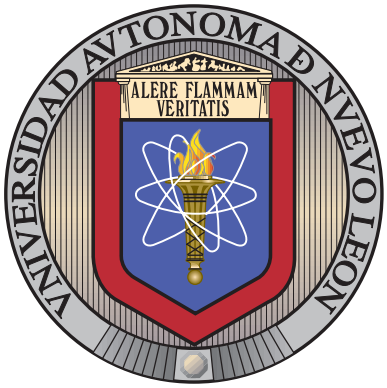
Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección de Estudios de Posgrado



Descubrimiento de conocimiento en pozos de petróleo basado en datos geológicos

por

Daniel Chong Sánchez

como requisito parcial para obtener el grado de

maestría en ciencias de la ingeniería

con Orientación en Sistemas

Octubre de 2020

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección de Estudios de Posgrado

Descubrimiento de conocimiento en pozos de petróleo basado en datos geológicos

por

Daniel Chong Sánchez

como requisito parcial para obtener el grado de

maestría en ciencias de la ingeniería

con Orientación en Sistemas

Octubre de 2020

Universidad Autónoma de Nuevo León
Facultad de Ingeniería Mecánica y Eléctrica
Subdirección de Estudios de Posgrado

Los miembros del Comité de Tesis recomendamos que la Tesis "Descubrimiento de conocimiento en pozos de petróleo basado en datos geológicos", realizada por el alumno Daniel Chong Sánchez, con número de matrícula 1985274, sea aceptada para su defensa como requisito para obtener el grado de Maestría en Ciencias de la Ingeniería con Orientación en Sistemas.

El Comité de Tesis

Dr. Fernando López Irarragorri
Director

Dr. Igor Semionovich Litvinchev
Revisor

Dr. César Medina Trejo
Revisor

Vo. Bo.

Dr. Simón Martínez Martínez
Subdirector de Estudios de Posgrado

SUBDIRECCIÓN DE ESTUDIOS DE POSGRADO
FIME
053

San Nicolás de los Garza, Nuevo León, octubre de 2020

Ciudad Universitaria Pedro de Alba s/n, C.P. 66455. A.P. 076 Suc. "F"
San Nicolás de los Garza, Nuevo León, México. Tels: (81) 8332 0903 / Conm.: 8329 4020 / Fax: (81) 8332 0904

A mis padres

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

Agradezco a las personas más importantes en mi vida. Mis padres Pedro David Chong y Maricela Sánchez quienes me enseñaron a luchar en la vida y a quienes les debo ser lo que soy. A mis tíos, y primos que me apoyaron.

Agradezco a la Facultad de Ingeniería Mecánica y Eléctrica y a la Universidad Autónoma de Nuevo León por el apoyo para la realización de este proyecto.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada durante mis estudios de maestría, sin la cual no hubiera sido posible realizar estos estudios y alcanzar el objetivo.

Al Posgrado de Ingeniería de Sistemas (PISIS) y a todo su colectivo de excelentes profesores quienes me ayudaron a mi formación académica y me brindaron su tiempo y paciencia, así como a los directivos y administrativos que apoyaron en la gestión de los trámites.

Agradezco en toda la extensión de la palabra a mi director de tesis el Dr. Fernando López Irarragori por brindarme su tiempo, conocimiento y mucha, pero mucha paciencia para la realización de este trabajo por brindarme su apoyo, gracias por su amistad.

Agradezco a mi comité de tesis al Dr. César Medina Trejo y al Dr. Igor Semionovich Litvinchev por el tiempo que le dedicaron, por sus observaciones y contribuciones a este trabajo.

Agradezco a mis amigos por escucharme, ayudarme, compartir conmigo sus conocimientos, transmitirme sus consejos y no dejarme caer en los momentos duros. Y a mis compañeros quienes hicieron mi estancia como estudiante de PISIS agradable.

A quienes confiaron en mí y que aportaron de una forma u otra un granito de arena para alcanzar este meta.

¡¡¡Muchas gracias a todos!!!

Resumen

Daniel Chong Sánchez. Candidato para obtener el grado de Maestría en Ciencias de Ingeniería con orientación en Sistemas. Universidad Autónoma de Nuevo León. Facultad de Ingeniería Mecánica y Eléctrica.

Título del estudio: Descubrimiento de conocimiento en pozos de petróleo basado en datos geológicos.

Número de páginas: 155.

Objetivos y método de estudio: En la presente investigación se ha establecido como objetivo general proponer una nueva metodología de apoyo a la decisión multicriterio y la metodología fundamental para la ciencia de datos, propuesta por IBM, que contribuya a la caracterización de yacimientos de petróleo a partir de datos geológicos.

Contribuciones y conclusiones: La metodología formulada puede aportar valor en los proyectos de prospección de petróleo. Además, se establece la relación entre las características de yacimientos de petróleo a partir de la aplicación de métodos de aprendizaje automatizado y MCDM para pozos heterogéneos.

Firma del asesor: _____

Dr. Fernando López Irarragorri

Capítulo 1

Introduccion

1.1 Introduccion

En el presente capítulo se describe el diseño de la investigación, que incluye el contexto, los antecedentes, la descripción del problema científico, la justificación, la necesidad, los objetivos principales y los objetivos secundarios, los resultados esperados, la novedad científica y eventualmente las tareas científicas y las conclusiones.

1.2 Contexto

La industria del petróleo tuvo su auge a partir de la década de los años 70 aplicando las técnicas convencionales de exploración que existían en ese momento, como son: los métodos sísmicos, los métodos geo-eléctricos, los métodos electromagnéticos [107] con el que se crean mapas de variaciones en el espesor de las principales cuencas sedimentarias, la técnica de gravimetría, la técnica de magnetometría [43] que se basa en la medición de propiedades físicas de la roca [87], entre otras técnicas, para identificar yacimientos que están más cerca de la superficie, así como identificar el potencial de áreas de hidrocarburos, entre otras. Con el paso de los años los yacimientos de petróleo en tierra se han agotado, obligando a los geólogos e inge-

nieros a buscar nuevos yacimientos en mar adentro y a mayor profundidad. Esto ha traído como consecuencia el aumento de los costos y una caracterización y evaluación compleja de nuevos yacimientos como parte de la etapa de exploración inicial.

En un período relativamente reciente, se han incorporado nuevas técnicas como son: modelos matemáticos, la técnica de segmentación litológica, la simulación de yacimientos en 3 dimensiones y técnicas de inteligencia artificial (*Artificial Intelligence, AI*). Estas últimas constituyen alternativas a considerar en el análisis de datos de yacimientos de petróleo para facilitar la identificación, caracterización y evaluación de nuevos reservorios a partir de datos geológicos [43].

