



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA
DIVISIÓN DE INGENIERÍA ELÉCTRICA

MODELADO DE DATOS DE FLUIDOS PETROLEROS: UN
ENFOQUE DE APRENDIZAJE AUTOMÁTICO

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

INGENIERA EN COMPUTACIÓN

PRESENTA:

ILSE ABRIL VÁZQUEZ SÁNCHEZ

DIRECTOR DE TESIS:

DR. ROBERTO GIOVANNI RAMÍREZ CHAVARRÍA



Ciudad Universitaria, CDMX 2021



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Este trabajo fue realizado en colaboración con José Enrique Leal Castillo, egresado de Ingeniería Petrolera.

Agradecimientos

Después de concluir esta etapa de mi vida académica con éxito, quiero agradecer a todas aquellas personas que, de alguna manera, me apoyaron para hacer posible este sueño.

Aunque sé que no hay palabras suficientes para expresar el infinito agradecimiento que tengo hacia ustedes por todo lo que me han dado, quiero mencionar a mis padres, porque a ustedes les debo todos mis logros. Sé que el camino aún es largo y faltan muchas metas por cumplir y sueños que alcanzar pero, gracias a ustedes dos, sé que podré realizarlos. Gracias por todo el amor, el cuidado, la protección, los desvelos y los sacrificios que han tenido siempre para mí. Espero que sepan cuánto los amo y, ojalá, algún día pueda recompensarles todo lo que hacen por mí.

A mi hermana, Grecia, a quien desde pequeña elegí como mi mayor fuente de inspiración. Quiero agradecerte porque, a pesar de la distancia, siempre has buscado la manera de seguir estando a mi lado. Gracias por cuidarme, escucharme y quererme tanto, pero sobre todo, por las enseñanzas tan valiosas que me has dado. El ver cómo te esfuerzas todos los días para alcanzar tus metas y lo lejos que has llegado, ha sido un aliciente para mí a lo largo de mi carrera universitaria y para los planes que tengo a futuro. Eres la persona más excepcional que he conocido y, espero que con el logro que alcanzo el día de hoy, yo pueda enorgullecerte, como tú siempre lo has hecho conmigo.

A mi tía Cecy, porque siempre has creído en mí y, me has brindado tu apoyo y cariño incondicionalmente. Nunca dejaré de valorar tu presencia en mi vida. Gracias por todo lo bueno que me has dado desde que era una niña, me siento dichosa de saber que cuento con una persona tan paciente y amorosa como tú.

A mi abuela Pilar, por estar siempre en los momentos importantes de mi vida. Gracias por tu paciencia, por tus consejos y, por el amor y el apoyo que me has ofrecido. También, quiero recordar a mis tres abuelos, José, Aurora y Juan, que ya no están en este mundo pero que significaron mucho para mí. Estén donde estén, espero que me vean y disfruten de este momento tan importante de mi vida. Sus recuerdos aún continúan en mi corazón.

A Diego y Emiliano, por recordarme siempre el lado divertido de la vida. Gracias por ser mis amigos, mis cómplices y mis hermanos, por contagiarme de su alegría infinita y por todas las anécdotas inigualables que hemos vivido juntos. Espero ser siempre un buen ejemplo para ustedes.

A Enrique, por compartir conmigo la emoción de realizar este trabajo. Te agradezco por haber confiado en mí durante este proceso, por haber recorrido conmigo este camino y motivarme cada que lo necesitaba. Que esta sea la primera de varias metas que alcanzaremos juntos.

Al Dr. Roberto Giovanni Ramírez Chavarría, por su paciencia, apoyo y asesoría en la realización de esta tesis. Siempre le estaré agradecida por haber contribuido en mi formación como ingeniera, por darme la oportunidad de trabajar con usted en este proyecto y por los conocimientos transmitidos durante las horas de clase.

A los miembros del jurado, por el tiempo que dedicaron para revisar y proporcionarme sus valiosas contribuciones al trabajo final. Además, por empeñarse en formar el perfil profesional de tantos estudiantes de la Facultad de Ingeniería, incluyendo el mío, al brindarnos con tanta vocación sus conocimientos y sabiduría.

A mis amigos, quienes formaron parte de esta aventura, porque cada uno de ustedes son una persona muy especial para mí. Gracias por todos los buenos momentos que vivimos juntos en la universidad, la vida me premió con su valiosa amistad y siempre quedarán en mis recuerdos.

Finalmente, me gustaría agradecer a la Universidad Nacional Autónoma de México, con la cual estaré eternamente agradecida por mi formación académica, profesional y personal. Gracias por otorgarme tantas satisfacciones como estudiante y persona, llevaré con orgullo tu nombre a todos lados.

Esta tesis fue realizada con apoyo del Programa UNAM-PAPIIT TA100221.

Índice general

1	Resumen	1
2	Introducción	3
2.1	Antecedentes y motivación	3
2.2	Planteamiento del problema	7
2.3	Contribución	8
2.4	Estructura de la tesis	8
3	Preliminares	10
3.1	Sistemas de aprendizaje automático	10
3.2	El ciclo del aprendizaje automático	14
3.3	Conjuntos de datos	15
3.4	Tipos de aprendizaje	17
3.4.1	Aprendizaje supervisado	18
3.4.2	Aprendizaje no supervisado	19
3.4.3	Aprendizaje semi-supervisado	21
3.4.4	Aprendizaje por refuerzo	22
3.5	Algoritmos de aprendizaje supervisado	23
3.5.1	Regresión lineal	23
3.5.2	Regresión logística	26

3.5.3	Árboles de decisión	28
3.5.4	Máquinas de vectores de soporte	31
3.5.5	Redes neuronales artificiales	34
4	Propiedades PVT de fluidos petroleros	41
4.1	Definición de las propiedades PVT de un aceite	41
4.1.1	Densidad del aceite ρ_o	41
4.1.2	Presión de burbuja P_b	42
4.1.3	Densidad relativa del aceite ρ_{ro}	42
4.1.4	Relación de solubilidad gas-aceite R_s	44
4.1.5	Factor de volumen del aceite B_o	45
4.1.6	Viscosidad del aceite μ_o	47
4.2	Clasificación de fluidos petroleros	49
5	Clasificación de fluidos petroleros	53
5.1	Regresión logística	56
5.2	Árboles de decisión	57
5.3	Redes neuronales artificiales	59
6	Estimación de propiedades de fluidos petroleros	62
6.1	Preparación de salidas de la región saturada	64
6.2	Preparación de salidas de la región bajo saturada	65
6.3	Estimación de P_b	67
6.3.1	Preparación de los datos para P_b	67
6.3.2	Estimación de P_b mediante regresión lineal	68
6.3.3	Estimación de P_b mediante SVR	69
6.3.4	Estimación de P_b mediante redes neuronales	69
6.3.5	Comparación de los tres métodos	70
6.4	Estimación de B_o en la región saturada	73
6.4.1	Preparación de los datos para B_o <i>sat</i>	73
6.4.2	Estimación de B_o en la región saturada mediante regresión lineal . .	74

6.4.3	Estimación de B_o en la región saturada mediante SVR	78
6.4.4	Estimación de B_o en región saturada mediante redes neuronales artificiales	79
6.4.5	Comparación de los tres métodos	80
6.5	Estimación de B_o en la región bajo saturada	81
6.5.1	Preparación de los datos para B_o <i>bajosat</i>	81
6.5.2	Estimación de B_o región bajo saturada mediante regresión lineal . .	82
6.5.3	Estimación de B_o región bajo saturada mediante SVR	83
6.5.4	Estimación de B_o <i>bajosat</i> mediante redes neuronales	83
6.5.5	Comparación de los tres métodos	84
6.5.6	Generación de la curva completa de B_o	85
6.6	Estimación de la ρ_{ro} en la región saturada	89
6.6.1	Preparación de los datos para estimar ρ_{rosat}	89
6.6.2	Estimación de ρ_{ro} en la región saturada mediante regresión lineal .	90
6.6.3	Estimación de ρ_{ro} en la región saturada mediante SVR	96
6.6.4	Estimación de ρ_o región saturada mediante redes neuronales artificiales	96
6.6.5	Comparación de los tres métodos	97
6.7	Estimación de ρ_{ro} en la región bajo saturada	98
6.7.1	Preparación de los datos para ρ_{ro} <i>bajosat</i>	98
6.7.2	Estimación de ρ_{ro} región bajo saturada mediante regresión lineal . .	99
6.7.3	Estimación de ρ_{ro} región bajo saturada mediante SVR	100
6.7.4	Estimación de ρ_{ro} <i>bajosat</i> mediante redes neuronales	100
6.7.5	Comparación de los tres métodos	101
6.7.6	Generación de la curva completa de ρ_{ro}	102
6.8	Estimación de R_s	106
6.8.1	Preparación de los datos para R_s	106
6.8.2	Estimación de R_s mediante regresión lineal	107
6.8.3	Estimación de R_s mediante SVR	110
6.8.4	Estimación de R_s mediante redes neuronales artificiales	110
6.8.5	Comparación de los tres métodos	111

6.8.6	Generación de la curva de R_s	112
6.9	Estimación de μ_o en la región saturada	116
6.9.1	Preparación de los datos para $\mu_o sat$	116
6.9.2	Estimación de μ_o en la región saturada mediante regresión lineal . .	117
6.9.3	Estimación de μ_o en la región saturada mediante SVR	123
6.9.4	Estimación de μ_o en la región saturada mediante redes neuronales artificiales	123
6.9.5	Comparación de los tres métodos	124
6.10	Estimación de μ_o en la región bajo saturada	125
6.10.1	Preparación de los datos para $\mu_o bajosat$	125
6.10.2	Estimación de μ_o región bajo saturada mediante regresión lineal . .	126
6.10.3	Estimación de μ_o región bajo saturada mediante SVR	127
6.10.4	Estimación de $\mu_o bajosat$ mediante redes neuronales	127
6.10.5	Comparación de los tres métodos	128
6.10.6	Generación de la curva completa de μ_o	129
7	Conclusiones y trabajo futuro	134
	Anexos	137
A	Comparaciones de resultados con correlaciones	138
A.1	Comparación de B_o	139
A.2	Comparación de R_s	143
A.3	Comparación de μ_o	145

Índice de figuras

3.1	Diagrama de bloques del funcionamiento de un sistema de aprendizaje automático	11
3.2	Aprendizaje supervisado y no supervisado	12
3.3	Aprendizaje por refuerzo	12
3.4	Aprendizaje en línea	13
3.5	Aprendizaje por lotes	13
3.6	Esquema del funcionamiento del aprendizaje supervisado	18
3.7	Esquema del funcionamiento del aprendizaje no supervisado	20
3.8	Esquema del funcionamiento del aprendizaje semi-supervisado	21
3.9	Esquema del funcionamiento del aprendizaje por refuerzo	22
3.10	Regresión lineal	24
3.11	Regresión logística	26
3.12	Árboles de decisión	28
3.13	Máquinas de vectores de soporte	31
3.14	Estructura de una red neuronal artificial	34
3.15	Esquema del funcionamiento del algoritmo de retropropagación	37
4.1	Comportamiento de la densidad relativa del aceite (ρ_{ro}) vs Presión	43
4.2	Comportamiento de la relación gas-aceite (R_s) vs Presión	44
4.3	Comportamiento del factor de volumen del aceite (B_o) contra presión	46
4.4	Comportamiento de la viscosidad del aceite (μ_o) vs Presión	48
4.5	Diagrama de fases generado para una mezcla de hidrocarburos.	50
4.6	Ubicación de los tipos de yacimiento en un diagrama de fases.	51
5.1	Mapa de correlación para elegir las variables de clasificación.	55

5.2	Árbol de decisión generado para clasificar aceites petroleros.	58
5.3	Red neuronal obtenida para clasificar aceites petroleros.	59
6.1	Mapa de correlación para elegir las variables de regresión.	63
6.2	Ejemplo de la gráfica a generar para cada propiedad.	64
6.3	Red neuronal generada para estimar el valor de P_b	70
6.4	Gráfica de resultados de P_b obtenidos con los modelos de aprendizaje automático.	72
6.5	Red neuronal generada para estimar el valor de $B_o sat$	79
6.6	Red neuronal generada para estimar el valor de $(B_o @(P_b + 200) - B_o @P_b[kg/cm^2])$	84
6.7	Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo A.	86
6.8	Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo B.	87
6.9	Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo C.	87
6.10	Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo D.	88
6.11	Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo E.	88
6.12	Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo F.	89
6.13	Red neuronal generada para estimar el valor de $\rho_{ro sat}$	97
6.14	Red neuronal generada para estimar el valor de $(\rho_{ro} @(P_b + 200) - \rho_{ro} @P_b[kg/cm^2])$	101
6.15	Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo A.	103
6.16	Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo B.	103

6.17 Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo C. 104

6.18 Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo D. 104

6.19 Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo E. 105

6.20 Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo F. 105

6.21 Red neuronal generada para estimar los valores de R_s 111

6.22 Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo A. 113

6.23 Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo B. 114

6.24 Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo C. 114

6.25 Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo D. 115

6.26 Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo E. 115

6.27 Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo F. 116

6.28 Red neuronal generada para estimar el valor de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$ 128

6.29 Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo A. 130

6.30 Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo B. 130

6.31 Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo C. 131

6.32 Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo D.	131
6.33 Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo E.	132
6.34 Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo F.	132
A.1 Curvas de B_o Pozo A	140
A.2 Curvas de B_o Pozo B	140
A.3 Curvas de B_o Pozo C	140
A.4 Curvas de B_o Pozo D	140
A.5 Curvas de B_o Pozo E	141
A.6 Curvas de B_o Pozo F	141
A.7 Curvas de ρ_{ro} Pozo A	142
A.8 Curvas de ρ_{ro} Pozo B	142
A.9 Curvas de ρ_{ro} Pozo C	142
A.10 Curvas de ρ_{ro} Pozo D	142
A.11 Curvas de ρ_{ro} Pozo E	143
A.12 Curvas de ρ_{ro} Pozo F	143
A.13 Curvas de R_s Pozo A	144
A.14 Curvas de R_s Pozo B	144
A.15 Curvas de R_s Pozo C	144
A.16 Curvas de R_s Pozo D	144
A.17 Curvas de R_s Pozo E	145
A.18 Curvas de R_s Pozo F	145
A.19 Curvas de μ_o Pozo A	146
A.20 Curvas de μ_o Pozo B	146
A.21 Curvas de μ_o Pozo C	146
A.22 Curvas de μ_o Pozo D	146
A.23 Curvas de μ_o Pozo E	147

A.24 Curvas de μ_o Pozo F 147

Índice de tablas

3.1	Funciones de activación para redes neuronales más comunes	39
5.1	Clasificación de Méndez para los tipos de aceites.	54
5.2	Número de muestras utilizadas para entrenar y validar los modelos de clasificación.	54
5.3	Rango de valores de las propiedades del conjunto de datos.	54
5.4	Coefficientes de la regresión logística.	56
5.5	Matriz de confusión de la clasificación mediante regresión logística.	57
5.6	Resultados del algoritmo de árbol de decisión	58
5.7	Matriz de confusión de la clasificación mediante árboles de decisión.	59
5.8	Matriz de confusión de la clasificación mediante redes neuronales artificiales.	60
5.9	Comparación de los resultados obtenidos con los modelos de clasificación entrenados.	60
6.1	Ejemplo de salidas recolectadas para la región saturada.	65
6.2	Ejemplo de salidas recolectadas para la región bajo saturada de la curva de B_o	66
6.3	Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de P_b	67
6.4	Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de P_b y normalizar las variables de entrada durante la etapa de procesamiento de datos.	67
6.5	Coefficientes de la regresión lineal para estimación de P_b	68
6.6	Indicadores de algoritmo de regresión lineal para estimación de P_b	69
6.7	Indicadores del algoritmo SVR para estimación de P_b	69

6.8 Indicadores del algoritmo de Redes Neuronales para estimación de P_b 70

6.9 Indicadores del algoritmo de Redes Neuronales para estimación de P_b 70

6.10 Conjunto de datos de prueba para P_b 71

6.11 Valores estimados de P_b con los tres modelos de aprendizaje automático 71

6.12 Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de $B_o\ sat$ 73

6.13 Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de $B_o\ sat$ y normalizar las variables de entrada durante la etapa de procesamiento de datos. 73

6.14 Coeficientes de la regresión lineal para estimación de $B_o\ sat$ en $\frac{1}{5}P_b$ 74

6.15 Indicadores del algoritmo de regresión lineal para estimación de $B_o\ sat$ en $\frac{1}{5}P_b$. 74

6.16 Coeficientes de la regresión lineal para estimación de $B_o\ sat$ en $\frac{2}{5}P_b$ 75

6.17 Indicadores del algoritmo de regresión lineal para estimación de $B_o\ sat$ en $\frac{2}{5}P_b$. 75

6.18 Coeficientes de la regresión lineal para estimación de $B_o\ sat$ en $\frac{3}{5}P_b$ 76

6.19 Indicadores del algoritmo de regresión lineal para estimación de $B_o\ sat$ en $\frac{3}{5}P_b$. 76

6.20 Coeficientes de la regresión lineal para estimación de $B_o\ sat$ en $\frac{4}{5}P_b$ 77

6.21 Indicadores del algoritmo de regresión lineal para estimación de $B_o\ sat$ en $\frac{4}{5}P_b$. 77

6.22 Indicadores del algoritmo de regresión lineal para estimación de $B_o\ sat$ en P_b . 77

6.22 Coeficientes de la regresión lineal para estimación de $B_o\ sat$ en P_b 78

6.24 Indicadores del algoritmo de SVR para estimación de $B_o\ sat$ 78

6.25 Indicadores del algoritmo de Redes Neuronales para estimación de $B_o\ sat$ 80

6.26 Indicadores del algoritmo de Redes Neuronales para estimación de $B_o\ sat$ 80

6.27 Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de $B_o\ bajosat$ 81

6.28 Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de $B_o\ bajosat$ y normalizar las variables de entrada durante la etapa de procesamiento de datos. 81

6.29 Coeficientes de la regresión lineal para estimación de $(B_o\ @(P_b + 200) - B_o\ @P_b[kg/cm^2])$ 82

6.30 Indicadores del algoritmo de regresión lineal para estimación de $B_o\ sat$ en $\frac{1}{5}P_b$. 82

6.31 Indicadores del algoritmo de SVR para estimación de $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])$ 83

6.32 Indicadores del algoritmo de Redes Neuronales para estimación de $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])$ 84

6.33 Errores en la estimación de B_{oy} 85

6.34 Comparación de resultados obtenidos con el set de datos de prueba para B_{oy} 85

6.35 Conjunto de datos de prueba para B_o 86

6.36 Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de ρ_{ro} 90

6.37 Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de ρ_{ro} y normalizar las variables de entrada durante la etapa de procesamiento de datos. 90

6.38 Coeficientes de la regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{1}{5}P_b$ 91

6.39 Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{1}{5}P_b$ 91

6.40 Coeficientes de la regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{2}{5}P_b$ 92

6.41 Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{2}{5}P_b$ 92

6.42 Coeficientes de la regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{3}{5}P_b$ 93

6.43 Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{3}{5}P_b$ 93

6.44 Coeficientes de la regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{4}{5}P_b$ 94

6.45 Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{4}{5}P_b$ 94

6.46 Coeficientes de la regresión lineal para estimación de $\rho_{ro sat}$ en P_b 95

6.47 Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro sat}$ en P_b . 95

6.48 Indicadores del algoritmo de SVR para estimación de $\rho_{ro sat}$ 96

6.49 Indicadores del algoritmo de Redes Neuronales para estimación de $\rho_o sat$. . 97

6.50 Indicadores del algoritmo de Redes Neuronales para estimación de $\rho_{ro sat}$. . 98

6.51 Número de muestras del conjunto de datos utilizado para entrenar y probar los modelos de estimación de $\rho_{ro \text{ bajosat}}$ 99

6.52 Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de $\rho_{ro \text{ bajosat}}$ y normalizar las variables de entrada durante la etapa de procesamiento de datos. 99

6.53 Coeficientes de la regresión lineal para estimación de $(\rho_{ro} @(P_b + 200) - \rho_{ro} @P_b[kg/cm^2])$ 100

6.54 Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro \text{ sat}}$ en $\frac{1}{5}P_b$ 100

6.55 Indicadores del algoritmo de SVR para estimación de $(\rho_{ro} @(P_b + 200) - \rho_{ro} @P_b[kg/cm^2])$ 100

6.56 Indicadores del algoritmo de Redes Neuronales para estimación de $(\rho_{ro} @(P_b + 200) - \rho_{ro} @P_b[kg/cm^2])$ 101

6.57 Indicadores del algoritmo de Redes Neuronales para estimación de $\rho_{ro \text{ sat}}$ 102

6.58 Comparación de resultados obtenidos con el conjunto de datos de prueba para ρ_{roy} 102

6.59 Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de R_s 106

6.60 Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de R_s y normalizar las variables de entrada durante la etapa de procesamiento de datos. 106

6.61 Coeficientes de la regresión lineal para estimación de R_s en $\frac{1}{5}P_b$ 107

6.62 Indicadores del algoritmo de regresión lineal para estimación de R_s en $\frac{1}{5}P_b$ 107

6.63 Coeficientes de la regresión lineal para estimación de R_s en $\frac{2}{5}P_b$ 108

6.64 Indicadores del algoritmo de regresión lineal para estimación de R_s en $\frac{2}{5}P_b$ 108

6.65 Coeficientes de la regresión lineal para estimación de R_s en $\frac{3}{5}P_b$ 109

6.66 Indicadores del algoritmo de regresión lineal para estimación de R_s en $\frac{3}{5}P_b$ 109

6.67 Coeficientes de la regresión lineal para estimación de R_s en $\frac{4}{5}P_b$ 110

6.68 Indicadores del algoritmo de regresión lineal para estimación de R_s en $\frac{4}{5}P_b$ 110

6.69 Indicadores del algoritmo de SVR para estimación de R_s 111

6.70 Indicadores del algoritmo de Redes Neuronales para estimación de R_S . . . 112

6.71 Error absoluto medio obtenido con los tres algoritmos utilizados para calcular R_S 112

6.72 Conjunto de datos de prueba para R_s 113

6.73 Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de $\mu_o sat$ 117

6.74 Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de $\mu_o sat$ y normalizar las variables de entrada durante la etapa de procesamiento de datos. 117

6.75 Coeficientes de la regresión lineal para estimación de $\mu_o sat$ en $\frac{1}{5}P_b$ 118

6.76 Indicadores del algoritmo de regresión lineal para estimación de $\mu_o sat$ en $\frac{1}{5}P_b$.118

6.77 Coeficientes de la regresión lineal para estimación de $\mu_o sat$ en $\frac{2}{5}P_b$ 119

6.78 Indicadores del algoritmo de regresión lineal para estimación de $\mu_o sat$ en $\frac{2}{5}P_b$.119

6.79 Coeficientes de la regresión lineal para estimación de $\mu_o sat$ en $\frac{3}{5}P_b$ 120

6.80 Indicadores del algoritmo de regresión lineal para estimación de $\mu_o sat$ en $\frac{3}{5}P_b$.120

6.81 Coeficientes de la regresión lineal para estimación de $\mu_o sat$ en $\frac{4}{5}P_b$ 121

6.82 Indicadores del algoritmo de regresión lineal para estimación de $\mu_o sat$ en $\frac{4}{5}P_b$.121

6.83 Coeficientes de la regresión lineal para estimación de $\mu_o sat$ en P_b 122

6.84 Indicadores del algoritmo de regresión lineal para estimación de $\mu_o sat$ en P_b . 122

6.85 Indicadores del algoritmo de SVR para estimación de $\mu_o sat$ 123

6.86 Indicadores del algoritmo de Redes Neuronales para estimación de $\mu_o sat$. . 124

6.87 Indicadores del algoritmo de Redes Neuronales para estimación de $\mu_o sat$. . 124

6.88 Número de muestras del conjunto de datos utilizado para entrenar y probar los modelos de estimación de $\mu_o bajosat$ 125

6.89 Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de $\mu_o bajosat$ y normalizar las variables de entrada durante la etapa de procesamiento de datos. 125

6.90 Coeficientes de la regresión lineal para estimación de $(\mu_o @(P_b + 200) - \mu_o @P_b[kg/cm^2])$ 126

6.91 Indicadores del algoritmo de regresión lineal para estimación de $\mu_o sat$ en $\frac{1}{5}P_b$.126

6.92 Indicadores del algoritmo de SVR para estimación de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$ 127

6.93 Indicadores del algoritmo de Redes Neuronales para estimación de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$ 128

6.94 Indicadores del algoritmo de Redes Neuronales para estimación de $\mu_o sat$ 129

6.95 Comparación de resultados obtenidos con el conjunto de datos de prueba para μ_{oy} 129

7.1 Modelos con menor porcentaje de E_a para cada propiedad PVT. 134

A.1 Resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos utilizando el conjunto de datos de validación 139

A.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. 139

A.3 Comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar B_o en el conjunto de validación. 140

A.4 Comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar ρ_{ro} en el conjunto de validación. 142

A.5 Comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar R_s en el conjunto de validación. 144

A.6 Comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar μ_o en el conjunto de validación. 146

Capítulo 1

Resumen

Conocer el comportamiento de fluidos petroleros es de gran importancia en la industria petrolera, específicamente para Ingeniería de Yacimientos e Ingeniería de Producción. Los cálculos de balance de materia, análisis de pruebas de presión, estimación 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 como presión, volumen y temperatura (PVT) permiten conocer las características físicas de los fluidos. Particularmente, en el caso de los fluidos petroleros, dichas propiedades permiten realizar un análisis exhaustivo del yacimiento petrolero que contiene a dicho fluido. Las propiedades PVT pueden ser determinadas de diferentes maneras, siendo las más exactas las pruebas de laboratorio, cuya desventaja es su alto costo debido a los equipos requeridos para realizarlas.

Actualmente, el desarrollo de técnicas de aprendizaje automático ha atraído el interés de científicos y tecnólogos debido a su capacidad de trabajar con datos y a su versatilidad para estimar parámetros, inferir comportamientos y modelar procesos. El aprendizaje automático es una rama de la inteligencia artificial que permite que las máquinas “aprendan”, 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 de tesis, se propone el desarrollo y puesta en funcionamiento de algoritmos de aprendizaje automático con el apoyo de herramientas de código abierto, como R y Python, para clasificar y modelar datos de fluidos petroleros, específicamente aceites negros y volátiles, a partir de parámetros que no requieren de pruebas especializadas para poder conocerse. La información necesaria para entrenar y validar los modelos será obtenida de reportes de pruebas PVT de una región petrolera de la República Mexicana.

El objetivo, es obtener con los modelos creados, las curvas que describen las diferentes propiedades PVT de los aceites conforme varía la presión a condiciones de yacimiento con la mayor precisión posible, así como clasificar a dicho fluido como negro o volátil.

Capítulo 2

Introducción

Esta parte del trabajo habla acerca de la relevancia de la determinación de las propiedades de presión, volumen y temperatura de los fluidos en la Ingeniería Petrolera y de los métodos empleados actualmente para realizar esta tarea. Además, se presenta una exhaustiva revisión al estado del arte sobre la aplicación de técnicas de aprendizaje automático para la estimación de dichas propiedades.

2.1. Antecedentes y motivación

Dindoruk y Christman (2001) definen las propiedades de presión, volumen y temperatura (PVT) de los aceites petroleros como una serie de propiedades físicas de un fluido en el yacimiento (petróleo, agua o gas) que relacionan presión, volumen y temperatura. Estas propiedades son necesarias siempre en los estudios de yacimientos, desde los cálculos de balance de materia hasta la simulación, para evaluar su desempeño y diseñar las instalaciones subterráneas y superficiales necesarias. Debido a esto, la determinación de las propiedades PVT es un factor clave para maximizar las ganancias en la explotación de un yacimiento petrolero.

Idealmente, deberían utilizarse datos PVT medidos en un laboratorio. Sin embargo, muchas veces estos datos no están disponibles y la realización de las pruebas de laborato-

rio pueden llegar a ser muy costosas por lo que, deben utilizarse otros métodos en su lugar.

Muchos investigadores han utilizado los resultados PVT de diversas pruebas de laboratorio y datos de campo para desarrollar correlaciones generalizadas para estimar las propiedades PVT de los fluidos de un yacimiento. Si bien, las correlaciones generalmente coinciden con los datos experimentales con una desviación promedio de menos de un pequeño porcentaje, puede llegar a ser difícil determinar qué correlación utilizar debido a que están diseñadas a partir de muestras regionales determinadas, lo cual puede llevar a obtener resultados con desviaciones con un orden de magnitud muy grande.

En años recientes, el aprendizaje automático ha llamado la atención dentro de diversas ramas de la Ingeniería Petrolera. Debido a ello, recientemente se ha comenzado a estudiar la estimación de las propiedades PVT de los fluidos petroleros por medio de técnicas de aprendizaje automático, sin embargo, es un tema que apenas se ha explorado vagamente. Desde hace más de dos décadas se han implementado diversos algoritmos para realizar estimaciones por diferentes métodos como redes neuronales artificiales (*ANN: Artificial Neural Networks*), máquinas de vectores de soporte (*SVM: Support Vector Machines*), lógica difusa e incluso sistemas híbridos.

Gharbi *et al.* (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 P_b con un porcentaje de error absoluto medio (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 *et al.* (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 (*American Petroleum Institute*) 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.

Al-Marhourn y Osman (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 de error absoluto medio de 0.52% con un coeficiente de correlación de 0.999, mejorando notablemente los resultados obtenidos anteriormente (Osman *et al.* (2001)).

Nagi *et al.* (2009) exploraron otras técnicas de aprendizaje automático para investigar la capacidad de las máquinas de vectores de soporte para modelar las propiedades PVT de los sistemas de petróleo crudo debido a los inconvenientes de las redes neuronales artificiales, las cuales, no operan con precisión ya que funcionan únicamente para un cierto rango de características del fluido del yacimiento y área geográfica 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 *et al.* (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 error 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 *et al.* (2015) propusieron el uso de 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 Selamat *et al.* (2012), llegaron a la conclusión de que la lógica difusa superaba a las redes neuronales en cuanto a precisión en la resolución de este problema.

A partir de información básica de pozos de una región de México, Camargo (2016) realizó la estimación de propiedades PVT para yacimientos de aceite negro y volátil mediante el uso de redes neuronal artificiales. Se obtuvieron con resultados satisfactorios la presión de burbuja (P_b), relación de solubilidad gas – aceite (R_{sb}), compresibilidad del aceite (C_{ob}) y factor de volumen del aceite a la presión de saturación (B_{ob}).

Oloso *et al.* (2017) 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.

Ramirez *et al.* (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: Principal Component Analysis*) 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.

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 logró reconstruir de forma satisfactoria la curva del B_o

contra presión para varias muestras de aceite, además se estimó con un E_a de 5.48 % la P_b de dichas muestras.

2.2. Planteamiento del problema

Como puede verse, en los trabajos mencionados anteriormente se estiman propiedades PVT en valores de presión específicos como en el punto de burbuja o en la presión del yacimiento. Estos valores son de gran utilidad para realizar cálculos de volúmenes de hidrocarburos en sitio, lo cual permite determinar la rentabilidad de un proyecto petrolero. Sin embargo, para la toma de decisiones más complicadas como el diseño de las instalaciones a utilizar para explotar un yacimiento, es necesario saber cómo varían las propiedades PVT respecto a la presión, por lo que se requiere conocer la curva completa de dichas propiedades.

Conocer el comportamiento de los fluidos petroleros es importante en 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.

Además, la clasificación correcta de los fluidos de los yacimientos petroleros es esencial en la Ingeniería Petrolera para determinar el proceso de administración y explotación del yacimiento.

2.3. Contribución

El presente trabajo propone el uso de técnicas de aprendizaje automático para obtener diferentes modelos que permitan clasificar los aceites petroleros de una región de la República Mexicana y estimar de forma fiable las curvas de sus propiedades PVT como la relación de solubilidad gas - aceite (R_s), el factor de volumen del aceite (B_o), la densidad (ρ_o) y la viscosidad (μ_o), así como el punto de burbuja (P_b).

Para ello, se plantea el uso de algoritmos como regresión logística, árboles de decisión y redes neuronales, para clasificar aceites a partir de propiedades físicas fáciles de obtener como la temperatura del yacimiento, densidad relativa del aceite y valor de la relación de solubilidad gas - aceite en el punto de burbuja. Posteriormente, a partir de las propiedades físicas mencionadas anteriormente, se analiza el uso de algoritmos como regresión lineal, máquinas de vectores de soporte y redes neuronales para obtener el punto de burbuja, y diferentes puntos de las curvas de las propiedades PVT de los aceites en sus partes saturadas y bajo saturadas para, finalmente, construir dichas curvas a partir de los resultados obtenidos mediante el método de ajuste polinomial por mínimos cuadrados.

2.4. Estructura de la tesis

En este primer capítulo se habló sobre la relevancia de estudiar las propiedades PVT en la Ingeniería Petrolera, y se revisó el estado del arte sobre la aplicación de técnicas de aprendizaje automático para realizar esta tarea como una alternativa a las pruebas clásicas de laboratorio.

