

**UNIVERSIDAD MAYOR DE SAN ANDRÉS**

**FACULTAD DE INGENIERÍA**

**CARRERA DE INGENIERÍA PETROLERA**



**PROYECTO DE GRADO**

**“DETECCIÓN DE ANOMALÍAS EN LOS SEPARADORES EN EL POZO CUR-X1001  
LL MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO”**

**POSTULANTE:**

Univ. Emanuel Chambi Quisbert

**TUTOR:**

MSc. Ing. Pedro Reynaldo Marín Domínguez

La Paz – Bolivia



**UNIVERSIDAD MAYOR DE SAN ANDRÉS  
FACULTAD DE INGENIERIA**



**LA FACULTAD DE INGENIERIA DE LA UNIVERSIDAD MAYOR DE SAN ANDRÉS AUTORIZA EL USO DE LA INFORMACIÓN CONTENIDA EN ESTE DOCUMENTO SI LOS PROPÓSITOS SON ESTRICTAMENTE ACADÉMICOS.**

**LICENCIA DE USO**

El usuario está autorizado a:

- a) Visualizar el documento mediante el uso de un ordenador o dispositivo móvil.
- b) Copiar, almacenar o imprimir si ha de ser de uso exclusivamente personal y privado.
- c) Copiar textualmente parte(s) de su contenido mencionando la fuente y/o haciendo la cita o referencia correspondiente en apego a las normas de redacción e investigación.

El usuario no puede publicar, distribuir o realizar emisión o exhibición alguna de este material, sin la autorización correspondiente.

**TODOS LOS DERECHOS RESERVADOS. EL USO NO AUTORIZADO DE LOS CONTENIDOS PUBLICADOS EN ESTE SITIO DERIVARA EN EL INICIO DE ACCIONES LEGALES CONTEMPLADAS EN LA LEY DE DERECHOS DE AUTOR.**

## **Dedicatoria**

**A Dios**, por guiarme y sustentarme con su infinito amor hasta el día de hoy.

**A mi papá en forma de mamá**, Ana Elizabeth Quisbert Chuquimia que me dio la valentía, valores y educación para ser día tras día mejor persona de la que fui ayer, la que me ayudo a cumplir con todos los requisitos para terminar un sueño más y nunca dejarme caer con las aflicciones que nos presenta este mundo.

## **Agradecimiento**

A Dios, por escuchar los deseos y cumplir los anhelos de mi corazón ayudándome a vencer cada obstáculo para llegar hasta el lugar donde estoy el día de hoy.

*Jehová es mi pastor; nada me faltará. En lugares de delicados pastos me hará descansar; Junto a aguas de reposo me pastoreará. Confortara mi alma; Me guiara por sendas de Justicia por amor de su nombre.*

*Salmos 23, 1 - 3*

Gracias a mi mamá Ana Elizabeth Quisbert Chuquimia, que, cumpliendo el rol de papá y mamá a la vez, tuvo las fuerzas para sacar adelante a su primogénito llevando el pan de cada día a casa con el sudor de su frente, con lágrimas en los ojos mi eterna gratitud a lo más importante en mi vida, mi madre.

A mi hermano Israel Alejandro Quisbert, por ser mi mano derecha, mi motivación el que no dudaba en alcanzarme un café en mis noches de desvelo.

A mi hermanito Miguel Ángel Quisbert, por ser mi alegría y demostrarme tanta bondad que por afanes del estudio iba perdiendo.

A mi novia Yessica Soto, por el amor y apoyo brindado en el tiempo que estuvo conmigo, tanto buenos como malos, demostrando que es una excelente mujer para mi vida.

A mi Tutor, MSc. Ing. Pedro Reynaldo Marín Domínguez, por su tiempo, por su paciencia, por su experiencia y apoyo brindado en el desarrollo de mi proyecto.

A mis docentes de la Universidad Mayor de San Andrés, de la Facultad de Ingeniería y mi querida carrera Ingeniería Petrolera que me brindaron tanto conocimiento y valores para ejercer una vida profesional culta y de responsabilidad.

A los amigos que me apoyaron, tales que hicieron la carrera más liviana por el apoyo mutuo brindado.

## Índice General

Dedicatoria.....	i
Agradecimiento.....	ii
Índice General.....	iii
Índice de Figuras .....	vii
Índice de Tablas .....	ix
Indicé de Ecuaciones.....	x
Glosario de abreviaturas:.....	xi
Glosario de términos:.....	xii
Resumen .....	xiii
1. CAPÍTULO 1 GENERALIDADES .....	1
1.1. INTRODUCTION.....	1
1.2. ANTECEDENTES.....	2
1.2.1. Características del reservorio. ....	4
1.2.2. Producción del pozo CUR-X1001 LL. ....	5
1.2.3. Sistema de separación. ....	6
1.3. PLANTEAMIENTO DEL PROBLEMA.....	7
1.3.1. Identificación del problema .....	8
1.3.2. Formulación del problema.....	8
1.4. OBJETIVOS.....	8
1.4.1. Objetivo general.....	8
1.4.2. Objetivos específicos y acciones de la investigación .....	8
1.5. JUSTIFICACIÓN. ....	9
1.5.1. Justificación técnica. ....	9
1.5.2. Justificación económica. ....	9
1.6. ALCANCE .....	9
1.6.1. Alcance temático.....	9
1.6.2. Alcance temporal .....	11
2. CAPÍTULO 2 MARCO TEÓRICO .....	12

2.1.	PRODUCCIÓN DE HIDROCARBUROS .....	12
2.1.1.	Mecanismos de producción .....	12
2.1.2.	Las Técnicas de producción.....	17
2.1.3.	El recorrido de los Crudos. ....	22
2.1.4.	Bombeo de hidrocarburos en pozo. ....	25
2.2.	SEPARACIÓN DE HIDROCARBUROS.....	29
2.2.1.	Principios de separación .....	29
2.2.1.1.	Separación por gravedad .....	29
2.2.1.2.	Numero de Reynolds.....	30
2.2.1.3.	Fuerza Centrifuga .....	30
2.2.1.4.	Choque de Partículas .....	30
2.2.2.	Función de un separador .....	31
2.2.2.1.	Funcionalidad en Separadores Gas-Líquido.....	32
2.2.2.2.	Funcionalidad en Separadores Gas-Petróleo-Agua. ....	32
2.2.3.	Explicación del proceso de Separación .....	32
2.2.4.	Etapas de la separación.....	33
2.2.5.	Variables que afectan la eficiencia de separación. ....	34
2.2.5.1.	Viscosidad dinámica del gas.....	34
2.2.5.2.	Temperatura. ....	35
2.2.5.3.	Densidad del líquido y gas.....	36
2.2.5.4.	Tamaño de las partículas de líquidos.....	36
2.2.5.5.	Presión de separación. ....	37
2.3.	POSIBLES ANOMALÍAS EN UN SEPARADOR. ....	37
2.3.1.	Formación de espumas dentro del separador.....	37
2.3.2.	Formación de vórtices:.....	38
3.	CAPÍTULO 3 CIENCIA DE DATOS.....	39
3.1.	INTELIGENCIA ARTIFICIAL.....	40
3.2.	MACHINE LEARNING .....	40
3.2.1.	Aprendizaje Supervisado.....	40
3.2.2.	Aprendizaje No Supervisado .....	48
3.3.	DEEP LEARNING .....	49
3.4.	MINERÍA DE DATOS.....	50
3.4.1.	Metodología KDD .....	50

3.4.2.	Fases de metodología KDD .....	50
3.4.2.1.	Selección .....	50
3.4.2.2.	Preprocesamiento.....	51
3.4.2.3.	Transformación .....	51
3.4.2.4.	Minería de datos .....	52
3.4.2.5.	Evaluación .....	53
3.5.	MÉTRICAS DE EVALUACIÓN. ....	53
3.5.1.	Coeficiente de Determinación. ....	54
3.5.2.	Error de Cuadrático Medio. ....	55
3.5.3.	Error porcentual absoluto medio.....	56
3.5.4.	Normalización .....	57
-	Normalización media. ....	57
-	Normalización Min-Max.....	57
-	Normalización de cuantiles.....	57
3.5.5.	Escala estandarizada. ....	57
4.	CAPÍTULO IV APLICACIÓN PRACTICA .....	58
4.1.	DETECCIÓN DE ANOMALÍAS EN EL SEPARADOR DE PRUEBA EN EL POZO CUR-X1001 MEDIANTE LA REGRESIÓN LINEAL SIMPLE Y REGRESIÓN LINEAL MÚLTIPLE .....	58
4.1.1.	Selección .....	58
4.1.2.	Preprocesado/Limpieza .....	59
4.1.3.	Transformación .....	61
4.1.4.	Minería de datos .....	61
4.1.4.1.	Método 1: Regresión lineal simple entre gas y condensado. ....	62
4.1.4.2.	Método 2: Regresión Lineal Múltiple entre gas, condensado y agua.....	65
4.1.4.3.	Método 3: Suprimiendo los outliers. ....	65
4.1.4.4.	Método 4: Criterio de la normalización con la regresión lineal simple. ....	67
4.1.4.5.	Método 5: Criterio de la normalización con la regresión lineal múltiple. ....	70
4.1.4.6.	Método 6: Criterio de la normalización con la regresión lineal Simple eliminando variables. 72	
4.1.4.7.	Método 7: Añadiendo más variables. ....	74
4.1.5.	Evaluación .....	75
4.2.	TABLA DE RESULTADOS.....	76
4.3.	ANÁLISIS DE RESULTADOS PARA LA APLICACIÓN. ....	78

4.3.1. Puesta en marcha del algoritmo .....	79
4.4. ANALISIS DE COSTOS.....	82
5. CAPÍTULO V CONCLUSIONES Y RECOMENDACIONES .....	83
5.1. conclusiones.....	83
5.2. RECOMENDACIONES.....	85
BIBLIOGRAFIA.....	87
ANEXOS .....	89



## Índice de Figuras

Figura 1.1: Corte del reservorio Peteca. ....	2
Figura 1.2: Importación de los datos. ....	3
Figura 1.3: Tamaño del DataFrame.....	4
Figura 1.4: Tamaño del DataFrame.....	5
Figura 1.5 Separadores del campo.....	6
Figura 2.1: Esquema de empuje por Agua .....	13
Figura 2.2: Esquema de empuje por Casquete de Gas .....	14
Figura 2.3: Esquema de empuje por gas disuelto .....	15
Figura 2.4: Esquema de empuje por gravedad. ....	16
Figura 2.5: Esquema de empuje por expansión de líquido.....	17
Figura 2.6: Esquema de inyección y levantamiento por Gas.....	19
Figura 2.7: Esquema de la técnica de recuperación secundaria.....	20
Figura 2.8: Esquema de la técnica de recuperación mejorada-inyección de agua.....	21
Figura 2.9: Sistema de Extracción de Hidrocarburos .....	22
Figura 2.10: Esquema de un campo de extracción de petróleo .....	23
Figura 2.11: Esquema de un campo de extracción de petróleo .....	24
Figura 2.12: Extractor de alambre .....	31
Figura 3.1: Ramas de la ciencia de datos .....	39
Figura 3.2: Función sigmoid. ....	46
Figura 3.3: Graficas los diferentes valores por regiones.....	48
Figura 4.1: Box plot para el análisis del condensado .....	60
Figura 4.2: Box plot para el análisis del gas. ....	60
Figura 4.3: Box plot para el análisis del agua.....	61
Figura 4.4: Correlación entre las variables.....	62
Figura 4.5: Gas Vs Condensado -Entrenamiento. ....	63

Figura 4.6: Gas Vs Condensado -Examen.....	64
Figura 4.7: Eliminación de outliers.....	66
Figura 4.8: Gas Vs Condensado -Entrenamiento-Normalizado. ....	67
Figura 4.9: Gas Vs Condensado -Examen-Normalizado.....	68
Figura 4.10: Métricas estadísticas método 4 con la librería regressors. ....	69
Figura 4.11: Grafica de residuales métodos 4. ....	70
Figura 4.12: Métricas estadísticas método 5 con la librería regressors. ....	71
Figura 4.13: Grafica de residuales métodos 4. ....	71
Figura 4.14: Métricas estadísticas método 6 con la librería regressors. ....	73
Figura 4.15: Grafica residual método 6.....	73
Figura 4.16: Añadiendo las variables de presión y temperatura.....	74
Figura 4.17: Correlación con las nuevas variables. ....	74
Figura 4.18: Un dato aleatorio del DataFrame: .....	79
Figura 4.19: Entrada de datos de prueba. ....	80
Figura 4.20: Entrada de datos de prueba gas excesivo.....	81
Figura 4.21: Entrada de datos de prueba gas disminuido. ....	81
Figura 5.1: Producción de Gas pozo Tacobo-PCD .....	85
Figura 5.2: Posible grafico para expresar la producción por pozo.....	86

## Índice de Tablas

Tabla 1.1: Características del POZO CURX-1001 .....	3
Tabla 1.2: Características del Reservorio Petaca. ....	4
Tabla 1.3: Características del separador de grupo V-101. ....	6
Tabla 1.4: Características del separador de prueba V-201. ....	6
Tabla 1.5: Promedio esperado de producción por pozo. ....	7
Tabla 2.1: Ventajas y desventajas del bombeo mecánico .....	26
Tabla 2.2: Ventajas y desventajas del bombeo electro sumergible .....	27
Tabla 2.3: Ventajas y desventajas del bombeo de cavidad progresiva .....	28
Tabla 2.4: Ventajas y desventajas del bombeo de cavidad progresiva. ....	29
Tabla 4.1: Numero de Filas y Columnas de nuestra Data-Original. ....	59
Tabla 4.2: Columnas por su número de datos no nulos y tipo de dato. ....	59
Tabla 4.3: Métricas del método 1. ....	64
Tabla 4.4: Métricas del método 2. ....	65
Tabla 4.5: Métricas del Método 3. ....	66
Tabla 4.6: Métricas del Método 4. ....	68
Tabla 4.7: Métricas del método 5. ....	70
Tabla 4.8: Métricas del método 5. ....	72
Tabla 4.9: Métricas del método 7. ....	75
Tabla 4.10: Resumen de resultados. ....	76

## Indicé de Ecuaciones

Ecuación 1: Regresión Lineal Simple.....	41
Ecuación 2: Determinación del coeficiente 'a' en la regresión lineal simple.....	42
Ecuación 3: Determinación del coeficiente 'b' en la regresión lineal simple .....	42
Ecuación 4: Regresión lineal Múltiple.....	44
Ecuación 5: Regresión Logística .....	46
Ecuación 6: Árbol de decisiones.....	47
Ecuación 7: Coeficiente de determinación .....	54
Ecuación 8: Formula del error cuadrático medio. ....	55
Ecuación 9: Formula del error porcentual absoluto medio. ....	56
Ecuación 10: Formula del error cuadrático medio. ....	57

## **Glosario de abreviaturas:**

**BBL:** Nos referimos a la abreviatura de Barril de condensado.

**Data:** Terminó usado para referirse a los datos, en el caso del proyecto se habla de una matriz de datos con filas y columnas.

**DF:** Nos referimos a la estructura del DataFrame.

**IA:** Nos referimos a la ciencia de la Inteligencia Artificial.

**IOT:** En su traducción al inglés ‘Internet Of Things’ es decir el ‘Internet de las cosas’ es la agrupación e interconexión de dispositivos y objetos a través de una red (bien sea privada o Internet, la red de redes), dónde todos ellos podrían ser visibles e interaccionar.

**KDD:** Es la abreviatura de Knowledge Discovery in Databases o también por su traducción al castellano es Descubrimiento de Conocimiento en Bases de Datos, se refiere a las secuencias de fases que se realiza para a identificación de patrones posiblemente útiles y entendibles que ayuden a la toma de decisiones.

**MPC:** Es la abreviatura de la unidad de volumen de miles de pies cúbicos.

**MSE:** Es el error cuadrático medio es el criterio de evaluación más usado para problemas de regresión. Se usa sobre todo cuando usamos aprendizaje automático supervisado. Para cada dato histórico podremos indicar el resultado correcto. Vamos a ver como se calcula.

**R<sup>2</sup>:** Es el coeficiente de correlación correspondiente a una regresión lineal, que presenta valores de 0 a 1, mientras más cercano este a la unidad indica mayor exactitud.

## **Glosario de términos:**

**DataFrame:** Es una estructura de almacenamiento de datos de 2 dimensiones.

**BOXPLOT:** También llamado como diagrama de caja, es un método estandarizado para representar una serie de datos numéricos a través de sus cuartiles.

**Examen:** Es la comparación que se hace de los datos, que se guardaron a para este fin con el valor real que presentan estos.

**Mecanismo:** Un grupo de partes o elementos que se ajustan entre sí y usan energía mecánica para hacer un trabajo o realizar una función.

**Métrica:** Cuando nos referiremos a ellas son valores que cuando se los compara presentan variación, por ende, se escoge al mejor:

**Modelo:** Nos referimos al modelo aplicado de Machine Learning.

**Outliers:** Valores fuera de los límites de datos del centro estadísticamente, donde su efecto es más notorio en la media y no en la mediana o la moda.

**Red neuronal:** Es un modelo de la inteligencia artificial que simplícidamente intenta emular al cerebro humano en la forma de procesar información. El conjunto de estas intenta simular el proceso de aprendizaje de las neuronas del cerebro.

**Revolución 4.0:** También es llamada como la cuarta revolución industrial esta cuarta etapa se caracteriza por una fusión de tecnologías actualmente en prueba o en desarrollo, lo que está desintegrando las fronteras entre las esferas física, digital, y biológica.

**Separador de control:** Separador usado para analizar los flujos de un pozo exclusivamente.

**Separador general:** Separador usado para separar los fluidos de un campo.

**Técnica:** Conjunto de programas o recursos para las artes, las ciencias o actividades específicas, especialmente cuando se adquieren mediante la práctica y requieren habilidades.

**Variable:** Es un símbolo que permite identificar a un elemento no especificado dentro de un determinado grupo. Este conjunto suele ser definido como el conjunto universal de la variable.

## **Resumen**

En el presente proyecto de Grado titulado: “DETECCIÓN DE ANOMALÍAS EN LOS SEPARADORES EN EL POZO CUR-X1001 MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO” se plantea la idea de realizar un algoritmo de aprendizaje automático capaz de predecir la cantidad de gas en la salida del separador de prueba, para compararlos con el valor real. Es decir que comparar la cantidad de Gas registrada y la predicha para ser capaz de detectar anomalías, estas se pueden ver afectadas ya sea por la antigüedad del pozo o dentro del separador algún imperfecto.

Se escogieron la metodología KDD para tratar los datos, se empezó con un dataframe de dimensiones 53 filas y columnas, se usaron 6 diferentes métodos y 2 modelos de aprendizaje automático que fueron la regresión lineal simple y la regresión lineal múltiple.

Los criterios de los cuales se partieron fueron usar algoritmos capaces de graficar si existe correlación entre las variables. Donde se vio una fuerte correlación entre la cantidad de gas de salida y también la cantidad de condensado.

Se entrenaron los modelos y métodos con el 70% de los datos y 30% para su examen correspondiente.

Para los métodos que se usaron fueron: criterio de eliminación de valores atípicos estadísticamente, normalización, p\_value. Al analizar las métricas se evidencio que el método 6 presenta mejores métricas.

Se crearon algoritmos capaces de alertar en caso de presentarse un valor anómalo en la salida del separador en la salida de gas.

# **1. CAPÍTULO 1 GENERALIDADES**

## **1.1. INTRODUCTION**

Durante décadas llevamos produciendo cantidades enormes de hidrocarburos, agua y otros diferentes compuestos contaminantes que vienen pegados en la producción de campo, la producción de los mismos nos ayuda a que la industria en general tenga movimiento, ya que estos compuestos mediante procesos de refinación, petroquímica o tecnológicos van dando diferentes usos a nuestro diario vivir.

La PRODUCCION de los hidrocarburos forma parte de la cadena de UPSTREAM desde ya hace poco tiempo (aproximadamente 10 años) diferentes empresas petroleras en las que se destacan las más importantes (Shell, PB, Lukoil, Baker Hughes, Schlumberger, Halliburton) están sacando provecho a los DATOS que se genera en los campos de producción. Dichas empresas ya están dentro de la industria 4.0 usando maneras masivas de almacenamiento de datos (Azure, AWS, Google Cloud).

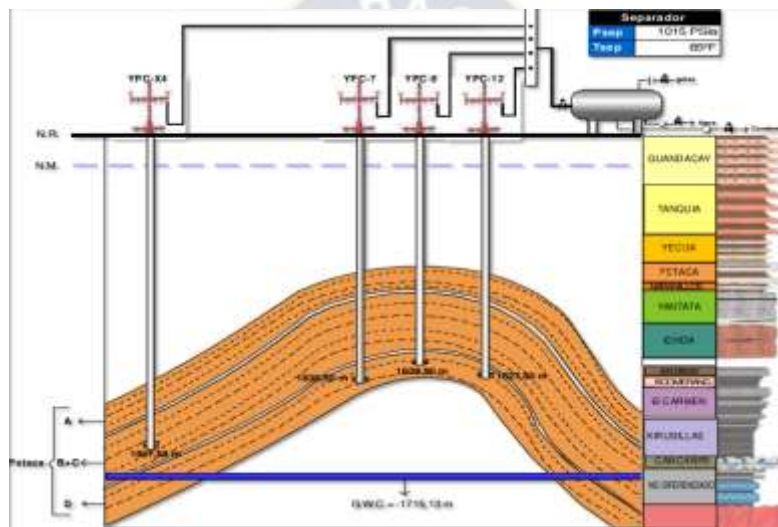
En el presente proyecto se usara los DATOS del separado de Control en el campo Tacobo-Curiche, el impacto que se quiere llegar a tener es que gracias a la creación de algoritmos usando el lenguaje de programación Python se podrá ser capaz de predecir cierta falla dentro del separador es decir que sea capaz de respondernos si la cantidad de salida de gas es correcta o no, esto nos llevaría a que el proceso sea más automatizado ya que si somos capaces de hallar el algoritmo en unos cuantos pozos en un campo, el algoritmo también sería capaz de satisfacer necesidades de otros campos de comprendan muchos más pozos. De esta manera la industria nacional daría los pasos correctos para sumarse a la INDUSTRIA 4.0.



## 1.2. ANTECEDENTES

La información con que se cuenta es la del pozo CUR-X1001 LL perteneciente reservorio YECUA/PETECA, el reservorio Petaca es un reservorio productor de gas, que aporta con la producción de gas y líquidos condensados en superficie en la figura 1.1 se presentara el corte del reservorio Petaca.

Figura 1.1: Corte del reservorio Peteca.



Fuente: C.N.I.H. YPFB

Este pozo se encuentra ubicado a 150 m. de la planta de acondicionamiento de gas, este presenta una Tubería de producción dual es decir presenta línea corta como una línea larga debido a las dos formaciones productoras. En la tabla 1.1 se muestran las características más importantes del pozo CURX-1001.

Tabla 1.1: Características del POZO CURX-1001

	<b>CONDICIONES EN LINEA CORTA</b>	<b>CONDICIONES EN LINEA LARGA</b>
<b>PROFUNDIDAD</b>	1332-1338 m	1518-1518 m
<b>FORMACION</b>	Chaco inferior	Yecua/Petaca
<b>PRESION</b>	1588 PSI	1640 PSI
<b>CHOKE</b>	26/64	22/64

Fuente: Informe técnico campos Tacobo-Curiche YPFB.

Los datos con los que contamos contienen la producción de petróleo, gas y agua además la fecha en que fueron tomados estos datos. La Data original es un archivo tipo Excel que cuenta con más de 50 filas.

En la figura 1.2 podemos ver como se ha transferido los datos del Excel exitosamente al entorno de desarrollo de Google Colaboratory creando el objeto DataFrame este procedimiento se lo hizo con el lenguaje de programación PYTHON y específicamente con su librería de PANDAS.

Figura 1.2: Importación de los datos.