Con el transcurso del tiempo, se ha acumulado experiencia en la aplicación de varias técnicas de inteligencia artificial utilizando redes neuronales artificiales (*Artificial Neural Networks*, ANN), máquina de soporte vectorial (*Support Vector Machine, SVM*), lógica difusa (*Fuzzy Logic*), algoritmos genéticos (*Genetic Algorithms*), árbol de decisión (*Decision tree*) y combinaciones de estas técnicas, entre otras, lo que ha permitido evaluar distintos atributos que caracterizan a los yacimientos como: la permeabilidad, la porosidad, la saturación, la profundidad, entre los más importantes. Estas técnicas permiten procesar grandes volúmenes de información que se generan de las mediciones que se realizan en los pozos de exploración, y hacer predicciones de los yacimientos que apoyan la toma de decisiones, incluso bajo condiciones de incertidumbre o cuando los datos están incompletos [114].

En un ámbito más amplio se han utilizado técnicas de máquina de soporte vectorial y redes neuronales para hacer predicciones de las propiedades petrofísicas como la porosidad, utilizando los datos de reservas de los campos del oeste de Australia con resultados prometedores en la predicción de este parámetro [113], por otro lado en las zonas fracturadas en el campo petrolífero de Hassi Messaoud, Argelia, se aplicó con éxito el método híbrido de inteligencia artificial formado por técnicas de redes neuronales y lógica difusa, para caracterizar y predecir reservorios naturalmente fracturados en base a los parámetros de porosidad, permeabilidad y volumen

de lutita, con muy buenos resultados [1].

Entre los ejemplos que se pueden mencionar existen en los campos de petróleo canadiense se han aplicado técnicas de inteligencia artificial, en específico redes neuronales y máquina de soporte vectorial, para predecir la viscosidad del petróleo crudo, obteniendo resultados exitosos con aceptable precisión [49]. Además, se ha utilizado en la costa afuera del este de Canadá, en el campo de gas de Venture, la técnica de Redes Neuronales de back propagation (*BP-ANN*) utilizando seis registros de pozos diferentes y la permeabilidad para modelar las interrelaciones entre la posición espacial, con el fin de estimar la permeabilidad [57].

Por otra parte, en el marco de la industria del petróleo en México, existe la voluntad de aplicar técnicas de inteligencia artificial junto con estrategias de minería robusta de datos, técnicas de modelación sísmica en 2D, 3D y 4D, caracterización avanzada de pozos, entre otras, para acceder a importantes reservas adicionales de hidrocarburos con que cuenta el país. Además, optimizar la producción, así como identificar el volumen de hidrocarburo remanente existente en yacimientos, entre otras actividades. Existen zonas en el territorio mexicano con potencial para aplicar estas técnicas como son: los campos Akal y Ku [47].

1.3 Antecedentes

1970s

Los primeros trabajos se remontan al año 1977 [30], en donde aparecen las primeras versiones de la aplicación en el campo de las geociencias de la Lógica Difusa como queda reflejado en el trabajo de Chapaz, a partir de los precedentes establecidos en el trabajo previo de Zadeh de 1965, donde quedaron plasmados las bases de la teoría de conjuntos difusos. En el trabajo de Chapaz, se propone el uso de la teoría de conjuntos difusos para la interpretación de secciones sísmicas.

1980s

Más adelante, en una época posterior en la década de los 80, se desarrollaron varios trabajos sobre la caracterización del proceso de exploración de pozos de petróleo aplicando técnicas de inteligencia artificial, como queda reflejado en los trabajos de Aminzadeh y Chatterjee de 1984 [9], en donde se aplica la técnica de clustering a la exploración sismológica. En 1989 en el trabajo de Liu et al. [71], se puso de manifiesto la aplicación de las redes neuronales para rastrear eventos sísmicos.

1990s

Un poco después, en la década de los 90 se realizaron varios trabajos entre los que está el de Holtz et al. de 1998 [53], en donde se propone una metodología para caracterizar el potencial de crecimiento de la reserva. En este trabajo a partir de la determinación de la arquitectura de la reserva, se realiza un análisis de las tendencias de flujo de fluido y como paso importante se identifica la correspondencia entre la arquitectura del reservorio y las tendencias del flujo de fluidos y así poder establecer cuál grupo actúa como compartimiento o unidad de flujo. Otros trabajos que aparecieron en esta década son el de Ahmed et al. de 1997 [4] y el de Mohaghegh et al. de 1996 [79]. En este último, a partir de diferentes propiedades del yacimiento, como la porosidad, la permeabilidad y la saturación de fluidos, en reservas altamente heterogéneas se propone una metodología para la caracterización de pozos de petróleo a partir de la aplicación de redes neuronales, utilizando registros de pozos geofísicos disponibles. Esta red permitió estimar los parámetros de formación utilizados en los simuladores de yacimientos. Además, este trabajo demuestra que puede haber una relación, no importa cuán compleja sea la naturaleza, entre los rayos gamma, la densidad aparente y las respuestas de registro de inducción profunda con la permeabilidad de la formación.

Otro de los trabajos de este período es el de Balch et al. de 1999 [20], en el que se propone una metodología, desarrollada y probada para relacionar las propiedades del yacimiento con conjuntos de atributos sísmicos, con el fin de predecir las propiedades

del yacimiento en dos zonas del campo Nash Draw en Nuevo México. Debido a que los 350 atributos no pueden ser utilizados computacionalmente, se utilizó la lógica difusa para seleccionar la mayoría de los atributos estadísticamente significativos y así desarrollar ecuaciones de regresión para propiedades de reservorios individuales. Para este trabajo los datos de salida fueron una propiedad del reservorio: la porosidad o saturación de agua o net pay; correspondientes a 19 pozos. Una parte sustancial de esta metodología es la red neuronal entrenada para estimar cada propiedad. Por otra lado, en 1996, el trabajo de Huang et al. [57], propone una metodología para predecir la permeabilidad a partir de registros de pozos aplicando redes neuronales.

En los últimos años, se ha dedicado una atención considerable al uso de enfoques de lógica difusa de red neuronal híbrida como alternativa para el reconocimiento de patrones, agrupamiento y modelado estadístico y matemático. Se ha demostrado que se pueden usar modelos de redes neuronales para construir modelos internos que capturan la presencia de reglas difusas.

2000s

En la primera década a partir del año 2000, continúa el interés por aplicar las técnicas de Inteligencia Artificial en la caracterización de pozos de petróleo analizando grandes volúmenes de datos geológicos como queda reflejado en varios trabajos como el Mohaghegh et al. de 2000 [80] y el trabajo de Mohaghegh et al. del 2005 [81]. Varias de estas técnicas de aprendizaje automatizado se han venido aplicando con el objetivo de determinar las relaciones entre varios de los atributos que caracterizan a los pozos de petróleo como queda reflejado en el trabajo de Nikravesh y Fred Aminzadeh del 2001 [75].