En el segundo capítulo se da una breve introducción al aprendizaje automático y al método que debe seguirse para crear un sistema capaz de aprender desde cero. Además, se explican los algoritmos de clasificación y regresión utilizados para el desarrollo de esta tesis.

En el tercer capítulo, se definen cada una de las propiedades PVT analizadas en fluidos petroleros y, se describe el método de clasificación de aceites utilizado actualmente por los ingenieros petroleros.

Posteriormente, en el capítulo cuatro se muestra el método seguido para clasificar fluidos petroleros en negros y volátiles a partir de técnicas de aprendizaje automático, así como los resultados obtenidos por cada uno de los algoritmos utilizados.

Luego, en el capítulo cinco se aborda el desarrollo de modelos de regresión para calcular diferentes puntos de las curvas de las propiedades PVT de los fluidos petroleros en sus partes saturadas y bajo saturadas. Además, se muestran los resultados obtenidos por cada uno de los algoritmos empleados al construir las curvas PVT con el método de ajuste polinomial por mínimos cuadrados a partir de los puntos calculados.

Finalmente, en el capítulo seis, se analizan los resultados obtenidos en los dos capítulos anteriores para realizar las conclusiones finales y se presentan las alternativas de trabajo futuro para darle continuidad a este trabajo.

Capítulo 3

Preliminares

En este capítulo, se da una breve introducción al aprendizaje automático y las diferentes categorías de aprendizaje automático existentes: supervisado, no supervisado, semi-supervisado y reforzado, así como el proceso a seguir para crear un sistema de aprendizaje automático. Además, se explican específicamente los algoritmos de aprendizaje supervisado utilizados para resolver los problemas de clasificación y regresión planteados en este trabajo.

3.1. Sistemas de 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 computacionales que detectan de forma automática 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 cuáles movimientos constituían las estrategias ganadoras para utilizarlos en partidas futuras. Samuel (1959) explicó el enfoque del aprendizaje automático utilizado en un artículo publicado 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 automatizar sus sistemas, obteniendo una mayor eficiencia y rendimiento en sus procesos.

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, como puede verse en la Figura 3.1.

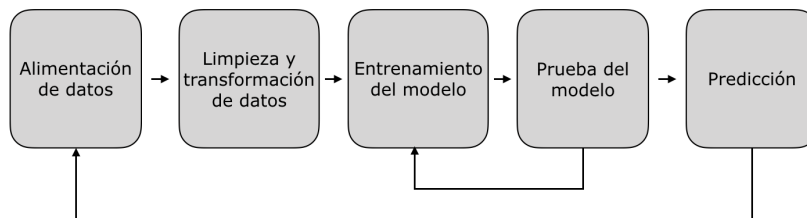


Figura 3.1: Diagrama de bloques del funcionamiento de un sistema de aprendizaje automático

Ayodele (2010) clasifica a los sistemas de aprendizaje automático dependiendo de su enfoque de aprendizaje en supervisados, no supervisados, semi-supervisados y por refuerzo. En la diferencia Figura 3.2 pueden observarse de manera gráfica las diferencias entre estos tipos de aprendizaje. 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, representadas por círculos y triángulos en la Figura 3.2,

mientras que en el aprendizaje no supervisado el algoritmo simplemente categoriza los datos en función de su estructura oculta. El aprendizaje semi-supervisado utiliza una combinación de datos etiquetados y no etiquetados para generar un clasificador o una función apropiada para el problema a resolver.

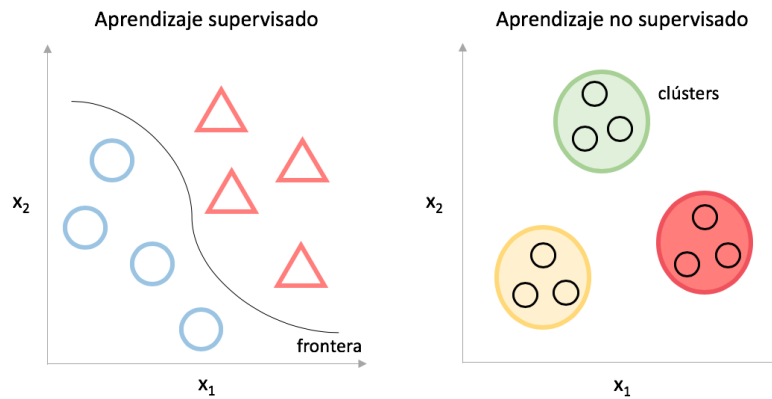


Figura 3.2: Aprendizaje supervisado y no supervisado

Por su parte, como puede observarse en la Figura 3.3, 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.

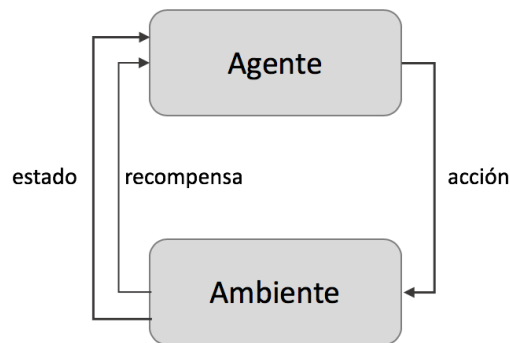


Figura 3.3: Aprendizaje por refuerzo

Otra forma de clasificar los sistemas de aprendizaje automático, según Dekel (2009), es con base en la forma en la que estos realizan el proceso de aprendizaje. En los sistemas de aprendizaje en línea, ilustrados en la Figura 3.4, (*online learning*) los modelos se entrenan

de forma incremental al alimentar instancias secuencialmente, ya sea individualmente o en grupos pequeños llamados mini-lotes. Por otro lado, en los sistemas de aprendizaje por lotes (*batch learning*), cuyo diagrama puede observarse en la Figura 3.5, los modelos son incapaces de aprender de forma incremental. Por ello, primero el modelo se entrena con todos los datos y luego es lanzado a producción.

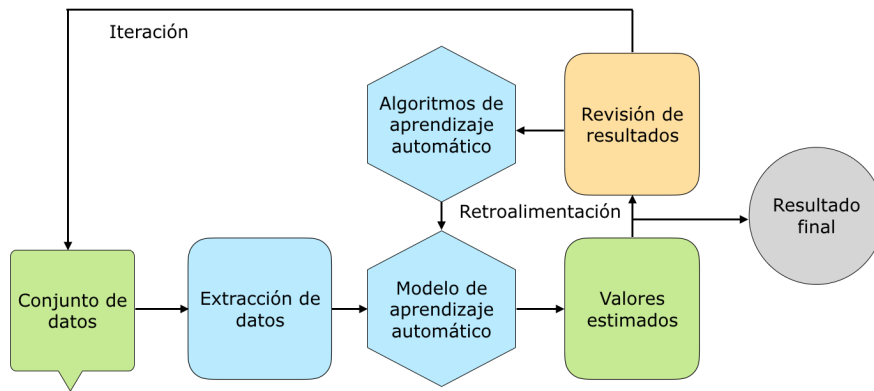


Figura 3.4: Aprendizaje en línea

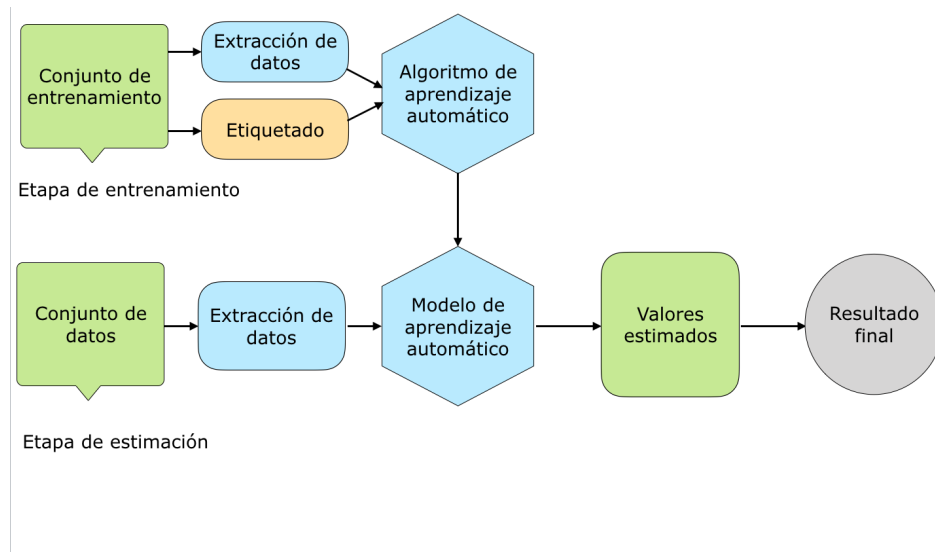


Figura 3.5: Aprendizaje por lotes

3.2. El ciclo del aprendizaje automático

El proceso de creación de un sistema de aprendizaje automático es iterativo ya que se obtiene nueva información cada día. Debido a ello, se debe mantener el modelo actualizado una vez que entra a producción.

Los pasos propuestos por Hurwitz y Kirsch (2018) para cumplir con el ciclo del aprendizaje automático, pueden resumirse en los siguientes ocho puntos:

1. **Identificación de las fuentes de datos:** este paso incluye la identificación de fuentes de datos 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 las predicciones:** la información recopilada del análisis de la validez de las nuevas predicciones es de utilidad para retroalimentar 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 un 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.

Por otro lado, 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.

La información con la que se trabajará puede ser obtenida de diversas fuentes que son clasificadas en dos grupos dependiendo del tipo de datos que almacenan. El primer grupo conforma a las fuentes de datos estructuradas que generalmente son bases de datos relacionales tradicionales que contienen datos con una longitud y formato definidos. La mayoría de las organizaciones tienen una gran cantidad de datos estructurados en sus centros de datos locales.

El segundo grupo hace referencia a las fuentes de datos no estructuradas. En ellas, se almacenan datos que no siguen un formato específico. Este tipo de datos incluyen

información generada en las plataformas de redes sociales, mensajes de texto, fotografías, videos, entre otros.

Una vez que se cuenta con el conjunto de datos, se debe considerar una etapa de pre-procesamiento que incluye las siguientes acciones:

- **Formato:** Si los datos están distribuidos en diferentes archivos, deben reunirse en uno solo para formar el conjunto de datos.
- **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.
- **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 rápidos y con bajo consumo de memoria.

Posteriormente, 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, dos investigadores de Microsoft, 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.

También, Alpaydin (2014) menciona que 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” y, al recibir datos nuevos, su desempeño es pobre. De forma contraria, 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 incluye 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. En la Figura 3.6 puede verse el procedimiento necesario para resolver un problema mediante aprendizaje supervisado.

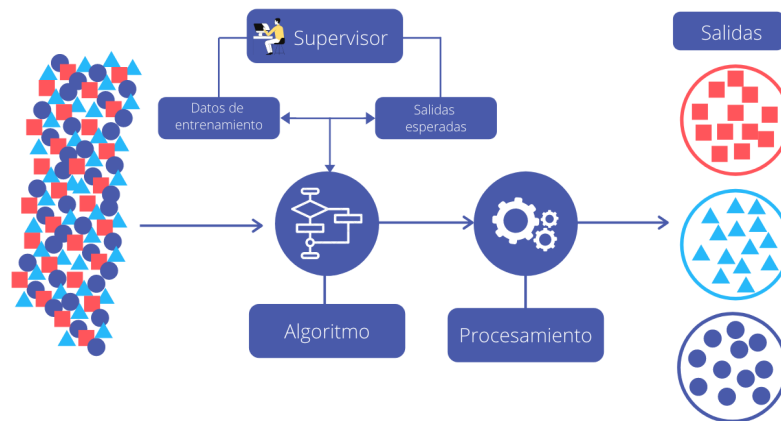


Figura 3.6: Esquema del funcionamiento del aprendizaje supervisado

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 discretos, 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.

Si el modelo obtenido es capaz de representar solamente los patrones existentes en el subconjunto de datos de entrenamiento, ocurre un problema de sobreajuste. En el caso del aprendizaje supervisado, para protegerse contra el sobreajuste, Ayodele (2010) propone realizar pruebas con datos etiquetados imprevistos o desconocidos para el modelo.

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. La Figura 3.7 muestra los pasos necesarios para entrenar un sistema mediante aprendizaje no supervisado.

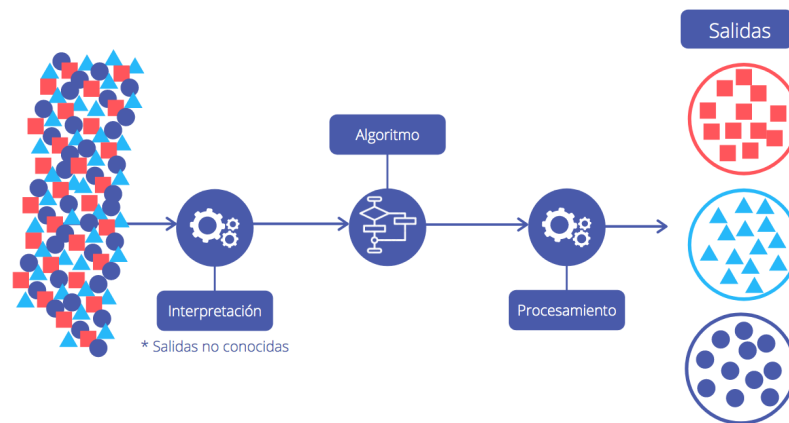


Figura 3.7: Esquema del funcionamiento del aprendizaje no supervisado

Ayodele (2010) describe dos enfoques en este tipo de aprendizaje. 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 agrupación (*clustering*). 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 grupos 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.

Los algoritmos de aprendizaje no supervisados, según Ghahramani (Ghahramani, 2008) 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

En la Figura 3.8, se puede observar que el aprendizaje semi-supervisado es una combinación de los dos enfoques de aprendizaje mencionados anteriormente pues, fue introducido para resolver los problemas encontrados en ambos. 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.

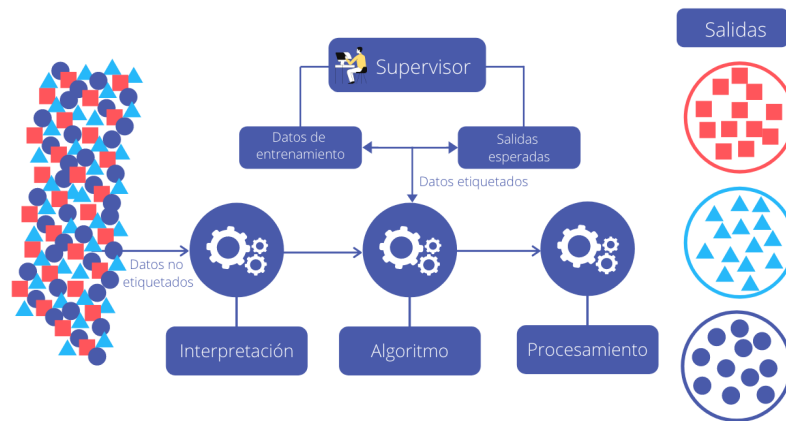


Figura 3.8: Esquema del funcionamiento del aprendizaje semi-supervisado

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. Chapelle *et al.* (2006) mencionan que 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.

El procedimiento básico, según Chapelle *et al.* (2006), que debe seguirse 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

El aprendizaje por refuerzo es descrito por Hurwitz y Kirsch (2018) como un modelo de aprendizaje conductual en el que se capacitan a los modelos (agentes) de aprendizaje automático para tomar una secuencia de decisiones. La Figura 3.9 muestra los elementos necesarios en un sistema de aprendizaje por refuerzo.

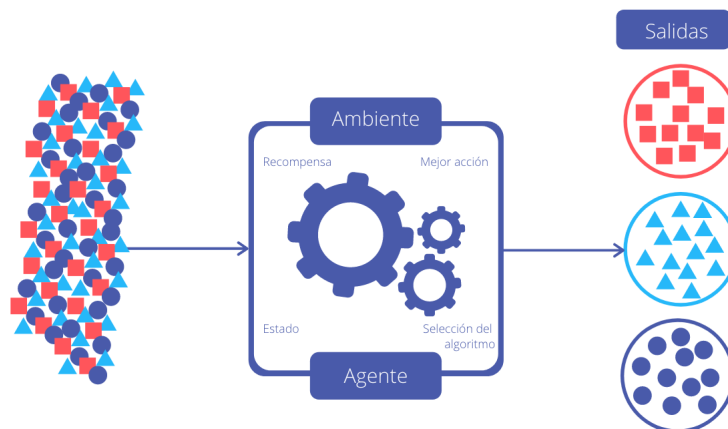


Figura 3.9: Esquema del funcionamiento del aprendizaje por refuerzo

Este tipo de aprendizaje difiere de los anteriores porque el sistema no está capacitado con el conjunto de datos de muestra, sino que 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, los cuales se describen a continuación.

3.5.1. Regresión lineal

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.

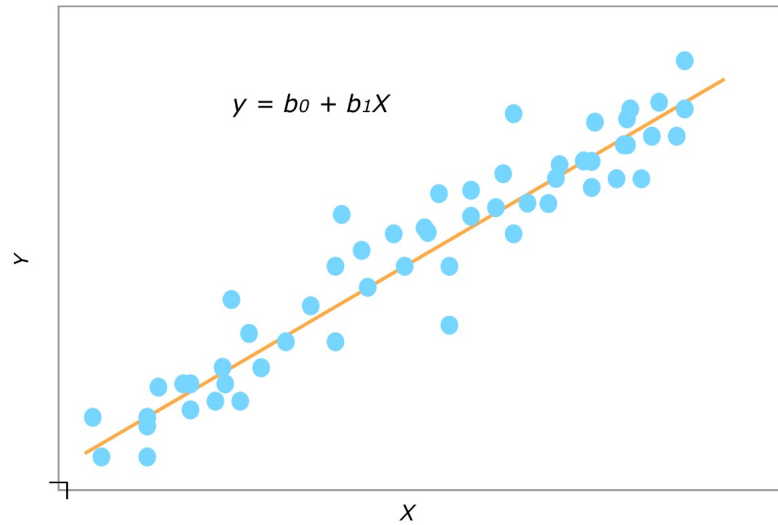


Figura 3.10: Regresión lineal

Como plantea Alpaydin (2014), el método de regresión lineal simple es un enfoque muy sencillo para predecir una respuesta cuantitativa Y sobre la base de una variable de predicción simple X , matemáticamente, la relación lineal entre X y Y está dada por

$$Y \approx \beta_0 + \beta_1 X, \quad (3.1)$$

donde β_0 y β_1 son dos constantes desconocidas que representan la intersección y la pendiente del modelo lineal, respectivamente y, se conocen como coeficientes o parámetros del modelo.

Una vez que los datos de entrenamiento son utilizados para producir valores estimados de los parámetros β_0 y β_1 . Es posible estimar nuevos valores a partir de entradas nuevas conocidas x haciendo uso del modelo obtenido mediante la ecuación $\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x$. De la ecuación anterior, \hat{y} indica una predicción de Y sobre la base de $X = x$. En la Figura 3.10 se puede apreciar que el objetivo de este algoritmo es encontrar las mejores estimaciones de los coeficientes para minimizar los errores en la predicción de \hat{y} a partir de x . En este caso, el símbolo de acento circunflejo es utilizado para denotar el valor estimado de un parámetro o coeficiente desconocido o, para denotar el valor predicho de la respuesta.

El enfoque más común para determinar los valores de estos parámetros es el criterio de mínimos cuadrados. Sea $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ la predicción para Y basada en el i -ésimo valor de X , entonces $e_i = y_i - \hat{y}_i$ 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: residual sum of squares*) es definida como

$$RSS = e_1^2 + e_2^2 + \dots + e_n^2. \quad (3.2)$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (3.3a)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \quad (3.3b)$$

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

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n, \quad (3.4)$$

en donde, los coeficientes de regresión $\beta_0, \beta_1, \dots, \beta_n$ son desconocidos y deben estimarse utilizando el mismo enfoque de mínimos cuadrados, es decir, deben elegirse valores de $\beta_0, \beta_1, \dots, \beta_n$ que minimicen la suma de los residuos al cuadrado. Los valores $\beta_0, \beta_1, \dots, \beta_n$

que minimizan la ecuación 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, como se muestra en la Figura 3.11.

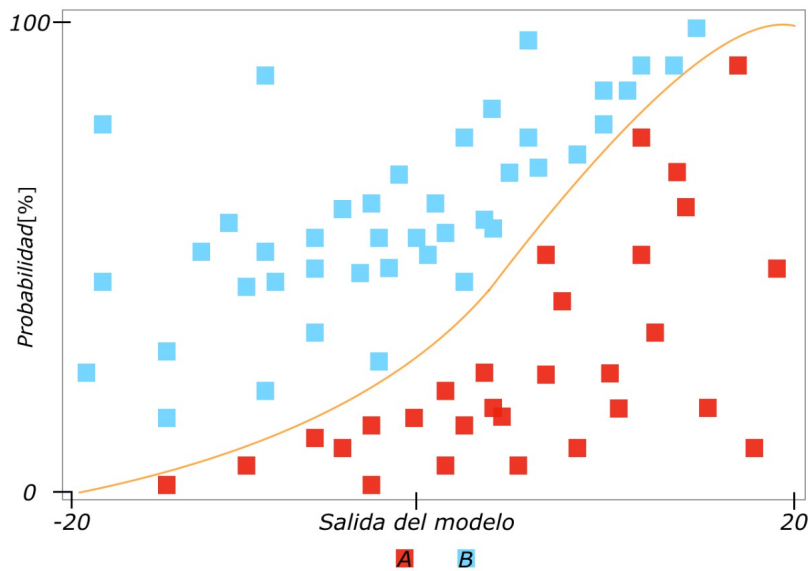


Figura 3.11: Regresión logística

Para ello, utiliza la misma estructura que la regresión lineal, pero transformando la variable respuesta Y en una probabilidad. En este caso, Smola y Vishwanathan (2008) explican que 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

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}, \quad (3.5)$$

donde, 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 $\hat{p}(x_i)$ 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:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})). \quad (3.6)$$

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 forma:

- Si la probabilidad $p(X)$ obtenida es igual o superior a 0.5, se le asigna la clase 1.
- Si la probabilidad $p(X)$ 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

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}}, \quad (3.7)$$

en donde $X = (X_1, \dots, X_n)$ son las n variables predictoras y, los parámetros β_1, \dots, β_n son desconocidos y deben estimarse mediante el método de máxima verosimilitud.

3.5.3. Árboles de decisión

Un árbol de decisión es definido por Alpaydin (2014) como 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.

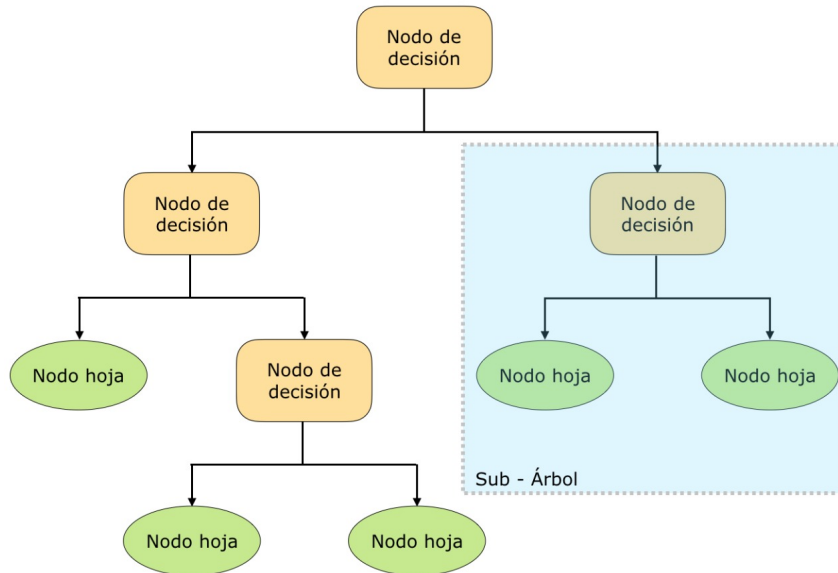


Figura 3.12: Árboles de decisión

Un árbol de decisión está compuesto por nodos de decisión internos y hojas terminales,

los cuales están representados en la Figura 3.12. 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, como plantea Shai y Shai (2014), 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.

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 $p_i (i = 1, \dots, m)$ 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

$$\sum_{i=1}^m -p_i \log_2(p_i), \quad (3.8)$$

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

$$I(S) = \sum_{v \in V(A)} \frac{|S_v|}{|S|} I(S_v). \quad (3.9)$$

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

De acuerdo con Shai y Shai (2014), una máquina de vectores de soporte (SVM: support vector machine) 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.

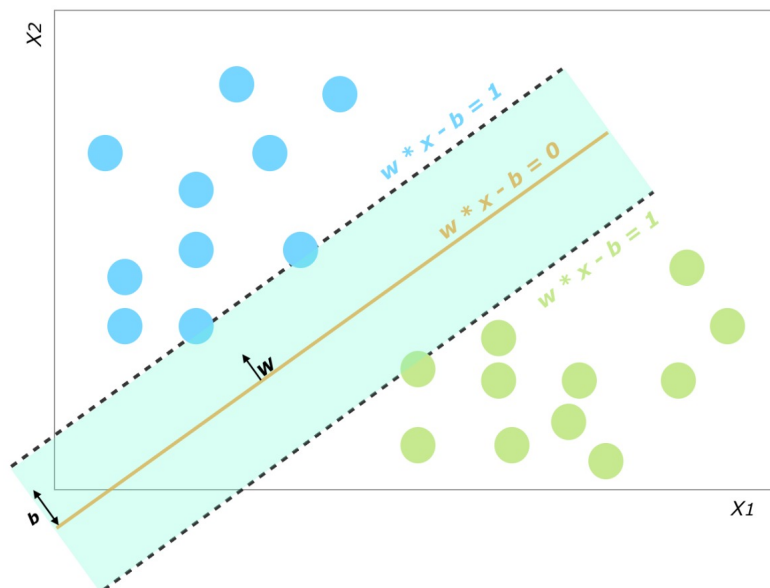


Figura 3.13: Máquinas de vectores de soporte

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, como se aprecia en la Figura 3.13.

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, ejemplo mostrado en la Figura 3.13, 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.

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 γ 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 γ 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. Entre los kernels más populares, según Alpaydin (2014), 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

$$K(x_i, x_{i'}) = \sum_{j=1}^p x_{ij}x_{i'j}. \quad (3.10)$$

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

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^p x_{ij}x_{i'j}\right)^d. \quad (3.11)$$

- **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 ecuación 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.

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^p (x_{ij} - x_{i'j})^2\right). \quad (3.12)$$

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

Una red neuronal artificial es descrita por Shai y Shai (2014) como 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.

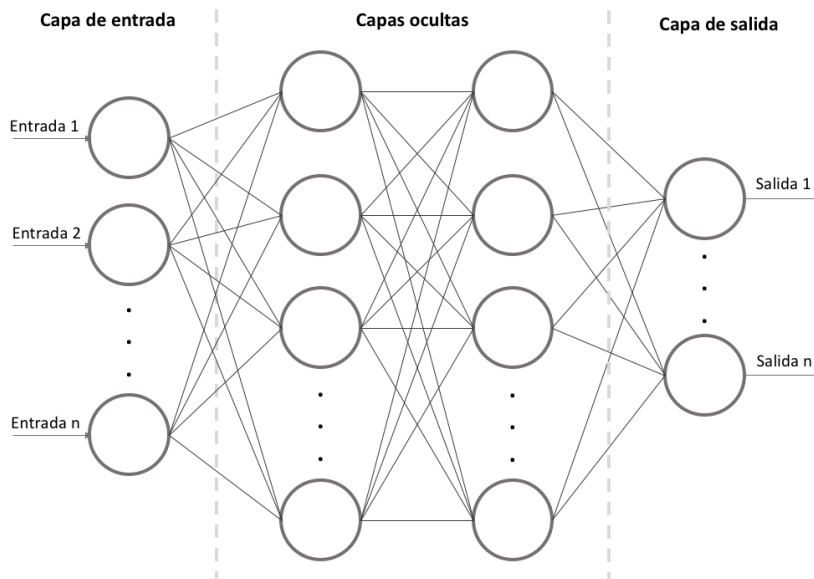


Figura 3.14: Estructura de una red neuronal artificial

Con base en 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.

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, como la mostrada en la Figura 3.14. Goodfellow *et al.* (2016) 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 retropropagación 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. Los pasos que realiza el algoritmo de retropropagación, descritos por Shai y Shai (2014) y representados de manera gráfica en la Figura 3.15, 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 retropropagación es un conjunto de pesos que minimizan la función de error.
4. Los pesos se actualizan.

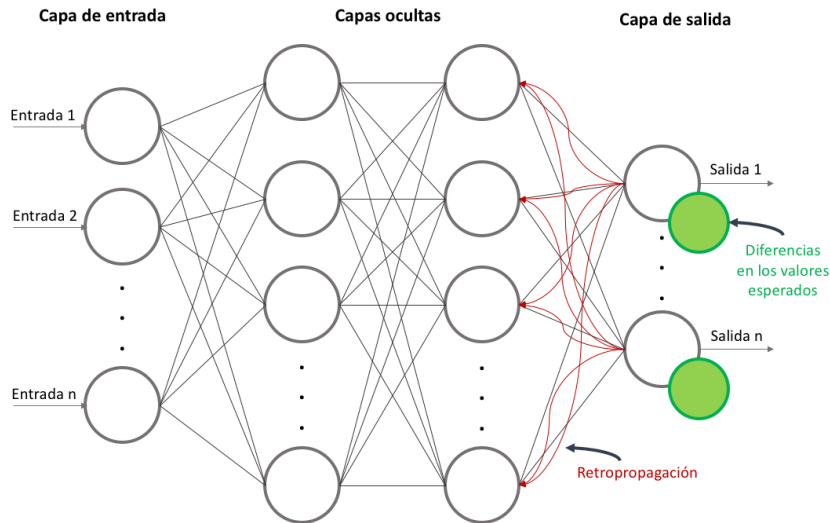


Figura 3.15: Esquema del funcionamiento del algoritmo de retropropagación

En general, el algoritmo de retropropagación 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.

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.

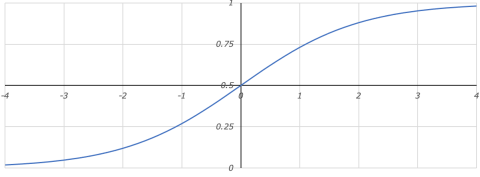
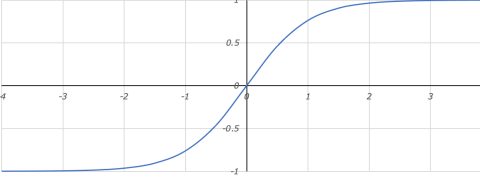
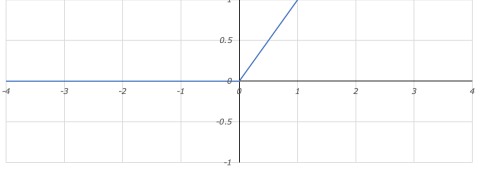
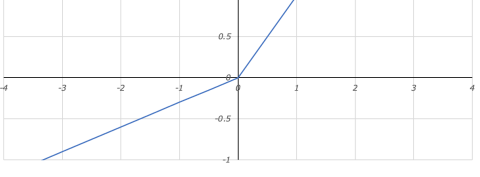
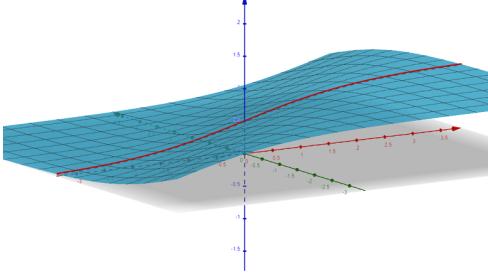
En una red neuronal, los valores de entrada se introducen en las neuronas de la red. Cada neurona tiene un peso y las entradas se multiplican por dicho peso para, posteriormente, alimentar a la función de activación.

La salida de cada neurona es, a su vez, 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 retropropagación explicado anteriormente.

La selección de una función de activación es crítica para la construcción y entrenamiento de la red. Las funciones de activación que Sharma *et al.* (2020) definen como las 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 cero 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.

Tabla 3.1: Funciones de activación para redes neuronales más comunes

Función	Ecuación	Curva
Sigmoide	$f(x) = \frac{1}{1+e^{-x}}$	
TanH	$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	
Relu	$f(x) = \begin{cases} 0, & \text{si } x \leq 0 \\ x, & \text{si } x > 0 \end{cases}$	
Leaky Relu	$f(x) = \begin{cases} a \times x, & \text{si } x \leq 0 \\ x, & \text{si } x > 0 \end{cases}$	
Softmax	$f(x) = \frac{e^{z_j}}{\sum_{k=1}^K e^{x_k}}$	

En este capítulo, se dió una introducción al aprendizaje automático y se explicaron los fundamentos de los algoritmos utilizados en el desarrollo de este trabajo. Para resolver el problema de clasificación de aceites, se utilizaron los algoritmos de regresión logística, árboles de decisión y redes neuronales. Mientras que para la estimación de curvas PVT, se decidió implementar los algoritmos de regresión lineal, máquinas de vectores de soporte y redes neuronales.