	<b>UNIQUEID</b>	<b>DATE</b>	<b>COND BBL</b>	<b>GAS MPC</b>	<b>WATER BBL</b>
<b>0</b>	CUR_1001_LL	2010-09-30	0.896667	7.796043	0.000000
<b>1</b>	CUR_1001_LL	2011-01-31	38.938710	169.537439	0.000000
<b>2</b>	CUR_1001_LL	2011-02-28	40.207143	195.866988	0.000000
<b>3</b>	CUR_1001_LL	2011-03-31	50.232258	211.505722	0.483871
<b>4</b>	CUR_1001_LL	2011-04-30	33.266667	154.263509	0.333333
<b>5</b>	CUR_1001_LL	2011-05-31	31.429032	141.630508	0.404194
<b>6</b>	CUR_1001_LL	2011-06-30	32.090000	153.329109	0.592000
<b>7</b>	CUR_1001_LL	2011-07-31	32.041935	147.264639	0.518710
<b>8</b>	CUR_1001_LL	2011-08-31	31.341935	144.663349	0.454839
<b>9</b>	CUR_1001_LL	2011-09-30	29.960000	144.203733	0.460667

Fuente: Elaboración propia usando Python con datos de Informe técnico de YPFB, 2012.

En la figura 1.3 describirá el total de los datos que ascienden a 265 que es el producto del número de filas con el número de las columnas, dicho número de filas es 53 y las columnas es de 5 como se puede ver en la siguiente figura.

Figura 1.3: Tamaño del DataFrame.

```
df.size
265

df.shape
(53, 5)
```

Fuente: Elaboración propia usando Python con datos de Informe técnico de YPF, 2012.

### 1.2.1. Características del reservorio.

Las características del reservorio Petaca se presentarán en la tabla 1.2.

Tabla 1.2: Características del Reservorio Petaca.

Características	Gas Condensado	Gas Húmedo	Gas Seco
GRP Inicial SCF/STB.	8.000-70.000	60.000-100.000	-
°API Inicial	>60	>60	-
C7+molar (%)	<12.5	-	-
Cambio de fase en el reservorio.	Punto de rocío	Sin cambio de Fase	Sin cambio de Fase.
Color del Fluido producido.	Colorado, marrón, anaranjado o transparente.	Transparente.	-

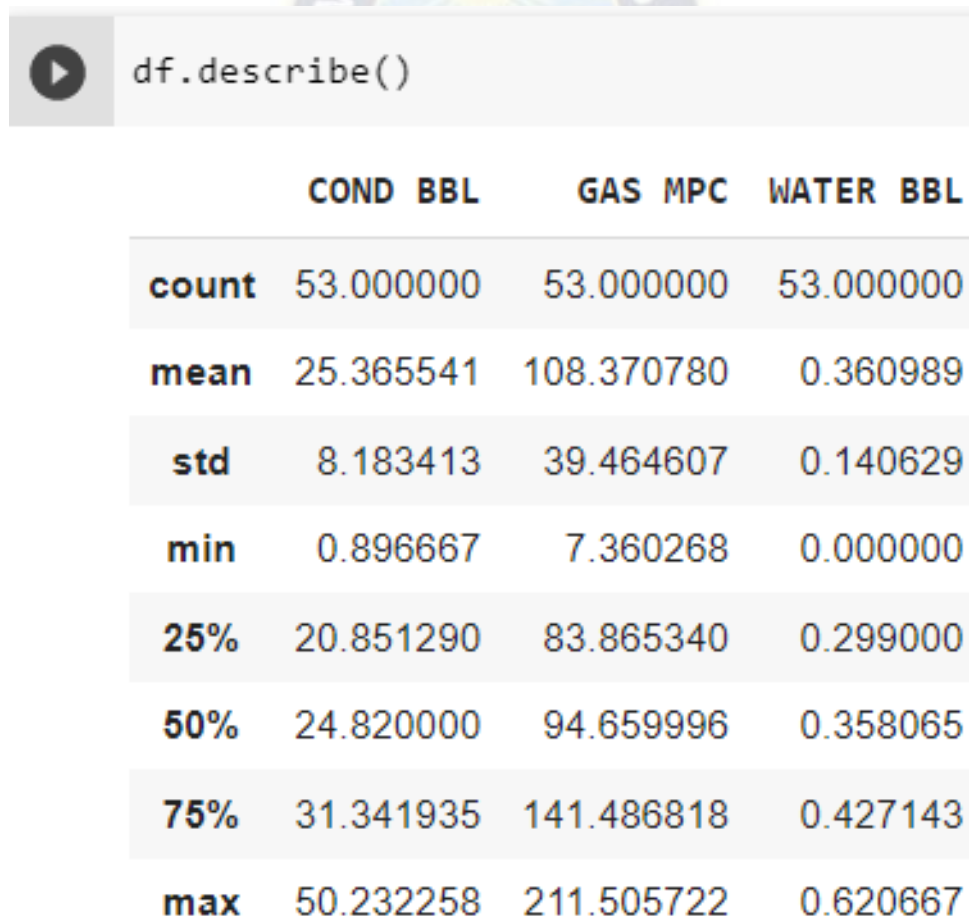
Fuente: Trabajo de grado: Explotación del reservorio Petaca del campo Yapacaní mediante la determinación de caudales óptimos de gas condensado.

### 1.2.2. Producción del pozo CUR-X1001 LL.

El lenguaje Python contiene herramientas que fácilmente realizan un análisis descriptivo de las columnas que contenga el DataFrame, cuya condición es que las columnas sean de tipo numérico ya sea como enteros o decimales.

En la figura 1.4 se puede observar la descripción de los datos en el siguiente orden: Contar, promedio, desviación estándar, valor mínimo, cuartil 25%, mediana, cuartil 75%, valor máximo en el mismo orden.

Figura 1.4: Tamaño del DataFrame.



```
df.describe()
```

	COND BBL	GAS MPC	WATER BBL
<b>count</b>	53.000000	53.000000	53.000000
<b>mean</b>	25.365541	108.370780	0.360989
<b>std</b>	8.183413	39.464607	0.140629
<b>min</b>	0.896667	7.360268	0.000000
<b>25%</b>	20.851290	83.865340	0.299000
<b>50%</b>	24.820000	94.659996	0.358065
<b>75%</b>	31.341935	141.486818	0.427143
<b>max</b>	50.232258	211.505722	0.620667

Fuente: Elaboración propia usando Python con datos de Informe técnico de YPF, 2012.

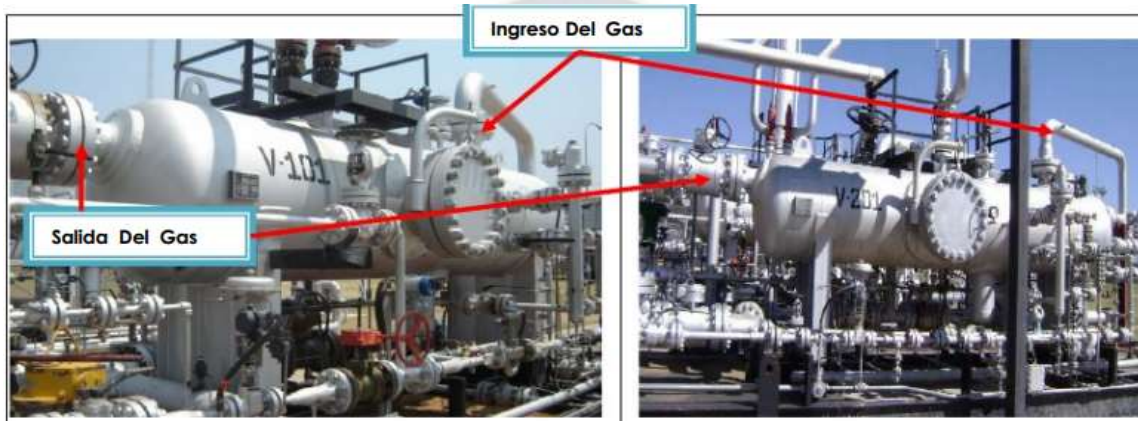
### 1.2.3. Sistema de separación.

La separación cuenta con 2 separadores horizontales trifásicos de alta presión 1440 y 1100 PSI.

Uno es un separador de Grupo (V-101) y el otro es un separador de Prueba (V-201).

En la figura 1.5 se puede evidencia a ambos separadores en el campo.

Figura 1.5 Separadores del campo.



Fuente: Informe técnico campos Tacobo-Curiche YPF, 2012.

En la tabla 1.3 se describirá las condiciones de operación en el separador de grupo.

Tabla 1.3: Características del separador de grupo V-101.

<b>Presión de operación.</b>	<b>Temperatura de operación.</b>	<b>Caudal de Gas MMPCS.</b>
1440 PSI	120 °F	43,8

Fuente: Informe técnico campos Tacobo-Curiche YPF, 2012.

En la tabla 1.4 se describirá las condiciones de operación en el separador de prueba.

Tabla 1.4: Características del separador de prueba V-201.

<b>Presión de operación.</b>	<b>Temperatura de operación.</b>	<b>Caudal de Gas MMPCS.</b>
1100	68 °F	12.4

Fuente: Informe técnico campos Tacobo-Curiche YPF, 2012.

En la Batería de producción se realizan de acuerdo a un cronograma de Pruebas cada 4 días se cambia una nueva línea para su prueba. A continuación, se presentará el promedio de la producción que tendría que cumplir el los pozos en Gas y Condensado en la tabla 1.5.

Tabla 1.5: Promedio esperado de producción por pozo.

POZOS	LINEA	WHP PSI	TWH °F	P.SEP PSI	T.SEP °F	GAS MMPCD	COND. BPD	API
CURX-1001	L/C	1505	100	1194	82.7	5.256	11.5	0.4
	L/L	1555	99.5	1186	83.1	3.934	24.22	0.4
CURX-1002	L/C	1550	104.5	1201	80	4.232	59.2	0.4
	L/L	1545	99	1150	79.6	5.601	65.2	0.6
CURX-1004	L/C	1540	100	1193	80.5	4.183	57.5	0.4
	L/L	1555	101.5	1189	80	3.941	65.2	0.5
CURX-1006	L/C	1520	91.5	1185	75.6	3.811	58.4	0.4
	Cerrado	0	0	0	0	0	0	0

Fuente: Informe técnico campos Tacobo-Curiche YPFB, 2012.

### 1.3. PLANTEAMIENTO DEL PROBLEMA

La existencia es casi segura de dos separadores en casi todo campo de gas condensado, uno general y otro de prueba. Con los resultados de la salida de separador de prueba se tiene los productos de un solo pozo y mediante esto podemos realizar un análisis de datos inclusive aplicar modelos de aprendizaje automático para ser capaces de predecir una cantidad anómala del gas de salida lo que nos puede ayudar más adelante a detectar alguna anomalía.

Como es de conocimiento general los equipos pueden presentar anomalías a lo largo de su vida productiva, la importancia de hallar los valores anómalos radicara en que valores de agua, gas y condensado que salen del separador deben estar correlacionados caso contrario puede indicar que se presentaría una anomalía.

### **1.3.1. Identificación del problema**

Con los datos obtenidos de la salida de los productos del separador de prueba del pozo CURX-1001 se puede detectar posibles anomalías en el separador.

Como es de conocimiento la función más importante de un separador es separar los hidrocarburos y agua. En caso de que una empresa petrolera no sea capaz de detectar la anomalía en la salida de sus fluidos arrastraría problemas más grandes posteriores ya sea en el procesamiento (Condensaciones no deseadas) y/o transporte de los hidrocarburos (menor capacidad de transporte por la condensación).

### **1.3.2. Formulación del problema**

Si no se reacciona a la anomalía de separación de los fluidos a la salida del separador de control se tendrá problemas en procesos posteriores que atraviesan nuestros hidrocarburos, con la aplicación de un algoritmo se desea una respuesta inmediata a una anomalía en el la salida de los fluidos en el separador.

El algoritmo debería ser capaz de responder si los valores de los fluidos son normales (según nuestro modelo desarrollado) o es un valor anómalo a la salida del separador.

## **1.4. OBJETIVOS.**

### **1.4.1. Objetivo general**

- Predecir anomalías en el separador de prueba del pozo CUR X1001 LL, basándonos en los valores que se registran en el separador de gas, condensado y agua, mediante modelos de aprendizaje automático.

### **1.4.2. Objetivos específicos y acciones de la investigación**

- Plantear una metodología de análisis de datos.

- Entrenar cada modelo con al menos 70% hasta el 80% para su posterior evaluación.
- Tener métricas correspondientes para la evaluación de cada modelo.
- Aplicar el mejor método y/o modelo de aprendizaje automático en el pozo CUR X1001 LL.

## **1.5. JUSTIFICACIÓN.**

### **1.5.1. Justificación técnica.**

El presente proyecto de grado ayudara a la ingeniería de producción a actualizarse a las tecnologías usadas por otras empresas (INDUSTRIA 4.0), a una pronta respuesta ante valores anómalos registrados en un separador. Ayudando a la industria petrolífera nacional, demostrando la importancia que se tiene en recolectar, almacenar y usar los datos registrados a lo largo del tiempo.

### **1.5.2. Justificación económica.**

Una de las consecuencias de una separación ineficiente más preocupantes es por ejemplo de una posible formación de hidratos en tuberías dedicadas al transporte de hidrocarburos directamente supondría un gran problema ya que podría llevarnos al paro de una planta, desgaste y/o corrosión de las válvulas y demás equipos que se tiene.

## **1.6. ALCANCE**

### **1.6.1. Alcance temático**

Los alcances se expondrán en el mismo orden que el objetivo general y específicos fueron descritos, son los siguientes:

- Al predecir nos referimos a que los modelos de aprendizaje automático que plantearemos se entrenaron con una parte de los datos y luego el modelo con la parte restante de los datos debe ser capaz de predecir la cantidad de salida de gas natural (variable dependiente)



usando las cantidades de salida de condensado y agua (variables independientes), con la parte de los datos reservada para la predicción se sacaran métricas para comparar los modelos. Ahora para la parte de anomalías se planteará si la salida del gas natural predicha es semejante al valor real, en caso contrario posiblemente se haya producido una anomalía.

- Se planea usar la metodología KDD usada en minería de datos, en caso de que no se puedan cumplir ciertas etapas de la metodología KDD, se harán posibles sugerencias de cómo se pueden realizar.
- Entrenar los datos con el 70% al 80% esto es una acción se la realiza por ser una buena costumbre en proyectos de implementación de aprendizaje automático supervisado, y se evaluara el modelo con el restante de los datos en base a las métricas que se expondrán más adelante.
- Se deberán tener en cuenta la variable a predecir la cantidad de gas de salida del separador en un tipo de numero continuo así que se plasmarán métricas que se adapten a números continuos como ser el coeficiente de determinación y el error cuadrático medio.
- Se observarán los datos de distintas maneras con el fin de obtener las mejores predicciones, lo que se traduce en obtener las mejores métricas. El modelo optimo será el que obtenga mejores métricas.
- Ser capaz de desarrollar algoritmos donde el usuario pueda ingresar los datos de gas, agua y condensado (real), donde el algoritmo sea capaz de retornar si es un valor normal o un valor que señale una anomalía. Ya sea posible para el desarrollo para pozo en ser posible y un analizando para el campo en general.

### 1.6.2. Alcance temporal

Los datos fueron sacados de un reporte de productos de salida del separador de prueba en un archivo Excel del pozo CUR X1001 LL que abarca la data de la producción desde el año 2012. El total de muestras es de 53 datos, cabe aclarar que estos datos son antes de la limpieza de datos correspondiente que nos siguiere la metodología KDD. Además, que el pozo CUR X10001 LL se conectó al separador de control cada 4 días, no fueron días seguidos.



## 2. CAPÍTULO 2 MARCO TEÓRICO

### 2.1. PRODUCCIÓN DE HIDROCARBUROS

Básicamente la producción del Petróleo se lleva a cabo por el desplazamiento del agua o gas que está dentro del yacimiento gracias a su presión natural del yacimiento. A medida que se va produciendo gas y petróleo con el tiempo esta presión natural va disminuyendo. Se denomina RECUPERACION al proceso de recolección del crudo y el gas, y que se puede clasificar en recuperación primaria, secundaria y terciaria. (PDVSA, 2016, p. 5)

#### 2.1.1. Mecanismos de producción

Existen básicamente cinco mecanismos de producción, la mayoría de los yacimientos se combinan entre los siguientes mecanismos descritos. Por otra parte, también es frecuente que un yacimiento durante su vida productora actúe los cinco mecanismos en diferentes periodos. (PDVSA, 2016, p. 9)

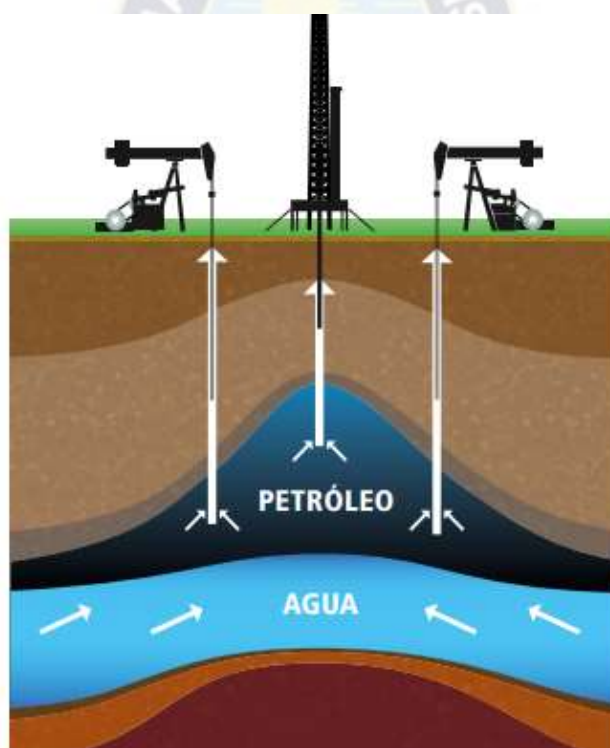
Estos mecanismos son:

- Empuje por agua
- Empuje por casquete de Gas
- Empuje por Gas Disuelto
- Empuje por gravedad
- Empuje por Expansión de líquido

**Empuje por agua:** Es cuando hay una capa de agua debajo de la capa de aceite, que puede provenir de un tanque o de un bache que lo alimenta desde la superficie. Aunque el agua no se puede comprimir mucho, la gran masa que hay en la tierra acumulará una gran cantidad de energía, actuando en forma de pistón que empuja el líquido de abajo hacia arriba a través de los poros de

la tierra. Las rocas son levantadas por tubería de producción desde el pozo y transportadas a través de la base superficial hasta la estación de recolección. Es uno de los mecanismos de propulsión más eficientes, con una tasa de recuperación de hasta el 80% del aceite base en el yacimiento. El estricto seguimiento y control de la producción del pozo es condición necesaria para la efectividad de este mecanismo, a fin de mantener la uniformidad de influencia mutua y evitar el desvío o bloqueo de agua del acuífero. En la figura 2.1 representa la explicación del procedimiento del empuje por agua. (PDVSA, 2016, p. 10)

Figura 2.1: Esquema de empuje por Agua



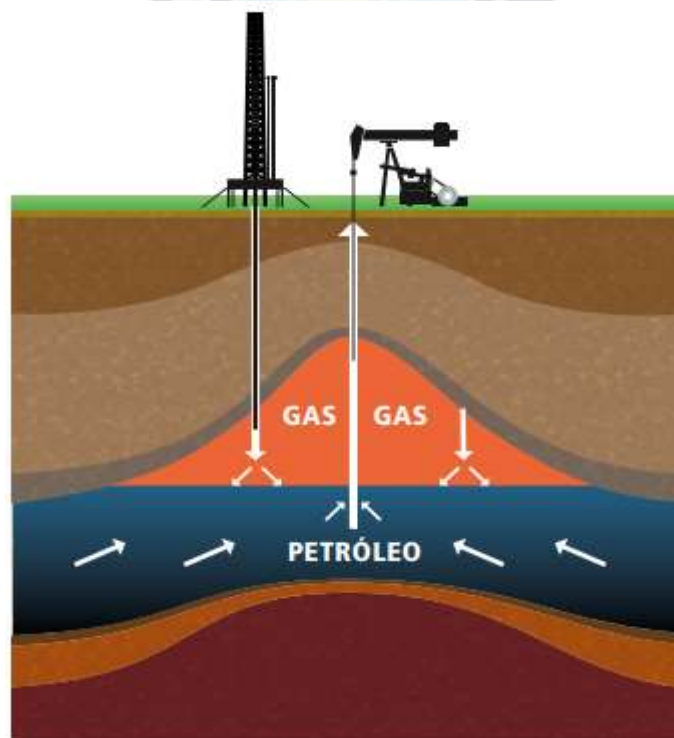
Fuente: Revista Soberanía Petrolera PDVSA

**Empuje por Casquete de Gas:** Los yacimientos donde el contenido de gas original es mayor que lo que se puede disolver en el aceite bajo las condiciones de presión y temperatura existentes, el

volumen de gas no disuelto forma una tapa sobre la capa de aceite que se expande a medida que el aceite fluye, aplicando un pistón de arriba hacia abajo reemplaza el papel del aceite.