Dentro de los trabajos revisados, las redes neuronales son las técnicas de uso más frecuentes, así como la técnica de SVM, entre otras como lo reflejan el trabajo de Aminian et al. del 2005 [6], Kotsiantis et al. del 2007 [64], entre otros, que se utilizan para el análisis, interpretación, estimación y predicción a partir de datos geológicos para determinar las relaciones entre los atributos que caracterizan a los

pozos, por ejemplo en la determinación de la relación entre las propiedades porosidad y permeabilidad como queda reflejado en el trabajao de Anifowose [13], no obstante aún existen limitaciones en estos algoritmos para determinar esta relación.

2010s

Más recientemente y hasta la fecha, concretamente en la última década, se han realizado otros trabajos como el de Anifowose et al. del año 2015 [13], Mirzaei et al. del 2012 [77] en el que se pone de manifiesto el interés en aplicar el paradigma basado en la técnica de aprendizaje de conjuntos (*Ensemble*) para resolver los problemas más desafiantes en esta industria, como es el caso del primer trabajo mencionado del 2015. En este se propone un modelo de conjunto de generalización apilado de SVM que incorpora diferentes opiniones de expertos sobre los valores óptimos de este parámetro en la predicción de la porosidad y la permeabilidad de los yacimientos de petróleo utilizando conjuntos de datos de diversas formaciones geológicas. En el segundo trabajo, se presenta una nueva metodología usando la técnica de red neuronal artificial para predecir el rango de flujo de producción de petróleo, cuando usualmente son utilizadas correlaciones empíricas, pero esto trae consigo valores imprecisos. Más recientemente, en el trabajo de Oloso et al. de 2017 [85] se propone un sistema híbrido usando agrupamiento a través de K-Means y una red funcional para predecir las propiedades PVT(Presión-Volumen-Temperatura) de una reserva de petróleo crudo.

A pesar de que se han aplicado varias de técnicas de inteligencia artificial, tanto de forma única o en conjunto con otros algoritmos formando sistemas híbridos, varias de estas técnicas aún presentan deficiencias para afrontar los problemas desafiantes en la caracterización de reservas a partir de datos geológicos, como es el caso de las ANN, en las que una de las dificultades es determinar el algoritmo de aprendizaje más idóneo para un desempeño óptimo del modelo [12], en el caso de la técnica de SVM, su desempeño está sujeto al refinamiento de sus parámetros, más especialmente la regularización del parámetro C como se refleja en [13], para otros algoritmos se les

dificulta afrontar situaciones donde existe desbalance de clases [], por todo lo anterior aún persiste la necesidad de técnicas que brinden resultados más precisos de forma generalizada tanto en situaciones de baja como de alta diversidad ya sea a partir del análisis de pequeñas muestras como de grandes volúmenes de datos.

1.4 Problema científico e hipótesis

Formulación del problema científico: ¿Es posible hacer caracterizaciones eficientes de pozos de petróleo con características heterogéneas a partir de información geológica?

Hipótesis: El desarrollo de una metodología de apoyo a la decisión basada en la ciencia de datos y la IA facilita la generación de soluciones razonables para proyectos de prospección de pozos de petróleo a partir de datos geológicos.

1.5 Justificación

La importancia fundamental de este trabajo radica en que aportará una nueva metodología en la detección de zonas geológicas de petróleo, lo cual constituirá una herramienta que facilite la identificación, clasificación y evaluación de yacimientos con posibilidades de prospección.

Existe un motivo principal por el que las técnicas de aprendizaje automatizado entre las que se encuentran las redes neuronales, máquina de soporte vectorial, entre otras, están evolucionando hacia una nueva etapa de avance, se debe a que en la actualidad existen grandes volúmenes de datos de los cuales se pueden extraer características y patrones que facilitarían una gran cantidad de tareas, las cuales en este punto se convierten en tareas imposibles de procesar contando únicamente con el esfuerzo humano. Las herramientas de aprendizaje automatizado son capaces de

obtener provecho de este bloque de información en poco tiempo y haciendo buen uso de recursos computacionales pueden llegar a ser de gran utilidad.

De manera particular en el ámbito de la industria del petróleo e hidrocarburos, como un tema más específico, se genera una gran cantidad de información que puede beneficiarse de las herramientas de aprendizaje automatizado para determinar patrones y conformar con mayor exactitud los perfiles de los pozos en exploración, añadiendo a esto que, estas actividades de exploración son costosas y para caracterizar, clasificar y elaborar perfiles de pozos se requiere de personal altamente especializado, por lo que las técnicas de AI pueden aportar un alto valor a la elaboración de estos perfiles, así como disminuir los costos de operación en estas actividades.

De forma general la mayoría de las perforaciones de pozos en la actualidad son cada vez más profundas, para poder extraer el hidrocarburo del reservorio en el subsuelo, debido a esto, cada vez se genera mucha más información que en décadas anteriores, creando inmensas bases de datos de pozos de exploración y esto propicia aún más el uso de las herramientas de AI en esta industria por parte de los especialistas. En este contexto, situaciones como el trabajo con datos incompletos de los pozos es cada vez más frecuente, debido a las condiciones difíciles en las que se realizan las mediciones, esto conlleva a que las tareas de clasificación y caracterización de los datos geológicos obtenidos se vuelve una tarea compleja, y esto puede ser aprovechado para aplicar las técnicas de AI.

La etapa de Exploración de pozos, constituye el punto de partida de cualquier proyecto de perforación, y es de vital importancia, por lo que la tarea de clasificación y caracterización de pozos adquiere relevancia toda vez que, la obtención de perfiles adecuados y precisos de pozos de exploración, entre otros factores, le antecede a continuar con las etapas de Evaluación, Desarrollo y Producción, por lo cual, una metodología que contribuya a agilizar este propósito tiene bastante relevancia.

1.6 Objetivo general

Establecer cómo influye la geología del terreno en la existencia o no del petróleo en una localización geográfica aplicando métodos de aprendizaje automatizado desde una perspectiva de apoyo a la decisión multicriterio.

1.7 Objetivos específicos

Objetivo Específico 1: Desarrollar una metodología de apoyo a la decisión para descubrir conocimiento en base a los datos geológicos de pozos de petróleo relacionados con la existencia o no de yacimientos de petróleo en los mismos.

Objetivo Específico 2: Aplicar una metodología de aprendizaje automatizado para establecer relaciones entre las características geológicas de pozos de petróleo y estimar la posible existencia o no de de petróleo en los mismos.

1.8 Resultados esperados

Con el desarrollo de este trabajo se establecerá cómo influye la geología del terreno en el resultado del payflag, contribuyendo al diseño de una nueva metodología para la caracterización y evaluación de yacimientos de petróleo basado en datos geológicos heterogéneos.