Como se menciona en esta sección, una parte importante del ciclo de vida del aprendizaje automático es la identificación de fuentes de datos, así como la preparación y selección de variables de entrada para entrenar a los modelos. En el siguiente capítulo, se abordarán los conceptos básicos de Ingeniería Petrolera para entender las variables utilizadas para el entrenamiento de los modelos de aprendizaje automático en este trabajo.

Capítulo 4

Propiedades PVT de fluidos petroleros

En este capítulo se dará una introducción a las propiedades PVT más relevantes en la industria petrolera. La abreviatura PVT significa presión, volumen y temperatura, y es utilizada en la industria petrolera para denotar los cambios que tienen las propiedades de los fluidos cuando varían las condiciones físicas del medio en el que se encuentran. Una estimación precisa de la variación de las propiedades físicas de los fluidos es de gran importancia en diversos campos de la Ingeniería Petrolera.

4.1. Definición de las propiedades PVT de un aceite

4.1.1. Densidad del aceite ρ_o

Ahmed (2006), define la densidad del petróleo crudo como la masa que tiene una unidad de volumen a condiciones de presión y temperatura designadas. Generalmente, esta propiedad se expresa con la unidad de kilogramo sobre metro cúbico [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}) o de yacimiento (ρ_{oy}).

A partir del valor de ρ_{ob} , es posible estimar ρ_{oy} en un yacimiento bajo saturado (presión por arriba del punto de burbuja) con la ecuación

$$\rho_{oy} = \rho_{ob} \exp [C_o (p_y - p_b)], \quad (4.1)$$

donde p_y hace referencia a la presión del yacimiento, p_b es la presión de burbuja y C_o representa la compresibilidad isotérmica del aceite.

4.1.2. Presión de burbuja P_b

Esta propiedad, 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 a 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.

4.1.3. Densidad relativa del aceite ρ_{ro}

Esta propiedad, también conocida como gravedad específica del aceite, se define como la densidad del aceite dividida entre la densidad del agua, ambas medidas, generalmente, a condiciones estándar.

$$\rho_{ro} = \frac{\rho_o}{\rho_w}, \quad (4.2)$$

donde ρ_{ro} representa la densidad relativa del aceite, ρ_o la densidad del aceite y ρ_w la densidad del agua.

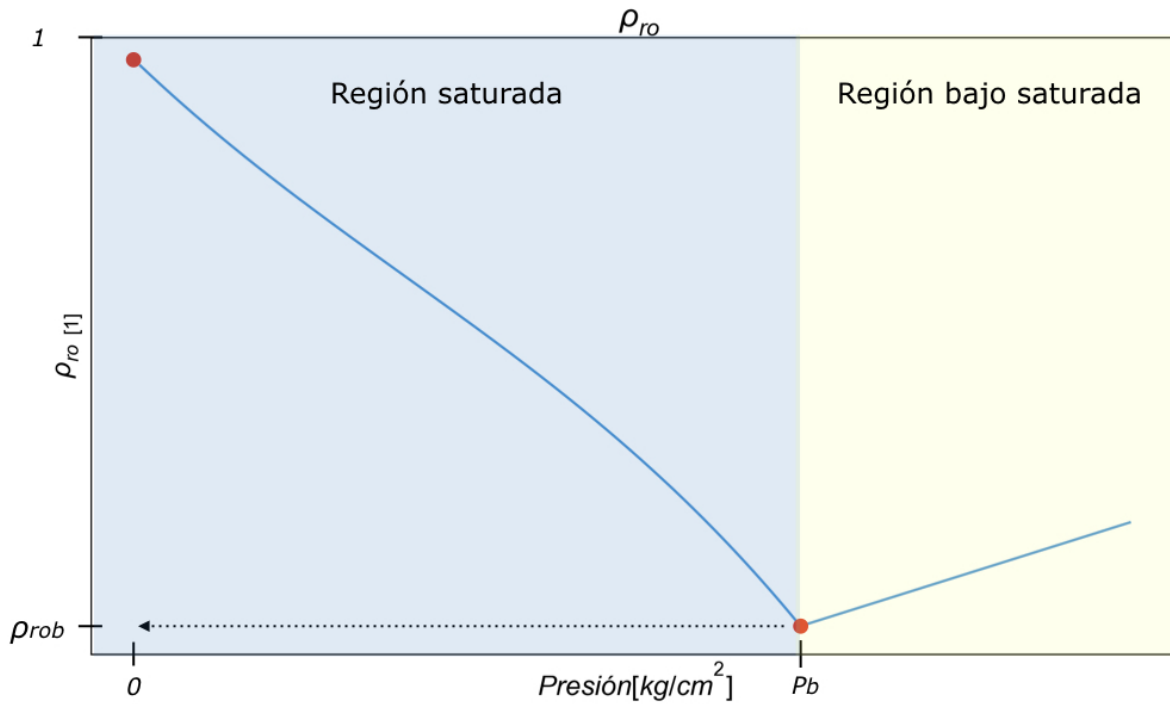


Figura 4.1: Comportamiento de la densidad relativa del aceite (ρ_{ro}) vs Presión

En la Figura 4.1 se muestra la curva típica del comportamiento de la densidad relativa 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 por lo que solo existe una sola fase en el sistema (aceite + gas disuelto). A medida que la presión se reduce, el volumen del aceite aumenta y por lo tanto, 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 este punto (región saturada), la densidad del aceite aumenta a medida que se libera el gas en solución y se pierden los elementos más ligeros.

4.1.4. 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 $[m^3/m^3]$.

$$R_s = \frac{V_{gd}}{(V_o)_{sc}}, \quad (4.3)$$

donde R_s hace referencia a la relación de solubilidad gas aceite, V_{gd} es el volumen de gas disuelto y $(V_o)_{sc}$ es el volumen de aceite a condiciones estándar.

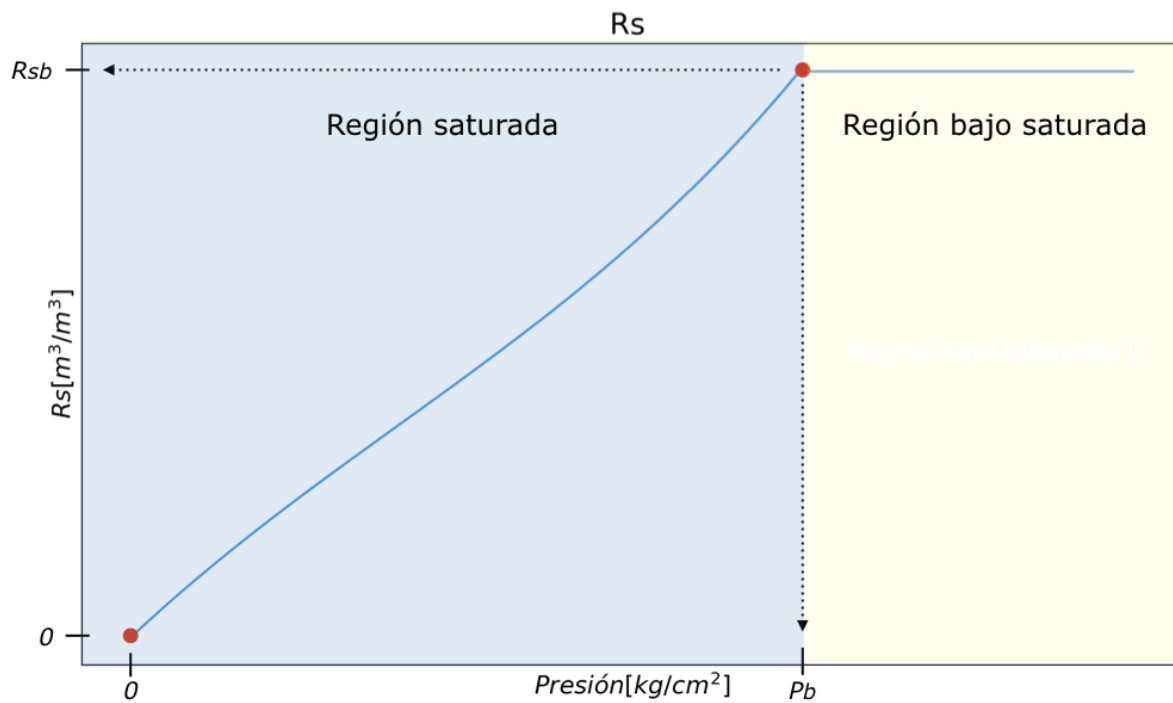


Figura 4.2: Comportamiento de la relación gas-aceite (R_s) vs Presión

La Figura 4.2 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 localice 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.

4.1.5. 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 del aceite a condiciones estándar. Por definición general, el factor de volumen del aceite es siempre mayor o igual que uno.

El factor de volumen del aceite se expresa matemáticamente como

$$B_o = \frac{(V_o)_{p,T}}{(V_o)_{sc}}, \quad (4.4)$$

donde B_o es el factor de volumen del aceite, $(V_o)_{p,T}$ es el volumen de aceite a la condición de presión y temperatura seleccionada y $(V_o)_{sc}$ es el volumen de aceite a condiciones estándar.

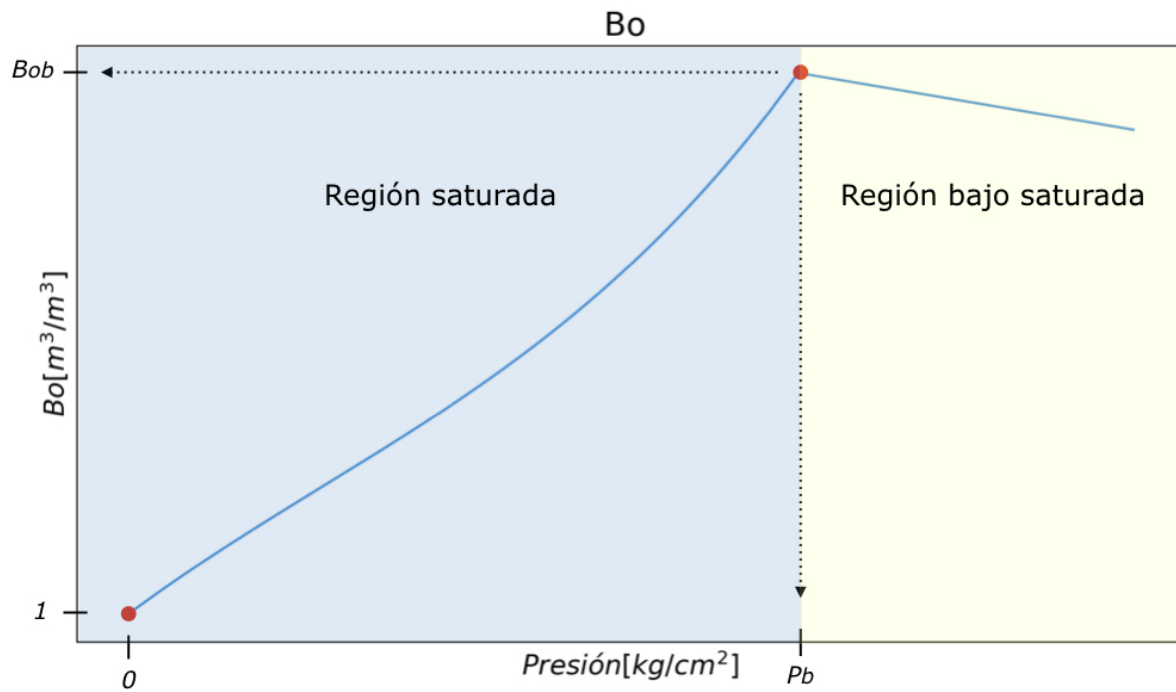


Figura 4.3: Comportamiento del factor de volumen del aceite (B_o) contra presión

La Figura 4.3 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 sólo existe una fase en el sistema.

A medida que la presión se reduce, el volumen del 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 y, a medida que la presión se sigue reduciendo por debajo de la P_b (región saturada), el volumen del aceite disminuye a

medida que se libera el gas en solución.

4.1.6. Viscosidad del aceite μ_o

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

Según las condiciones de presión y temperatura, la viscosidad puede clasificarse en tres categorías:

- **Viscosidad del aceite muerto (μ_{od}).** Es la viscosidad del petróleo crudo observada a presión atmosférica y a 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 a temperatura del yacimiento.
- **Viscosidad del aceite bajo saturado (μ_{ou}).** Hace referencia a la viscosidad del petróleo crudo registrada a una presión por arriba del punto de burbuja y a la temperatura de yacimiento.

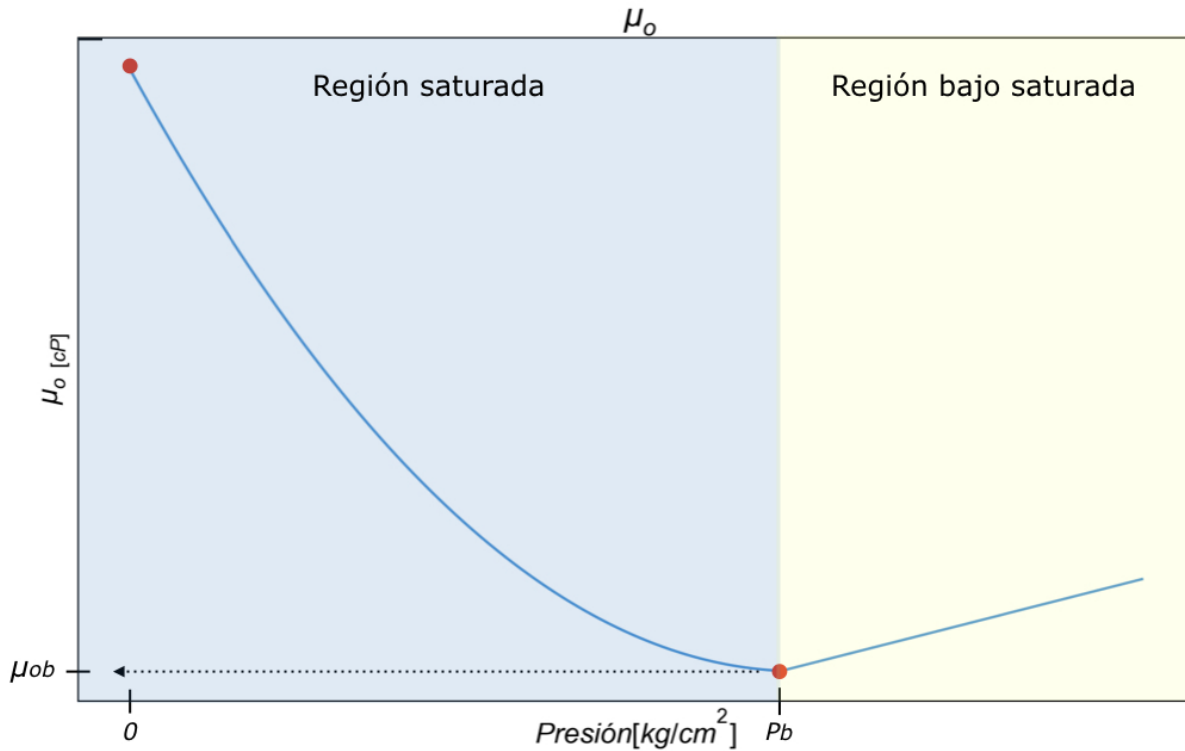


Figura 4.4: Comportamiento de la viscosidad del aceite (μ_o) vs Presión

En la Figura 4.4 se puede observar la curva típica del comportamiento de la viscosidad del aceite en función de la presión a temperatura constante.

Iniciando en una presión por arriba de la presión de burbuja, el aceite se ubica en la región bajo saturada y, al reducir el valor de la presión, la viscosidad también se reduce debido al efecto de expansión del aceite.

En la P_b , el aceite llega a su valor mínimo de viscosidad y, al continuar la reducción de la presión por debajo de la P_b (región saturada), la viscosidad del aceite aumenta a medida que se libera el gas en solución y se pierden los elementos más ligeros.

4.2. Clasificación de fluidos petroleros

Los hidrocarburos que se encuentran de forma natural en los yacimientos son mezclas de compuestos orgánicos que tienen un comportamiento multifásico cambiante en amplios intervalos de presiones y temperaturas. Las diferencias en el comportamiento de las fases dan como resultado la existencia de diversos tipos de yacimientos de hidrocarburos.

El diagrama de fases, es descrito por Ahmed (2006), como una herramienta útil para determinar el tipo de yacimiento de acuerdo a los fluidos que contiene. En este diagrama se representan de forma gráfica los estados físicos de una sustancia o mezcla de compuesto bajo diferentes condiciones de presión y temperatura.

La interpretación de un diagrama de fase es simple: cuando se cruzan las curvas, se origina un cambio de fase. Sin embargo, es necesario identificar los elementos clave indicados en la Figura 4.5 para una comprensión correcta.

- **Cricondenterma:** temperatura máxima por encima de la cual no se puede formar líquido independientemente de las condiciones de presión.
- **Cricondenbara:** presión máxima por encima de la cual no se puede formar gas independientemente de la temperatura.
- **Punto crítico:** para una mezcla multicomponente, es el punto de presión y temperatura en el que todas las propiedades intensivas de las fases líquida y gaseosa son iguales.
- **Envolverte de fase:** región delineada por las curvas de punto de burbuja y punto de rocío. Dentro de ella, el gas y el líquido coexisten en equilibrio.
- **Líneas de calidad:** líneas punteadas dentro del diagrama que describen las condiciones de presión y temperatura para obtener volúmenes iguales de líquidos.
- **Curva de puntos de burbuja:** curva que divide la región de la fase líquida con la región de dos fases.

- **Curva de punto de rocío:** curva que divide la región de la fase de vapor con la región de dos fases.

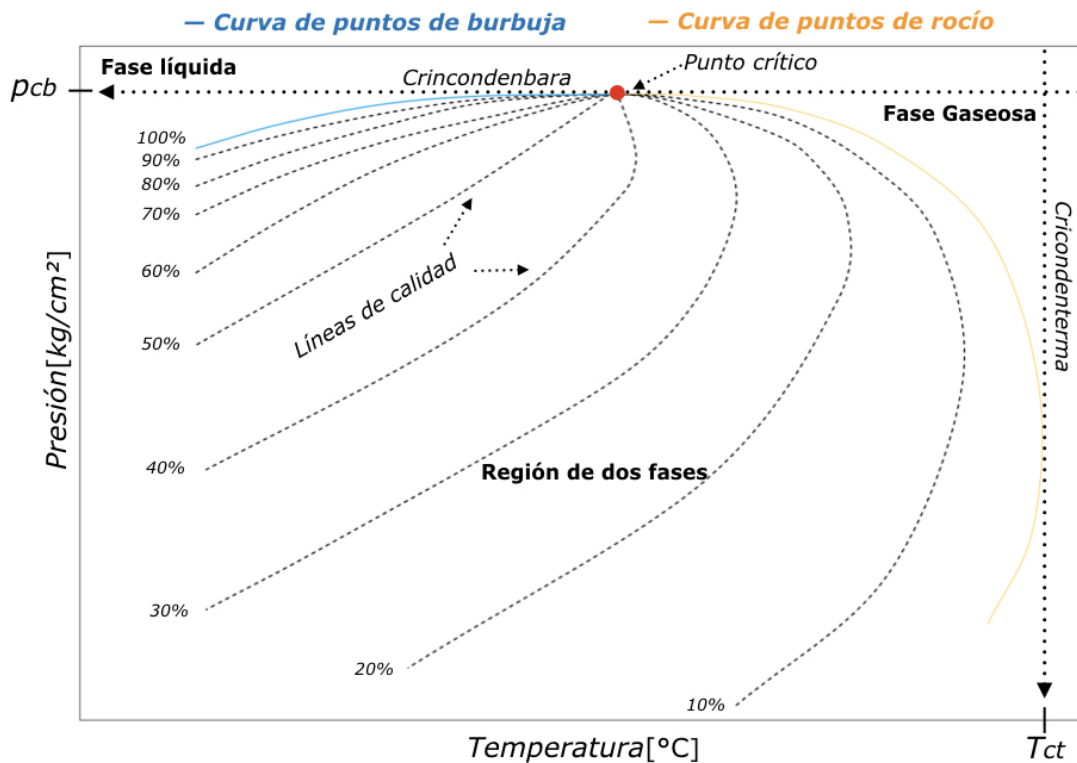


Figura 4.5: Diagrama de fases generado para una mezcla de hidrocarburos.

Según la ubicación del punto que representa la presión y temperatura del yacimiento en el diagrama de fases generado a partir del fluido que contienen, los yacimientos pueden ser clasificados en los siguientes tipos McCain (1990):

- **Aceite negro.** Consiste en una gran variedad de compuestos químicos que incluyen cadenas pesadas y largas de moléculas no volátiles.
- **Aceite volátil.** Comparado con los aceites negros, contiene menos moléculas pesadas y más cadenas intermedias (etanos a hexanos) de hidrocarburos.
- **Gas y condensado.** En este tipo de yacimientos, el hidrocarburo existe en forma de gas pero, puede existir la presencia de líquidos livianos (condensados) precipitados

en la vecindad del pozo. A medida que la presión decrece en el trayecto rumbo a la superficie, se precipitan cantidades considerables de líquidos condensados.

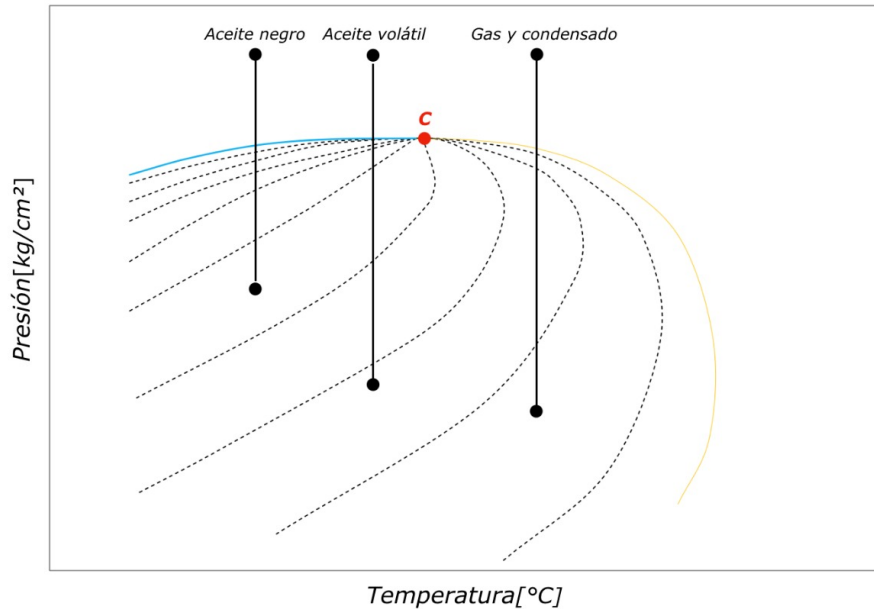


Figura 4.6: Ubicación de los tipos de yacimiento en un diagrama de fases.

Como puede verse en la Figura 4.6, en un diagrama de fases, los yacimientos de aceite negro se encuentran alejados a la izquierda del punto crítico, teniendo su fase estable en forma líquida dentro del yacimiento.

Por otro lado, los yacimientos de aceite volátil se encuentran también a la izquierda del punto crítico en el diagrama de fases pero, a una menor distancia en comparación con los yacimientos de aceite negro.

Finalmente, las condiciones de los yacimientos de gas y condensado se encuentran entre el punto crítico y la cricondenterma, obteniendo una fase gaseosa en el yacimiento. Sin embargo, conforme se reducen la presión y la temperatura, se entra en la envolvente de fases por su parte superior en donde se empieza a condensar parte del líquido.

En el siguiente capítulo, se mostrarán algunos métodos de aprendizaje automático a

partir de los cuales es posible llevar a cabo la clasificación y la estimación de la variación de las propiedades PVT de los aceites petroleros basándose en algunas de sus propiedades físicas mencionadas en este capítulo.

Capítulo 5

Clasificación de fluidos petroleros

En este capítulo se explicará la metodología utilizada para resolver el problema de clasificación de aceites petroleros mediante algoritmos de aprendizaje automático como regresión logística, árboles de decisión y redes neuronales. Para ello, se abordará desde la obtención de datos hasta los resultados obtenidos a partir de cada uno de los modelos entrenados.

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 20° [C] y la relación gas – aceite disuelto 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 L.T. y Tayssier (1979), mostrados en la Tabla 5.1, para determinar a qué tipo de yacimiento pertenecía cada muestra de los pozos elegidos.

Tabla 5.1: Clasificación de Méndez para los tipos de aceites.

Propiedad	Aceite negro	Aceite volátil
$\rho_{ro}[1]$	> 0.85	$0.75 - 0.85$
$R_s[m^3/m^3]$	< 200	$200 - 1000$
$B_{ob}[m^3/m^3]$	< 2.0	> 2.0

Como se puede observar en la Tabla 5.2, 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 5.3 pueden verse los valores utilizados para normalizar las variables de entrada del conjunto de datos.

Tabla 5.2: Número de muestras utilizadas para entrenar y validar los modelos de clasificación.

Tipo	Número de pozos
Aceites negros	53
Aceites volátiles	53
Total de muestras	106

Tabla 5.3: Rango de valores de las propiedades del conjunto de datos.

Propiedad	Valor mínimo	Valor máximo	Promedio
$\rho_{ro}[1]$	0.798	0.933	0.854
$R_s[m^3/m^3]$	11.1	783.1	221.8
$T_y[^\circ C]$	42.6	162.8	116.4

Posteriormente, a partir de los resultados obtenidos mediante el mapa de correlación mostrado en la Figura 5.1, 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 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 (T_y), la densidad relativa del aceite (ρ_{ro}) y la relación gas – aceite disuelto (R_s).

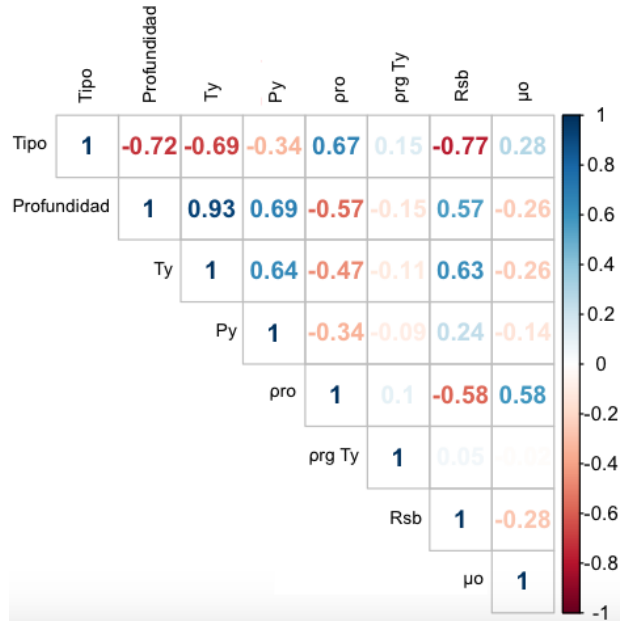


Figura 5.1: Mapa de correlación para elegir las variables de clasificación.

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 *et al.*, 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 ecuación 4.1.

$$v' = \frac{v - \min_A}{\max_A - \min_A}, \tag{5.1}$$

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.

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.

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.

5.1. 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 5.4, se construyó la ecuación 4.2, a partir de la cual es posible realizar la clasificación de los aceites.

$$Tipo = -7.196 - 7.165(T) + 29.531(\rho_{ro}). \quad (5.2)$$

Tabla 5.4: Coeficientes de la regresión logística.

Variable	Coefficiente
Intersección	-7.196
T_y	-7.165
ρ_{ro}	29.531

La ecuación 4.2 puede ser utilizada para predecir la probabilidad de que el fluido

petrolero sea negro (1) o volátil (0), para lo que toma en cuenta los valores que asumen las variables T_y y ρ_{ro} . Con base en ello, el modelo calculará una determinada probabilidad de que el fluido sea negro (es decir, el valor de la variable dependiente es cercano a 1) o volátil (el valor de la variable dependiente es cercano a 0).

A partir de la ecuación 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 mostrados en la matriz de confusión.

Tabla 5.5: Matriz de confusión de la clasificación mediante regresión logística.

Modelo Logístico		
	Volátil	Negro
Volátil	10	0
Negro	3	7

De la matriz de confusión mostrada en la Tabla 5.5, 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 15%.

5.2. Árboles de decisión

Para el caso de los árboles de decisión, las reglas de inducción a las que se llegaron mediante este algoritmo se muestran en la Figura 5.2.

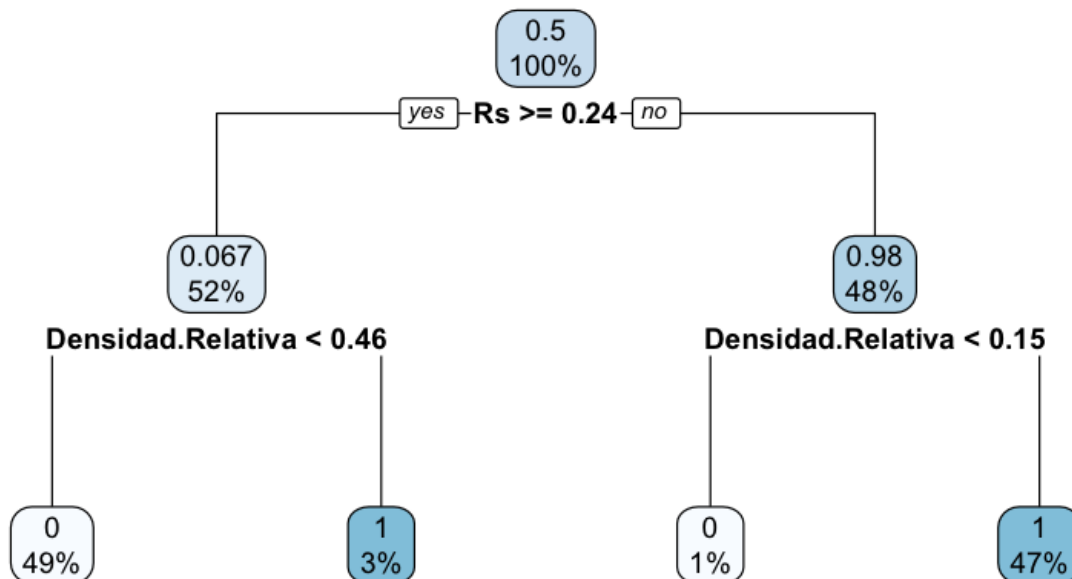


Figura 5.2: Árbol de decisión generado para clasificar aceites petroleros.

Tabla 5.6: Resultados del algoritmo de árbol de decisión

Condición	Resultado
Si $R_s \geq 0.24$ y $\rho_{ro} < 0.46$	<i>Tipo = volatil</i>
Si $R_s \geq 0.24$ y $\rho_{ro} \geq 0.46$	<i>Tipo = negro</i>
Si $R_s < 0.24$ y $\rho_{ro} < 0.15$	<i>Tipo = volatil</i>
Si $R_s < 0.24$ y $\rho_{ro} \geq 0.15$	<i>Tipo = negro</i>

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 siguiente matriz de confusión.

Tabla 5.7: Matriz de confusión de la clasificación mediante árboles de decisión.

Árboles de decisión		
	Volátil	Negro
Volátil	10	0
Negro	0	10

La Tabla 5.7, 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 %.

5.3. 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.

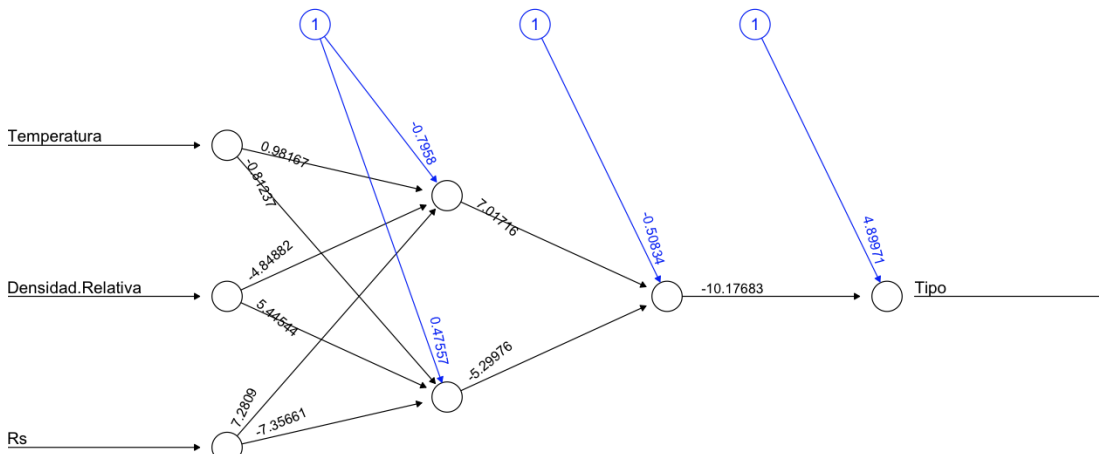


Figura 5.3: Red neuronal obtenida para clasificar aceites petroleros.