A través de este mecanismo, el factor de recuperación del petróleo original en el yacimiento puede llegar al 60%. La condición ideal es mantener uniforme el efecto pistón de la tapa mediante el control efectivo de la producción del pozo, evitando que escape el gas libre y, por lo tanto, pierda la energía impulsora natural. La figura 2.2 explicara el procedimiento de empuje por casquete de gas. (PDVSA, 2016, p. 11)

Figura 2.2: Esquema de empuje por Casquete de Gas

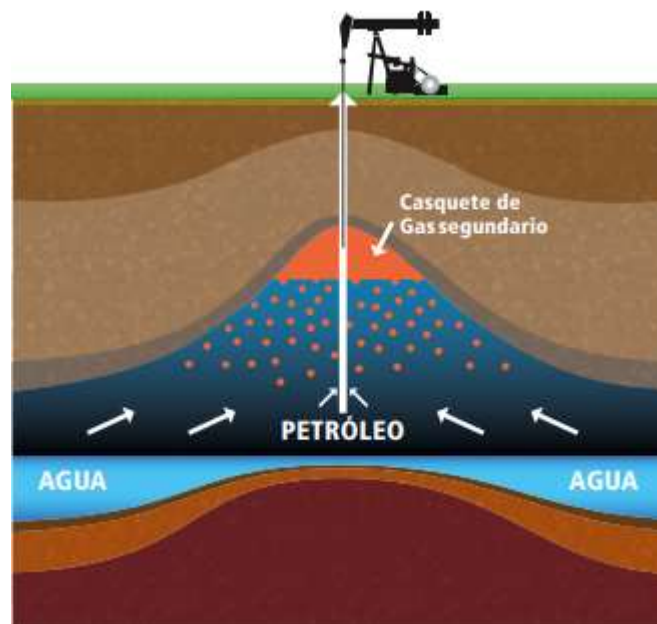


Fuente: Revista Soberanía Petrolera PDVSA

**Empuje por Gas Disuelto:** Este ocurre en reservorios donde todo el gas se disuelve en el petróleo, por lo que no hay capa de gas libre. A medida que se bombean fluidos desde los pozos de producción, la presión del yacimiento disminuye y el gas disuelto se expande, creando una fuerza

que empuja el petróleo. Es el mecanismo de recuperación más común, con una tasa de recuperación de aproximadamente el 25% del petróleo original en el yacimiento. Cuando están presentes las condiciones adecuadas de roca y fluido, se produce un casquete de gas secundario que ayuda a mejorar la recuperación de petróleo. En este tipo de mecanismos, es muy importante controlar la relación entre las cantidades de gas y petróleo producidas, para no desperdiciar energía y promover la formación de casquetes de gas secundarios donde sea razonable. La figura 2.3 explicara el procedimiento de empuje por gas disuelto. (PDVSA, 2016, p. 11)

Figura 2.3: Esquema de empuje por gas disuelto

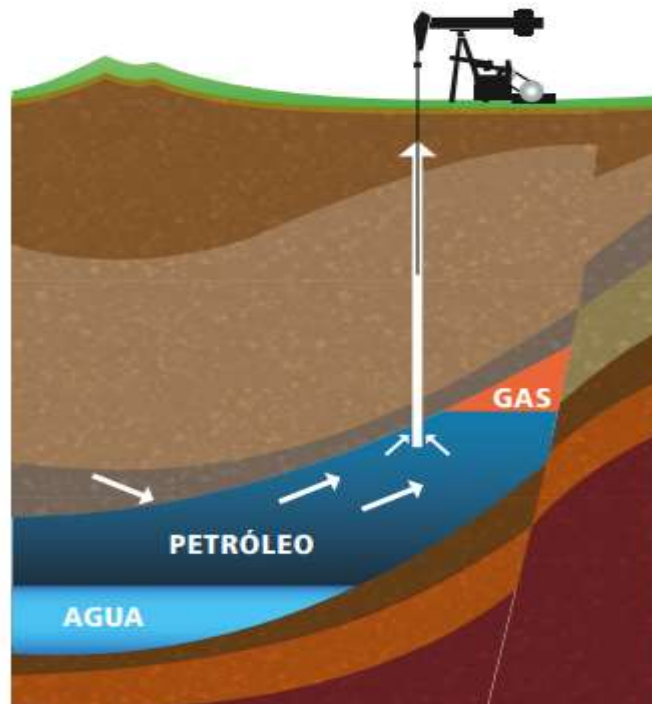


Fuente: Revista Soberanía Petrolera PDVSA

**Empuje por Gravedad:** en algunos depósitos Utilizando un mecanismo de impulsión de gas disuelto, la pendiente rocosa es bastante empinada, este factor facilita que el petróleo se mueva mediante la acción de la gravedad hacia el fondo para llegar al mismo tiempo, el gas libre migra hacia la parte superior, formando una tapa de aire menor causando un efecto de doble

desplazamiento ayuda para la formación de aceite y capa aumentar la tasa de recuperación, es decir, En condiciones óptimas puede llegar al 40% Petróleo presente en un yacimiento. La figura 2.4 explicara el procedimiento de empuje por gravedad. (PDVSA, 2016, p. 11)

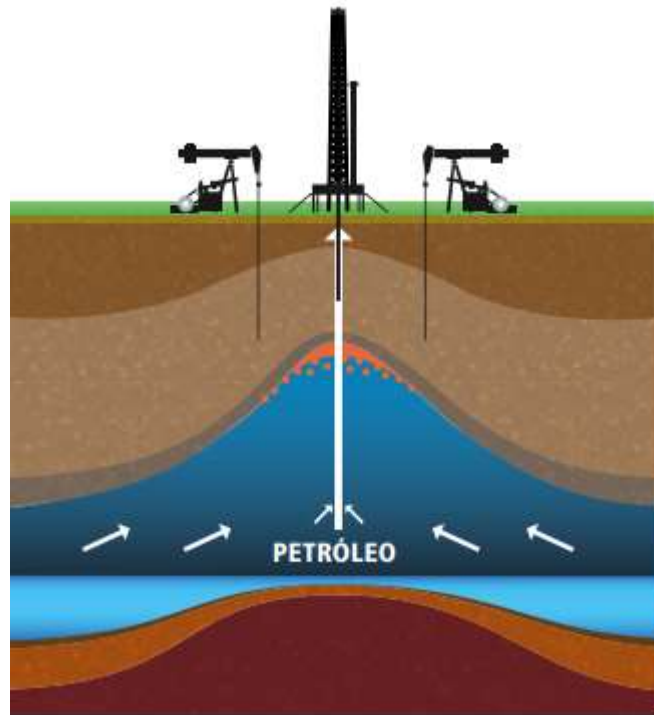
Figura 2.4: Esquema de empuje por gravedad.



Fuente: Revista Soberanía Petrolera PDVSA

**Empuje por expansión de Líquido:** en algunos yacimientos sin capa acuífera y con poco gas disuelto en el petróleo, la producción inicial se obtiene solamente por expansión del petróleo; y bajo estas condiciones la presión declina rápidamente a medida que se extrae el petróleo. La recuperación por expansión de líquido está en el orden de 3% del petróleo original existente en el yacimiento. La figura 2.5 explicara el procedimiento de empuje por expansión de líquido. (PDVSA, 2016, p. 12)

Figura 2.5: Esquema de empuje por expansión de líquido.



Fuente: Revista Soberanía Petrolera PDVSA

### 2.1.2. Las Técnicas de producción

Las técnicas corresponden a los siguientes:

- Flujo natural
- Levantamiento por Gas
- Bombeo
- Recuperación Adicional
- Técnicas de recuperación secundaria
- Técnicas de recuperación Mejorada



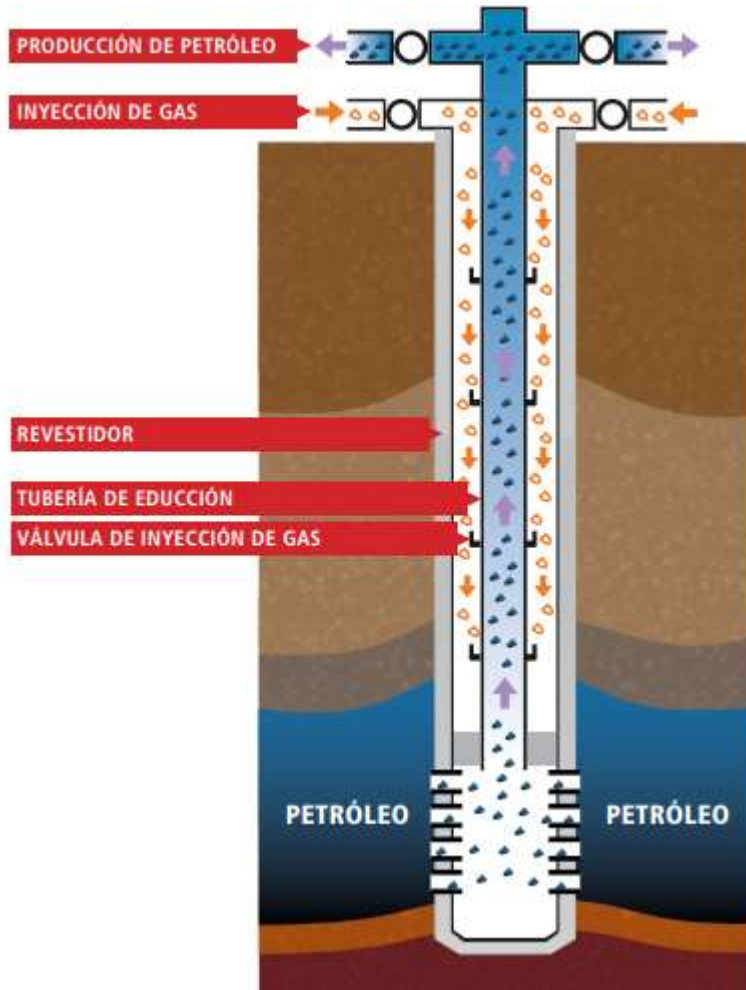
### **Flujo natural:**

En la mayoría de los yacimientos, el pozo comienza a producir por flujo natural, ya que inicialmente la presión existente en el subsuelo es suficiente para impulsar los fluidos en su trayectoria a través de los poros y canales de la roca, en su ascenso por la tubería vertical o inclinada del pozo y en el desplazamiento por la tubería de flujo hasta la estación de recolección. Esa presión natural, sin embargo, se va debilitando a medida que se extraen los fluidos del yacimiento y pierde paulatinamente su fuerza impulsora, entonces se hace necesario aplicar técnicas de levantamiento artificial, con el objeto de mantener los pozos en producción. Entre las técnicas más utilizadas se encuentran el levantamiento por gas y el levantamiento con bombeo mecánico. La escogencia se hace aplicando criterios técnicos y económicos que consideren: las condiciones del yacimiento, las instalaciones mecánicas, las propiedades y características de los fluidos, el mecanismo de impulso del yacimiento, y la disponibilidad de fuerza eléctrica y gas de alta presión. (PDVSA, 2016, p. 14)

### **Levantamiento por Gas:**

Consiste en introducir, mediante compresores, gas de alta presión de una fuente externa en la tubería productora, ya sea directamente por su extremo inferior o mediante válvulas especiales colocadas a diferentes profundidades en dicha tubería, las cuales abren y cierran respondiendo al nivel de la columna de petróleo y la presión del gas. El gas inyectado por el anillo se mezcla con la columna de petróleo y la impulsa hacia la superficie, por la fuerza de expansión del gas. A continuación, la figura 2.6 representa la inyección y levantamiento por Gas. (PDVSA, 2016, p. 15)

Figura 2.6: Esquema de inyección y levantamiento por Gas.



Fuente: Revista Soberanía Petrolera PDVSA

### Recuperación Adicional:

En los primeros tiempos de la Industria Petrolera, los operadores tenían escasos conocimientos del subsuelo. Extraían petróleo sin mayor control hasta agotar los mecanismos de producción primaria por flujo natural y por levantamiento artificial.

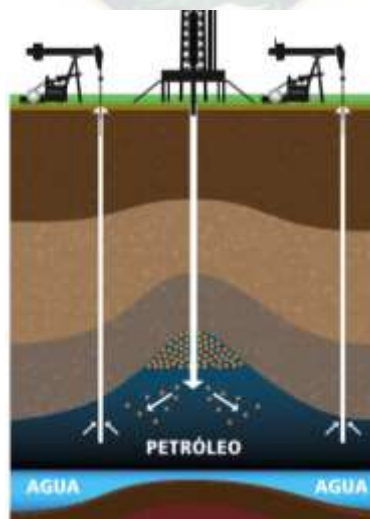
Después abandonaban los pozos y el campo. Se ha podido comprobar que con ese procedimiento se producía sólo una pequeña fracción del petróleo existente en los poros de las rocas y quedaba

un volumen considerable sin recuperar, que en muchos casos representaba hasta 80% del petróleo original acumulado en el yacimiento. (PDVSA, 2016, p. 15)

### **Técnicas de recuperación secundaria:**

Implican la inyección de agua o gas en el yacimiento bajo presión, cuyo efecto de desplazamiento ha demostrado ser muy efectivo para mejorar la recuperación de petróleo. Estas tecnologías son ahora una parte integral del proceso de producción, a partir de la extracción de los yacimientos minerales. La inyección de agua o gas se realiza a través de pozos inyectores de agua, cuya ubicación depende del tipo de yacimiento y de las propiedades de la roca y del fluido. Para ello, se apoya en modelos de simulación computarizados y experimentales que pueden predecir con mucha precisión el comportamiento y evolución de los diferentes parámetros implicados. En algunos casos, la inyección de agua o gas no es adecuada o satisfactoria, lo que ha llevado al uso de técnicas de recuperación mejorada de petróleo. La siguiente figura 2.7 describe la técnica de recuperación mejorada. (PDVSA, 2016, p. 16)

Figura 2.7: Esquema de la técnica de recuperación secundaria

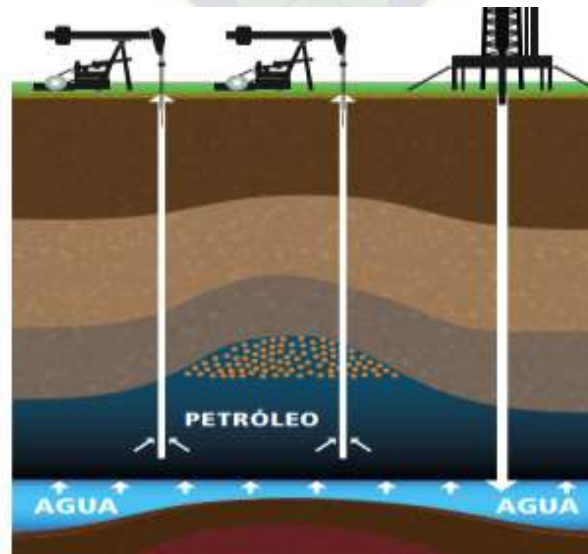


Fuente: Revista Soberanía Petrolera PDVSA

## Técnicas de Recuperación Mejorada:

En esta aplicación, se modifican algunas características del aceite del tanque o del líquido bombeado para aumentar su eficiencia de desplazamiento. En campos de petróleo pesado y viscoso, se ha logrado un aumento significativo en la producción mediante el uso de calor con agua caliente, vapor y combustión directa de una llama subterránea, que enciende el petróleo y ayuda al proceso de combustión por inyección de aire. En campos de petróleo mediano a ligero, se inyectan solventes y productos químicos en el gas o el agua para reducir la fuerza capilar que ayuda a que el petróleo se adhiera a la roca, reducir la viscosidad del petróleo para facilitar la fluidización o aumentar la viscosidad del petróleo, se inyecta líquido para aumentar su eficiencia de empuje. Muchos estudios recientes han abierto nuevas oportunidades para aumentar la recuperación de petróleo, como la inyección controlada de microorganismos (bacterias) en los tanques de almacenamiento, lo que mejora el flujo de petróleo a los pozos de producción, así como su calidad. La figura 2.8 representa la técnica de inyección de agua. (PDVSA, 2016, p. 17)

Figura 2.8: Esquema de la técnica de recuperación mejorada-inyección de agua



Fuente: Revista Soberanía Petrolera PDVSA

### 2.1.3. El recorrido de los Crudos.

Una vez que el campo petrolero se pone en producción, el petróleo, el gas y el agua pasan a través de los pequeños poros y canales en la roca del yacimiento, ingresan al pozo a través de la pistola de perforación, ascienden a través de la tubería de producción y alcanzan la superficie donde se encuentra el cabezal del pozo. Instalar.

La cabeza del pozo de flujo natural se llama "árbol de Navidad" porque su forma tiene cierta semejanza con las decoraciones navideñas. Es el conjunto de válvulas, conexiones utilizadas para cerrar y abrir pozos, tomar muestras, registrar presiones, controlar caudales e introducir ciertas herramientas de trabajo requeridas ocasionalmente. Un elemento importante en un pozo de flujo natural es un reductor o estrangulador, conocido en la industria por el término inglés choque, un aditamento desmontable que regula la extracción de crudo a través de un pequeño orificio, cuyo diámetro está determinado por normas técnicas. La siguiente figura 2.9 representa el sistema de extracción de hidrocarburos para ser específico un árbol de navidad. (PDVSA, 2016, p. 19)

Figura 2.9: Sistema de Extracción de Hidrocarburos

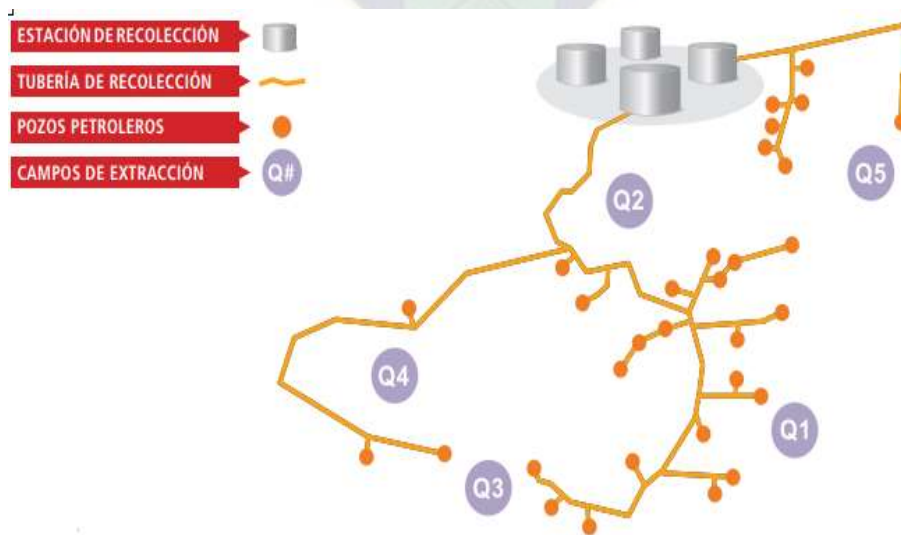


Fuente: Revista Soberanía Petrolera PDVSA

### Estación colectiva:

Después de pasar por la unidad de estrangulamiento, el líquido es conducido a través de una tubería presurizada a una estación de recolección que consta de separador, tanque de almacenamiento, bomba y motor, donde se generan varios pozos cercanos numerados, la cantidad depende de la longitud de la mina entregado y el número de ejes de producción y volumen de producción. En estas estaciones, cada pozo puede ser inspeccionado individualmente para verificar periódicamente el estado de su producción de petróleo, gas y agua, y muestrear estos fluidos para identificar variables importantes para la planificación, la gestión del yacimiento y el plan de operación. Luego de llegar a la estación de recolección, los líquidos van al separador, aumentando la turbulencia del flujo y facilitando la separación de gases que serán dirigidos a los compresores, plantas de tratamiento y plantas de proceso; descomponerlo en productos de mayor valor comercial como gas doméstico, anticontaminación y más. La figura 2.10 representa un esquema tipo de extracción de hidrocarburos. (PDVSA, 2016, p. 19)

Figura 2.10: Esquema de un campo de extracción de petróleo

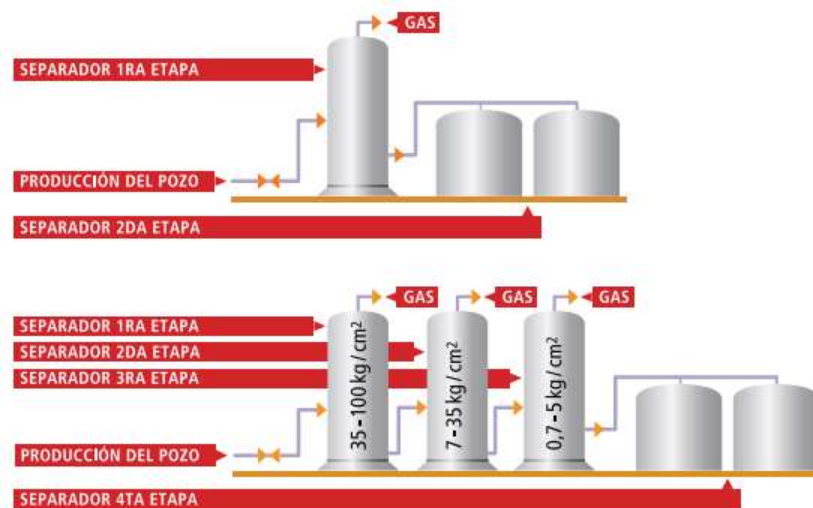


Fuente: Revista Soberanía Petrolera PDVSA

## Tanque:

El aceite y el agua, luego de completar su recorrido, pasan del separador a los tanques de almacenamiento. El agua se deposita bajo la acción de la gravedad y sale. Desde allí, el aceite se bombea a través de las tuberías de recolección al tanque de almacenamiento. En la figura 2.11 se tiene el esquema tipo de extracción de petróleo (PDVSA, 2016, p. 20)

Figura 2.11: Esquema de un campo de extracción de petróleo



Fuente: Revista Soberanía Petrolera PDVSA

## Patio de tanques:

Este es el punto final en el proceso de producción, que generalmente incluye una gran área de terreno en la que se construyen muchos tanques de gran capacidad, rodeados de altos muros de tierra y asfalto compactado que es probable que gotee o se oxide. Además, los tanques cuentan con potentes bombas y motores, un laboratorio principal, pozos o equipos para la separación adicional de agua libre y, en caso de ser necesario, instalación para la deshidratación y desalinización del aceite. Desde el parque reservorio, el petróleo se dirige a las refinerías para su

procesamiento los muelles para su transporte en tanqueros a otras partes del país o al exterior.  
(PDVSA, 2016, p. 21)

#### **2.1.4. Bombeo de hidrocarburos en pozo.**

Los principales tipos de bombeo son los siguientes:

- Bombeo mecánico
- Bombeo Electro sumergible
- Bombeo de cavidad progresiva
- Bombeo hidráulico

##### **Bombeo mecánico:**

La bomba se baja hasta el fondo de la tubería de producción, y se acciona por varillas movidas por un balancín, al que se imprime un movimiento de vaivén. Sus características son:

- Más usado en el mundo.
- Balancín de producción.
- Proceso de succión y transferencia.
- Válvulas fijas y viajeras.
- Válvulas de no retorno.

Componentes principales balancín de producción:

Motor: combustión interna o eléctrico.

Bomba de Sub-suelo: bomba de pistón de desplazamiento positivo, desde su profundidad de instalación hasta la superficie, que funciona por diferenciales de presión mediante bolas y asientos,



para permitir la entrada y sello de fluido en ciclos periódicos sincronizados. En la tabla 2.1 se expondrán las ventajas y desventajas del bombeo mecánico. (Berasategui Moreno, 2018, p. 44)

Tabla 2.1: Ventajas y desventajas del bombeo mecánico

Ventajas	Desventajas
El diseño es simple.	La efectividad del sistema puede verse afectada por la presencia del gas.
El sistema es eficiente y fácil de operar por el personal de campo.	La presencia de arenas ocasiona el desgaste severo del equipo.
Es aplicado en crudo pesado y altamente viscoso.	Requiere altos costos e mantenimiento
Puede utilizar combustible o electricidad como fuente de energía.	Profundidades limitadas.
El equipo puede operar a temperatura elevadas.	El equipo es pesado y ocupa mucho espacio.
Permite variar la velocidad y longitud de carrera para el control de la producción	La producción declina rápidamente.

Fuente: Presentación de la Universidad de Cantabria – Explotación del Petróleo.

### **Bombeo electro sumergible (BES):**

Es una bomba axial multipala montada sobre un eje vertical conectado a un motor eléctrico. Todo se baja al pozo a través de una tubería especial con un cable que transmite electricidad al motor.

(Berasategui Moreno, 2018, p. 47)

Características:

- Consiste en una serie de bombas centrífugas subterráneas para levantar líquidos del fondo del pozo.
- Transmite energía eléctrica, transmite energía mecánica (torque).
- Se basa en el principio de rotación de fluidos.
- Que a su vez genera energía potencial (a través de pasos).

En la tabla 2.2 se expondrán las ventajas y desventajas del bombeo electro sumergible.

Tabla 2.2: Ventajas y desventajas del bombeo electro sumergible

<b>Ventajas</b>	<b>Desventajas</b>
Puede levantar altos volúmenes de fluidos	Inversión inicial muy alta.
Maneja altos cortes de agua (aplicables off-shore)	Alto consumo de potencia.
Su vida útil puede ser muy larga.	No es rentable en pozos de baja producción.
Trabaja bien en pozos desviados	Los cables se deterioran al estar expuestos a temperaturas elevadas.
Puede usarse para inyectar fluidos a la formación.	Válidas para producción de gas y presencia de arena.
Fácil aplicación de tratamientos contra la corrosión y formaciones de escamas.	Las bombas y motor son susceptibles a fallos.

Fuente: Presentación de la Universidad de Cantabria – Explotación del Petróleo.

### **Bombeo de cavidad progresiva**

Las funciones son las siguientes:

- La operación de la bomba es simple, debido a que el rotor gira excéntricamente dentro del estator, se forman cavidades selladas entre las superficies de ambos estatores, conduciendo el líquido desde la bomba de succión hasta la descarga de la bomba.
- Este es un método de levantamiento artificial que se puede usar para producir líquidos que son muy viscosos y tienen pocas partes móviles.
- Son relativamente fáciles de mantener. (Berasategui Moreno, 2018, p. 50)

En la tabla 2.3 se expondrán las ventajas y desventajas del bombeo electro sumergible.

Tabla 2.3: Ventajas y desventajas del bombeo de cavidad progresiva

Ventajas	Desventajas
Bajo costo de instalación.	Su profundidad de operaciones recomendada es de 4.000 pies.
Bombea crudo de alta y baja gravedad API.	Requiere suministro de energía eléctrica.
Puede manejar hasta 100% de agua.	No se recomienda en pozos de más de 180°F.
El equipo de superficie puede ser transportado e instalado fácilmente.	El material elastómero es afectado por crudos con aromáticos
Bajo consumo de energía eléctrica.	
Bajos costos de mantenimiento.	