1.9 Novedad científica

Propuesta de una metodología de apoyo a la decisión multicriterio unido al aprendizaje automatizado para relacionar varias características de yacimientos ba-

sado en datos geológicos, que posibilite clasificar, evaluar y finalmente determinar si existe o no petróleo en los yacimientos.

1.10 Estructura de la tesis

El presente documento está organizado de la siguiente manera:

- En el Capítulo 2, se presenta el marco teórico, abarcando los temas relacionados con el problema, y la perspectiva desde la cual se aborda el problema. Se exponen temas como la optimización monoobjetivo, la optimización multiobjetivo, el ciclo y fases de un pozo de petróleo, entre otros temas. También, se realiza una revisión bibliográfica y se presenta el estado del arte, en donde se abordan los trabajos más recientes relacionados con el tema de este trabajo.
- En el Capítulo 3 se plantea detalladamente el problema, presentando las variables a considerar en la caracterización de pozos y los supuestos considerados.
- En el Capítulo 4 se describe la metodología general que se sigue para resolver la problemática de este trabajo.
- En el Capítulo 5 se presentan las conclusiones y recomendaciones.

1.11 Conclusiones

En este capítulo queda presentado de manera formal el problema y además se describe la estructura general del trabajo.

Capítulo 2

Marco Teórico y estado del arte

2.1 Introduccion

En este capítulo se presentan entre otros temas las bases teóricas sobre las cuales se desarrolla este trabajo de tesis, así como las técnicas de aprendizaje automatizado y se aborda además el método de aprendizaje de conjuntos (*Ensemble*). Además, se presentan las fases de desarrollo de un pozo de petróleo y la revisión de la literatura existente.

2.2 Ciclo y fases de un pozo de petróleo

Un proyecto de petróleo está dividido en varias fases, como queda reflejado en la figura 2.1.

Fases del ciclo de vida de un pozo de petróleo [58]:

- Obtener de acceso (*Gaining Access*)

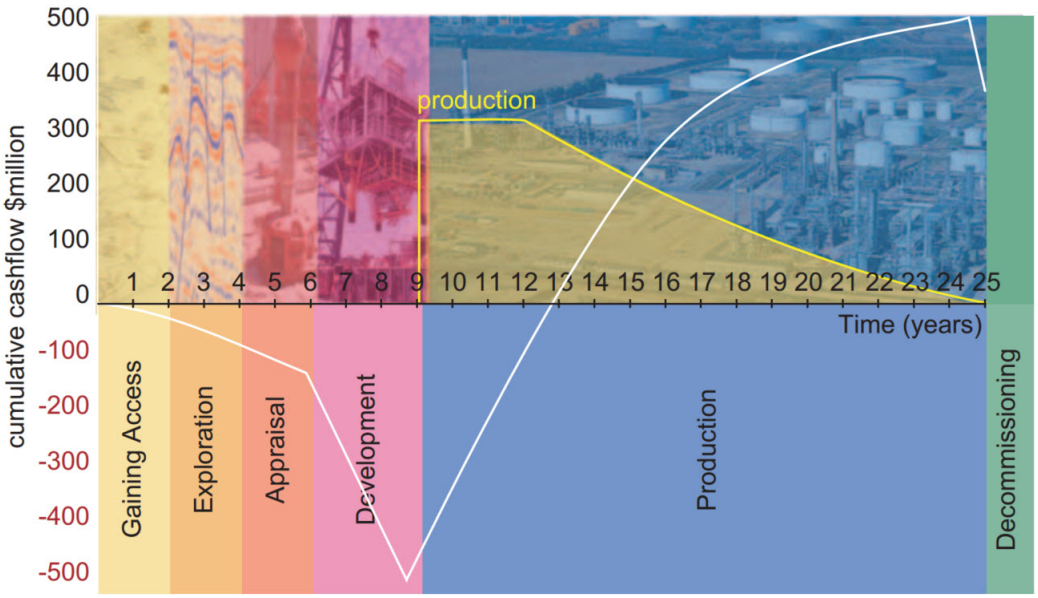


Figura 2.1: Ciclo de vida de un campo de petróleo [58].

- Exploración (*Exploration*)
- Evaluación (*Appraisal*)
- Desarrollo (*Development*)
- Producción (*Production*)
- Desmantelamiento (*Decommissioning*)

El ciclo de vida comienza por la fase de Obtención de acceso, es aquí donde se realizan varios estudios necesarios e imprescindibles de la región de interés que involucrarán una evaluación de aspectos de índole político, legal, social, técnico, económico y medio ambiental. Una vez superada esta fase, le sucede la Exploración del área de interés.

Es durante la fase de Exploración donde se estudia la historia geológica del área y donde la probabilidad de la existencia de hidrocarburos es cuantificada. Trabajo de campo, pruebas magnéticas, pruebas de gravimetría y sísmicas, perforación de pozos

exploratorios, son algunas de las pruebas y actividades tradicionales que se emplean en este período para, en definitiva, determinar a partir de pozos de exploración la existencia de hidrocarburos. Esta es una fase que permanece identificada como actividad de alto riesgo. Es aquí donde está enmarcada la caracterización de pozos de exploración, como queda reflejado en la figura 2.2. Esta fase culmina cuando existe la certeza de haber encontrado evidencia de la existencia de acumulación de hidrocarburo [58].

Figura 2.2: Sub etapas de un proyecto de exploración de petróleo [32].

A continuación, se muestra una imagen de las fases de un programa de exploración típico y el tiempo de duración aproximado de cada una.

Una vez en la fase de evaluación y con la certeza de que se ha encontrado petróleo, aún no existe información sobre el tamaño de la estructura o cavidad geológica, la forma y la productividad de la acumulación y son considerados 4 posibles escenarios:

- Proceder con el desarrollo y así generar ingresos dentro de un corto periodo de

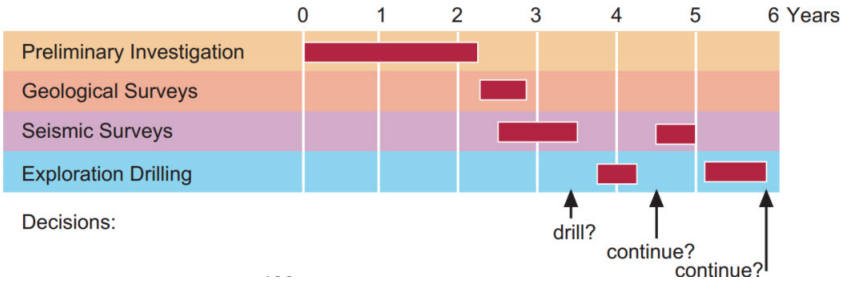


Figura 2.3: Fases de un programa de exploración típico [58].

tiempo. El riesgo es que el campo resulte ser más grande o más pequeño de lo previsto, las instalaciones serán más grandes o más pequeñas y la rentabilidad del proyecto pueden verse afectadas.