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 la Tabla 5.8.

Tabla 5.8: Matriz de confusión de la clasificación mediante redes neuronales artificiales.

Redes neuronales artificiales		
	Volátil	Negro
Volátil	10	0
Negro	0	10

Como se puede observar, la red neuronal artificial logró clasificar correctamente 10 aceites volátiles 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 con los modelos de aprendizaje automático creados para clasificar los yacimientos de acuerdo al tipo de aceite contenido pueden resumirse en la tabla mostrada a continuación.

Tabla 5.9: Comparación de los resultados obtenidos con los modelos de clasificación entrenados.

Algoritmo	Exactitud	Tasa de error
Regresión logística	85 %	15 %
Árboles de decisión	100 %	0 %
Redes neuronales artificiales	100 %	0 %

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 fiable 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.

En el siguiente capítulo se abordará la aplicación de métodos de regresión de aprendizaje automático para la estimación de curvas de las propiedades PVT de fluidos petroleros. Para ello, se hará uso de propiedades físicas de fácil obtención, al igual que en este capítulo. El objetivo es construir la curva completa de diferentes propiedades PVT en su región saturada y bajo saturada.

Capítulo 6

Estimación de propiedades de fluidos petroleros

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 físicas 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 buscan estimar de los aceites a condiciones de temperatura de yacimiento conforme varía la presión son las siguientes:

- Presión de burbuja (P_b)
- Curva de factor de volumen del aceite (B_o)
- Curva de densidad del aceite (ρ_o)
- Curva de relación de solubilidad gas-aceite (R_s)
- Curva de viscosidad del aceite (μ_o)

Para armar el conjunto de datos 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, la temperatura del yacimiento, la profundidad del yacimiento, la presión del pozo, viscosidad del aceite a 20° [C] y el Rs de pozos de una región de México.

Estas propiedades de entrada fueron elegidas debido a su alto valor de correlación con las propiedades que se desean calcular. En la Figura 6.1 pueden observarse los valores de correlación entre cada una de las propiedades PVT que conforman al conjunto de datos. En este mapa, un coeficiente positivo y alto indica que ambas variables crecen o decrecen de forma simultánea y, por lo tanto, presentan una fuerte correlación directa. Por otro lado, un coeficiente alto y negativo indica que, cuando una de las variable crece, la otra decrece y viceversa, es decir, presentan una fuerte correlación inversa. Finalmente, si el coeficiente es igual a cero o se aproxima a cero, indica que no existe relación entre ambas variables.

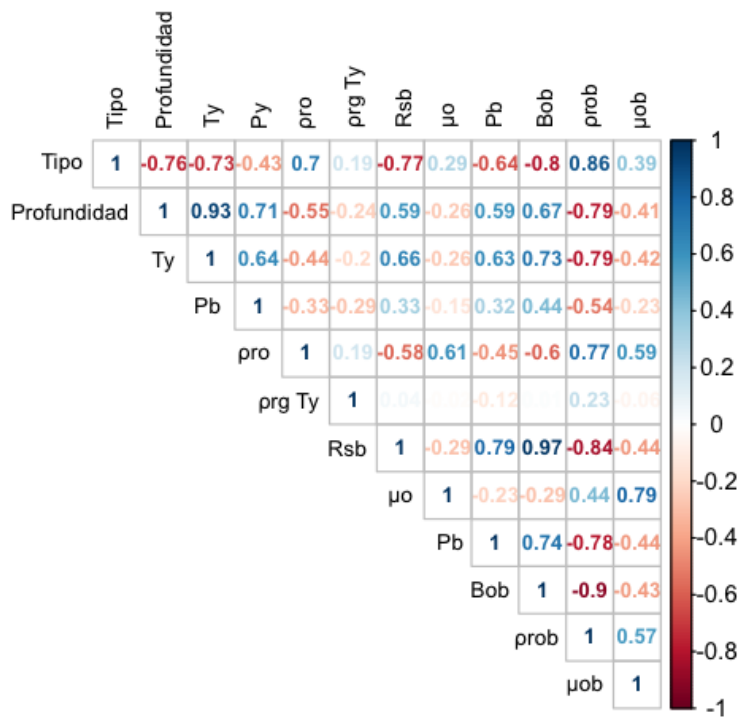


Figura 6.1: Mapa de correlación para elegir las variables de regresión.

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).

Es importante puntualizar 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 cuyos resultados pueden ser combinados posteriormente para generar la curva completa del comportamiento del fluido con la variación de presión.

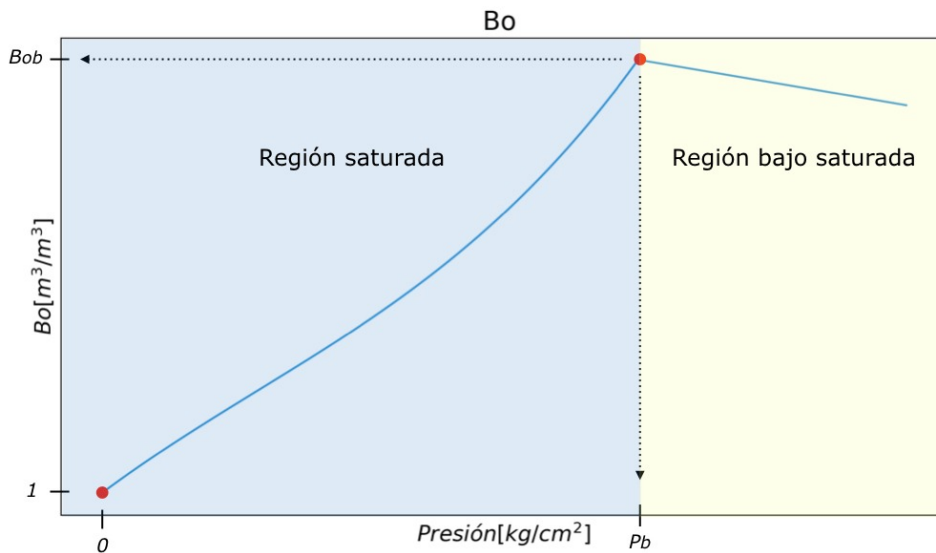


Figura 6.2: Ejemplo de la gráfica a generar para cada propiedad.

6.1. 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 calcular los valores de las propiedades a dichas presiones es necesario encontrar, en primer lugar, la P_b registrada en los reportes PVT para, posteriormente, encontrar la función matemática del grado adecuado que describe 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.

$$f(x) = A_1x^3 + A_2x^2 + A_3x + A_4, \tag{6.1}$$

Por ejemplo, para un pozo cuyo B_o en la región bajo saturada es expresado mediante

$$B_{o\ sat}(x) = (3.16 \cdot 10^{-8})x^3 - (1.49 \cdot 10^{-5})x^2 + (3.4 \cdot 10^{-3})x + 1.057,$$

y de la cual, la P_b reportada es 293 $[\frac{kg}{cm^2}]$, se obtuvieron los puntos mostrados en la Tabla 6.1, con los cuales es posible entrenar el modelo de la región saturada.

Tabla 6.1: Ejemplo de salidas recolectadas para la región saturada.

Propiedad	$\frac{1}{5}P_b$	$\frac{2}{5}P_b$	$\frac{3}{5}P_b$	$\frac{4}{5}P_b$	P_b
Presión $[\frac{kg}{cm^2}]$	58.6	117.2	175.8	234.4	293
$B_o[m^3/m^3]$	1.214	1.306	1.372	1.450	1.579

6.2. 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 que pasa entre los puntos de los valores registrados en la P_b y los valores registrados en la P_y .

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 con el valor de P_b estimado con el modelo de la región saturada.

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

$$B_o \text{ bajosat}(x) = (-3.29 \cdot 10^{-4})x + 1.668, \quad (6.2a)$$

$$P_b = 293 \left[\frac{kg}{cm^2} \right], \quad (6.2b)$$

$$P_y = 309.92 \left[\frac{kg}{cm^2} \right], \quad (6.2c)$$

los valores calculados para entrenar al modelo de regresión son los mostrados en la Tabla 6.2.

Tabla 6.2: Ejemplo de salidas recolectadas para la región bajo saturada de la curva de B_o .

Propiedad	P_b	$P_b + 200$	ΔP
Presión $\left[\frac{kg}{cm^2} \right]$	293	493	200
$B_o [m^3/m^3]$	1.572	1.506	0.066

Una vez obtenidos estos puntos, se hace uso de la ecuación

$$B_o(P_y) = B_o(P_b) - \frac{B_o(P_b) - B_o(P_b + 200)}{200} \times (P_y - P_b) \left[\frac{m^3}{m^3} \right], \quad (6.3)$$

para encontrar el valor de B_o en P_y . De esta manera, con los valores mostrados en la Tabla 6.2, el valor de $B_o(P_y)$ se calcula como

$$B_o(P_y) = 1.572 - \frac{1.572 - 1.596}{200} \times (309.92 - 293) \left[\frac{m^3}{m^3} \right] \quad (6.4a)$$

$$B_o(P_y) = 1.566 \left[\frac{m^3}{m^3} \right]. \quad (6.4b)$$

Finalmente, con los valores de 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.

6.3. Estimación de P_b

6.3.1. Preparación de los datos para P_b

A partir del mapa de correlación mostrado en la Figura 6.1, 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, a pesar de presentar una alta correlación con la presión de burbuja, fue descartada debido a su alta dependencia con la temperatura del pozo.

El conjunto de datos utilizado para el entrenamiento, prueba y validación de los modelos de regresión para calcular la P_b se compone de muestras de 106 pozos con las características mostradas en la Tabla 6.4.

Tabla 6.3: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de P_b .

Tipo de yacimiento	Cantidad
Aceites negros	53
Aceites volátiles	53
Total	106

Tabla 6.4: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de P_b y normalizar las variables de entrada durante la etapa de procesamiento de datos.

Propiedad	Valor mínimo	Valor máximo	Promedio
$\rho_{ro}[1]$	0.798	0.933	0.854
$R_s[\frac{m^3}{m^3}]$	11.1	783.1	221.8
$T_y[^\circ C]$	42.6	162.8	116.4
$P_b[\frac{kg}{cm^2}]$	31.6	408.34	234

A partir de los datos de la Tabla 6.4, se normalizaron todos los datos de entrada a partir del método mín-máx. Es decir, para un pozo cuya temperatura es de $102^\circ C$, este

dato se normalizó, a partir de las características identificadas en la tabla anterior, con la ecuación

$$T_y' = \frac{102 - 42.6}{162.8 - 42.6} = 0.4941, \quad (6.5)$$

de donde, el valor $T_y' = 0.4941$ corresponde al valor normalizado de $T_y = 102^\circ C$.

6.3.2. Estimación de P_b mediante 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 ecuación

$$P_b' = 0.1003T_y + 0.9478R_s + 0.2105, \quad (6.6)$$

obtenida con el conjunto de datos de entrenamiento. Una vez obtenido el valor de P_b es necesario desnormalizar para, posteriormente, utilizarlo para graficar las demás propiedades del aceite, con la fórmula

$$P_b = P_b'(408.34 - 31.6) + 31.6. \quad (6.7)$$

Tabla 6.5: Coeficientes de la regresión lineal para estimación de P_b .

Variable	Coefficiente	Descripción
Intersección	0.2105	Corresponde a la ordenada al origen, es decir, es el valor de la presión de burbuja cuando las demás variables predictoras tienen un valor igual a cero.
T_y	0.1003	Este coeficiente indica que, para cada unidad adicional de temperatura, la presión de burbuja aumenta en una media de $0.1003 \left[\frac{kg}{cm^2} \right]$.
R_s	0.9478	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, la presión de burbuja aumentará en una media de $0.9478 \left[\frac{m^3}{m^3} \right]$.

Las métricas de error obtenidas para este algoritmo son las mostradas en la Tabla 6.6.

Tabla 6.6: Indicadores de algoritmo de regresión lineal para estimación de P_b .

Indicador	Valor
E_a	0.1739
R^2	0.8339

6.3.3. Estimación de P_b mediante SVR

El modelo de 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 parámetros de error mostrados en la Tabla 6.7.

Tabla 6.7: Indicadores del algoritmo SVR para estimación de P_b .

Indicador	Valor
E_a	0.1419
R^2	0.935

6.3.4. Estimación de P_b mediante redes neuronales

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:

1. Capa de entrada: 3 neuronas correspondientes a las propiedades de entrada para calcular P_b .
2. Dos capas ocultas de tres y dos neuronas, respectivamente.
3. Una capa de salida con una neurona equivalente al valor de P_b calculado.

La función de activación utilizada en todas las neuronas fue sigmoideal.

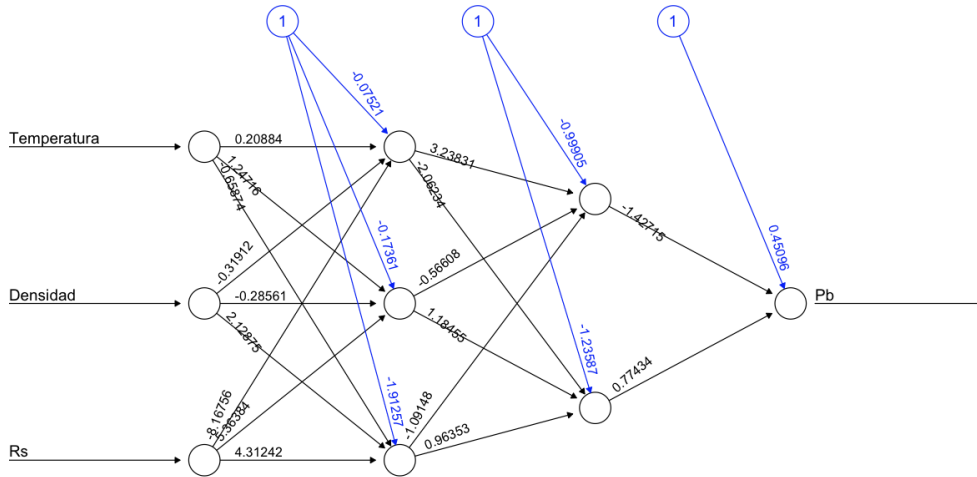


Figura 6.3: Red neuronal generada para estimar el valor de P_b .

A continuación, en la Tabla 6.8, se muestran las métricas obtenidas con el modelo redes neuronales creado para estimar la P_b .

Tabla 6.8: Indicadores del algoritmo de Redes Neuronales para estimación de P_b .

Indicador	Valor
E_a	0.0953
R^2	0.9263

6.3.5. Comparación de los tres métodos

La Tabla 6.9 muestra una comparación de los tres modelos propuestos para la estimación de la presión de burbuja a partir de los datos de entrada mencionados.

Tabla 6.9: Indicadores del algoritmo de Redes Neuronales para estimación de P_b .

Algoritmo	Porcentaje de error absoluto medio [%]
Regresión lineal	17.39
SVR	14.19
RNA	9.53

Como se mencionó anteriormente, el conjunto de datos fue dividido en dos partes: una de entrenamiento y otra de validación. Para observar gráficamente los resultados

obtenidos con los modelos de aprendizaje automático utilizados, se hizo uso del conjunto de datos de validación, mostrado en la Tabla 6.10.

Tabla 6.10: Conjunto de datos de prueba para P_b

Pozo	$T_y[^\circ C]$	$\rho[1]$	$R_s[m^3/m^3]$	$\mu@20^\circ C[cP]$
A	127	0.8995	96.397	18.57
B	102	0.8567	175.7	9.341
C	47	0.8596	80.6	10.9581
D	126	0.8514	401.1	7.0465
E	147	0.8338	233.466	5.5465
F	124	0.8497	373.7	4.8473

Los valores de P_b calculados con los tres algoritmos mencionados anteriormente, así como el valor de P_b registrado en los reportes PVT, pueden apreciarse en la Tabla 6.11.

Tabla 6.11: Valores estimados de P_b con los tres modelos de aprendizaje automático

Pozo	$P_bPVT[kg/cm^2]$	$P_bReg.Lineal[kg/cm^2]$	$P_bSVR[kg/cm^2]$	$P_bRNA[kg/cm^2]$
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	272	246.558	244.205	256.02
F	318.5	304.192	318.274	323.542

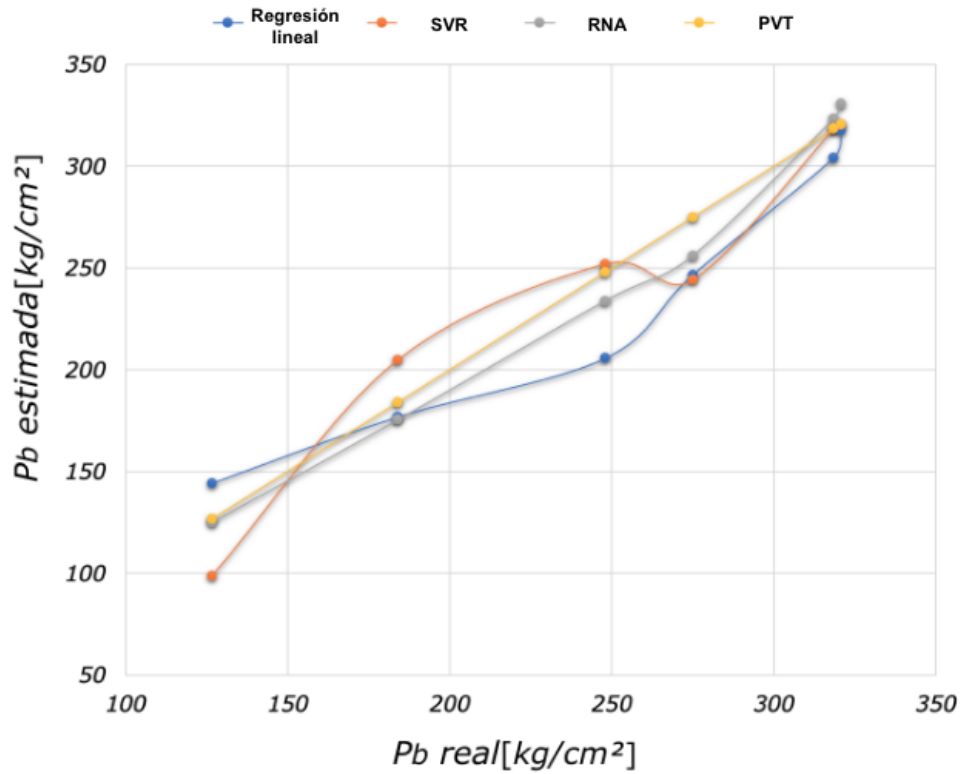


Figura 6.4: Gráfica de resultados de P_b obtenidos con los modelos de aprendizaje automático.

La Figura 6.4 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 PVT 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 artificiales, con un valor de E_a del 9.53%, por lo que, para obtener esta propiedad, puede considerarse como el modelo más adecuado.

6.4. Estimación de B_o en la región saturada

6.4.1. Preparación de los datos para $B_{o\ sat}$

A partir del mapa de correlación mostrado en la Figura 6.1, 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}).

El conjunto de datos utilizado para el entrenamiento, prueba y validación de los modelos de regresión para calcular $B_{o\ sat}$ está conformado por muestras de 101 pozos con las características mostradas en la Tabla 6.13. A partir de esta información, se normalizaron los datos de todo el conjunto de datos utilizado a partir del método mín-máx.

Tabla 6.12: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de $B_{o\ sat}$.

Tipo de yacimiento	Cantidad
Aceites negros	51
Aceites volátiles	50
Total	101

Tabla 6.13: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de $B_{o\ sat}$ y normalizar las variables de entrada durante la etapa de procesamiento de datos.

Propiedad	Valor mínimo	Valor máximo	Promedio
$\rho_{ro}[1]$	0.798	0.933	0.855
$R_s[\frac{m^3}{m^3}]$	11.1	783.1	221.8
$T_y[^\circ C]$	43.6	162.8	116.5
$B_{ob}[\frac{m^3}{m^3}]$	1.10	3.90	1.82

6.4.2. Estimación de B_o en la región saturada mediante regresión lineal

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

Primer punto ($\frac{1}{5}P_b$)

$$B'_{o \text{ sat}} @ \frac{1}{5}P_b = 0.16209T_y - 0.09263\rho_o + 0.73161R_s + 0.10266 \quad (6.8)$$

Tabla 6.14: Coeficientes de la regresión lineal para estimación de B_o *sat* en $\frac{1}{5}P_b$.

Variable	Coeficiente	Descripción
Intersección	0.10266	Corresponde a la ordenada al origen, es decir, es el valor del primer punto del factor de volumen del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	0.16209	Este coeficiente indica que, para cada unidad adicional de temperatura del yacimiento, el factor de volumen del aceite en el primer punto de la región saturada aumenta en una media de $0.16209 \left[\frac{m^3}{m^3}\right]$.
ρ_o	-0.09263	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, el factor de volumen del aceite en el primer punto de la región saturada disminuye en una media de $0.09263 \left[\frac{m^3}{m^3}\right]$.
R_s	0.73161	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, el factor de volumen del aceite en el primer punto de la región saturada aumenta en una media de $0.73161 \left[\frac{m^3}{m^3}\right]$.

En la Tabla 6.15 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de B_o *sat* en el primer punto de presión propuesto.

Tabla 6.15: Indicadores del algoritmo de regresión lineal para estimación de B_o *sat* en $\frac{1}{5}P_b$.

Indicador	Valor
E_a	0.0268
R^2	0.9777

Segundo punto ($\frac{2}{5}P_b$)

$$B'_{o \text{ sat}} @ \frac{2}{5}P_b = 0.1866T_y - 0.1486\rho_o + 0.6073R_s + 0.1435 \quad (6.9)$$

Tabla 6.16: Coeficientes de la regresión lineal para estimación de $B_{o \text{ sat}}$ en $\frac{2}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.1435	Corresponde a la ordenada al origen, es decir, es el valor del segundo punto del factor de volumen del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	0.1866	Este coeficiente indica que, para cada unidad adicional de temperatura del yacimiento, el factor de volumen del aceite en el segundo punto de la región saturada aumenta en una media de 0.1866 $[\frac{m^3}{m^3}]$.
ρ_o	-0.1486	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, el factor de volumen del aceite en el segundo punto de la región saturada disminuye en una media de 0.1486 $[\frac{m^3}{m^3}]$.
R_s	0.6073	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, el factor de volumen del aceite en el segundo punto de la región saturada aumenta en una media de 0.6073 $[\frac{m^3}{m^3}]$.

En la Tabla 6.17 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $B_{o \text{ sat}}$ en el segundo punto de presión propuesto.

Tabla 6.17: Indicadores del algoritmo de regresión lineal para estimación de $B_{o \text{ sat}}$ en $\frac{2}{5}P_b$.

Indicador	Valor
E_a	0.0363
R^2	0.9107

Tercer punto ($\frac{3}{5}P_b$)

$$B'_{o \text{ sat}} @ \frac{3}{5}P_b = 0.2115T_y - 0.1591\rho_o + 0.6571R_s + 0.1398 \quad (6.10)$$

En la Tabla 6.19 se muestran las métricas de error obtenidas con el modelo de regresión

Tabla 6.18: Coeficientes de la regresión lineal para estimación de $B_{o\ sat}$ en $\frac{3}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.1398	Corresponde a la ordenada al origen, es decir, es el valor del tercer punto del factor de volumen del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	0.2115	Este coeficiente indica que, para cada unidad adicional de temperatura del yacimiento, el factor de volumen del aceite en el tercer punto de la región saturada aumenta en una media de 0.2115 $[\frac{m^3}{m^3}]$.
ρ_o	-0.1591	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, el factor de volumen del aceite en el tercer punto de la región saturada disminuye en una media de 0.1591 $[\frac{m^3}{m^3}]$.
R_s	0.6571	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, el factor de volumen del aceite en el tercer punto de la región saturada aumenta en una media de 0.6571 $[\frac{m^3}{m^3}]$.

lineal para la estimación de $B_{o\ sat}$ en el tercer punto de presión propuesto.

Tabla 6.19: Indicadores del algoritmo de regresión lineal para estimación de $B_{o\ sat}$ en $\frac{3}{5}P_b$.

Indicador	Valor
E_a	0.0379
R^2	0.9453

Cuarto punto ($\frac{4}{5}P_b$)

$$B'_{o\ sat} @ \frac{4}{5}P_b = 0.18073T_y - 0.10739\rho_o + 0.86433R_s + 0.05474 \quad (6.11)$$

En la Tabla 6.21 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $B_{o\ sat}$ en el cuarto punto de presión propuesto.

Tabla 6.20: Coeficientes de la regresión lineal para estimación de $B_{o\ sat}$ en $\frac{4}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.05474	Corresponde a la ordenada al origen, es decir, es el valor del cuarto punto del factor de volumen del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	0.18073	Este coeficiente indica que, para cada unidad adicional de temperatura del yacimiento, el factor de volumen del aceite en el cuarto punto de la región saturada aumenta en una media de 0.18073 $[\frac{m^3}{m^3}]$.
ρ_o	-0.10739	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, el factor de volumen del aceite en el cuarto punto de la región saturada disminuye en una media de 0.10739 $[\frac{m^3}{m^3}]$.
R_s	0.86433	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, el factor de volumen del aceite en el cuarto punto de la región saturada aumenta en una media de 0.86433 $[\frac{m^3}{m^3}]$.

Tabla 6.21: Indicadores del algoritmo de regresión lineal para estimación de $B_{o\ sat}$ en $\frac{4}{5}P_b$.

Indicador	Valor
E_a	0.0268
R^2	0.9777

Quinto punto (P_b)

$$B'_{o\ sat} @ P_b = 0.10124T_y - 0.03649\rho_o + 0.92621R_s - 0.03973 \quad (6.12)$$

En la Tabla 6.23 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $B_{o\ sat}$ en el quinto y último punto de presión propuesto.

Tabla 6.23: Indicadores del algoritmo de regresión lineal para estimación de $B_{o\ sat}$ en P_b .

Indicador	Valor
E_a	0.036
R^2	0.99

Tabla 6.22: Coeficientes de la regresión lineal para estimación de $B_{o\ sat}$ en P_b .

Variable	Coefficiente	Descripción
Intersección	-0.03973	Corresponde a la ordenada al origen, es decir, es el valor del quinto punto del factor de volumen del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	0.10124	Este coeficiente indica que, para cada unidad adicional de temperatura del yacimiento, el factor de volumen del aceite en el quinto punto de la región saturada aumenta en una media de $0.10124 \left[\frac{m^3}{m^3}\right]$.
ρ_o	-0.03649	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, el factor de volumen del aceite en el quinto punto de la región saturada disminuye en una media de $0.03649 \left[\frac{m^3}{m^3}\right]$.
R_s	0.92621	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, el factor de volumen del aceite en el quinto punto de la región saturada aumenta en una media de $0.92621 \left[\frac{m^3}{m^3}\right]$.

6.4.3. Estimación de B_o en la región saturada mediante SVR

El modelo de SVR fue entrenado con la función Kernel radial y, haciendo uso del mismo conjunto de datos que el algoritmo de regresión lineal explicado anteriormente. Con este algoritmo, se obtuvieron los siguientes parámetros de error mostrados en la Tabla 6.24.

Tabla 6.24: Indicadores del algoritmo de SVR para estimación de $B_{o\ sat}$.

Punto	Indicador	Valor
$\frac{1}{5}P_b$	E_a	0.0284
	R^2	0.8987
$\frac{2}{5}P_b$	E_a	0.0321
	R^2	0.9347
$\frac{3}{5}P_b$	E_a	0.0325
	R^2	0.9608
$\frac{4}{5}P_b$	E_a	0.0321
	R^2	0.9775
$\frac{5}{5}P_b$	E_a	0.0336
	R^2	0.9807

6.4.4. Estimación de B_o en región saturada mediante 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 es la siguiente:

1. Capa de entrada: 3 neuronas correspondientes a las propiedades de entrada para calcular $B_o sat$.
2. Dos capas ocultas de cuatro y tres neuronas, respectivamente.
3. Una capa de salida con cinco neuronas, cada una correspondiente al valor de $B_o sat$ en cada punto de presión propuesto.

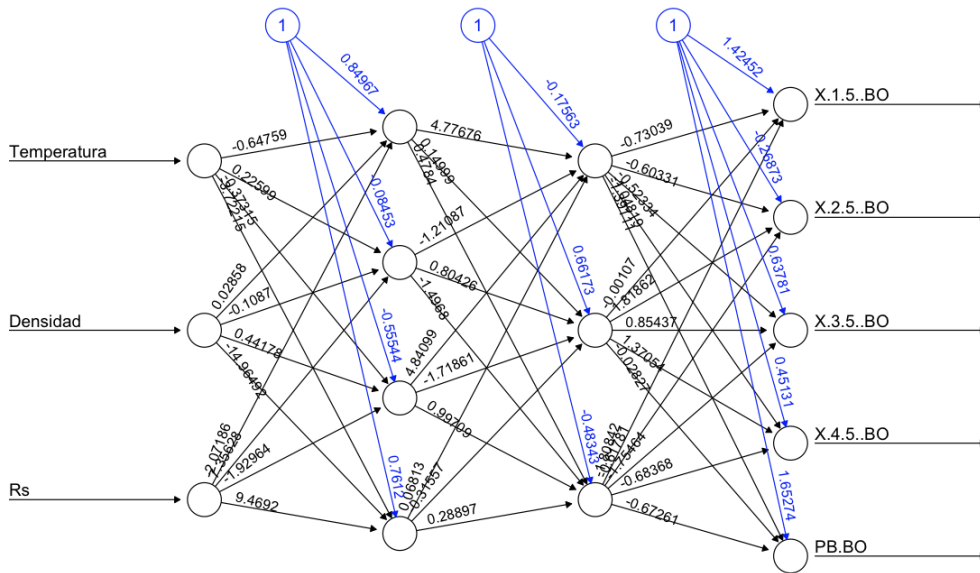


Figura 6.5: Red neuronal generada para estimar el valor de $B_o sat$.

A continuación, en la Tabla 6.25, se muestran las métricas obtenidas con el modelo redes neuronales creado para estimar el $B_o sat$.

Tabla 6.25: Indicadores del algoritmo de Redes Neuronales para estimación de $B_o sat.$

Punto	Indicador	Valor
$\frac{1}{5}P_b$	E_a R^2	0.026 0.8882
$\frac{2}{5}P_b$	E_a R^2	0.0277 0.9391
$\frac{3}{5}P_b$	E_a R^2	0.0275 0.966
$\frac{4}{5}P_b$	E_a R^2	0.0221 0.9842
P_b	E_a R^2	0.0216 0.9905

6.4.5. Comparación de los tres métodos

La Tabla 6.26 muestra una comparación de los tres modelos propuestos para la estimación del factor de volumen del aceite en su región saturada a partir de los datos de entrada propuestos. Se utiliza el promedio del porcentaje de error absoluto medio obtenido para los cinco puntos para comparar los tres modelos.

Tabla 6.26: Indicadores del algoritmo de Redes Neuronales para estimación de $B_o sat.$

Algoritmo	Porcentaje de error absoluto medio [%]
Regresión lineal	3.31
SVR	3.17
RNA	2.5

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 la red neuronal artificial 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.

6.5. Estimación de B_o en la región bajo saturada

6.5.1. Preparación de los datos para B_o *bajosat*

A partir del mapa de correlación mostrado en la Figura 6.1, 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}).

El conjunto de datos utilizado para el entrenamiento, prueba y validación de los modelos de regresión para calcular B_o *bajosat* se compone de muestras de 98 pozos con las características mostradas en la Tabla 6.28. A partir de esta información, se normalizaron los datos de todo el conjunto de datos utilizado a partir del método mín-máx.

Tabla 6.27: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de B_o *bajosat*.

Tipo de yacimiento	Cantidad
Aceites negros	50
Aceites volátiles	48
Total	98

Tabla 6.28: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de B_o *bajosat* y normalizar las variables de entrada durante la etapa de procesamiento de datos.

Propiedad	Valor mínimo	Valor máximo	Promedio
$\rho_{ro}[1]$	0.798	0.933	0.855
$R_s[\frac{m^3}{m^3}]$	11.1	783.1	217.7
$T_y[^\circ C]$	42.6	162.8	114.9
$(B_o @ (P_b + 200) - B_o @ P_b)[kg/cm^2][\frac{m^3}{m^3}]$	0.0142	0.5808	0.1102

6.5.2. Estimación de B_o región bajo saturada mediante regresión lineal

Mediante este algoritmo es posible estimar el valor normalizado de $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2]) [\frac{m^3}{m^3}]$ a partir de la temperatura del yacimiento y la relación gas-aceite disuelto con la ecuación

$$(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])' = 0.05105T_y + 0.80606R_s - 0.07659. \quad (6.13)$$

Tabla 6.29: Coeficientes de la regresión lineal para estimación de $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])$.

