superficiales del aceite.
- La creación de modelos para la estimación de las propiedades PVT de otros tipos de fluidos como los son aceites pesados, extrapesados, gas seco, gas condensado y gas húmedo.
- La exploración y uso de otros algoritmos de aprendizaje automático no utilizados en este trabajo con el fin de obtener una mejor eficiencia como son los algoritmos de lógica difusa o redes neuronales recurrentes entre otros.

Anexo A

Parámetro	*Símbolo*	*Definición*	*Fórmula*
Error relativo	E_i	*Es la medida de la desviación relativa entre un valor estimado y un dato experimental*	$E_i = \left[\frac{(X)_{exp} - (X)_{est}}{(X)_{exp}}\right] x\ 100$
Error relativo promedio	E_r	*Es la medida de la desviación relativa entre los valores estimados y los datos experimentales*	$E_r = \frac{1}{n}\sum_{i=1}^{N} E_i$
Error relativo absoluto promedio	E_a	*Es la medida del valor absoluto de la desviación relativa entre los valores estimados y los datos experimentales*	$E_a = \frac{1}{n}\sum_{i=1}^{N} \lvert E_i \rvert$
Coeficiente de Correlación	R^2	*Es la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, también llamado R cuadrado, refleja la bondad del ajuste de un modelo a la variable que pretender explicar.* *El R-cuadrado siempre está entre 0 y 1:* • *El 0 indica que el modelo no explica ninguna porción de la variabilidad de los datos de respuesta en torno a su media.* • *El 1 indica que el modelo explica toda la variabilidad de los datos de respuesta en torno a su media.*	$R^2 = \frac{\sum_{i=1}^{N}\left((X)_{est} - (\bar{x})_{exp}\right)^2}{\sum_{i=1}^{N}\left((X)_{exp} - (\bar{x})_{exp}\right)^2}$ *Donde:* $(X)_{est}$ *= Valor estimado* $(X)_{exp}$ *= Valor experimental* *N,n = Número de valores* $(\bar{x})_{exp}$ *= Valor promedio*

Anexo A: Métricas utilizadas para evaluar los resultados de los modelos empleados en este trabajo.

Referencias

Ahmed, T. (2006). Reservoir Engineering Handbook. Elsevier, pp. 1 – 182.

Al-Marhourn, M. & Osman, E. (2002). Using Artificial Neural Networks to Develop New PVT Correlations for Saudi Crude Oils. Society of Petroleum Engineers, SPE - 78592.

Alpaydin, E. (2014).Introduction to Machine Learning. Massachusetts Institute of Technology.

Ayodele, T. (2010). New Advances in Machine Learning. IntechOpen.

Baarimah, S., Gawish, A. & BinMerdhah, A. (2015). Artificial Intelligence Techniques for Predicting the Reservoir Fluid Properties of Crude Oil Systems. International Research Journal of Engineering and Technology, Volume: 02 Issue: 07.

Banko, M. & Bill, E. (2001). Scaling to very very large corpora for natural language disambiguation. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics.

Banzer,C. (1996). Correlacion Numéricas P.V.T. Universidad del Zulia, pp. 48 – 110.

Camargo, A. (2016). Estimación de propiedades PVT para yacimientos de aceite negro y volátil mediante una red neuronal artificial. Universidad Nacional Autónoma de México.

Carlson,M. (2006). Practical Reservoir Simulation: Using, Assessing, and Developing Results. PenWell Publishing Company, pp. 121.

Chapelle, O., Schölkopf, B. & Zien, A. (2006).Semi-Supervised Learning. The MIT Press.Dekel, O. (2009). From Online to Batch Learning with Cutoff-Averaging. Microsoft Research.

Dandekar,A (2013). Petroleum Reservoir Rock and Fluid Properties. CRC Press, pp. 313.

Dake,L. (1998). Fundamentals of Reservoir Engineering. Elsevier, pp. 52.

Dekel, O. (2009). From Online to Batch Learning with Cutoff-Averaging. Microsoft Research.