- Realizar un programa de evaluación con el objetivo de optimizar el desarrollo técnico. Esto puede retrasar la producción del "primer petróleo"del campo por varios años y puede aumentar la inversión inicial requerida. Sin embargo, en general la rentabilidad del proyecto puede mejorarse.

- Vender el descubrimiento, en cuyo caso se requerirá una valoración. Algunas compañías se especializan en aplicar sus habilidades de exploración, sin intención de invertir en la fase de desarrollo. Estas empresas crean valor al vender el descubrimiento y luego seguir con la exploración de una nueva oportunidad.

- No ejecutar ninguna acción. Esta es siempre una opción, aunque débil, y puede conducir a frustración en nombre del gobierno de la nación anfitriona, que puede forzar una renuncia si la compañía petrolera continúa retrasando la acción.

El propósito de esta etapa es por lo tanto reducir las incertidumbres, en particular las relacionadas con los volúmenes producibles contenidos dentro de la estructura. En consecuencia, el propósito de la evaluación en el contexto del desarrollo del campo no es encontrar volúmenes adicionales de petróleo o gas. Habiendo definido y recopilado información adecuada para las estimaciones iniciales de la reserva, junto

con la búsqueda de diferentes opciones para desarrollar el campo. El objetivo de esta fase es ejecutar un estudio de factibilidad para documentar las diferentes opciones técnicas, en la cual al menos una debe ser económicamente viable [58].

A continuación, en la fase de desarrollo, a partir del estudio de factibilidad previamente realizado, se conforman los objetivos de desarrollo, así como un plan de desarrollo del campo para ser ejecutado. En esta fase, además, quedan establecidos los principios de operación y mantenimiento, se determinan los costos de mano de obra, se planea el proyecto, se conforma una propuesta de presupuesto, y describen las instalaciones de ingeniería [58].

La fase de producción comienza cuando ya se tienen las primeras cantidades comerciales de hidrocarburo. Esto marca el punto de inflexión desde el punto de vista del flujo de caja, ya que a partir de este momento se genera efectivo y puede utilizarse para pagar las inversiones anteriores.

La fase de desmantelamiento del pozo usualmente termina una vez que los beneficios obtenidos se vuelven permanentemente negativos [58].

2.3 El proceso de la toma de decisión

Personalidades importantes como el premio Nobel, Herbert Simon y otros como Bernard Roy, entre otros, han investigado de manera amplia y extensa, la modelación y materialización de los problemas de las decisiones humanas, que implican un sin número de posibilidades y que han requerido el trabajo combinado de diversas disciplinas como: profesionales estadísticos, licenciados en matemática y economía, ingenieros informáticos, otros profesionales del sector de humanidades como psicólogos, también del sector de la salud como médicos han colaborado y aportado al desarrollo de modelos que tomen en cuenta las decisiones humanas en los diferentes contextos y ámbitos de la sociedad. Estos modelos, en los que progresivamente se ha avanzado, son todavía insuficientes y están sujetos a nuevos avances en las distintas

ramas [25].

La teoría de la decisión está ligada a la toma de decisiones monocriterio y multicriterio. Existen otros elementos que siempre han influido, pero que no se tomaban en cuenta, como son: la incertidumbre, más de un criterio a tener en cuenta o en otro caso varios decisores, etc [76].

Se puede decir que el problema de toma de decisión consiste en elegir alternativas de un grupo de solución para lograr ciertos objetivos formulados por el tomador de decisiones tomando en cuenta las preferencias del decisor [95].

Este proceso consta de varias fases que se presentan a continuación en la figura 2.4

Existen 3 roles con un papel importante en el proceso de toma de decisión [95]:

-el tomador de decisiones (*Decision maker* o *stakeholder*).

-el analista (*the analist*).

-el cliente (*the client*).

El tomador de decisiones (*Decision maker* o *stakeholder*), es la persona o grupo de personas encargados de tomar una decisión, según sus propias preferencias expresadas con respecto a los objetivos a alcanzar y considerando la importancia relativa de cada criterio, puede venir de una firma o un grupo comunitario o de una administración del gobierno. Es importante que actúe libremente y bajo ninguna coacción ya que él, es el protagonista de la toma de decisión. Un problema puede ser planteado por más de un decisor, en este caso se obtienen distintas soluciones según la visión de cada uno y existirán métodos para llegar a una solución de compromiso entre todos en caso de que fuera necesario [95].

El analista, es aquella persona encargada de seleccionar el método cuantitativo de decisión a emplear, y que conociendo las preferencias del decisor o grupo de decisores, extrae del mismo unas conclusiones que permitan tomar la decisión o

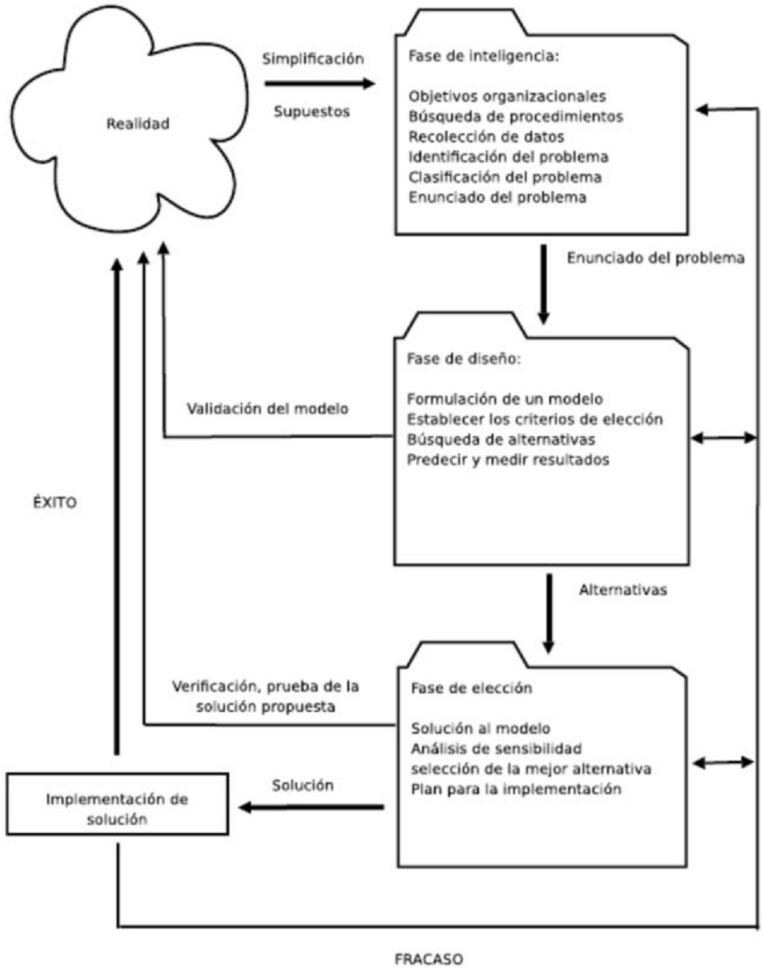


Figura 2.4: Fases del proceso de toma de decisión [119].

decisiones correspondientes. Es el encargado de explicar, justificar, recomendar y modelar el problema y de hacer recomendaciones a los decisores, pero en ningún caso se encarga de ejecutar la decisión final o de expresar opiniones personales.