Variable	Coefficiente	Descripción
Intersección	-0.07659	Corresponde a la ordenada al origen, es decir, es el valor del factor de volumen del aceite en la región bajo saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	0.05105	Este coeficiente indica que, para cada unidad adicional de temperatura del yacimiento, el factor de volumen del aceite en la región bajo saturada aumenta en una media de $0.05105 [\frac{m^3}{m^3}]$.
ρ_o	-0.03649	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, el factor de volumen del aceite en la región bajo saturada disminuye en una media de $0.03649 [\frac{m^3}{m^3}]$.
R_s	0.80606	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, el factor de volumen del aceite en la región bajo saturada aumenta en una media de $0.80606 [\frac{m^3}{m^3}]$.

En la Tabla 6.30 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])$.

Tabla 6.30: Indicadores del algoritmo de regresión lineal para estimación de B_o_{sat} en $\frac{1}{5}P_b$.

Indicador	Valor
E_a	0.0268
R^2	0.9777

6.5.3. Estimación de B_o región bajo saturada mediante SVR

El modelo de SVR fue entrenado con la función Kernel radial. Con este algoritmo, se obtuvieron los siguientes parámetros de error mostrados en la Tabla 6.31.

Indicador	Valor
E_a	0.2636
R^2	0.9522

Tabla 6.31: Indicadores del algoritmo de SVR para estimación de $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])$.

6.5.4. Estimación de B_o *bajosat* mediante redes neuronales

El modelo de red neuronal artificial fue entrenado con el mismo conjunto de datos que los dos modelos anteriores y, a partir del algoritmo de retropropagación. En este caso, se propuso la siguiente estructura de red:

1. Capa de entrada: 3 neuronas correspondientes a las propiedades de entrada para calcular $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])$.
2. Dos capas ocultas de seis neuronas cada una.
3. Una capa de salida con una neurona correspondiente al valor de $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])$ calculado.

La función de activación utilizada en todas las neuronas fue sigmoideal.

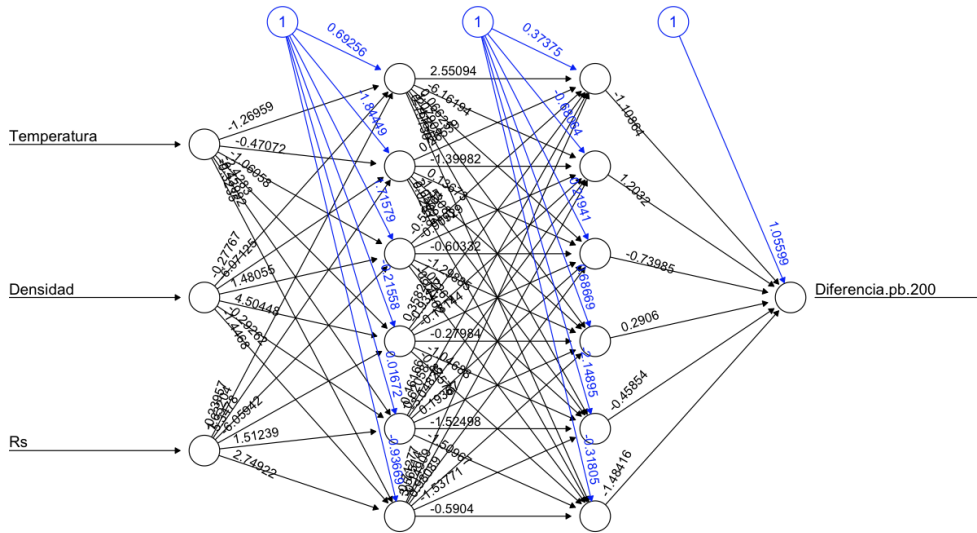


Figura 6.6: Red neuronal generada para estimar el valor de $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])$.

A continuación, en la Tabla 6.32, se muestran las métricas obtenidas con el modelo redes neuronales creado para estimar la $\Delta_{(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])}$.

Tabla 6.32: Indicadores del algoritmo de Redes Neuronales para estimación de $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])$.

Indicador	Valor
E_a	0.1199
R^2	0.9423

6.5.5. Comparación de los tres métodos

Hay que recordar que, una vez obtenido el valor de $(B_o @ (P_b + 200) - B_o @ P_b [kg/cm^2])$, 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.

Tabla 6.33: Errores en la estimación de B_{oy} .

Algoritmo	Porcentaje de error absoluto medio [%]
Regresión lineal	3.13
SVR	2.22
RNA	1.91

En la Tabla 6.33, 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 6.34.

Tabla 6.34: Comparación de resultados obtenidos con el set de datos de prueba para B_{oy}

Pozo	$B_{oy}PVT[m^3/m^3]$	$B_{oy}Reg.Lineal[m^3/m^3]$	$B_{oy}SVR[m^3/m^3]$	$B_{oy}RNA[m^3/m^3]$
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

6.5.6. Generación de la curva completa de B_o

La Tabla 6.35 muestra las características del conjunto de datos de validación, el cual está hecho a partir de una muestra tomada del conjunto de datos de prueba.

Tabla 6.35: Conjunto de datos de prueba para B_o

Pozo	$T_y[^\circ C]$	$\rho[1]$	$R_s[m^3/m^3]$	$\mu@20^\circ C[cP]$	$P_b[kg/cm^2]$	$P_y[kg/cm^2]$
A	127	0.8995	96.397	18.57	184	163
B	102	0.8567	175.7	9.341	248	322
C	47	0.8596	80.6	10.9581	126.6	141
D	126	0.8514	401.1	7.0465	320.7	456
E	147	0.8338	233.466	5.5465	275	374
F	124	0.8497	373.7	4.8473	318.5	427

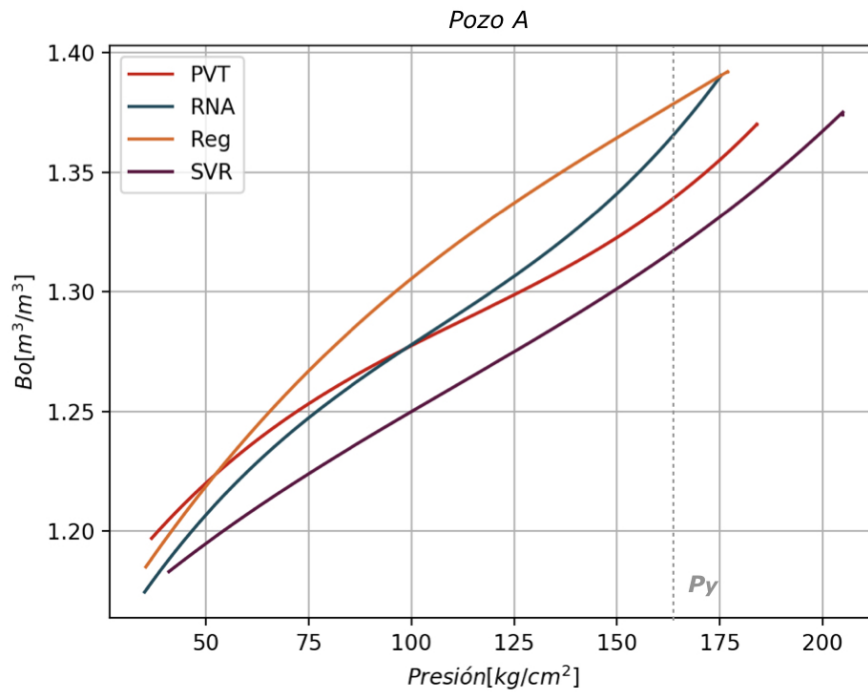


Figura 6.7: Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo A.

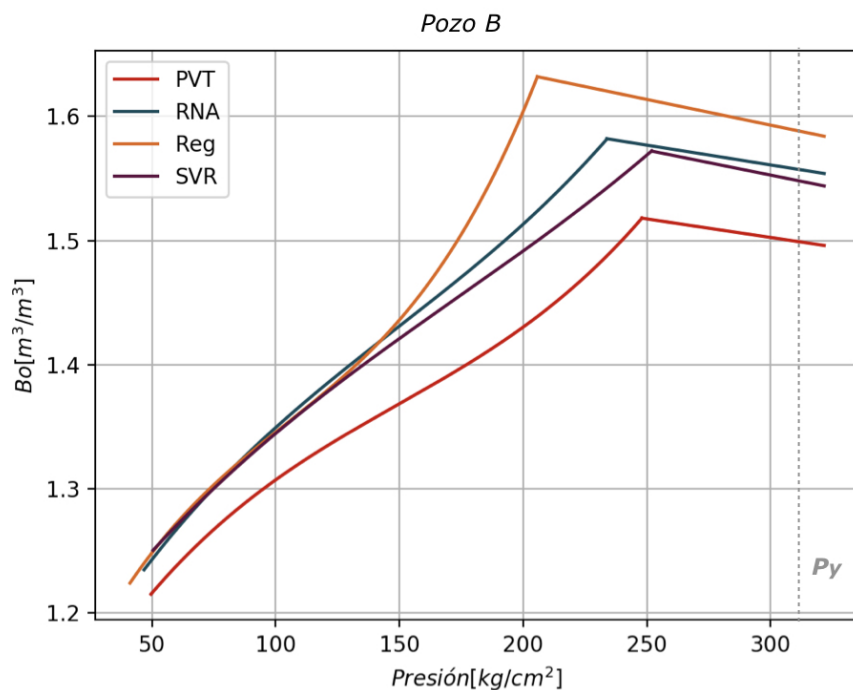


Figura 6.8: Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo B.

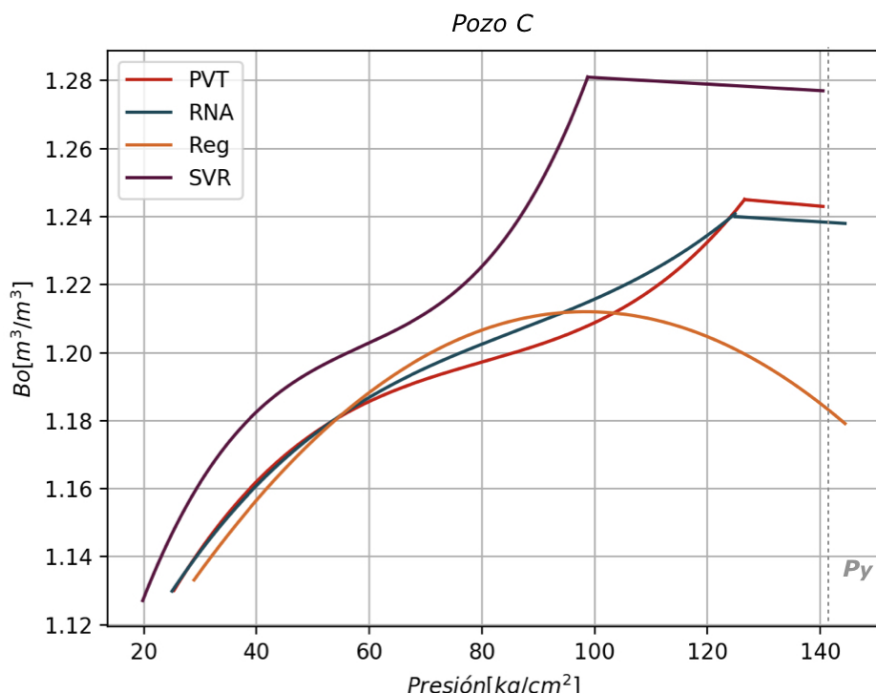


Figura 6.9: Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo C.

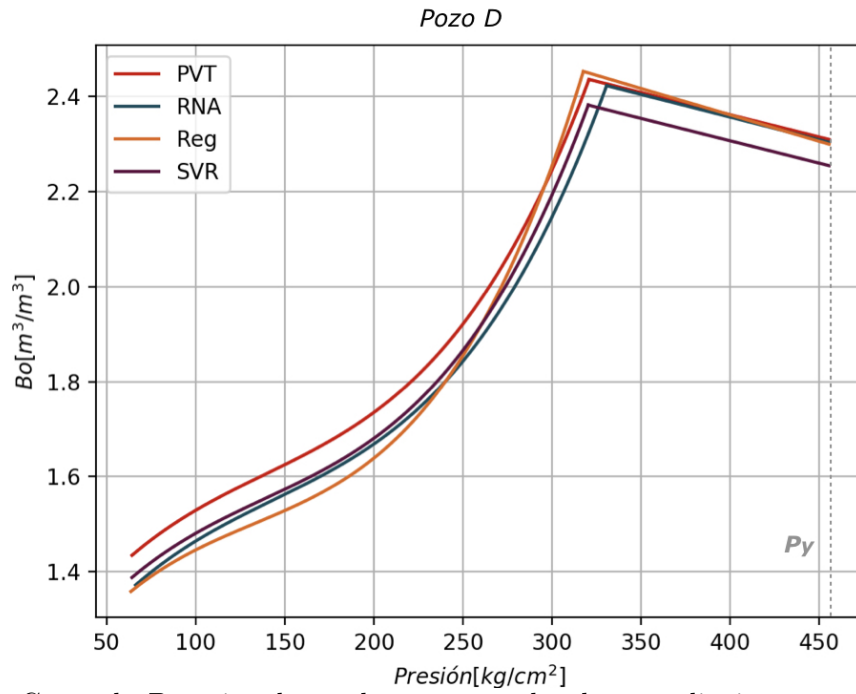


Figura 6.10: Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo D.

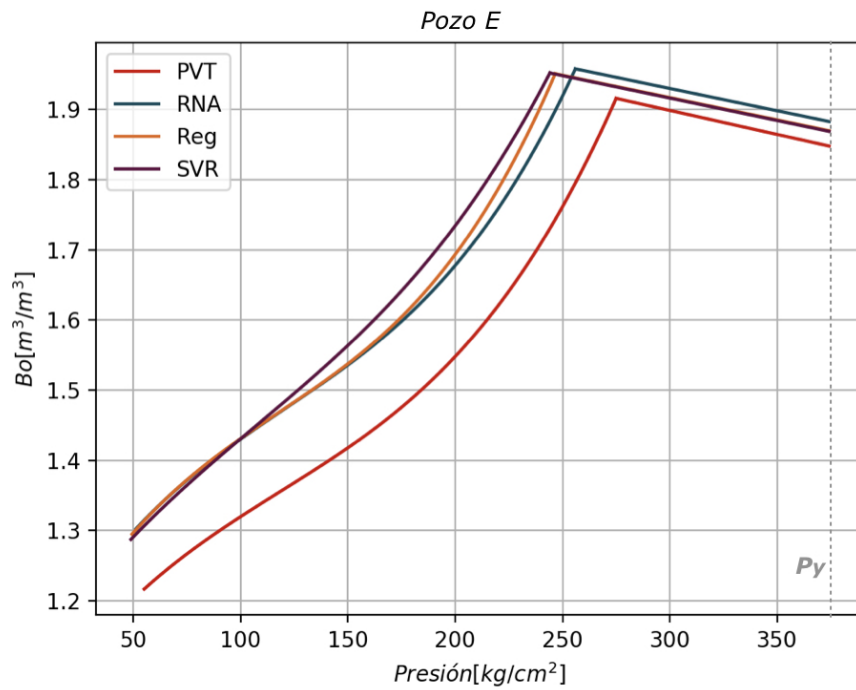


Figura 6.11: Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo E.

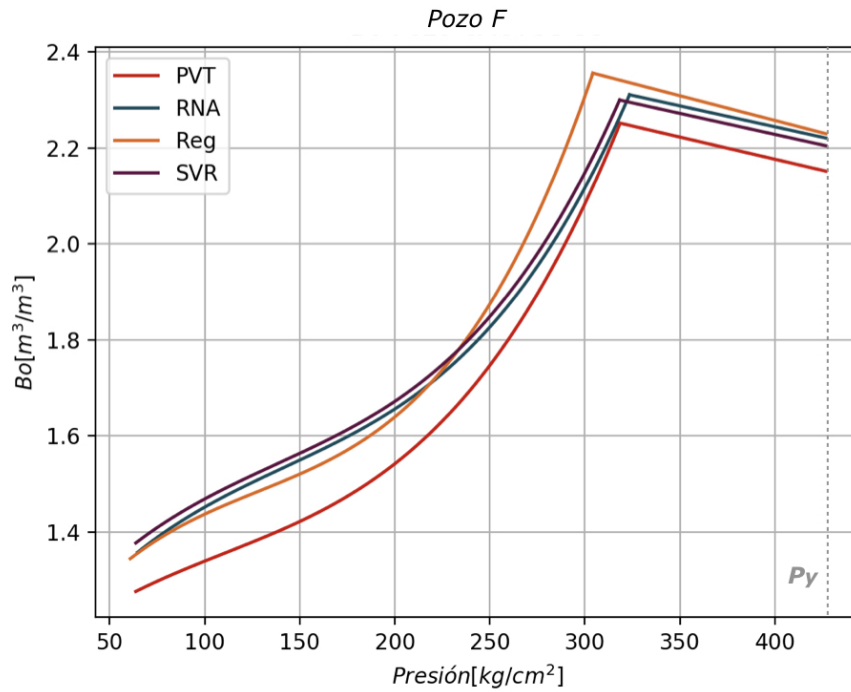


Figura 6.12: Curva de B_o estimada con los tres métodos de aprendizaje automático para el pozo F.

Las Figuras 6.7, 6.8, 6.9, 6.10, 6.11 y 6.12 nos permiten observar de manera gráfica las curvas generadas por la unión de los resultados obtenidos con los modelos de B_o en la región saturada y bajo saturada elaborados en este trabajo. Con las figuras anteriores, podemos confirmar visualmente que, para estimar B_o , el modelo de aprendizaje automático que tiene una mejor aproximación es la red neuronal artificial.

6.6. Estimación de la ρ_{ro} en la región saturada

6.6.1. Preparación de los datos para estimar ρ_{rosat}

A partir del mapa de correlación mostrado en la Figura 6.1, 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), la densidad relativa del aceite (ρ_{ro}) y la viscosidad relativa del aceite a $20^\circ C$ ($\mu_o @ 20^\circ C$).

El conjunto de datos utilizado para el entrenamiento, prueba y validación de los modelos de regresión para calcular ρ_{ro} en su región saturada está conformado por muestras de 103 pozos con las características mostradas en la Tabla 6.37. A partir de esta información, se normalizaron los datos de todo el conjunto de datos utilizado con el método mín-máx.

Tabla 6.36: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de ρ_{ro} .

Tipo de yacimiento	Cantidad
Aceites negros	52
Aceites volátiles	51
Total	103

Tabla 6.37: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de ρ_{ro} y normalizar las variables de entrada durante la etapa de procesamiento de datos.

Propiedad	Valor mínimo	Valor máximo	Promedio
$\rho_{ro}[1]$	0.798	0.933	0.855
$R_s[\frac{m^3}{m^3}]$	11.1	783.1	220.6
$T_y[^\circ C]$	43.6	162.8	116.6
$\mu_o @ 20^\circ C[cP]$	2.84	302.74	18.91
$\rho_{rob}[1]$	0.396	0.874	0.629

6.6.2. Estimación de ρ_{ro} en la región saturada mediante regresión lineal

Con este algoritmo se estimaron los valores normalizados de ρ_{ro} en los 5 puntos de presión propuestos, haciendo uso de las 4 entradas mencionadas anteriormente, a partir de las expresiones mostradas a continuación.

Primer punto ($\frac{1}{5}P_b$)

$$\rho'_{ro\ sat} @ \frac{1}{5}P_b = -0.20232T_y + 0.35113\rho_{ro} - 0.15774R_s + 0.04018\mu_o @ 20^\circ C + 0.63691, \quad (6.14)$$

Tabla 6.38: Coeficientes de la regresión lineal para estimación de $\rho_{ro\ sat}$ en $\frac{1}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.63691	Corresponde a la ordenada al origen, es decir, es el valor del primer punto de la densidad relativa del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.20232	Este coeficiente indica que, para cada unidad adicional de temperatura, la densidad relativa del aceite en el primer punto de la región saturada disminuye en una media de 0.20232 [1].
ρ_o	0.35113	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, la densidad relativa del aceite en el primer punto de la región saturada aumenta en una media de 0.35113 [1].
R_s	-0.15774	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, la densidad relativa del aceite en el primer punto de la región saturada disminuye en una media de 0.15774 [1].
$\mu_o @ 20^\circ C$	0.04018	Este coeficiente indica que, para cada unidad adicional de viscosidad, la densidad relativa del aceite en el primer punto de la región saturada aumenta en una media de 0.04018 [1].

En la Tabla 6.39 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $\rho_{ro\ sat}$ en el primer punto de presión propuesto.

Tabla 6.39: Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro\ sat}$ en $\frac{1}{5}P_b$.

Indicador	Valor
E_a	0.0204
R^2	0.9242

Segundo punto ($\frac{2}{5}P_b$)

$$\rho'_{ro\ sat\ @\ \frac{2}{5}P_b} = -0.2420T_y + 0.3846\rho_{ro} - 0.1946R_s + 0.0388\mu_o@20^\circ C + 0.6086, \quad (6.15)$$

Tabla 6.40: Coeficientes de la regresión lineal para estimación de $\rho_{ro\ sat}$ en $\frac{2}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.6086	Corresponde a la ordenada al origen, es decir, es el valor del segundo punto de la densidad relativa del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.2420	Este coeficiente indica que, para cada unidad adicional de temperatura, la densidad relativa del aceite en el segundo punto de la región saturada disminuye en una media de 0.2420 [1].
ρ_{ro}	0.3846	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, la densidad relativa del aceite en el segundo punto de la región saturada aumenta en una media de 0.3846 [1].
R_s	-0.1946	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, la densidad relativa del aceite en el segundo punto de la región saturada disminuye en una media de 0.1946 [1].
$\mu_o@20^\circ C$	0.0388	Este coeficiente indica que, para cada unidad adicional de viscosidad, la densidad relativa del aceite en el segundo punto de la región saturada aumenta en una media de 0.0388 [1].

En la Tabla 6.41 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $\rho_{ro\ sat}$ en el segundo punto de presión propuesto.

Tabla 6.41: Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro\ sat}$ en $\frac{2}{5}P_b$.

Indicador	Valor
E_a	0.0204
R^2	0.9242

Tercer punto ($\frac{3}{5}P_b$)

$$\rho'_{ro\ sat\ @\ \frac{3}{5}P_b} = -0.26338T_y + 0.38907\rho_{ro} - 0.24570R_s + 0.04755\mu_o@20^\circ C + 0.60618, \quad (6.16)$$

Tabla 6.42: Coeficientes de la regresión lineal para estimación de $\rho_{ro\ sat}$ en $\frac{3}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.60618	Corresponde a la ordenada al origen, es decir, es el valor del tercer punto de la densidad relativa del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.26338	Este coeficiente indica que, para cada unidad adicional de temperatura, la densidad relativa del aceite en el tercer punto de la región saturada disminuye en una media de 0.26338 [1].
ρ_{ro}	0.38907	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, la densidad relativa del aceite en el tercer punto de la región saturada aumenta en una media de 0.38907 [1].
R_s	-0.24570	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, la densidad relativa del aceite en el tercer punto de la región saturada disminuye en una media de 0.24570 [1].
$\mu_o@20^\circ C$	0.04755	Este coeficiente indica que, para cada unidad adicional de viscosidad, la densidad relativa del aceite en el tercer punto de la región saturada aumenta en una media de 0.04755 [1].

En la Tabla 6.43 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $\rho_{ro\ sat}$ en el tercer punto de presión propuesto.

Tabla 6.43: Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro\ sat}$ en $\frac{3}{5}P_b$.

Indicador	Valor
E_a	0.0298
R^2	0.9319

Cuarto punto ($\frac{4}{5}P_b$)

$$\rho'_{sat @ \frac{4}{5}P_b} = -0.27755T_y + 0.39801\rho_{ro} - 0.35589R_s + 0.04574\mu_o@20^\circ C + 0.61174, \quad (6.17)$$

Tabla 6.44: Coeficientes de la regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{4}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.61174	Corresponde a la ordenada al origen, es decir, es el valor del cuarto punto de la densidad relativa del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.27755	Este coeficiente indica que, para cada unidad adicional de temperatura, la densidad relativa del aceite en el cuarto punto de la región saturada disminuye en una media de 0.27755 [1].
ρ_{ro}	0.39801	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, la densidad relativa del aceite en el cuarto punto de la región saturada aumenta en una media de 0.39801 [1].
R_s	-0.35589	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, la densidad relativa del aceite en el cuarto punto de la región saturada disminuye en una media de 0.35589 [1].
$\mu_o@20^\circ C$	0.04574	Este coeficiente indica que, para cada unidad adicional de viscosidad, la densidad relativa del aceite en el cuarto punto de la región saturada aumenta en una media de 0.04574 [1].

En la Tabla 6.45 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $\rho_{ro sat}$ en el cuarto punto de presión propuesto.

Tabla 6.45: Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{4}{5}P_b$.

Indicador	Valor
E_a	0.0315
R^2	0.9503

Quinto punto (P_b)

$$\rho'_{ro\ sat\ @\ P_b} = -0.28408T_y + 0.42414\rho_{ro} - 0.54034R_s + 0.02232\mu_o@20^\circ C + 0.62248, \quad (6.18)$$

Tabla 6.46: Coeficientes de la regresión lineal para estimación de $\rho_{ro\ sat}$ en P_b .

Variable	Coefficiente	Descripción
Intersección	0.62248	Corresponde a la ordenada al origen, es decir, es el valor del quinto punto de la densidad relativa del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.28408	Este coeficiente indica que, para cada unidad adicional de temperatura, la densidad relativa del aceite en el quinto punto de la región saturada disminuye en una media de 0.28408 [1].
ρ_{ro}	0.42414	Este coeficiente indica que, para cada unidad adicional de densidad del aceite, la densidad relativa del aceite en el quinto punto de la región saturada aumenta en una media de 0.42414 [1].
R_s	-0.54034	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, la densidad relativa del aceite en el quinto punto de la región saturada disminuye en una media de 0.54034 [1].
$\mu_o@20^\circ C$	0.02232	Este coeficiente indica que, para cada unidad adicional de viscosidad, la densidad relativa del aceite en el quinto punto de la región saturada aumenta en una media de 0.02232 [1].

En la Tabla 6.47 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $\rho_{ro\ sat}$ en el quinto punto de presión propuesto.

Tabla 6.47: Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro\ sat}$ en P_b .

Indicador	Valor
E_a	0.0432
R^2	0.9549

6.6.3. Estimación de ρ_{ro} en la región saturada mediante SVR

Para el modelo de SVM, se hizo uso de la función kernel radial para entrenar dicho modelo. En este caso, se obtuvieron los indicadores de error mostrados a continuación al evaluar $\rho_{ro\ sat}$ en los cinco puntos de presión con el conjunto de datos de prueba.

Tabla 6.48: Indicadores del algoritmo de SVR para estimación de $\rho_{ro\ sat}$.

Punto	Indicador	Valor
$\frac{1}{5}P_b$	E_a	0.0244
	R^2	0.8995
$\frac{2}{5}P_b$	E_a	0.0297
	R^2	0.9037
$\frac{3}{5}P_b$	E_a	0.0281
	R^2	0.9306
$\frac{4}{5}P_b$	E_a	0.0266
	R^2	0.9569
$\frac{5}{5}P_b$	E_a	0.0377
	R^2	0.9633

6.6.4. Estimación de ρ_o región saturada mediante redes neuronales artificiales

La estructura de red propuesta para estimar ρ_{ro} en su región saturada en los 5 puntos de presión planteados es la siguiente:

1. Capa de entrada: 4 neuronas correspondientes a las propiedades de entrada para calcular $\rho_{ro\ sat}$.
2. Dos capas ocultas de cuatro y tres neuronas, respectivamente.
3. Una capa de salida con cinco neuronas, cada una correspondiente al valor de $\rho_{ro\ sat}$ en cada punto de presión propuesto.

La función de activación utilizada en todas las neuronas fue sigmoïdal.

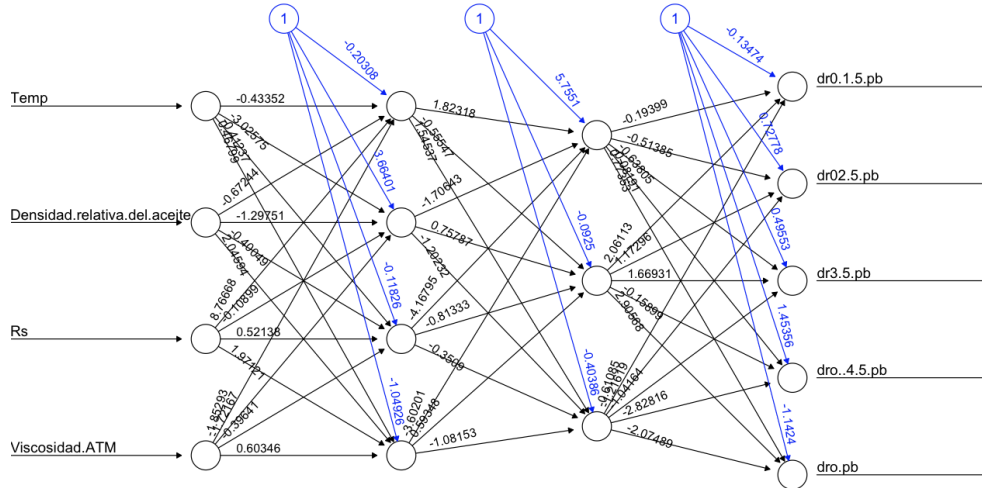


Figura 6.13: Red neuronal generada para estimar el valor de $\rho_{ro sat}$.

A continuación, en la Tabla 6.49, se muestran las métricas obtenidas con el modelo redes neuronales creado para estimar la $\rho_{ro sat}$ en los 5 puntos de presión propuestos.

Tabla 6.49: Indicadores del algoritmo de Redes Neuronales para estimación de $\rho_o sat$.

Punto	Indicador	Valor
$\frac{1}{5} P_b$	E_a	0.0208
	R^2	0.9166
$\frac{2}{5} P_b$	E_a	0.0268
	R^2	0.9119
$\frac{3}{5} P_b$	E_a	0.0282
	R^2	0.9349
$\frac{4}{5} P_b$	E_a	0.0295
	R^2	0.9533
$\frac{5}{5} P_b$	E_a	0.0442
	R^2	0.9542

6.6.5. Comparación de los tres métodos

La Tabla 6.50 muestra una comparación de los tres modelos propuestos para la estimación de la densidad del aceite en su región saturada a partir de los datos de entrada seleccionados. Para realizar dicha comparación, se hizo uso del promedio del porcentaje de error absoluto medio obtenido para los cinco puntos.

Tabla 6.50: Indicadores del algoritmo de Redes Neuronales para estimación de $\rho_{ro\ sat}$.

Algoritmo	Porcentaje de error absoluto medio [%]
Regresión lineal	3.06
SVR	2.93
RNA	2.99

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 2.93 %, compitiendo contra las redes neuronales que obtienen un E_a igual a 2.99 %.

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

6.7.1. Preparación de los datos para $\rho_{ro\ bajosat}$

Conforme al mapa de correlación de la Figura 6.1, 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}).

El conjunto de datos utilizado para el entrenamiento y prueba de los modelos de regresión para calcular $\rho_{ro\ bajosat}$ se compone de una muestra de 88 pozos con las características mostradas en la Tabla 6.52. A partir de esta información, se normalizaron los datos de todo el conjunto de datos utilizado a partir del método mín-máx.

Tabla 6.51: Número de muestras del conjunto de datos utilizado para entrenar y probar los modelos de estimación de ρ_{ro} *bajosat*.

Tipo de yacimiento	Cantidad
Aceites negros	42
Aceites volátiles	46
Total	88

Tabla 6.52: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de ρ_{ro} *bajosat* y normalizar las variables de entrada durante la etapa de procesamiento de datos.

Propiedad	Valor mínimo	Valor máximo	Promedio
$\rho_{ro}[1]$	0.798	0.9332	0.853
$R_s[\frac{m^3}{m^3}]$	11.1	783.1	227
$T_y[^\circ C]$	42.6	162.8	117
$\Delta(\rho_{ro} @(P_b+200) - \rho_{ro} @P_b[kg/cm^2])[\frac{m^3}{m^3}]$	0.0030	0.0767	0.0330

6.7.2. Estimación de ρ_{ro} región bajo saturada mediante regresión lineal

Mediante este algoritmo es posible estimar el valor normalizado de $\rho_{ro} @(P_b + 200) - \rho_{ro} @P_b[kg/cm^2]$ a partir de la temperatura del yacimiento y la relación gas-aceite disuelto con la ecuación

$$(\rho_{ro} @(P_b + 200) - \rho_{ro} @P_b[kg/cm^2])' = 0.08028T_y + 0.63891R_s + 0.16392. \quad (6.19)$$

En la Tabla 6.54 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $(\rho_{ro} @(P_b + 200) - \rho_{ro} @P_b[kg/cm^2])$.

Tabla 6.53: Coeficientes de la regresión lineal para estimación de $(\rho_{ro} @ (P_b + 200) - \rho_{ro} @ P_b [kg/cm^2])$.