Fath, A. Pouranfard, A. & Foroughizadeh, P. (2018). Development of an artificial neural network model for prediction of bubble point pressure of crude oils. Ke Ai, Petroleum 4 pp. 281 – 291.

Ghahramani, Z. (2004). Unsupervised Learning. University College London.

Gharbi, R., Elsharkawy, A. & Karkboub, M. (1999). Universal Neural Network-Based Model for Estimating the PVT Properties of Crude Oil Systems. Energy and Fuels, 13, pp. 454 – 458.

Goodfellow, I., Bengio, Y. & Courville, A. (2016).Deep Learning. The MIT Press.

Hassn, O & Sadiq, D. (2009). New Correlation For Oil Formation Volume Factor At And Below Bubble Point Pressure. Journal of Engineering. Número 4, Volumen 15.

Hernández, I. (2017). Aplicación de redes neuronales en la Ingeniería Petrolera. Universidad Nacional Autónoma de México.

Hurwitz, J. & Kirsch, D. (2018).Machine Learning para principiantes.

McCain,W. (1990). The Properties of Petroleum Fluids, Second Edition. PenWell Publishing Company, pp. 46 – 159.

McKinnell, M. Verhein,J. & Yu, P. (2020). Phase Diagram. url: www.chem.libretexts.org.

Mendez, L. Teyssier, S. (1979). Caracterización de Fluidos de Yacimientos Petroleros. Revista Instituto Mexicano del Petróleo, Vol X, Número: 4.

Nagi, J., Kiong, T., Ahmed, S. & Nagi, F. (2009). Prediction of PVT Properties in crude oil systems using support vector machines. 3rd International Conference of Energy and Environment.

Oloso, M., Hassan, M., Bader-El-Den, M. & Buick, J. (2017). Ensemble SVM for characterization of crude oil viscosity. KACST.

Osman, E., Abdel-Wahhab, O. & Al-Marhoun, M. (2001). Prediction of Oil PVT Properties Using Neural Networks. SPE Middle East Oil Show held in Bahrain, SPE - 68233.

Ramírez, A., Valle, G., Romero, F. & Jaimes, M. (2017). Production of PVT Properties in Crude Oil Using Machine Learning Techniques. SPE Latin America and Caribbean, SPE – 185536 – MS.

Redacción APD. (2019). "¿Qué es Machine Learning y cómo funciona?", url: www.APD.es.

Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development.

Santos,P. (2013). Caracterización Integral de Fluidos de los Yacimientos Petroleros. Universidad Nacional Autónoma De México.

Schlumberger. (Consultado el 06 – 2020). Schlumberger Oilfield Glossary. url: www. glossary.oilfield.slb.com/.

Selamat, A., Olatunji, S. & Raheem, A. (2012). Modeling PVT Properties of Crude Oil Systems based on Type-2 Fuzzy Logic Approach and Sensivity Based Linear Learning Method. Springer-Verlag Berlin Heidelberg.

Shai, S. & Shai, B. (2014).Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

Sharma, S., Sharma, S. & Athaiya, A. (2020). Activation Functions in Neural Networks. IJEAS.

Shizhen,T. Yuan, X. & Hou L. (2016). Play types, geologic characteristics and exploration domains of lithological reservoirs in China. Petroleum Exploration And Development, Volumen 43, Número 6.

Smola, A. & Vishwanathan, S. (2008).Introduction to Machine Learning. Cambridge University Press.

Wood, D., Choubineh, A. (2019). Reliable predictions of oil formation volume factor based on transparent and auditable machine learning approaches. Advances in Geo-Energy Research, Volumen 3, Número 3, pp. 225 - 241.

Wu, R. & Rosenegger, L. (1997). EOS Oil Characterization Aids Integrated Reservoir Studies. Petroleum Society of Canada, Annual Technical Meeting.

Yee, K. (2011). Simulation of Phase Behavior of Crude Oil using Different Equations of State (EOS) Universiti Teknologi Petronas.