Fuente: Presentación de la Universidad de Cantabria – Explotación del Petróleo.

### Bombeo Hidráulico:

Consiste en bombas que son accionadas hidráulicamente por un líquido, generalmente agua, y crudo liviano proveniente del mismo pozo, conocido como fluido matriz. Las bombas se bajan a la tubería y se conducen desde la subestación. (Berasategui Moreno, 2018, p. 52)

Característica:

- Alcanza una profundidad de aprox. 18.000 conjuntos.
- Su potencia se transmite a través del líquido a presión bombeado a través de la tubería.
- Este fluido es utilizado por una bomba subterránea que actúa como transformador.
- Convertir la energía del fluido en energía potencial o energía de presión.

En la tabla 2.4 se expondrán las ventajas y desventajas del bombeo hidráulico.

Tabla 2.4: Ventajas y desventajas del bombeo de cavidad progresiva.

<b>Ventajas</b>	<b>Desventajas</b>
Puede ser usados en pozos profundos (+/- 18.000 pies).	Costo inicial alto
Puede ser utilizado en pozos desviados, direccionales y sitios inaccesibles.	Las instalaciones de superficie presentan mayor riesgo, por la presencia de altas presiones.
Varios pozos pueden ser controlados y operados desde una instalación central de control.	Altos costos en la reparación del equipo.

Fuente: Presentación de la Universidad de Cantabria – Explotación del Petróleo.

## **2.2. SEPARACIÓN DE HIDROCARBUROS**

### **2.2.1. Principios de separación**

Es importante tener en cuenta los diferentes estados en los que pueden estar presentes los líquidos y el efecto que las diferentes fuerzas o principios físicos pueden tener sobre ellos. Los principios fundamentales de la separación física de vapores, líquidos o sólidos son: gravedad, fuerza centrífuga y colisión o aglomeración de partículas. Cada separación puede utilizar una o más de estas bases, pero las fases líquidas siempre deben ser inmiscibles y de diferentes densidades para que se produzca la separación. (Rodríguez, 2006, p. 24)

#### **2.2.1.1. Separación por gravedad**

Este es el mecanismo de separación más utilizado porque el material requerido es muy simple. Cada extensión de flujo actúa como un depósito bajo la acción de la gravedad de las gotas de líquido suspendidas en la corriente de gas. La deposición ocurre como resultado de una reducción en el caudal, en los separadores, la sedimentación por gravedad ocurre principalmente en la sección secundaria, también llamada sección de sedimentación por gravedad. Si el flujo es verticalmente ascendente, como en los separadores verticales, las partículas de líquido separadas

caen contra el flujo de gas. Estas partículas líquidas, al caer bajo la influencia de la gravedad, se acelerarán hasta que la resistencia sea equilibrada por la fuerza de la gravedad. Después de este punto, las partículas continúan descendiendo a una velocidad constante, conocida como velocidad de sedimentación o velocidad posterior. La tasa de caída calculada de una gota de líquido de un diámetro dado indica la velocidad máxima que debe tener el gas para que las gotas de ese diámetro o más grandes se separen. (Rodríguez, 2006, p. 24)

#### **2.2.1.2. Numero de Reynolds**

Permite identificar la naturaleza del flujo, es decir, si es laminar o turbulento, y su posición relativa en una escala que muestra la importancia relativa de las tendencias turbulentas a laminares. (Rodríguez, 2006, p. 28)

#### **2.2.1.3. Fuerza Centrifuga**

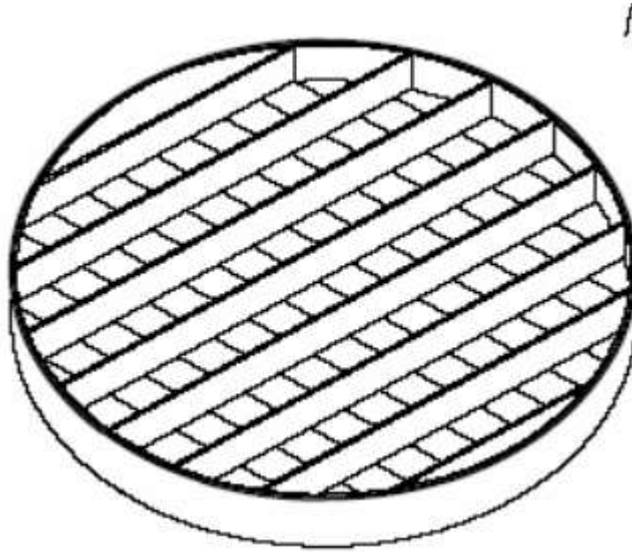
La fuerza centrífuga sobre las partículas líquidas suspendidas en una corriente de gas puede ser mucho mayor que la fuerza gravitatoria sobre las propias partículas. Este principio de separación mecánica se utiliza en separadores tanto en separadores primarios como en algunos tipos de separadores de gases, por ejemplo, extractores ciclónicos. Sin embargo, la velocidad del gas en la pared del tubo ciclónico. Se recomienda que la velocidad del gas no exceda un cierto valor crítico. (Rodríguez, 2006, p. 28)

#### **2.2.1.4. Choque de Partículas**

Este mecanismo de separación es quizás el más utilizado para eliminar partículas líquidas finas suspendidas en un flujo. Las partículas de fluido que se mueven en la corriente de gas chocan con los obstáculos sobre los que se depositan. La separación por colisión se utiliza principalmente en los deshumidificadores de tipo agitador y cordel, se denomina distancia de retención, la distancia

que recorre una partícula de cierto diámetro a través de la trayectoria del flujo de gas. La figura 2.12 es una ampliación de un extractor de alambre. (Rodríguez, 2006, p. 28)

Figura 2.12: Extractor de alambre



Fuente: Trabajo de Grado-Diseño y evaluación de separadores Bifásicos y Trifásicos

La colisión de partículas durante la separación de una mezcla depende del tiempo de separación de dos líquidos inmiscibles, proceso que ocurre cuando chocan dos gotas de líquido diferentes.

Si se somete un par de gotas a una determinada presión y turbulencia del medio, la energía cinética de este par de gotas hace que se agrupen de tal forma que formen un todo, es decir, se genera energía de cohesión. Por lo tanto, cuando se rompe este contacto, el proceso se completa y se denomina aglomeración. (Rodríguez, 2006, p. 28)

### **2.2.2. Función de un separador**

Los separadores de mezcla gas/líquido forman parte de un grupo mayor de equipos involucrados en el proceso físico de separación de fases: sólidos, líquidos y gases. El equipo está diseñado para separar diferentes fases: gas-líquido, sólido-sólido, sólido-líquido-gas. El propósito básico de

cualquier separador es separar la fase deseada de las otras fases lo más completamente posible, para lo cual se debe elegir un proceso físico apropiado.

#### **2.2.2.1. Funcionalidad en Separadores Gas-Líquido**

El término "separador de gas-líquido" en términos de campo se refiere a un recipiente presurizado utilizado para separar fluidos de pozos de producción de petróleo y gas en componentes gaseosos y líquidos a una presión y temperatura. Cuando se trata de mezclas de hidrocarburos depositadas en el fondo del recipiente, a la temperatura a la que se produce la separación, la presión de vapor del fluido es igual a la presión de operación. (Rodríguez, 2006, p. 31)

#### **2.2.2.2. Funcionalidad en Separadores Gas-Petróleo-Agua.**

Se refiere a la separación de gas, petróleo y agua en un solo separador, todos los separadores básicos (verticales y horizontales) se pueden utilizar para la separación trifásica, independientemente de su forma. En este caso, el agua es una impureza omnipresente en la mezcla que debe eliminarse antes de liberar el producto final. Debe evitarse la emulsificación del agua con cualquier otro fluido. Si se produce emulsificación, debe descomponerse utilizando productos químicos, calor u otros métodos. (Rodríguez, 2006, p. 31)

#### **2.2.3. Explicación del proceso de Separación**

En el caso de una mezcla gas-líquido, la mezcla de fases ingresa al separador y, si está presente, Colisión con el conector interno en la entrada, lo que resulta en cantidad de movimiento de la mezcla, lo que conduce a la separación de fases. Próximo, En la sección de sedimentación (espacio libre) del separador, la gravedad actúa Por encima del fluido, se permite que el líquido abandone la fase gaseosa y caiga a Fondo del separador (sección de recogida de líquidos). Esta sección

proporciona Tiempo de retención suficiente para que funcione el equipo aguas abajo satisfactoriamente, si se especifica, liberar el líquido Burbujas atrapadas. (Rodriguez, 2006, p. 32)

#### **2.2.4. Etapas de la separación**

Son las siguientes:

- Sección primaria
- Sección secundaria
- Sección de extracción de Neblina
- Segregación Final

Se las describirá cada etapa, son las siguientes:

##### **Sección primaria:**

El flujo de fluido que ingresa al separador ingresa a alta velocidad, lo que resulta en un flujo turbulento entre la fase gaseosa y la fase líquida. Por lo tanto, se debe disipar el gran impulso que tiene el flujo de fluido en la entrada del separador. Para reducir el impulso y reducir la turbulencia, se pueden utilizar deflectores o cualquier otra técnica que cree una fuerza centrífuga en el fluido, separando así un volumen de líquido del gas. (Rodriguez, 2006, p. 32)

##### **Sección secundaria:**

El principio de separación más importante en esta sección es separar la gravedad del líquido de la corriente de gas después de que se reduce su velocidad. El desempeño en esta área depende de las propiedades del gas y del líquido, el tamaño de las partículas y el grado de turbulencia del gas. El grado de turbulencia debe mantenerse al mínimo, medido por el número de Reynolds, y algunos



diseños incluyen deflectores internos para reducir la turbulencia y dispersar la espuma. Los desviadores también pueden actuar como eliminadores de gotas. (Rodríguez, 2006, p. 33)

### **Sección de extracción de Neblina:**

De esta forma, se separarán las partículas más pequeñas de líquido que contenga gas. En ambos métodos, las pequeñas gotas de líquido se separan de la corriente de gas en forma de gotas grandes (aglomeración), que luego caen en la zona de recolección de líquido. (Rodríguez, 2006, p. 33)

### **Segregación Final:**

En esta etapa se descargan los diferentes líquidos, el líquido libre de gases y el líquido libre de gases, en las condiciones de operación especificadas, evitando el reordenamiento de las partículas con diferentes fases y la formación de espuma. Esto requiere una mínima retención de líquidos y una mínima ingesta de alimentos. El amortiguador de vórtice se puede ubicar sobre los puertos de salida de líquido y/o los puertos para evitar que el gas o el aceite entren en el líquido residual. (Rodríguez, 2006, p. 33)

#### **2.2.5. Variables que afectan la eficiencia de separación.**

Hay muchos factores que pueden afectar la eficiencia y eficacia de la separación en el proceso de separación. Saber cuál de estos se considera el más importante y esencial para desarrollar un diseño de divisor efectivo. (Rodríguez, 2006, p. 33) Los factores son:

##### **2.2.5.1. Viscosidad dinámica del gas.**

El efecto de la viscosidad del gas en la separación se puede ver en las fórmulas que determinan la velocidad de caída de las partículas líquidas. La viscosidad del gas se utiliza en el parámetro NRE, que determina el valor del coeficiente de arrastre de la ley de Stokes para determinar la tasa de

caída de partículas de un diámetro determinado. En consecuencia, a medida que aumenta la viscosidad del gas, disminuye la tasa de sedimentación y, por lo tanto, la tasa de flujo de gas del separador. (Rodriguez, 2006, p. 34)

#### **2.2.5.2. Temperatura.**

Agregar calor a la corriente de entrada de aceite/agua es el método de separación tradicional. Un aumento en la temperatura reduce la viscosidad de la fase de aceite, teniendo. También funciona para disolver cristales finos de parafina y asfalto, anulando su acción como potencial emulsionante. Las temperaturas generalmente están en el rango de 100-160 ° F. Cuando se procesa crudo pesado, las temperaturas pueden estar por encima de los 300 ° F. La adición de calor puede causar una pérdida significativa de hidrocarburos livianos si se alcanza el punto de ebullición de la mezcla. Este resultado se conoce como "contracción" del aceite. Las partículas que salen de la fase de aceite se pueden eliminar o comprimir y vender con el gas. Incluso cuando se vende junto con el gas, todavía existe la posibilidad de pérdida de ganancias netas debido al proceso de introducción de volumen de líquido en el gas.

Al incrementar la temperatura a la cual ocurre el proceso, se presenta la desventaja de hacer que el petróleo crudo que se recupera en los tanques de almacenamiento se vuelva más pesado, y pierda valor comercial. Los compuestos livianos salieron de la mezcla debido que se alcanzó el punto de ebullición de la misma y el líquido remanente tiene una gravedad API más baja.

El aumento de la temperatura puede disminuir la gravedad específica y la presión de operación a la que el aceite se separa del agua, por lo que dependiendo de las características del aceite puede aumentar o disminuir la gravedad específica. Solo cuando la temperatura está por debajo de los 200 °F, el cambio en la gravedad específica es insignificante. Los gases que se liberan cuando el

aceite se expone al calor pueden dañar el separador si no se diseña correctamente debido a la mayor presencia de gas en la aglomeración. Si se libera demasiado gas, puede crear suficiente turbulencia para evitar la aglomeración, es decir, separar el gas de la mezcla. (Rodríguez, 2006, p. 34)

#### **2.2.5.3. Densidad del líquido y gas.**

La densidad del líquido y el gas afectan la capacidad de los separadores. La capacidad de manejo de gas de un separador, es directamente proporcional a la diferencia de densidades del líquido y del gas e inversamente a la densidad del gas. (Rodríguez, 2006, p. 36)

#### **2.2.5.4. Tamaño de las partículas de líquidos.**

El tamaño de las partículas en la corriente de gas es un factor importante para determinar la tasa de deposición durante la separación por gravedad y la separación centrífuga cuando la separación se realiza por fuerza. La velocidad promedio del gas en la separación secundaria corresponde a la velocidad de sedimentación de una gota de líquido de un diámetro dado, que se puede considerar como el diámetro del fondo. Teóricamente, todas las gotas más grandes que el sustrato deben eliminarse. De hecho, las partículas más pequeñas que el diámetro básico se separan, mientras que las partículas con un diámetro mayor no se separan. Esto se debe a la turbulencia del flujo y al hecho de que algunas partículas líquidas tienen una velocidad inicial mayor que la velocidad promedio de la corriente de gas. La separación del flujo horizontal también está sujeta a efectos similares. El tamaño de partícula del líquido formado en la corriente de líquido y gas en ausencia de una fuerte agitación es lo suficientemente grande como para lograr un buen rendimiento de separación. (Rodríguez, 2006, p. 36).

#### **2.2.5.5. Presión de separación.**

Es uno de los elementos de separación más importantes en el campo de la recuperación de líquidos. Siempre hay una presión de separación óptima para cada situación particular. A veces, cuando se baja la presión de separación, principalmente para separar el gas y el condensado, aumenta la recuperación de líquido. Sin embargo, es importante considerar la economía del volumen de líquido frente a la compresión adicional que puede ser necesaria para transportar el gas. La presión de separación también afecta la eficiencia del separador; Un aumento de presión aumenta la separación de gases y viceversa. (Rodríguez, 2006, p. 37)

### **2.3. POSIBLES ANOMALÍAS EN UN SEPARADOR.**

#### **2.3.1. Formación de espumas dentro del separador.**

Cuando se quiere mezclar 2 sustancias en cualquier estado pueden suceder dos cosas:

- Que ambas se disuelvan mutuamente constituyendo una sola fase, es decir una solución homogénea.
- Que ambas no se disuelvan entre si bien lo hagan solo parcialmente, en este caso se forma una dispersión.

Una espuma es en muchos aspectos análoga a una emulsión, está constituida por glóbulos dispersos de gas en vez de líquido, en fase líquida.

No es usual que las sustancias químicamente puras formen emulsiones o espumas estables, excepto cuando se encuentra presente una tercera sustancia capaz de generar una interfase de características diferentes a las fases dispersas. (S.R.L.)

**Razones para la formación de espuma en los separadores:** La inexistencia de formación de espumas en el separador garantiza una eficiente separación del gas y condensado.

La formación de espuma dentro del separador muy posiblemente sea por:

- Porque un arrastre líquido en la corriente gaseosa. Denotando aumento en la salida del gas.
- Aumento brusco de los hidrocarburos a separar, es decir el aumento brusco en la velocidad.
- Presencia de ácidos y parafinas.
- Aumento de temperatura favorece a la formación de espumas ya que reduce la tensión gas líquido y la viscosidad.

### **2.3.2. Formación de vórtices:**

Un rompedor de vórtices es un dispositivo utilizado en ingeniería para detener la formación de un vórtice cuando se drena un fluido (líquido o gas) de un recipiente como un tanque o un separador de vapor-líquido. La formación de vórtices puede arrastrar vapor en la corriente líquida, lo que conduce a una mala separación en las etapas del proceso, como la destilación o una caída de presión excesiva, o provoca la cavitación de las bombas aguas abajo. (hmong, 2017)

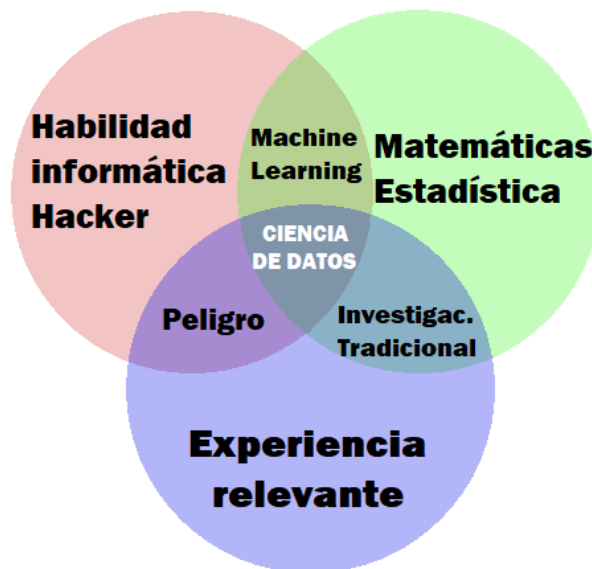
### 3. CAPÍTULO 3 CIENCIA DE DATOS.

La ciencia de datos es un paso evolutivo en campos interdisciplinarios como la inteligencia empresarial que integra la informática, el modelado, las estadísticas, el análisis y las matemáticas en un solo proceso.

Sin duda la ciencia de datos hoy por hoy se utiliza en diferentes áreas como ser: medicina, energía, fraudes bancarios.

Donde el estado del arte del mismo son las redes neuronales. En la figura 3.1 se expresa los conocimientos y áreas necesarias para llevar a cabo un proyecto en ciencia de datos. Las cuales son: Conocimiento del tema, programación y estadística. Y en la intersección de estas esta la ciencia de datos. (Mora)

Figura 3.1: Ramas de la ciencia de datos



Fuente: Presentación de Universidad Javeriana-MS in Urban Data Science.

### **3.1. INTELIGENCIA ARTIFICIAL**

Se han desarrollado muchas definiciones a lo largo de los años, pero la comunidad científica aún no puede encontrar un solo concepto. Sin embargo, podemos llegar a saber lo que realmente significa AI rastreando algunos de sus significados:

“La inteligencia artificial es la ciencia y la tecnología que permitirá a las máquinas replicar la inteligencia humana. (López, González, Meseguer, 2017)”

“La IA es la ciencia y tecnología que posibilita el diseño y la programación de computadoras para realizar tareas que requieren inteligencia” (López, González, Meseguer, 2017)”

### **3.2. MACHINE LEARNING**

El aprendizaje automático, también conocido como aprendizaje automático, es una rama de la inteligencia artificial que tiene como objetivo desarrollar algoritmos para organizar y codificar datos utilizados como base del aprendizaje automático. Gracias a los conocimientos adquiridos, serás capaz de predecir y dar respuesta a problemas específicos en diferentes áreas del conocimiento. Los programadores deben crear algoritmos cada vez más precisos para que la máquina pueda aprender de un conjunto de variables y lograr un mayor porcentaje de certeza. (Ligdi, n.d.)

Dentro del Machine Learning existen 2 tipos de aprendizajes: Supervisado y no supervisado.

#### **3.2.1. Aprendizaje Supervisado**

Los algoritmos de aprendizaje supervisado se basan en un conjunto pre-marcado de datos de entrenamiento. Por etiquetado, queremos decir que, para cada instancia del conjunto de datos de entrenamiento, conocemos el valor de su atributo de destino. Esto permitirá que el algoritmo

"aprenda" una función capaz de predecir un atributo objetivo para el nuevo conjunto de datos. Se supervisan dos grandes grupos de algoritmos:

Algoritmo de regresión cuando el resultado de la predicción es un atributo numérico. Algoritmos de clasificación cuando el resultado de la predicción es un atributo categórico.

Ejemplos de este tipo de algoritmo son modelos:

- Regresión lineal simple.
- Regresión lineal múltiple.
- Regresión logística Árboles de decisión

### **Regresión lineal**

El modelo de Regresión Lineal es tan simple que muchos argumentan que no es digno de ser clasificado como Machine Learning. Este algoritmo es un método estadístico que nos permite resumir y estudiar las relaciones entre dos variables continuas cuantitativas.

La Regresión Lineal es una técnica paramétrica utilizada para predecir variables continuas, dependientes, dado un conjunto de variables independientes. Es de naturaleza paramétrica porque hace ciertas suposiciones basadas en el conjunto de datos. Si el conjunto de datos sigue esas suposiciones, la regresión arroja resultados increíbles, de lo contrario, tiene dificultades para proporcionar una precisión convincente. (Abuin, 2007)

Matemáticamente, la regresión usa una función lineal para aproximar o predecir la variable dependiente dada como la cual esta representa en la ecuación 1:

Ecuación 1: Regresión Lineal Simple

$$Y = a + bX \dots \dots (1)$$



Donde:

$Y$ = Es la variable dependiente o variable a predecir.

$X$ = Es la variable independiente o la variable que usamos para hacer una predicción.

$a$ = Es la pendiente o el valor que debe ser determinado, se le conoce como coeficiente y es una especie de magnitud de cambio que pasa por  $y$  cuando  $x$  cambia.

$b$ = Es la constante que debe ser determinada, se le conoce como intercepto porque cuando  $x$  es igual a 0, entonces  $y = b$ .

El cálculo del coeficiente que choca con la ordenada esta expresada en la ecuación 2:

Ecuación 2: Determinación del coeficiente 'a' en la regresión lineal simple

$$a = \frac{\sum(x_i - x_{media})(y_i - y_{media})}{\sum(x_i - x_{media})^2} \dots \dots (2)$$

Donde:

$x_i$ = Es el valor que toma la variable  $X$ . Según los datos.

$x_{media}$ = Es el promedio de los valores de la variable  $X$ .

$y_i$ = Es el valor que toma la variable  $Y$ . Según los datos.

$y_{media}$ = Es el promedio de los valores de la variable  $Y$ .

El cálculo del coeficiente que explica el valor de la pendiente es explica en la ecuación 3:

Ecuación 3: Determinación del coeficiente 'b' en la regresión lineal simple

$$b = y_{media} - a(x_{media}) \dots \dots (3)$$

Donde:

$x_{media}$  = Es el promedio de los valores de la variable X.

$y_{media}$  = Es el promedio de los valores de la variable Y.