El cliente, cumple una función entre el *Decision maker* y el analista, de modo que solicita el estudio y es el responsable para asignar o distribuir el significado necesario para conducirla. Este rol aparece cuando el analista y el *Decision maker* son personas diferentes. A veces el cliente puede adoptar el rol del analista.

Los métodos de decisión multicriterio se pueden aplicar a diversos sectores como la salud, la educación, el sector empresarial, entre otros. Son una base sustentada en

elementos científicos, que aporta mejoras distintivas para asumir una decisión como lo han estudiado los investigadores Hammond, Keeney y Raiffa [110].

2.4 Teoría de la Decisión multicriterio

Los problemas de decisión de acuerdo a Arratia [18], se clasifican en:

1. Según la naturaleza de las consecuencias asociadas a las alternativas.
   - Decisión bajo certeza
   - Decisión bajo incertidumbre
   - Decisión bajo estricta incertidumbre
2. Según la cardinalidad del conjunto de alternativas.
   - Discreto (finito, no muy grande).
   - Continuo (infinito o muy grande).
3. Según las operaciones que se realizan sobre el conjunto de alternativas [95].
   - Selección ($P.\alpha$): esta clase de problemas se presenta cuando el tomador de decisiones escoge las mejores alternativas de un conjunto en base a sus preferencias.
   - Jerarquización ($P.\gamma$): esta clase de problemas se presenta cuando el tomador de decisiones ordena, según una relación de preferencia, las alternativas a través de una jerarquización.
   - Clasificación ($P.\beta$).
4. Según la cantidad de tomadores de decisiones:
   - Una sola persona: las decisiones recaen en una sola persona.
   - Un grupo de personas: las decisiones son tomadas por dos o más personas.

Dentro de los problemas de decisión bajo certeza se utiliza frecuentemente una relación denominada de sobreclasificación para representar las preferencias del tomador de decisiones. Esta relación puede interpretarse de la siguiente manera: sean a y b dos alternativas del problema de decisión en cuestión, si a sobreclasifica a b *(aSb)* puede interpretarse como que existen argumentos claros o razones suficientes para considerar que a es al menos tan buena como b.

Es típico que las preferencias del tomador de decisiones sean descritas por un sistema de relaciones compuesto por las relaciones P, Q, I y R donde [95]:

- P denota preferencia estricta entre dos alternativas (el tomador de decisiones elige sin dudar una de las dos alternativas que le son presentadas).
- Q representa la preferencia débil (el tomador de decisiones elige una de las dos alternativas, pero tiene algunas dudas sobre la preferencia).
- I denota una relación de indiferencia (el tomador de decisiones elegiría ambas alternativas sin mostrar preferencia a favor de una o de la otra).
- R representa la incomparabilidad (el tomador de decisiones rehúsa elegir cualquiera de las dos alternativas considerando que no puede emitir preferencia entre ellas).

2.5 Metodología de ciencia de datos

Ciencia de datos (CD): es el área de estudio que comprende desde el desarrollo de software, la inteligencia artificial, el manejo de los datos y la estadística [34]. La ciencia de los datos involucra principios, procesos y técnicas para entender un fenómeno vía el análisis de los datos [90]. Los proyectos de ciencia de datos se enfocan en descubrir nuevos conocimientos de los datos, identificar correlaciones, relaciones causales, clasificar y predecir eventos, identificar patrones y anomalías e inferir probabilidades y para ello se relacionan áreas relevantes del conocimiento, entre ellas:

Cálculo, Álgebra, Teoría de Probabilidades y Estadística, Complejidad y Diseño de Algoritmos y Desarrollo de software. La ciencia de los datos cubre el proceso de la recopilación de los datos, su análisis y los resultados de los experimentos.

2.6 Etapas de la metodología de ciencia de datos propuesta por IBM

En la figura 2.5 se presentan las fases de la metodología propuesta IBM (International business machines) para la aplicación de la ciencia de datos a nivel industrial y científico [74].

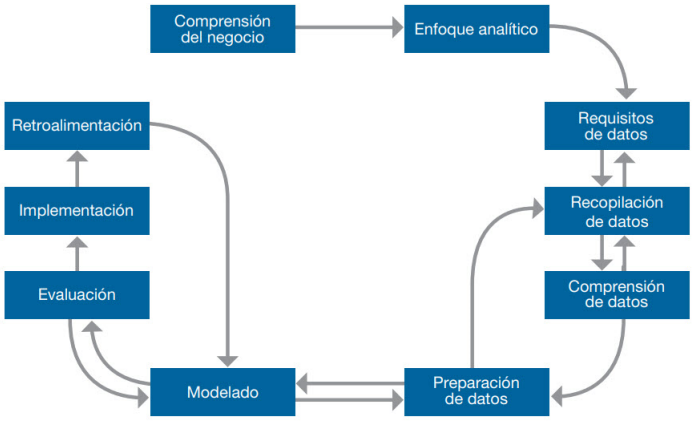


Figura 2.5: Metodología fundamental para la Ciencia de Datos [93]

Esta metodología consiste en 10 etapas de modo general, como se puede apreciar en la imagen anterior [93].

Etapa 1 Comprensión del proyecto: en esta etapa se definen tanto los problemas y los objetivos del proyecto como también los requisitos de la solución desde un punto de vista comercial. Por eso, esta etapa se considera como la más difícil.

Etapa 2 Acercamiento analítico: en esta etapa, el científico de datos define el enfoque analítico que utilizará para resolver el problema. Esto implica saber expresar

el problema en el contexto de técnicas estadísticas y de aprendizaje automático para que pueda identificar los mecanismos adecuados para lograr un resultado exitoso.

Etapa 3 Requerimiento de datos: la elección de la aproximación analítica se determina a partir de los requisitos de los datos para poder conocer el método analítico que se utilizará. Las características de los datos son particulares en diversos formatos y representaciones; pueden ser estructurados, semi-estructurados y no estructurados.

Etapa 4 Recolección de datos. En esta etapa, el científico de datos identifica y reúne datos estructurados, semi-estructurados o no estructurados, relevantes en el dominio del problema.