Variable	Coficiente	Descripción
Intersección	0.16392	Corresponde a la ordenada al origen, es decir, es el valor de la densidad relativa del aceite en la región bajo saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	0.08028	Este coeficiente indica que, para cada unidad adicional de temperatura, la densidad relativa del aceite en la región bajo saturada aumenta en una media de 0.08028 [1].
R_s	0.63891	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite, la densidad relativa del aceite en la región bajo saturada aumenta en una media de 0.63891 [1].

Tabla 6.54: Indicadores del algoritmo de regresión lineal para estimación de $\rho_{ro sat}$ en $\frac{1}{5}P_b$.

Indicador	Valor
E_a	0.1238
R^2	0.9509

6.7.3. Estimación de ρ_{ro} región bajo saturada mediante SVR

El modelo de SVR fue entrenado con la función Kernel radial. Con este algoritmo, las métricas de error obtenidas fueron las mostradas en la Tabla 6.55.

Tabla 6.55: Indicadores del algoritmo de SVR para estimación de $(\rho_{ro} @ (P_b + 200) - \rho_{ro} @ P_b [kg/cm^2])$.

Indicador	Valor
E_a	0.1230
R^2	0.9622

6.7.4. Estimación de $\rho_{ro bajosat}$ mediante redes neuronales

El modelo de red neuronal artificial con el algoritmo de retropropagación. En este caso, se propuso la siguiente estructura de red:

1. Capa de entrada: 3 neuronas correspondientes a las propiedades de entrada para calcular $(\rho_{ro} @ (P_b + 200) - \rho_{ro} @ P_b [kg/cm^2])$.
2. Dos capas ocultas de seis neuronas cada una.
3. Una capa de salida con una neurona equivalente al valor de $(\rho_{ro} @ (P_b + 200) - \rho_{ro} @ P_b [kg/cm^2])$ calculado.

La función de activación utilizada en todas las neuronas fue sigmoideal.

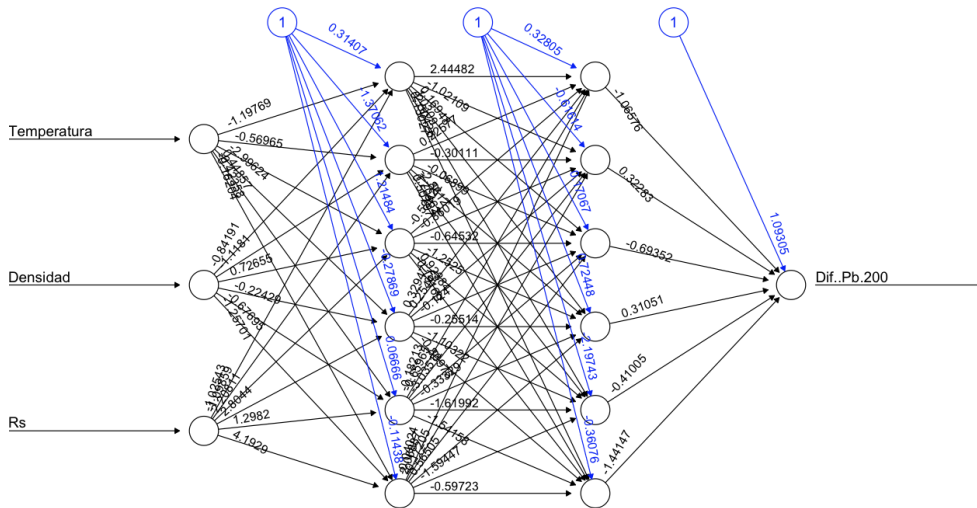


Figura 6.14: Red neuronal generada para estimar el valor de $(\rho_{ro} @ (P_b + 200) - \rho_{ro} @ P_b [kg/cm^2])$.

En la Tabla 6.56 se muestran las métricas de error obtenidas con el modelo redes neuronales creado para estimar la $(\rho_{ro} @ (P_b + 200) - \rho_{ro} @ P_b [kg/cm^2])$.

Tabla 6.56: Indicadores del algoritmo de Redes Neuronales para estimación de $(\rho_{ro} @ (P_b + 200) - \rho_{ro} @ P_b [kg/cm^2])$.

Indicador	Valor
E_a	0.1421
R^2	0.9253

6.7.5. Comparación de los tres métodos

Una vez obtenido el valor de $(\rho_{ro} @ (P_b + 200) - \rho_{ro} @ P_b [kg/cm^2])$, es necesario normalizar el resultado llevándolo a condiciones de presión de yacimiento (ρ_{roy}) . En la Tabla

6.57, se muestra una comparación del porcentaje de error absoluto medio obtenido con los valores de ρ_{ro} normalizados a condiciones de presión de yacimiento.

Tabla 6.57: Indicadores del algoritmo de Redes Neuronales para estimación de $\rho_{ro sat}$.

Algoritmo	Porcentaje de error absoluto medio [%]
Regresión lineal	2.32
SVR	2.47
RNA	2.13

En la Tabla 6.57, se puede observar que para estimar el ρ_{roy} , el modelo de aprendizaje automático que menos error obtiene con el conjunto de pruebas es la red neuronal artificial con un porcentaje de error absoluto medio del 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 Tabla 6.58.

Tabla 6.58: Comparación de resultados obtenidos con el conjunto de datos de prueba para ρ_{roy}

Pozo	$\rho_{roy}PVT[1]$	$\rho_{roy}Reg.Lineal[1]$	$\rho_{roy}SVR[1]$	$\rho_{roy}RNA[1]$
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

6.7.6. Generación de la curva completa 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} .

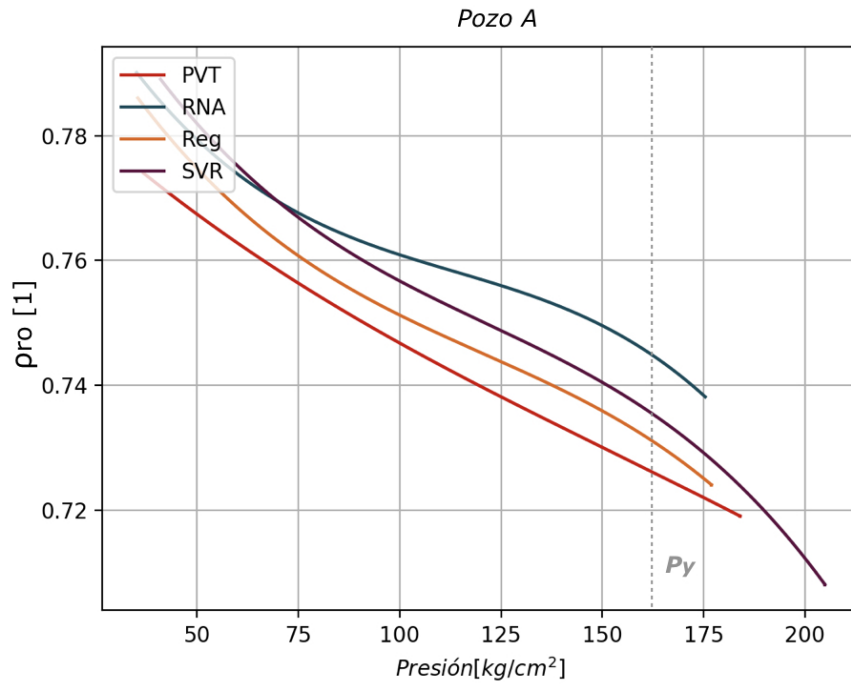


Figura 6.15: Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo A.

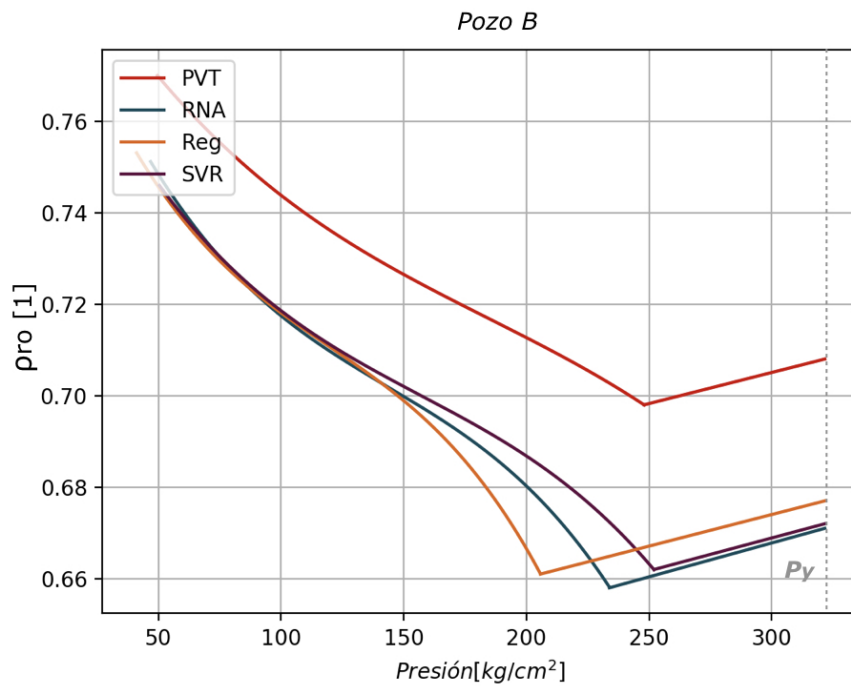


Figura 6.16: Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo B.

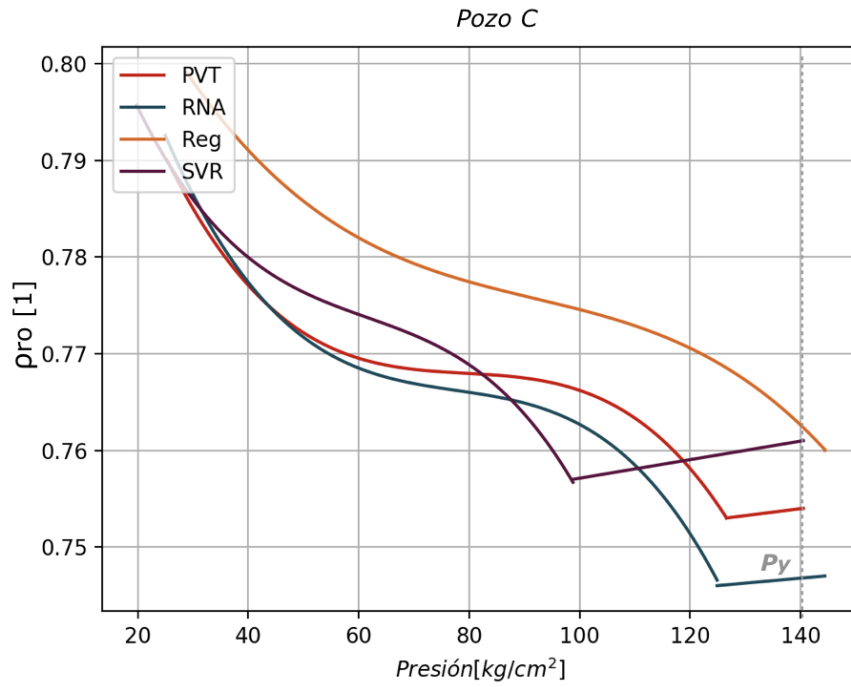


Figura 6.17: Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo C.

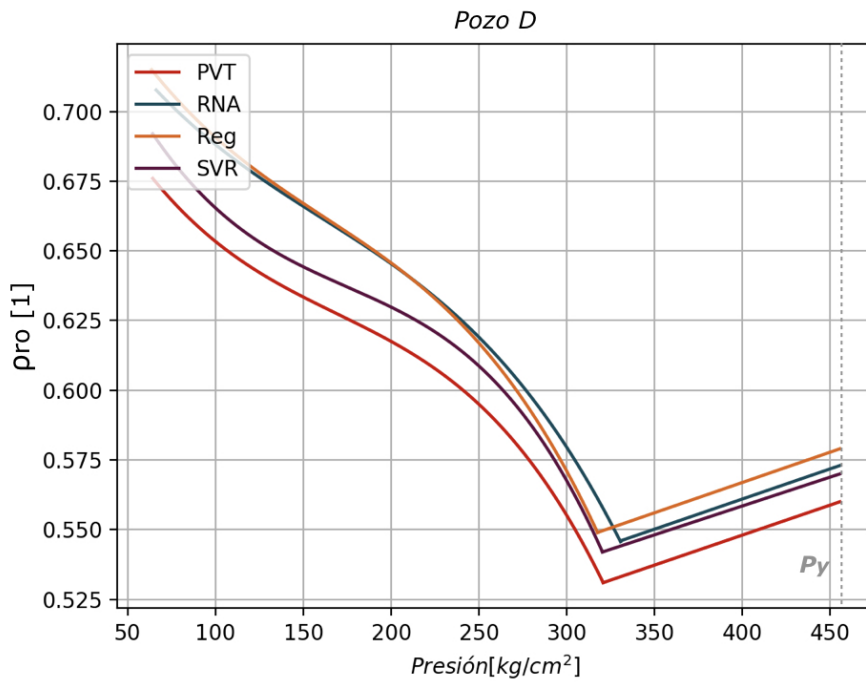


Figura 6.18: Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo D.

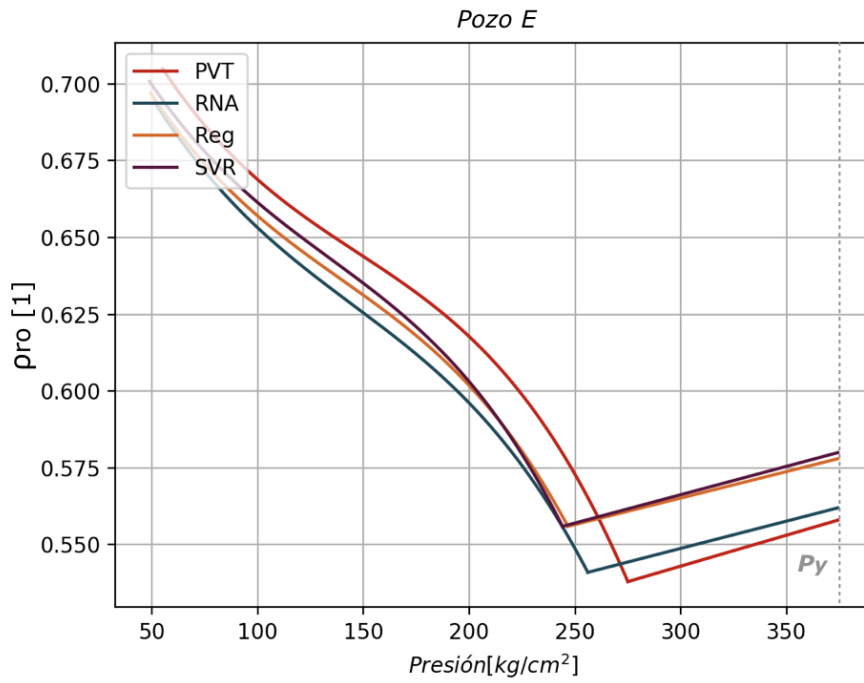


Figura 6.19: Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo E.

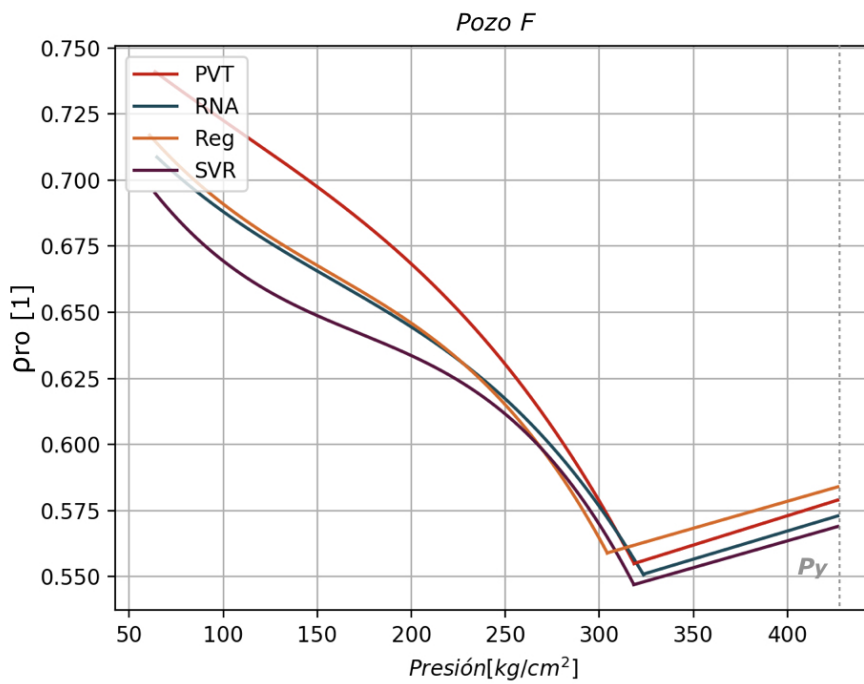


Figura 6.20: Curva de ρ_{ro} estimada con los tres métodos de aprendizaje automático para el pozo F.

Con las Figuras 6.15, 6.16, 6.17, 6.18, 6.19 y 6.20 es posible observar de manera gráfica que, para estimar 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.

6.8. Estimación de R_s

6.8.1. Preparación de los datos para R_s

Se utilizó el mapa de correlación mostrado en la Figura 6.1 para elegir como entradas para los algoritmos de aprendizaje automático las propiedades con valores de correlación más alto. En este caso, las variables a utilizar son la temperatura del pozo (T_y), la relación gas-aceite disuelto en el punto de burbuja (R_{sb}) y la densidad relativa del aceite (ρ_{ro}).

El conjunto de datos utilizado para el entrenamiento, prueba y validación de los modelos de regresión para calcular R_s se compone de muestras de 101 pozos con las características mostradas en la Tabla 6.60. A partir de estos datos, se normalizaron todos los datos de con el método mín-máx.

Tabla 6.59: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de R_s .

Tipo de yacimiento	Cantidad
Aceites negros	51
Aceites volátiles	50
Total	101

Tabla 6.60: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de R_s y normalizar las variables de entrada durante la etapa de procesamiento de datos.

Propiedad	Valor mínimo	Valor máximo	Promedio
$\rho_{ro}[1]$	0.798	0.9332	0.855
$R_{s@P_b}[\frac{m^3}{m^3}]$	11.1	783.1	221.8
$T_y[^\circ C]$	43.6	162.8	116.5

6.8.2. Estimación de R_s mediante 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.

Primer punto ($\frac{1}{5}P_b$)

$$R'_s @ \frac{1}{5}P_b = -0.08053T_y - 0.09973\rho_{ro} + 1.06831R_s + 0.27982 \quad (6.20)$$

Tabla 6.61: Coeficientes de la regresión lineal para estimación de R_s en $\frac{1}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.27982	Corresponde a la ordenada al origen, es decir, es el valor del primer punto de la relación de solubilidad gas-aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.08053	Este coeficiente indica que, para cada unidad adicional de temperatura, la relación de solubilidad gas-aceite en el primer punto de la región saturada disminuye en una media de 0.08053 [1].
ρ_o	-0.09973	Este coeficiente indica que, para cada unidad adicional de densidad, la relación de solubilidad gas-aceite en el primer punto de la región saturada disminuye en una media de 0.09973 [1].
$R_{s@P_b}$	1.06831	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite en el punto de burbuja, la relación de solubilidad gas-aceite en el primer punto de la región bajo saturada aumenta en una media de 1.06831 [1].

En la Tabla 6.62 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de R_s en el primer punto de presión propuesto.

Tabla 6.62: Indicadores del algoritmo de regresión lineal para estimación de R_s en $\frac{1}{5}P_b$.

Indicador	Valor
E_a	0.1121
R^2	0.9168

Segundo punto ($\frac{2}{5}P_b$)

$$R'_s @ \frac{2}{5}P_b = -0.01702T_y - 0.11366\rho_{ro} + 1.10395R_s + 0.30148 \quad (6.21)$$

Tabla 6.63: Coeficientes de la regresión lineal para estimación de R_s en $\frac{2}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.30148	Corresponde a la ordenada al origen, es decir, es el valor del segundo punto de la relación de solubilidad gas-aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.01702	Este coeficiente indica que, para cada unidad adicional de temperatura, la relación de solubilidad gas-aceite en el segundo punto de la región saturada disminuye en una media de 0.01702 [1].
ρ_o	-0.11366	Este coeficiente indica que, para cada unidad adicional de densidad, la relación de solubilidad gas-aceite en el segundo punto de la región saturada disminuye en una media de 0.11366 [1].
$R_{s@P_b}$	1.10395	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite en el punto de burbuja, la relación de solubilidad gas-aceite en el segundo punto de la región bajo saturada aumenta en una media de 1.10395 [1].

En la Tabla 6.64 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de R_s en el segundo punto de presión propuesto.

Tabla 6.64: Indicadores del algoritmo de regresión lineal para estimación de R_s en $\frac{2}{5}P_b$.

Indicador	Valor
E_a	0.1079
R^2	0.8868

Tercer punto ($\frac{3}{5}P_b$)

$$R'_s @ \frac{3}{5}P_b = 0.02671T_y - 0.07784\rho_{ro} + 1.07759R_s + 0.20631, \quad (6.22)$$

Tabla 6.65: Coeficientes de la regresión lineal para estimación de R_s en $\frac{3}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.20631	Corresponde a la ordenada al origen, es decir, es el valor del tercer punto de la relación de solubilidad gas-aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	0.02671	Este coeficiente indica que, para cada unidad adicional de temperatura, la relación de solubilidad gas-aceite en el tercer punto de la región saturada aumenta en una media de 0.02671 [1].
ρ_o	-0.07784	Este coeficiente indica que, para cada unidad adicional de densidad, la relación de solubilidad gas-aceite en el tercer punto de la región saturada disminuye en una media de 0.07784 [1].
$R_{s@P_b}$	1.07759	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite en el punto de burbuja, la relación de solubilidad gas-aceite en el tercer punto de la región bajo saturada aumenta en una media de 1.07759 [1].

En la Tabla 6.66 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de R_s en el tercer punto de presión propuesto.

Tabla 6.66: Indicadores del algoritmo de regresión lineal para estimación de R_s en $\frac{3}{5}P_b$.

Indicador	Valor
E_a	0.1221
R^2	0.8801

Cuarto punto ($\frac{4}{5}P_b$)

$$R'_s @ \frac{4}{5}P_b = 0.02223T_y - 0.02847\rho_{ro} + 1.00876R_s + 0.07854 \quad (6.23)$$

En la Tabla 6.68 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de R_s en el cuarto punto de presión propuesto.

Tabla 6.67: Coeficientes de la regresión lineal para estimación de R_s en $\frac{4}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.07854	Corresponde a la ordenada al origen, es decir, es el valor del cuarto punto de la relación de solubilidad gas-aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	0.02223	Este coeficiente indica que, para cada unidad adicional de temperatura, la relación de solubilidad gas-aceite en el cuarto punto de la región saturada aumenta en una media de 0.02223 [1].
ρ_o	-0.02847	Este coeficiente indica que, para cada unidad adicional de densidad, la relación de solubilidad gas-aceite en el cuarto punto de la región saturada disminuye en una media de 0.02847 [1].
$R_{s@P_b}$	1.00876	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite en el punto de burbuja, la relación de solubilidad gas-aceite en el cuarto punto de la región bajo saturada aumenta en una media de 1.00876 [1].

 Tabla 6.68: Indicadores del algoritmo de regresión lineal para estimación de R_s en $\frac{4}{5}P_b$.

Indicador	Valor
E_a	0.0831
R^2	0.9581

6.8.3. Estimación de R_s mediante SVR

El modelo de SVR para estimar la curva de R_s fue entrenado con la función kernel radial. A continuación se muestran las métricas de error obtenidas al evaluar R_s en los cuatro puntos de presión con el conjunto de datos de prueba.

6.8.4. Estimación de R_s mediante redes neuronales artificiales

La estructura de red propuesta para estimar R_s en los cuatro puntos de presión planteados es la siguiente:

1. Capa de entrada: 3 neuronas correspondientes a las propiedades de entrada para calcular la curva de R_s .

Tabla 6.69: Indicadores del algoritmo de SVR para estimación de R_s .

Punto	Indicador	Valor
$\frac{1}{5}P_b$	E_a	0.0934
	R^2	0.9577
$\frac{2}{5}P_b$	E_a	0.0667
	R^2	0.9666
$\frac{3}{5}P_b$	E_a	0.0648
	R^2	0.9614
$\frac{4}{5}P_b$	E_a	0.0599
	R^2	0.9826

2. Dos capas ocultas de cuatro y tres neuronas, respectivamente.
3. Una capa de salida con cuatro neuronas, cada una equivalente al valor de R_s en cada punto de presión propuesto.

La función de activación utilizada en todas las neuronas fue sigmoideal.

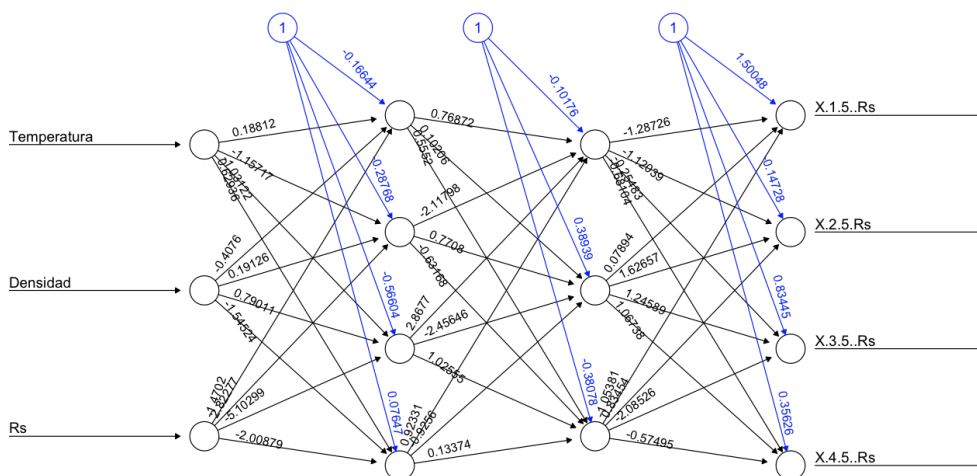


Figura 6.21: Red neuronal generada para estimar los valores de R_s .

A continuación, en la Tabla 6.70, se muestran las métricas obtenidas con el modelo redes neuronales creado para estimar la R_s en los 4 puntos de presión propuestos.

6.8.5. Comparación de los tres métodos

La Tabla 6.71 muestra una comparación de los tres modelos propuestos para la estimación de R_s a partir de los datos de entrada seleccionados. Para realizar dicha

Tabla 6.70: Indicadores del algoritmo de Redes Neuronales para estimación de R_S .

Punto	Indicador	Valor
$\frac{1}{5}P_b$	E_a	0.1096
	R^2	0.9311
$\frac{2}{5}P_b$	E_a	0.0722
	R^2	0.9571
$\frac{3}{5}P_b$	E_a	0.0573
	R^2	0.9534
$\frac{4}{5}P_b$	E_a	0.0679
	R^2	0.9783

comparación, se hizo uso del promedio del porcentaje de error absoluto medio obtenido para los cinco puntos.

Tabla 6.71: Error absoluto medio obtenido con los tres algoritmos utilizados para calcular R_S .

Algoritmo	Porcentaje de error absoluto medio [%]
Regresión lineal	10.63
SVR	7.12
RNA	7.68

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%.

6.8.6. Generación de la curva de R_s

La Tabla 6.72 muestra las características del conjunto de datos de validación, el cual está hecho a partir de una muestra tomada del conjunto de datos de prueba.

Tabla 6.72: Conjunto de datos de prueba para R_s

Pozo	$T_y [^{\circ}C]$	$\rho [1]$	$R_s [m^3/m^3]$	$P_b [kg/cm^2]$
A	127	0.8995	96.397	184
B	102	0.8567	175.7	248
C	47	0.8596	80.6	126.6
D	126	0.8514	401.1	320.7
E	147	0.8338	233.466	275
F	124	0.8497	373.7	318.5

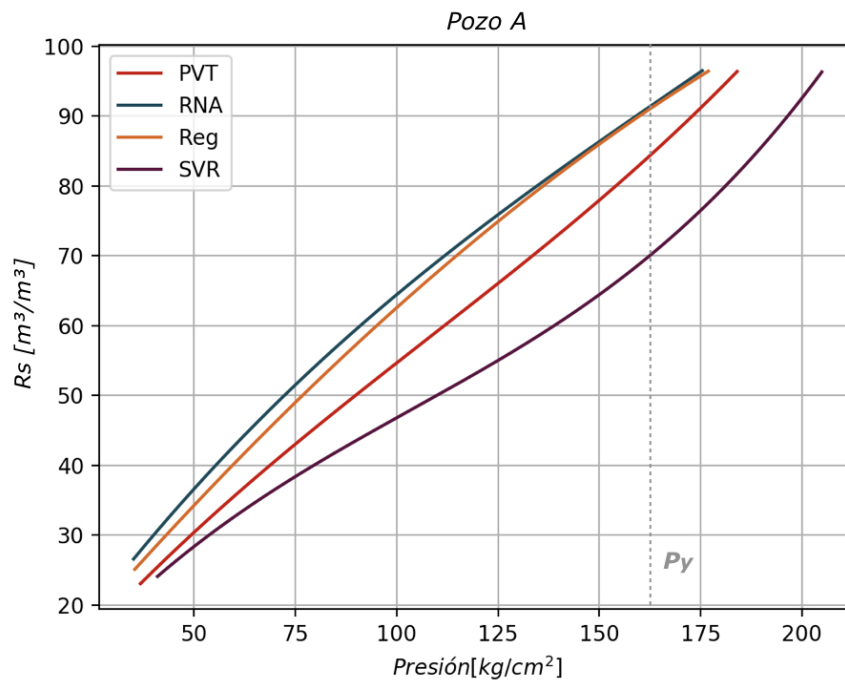


Figura 6.22: Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo A.

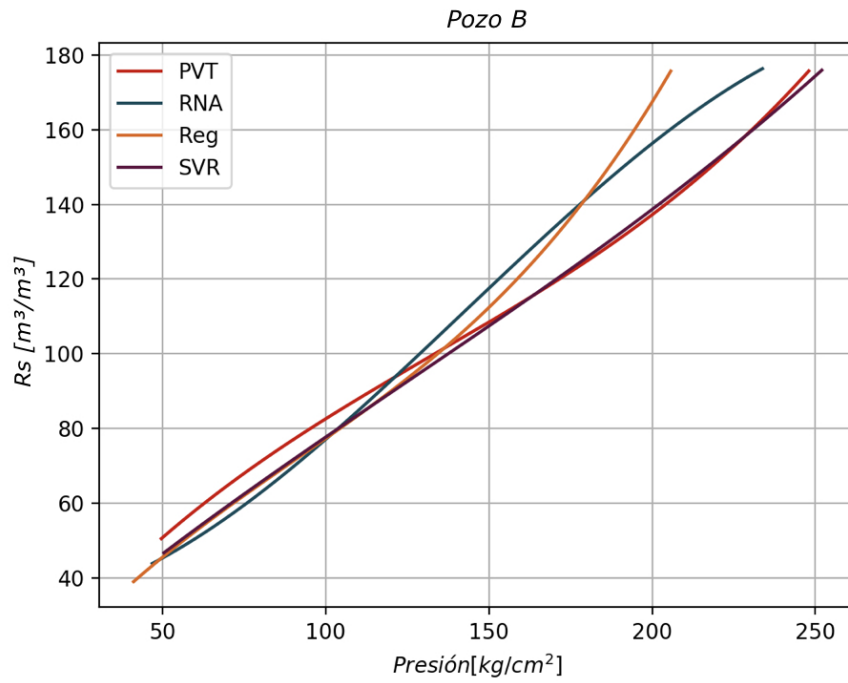


Figura 6.23: Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo B.

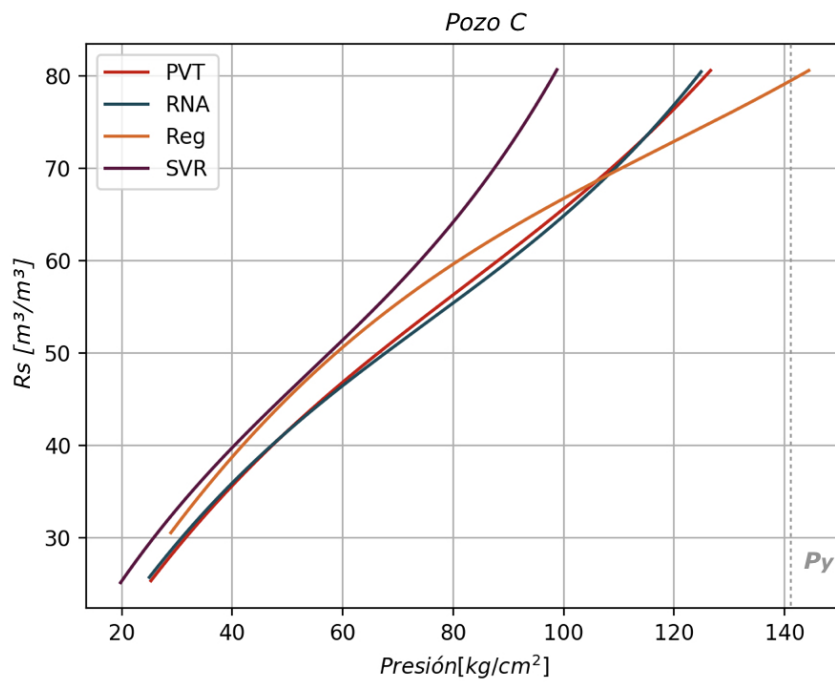


Figura 6.24: Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo C.