### **Suposiciones en la regresión lineal**

La regresión es un método paramétrico, por lo que hace suposiciones, veamos algunas de ellas:

- Existe una relación lineal y aditiva entre las variables dependientes e independientes. Linealmente, es decir, el cambio en la variable dependiente por unidad de cambio en la variable independiente es constante. Además, se refiere al hecho de que "x" e "y" son independientes de otras variables.
- No debe haber correlación entre las variables independientes. La presencia de correlaciones en las variables independientes conduce a la multifinalidad. Si las variables están correlacionadas, será difícil que el modelo determine la verdad de las variables independientes para las variables dependientes. Las condiciones de error deben tener una varianza constante y no se pueden correlacionar.
- La presencia de correlación en términos de errores se denomina auto correlación y tiene una gran influencia en los coeficientes de regresión y los valores de error estándar porque se basan en la suposición de términos de error no correlacionados.
- La variable dependiente y las condiciones de error deben estar normalmente distribuidas.

La presencia de estos supuestos hace que la regresión sea muy limitada, es decir, la efectividad del modelo de regresión depende de la implementación de estos supuestos. Cuando se violan estos supuestos, la regresión producirá predicciones sesgadas y engañosas, por lo que esto debe tenerse en cuenta al trabajar con este algoritmo. (Ligdi, n.d.)

## Regresión lineal múltiple

La Regresión Lineal Múltiple es muy parecida a la simple, pero acá se manejan múltiples variables independientes que contribuyen a la variable dependiente. Se manejan múltiples coeficientes y a su vez es computacionalmente más compleja debido a las variables añadidas.

Ecuación 4: Regresión lineal Múltiple

$$Y = a_1X_1 + a_2X_2 + \dots \dots a_nX_n + b \dots \dots (4)$$

Donde:

$X_{1,\dots,n}$  = Variables independientes.

$Y_{1,\dots,n}$  = Variable dependiente.

$a_{1,\dots,n}$  = Coeficientes.

$b$  = Intersección.

Como puedes ver, esto es muy similar a una regresión lineal simple, solo que aquí incluimos  $n$  variables explicativas con sus respectivos coeficientes.

En general, cuando se trata de regresiones lineales múltiples, no tomamos en cuenta todas las variables explicativas a la vez y luego comenzamos a minimizar la función de error. Lo primero que debe hacer es concentrarse en seleccionar las variables independientes que mejor pueden contribuir a la variable dependiente. Para hacer esto, necesitamos construir una matriz de correlación para todas las variables independientes e incluir la variable dependiente. El valor de la correlación nos da una idea de qué variables son importantes y por qué factores. A partir de esta matriz, seleccionamos las variables independientes en orden decreciente de valor de correlación

y ejecutamos un modelo de regresión para estimar los coeficientes minimizando la función de error. Nos detuvimos cuando no hubo una mejora notable en la función de revisión, incluida otra función independiente. Este método aún puede complicarse cuando hay un gran número de características independientes que contribuyen significativamente a la determinación de nuestra variable dependiente. Este método está mucho mejor explicado en otro post del mismo blog.

Al usar este método, tenga en cuenta que agregar más variables explicativas no mejora la regresión ni hace mejores predicciones. La regresión lineal simple y múltiple tiene diferentes casos de uso, uno de los cuales no es mejor. En algunos casos, agregar más variables explicativas puede empeorar las cosas, esto se conoce como sobreajuste. Por otro lado, cuando agregas más variables independientes, se crean relaciones entre ellas. Por lo tanto, las variables independientes no solo están potencialmente relacionadas con la variable dependiente sino también potencialmente relacionadas entre sí, lo que se denomina colinealidad. El escenario óptimo es que todas las variables independientes estén correlacionadas con la variable dependiente pero no entre sí. Recuerde que muchas de las suposiciones explicadas para la regresión lineal simple se aplican aquí, por lo que lo animo a leer esta guía para complementar la información que se presenta aquí. (Carrasquilla Batista, 2016)

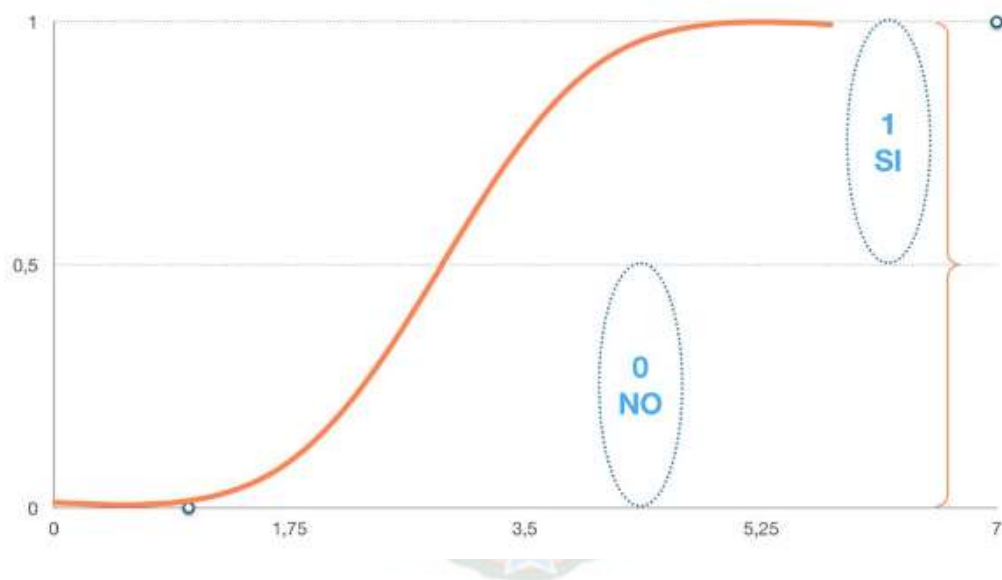
### **Regresión logística**

La regresión logística es un método estadístico que predice clases binarias. La variable de resultado o la variable objetivo no es dicotómica. Dicotómico significa que solo hay dos clases posibles. Por ejemplo, se puede utilizar para tareas relacionadas con la investigación del cáncer o para calcular la probabilidad de que ocurra un evento. La regresión logística es uno de los algoritmos de aprendizaje automático más simples y más utilizados para la clasificación de dos clases. Es fácil de implementar y puede usarse como base para cualquier problema de clasificación

binaria. La regresión logística describe y evalúa la relación entre una variable dependiente binaria y las variables independientes. (Ligdi, n.d.)

En general, este algoritmo se puede usar para varias tareas de clasificación que tengan como respuesta dependiente ser VERDADERO o FALSO, en la figura 3.2 se explica gráficamente cómo funciona esta función sigmoide. También, se conoce como regresión logística a la función utilizada en el núcleo del método, la función logística también se conoce como función sigmoide. Esta función es una curva en S que puede tomar cualquier número de valores reales y asignar valores de 0 a 1. (Ligdi, n.d.)

Figura 3.2: Función sigmoide.



Fuente: Pagina web 'Aprende IA con Ligdi González'.

La ecuación de una regresión logística será descrita en la ecuación 5:

Ecuación 5: Regresión Logística

$$p = \frac{1}{1 + e^{-y}} \dots \dots (5)$$

Donde:

$p$  = Probabilidad de pertenecer a SI o NO entre 0 y 1.

$e$  = Exponente logarítmico.

$y$  = Este valor nace de la ecuación (4).

### Arboles de decisión

Un árbol de decisión es una técnica de aprendizaje supervisado que predice valores de respuesta aprendiendo reglas de decisión derivadas de características. Se pueden utilizar tanto en el contexto de la regresión como en el contexto de la clasificación. Los árboles de decisión funcionan dividiendo el espacio funcional en varias áreas rectangulares simples separadas por divisores de ejes paralelos. Para obtener la predicción para un caso particular, se utiliza el método promedio o el modo de respuesta de los casos de entrenamiento en la partición que contiene el nuevo caso.

Veamos matemáticamente las funciones del árbol de decisión (Ligdi, n.d.):

Ecuación 6: Árbol de decisiones

$$f(x) = \sum_{m=1}^M w_m \theta(x; v_m) \dots \dots (6)$$

Donde:

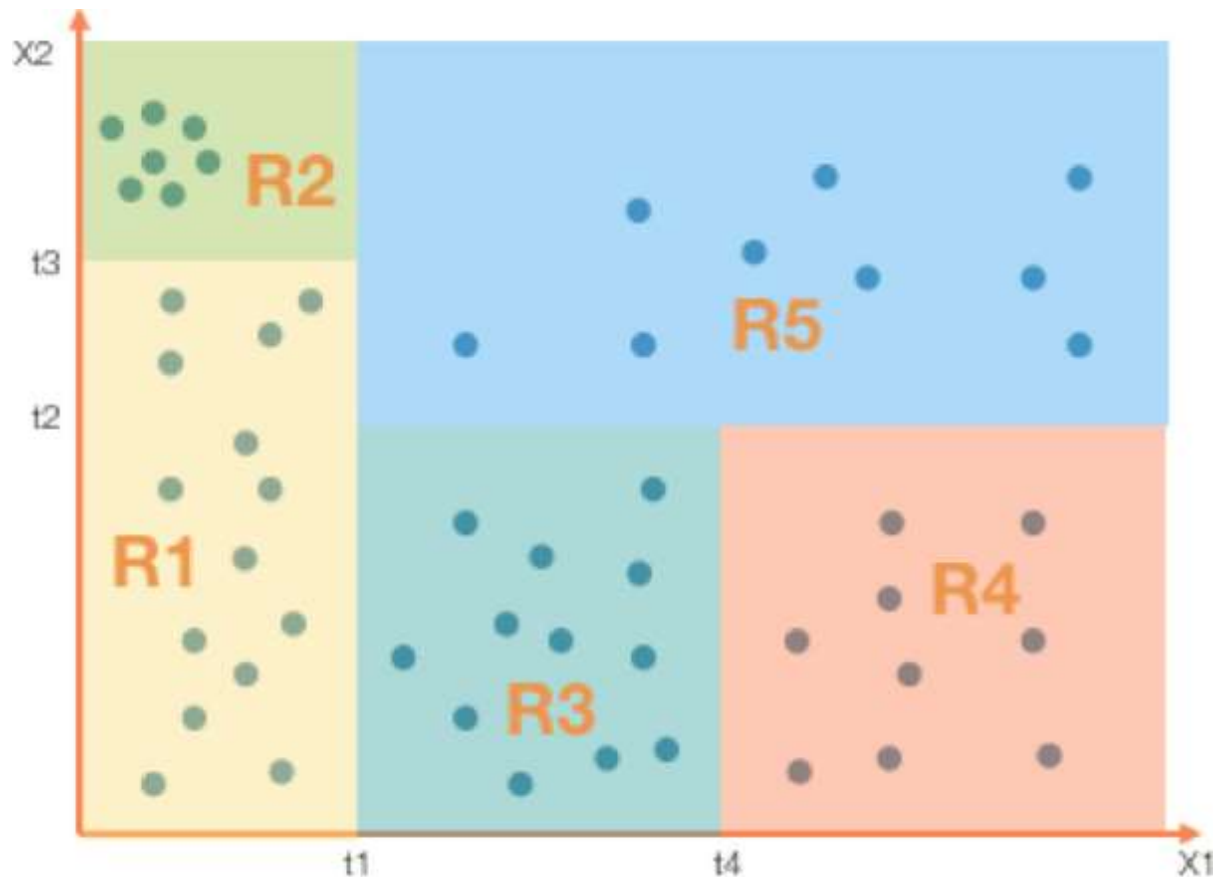
$W_m$  = Es la respuesta media en una región partícula.

$V_m$  = Representa cómo se divide cada variable en un valor de umbral particular.

$y$  = Este valor nace de la ecuación (4).

En la figura 3.3 considera un ejemplo con dos variables de características ( $X_1$  y  $X_2$ ) y una respuesta numérica “ $y$ ” que sería la región limitada con las letras mayúsculas R1, R2, R3, R4, R5.

Figura 3.3: Graficas los diferentes valores por regiones.



Fuente: Pagina web ‘Aprende IA con Ligdi González’.

### 3.2.2. Aprendizaje No Supervisado

La pregunta que se debe hacer, ¿Que se hace cuando los datos no tienen etiquetas?

En este punto, se debe aplicar el aprendizaje no supervisado, que es un grupo de algoritmos y enfoques de aprendizaje automático que funcionan con datos de "hechos sin fundamento".

En la serie anterior de publicaciones, explicamos los algoritmos de aprendizaje automático supervisado. Los algoritmos de aprendizaje no supervisados toman muestras del conjunto de datos sin referencia a resultados conocidos o etiquetados. A diferencia del aprendizaje supervisado, el aprendizaje no supervisado no se puede aplicar directamente a un problema de regresión o clasificación porque no se sabe cuál podría ser el valor inicial, lo que impide que el algoritmo se entrene de la forma habitual. En su lugar, puede utilizar el aprendizaje no supervisado para explorar la estructura subyacente de sus datos. Los algoritmos de aprendizaje no supervisados permiten tareas de procesamiento más complejas que el aprendizaje supervisado. Sin embargo, el aprendizaje no supervisado puede ser más impredecible que otros métodos de aprendizaje natural. Los algoritmos de aprendizaje no supervisados se utilizan para agrupar datos no estructurados en función de sus similitudes y diferencias en el conjunto de datos. El término "no supervisado" se refiere al hecho de que un algoritmo no supervisado es lo mismo que un algoritmo de aprendizaje supervisado (Ligdi, n.d.).

### **3.3. DEEP LEARNING**

El aprendizaje profundo es un tema que cada vez adquiere mayor relevancia en el campo de la inteligencia artificial (IA). Siendo una subcategoría del aprendizaje automático, el aprendizaje profundo trata del uso de redes neuronales para mejorar cosas tales como el reconocimiento de voz, la visión por ordenador y el procesamiento del lenguaje natural. Rápidamente se está convirtiendo en uno de los campos más solicitados en informática. En los últimos años, el aprendizaje profundo ha ayudado a lograr avances en áreas tan diversas como la percepción de objetos, la traducción automática y el reconocimiento de voz (todas ellas áreas especialmente complejas para los investigadores en IA). (BBVA).



### **3.4. MINERÍA DE DATOS**

La minería de datos es el conjunto de técnicas encaminadas a la extracción y tratamiento de conocimiento procesable de gran cantidad de datos, implícito de las bases de datos.

#### **3.4.1. Metodología KDD**

La abreviatura KDD quiere decir Knowledge Discovery in Database por su traducción al castellano es ‘Descubrimiento de Conocimiento en Bases de Datos’, este define como:

La metodología KDD es un proceso secuencial que en el busca encontrar patrones, para comer la situación que se actual o se puede plantear en un futuro de una organización por medio de modelo matemáticos. (minerva, n.d.)

#### **3.4.2. Fases de metodología KDD**

Las Fases que corresponde a la metodología KDD son:

- Selección.
- Preprocesamiento/Limpieza.
- Transformación/Reducción.
- Minería de datos (Data Mining).
- Interpretación/evaluación.

##### **3.4.2.1. Selección**

Durante la fase de selección, tras determinar los conocimientos relevantes y priorizados y definir el objetivo del proceso KDD, se genera un conjunto de datos objetivo desde el punto de vista del usuario final seleccionando todo el conjunto de datos o su muestra representativa que explora el

proceso de tener lugar y hecha ya la selección de datos varía según sus objetivos comerciales (minerva, n.d.).

En resumen:

- Determinar los objetivos que se deseas lograr.
- Debe ser preciso y medible.

#### **3.4.2.2. Preprocesamiento**

El paso de pre procesamiento/limpieza (limpieza de datos) analiza la calidad de los datos, aplica operaciones básicas como la eliminación de datos ruidosos y la elección de una estrategia para tratar con datos desconocidos (faltantes y vacíos), datos cero, datos duplicados y datos estadísticos métodos alternativos. En esta etapa, lo más importante es interactuar con el usuario o analista. Los datos de ruido son valores que están fuera del rango esperado; Se relacionan principalmente con errores humanos, cambios en el sistema, disponibilidad intempestiva de información y fuentes de datos heterogéneas. Los datos vacíos desconocidos son datos que no tienen valor en el mundo real, y los datos faltantes son datos con valor no capturado. Los datos vacíos son datos indefinidos permitidos por el sistema de gestión de bases de datos relacionales. Durante la limpieza, todos estos valores se ignoran, se reemplazan con el valor predeterminado o el más cercano, es decir, se utilizan índices de tipo estadístico como media, moda, mínimo y máximo para anularlos (minerva, n.d.).

#### **3.4.2.3. Transformación**

Durante la fase de transformación/reducción de datos, se buscan funciones útiles para la presentación de datos según el propósito del proceso. Las técnicas de transformación o reducción dimensional se utilizan para reducir el número efectivo de variables para incluir o encontrar

representaciones de datos invariantes. Las técnicas de reducción de dimensiones pueden simplificar una tabla de base de datos horizontal o verticalmente. La reducción horizontal es la eliminación de tuplas idénticas reemplazando el valor del atributo con otros valores de un nivel superior, en alguna jerarquía de valores categóricos, o por valores arbitrarios continuos (por ejemplo, edad por diferentes edades). La reducción vertical implica la eliminación de atributos superfluos o redundantes relacionados con el problema, como la eliminación de llaves, la eliminación de columnas funcionalmente dependientes (por ejemplo, edad y fecha de nacimiento). Se utilizan métodos de mitigación como agregación, compresión de datos, histograma, segmentación, personalización basada en entropía, muestreo y más (minerva, n.d.).

#### **3.4.2.4. Minería de datos**

Las técnicas de minería de datos generan modelos predictivos o descriptivos. Los modelos predictivos tienen como objetivo estimar valores futuros o desconocidos de variables de interés denominadas variable objetivo, variable dependiente o variable de clase mediante el uso de otras variables conocidas como variable independiente o predicciones de variables, como predecir nuevos clientes si son buenos o malos en función de su historia familiar. Estado, edad, género y ocupación, o para determinar si están abandonando los estudios para nuevos estudiantes en función de la región de origen, la facultad, la clase, el género, la edad y el promedio de calificaciones. Las tareas de predicción incluyen clasificación y regresión. Los modelos descriptivos revelan patrones que explican o resumen datos; se utilizan para estudiar las propiedades de los datos que se están descubriendo, no para predecir nuevos datos, como identificar grupos de personas con gustos similares o identificar patrones de compra de clientes en un área particular de la ciudad. Las tareas descriptivas incluyen reglas de asociación, patrones de secuencia, grupos y correlaciones. Por lo tanto, la selección de un algoritmo de minería de

datos implica elegir los métodos que se utilizarán al buscar patrones en los datos, así como decidir los modelos y parámetros más apropiados según el tipo de datos (categoría, número) usar (minerva, n.d.).

#### **3.4.2.5. Evaluación**

Los patrones encontrados se interpretan y pueden volver a los pasos anteriores en iteraciones posteriores. Este paso puede incluir la representación de las plantillas extraídas, la eliminación de plantillas innecesarias o innecesarias y la traducción de plantillas útiles en términos fáciles de usar. De lo contrario, el conocimiento encontrado se consolida para su integración en otro sistema de seguimiento o simplemente para documentar y comunicar a las partes interesadas; también para examinar y resolver posibles conflictos con conocimientos previamente descubiertos (minerva, n.d.).

### **3.5. MÉTRICAS DE EVALUACIÓN.**

Las métricas son aquellos datos expresados numéricamente que nos sirven para analizar el rendimiento de un determinado modelo. Digamos que, gracias a las métricas, podemos saber si estamos cumpliendo un objetivo. Y, en caso contrario, podremos plantear nuevos modelos (oleshop, n.d.).

Se planea usar 2 modelos estos son:

- Regresión lineal simple
- Regresión lineal múltiple

Se escoge estos modelos ya que los valores que tenemos tanto en producción de Oíl, Gas y Agua son tipo de DATOS NUMERICOS CONTINUOS.

Las métricas usadas en estos modelos serán principalmente 2, las cuales son:

- Coeficiente de Determinación.
- Error de Cuadrático medio.

### 3.5.1. Coeficiente de Determinación.

El coeficiente de determinación es la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, también conocido como R-cuadrado, refleja la bondad de ajuste del modelo a las variables que intenta explicar. Es importante saber que el resultado del coeficiente de determinación está entre 0 y 1. Cuanto más cercano sea su valor a 1, mejor será el ajuste del modelo a la variable que intentamos explicar. Por el contrario, cuanto más cerca está de cero, peor es el ajuste del modelo y, por lo tanto, menos fiable es (Rodríguez E. ).

Ecuación 7: Coeficiente de determinación

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - y_{pred})^2}{\sum_{i=1}^N (y_i - y_{promedio})^2} \dots \dots (7)$$

Donde:

$y_i$  = Representa al valor de la variable dependiente.

$y_{pred}$  = Es el valor que predice el modelo.

$y_{promedio}$  = Es el promedio de la variable dependiente.

### Problemas con coeficiente de determinación.

El problema del coeficiente de determinación, y la razón del coeficiente de determinación ajustado, es que no penaliza la inclusión de variables explicativas insignificantes.

Hay algunos problemas con R-cuadrado que se resuelven con R-cuadrado ajustado. Un R-cuadrado ajustado considerará la mejora marginal agregada por un término adicional en el modelo. Por lo tanto, aumentará si agrega los términos útiles y disminuirá si agrega predictores menos útiles. Sin embargo, R-cuadrado aumenta con el aumento de los términos a pesar de que el modelo en realidad no está mejorando (Rodríguez E. ).

### 3.5.2. Error de Cuadrático Medio.

El error cuadrático medio (MSE) de un estimador mide el promedio de los errores cuadráticos, la diferencia entre el estimador y el valor estimado. ECM es una función de riesgo que corresponde al valor esperado de la pérdida por error cuadrático o pérdida cuadrática. La discrepancia se produce debido a la aleatoriedad o porque el estimador no tiene en cuenta la información que podría producir una estimación más precisa.

Esta es una cantidad conocida, calculada dada una muestra particular (y por lo tanto dependiente de la muestra). Un valor MSE más bajo indica un mejor ajuste. MSE es una buena medida de qué tan bien un modelo predice una respuesta y es el criterio más importante para el ajuste si el propósito principal del modelo es predecir (Wikipedia, Wikipedia, 2014).

Ecuación 8: Formula del error cuadrático medio.

$$MSE = \frac{1}{N} * \sum_{i=1}^N (y_i - y_{pred})^2 \dots \dots (8)$$

Donde:

$MSE$ = Error cuadrático medio.

$N$ = Numero de datos.

$y_i$ = Representa al valor de la variable dependiente.

$y_{pred}$  = Es el valor que predice el modelo.

### 3.5.3. Error porcentual absoluto medio

También conocido como por MAPE que es una medida relativa que escala esencialmente para que se muestre en unidades de porcentaje en lugar de en unidades de la variable. El error de porcentaje medio absoluto es una medida de error relativa que utiliza valores absolutos para evitar que los errores positivos y negativos se cancelen entre sí y utiliza errores relativos para permitirle comparar la precisión de previsión entre métodos de serie de tiempo. (Wikipedia, Wikipedia, 2014).

Cuya formula es la siguiente:

Ecuación 9: Formula del error porcentual absoluto medio.

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{Y_{test} - Y_{pred}}{Y_{test}} \right| \dots \dots (9)$$

Donde:

$MAPE$  = Error porcentual absoluto medio.

$Y_{test}$  = Valor real.

$Y_{pred}$  = Valor predicho.

$n$  = Numero de datos.

Ten en cuenta que no representamos la salida como un porcentaje en el rango [0, 100]. En su lugar, la representamos en el rango [0, 1]. (SKLEARN, 2022)

### 3.5.4. Normalización

La estandarización o normalización de los datos es el paso inicial de la ingeniería de características. La normalización de las columnas implicará llevar los valores de las columnas a una escala común, lo que se hace principalmente para las columnas con rangos variados. En Pandas, las columnas de los Dataframes pueden ser normalizadas por una variedad de funciones (SKLEARN, 2022).

Hay dos métodos de normalización de datos más utilizados.

- Normalización media.
- Normalización Min-Max.
- Normalización de cuantiles.

### 3.5.5. Escala estandarizada.

Estandarice las características eliminando la media y escalando a la varianza de la unidad (SKLEARN, 2022).