Etapa 5 Comprensión de los datos. En esta etapa, la estadística descriptiva y las técnicas de visualización pueden ayudar a entender al científico de datos el contenido de los datos, evaluar su calidad y descubrir una idea inicial de estos.

Etapa 6 Preparación de los datos. Esta etapa comprende actividades enfocadas en la preparación del conjunto de datos que serán usados en la etapa de modelado. Esto incluye la limpieza de los datos, la combinación de los datos de múltiples fuentes y la transformación de los datos.

Etapa 7 Modelamiento. En esta etapa, se generan modelos predictivos mediante algoritmos de aprendizaje automático.

Etapa 8 Evaluación. En esta etapa, el científico de datos evalúa la calidad de los modelos generados.

Etapa 9 Despliegue. Después de haber elegido un modelo satisfactorio que es aprobado por los patrocinadores del proyecto, el modelo se implementa en el ambiente de producción o de prueba para su utilización.

Etapa 10 Retroalimentación. Mediante la recopilación de los resultados del modelo implementado, la organización recibe información sobre el rendimiento del modelo y se observa la forma en que este afecta su entorno. El análisis de esta infor-

mación le muestra al científico de datos si se debe refinar el modelo. Esto aumenta su precisión y, por tanto, su utilidad.

2.7 Aprendizaje automatizado

Aprendizaje automatizado (*Machine Learning*): actualmente se considera una disciplina distinta de la informática, y es en efecto un programa o sistema que puede aprender tareas específicas como la discriminación o clasificación sin ser programado explícitamente para hacerlo [3], es decir se emplean los algoritmos y heurísticas sobre datos tabulados convirtiéndolos en programas de computadora, sin necesidad de escribir los últimos de forma explícita. Cuando se obtienen modelos, estos deben ser capaces de generalizar comportamientos e inferencias para una muestra mucho más amplia de datos. Dentro del aprendizaje automatizado están relacionadas otras disciplinas, como la estadística inferencial, la programación en computadora, entre otras.

Esta disciplina ha progresado de forma dramática en las últimas décadas con diversas aplicaciones en áreas como el reconocimiento del habla, procesamiento natural del lenguaje, control de robot y otras aplicaciones [60], en el análisis y diagnósticos médicos [63], en el análisis y predicción del mercado de valores con sistemas híbridos basados en algoritmos genéticos [31], en la clasificación de secuencias de ADN usando SVM para predecir proteínas de unión ARNr, ARN y ADN a partir de la secuencia de aminoácidos [118], en los motores de búsqueda en dominios específicos [10], en la detección de fraude con el empleo de tarjetas de crédito [23], entre otras aplicaciones. Dentro de los algoritmos de aprendizaje automatizado, existen dos grupos importantes de algoritmos: Aprendizaje Supervisado (*Supervised machine learning*) y Aprendizaje No supervisado (*Unsupervised machine learning*).

A continuación, se presenta en la figura 2.6, la relación del aprendizaje automatizado con la Ciencia de los Datos (*Data Science*), la Inteligencia Artificial (*Artificial*

Intelligence) y la Ciencia de la computación (*Computer Science*).

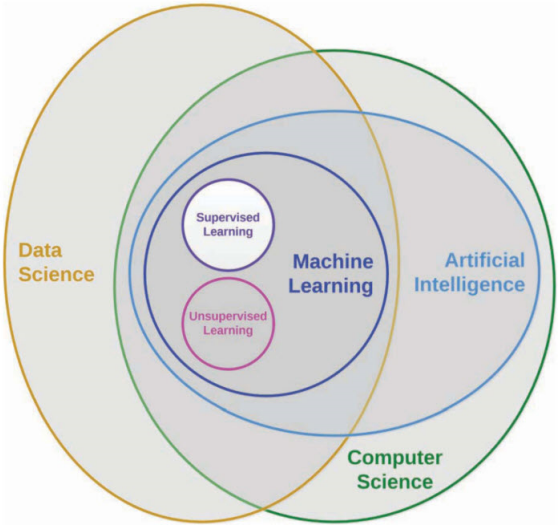


Figura 2.6: El aprendizaje automatizado relacionado con otras disciplinas [3].

2.7.1 Métodos de Aprendizaje supervisado

Aprendizaje supervisado: en estos algoritmos se genera un modelo predictivo, basado en datos de entrada y salida. En este caso se tiene un conjunto de datos etiquetados y clasificados, es decir tener un conjunto de muestra que ya se sabe a qué grupo, valor, o categoría pertenecen los ejemplos. Con este grupo de datos, que son llamados "datos de entrenamiento", se realiza el ajuste al modelo inicial planteado. De esta forma el algoritmo va aprendiendo a clasificar las muestras de entrada comparando el resultado del modelo, y la etiqueta real de la muestra, realizando compensaciones respectivas al modelo de acuerdo a cada error en la estimación del resultado [64].

Máquina de soporte vectorial (SVM): es un conjunto de métodos de aprendizaje supervisado relacionados utilizados para la clasificación y regresión. Un modelo SVM construye un hiperplano o un conjunto de hiperplanos en un espacio de alta dimensión, llamado espacio de características, que puede usarse para clasificación o

regresión [39]. Los objetos de diferentes clases están separados por el hiperplano con el margen más grande. En 2D, dos clases de objetos están separadas por la línea recta cuando la distancia desde el punto a la línea es máxima [96].

2.7.2 Regresión logística (RL)

El algoritmo de regresión logística (*Logistic regression*), es un método estadístico con el objetivo de predecir clases binarias. El resultado o la variable objetivo es de naturaleza binaria. Este es un método atractivo para predecir la probabilidad de ocurrencia de un evento porque está matemáticamente restringido a producir probabilidades en el rango [0, 1] [111]. Este algoritmo formula un análisis utilizado para predecir una variable categórica en función de una o varias variables predictoras [54]. La regresión logística predice la probabilidad de ocurrencia de un evento binario utilizando una función logit.

En este algoritmo los datos se distribuyen binomialmente siguiendo la forma:

$$Y_i \sim B(p_i, n_i), \qquad para \qquad i = 1, ..., m \tag{2.1}$$

En la expresión anterior:

- n_i son los números de ensayos Bernoulli conocidos
- p_i son las probabilidades de éxito desconocidas

Estos valores de probabilidades binomiales desconocidas (logit) conforman el modelo general del agoritmo de regresión logística, que tiene como expresión:

$$logit(p) = log\left(\frac{p}{1-p}\right) = x_i^T\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p = \eta \tag{2.2}$$

La regresión logística se emplea para correlacionar la probabilidad de ocurrencia de una variable cualitativa binaria con un conjunto de variables escalares. La probabilidad aproximada de pertenencia a cualesquiera de las dos categorías en el suceso se aproxima a través de una función logística dada por la expresión:

$$p_i = logit^{-1}(\eta) = \frac{e^n}{1+e^n} = \frac{1}{e^{-n}+1} \tag{2.3}$$

Propiedades de la regresión logística:

- La variable dependiente en la regresión logística sigue la distribución de Bernoulli.
- La estimación se realiza a través de la máxima probabilidad.