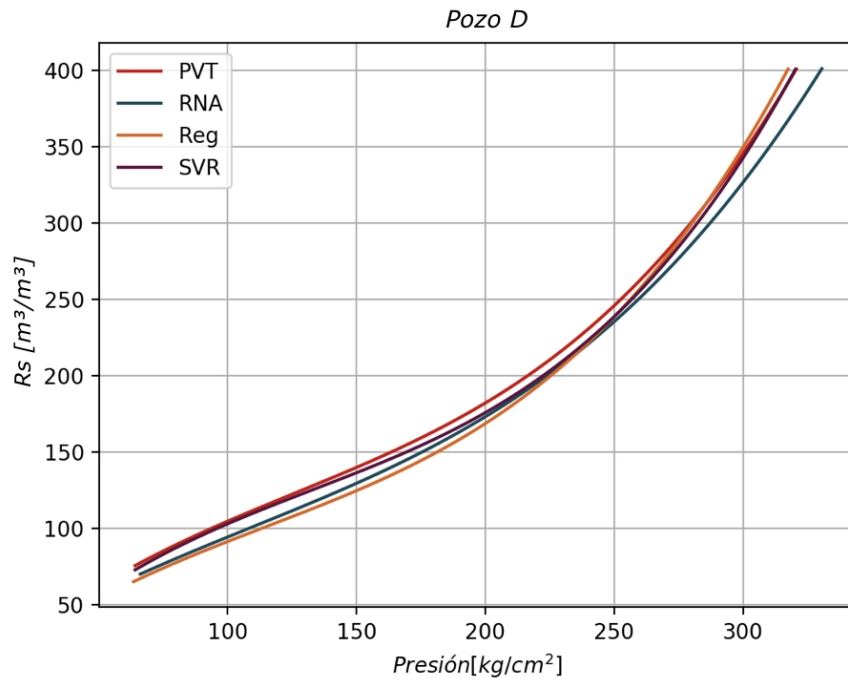


Figura 6.25: Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo D.

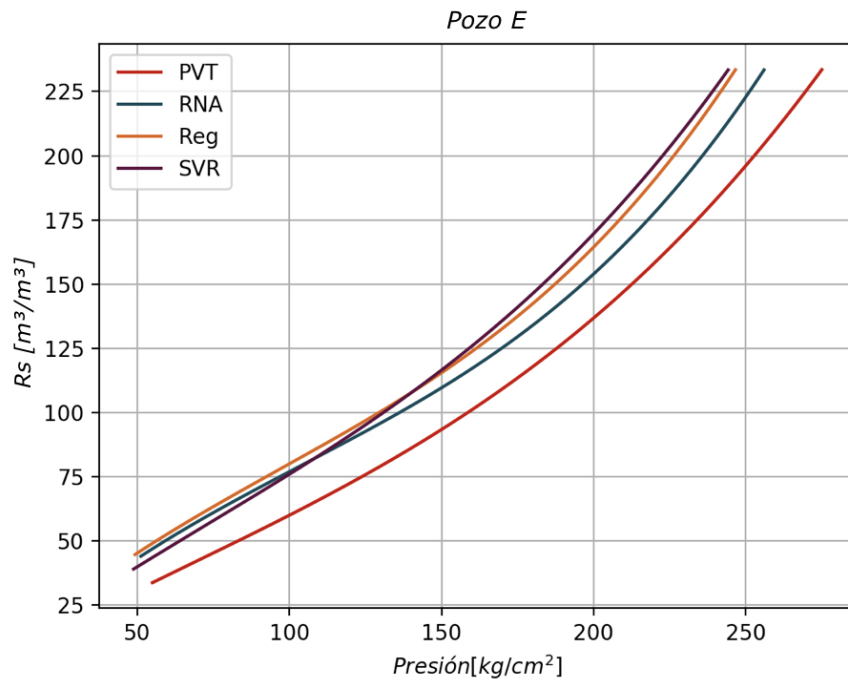


Figura 6.26: Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo E.

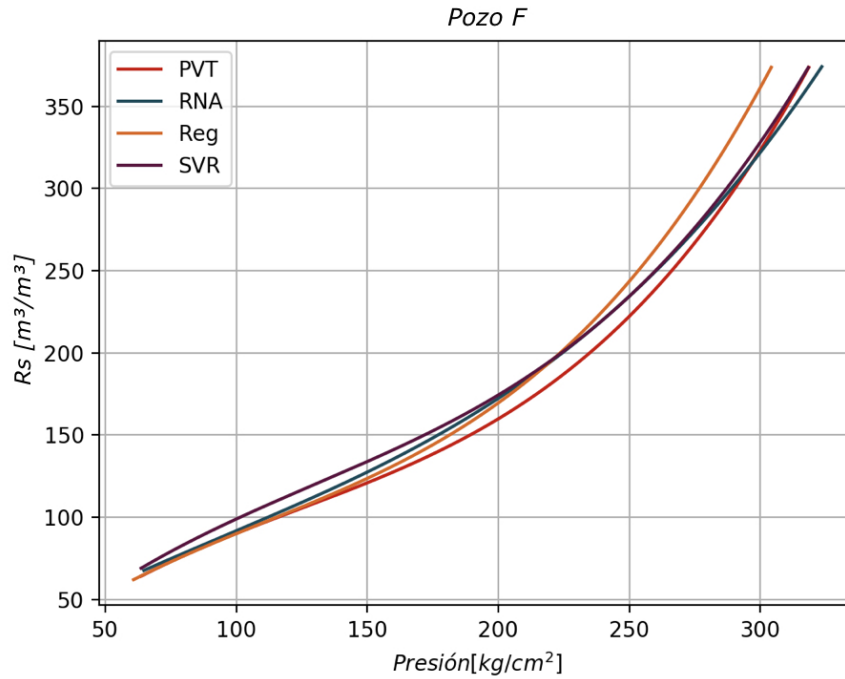


Figura 6.27: Curva de R_s estimada con los tres métodos de aprendizaje automático para el pozo F.

Las Figuras 6.22, 6.23, 6.24, 6.25, 6.26 y 6.27 nos permiten observar de manera gráfica las curvas generadas por la unión de los resultados obtenidos con los modelos de R_s en este trabajo. Con las figuras anteriores, 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.

6.9. Estimación de μ_o en la región saturada

6.9.1. Preparación de los datos para $\mu_o \text{ sat}$

Con los resultados obtenidos en el mapa de correlación mostrado en la Figura 6.1, 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), la densidad relativa del aceite (ρ_{ro}) y la viscosidad del aceite a $20^\circ C$ ($\mu_o @ 20^\circ C$).

El conjunto de datos utilizado para el entrenamiento, prueba y validación de los modelos de regresión para calcular μ_o en su región saturada está conformado por muestras de 92 pozos con las características mostradas en la Tabla 6.74. A partir de esta información, se normalizaron los datos de todo el conjunto de datos utilizado con el método mín-máx.

Tabla 6.73: Número de muestras del conjunto de datos utilizado para entrenar y validar los modelos de estimación de μ_o sat.

Tipo de yacimiento	Cantidad
Aceites negros	46
Aceites volátiles	46
Total	92

Tabla 6.74: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de μ_o sat y normalizar las variables de entrada durante la etapa de procesamiento de datos.

Propiedad	Valor mínimo	Valor máximo	Promedio
$\rho_{ro}[1]$	0.798	0.933	0.854
$R_s[\frac{m^3}{m^3}]$	11.1	503.3	220.1
$T_y[^\circ C]$	43.6	160	116.9
$\mu_o @ 20^\circ C[cP]$	2.84	302.74	19.35
$\mu_{ob}[cP]$	0.07	14.19	0.83

6.9.2. Estimación de μ_o en la región saturada mediante regresión lineal

Con este algoritmo se estimaron los valores normalizados de μ_o en los 5 puntos de presión propuestos, haciendo uso de las 4 entradas mencionadas anteriormente, a partir de las expresiones mostradas a continuación.

Primer punto ($\frac{1}{5}P_b$)

$$\mu'_{o\ sat} @ \frac{1}{5}P_b = -0.04988T_y + 0.04280\rho_{ro} - 0.01361R_s + 0.56470\mu_o@20^\circ C + 0.03933 \quad (6.24)$$

Tabla 6.75: Coeficientes de la regresión lineal para estimación de $\mu_{o\ sat}$ en $\frac{1}{5}P_b$.

Variable	Coficiente	Descripción
Intersección	0.03933	Corresponde a la ordenada al origen, es decir, es el valor del primer punto de la viscosidad del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.04988	Este coeficiente indica que, para cada unidad adicional de temperatura, la viscosidad del aceite en el primer punto de la región saturada disminuye en una media de 0.04988 [1].
ρ_o	0.04280	Este coeficiente indica que, para cada unidad adicional de densidad, la viscosidad del aceite en el primer punto de la región saturada aumenta en una media de 0.04280 [1].
$R_{s@P_b}$	-0.01361	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite en el punto de burbuja, la viscosidad del aceite en el primer punto de la región bajo saturada disminuye en una media de 0.01361 [1].
$\mu_o@20^\circ C$	0.56470	Este coeficiente indica que, para cada unidad adicional de viscosidad a $20^\circ C$, la viscosidad del aceite en el primer punto de la región bajo saturada aumenta en una media de 0.56470 [1].

En la Tabla 6.76 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $\mu_{o\ sat}$ en el primer punto de presión propuesto.

Tabla 6.76: Indicadores del algoritmo de regresión lineal para estimación de $\mu_{o\ sat}$ en $\frac{1}{5}P_b$.

Indicador	Valor
E_a	0.6284
R^2	0.7623

Segundo punto ($\frac{2}{5}P_b$)

$$\mu'_{o \text{ sat}} @ \frac{2}{5}P_b = -0.05330T_y + 0.03840\rho_{ro} - 0.02092R_s + 0.55183\mu_o@20^\circ C + 0.04692 \quad (6.25)$$

Tabla 6.77: Coeficientes de la regresión lineal para estimación de $\mu_{o \text{ sat}}$ en $\frac{2}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.04692	Corresponde a la ordenada al origen, es decir, es el valor del segundo punto de la viscosidad del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.05330	Este coeficiente indica que, para cada unidad adicional de temperatura, la viscosidad del aceite en el segundo punto de la región saturada disminuye en una media de 0.05330 [1].
ρ_o	0.03840	Este coeficiente indica que, para cada unidad adicional de densidad, la viscosidad del aceite en el segundo punto de la región saturada aumenta en una media de 0.03840 [1].
$R_{s@P_b}$	-0.02092	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite en el punto de burbuja, la viscosidad del aceite en el segundo punto de la región bajo saturada disminuye en una media de 0.02092 [1].
$\mu_o@20^\circ C$	0.55183	Este coeficiente indica que, para cada unidad adicional de viscosidad a $20^\circ C$, la viscosidad del aceite en el segundo punto de la región bajo saturada aumenta en una media de 0.55183 [1].

En la Tabla 6.78 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $\mu_{o \text{ sat}}$ en el segundo punto de presión propuesto.

Tabla 6.78: Indicadores del algoritmo de regresión lineal para estimación de $\mu_{o \text{ sat}}$ en $\frac{2}{5}P_b$.

Indicador	Valor
E_a	0.6387
R^2	0.7490

Tercer punto ($\frac{3}{5}P_b$)

$$\mu'_{o \text{ sat}} @ \frac{3}{5}P_b = -0.05576T_y + 0.04457\rho_{ro} - 0.03261R_s + 0.55512\mu_o@20^\circ C + 0.05360 \quad (6.26)$$

Tabla 6.79: Coeficientes de la regresión lineal para estimación de $\mu_{o \text{ sat}}$ en $\frac{3}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.05360	Corresponde a la ordenada al origen, es decir, es el valor del tercer punto de la viscosidad del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.05576	Este coeficiente indica que, para cada unidad adicional de temperatura, la viscosidad del aceite en el tercer punto de la región saturada disminuye en una media de 0.05576 [1].
ρ_o	0.04457	Este coeficiente indica que, para cada unidad adicional de densidad, la viscosidad del aceite en el tercer punto de la región saturada aumenta en una media de 0.04457 [1].
$R_{s@P_b}$	-0.03261	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite en el punto de burbuja, la viscosidad del aceite en el tercer punto de la región bajo saturada disminuye en una media de 0.03261 [1].
$\mu_o@20^\circ C$	0.55512	Este coeficiente indica que, para cada unidad adicional de viscosidad a 20 °C, la viscosidad del aceite en el tercer punto de la región bajo saturada aumenta en una media de 0.55512 [1].

En la Tabla 6.80 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $\mu_{o \text{ sat}}$ en el tercer punto de presión propuesto.

Tabla 6.80: Indicadores del algoritmo de regresión lineal para estimación de $\mu_{o \text{ sat}}$ en $\frac{3}{5}P_b$.

Indicador	Valor
E_a	0.6522
R^2	0.7574

Cuarto punto ($\frac{4}{5}P_b$)

$$\mu'_{o\ sat\ @\ \frac{4}{5}P_b} = -0.05837T_y + 0.04399\rho_{ro} - 0.03887R_s + 0.55392\mu_o@20^\circ C + 0.05839 \quad (6.27)$$

Tabla 6.81: Coeficientes de la regresión lineal para estimación de $\mu_o\ sat$ en $\frac{4}{5}P_b$.

Variable	Coefficiente	Descripción
Intersección	0.05839	Corresponde a la ordenada al origen, es decir, es el valor del cuarto punto de la viscosidad del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.05837	Este coeficiente indica que, para cada unidad adicional de temperatura, la viscosidad del aceite en el cuarto punto de la región saturada disminuye en una media de 0.05837 [1].
ρ_o	0.04399	Este coeficiente indica que, para cada unidad adicional de densidad, la viscosidad del aceite en el cuarto punto de la región saturada aumenta en una media de 0.04399 [1].
$R_{s@P_b}$	-0.03887	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite en el punto de burbuja, la viscosidad del aceite en el cuarto punto de la región bajo saturada disminuye en una media de 0.03887 [1].
$\mu_o@20^\circ C$	0.55392	Este coeficiente indica que, para cada unidad adicional de viscosidad a 20 °C, la viscosidad del aceite en el cuarto punto de la región bajo saturada aumenta en una media de 0.55392 [1].

En la Tabla 6.82 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $\rho_o\ sat$ en el cuarto punto de presión propuesto.

Tabla 6.82: Indicadores del algoritmo de regresión lineal para estimación de $\mu_o\ sat$ en $\frac{4}{5}P_b$.

Indicador	Valor
E_a	0.6922
R^2	0.7566

Quinto punto (P_b)

$$\rho'_{o \text{ sat @ } P_b} = -0.06127T_y + 0.04330\rho_{ro} - 0.04498R_s + 0.55394\mu_o@20^\circ C + 0.06289 \quad (6.28)$$

Tabla 6.83: Coeficientes de la regresión lineal para estimación de $\mu_{o \text{ sat}}$ en P_b .

Variable	Coeficiente	Descripción
Intersección	0.06289	Corresponde a la ordenada al origen, es decir, es el valor del quinto punto de la viscosidad del aceite en la región saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.06127	Este coeficiente indica que, para cada unidad adicional de temperatura, la viscosidad del aceite en el quinto punto de la región saturada disminuye en una media de 0.06127 [1].
ρ_o	0.04330	Este coeficiente indica que, para cada unidad adicional de densidad, la viscosidad del aceite en el quinto punto de la región saturada aumenta en una media de 0.04330 [1].
$R_{s@P_b}$	-0.04498	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite en el punto de burbuja, la viscosidad del aceite en el quinto punto de la región bajo saturada disminuye en una media de 0.04498 [1].
$\mu_o@20^\circ C$	0.55394	Este coeficiente indica que, para cada unidad adicional de viscosidad a 20 °C, la viscosidad del aceite en el quinto punto de la región bajo saturada aumenta en una media de 0.55394 [1].

En la Tabla 6.84 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $\mu_{o \text{ sat}}$ en el quinto punto de presión propuesto.

Tabla 6.84: Indicadores del algoritmo de regresión lineal para estimación de $\mu_{o \text{ sat}}$ en P_b .

Indicador	Valor
E_a	0.7566
R^2	0.7729

6.9.3. Estimación de μ_o en la región saturada mediante SVR

Para el modelo de SVM, se hizo uso de la función kernel radial para entrenar dicho modelo. En este caso, se obtuvieron los indicadores de error mostrados a continuación al evaluar μ_o *sat* en los cinco puntos de presión con el conjunto de datos de prueba.

Tabla 6.85: Indicadores del algoritmo de SVR para estimación de μ_o *sat*.

Punto	Indicador	Valor
$\frac{1}{5}P_b$	E_a R^2	0.3292 0.9277
$\frac{2}{5}P_b$	E_a R^2	0.4302 0.8673
$\frac{3}{5}P_b$	E_a R^2	0.3433 0.9114
$\frac{3}{5}P_b$	E_a R^2	0.3186 0.9294
$\frac{3}{5}P_b$	E_a R^2	0.2855 0.9587

6.9.4. Estimación de μ_o en la región saturada mediante redes neuronales artificiales

La estructura de red propuesta para estimar μ_o en su región saturada en los 5 puntos de presión planteados es la siguiente:

1. Capa de entrada: 4 neuronas correspondientes a las propiedades de entrada para calcular μ_o *sat*.
2. Dos capas ocultas de cuatro y tres neuronas, respectivamente.
3. Una capa de salida con cinco neuronas, cada una correspondiente al valor de μ_o *sat* en cada punto de presión propuesto.

La función de activación utilizada en todas las neuronas fue sigmoideal.

A continuación, en la Tabla 6.86, se muestran las métricas obtenidas con el modelo redes neuronales creado para estimar la μ_o *sat* en los 5 puntos de presión propuestos.

Tabla 6.86: Indicadores del algoritmo de Redes Neuronales para estimación de $\mu_o sat$.

Punto	Indicador	Valor
$\frac{1}{5}P_b$	E_a	0.4584
	R^2	0.7587
$\frac{2}{5}P_b$	E_a	0.4091
	R^2	0.8787
$\frac{3}{5}P_b$	E_a	0.3647
	R^2	0.9156
$\frac{3}{5}P_b$	E_a	0.2776
	R^2	0.9517
$\frac{3}{5}P_b$	E_a	0.2403
	R^2	0.9610

6.9.5. Comparación de los tres métodos

La Tabla 6.87 muestra una comparación de los tres modelos propuestos para la estimación de la densidad del aceite en su región saturada a partir de los datos de entrada seleccionados. Para realizar dicha comparación, se hizo uso del promedio del porcentaje de error absoluto medio obtenido para los cinco puntos.

Tabla 6.87: Indicadores del algoritmo de Redes Neuronales para estimación de $\mu_o sat$.

Algoritmo	Porcentaje de error absoluto medio [%]
Regresión lineal	67.40
SVR	34.74
RNA	37.00

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%.

6.10. Estimación de μ_o en la región bajo saturada

6.10.1. Preparación de los datos para μ_o *bajosat*

Conforme al mapa de correlación de la Figura 6.1, 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), la densidad relativa del aceite (ρ_{ro}) y la viscosidad relativa del aceite a $20^\circ C$ ($\mu_o @ 20^\circ C$).

El conjunto de datos utilizado para el entrenamiento y prueba de los modelos de regresión para calcular μ_o *bajosat* se compone de una muestra de 99 pozos con las características mostradas en la Tabla 6.89. A partir de esta información, se normalizaron los datos de todo el conjunto de datos utilizado a partir del método mín-máx.

Tabla 6.88: Número de muestras del conjunto de datos utilizado para entrenar y probar los modelos de estimación de μ_o *bajosat*.

Tipo de yacimiento	Cantidad
Aceites negros	51
Aceites volátiles	48
Total	99

Tabla 6.89: Rango de valores de las propiedades del conjunto de datos utilizado para entrenar, validar los modelos de estimación de μ_o *bajosat* y normalizar las variables de entrada durante la etapa de procesamiento de datos.

Propiedad	Valor mínimo	Valor máximo	Promedio
$\rho_{ro}[1]$	0.798	0.933	0.854
$R_s[\frac{m^3}{m^3}]$	11.1	783.1	220
$T_y[^\circ C]$	42.6	162.8	116.1
$(\rho_o @(P_b + 200) - \rho_o @P_b[kg/cm^2])[\frac{m^3}{m^3}]$	0.001	4.972	0.221

6.10.2. Estimación de μ_o región bajo saturada mediante regresión lineal

Mediante este algoritmo es posible estimar el valor normalizado de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$ a partir de la temperatura del yacimiento y la relación gas-aceite disuelto con la ecuación

$$(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])' = -0.07838T_y - 0.15239R_s + 0.13595 \quad (6.29)$$

Tabla 6.90: Coeficientes de la regresión lineal para estimación de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$.

Variable	Coeficiente	Descripción
Intersección	0.13595	Corresponde a la ordenada al origen, es decir, es el valor de la viscosidad del aceite en la región bajo saturada cuando las demás variables predictoras tienen un valor igual a cero.
T_y	-0.07838	Este coeficiente indica que, para cada unidad adicional de temperatura, la viscosidad del aceite en la región bajo saturada disminuye en una media de 0.07838 [1].
$R_{s@P_b}$	-0.15239	Este coeficiente indica que, para cada unidad adicional de relación de solubilidad gas-aceite en el punto de burbuja, la viscosidad del aceite en la región bajo saturada disminuye en una media de 0.15239 [1].

En la Tabla 6.91 se muestran las métricas de error obtenidas con el modelo de regresión lineal para la estimación de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$.

Tabla 6.91: Indicadores del algoritmo de regresión lineal para estimación de $\mu_o sat$ en $\frac{1}{5}P_b$.

Indicador	Valor
E_a	2.9893
R^2	0.7504

6.10.3. Estimación de μ_o región bajo saturada mediante SVR

El modelo de SVR fue entrenado con la función Kernel radial. Con este algoritmo, las métricas de error obtenidas fueron las mostradas en la Tabla 6.92.

Tabla 6.92: Indicadores del algoritmo de SVR para estimación de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$.

Indicador	Valor
E_a	5.3204
R^2	0.7843

6.10.4. Estimación de μ_o *bajosat* mediante redes neuronales

El modelo de red neuronal artificial con el algoritmo de retropropagación. En este caso, se propuso la siguiente estructura de red:

1. Capa de entrada: 3 neuronas correspondientes a las propiedades de entrada para calcular $(\mu_o @ (P_b + 200) - \mu_o @ P_b [m^3/m^3])$.
2. Dos capas ocultas de seis neuronas cada una.
3. Una capa de salida con una neurona equivalente al valor de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [m^3/m^3])$ calculado.

La función de activación utilizada en todas las neuronas fue sigmoideal.

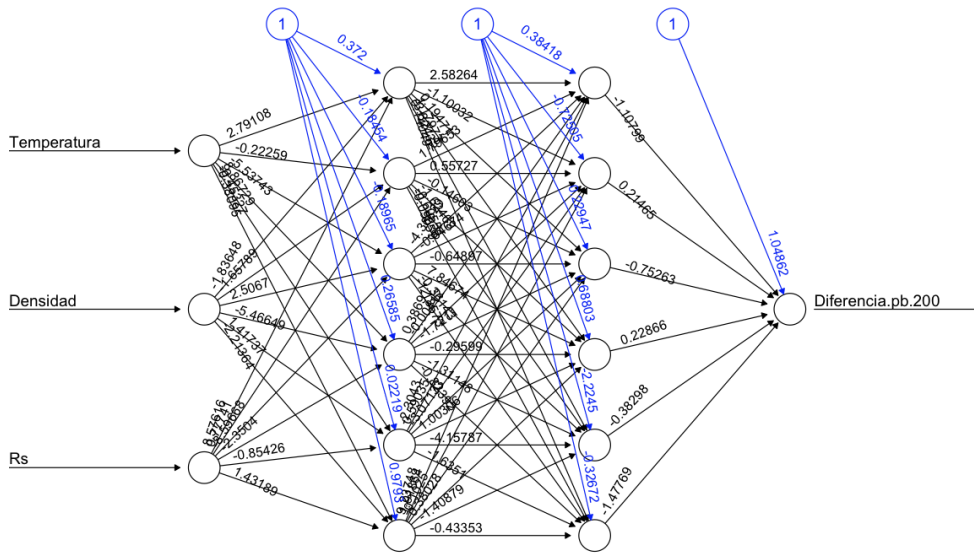


Figura 6.28: Red neuronal generada para estimar el valor de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$.

En la Tabla 6.93 se muestran las métricas de error obtenidas con el modelo redes neuronales creado para estimar $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$.

Tabla 6.93: Indicadores del algoritmo de Redes Neuronales para estimación de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$.

Indicador	Valor
E_a	2.1616
R^2	0.5903

6.10.5. Comparación de los tres métodos

Una vez obtenido el valor de $(\mu_o @ (P_b + 200) - \mu_o @ P_b [kg/cm^2])$, es necesario normalizar el resultado llevándolo a condiciones de presión de yacimiento (μ_{oy}). En la Tabla 6.94, se muestra una comparación del porcentaje de error absoluto medio obtenido con los valores de μ_o normalizado a condiciones de presión de yacimiento.

Tabla 6.94: Indicadores del algoritmo de Redes Neuronales para estimación de μ_o sat.

Algoritmo	Porcentaje de error absoluto medio [%]
Regresión lineal	43.36
SVR	44.74
RNA	21.17

En la Tabla 6.94, se puede observar que para estimar el μ_{oy} , el modelo de aprendizaje automático que menos error obtiene con el conjunto de pruebas es la red neuronal artificial con un porcentaje de error absoluto medio del 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 Tabla 6.95.

Tabla 6.95: Comparación de resultados obtenidos con el conjunto de datos de prueba para μ_{oy}

Pozo	$\mu_{oy}PVT[1]$	$\mu_{oy}Reg.Lineal[1]$	$\mu_{oy}SVR[1]$	$\mu_{oy}RNA[1]$
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

6.10.6. Generación de la curva completa de μ_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 μ_o .

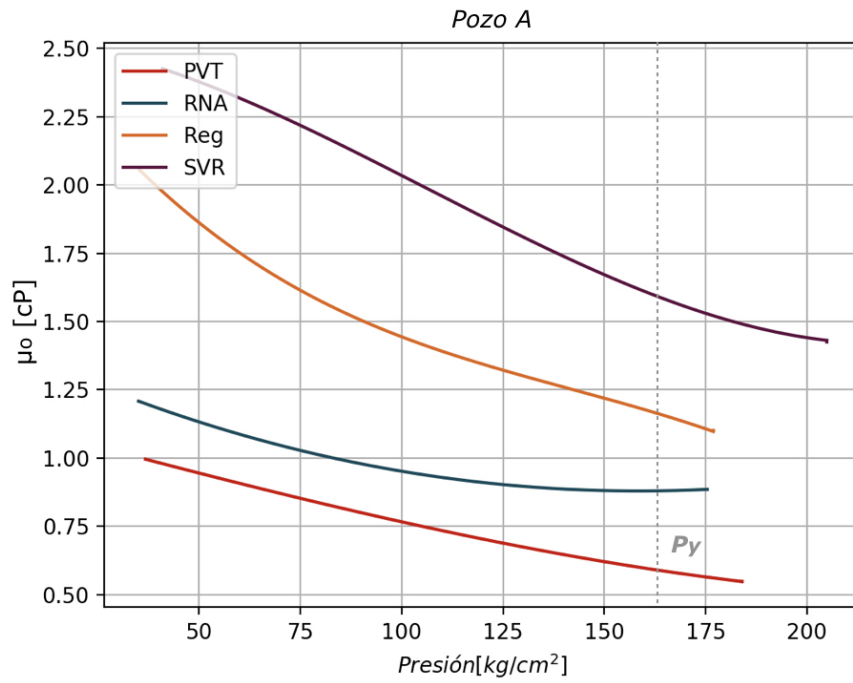


Figura 6.29: Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo A.

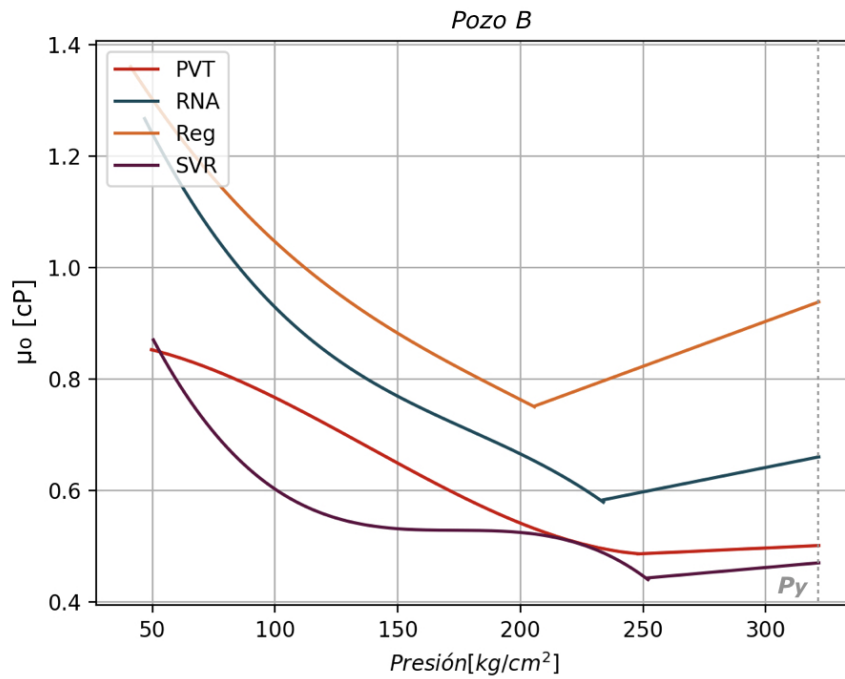


Figura 6.30: Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo B.

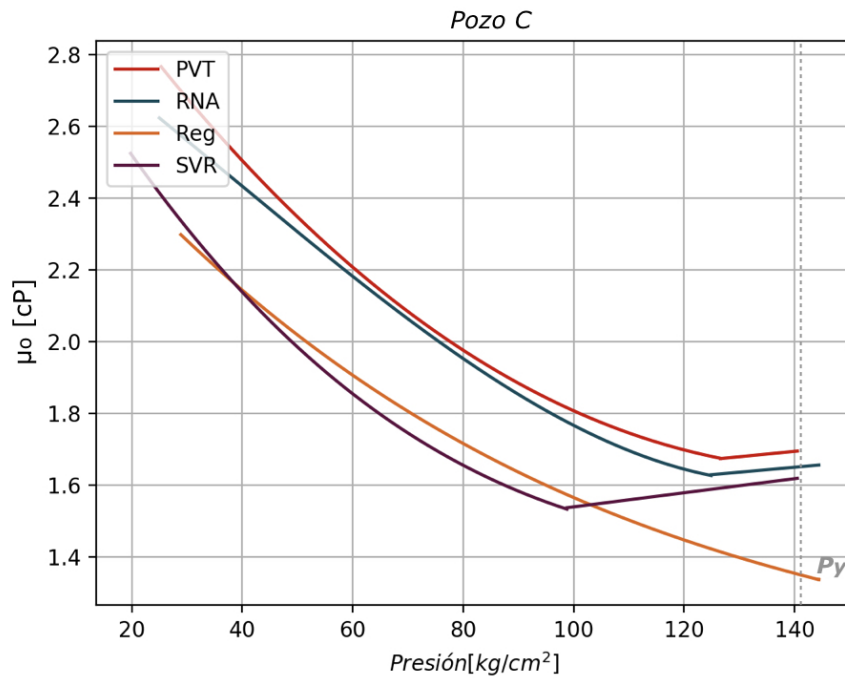


Figura 6.31: Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo C.