La puntuación estándar de una muestra  $x$  se calcula como:

Ecuación 10: Formula del error cuadrático medio.

$$Z = \frac{X - U}{S} \dots \dots (10)$$

Donde:

$Z$ = Valor calculado.

$X$ = Valor original.

$U$ = Media.

$S$ = Desviación estándar.



## **4. CAPÍTULO IV APLICACIÓN PRACTICA**

### **4.1. DETECCIÓN DE ANOMALÍAS EN EL SEPARADOR DE PRUEBA EN EL POZO CUR-X1001 MEDIANTE LA REGRESIÓN LINEAL SIMPLE Y REGRESIÓN LINEAL MÚLTIPLE**

Los algoritmos se los subió paso a paso con la explicación necesaria a un Google Colab, no se tiene ningún requisito para abrir los links (solo se habilito para que cualquier persona dando click a los mismo pueden ver los códigos y su desarrollo correspondiente).

[https://colab.research.google.com/drive/1ZOfSrHsyZLrJ\\_vqM1IYnCBwuYBgn\\_WMB](https://colab.research.google.com/drive/1ZOfSrHsyZLrJ_vqM1IYnCBwuYBgn_WMB)

- Link de la metodología KDD (Aquí se hizo la correspondiente limpieza, graficas, entrenamiento y evaluaciones de los modelos, se obtuvo las ecuaciones, se obtuvo las métricas señaladas).

#### **4.1.1. Selección**

Se precisa hallar al menos un modelo que sea capaz de predecir los posibles fallos que pueda tener un equipo de separación con una exactitud del 70% o más. Los datos fueron extraídos de un archivo Excel perteneciente a la producción del pozo CUR-X-1001 LL.

Los datos puntualmente describen el producto de la separación registrada de los productos de Agua, condensado y gas.

El pozo CUR-X-1001 LL es el sujeto de estudio, por ende, su producto estrella es la producción de gas, por lo tanto, se plantea usar esta cantidad de condensado producida como VARIABLE DEPENDIENTE que será usada como un indicador de un posible problema en los separadores.

Los datos de agua y condensado se los tratara como VARIABLES INDEPENDIENTES.

En la tabla 4.1 se expondrá la cantidad de datos que se tiene del archivo:

Tabla 4.1: Numero de Filas y Columnas de nuestra Data-Original.

Numero de Filas	Numero de Columnas
53	5

Fuente: Elaboración propia con datos de Informe técnico de YPF, 2012.

La corroboración de la Tabla 4.1 se encuentra en la parte de ANEXOS ('Exportación de los datos a Google Colab')

#### 4.1.2. Preprocesado/Limpieza

- Se pasó a analizar por columnas la cantidad de valores nulos que estos presentan. Se encuentra en ANEXOS ('Conteo de valores nulos por columna').

En la tabla 4.2 se describe el tipo de dato y además la cantidad de valores no nulos de cada columna dentro del DataFrames.

Tabla 4.2: Columnas por su número de datos no nulos y tipo de dato.

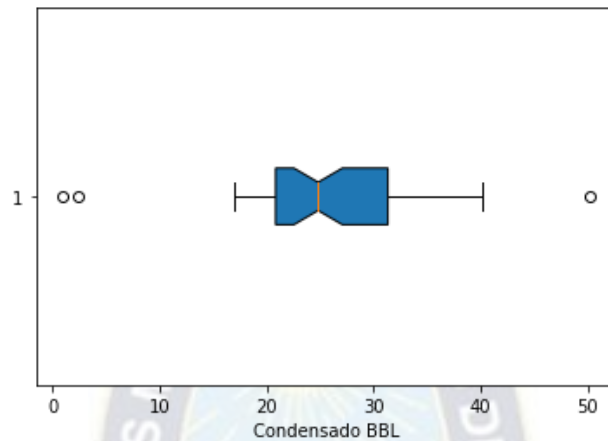
Columna	No nulos	Tipo de dato
UNIQUEID	53	Object
DATE	53	Datetime64[ns]
COND BBL	53	Float64
GAS MPC	53	Float64
WATER BBL	53	Float64

Fuente: Elaboración propia en Python con datos de Informe técnico de YPF, 2012.

- Como se puede ver en la tabla 4.2 no existe valores, pero si existen columnas que no nos ayudaran en nuestro análisis, posteriormente se retirará las que no serán necesarias como ser: UNIQUEID, DATE.
- Se procederá a encontrar valores atípicos mediante la graficas de Box Plot, tanto para el Condensado, Gas y Agua.

La figura 4.1 representa a la gráfica de boxplot para las cantidades de condensado a la salida del separador.

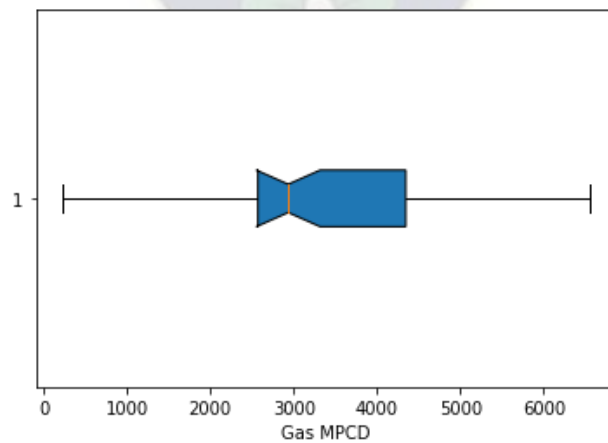
Figura 4.1: Box plot para el análisis del condensado.



Fuente: Elaboración propia usando Python con datos de Informe técnico de YPF, 2012.

La figura 4.2 representa a la gráfica de boxplot para las cantidades de gas a la salida del separador.

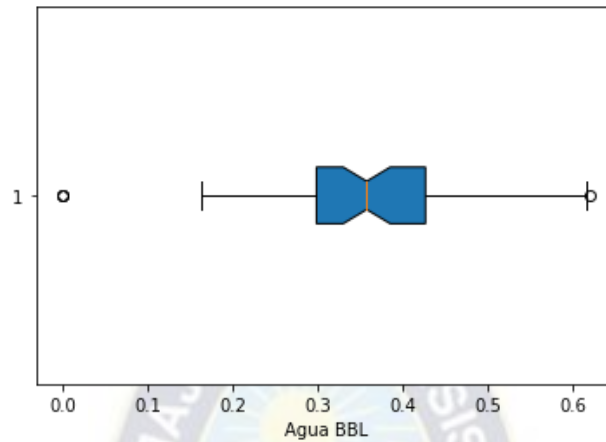
Figura 4.2: Box plot para el análisis del gas.



Fuente: Elaboración propia usando Python con datos de Informe técnico de YPF, 2012.

La figura 4.3 representa a la gráfica de boxplot para las cantidades de agua a la salida del separador.

Figura 4.3: Box plot para el análisis del agua.



Fuente: Elaboración propia usando Python con datos de Informe técnico de YPF, 2012.

En las figuras 4.1, 4.2 y 4.3 se ve la existencia de valores atípicos, posteriormente en la parte de minería de datos se verá si es necesario la eliminación de estos datos o no, ya que son bastante pocos.

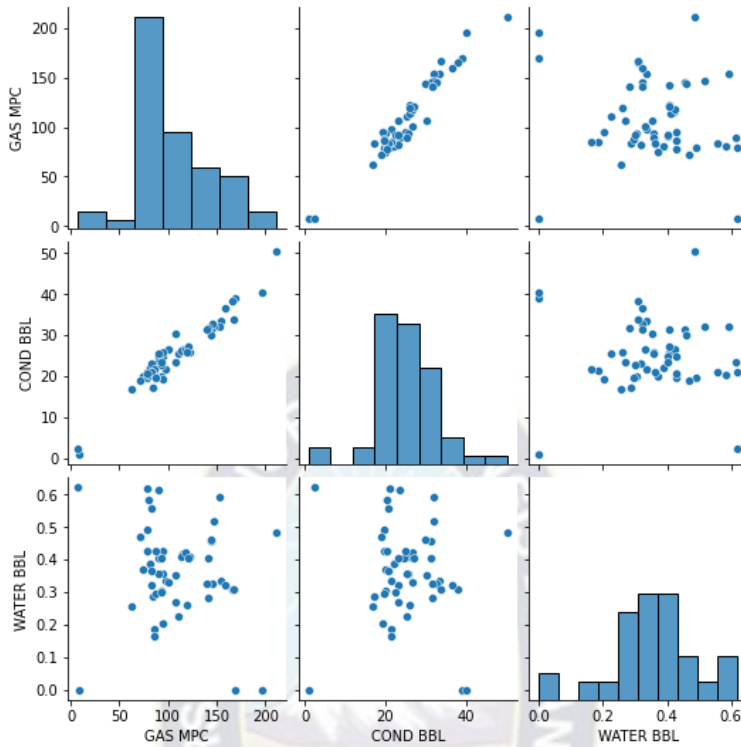
#### 4.1.3. Transformación

No se deben procese a realizar este paso por el momento, ya que las variables vienen dadas como tipos de dato numérico como se puede evidenciar en la tabla 4.2, en análisis posteriores puede ser necesaria una transformación de los datos en valores estandarizados para que estén en una misma escala, en caso de ser necesaria esta normalización se la describirá en los métodos empleados para hallar las mejores métricas.

#### 4.1.4. Minería de datos

En esta parte se procederá a ver como se relacionan estos datos, si tienen una correlación positiva, negativa o ninguna correlación. En la figura 4.4 se ve cómo se relaciones las variables.

Figura 4.4: Correlación entre las variables.



Fuente: Elaboración propia usando Python con datos de Informe técnico de YPF, 2012.

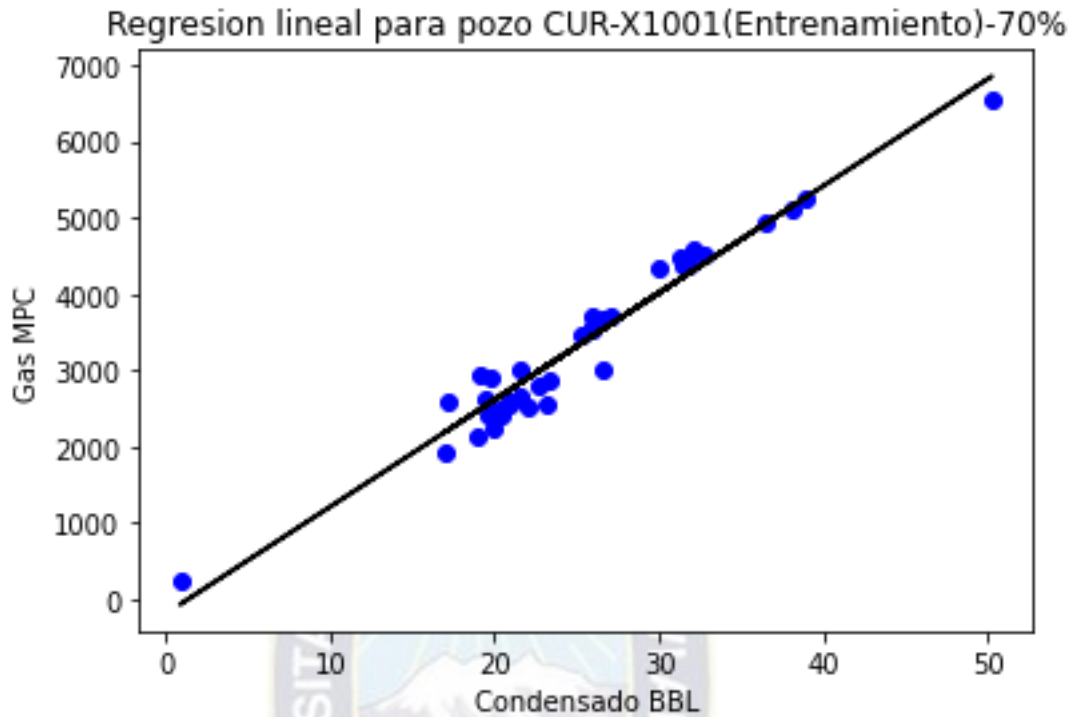
En la figura 4.4 se puede evidenciar la existencia de una fuerte correlación entre el Gas y el Condensado.

Se reconocerá como variables dependientes a la columna referida al Gas, las variables independientes serán la cantidad de condensado y de agua a la salida del separador.

#### 4.1.4.1. Método 1: Regresión lineal simple entre gas y condensado.

- La regresión lineal simple se la realizó tomando como variable dependiente el gas y variable independiente el condensado, se tomó el 70% de los datos como datos de ENTRENAMIENTO y 30% de los datos como EXAMEN. La figura 4.5 representa la cantidad de gas de salida y condensado de salida para el 70% de los datos.

Figura 4.5: Gas Vs Condensado -Entrenamiento.

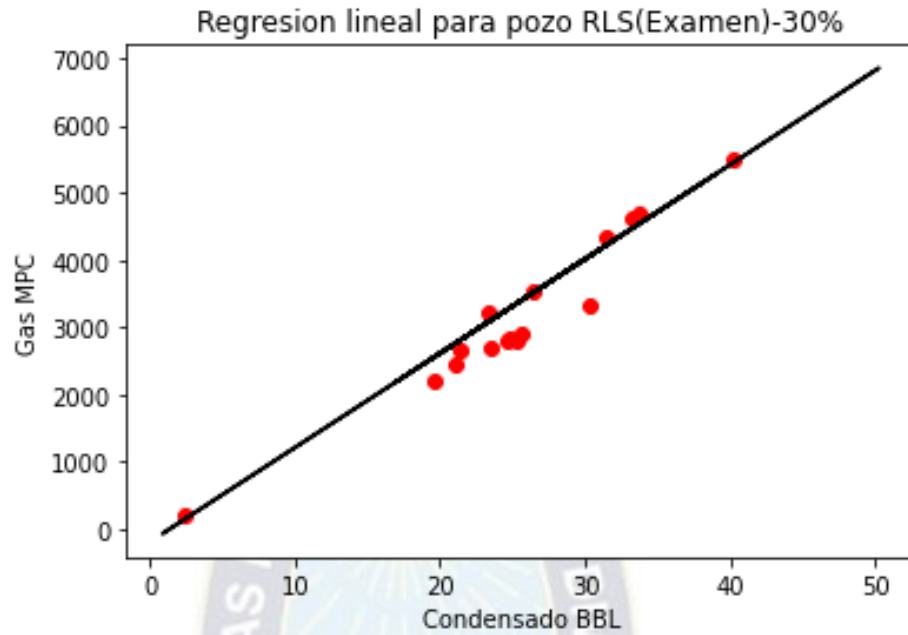


Fuente: Elaboración propia usando Python-Matplotlib-sklearn con datos de Informe técnico de YPFB, 2012.

La figura 4.5 representa el modelo de regresión lineal simple en el pozo CURX-1001 con el 70% de los datos para el entrenamiento, cabe recalcar que la línea de color negro representa el modelo generado por el algoritmo y los puntos azules son los datos reales. Las métricas para evaluar exactitud del modelo se podrán a continuación en la tabla 4.3.

La manera en que el modelo de regresión lineal simple es capaz de generalizar el 30% de datos faltantes se verá en la figura 4.6 en la etapa del examen del modelo.

Figura 4.6: Gas Vs Condensado -Examen.



Fuente: Elaboración propia usando Python-Matplotlib-sklearn con datos de Informe técnico de YPFB, 2012.

De la misma manera la línea negra corresponde al modelo, los puntos rojos representan a los datos que fueron separados (30% para el EXAMEN), se evidencia gráficamente que el modelo se correlaciona. En la tabla 4.3 representa las métricas generadas con el examen del modelo.

Tabla 4.3: Métricas del método 1.

Métricas	Valor
Error cuadrático medio	84924.72
Coefficiente de correlación	0.932
Error absoluto porcentual medio	0.0735

Fuente: Elaboración propia con datos de Informe técnico de YPFB, 2012.

De la tabla 4.3 representa las métricas del modelo lineal simple en el pozo, la codificación y evidencia se encuentra en ANEXOS (‘Métricas del método 1’).

**4.1.4.2. Método 2: Regresión Lineal Múltiple entre gas, condensado y agua.**

- La regresión lineal múltiple se la realizo para el gas, condensado y agua, se tomó el 70% de los datos como datos de ENTRENAMIENTO y 30% de los datos como EXAMEN. En la tabla 4.4 representa las métricas generadas con el examen del modelo.

Tabla 4.4: Métricas del método 2.

<b>Métricas</b>	<b>Valor</b>
Error cuadrático medio	89388.55
Coefficiente de correlación	0.929
Error absoluto porcentual medio	0.070

Fuente: Elaboración propia con datos de Informe técnico de YPF, 2012.

En la tabla 4.4 se puede evidenciar que mejora tanto el coeficiente de relación y por otra parte el error cuadrático medio disminuye. Se incluirá el respaldo del código junto con el output en ANEXOS (‘Métricas del modelo de regresión lineal múltiple’).

**4.1.4.3. Método 3: Suprimiendo los outliers.**

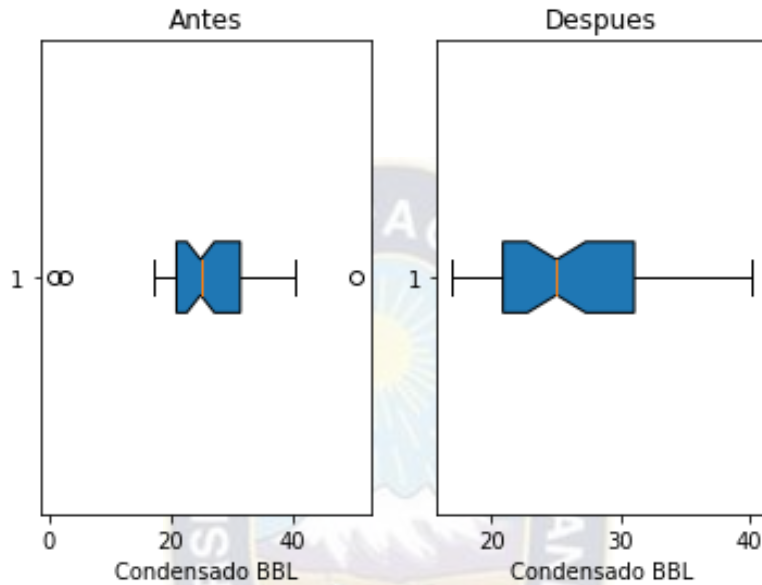
Como se puede evidenciar en la figura 4.1: Box plot para el análisis del condensado se procederá a eliminar los outliers, se pondrá de condición que se deben ser mayores a 10 BBL y menores a 45 BBL. En total se eliminaron 3 datos que nos lleva a tener solo 50 datos.

De aquí en adelante se usará el modelo de regresión lineal simple para ver si se mejoraron las métricas también como se ajustaron las gráficas de relación gas y condensado.



En la figura 4.7 se apreciará el antes y el después de la eliminación de los outliers se creó un nuevo DataFrame que solo tome valores mayores a 10 BBL y menores a 45 BBL.

Figura 4.7: Eliminación de outliers.



Fuente: Elaboración propia usando Python-Matplotlib-sklearn con datos de Informe técnico de YPF, 2012.

La última grafica representa la eliminación de los outliers, de la misma manera se procede a aplicar un modelo de regresión lineal simple, los resultados estarán en la tabla 4.5.

Tabla 4.5: Métricas del Método 3.

Métricas	Valor
Error cuadrático medio	84015.59
Coefficiente de correlación	0.896
Error absoluto porcentual medio	0.069

Fuente: Elaboración propia usando Python-sklearn con datos de Informe técnico de YPF, 2012.

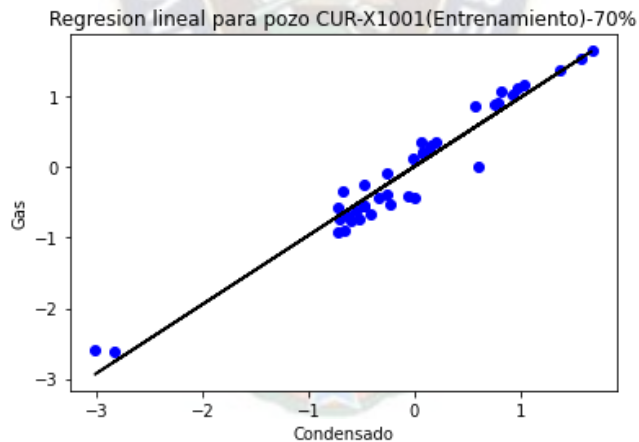
De la tabla 4.4 representa las métricas del modelo lineal simple eliminando los valores atípicos en el pozo, la codificación y evidencia se encuentra en ANEXOS ('Métricas del método 3').

#### 4.1.4.4. Método 4: Criterio de la normalización con la regresión lineal simple.

Como las unidades de la salida del separador se nota obviamente que no están en los mismos rangos ejemplo el promedio del gas es de 3.3 MMPCD y del condensado es de aproximadamente 25.36 BBLD. Con el criterio de la normalización se podrá llevar al mismo rango en la parte de teoría se puede evidenciar que formula se inició para dicha estandarización, además cabe mencionar y la librería de sklearn ya tiene carga la función para retornar valores normalizados.

La figura 4.8 se tendrá la gráfica de entrenamiento al 70% pero nótese que los ejes ya están estandarizados.

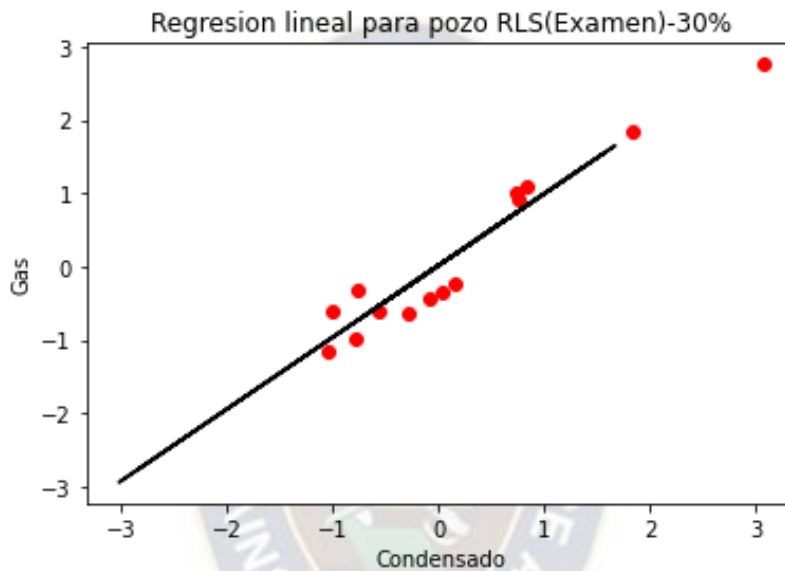
Figura 4.8: Gas Vs Condensado -Entrenamiento-Normalizado.



Fuente: Elaboración propia usando Python-matplotlib-sklearn con datos de Informe técnico de YPFB, 2012.

Los valores negativos de la figura 4.8 como tal no influyen ya que esto depende de la normalización que añadimos anteriormente, además de presentar unidades adimensionales lo cual esta explicado en la ecuación 9. También se realizará la gráfica para el examen que será expuesta en la figura 4.9 con 30% de los datos normalizado.

Figura 4.9: Gas Vs Condensado -Examen-Normalizado.



Fuente: Elaboración propia usando Python-matplotlib-sklearn, Informe técnico de YPFB, 2012.

Los resultados obtenidos se observarán en la Tabla 4.6.

Tabla 4.6: Métricas del Método 4.

Métricas	Valor
Error cuadrático medio	0.062
Coefficiente de correlación	0.931
Error absoluto porcentual medio	0.0693

Fuente: Elaboración propia usando Python-sklearn con el Informe técnico de YPFB, 2012.

La evidencia de los resultados obtenidos en la tabla 4.6 se podrán en la parte de ANEXOS ('Métricas del Método 4 regresión lineal simple con normalización')

Al normalizar también podemos usar criterios estadísticos como ser el coeficiente de correlación ajustado como tal, P-value, el intercepto.