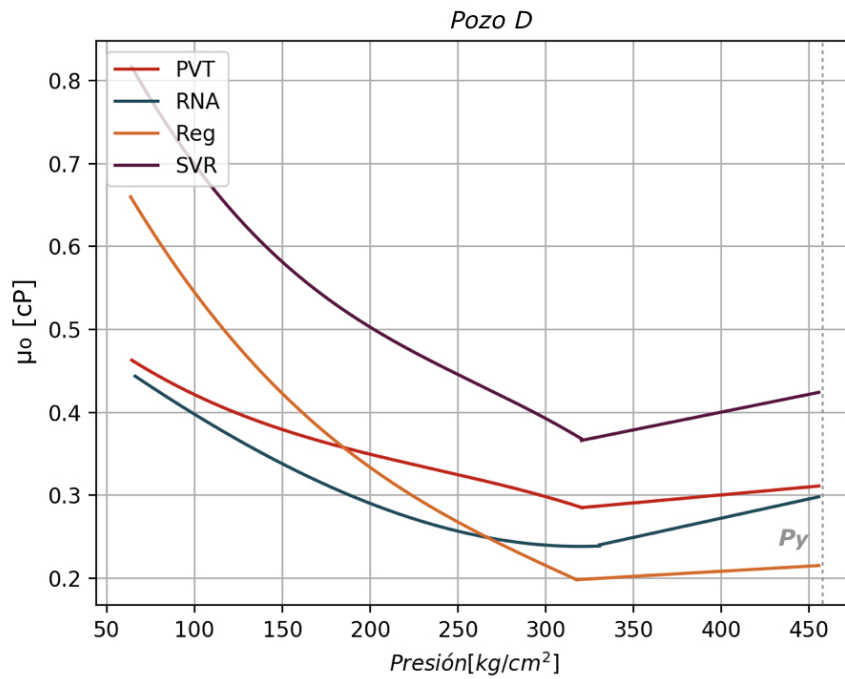


Figura 6.32: Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo D.

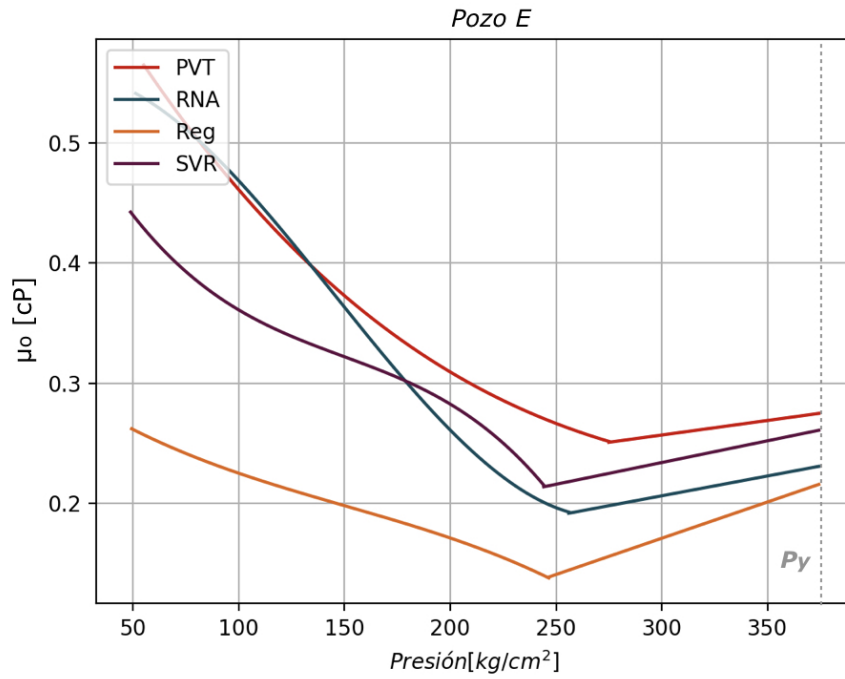


Figura 6.33: Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo E.

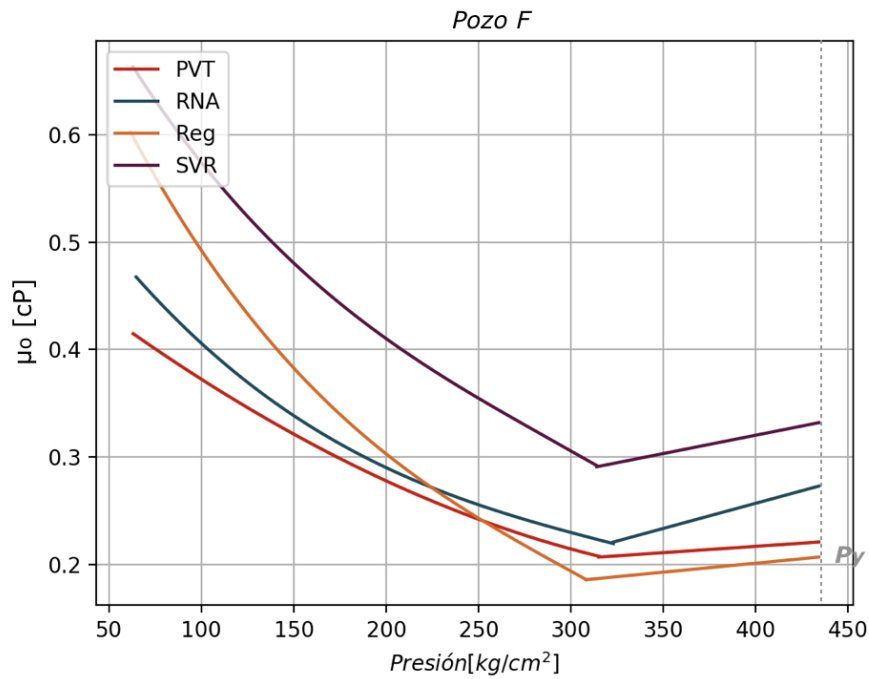


Figura 6.34: Curva de μ_o estimada con los tres métodos de aprendizaje automático para el pozo F.

Con las Figuras 6.29, 6.30, 6.31, 6.32, 6.33 y 6.34 es posible observar de manera gráfica que, para estimar μ_o , el modelos de aprendizaje automático que tiene una mejor aproximación es la red neuronal artificial.

En este capítulo se mostraron los resultados obtenidos por cada algoritmo de aprendizaje automático propuesto y, se puede observar que, en general, las redes neuronales artificiales y las máquinas de vectores de soporte son los más competitivos para estimar los puntos que conforman a las curvas de las propiedades PVT.

Sin embargo, para el caso de la viscosidad, se puede ver que los porcentajes de error obtenidos son muy altos. Probablemente, esto puede deberse a que las variables elegidas no son suficientes para modelar esta propiedad o bien, es necesario hacer uso de otros métodos de aprendizaje automático para estimar los valores de dicha curva.

En el siguiente capítulo, se analizarán los resultados obtenidos en los dos capítulos anteriores con la finalidad de obtener las conclusiones finales de este trabajo y, se darán algunas alternativas para mejorar este estudio a futuro.

Capítulo 7

Conclusiones y trabajo futuro

Con el desarrollo de este trabajo puede comprobarse que el uso de métodos de aprendizaje automático es de gran utilidad para la clasificación y estimación de propiedades PVT de fluidos petroleros.

Para cada una de las propiedades estudiadas en este trabajo, los mejores resultados pueden observarse a continuación en la Tabla 7.1.

Tabla 7.1: Modelos con menor porcentaje de E_a para cada propiedad PVT.

Propiedad	Modelo	E_a [%]
P_b	RNA	3.89
B_o	RNA	2.96
ρ_{ro}	SVR	2.26
R_s	SVR	3.89
μ_o	RNA	17.00

Como puede observarse en la Tabla 7.1, las redes neuronales y las máquinas de vectores de soporte son los modelos cuyos resultados al estimar propiedades como el punto de burbuja (P_b), la densidad relativa del aceite (ρ_{ro}), el factor de volumen del aceite (B_o) y la relación gas aceite (R_s) presentan errores pequeños, de menos del 10 %, superando incluso los resultados obtenidos por las técnicas clásicas de estimación de propiedades PVT disponibles en la literatura, que puede verse en el Anexo A.

Los buenos resultados obtenidos para las propiedades mencionadas anteriormente se deben a que, todas estas propiedades tienen un comportamiento más estable ante la variación de la presión del yacimiento. Además, el número de registros recopilados en el conjunto de datos es mayor a 100, por lo que puede considerarse un conjunto de datos grande.

Por otro lado, para el caso de la curva de viscosidad, los valores estimados tuvieron errores absolutos de más del 30% debido a que, además de que es una propiedad física compleja de modelar, el conjunto de datos con el que se contaba para entrenar los algoritmos era escaso.

También, es necesario aclarar que los modelos entrenados tendrán buenos resultados específicamente para la región estudiada en este trabajo, es decir, si se evalúa con aceites de otras regiones petroleras es probable que muestren resultados pobres. Por lo que, si se desea aplicar estas técnicas en proyectos reales, es necesario crear modelos específicos para cada región, campo o yacimiento. Debido a esto, es recomendable aplicar los métodos de aprendizaje automático en campos maduros que cuentan con suficiente información de pruebas PVT tomadas con anterioridad. Los modelos de aprendizaje automático pueden representar una reducción en el número de pruebas de laboratorio PVT a realizar en el futuro y, por lo tanto, significan un ahorro en el presupuesto de proyectos posteriores.

Como trabajo futuro, se proponen las siguientes acciones:

- Elaborar un conjunto de datos más abundante al utilizado para este trabajo con la finalidad de mejorar el entrenamiento de los modelos. Se espera que, a mayor número de muestras, mayor sea la precisión de los resultados obtenidos por los modelos de aprendizaje automático.
- Investigar sobre otros algoritmos de aprendizaje automático para mejorar los resultados obtenidos al estimar los puntos de la curva PVT de la viscosidad de los aceites.

Además, revisar qué otras variables pueden ser utilizadas para modelar dicha propiedad.

- Elaborar de una interfaz gráfica que mejore la experiencia de usuario y permita ingresar los datos de nuevos fluidos para obtener su clasificación y la estimación de sus curvas PVT de las propiedades físicas estudiadas en este trabajo.

Anexos

Anexo A

Comparaciones de resultados con correlaciones

Con el fin de obtener una referencia de los resultados obtenidos en este trabajo, se tomaron los modelos de aprendizaje automático con mejores resultados de cada propiedad para compararlo contra los resultados obtenidos de las correlaciones de de Standing, Vázquez y Kartoatmojdo, que son las más comúnmente utilizadas y recomendadas en la literatura.

Comparación de P_b

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

Tabla A.1: Resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos utilizando el conjunto de datos de validación

Pozo	A	B	C	D	E	F
$P_b PVT [kg/cm^2]$	184.0	248.0	126.6	320.7	275.0	318.5
$P_b RNA [kg/cm^2]$	175.4	233.8	124.9	330.6	256.0	323.5
$P_b Standing [kg/cm^2]$	239.4	252.6	129.8	511.4	322.2	467.4
$P_b Vázquez [kg/cm^2]$	224.1	231.5	148.5	428.9	265.7	396.6
$P_b Kartoatmojdo [kg/cm^2]$	237.3	255.9	141.6	516.0	303.8	475.4

Tabla A.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.

Método	$E_a de P_b estimado [\%]$
RNA	3.89
Standing	26.31
Vázquez	17.89
Kartoatmojdo	27.44

En las Tablas A.1 y A.2 se puede ver que el modelo de RNA elaborado supera el rendimiento obtenido por las correlaciones de uso tradicional en el conjunto de validación.

A.1. Comparación de 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 (temperatura del yacimiento). Debido a su buen rendimiento, se seleccionaron los resultados del modelo de redes neuronales artificiales como punto de comparación con las correlaciones implementadas en el conjunto de datos de validación.

Tabla A.3: Comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar B_o en el conjunto de validación.

Método	$E_a de B_o$ [%]
RNA	2.96
Standing	5.66
Vázquez	4.86
Kartoatmojdo	5.14

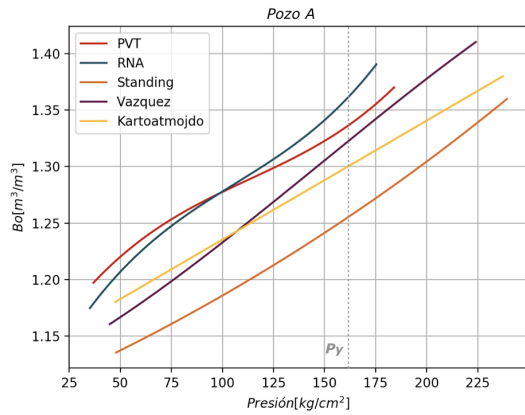


Figura A.1: Curvas de B_o Pozo A

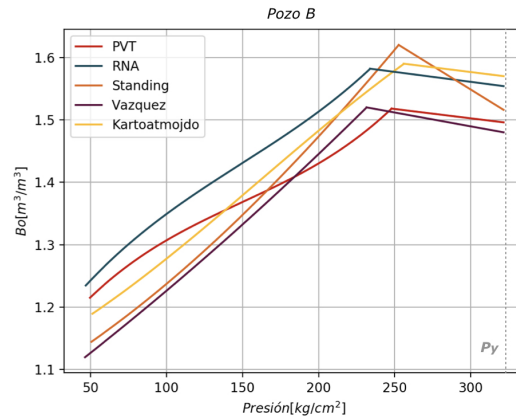


Figura A.2: Curvas de B_o Pozo B

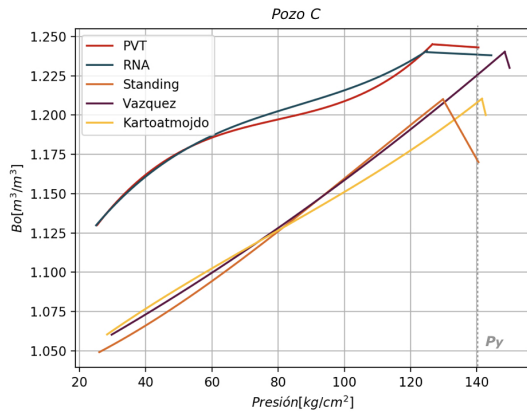


Figura A.3: Curvas de B_o Pozo C

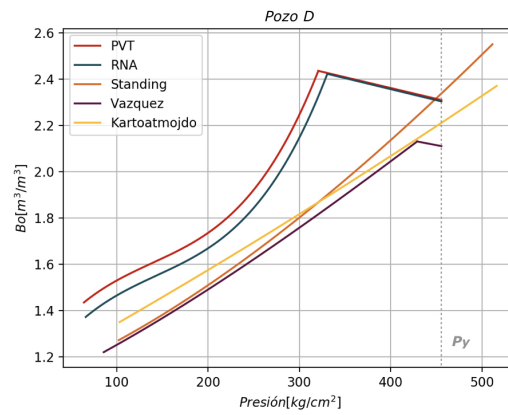


Figura A.4: Curvas de B_o Pozo D

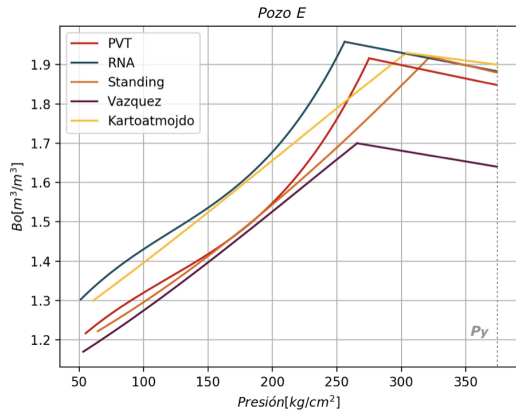


Figura A.5: Curvas de B_o Pozo E

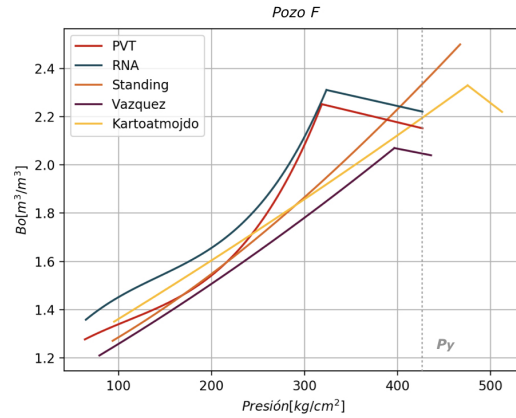


Figura A.6: Curvas de B_o Pozo F

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.

ρ_{ro}

A continuación, se hará la comparación del factor de volumen 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, se seleccionaron los resultados del modelo de máquinas de vectores de soporte como punto de comparación con las correlaciones implementadas en el conjunto de datos de validación.

Tabla A.4: Comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar ρ_{ro} en el conjunto de validación.

Método	$E_a de B_o$ [%]
SVR	2.26
Standing	5.29
Vázquez	3.92
Kartoatmojdo	4.55

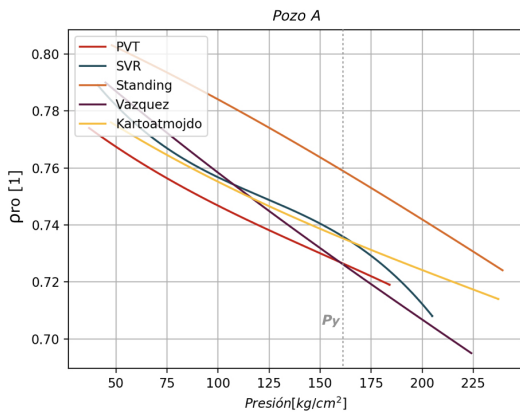


Figura A.7: Curvas de ρ_{ro} Pozo A

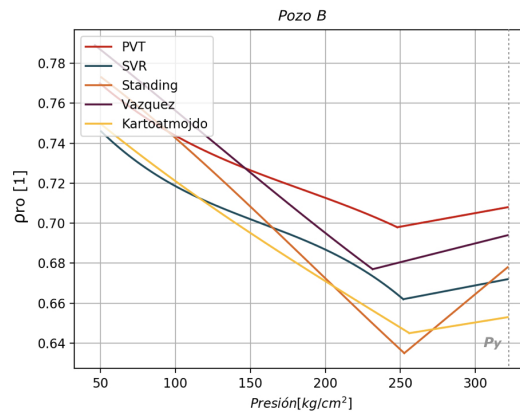


Figura A.8: Curvas de ρ_{ro} Pozo B

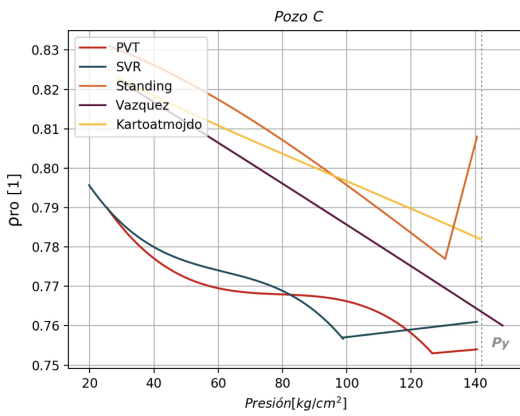


Figura A.9: Curvas de ρ_{ro} Pozo C

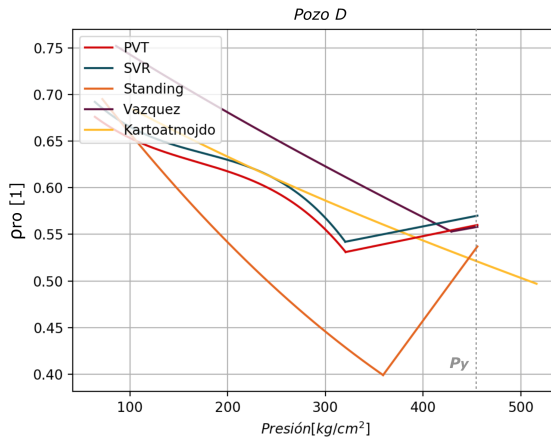


Figura A.10: Curvas de ρ_{ro} Pozo D

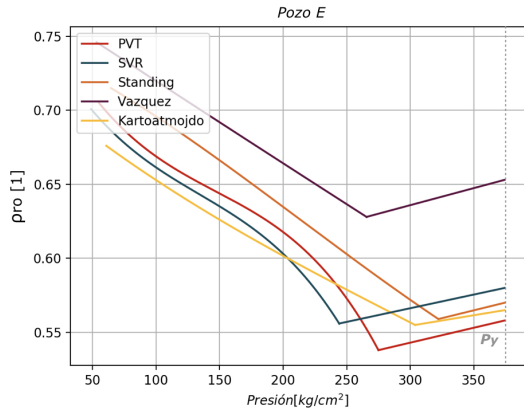


Figura A.11: Curvas de ρ_{ro} Pozo E

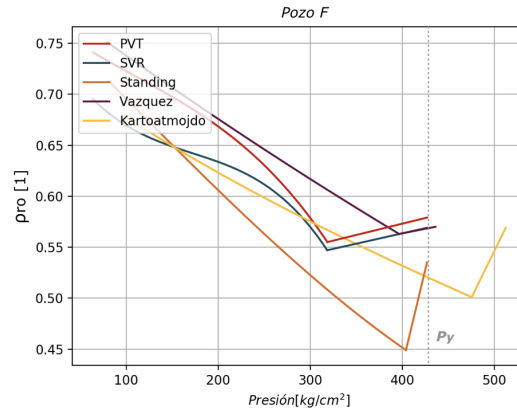


Figura A.12: Curvas de ρ_{ro} Pozo F

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 con el rendimiento obtenido por las correlaciones de uso tradicional en la mayoría de los pozos de el conjunto de validación, compitiendo muy de cerca con el modelo de Vázquez.

A.2. Comparación de R_s

En esta sección, se compara la relación gas-aceite, específicamente, la curva generada conforme varía la presión del sistema y la temperatura se mantiene constante (temperatura del yacimiento). Debido al buen funcionamiento del modelo de las máquinas de vectores de soporte, fue seleccionado como punto de comparación con las correlaciones implementadas en el conjunto de datos de validación.

Tabla A.5: Comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar R_s en el conjunto de validación.

Método	$E_a de B_o$ [%]
SVR	3.89
Standing	26.31
Vázquez	17.89
Kartoatmojdo	27.44

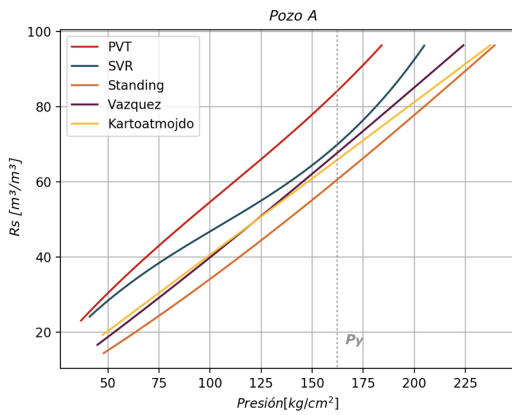


Figura A.13: Curvas de R_s Pozo A

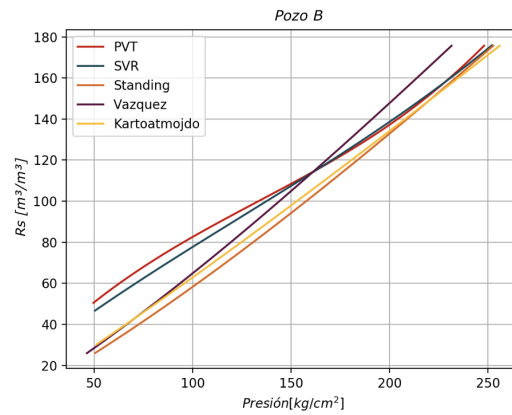


Figura A.14: Curvas de R_s Pozo B

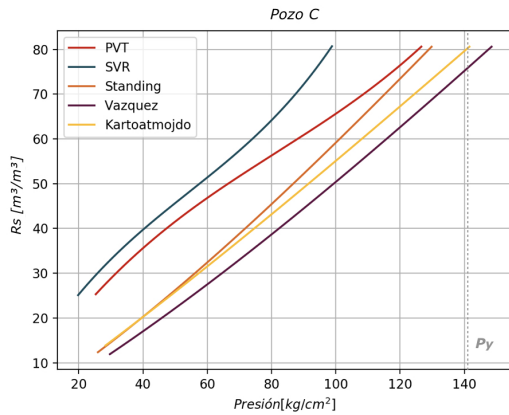


Figura A.15: Curvas de R_s Pozo C

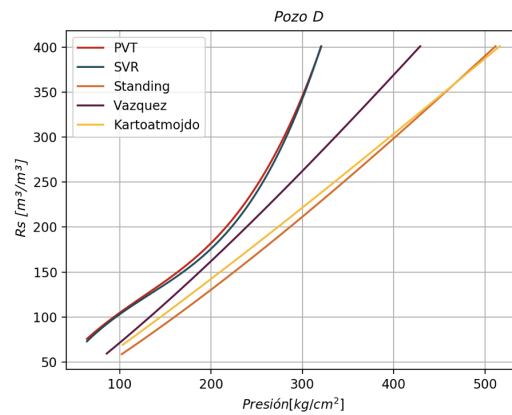


Figura A.16: Curvas de R_s Pozo D

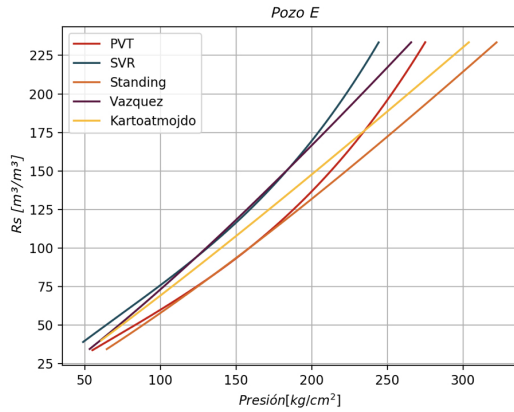


Figura A.17: Curvas de R_s Pozo E

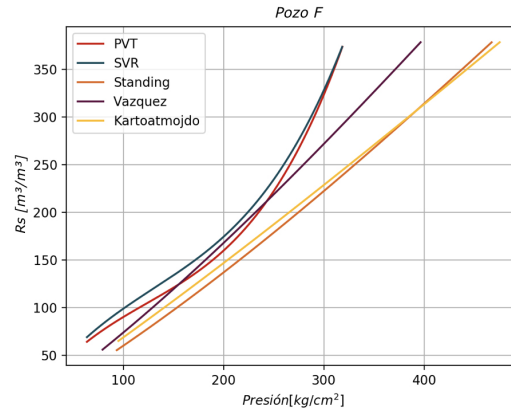


Figura A.18: Curvas de R_s Pozo F

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 SVR elaborado supera con el rendimiento obtenido por las correlaciones de uso tradicional en la mayoría de los pozos de el conjunto de validación.

A.3. Comparación de μ_o

La última propiedad a comparar es la viscosidad del aceite, específicamente la curva generada conforme varía la presión del sistema y la temperatura del yacimiento se mantiene constante. Debido a su buen rendimiento, se eligieron los resultados del modelo de RNA como punto de comparación con las correlaciones implementadas en el conjunto de datos de validación.

Tabla A.6: Comparación de los resultados obtenidos con las pruebas PVT y los modelos de estimación propuestos para estimar μ_o en el conjunto de validación.

Método	$E_a de B_o$ [%]
RNA	17.00
Vázquez	15.74
Kartoatmojdo	38.62

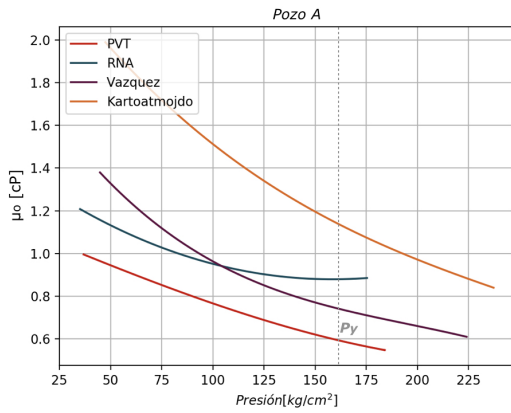


Figura A.19: Curvas de μ_o Pozo A

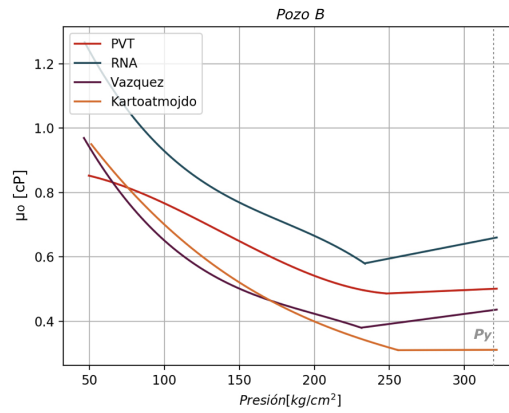


Figura A.20: Curvas de μ_o Pozo B

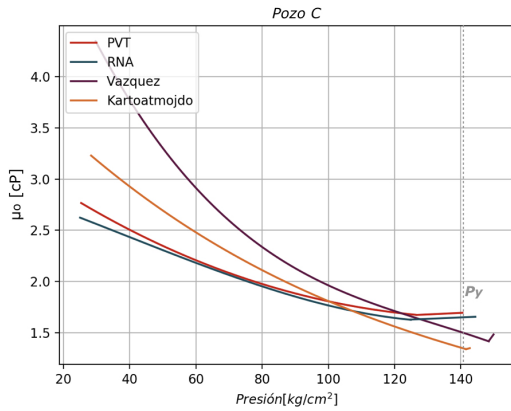


Figura A.21: Curvas de μ_o Pozo C

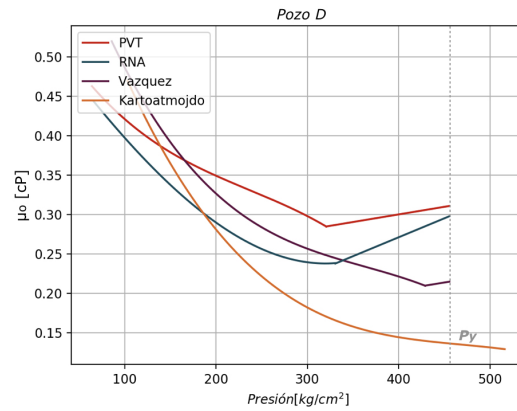
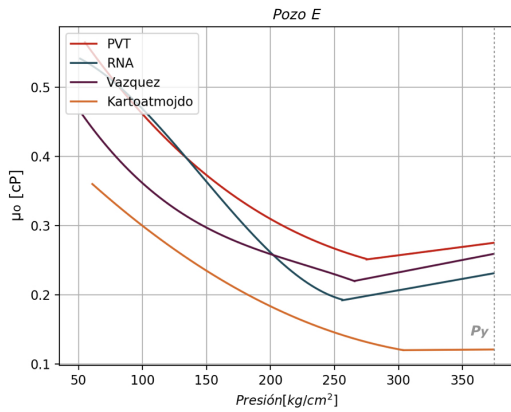
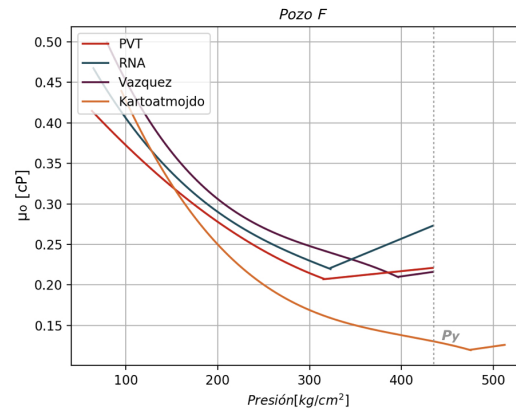


Figura A.22: Curvas de μ_o Pozo D

Figura A.23: Curvas de μ_o Pozo EFigura A.24: Curvas de μ_o Pozo F

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 supera con el rendimiento obtenido por las correlaciones de uso tradicional en los pozos de el conjunto de validación, sin embargo, el modelo de Vázquez es más competitivo en el cálculo de la viscosidad en la mayoría de los pozos.

Bibliografía

- Ahmed, T. (2006). *Reservoir Engineering Handbook*. Elsevier.
- Al-Marhourn, M. & Osman, E. (2002). Using Artificial Neural Networks to Develop New PVT Correlations for Saudi Crude Oils. *Society of Petroleum Engineers*.
- 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*.
- Banko, M. & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*.
- Camargo, A. (2016). Estimación de propiedades PVT para yacimientos de aceite negro y volátil mediante una red neuronal artificial. *Universidad Nacional Autónoma de México*.
- 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*.
- Dindoruk, B. & Christman, P. G. (2001). PVT Properties and Viscosity Correlations for Gulf of Mexico Oils. *Society of Petroleum Engineers*.
- 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(2).

- Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. The MIT Press.
- 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*.
- L.T., M. & Tayssier, S. (1979). Caracterización de Fluidos de Yacimientos Petroleros. *Revista Instituto Mexicano del Petróleo*, 11(4).
- Mcain, W. (1990). *The Properties of Petroleum Fluids*. PenWell Publishing Company, Second Edition.
- 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 characterisation 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*.
- Ramirez, A., Valle, G., Romero, F. & Jaimes, M. (2017). Production of PVT Properties in Crude Oil Using Machine Learning Techniques. *SPE Latin America and Caribbean*.
- Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3).
- Salamat, 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. *IJEAST*, 4(12).
- Shizhen, T., Xuanjun, Y. & Lianhua, H. (2016). Play types, geologic characteristics and exploration domains of lithological reservoirs in China. *Petroleum Exploration And Development*, 43(6).

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