En la figura 4.10 se verá los resultados obtenidos, usando criterios estadísticos, también se procede a usar una librería de Python dedicada específicamente a usar la regresión lineal.

Figura 4.10: Métricas estadísticas método 4 con la librería regressors.

```
▶ from regressors import stats
  model.intercept_ = model.intercept_[0]
  model.coef_ = model.coef_.reshape(-1)

] y_test = y_test.reshape(-1)

print("=====Summary=====")
stats.summary(model, X_test, y_test, X_cols)

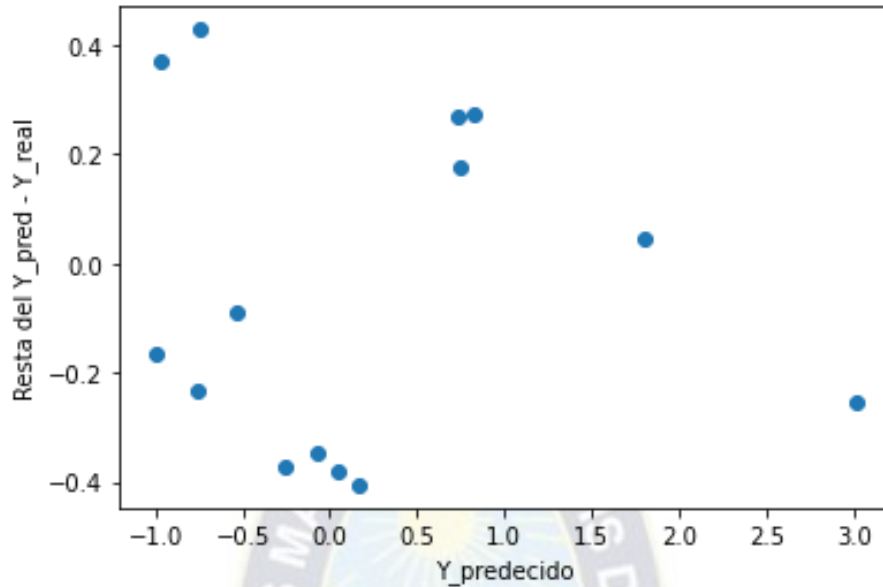
=====Summary=====
Residuals:
   Min    1Q  Median    3Q   Max
-0.3356 -0.224 -0.1578  0.0729  0.5577

Coefficients:
              Estimate  Std. Error  t value  p value
_intercept -0.013054    0.068856  -0.1896  0.852564
COND BBL    0.976925    0.067221  14.5331  0.000000
---
R-squared:  0.93167,    Adjusted R-squared:  0.92597
F-statistic: 163.61 on 1 features
```

Fuente: Elaboración propia usando Python y regressors.

También se usará las gráficas de residual, para ver visualmente que tan buenos son nuestros criterios la cual será expresada en la figura 4.11.

Figura 4.11: Grafica de residuales métodos 4.



Fuente: Elaboración propia usando Python y matplotlib.

**4.1.4.5. Método 5: Criterio de la normalización con la regresión lineal múltiple.**

- La regresión lineal múltiple se la realizo para el gas, condensado y agua, se tomó el 70% de los datos como datos de ENTRENAMIENTO y 30% de los datos como EXAMEN. En la tabla 4.7 se presentarán las métricas obtenidas con el 30% de datos del examen.

Tabla 4.7: Métricas del método 5.

Métricas	Valor
Error cuadrático medio	0.066
Coefficiente de correlación	0.927
Error absoluto porcentual medio	0.0673

Fuente: Elaboración propia con datos de Informe técnico de YPF, 2012.

En la figura 4.12 se verá los resultados obtenidos, usando criterios estadísticos, también se procede a usar una librería de Python dedicada específicamente a usar la regresión lineal.

Figura 4.12: Métricas estadísticas método 5 con la librería regressors.

```
y_test = y_test.reshape(-1)

print("====Summary====")
stats.summary(model, X_test, y_test, X_cols)

====Summary====
Residuals:
   Min    1Q  Median    3Q   Max
-0.4233 -0.2356 -0.1305  0.0658  0.5471

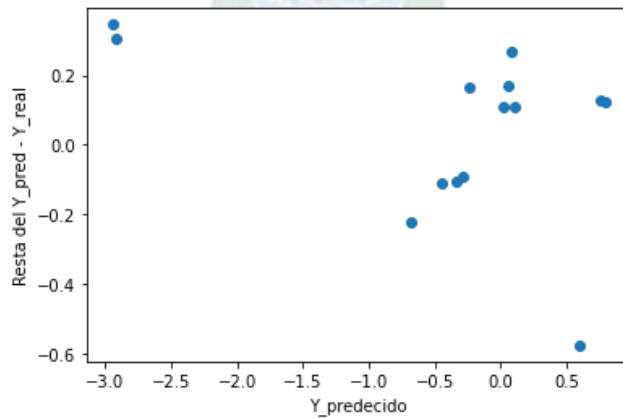
Coefficients:
              Estimate  Std. Error  t value  p value
_intercept -0.021008    0.073034  -0.2877  0.778148
COND BBL    0.965097    0.070057  13.7758  0.000000
WATER BBL  -0.062341    0.076705  -0.8127  0.431003
---
R-squared:  0.92770,    Adjusted R-squared:  0.91455
F-statistic: 70.57 on 2 features
```

Fuente: Elaboración propia usando Python y regressors.

Observando la figura 4.12 y por el criterio del P\_value nos dice que las variables no son significativas deben ser mayor a 0.05, entonces lo que se planea es eliminar esas variables las cuales son: WATER BBL y INTERCEPTO.

También se usará las gráficas de residual, la cual se presentará en la figura 4.13 para ver visualmente que tan buenos son nuestros criterios.

Figura 4.13: Grafica de residuales métodos 4.



Fuente: Elaboración propia usando Python y matplotlib.

#### 4.1.4.6. Método 6: Criterio de la normalización con la regresión lineal Simple eliminando variables.

La regresión lineal simple se la realizo para el gas y condensado. También ya se pasó a eliminar la variable de la cantidad de salida del agua. Se tomó el 70% de los datos como datos de ENTRENAMIENTO y 30% de los datos como EXAMEN. En la tabla 4.8 representara las métricas del método 6.

Tabla 4.8: Métricas del método 5.

<b>Métricas</b>	<b>Valor</b>
Error cuadrático medio	0.0618
Coefficiente de correlación	0.933
Error absoluto porcentual medio	0.0663

Fuente: Elaboración propia con datos de Informe técnico de YPFB, 2012.

Los resultados y evidencia de la tabla 4.8 se puede ver en ANEXOS (‘Métricas del método 6 regresión lineal simple, normalizando y eliminando variables.’)

En la figura 4.14 se verá los resultados obtenidos, usando criterios estadísticos, también se procede a usar una librería de Python dedicada específicamente a usar la regresión lineal.

Figura 4.14: Métricas estadísticas método 6 con la librería regressors.

```
[ ] y_test = y_test.reshape(-1)

print("====Summary====")
stats.summary(model, X_test, y_test, X_cols)

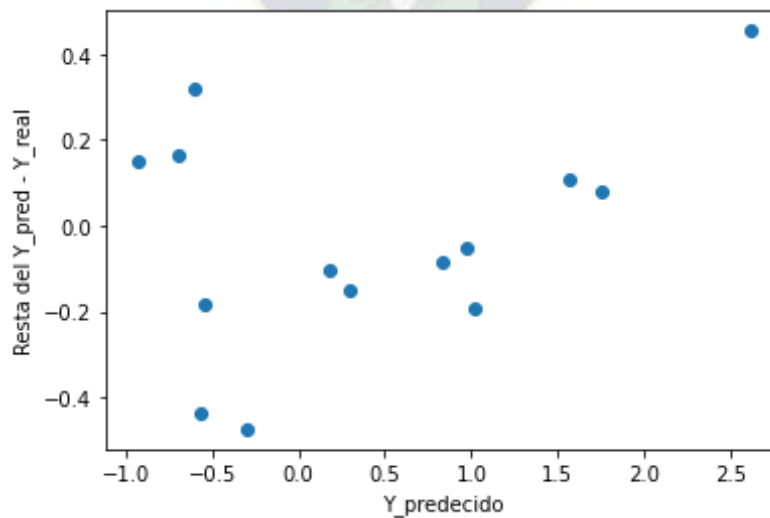
====Summary====
Residuals:
    Min      1Q   Median      3Q      Max
-0.3235 -0.2112 -0.1446  0.0853  0.5715

Coefficients:
              Estimate  Std. Error  t value  p value
_intercept    0.000000    0.068279   0.0000    1.0
COND BBL      0.978215    0.066657  14.6753    0.0
---
R-squared:  0.93281,    Adjusted R-squared:  0.92721
F-statistic: 166.59 on 1 features
```

Fuente: Elaboración propia usando Python y regressors.

También se usará las gráficas de residual, para ver visualmente que tan buenos son nuestros criterios esto en la figura 4.15.

Figura 4.15: Grafica residual método 6.



Fuente: Elaboración propia con Python.

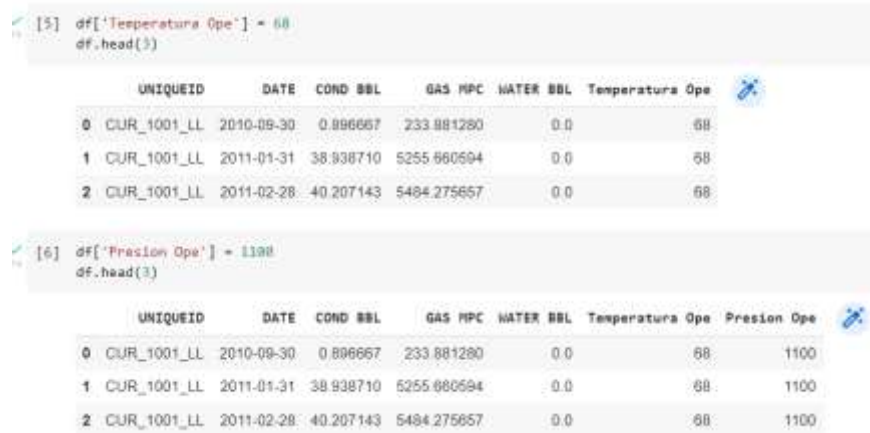


#### 4.1.4.7. Método 7: Añadiendo más variables.

Observando la tabla 1.4 que representa los valores de presión y temperatura que son constantes.

En la figura 4.16 se verá que se añadieron dichos valores a nuestro DataFrame.

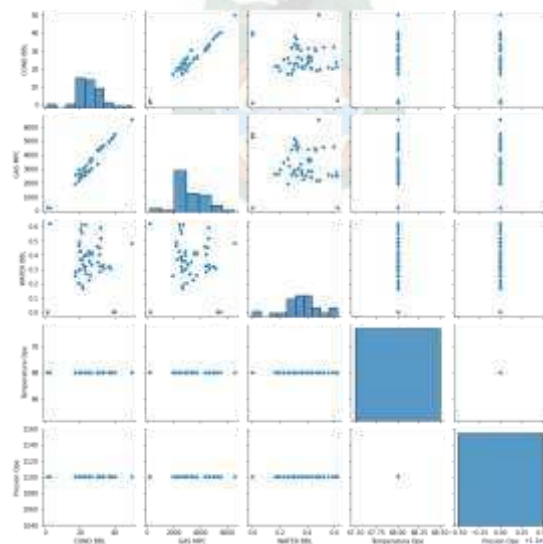
Figura 4.16: Añadiendo las variables de presión y temperatura.



Fuente: Elaboración propia usando Python y Pandas.

En la figura 4.17 se procederá a visualizar como se correlacionan las nuevas variables de presión y temperatura.

Figura 4.17: Correlación con las nuevas variables.



Fuente: Elaboración propia usando Python y Seaborn.

Se observa en la figura 4.17 que los datos del separador de prueba como son la presión y temperatura son constantes, si bien en la vida real no son constantes lastimosamente no se cuentan con los valores de registrados en cada vez que se registraron los datos, razón por la cual se los tomara como constantes.

Se procederá a analizarlos como una regresión lineal simple y las métricas se las presentará en la tabla 4.9.

Tabla 4.9: Métricas del método 7.

<b>Métricas</b>	<b>Valor</b>
Error cuadrático medio	89388.55
Coefficiente de correlación	0.929
Error absoluto porcentual medio	0.0735

Fuente: Elaboración propia usando Python y Sklearn.

En la tabla 4.9 se puede ver las métricas obtenidas añadiendo las variables de presión y temperatura, la evidencia de los mismos estará en la parte de ANEXOS (‘Métricas método 7’).

#### **4.1.5. Evaluación**

El método 6 se pudo observar que, usando la data Normalizada, eliminando de las variables no significativas como el intercepto y la cantidad de agua. Además, la figura 4.15 se evidencia que los valores están más próximos de los predichos y reales.

## 4.2. TABLA DE RESULTADOS

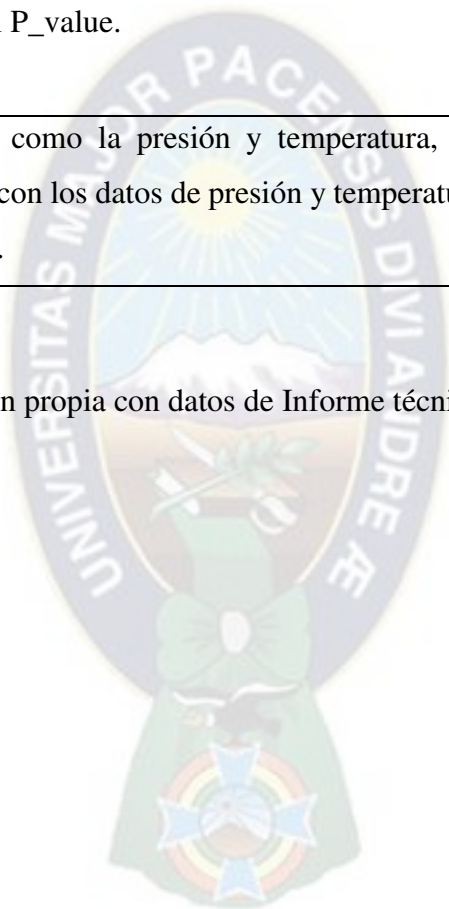
La tabla 4.10 representa un resumen de todas las métricas que se recolectaron a lo largo de los métodos empleados.

Tabla 4.10: Resumen de resultados.

METODO	DESCRIPCION	Métricas	
1	Se usó una regresión lineal simple usando el gas y condensado. Sin realizar una eliminación de valores atípicos ni estandarización.	MSE	84924.7
		R <sup>2</sup>	0.932
		MAPE	7.36%
2	Se usó una regresión lineal múltiple usando el gas, condensado y agua. Con outliers y sin normalizar.	MSE	89388.5
		R <sup>2</sup>	0.929
		MAPE	7.02%
3	Se realizó una regresión lineal simple, eliminando los datos outliers del condensado.	MSE	84015.5
		R <sup>2</sup>	0.896
		MAPE	6.9%
4	Se realizará la normalización en una regresión lineal simple para que este en la misma escala los datos, además también se puede sacar nuevas métricas como el coeficiente de correlación ajustado.	MSE	0.0628
		R <sup>2</sup>	0.931
		R <sup>2</sup> _Ajus	0.925
		MAPE	6.93%
5	Se realizará la normalización en una regresión lineal múltiple para que este en la misma escala los datos, además también se puede sacar nuevas métricas como el coeficiente de correlación ajustado.	MSE	0.066
		R <sup>2</sup>	0.927
		R <sup>2</sup> _Ajus	0.914

		MAPE	6.73%
6	Se realizará la normalización en una regresión lineal simple, se eliminó las variables no significativas con el criterio del P_value.	MSE	0.061
		$R^2$	0.933
		$R^2_{Ajus}$	0.927
		MAPE	6.63%
7	Se añadieron variables extras como la presión y temperatura, estos fueron valores constantes ya que no se cuenta con los datos de presión y temperatura de cada día que se analizó la salida del pozo CUR.	MSE	89388.5
		$R^2$	0.929
		MAPE	7.36%

Fuente: Elaboración propia con datos de Informe técnico de YPFB, 2012.



### **4.3. ANALISIS DE RESULTADOS PARA LA APLICACIÓN.**

En total se tomó 6 métodos diferentes para obtener el método que más exactitud nos marque, en este caso, se observa que gracias a la normalización se tienen mejor coeficiente de determinación, además gracias a las gráficas residuales, la figura perteneciente al método 6 tienen valores más próximos en el eje Y. Donde este eje Y significa la diferencia que existen en valor de gas predicho y el valor de gas real.

Se obtuvo mejores métricas en el método 6 que es una regresión lineal simple, normalizada. Inclusive se obtuvo mejores métricas que la regresión lineal múltiple, eso demuestra que se puede tener mejores métricas usando menos variables. También este método 6 se dejó del lado el intercepto y variable de agua ya que indica el criterio del p\_value si se presenta un valor mayor a 0.05 se debe eliminar dicha variable. Además, cabe mencionar que el coeficiente de correlación ajustado presenta la mínima diferencia con el coeficiente de determinación normal, esta diferencia es de apenas de 0.06 la mínima entre todos los métodos.

Como es bien sabido se sabe que la producción de un campo cuenta con 2 separadores, estos separadores son: El Separador General y el Separador de Prueba. La labor que realiza el Separador de Control es separar el gas, condensado y agua exclusivamente para un pozo, esto con el fin de supervisar la producción periódica. El Separador General por otra parte separa el gas, condensado y agua de los demás pozos. Posterior a cierto tiempo (cada 4 días) mediante el uso válvulas se cambia la alimentación del separador de control por la producción de otro pozo (esto con el fin de tener una evaluación paulatina de cada pozo).

Cuando se trabaje con el separador de prueba se debe usar nuestro algoritmo para la detección inmediata. Se planteó algoritmos capaces de decir si el valor de gas es un valor normal o un valor atípico. En caso de presentarse un valor atípico en la salida de gas en el separador de control, nos

daría una indicación que este valor podría ser una potencial anomalía en dicho pozo lo cual traería problemas posteriores procesos que se tiene en el tratamiento de los hidrocarburos.

#### 4.3.1. Puesta en marcha del algoritmo

Se procederá a realizar una simulación de valores de salida de condensado y gas, se tomó estas dos variables porque se evidencio que el método 6 se obtuvo mejores métricas, así que se procederá a que el usuario sea capaz de introducir las cantidades de gas y condensado que se obtuvieron en la salida del separador de prueba. El algoritmo devolverá el valor del Gas que el modelo predice. Con los siguientes criterios:

- En caso de obtener un dato anómalo de gas es decir una cantidad excesiva que el algoritmo predijo debe ser capaz de mostrar por pantalla que posiblemente se produjo una formación de espumas.
- En caso de obtener un valor menor al valor predicho por el modelo debe ser capaz de mostrar por pantalla que posiblemente se halla dado una formación de vórtices.

En los puntos anteriores se expandieron por el arrastre de condensado en la formación de espuma y el arrastre de gas en la formación de vórtices.

La figura 4.18 representa a que se tomaran 1 dato aleatorios de nuestro DataFrame.

Figura 4.18: Un dato aleatorio del DataFrame:

```
[35] df.sample(1)
```

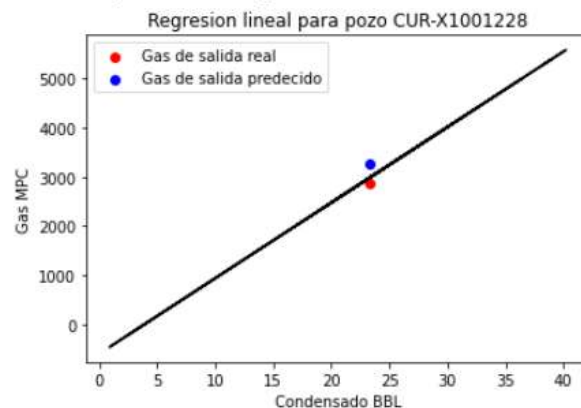
	GAS MPC	OIL BBL	WATER BBL
48	2862.26327	23.270645	0.403548

Fuente: Elaboración propia usando Python y Pandas.

En la figura 4.18 según el valor real de salida del Gas debería ser de 2862.26 MPCD para un condensado de 23.27 BBL. Si colocamos estos datos en el algoritmo este debería señalar que no presencia de datos anómalos, además de graficarlos esto se presentara en la figura 4.19.

Figura 4.19: Entrada de datos de prueba.

```
↳ Introduzca la cantidad de condensado en la salida del separador: 23.27
Introduzca la cantidad de gas en MPC registrada en la salida del separador: 2870
****NO SE PRODUCE ANOMALIAS****
El valor prededido de gas a la salida es de: 3283.6433743990515
```

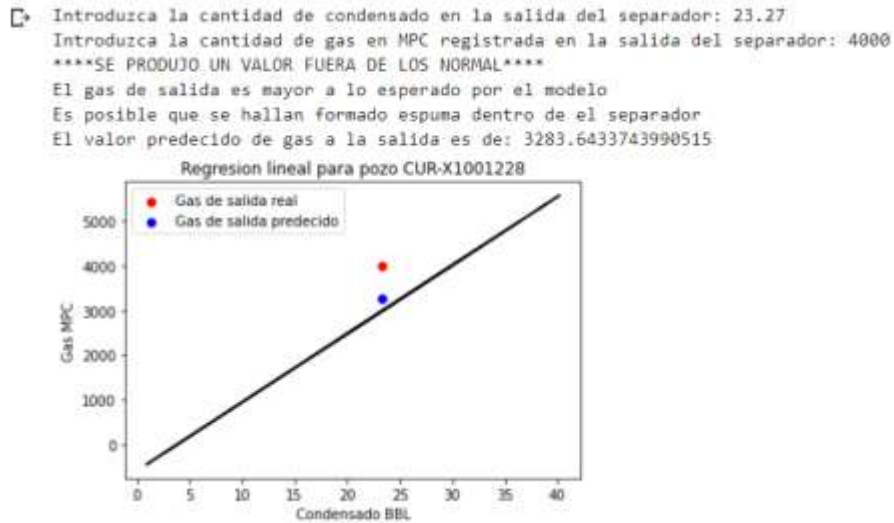


Fuente: Elaboración propia usando Python, sklearn, Matplotlib.

En la figura 4.19 la línea negra representa a los valores presentados por el modelo, el punto azul representa al valor de salida de gas precedido por el modelo, mientras que el punto rojo es el valor registrado a la salida del separador (valor real).

Ahora se procederá a añadir datos en la salida de gas que podrían ser anómalos, en la figura 4.20 se pondrá una cantidad excesiva del gas lo cual puede significar que se halla dado una posible formación de espumas dentro del separador.

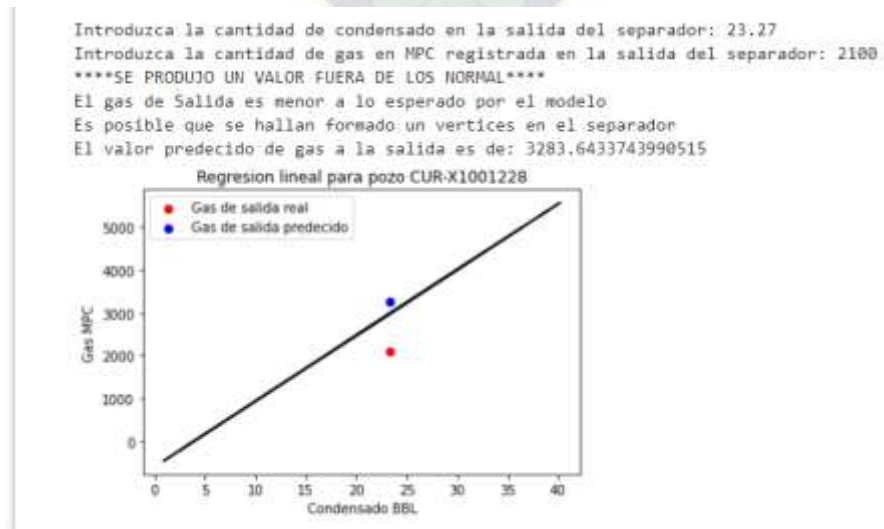
Figura 4.20: Entrada de datos de prueba gas excesivo.



Fuente: Elaboración propia usando Python, sklearn, Matplotlib.

Nuevamente se procederá a añadir una cantidad más pequeña en la salida del gas, el algoritmo debería ser capaz de decir que se encontró un valor anómalo, graficarlo y posiblemente una formación de vértices esto se verá reflejado en la figura 4.21.

Figura 4.21: Entrada de datos de prueba gas disminuido.



Fuente: Elaboración propia usando Python, sklearn, Matplotlib.



Como se pudo ver en las figuras 4.19, 4.20 y 4.21 el algoritmo desarrollado es capaz de responder si la salida del Gas en el separador es un valor normal o estamos en presencia de una anomalía. Las evidencias del código correspondiente se encuentran en ANEXOS ('Puesta en marcha del algoritmo')

#### **4.4. ANALISIS DE COSTOS**

Se supondría un escenario donde debido a un arreglo no programado con un separador se detiene la producción hasta arreglar dicho problema, imaginemos un 1 día de producción.

En la figura 1.4 se puede ver como el valor promedio en gas es de aproximadamente 3.4 MMPCD ('Millones de pies cúbicos días') el precio de la producción de un día, en el mercado interno se paga aproximadamente 6.2 \$US por MPC, lo que costaría un día promedio de producción sería de más de 21000 \$US, lo cual dañaría la economía de la empresa, mucho ojo que solo es la pérdida que se tendría del pozo CURX-1001. Los demás pozos tienen valores más altos de producción.

## 5. CAPÍTULO V CONCLUSIONES Y RECOMENDACIONES

### 5.1. conclusiones

- Con los datos obtenidos del pozo CURX-1001 LL se logró desarrollar un algoritmo capaz de predecir valores de cantidades de gas en base al condensado a la salida del separador (véase figuras 4.20, 4.21 y 4.22), este algoritmo es capaz de responder si se tienen valores excesivamente pequeños (formación de vértices) o grandes (formación de espumas) de gas lo cual se puede traducir como posibles anomalías, los modelos usados de aprendizaje automático usados fueron los de regresión lineal simple y regresión lineal múltiple, cada uno con diferentes métodos para tratar a los datos como ser la normalización de estos y/o eliminación de valores atípicos (outlayers).
- Se planteo la metodología KDD con su secuencia de etapas, se la logró definir y resolver cada etapa de esta metodología (los cuales están descritos desde el punto 4.1.1 hasta el 4.1.5) para los datos del separador de prueba para el pozo CURX-1001 LL, tomando como variable dependiente el gas de salida y como variables independientes el condensado y el agua a la salida de este.
- Se obtuvo por una distribución 70-30, los modelos entrenaron con un 70% de datos, ya no se corrió con 80% de los datos ya que mostraba buenas métricas ya con el 70% de entrenamiento y 30% para el examen.
- Se planteo las métricas de coeficiente de determinación, coeficiente de determinación ajustado y el error cuadrático medio. Esto debido a tipo de datos que presento son de

carácter son números continuos tanto para la variable dependiente (gas) y variables independientes (agua y condensado).

- Se aplicó el mejor método que es el método 6 que presentaba mejores métricas, el coeficiente de determinación fue de 0.933 (el más cercano a la unidad) y su error cuadrático medio obtenido fue de 0.061 (el más pequeño) razón por la cual fue el método elegido.

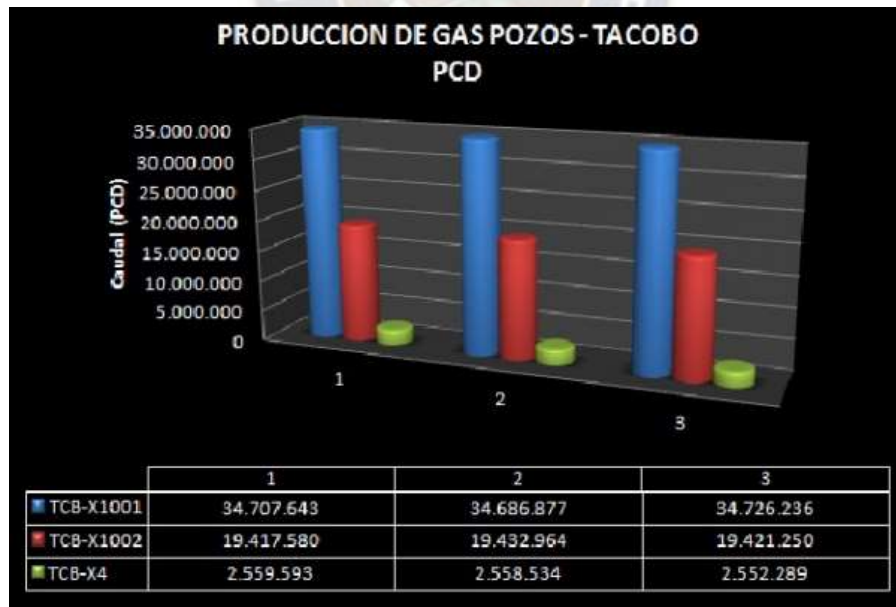


## 5.2. RECOMENDACIONES

Las recomendaciones son las siguientes:

- Si bien este análisis se aplicó solo para un pozo, también podría aplicarse para 1000 o más pozos y es ahí donde surge la necesidad de usar nuevas tecnologías.
- Mediante una investigación sobre el tema se determinó que la presión y temperatura son variables de vital importancia ya que van muy relacionadas con posibles anomalías que podrían surgir en el separador, lastimosamente no se cuentan con los correspondientes, así que se asumió como constante estos datos.
- Se evidenció que la empresa estatal del país no usaba técnicas de visualización de datos correctas, en el informe técnico se encontró gráficas para transmitir información que no son las adecuadas. El siguiente ejemplo fue extraído de dicho informe.

Figura 5.1: Producción de Gas pozo Tacobo-PCD

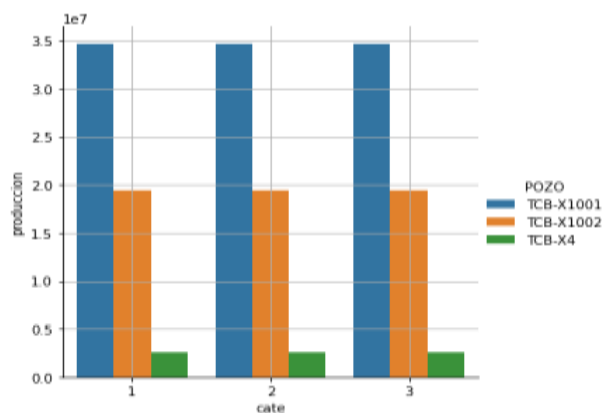


Fuente: Informe técnico campos Tacobo-Curiche YPFB.

El anterior grafico es muy confuso ya que presenta valores expresados en un fondo en 3D esta es una técnica mala para contar historia con datos. El efecto 3D está por demás en la gráfica y además que usa cilindro. Esta sobre trabajado y lo que puede provocar es esfuerzo cognitivo extra en el lector de la imagen. Una buena manera de contar historias con los datos es el grafico de barras.

La solución planteada al anterior grafico en matplotlib es la siguiente:

Figura 5.2: Posible grafico para expresar la producción por pozo.



Fuente: Elaboración propia usando Python-Seaborn-Matplotlib.

- Usar aparatos de medición que usen tecnología de la industria IOT esto para un correcto almacenamiento de los datos, para posteriormente aplicar técnicas de inteligencia artificial para mejores procesos en toda la industria petrolera en el país. Con la implementación de IOT e inteligencia artificial sin duda ayudara a una pronta respuesta a problemas que afronta la industria cada día.
- Ser conscientes que la inteligencia artificial reemplazara en tareas que realizan hoy por hoy ingenieros, con esto no señalo un reemplazo, pero si sugieren que las herramientas de inteligencia artificial debería ser pieza fundamental en la formación de futuros y actuales ingenieros en el país.

## **BIBLIOGRAFIA**

- Abuín, J. R. (2007). Regresión lineal múltiple. Estadística, 32.
- Berastegui Moreno Jose, Beatriz Malagon Picon -Departamento de transportes y tecnología de proyectos y procesos- Universidad de Cantabria.
- Cartagena, L. (2016). Explotación del reservorio petaca del campo Yapacani mediante la determinación de caudales óptimos de gas condensado. Santa Cruz de la Sierra: EMI.
- Carrasquilla-Batista, A., Chacón-Rodríguez, A., Núñez-Montero, K., Gómez-Espinoza, O., Valverde, J., & Guerrero-Barrantes, M. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. Revista Tecnología en Marcha.
- Cesar Pérez López (2007), “Econometría básica. Técnicas y herramientas”
- De la Cruz, L. (2003). Producción de petróleo. Manual. Centro de Ingenieros del estado Carabobo, Valencia. (Nuñez Castaño, 2019)
- Díaz Galarza, A. A., Paguay Loor, Á. L., & Tomalá De La Cruz, H. M. (2011). Parámetros Requeridos en el Diseño de Separadores para Manejar Mezclas Gas-Líquido (Bachelor's thesis, ESPOL. FICT).
- G.P.A. Estudios y Servicios Petroleros S.R.L. ESPUMAS EN SISTEMAS DE HIDROCARBUROS: ORIGEN, CONSECUENCIAS Y SOLUCIONES. Ciudad Autónoma de Buenos Aires.
- Granados, R. M. (2016). Modelos de regresión lineal múltiple. Granada, España: Departamento de Economía Aplicada, Universidad de Granada.
- Laguna, C. (2014). Correlación y regresión lineal. Instituto Aragonés de Ciencias de la Salud, 4, 1-18.
- Ligdi, G. (s.f.). aprendeIA. Obtenido de [https://escuela.aprendeia.com/author/escuela\\_aprendeia/](https://escuela.aprendeia.com/author/escuela_aprendeia/)
- MATPLOTLIB. (2022). matplotlib.org. Obtenido de [https://matplotlib.org/stable/plot\\_types/index.html](https://matplotlib.org/stable/plot_types/index.html)

- Méndez Arcos, P. (s.f.). Minera de Datos. Extracción no trivial de información.
- minerva. (s.f.). knowledge Discovery in Databases. Obtenido de <https://mnrva.io/kdd-platform.html>
- Mora, J. P. (s.f.). ¿Qué es la Ciencia de Datos, el aprendizaje automático, el Big Data y cuáles son sus usos? Universidad Javeriana.
- oleshop. (s.f.). oleshop sell better. Obtenido de <https://www.oleoshop.com/blog/que-es-una-metrica>.
- PANDAS. (2022). pandas.pydata.org. Obtenido de <https://pandas.pydata.org/pandas-docs/version/1.4.3/>
- PDVSA. (2016). Cuadernos de Soberanía Petrolera. Obtenido de <http://sch.pdvsa.com/interface.sp/database/fichero/publicacion/2619/193.PDF>
- Peláez, I. M. (2016). Modelos de regresión: lineal simple y regresión logística. Revista Seden, 14, 195-214.
- Rodríguez, E. M. (2005). Errores frecuentes en la interpretación del coeficiente de determinación lineal. Anuario jurídico y económico escurialense, (38), 315-331.
- Rodríguez Jose. (2006). Diseño y evaluación de separadores bifásicos y trifásicos. Caracas: Universidad Central de Venezuela.
- SEABORN. (2022). seaborn.pydata.org. Obtenido de [https://seaborn.pydata.org/examples/scatterplot\\_matrix.html](https://seaborn.pydata.org/examples/scatterplot_matrix.html)
- Silva Z, C., & Salinas, M. (2006). Modelos de regresión y correlación. Cienc. Trab, 185-189.
- SCIKIT-LEARN. (2022). sklearn. Obtenido de [https://scikitlearn.org/stable/supervised\\_learning.html#supervised-learning](https://scikitlearn.org/stable/supervised_learning.html#supervised-learning)
- “The Data Science Handbook: Advice and Insights from 25 Amazing Data Scientists” de Carl Shan, William Chen, Henry Wang y Max Song.
- Wikipedia. (2015). hmong. Obtenido de [https://hmong.es/wiki/Vortex\\_breaker](https://hmong.es/wiki/Vortex_breaker).
- YPFB. (2012). Informe técnico campos Tacobo-Curiche

## ANEXOS

### Exportación de los datos a Google Colab.

5	CUR_1001_LL	2011-05-31	31.429032	4390.545755	0.404194
6	CUR_1001_LL	2011-06-30	32.090000	4599.873280	0.592000
7	CUR_1001_LL	2011-07-31	32.041935	4565.203819	0.518710
8	CUR_1001_LL	2011-08-31	31.341935	4484.563819	0.454839
9	CUR_1001_LL	2011-09-30	29.960000	4326.112000	0.460667
10	CUR_1001_LL	2011-10-31	26.400000	3549.073548	0.418387
11	CUR_1001_LL	2011-11-30	23.303333	3209.264000	0.268667
12	CUR_1001_LL	2011-12-31	25.309677	3449.481290	0.225806
13	CUR_1001_LL	2012-01-31	21.541935	3015.437419	0.334194
14	CUR_1001_LL	2012-02-29	25.979310	3563.437241	0.407586
15	CUR_1001_LL	2012-03-31	26.012903	3530.492903	0.409032
16	CUR_1001_LL	2012-04-30	26.587097	3674.678710	0.422258
17	CUR_1001_LL	2012-05-31	27.022581	3724.846452	0.405161
18	CUR_1001_LL	2012-06-30	17.256774	2599.825548	0.284839
19	CUR_1001_LL	2012-07-31	19.839355	2897.338839	0.304516
20	CUR_1001_LL	2012-08-31	32.820323	4512.514065	0.324194
21	CUR_1001_LL	2012-09-30	38.117419	5122.739613	0.309355



## Eliminación de las 3 primeras filas originales fila.

```
df = pd.read_excel('/content/CUR_1001_LL - HISTORIA DE PRODUCCIÓN.xlsx', header=4)
df.head(10)
```

```
UNIQUEID    DATE  DAYS  COND BBL    GAS MPC  WATER BBL
0  CUR_1001_LL 2010-09-30   30    26.9   7016.4384    0.00
1  CUR_1001_LL 2011-01-31   31   1207.1  162925.4784    0.00
2  CUR_1001_LL 2011-02-28   28   1125.8  153559.7184    0.00
3  CUR_1001_LL 2011-03-31   31   1557.2  203256.9984   15.00
4  CUR_1001_LL 2011-04-30   30    998.0  138837.1584   10.00
5  CUR_1001_LL 2011-05-31   31    974.3  136106.9184   12.53
6  CUR_1001_LL 2011-06-30   30    962.7  137996.1984   17.76
7  CUR_1001_LL 2011-07-31   31    993.3  141521.3184   16.08
8  CUR_1001_LL 2011-08-31   31    971.6  139021.4784   14.10
9  CUR_1001_LL 2011-09-30   30    898.8  129783.3600   13.82
```

## Conteo de valores nulos por columna.

```
[ ] df.isnull().sum()
```

```
UNIQUEID    0
DATE        0
COND BBL    0
GAS MPC     0
WATER BBL   0
dtype: int64
```

Ver los tipos de datos.

```
[ ] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53 entries, 0 to 52
Data columns (total 5 columns):
#   Column      Non-Null Count  Dtype
---  -
0   UNIQUEID    53 non-null     object
1   DATE        53 non-null     datetime64[ns]
2   COND BBL    53 non-null     float64
3   GAS MPC     53 non-null     float64
4   WATER BBL   53 non-null     float64
dtypes: datetime64[ns](1), float64(3), object(1)
memory usage: 2.2+ KB
```

Métricas del método 1 modelo de regresión lineal simple

## ▼ Metodo 1

regresion lineal simple

```
✓ [70] print('-----METRICAS DE REGRESION LINEAL SIMPLE EN EL POZO-----')
0s y_pred=modelo_rls.predict(X_test1)
r_2 = r2_score(y_test1,y_pred)
mse = mean_squared_error(y_test1,y_pred)
mape = mean_absolute_percentage_error(y_test1,y_pred)
print('La medida del Ceficiente de Determinacion:',r_2)
print('La medida del Error cuadratico medio es:',mse)
print('La medida del Error porcental absoluto medio es:', mape)
```

```
-----METRICAS DE REGRESION LINEAL SIMPLE EN EL POZO-----
La medida del Ceficiente de Determinacion: 0.932640197775592
La medida del Error cuadratico medio es: 84924.72860018445
La medida del Error porcental absoluto medio es: 0.07357111667842409
```

## Métricas del método 2 modelo de regresión lineal múltiple

### ▼ Metodo 2

regresion lineal multiple

```
[73] X = df.iloc[:,[1,2]].values
      y = df.iloc[:,[0]].values
      X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.30, random_state=42)

#Creo el modelo
modelo_rlm = LinearRegression()
modelo_rlm.fit(X_train, y_train )

LinearRegression()

[75] print('-----METRICAS DE REGRESION LINEAL MULTIPLE EN EL POZO 12-----')
      y_pred=modelo_rlm.predict(X_test)
      r_2 = r2_score(y_test,y_pred)
      mse = mean_squared_error(y_test,y_pred)
      mape = mean_absolute_percentage_error(y_test,y_pred)
      print('La medida del Ceficiente de Determinacion:',r_2)
      print('La medida del Error cuadratico medio es:',mse)
      print('La medida del Error porcental absoluto medio es:', mape)

-----METRICAS DE REGRESION LINEAL MULTIPLE EN EL POZO 12-----
La medida del Ceficiente de Determinacion: 0.9290996212917767
La medida del Error cuadratico medio es: 89388.55549763433
La medida del Error porcental absoluto medio es: 0.07019054285028554
```

## Métricas del método 3 modelo de regresión lineal simple eliminando los outliers.

```
#Entrenaran con el 70% y divido los datos
X = df_out.iloc[:,[1]].values
y = df_out.iloc[:,[0]].values
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.30, random_state=42)

[82] #Creo el modelo
      modelo_rls = LinearRegression()
      modelo_rls.fit(X_train, y_train )

LinearRegression()

[83] print('-----METRICAS DE REGRESION LINEAL SIMPLE EN EL POZO-----')
      y_pred=modelo_rls.predict(X_test)
      r_2 = r2_score(y_test,y_pred)
      mse = mean_squared_error(y_test,y_pred)
      mape = mean_absolute_percentage_error(y_test,y_pred)
      print('La medida del Ceficiente de Determinacion:',r_2)
      print('La medida del Error cuadratico medio es:',mse)
      print('La medida del Error porcental absoluto medio es:', mape)

-----METRICAS DE REGRESION LINEAL SIMPLE EN EL POZO-----
La medida del Ceficiente de Determinacion: 0.8969521524015417
La medida del Error cuadratico medio es: 84015.59144431516
La medida del Error porcental absoluto medio es: 0.06904308099807413
```

## Métricas del Método 4 modelo regresión lineal simple con normalización.

```
▶ import sklearn.metrics as metrics
mse = metrics.mean_squared_error(y_test,y_pred)
r2 = metrics.r2_score(y_test, y_pred)
mape = metrics.mean_absolute_percentage_error(y_test,y_pred) / 40.5

print("r2 ", r2.round(4))
print("mse: ", mse.round(4))
print('mape: ', mape.round(4))
```

```
▶ r2  0.9317
mse:  0.0628
mape:  0.0693
```

## Métricas del método 5 regresión lineal múltiple y normalizando.

```
▶ X_cols = ['COND BBL', 'WATER BBL']
y_col = ['GAS MPC']

X = df[X_cols].values
y = df[y_col].values

X_train, X_test, y_train, y_test = train_test_split(X,y, random_state=42)
sc_x = StandardScaler().fit(X)
sc_y = StandardScaler().fit(y)

X_train = sc_x.transform(X_train)
X_test = sc_x.transform(X_test)
y_train = sc_y.transform(y_train)
y_test = sc_y.transform(y_test)

model = LinearRegression()
model.fit(X_train,y_train)
y_pred = model.predict(X_test)
```

```
▶ import sklearn.metrics as metrics
mse = metrics.mean_squared_error(y_test,y_pred)
r2 = metrics.r2_score(y_test, y_pred)
mape = metrics.mean_absolute_percentage_error(y_test, y_pred)
print("r2 ", r2.round(4))
print("mse: ", mse.round(4))
print('mape:', mape)
```

```
r2  0.9277
mse:  0.0665
mape:  0.06726842285856159
```

## Métricas del método 6 regresión lineal simple, normalizando y eliminando variables.

eliminando el intercepto en la regresion lineal simple

```
▶ X_cols = ['COND BBL']
  y_col = ['GAS MPC']

  X = df[X_cols].values
  y = df[y_col].values

  X_train, X_test, y_train, y_test = train_test_split(X,y, random_state=42)
  sc_x = StandardScaler().fit(X)
  sc_y = StandardScaler().fit(y)

  X_train = sc_x.transform(X_train)
  X_test = sc_x.transform(X_test)
  y_train = sc_y.transform(y_train)
  y_test = sc_y.transform(y_test)

  model = LinearRegression(fit_intercept=False)
  model.fit(X_train,y_train)
  y_pred = model.predict(X_test)
```

```
[297] mse = metrics.mean_squared_error(y_test,y_pred)
      r2 = metrics.r2_score(y_test, y_pred)
      mape = metrics.mean_absolute_percentage_error(y_test,y_pred)

      print("r2 ", r2.round(4))
      print("mse: ", mse.round(4))
      print('mape: ', mape.round(4))
```

```
r2 0.9328
mse: 0.0618
mape: 0.0663
```

## Métricas del método 7

```
[21] X = df.iloc[:,[0,2,3,4]].values
      y = df.iloc[:,[1]].values
      X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.30, random_state=42)

[22] #Creo el modelo
      modelo_rls = LinearRegression()
      modelo_rls.fit(X_train, y_train )

      LinearRegression()

      print('-----METRICAS DE REGRESION LINEAL SIMPLE EN EL POZO-----')
      y_pred=modelo_rls.predict(X_test)
      r_2 = r2_score(y_test,y_pred)
      mse = mean_squared_error(y_test,y_pred)
      print('La medida del Ceficiente de Determinacion:',r_2)
      print('La medida del Error cuadratico medio es:',mse)

      -----METRICAS DE REGRESION LINEAL SIMPLE EN EL POZO-----
      La medida del Ceficiente de Determinacion: 0.9290996212917767
      La medida del Error cuadratico medio es: 89388.55549763433
```

## Puesta en marcha del algoritmo

```
condensado = float(input('Introduzca la cantidad de condensado en la salida del separador: '))
gas_predict = pendiente*condensado
gas_real = float(input('Introduzca la cantidad de gas en MPC registrada en la salida del separador: '))

if (gas_real <= gas_predict*1.15) and (gas_real >= gas_predict*0.85):
    print('****NO SE PRODUCE ANOMALIAS****')
else:
    print('****SE PRODUJO UN VALOR FUERA DE LOS NORMAL****')
    if gas_real < gas_predict:
        print('El gas de Salida es menor a lo esperado por el modelo')
        print('Es posible que se hallan formado un vertices en el separador')
    elif gas_real > gas_predict:
        print('El gas de salida es mayor a lo esperado por el modelo')
        print('Es posible que se hallan formado espuma dentro de el separador')

print('El valor prededido de gas a la salida es de:', gas_predict)
viz_train = plt
viz_train.plot(X_train1, modelo_rls.predict(X_train1), color = 'black')
viz_train.scatter(condensado,gas_real,color='red',label='Gas de salida real')
viz_train.scatter(condensado,gas_predict,color='blue',label='Gas de salida prededido')
viz_train.title('Regresion lineal para pozo CUR-X1001228')
viz_train.xlabel('Condensado')
viz_train.ylabel('Gas')
viz_train.legend()

viz_train.show()
```