Suposiciones de regresión logística:

- La regresión logística binaria requiere que la variable dependiente sea binaria.
- Para una regresión binaria, el nivel de factor 1 de la variable dependiente debe representar el resultado deseado.
- Solo se deben incluir variables significativas.
- Las variables independientes deben ser independientes entre sí. Es decir, el modelo debe tener poca o ninguna multicolinealidad.
- Las variables independientes están relacionadas linealmente con las probabilidades de registro.
- La regresión logística requiere tamaños de muestra bastante grandes.

Como aspecto importante de este método de RL, es que utiliza una función logística que es usada para hallar la probabilidad de pertenencia de uno u otro grupo, sin embargo esta función es óptima cuando los datos son simétricos, es decir cuando la

variable dependiente categórica tiene una cantidad equilibrada de “ceros” y “unos”, aspecto que está presente en esta investigación, reflejado en que por encima del 95 % de los registros de cada dataset pertenecen a una clase con PayFlag en “cero”, y apenas el 4 % del total pertenecen a la clase con PayFlag en “uno”.

Ventajas:

- Es una técnica ampliamente utilizada porque es muy eficiente y no requiere demasiados recursos computacionales.
- Es altamente interpretable.
- Esta técnica no requiere funciones de entrada para escalar y no requiere ningún ajuste.
- Es fácil de regularizar y produce probabilidades pronosticadas bien calibradas.
- La regresión logística funciona mejor cuando elimina los atributos que no están relacionados con la variable de salida, así como los atributos que son muy similares (correlacionados) entre sí.

Desventajas:

- La regresión logística no es uno de los algoritmos más poderosos que existen y puede ser superado por otros algoritmos más complejos.
- No se pueden resolver problemas no lineales con regresión logística ya que su superficie de decisión es lineal.

2.7.3 Métodos de Aprendizaje No supervisado

Aprendizaje no supervisado: estos algoritmos trabajan de forma similar a los del tipo supervisado, con la diferencia de que estos solo ajustan su modelo predictivo

tomando en cuenta los datos de entrada, sin importar la salida. Es decir, a diferencia del supervisado, los datos de entrada no están clasificados ni etiquetados, y no son necesarias estas características para entrenar el modelo [64]. En otras palabras, en estos algoritmos solo hay datos en bruto sin ninguna cosa en particular que deba predecirse. Estos algoritmos son usados para descubrir patrones en datos generales. Los algoritmos de agrupamiento (*clustering*), que intentan dividir un conjunto de datos en grupos "naturales"son un ejemplo de prototipo de algoritmo no supervisado, ejemplo de esto son el algoritmo *K-Means* y *DBSCAN*. Se hará una referencia a estos algoritmos a continuación.

La agrupación (*Clustering*), es una división de datos en grupos de objetos similares. Cada grupo, llamado clúster, consta de objetos que son similares entre sí y diferentes a los objetos de otros grupos. La representación de datos por menos clústeres necesariamente pierde ciertos detalles finos, pero logra la simplificación. Representa muchos objetos de datos por pocos grupos y, por lo tanto, modela datos por sus grupos. El modelado de datos coloca el agrupamiento en una perspectiva histórica arraigada en matemáticas, estadísticas y análisis numérico. Desde una perspectiva de aprendizaje automático, los clústeres corresponden a patrones ocultos, la búsqueda de clústeres es un aprendizaje no supervisado y el sistema resultante representa un concepto de datos. Por lo tanto, la agrupación es el aprendizaje no supervisado de un concepto de datos ocultos. La minería de datos trata con grandes bases de datos que imponen al análisis de agrupamiento requisitos informáticos severos adicionales [24].

Clustering: la técnica de clustering o agrupamiento es el proceso que consiste en la división de los datos en grupos de objetos similares. Para medir la similaridad entre objetos se suelen utilizar diferentes formas de distancia: distancia euclidiana, de Manhattan, de Mahalanobis, etc. El representar los datos por una serie de clusters, conlleva a una pérdida de detalles, pero consigue la simplificación de los mismos [46].

2.7.4 DBSCAN

El algoritmo *DBSCAN* es una técnica de agrupamiento (*clustering*) que pertenece al grupo de algoritmos no supervisados [36]. Este algoritmo está basado en la densidad, y analiza los clústeres como áreas de alta densidad separadas por áreas de baja densidad. Debido a esta visión bastante genérica, los grupos encontrados por *DBSCAN* pueden tener cualquier forma, a diferencia del algoritmo *K-Means*, que supone que los grupos tienen forma convexa. El componente central de *DBSCAN* es el concepto de muestras de núcleo, que son muestras que se encuentran en áreas de alta densidad. Por lo tanto, un grupo es un conjunto de muestras de núcleo, cada una cerca de la otra (calculada por alguna medida de distancia) y un conjunto de muestras de núcleo que están cerca de una muestra de núcleo (pero no son muestras de núcleo). Hay dos parámetros para el algoritmo, *mindsamples* y *eps*, que definen formalmente el significado de denso. Mayor *mindsamples* o menor *eps* que indica mayor densidad necesaria para formar un grupo.

Ventajas del algoritmo:

- Este algoritmo utiliza la distancia especificada para separar los clústeres densos del ruido más disperso.
- Es el método de clustering más rápido.
- Funciona bien con todos los clústers potenciales, para ello se requiere que todos los clústers significativos presenten densidades similares.
- Aplicado para instancias con un gran número de muestras.

Desventajas:

- Solo es apropiado si se puede utilizar la distancia de búsqueda muy clara.

2.7.5 *K-Means*

El algoritmo *K-Means* es una técnica de agrupamiento que pertenece al grupo de algoritmos no supervisados [36]. Esta técnica, agrupa los datos a modo de separar las muestras en n grupos de igual varianza, minimizando un criterio conocido como la inercia o la suma de cuadrados dentro del grupo. Este algoritmo, requiere que se especifique el número de clústeres. Además, se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes.

La agrupación de *K-Means* se utiliza para generar grupos del conjunto de datos de entrada antes de utilizar redes funcionales para realizar la predicción de las variables objetivo reales [82].

El algoritmo *K-Means*, divide un conjunto de N muestras en X dentro de K racimos disjuntos C, cada uno descrito por la media de las muestras en el grupo. Los medios se denominan comúnmente el grupo *centroides*, luego el algoritmo *K-Means* tiene como objetivo elegir centroides que minimicen la inercia o el criterio de suma de cuadrados dentro del clúster.

El algoritmo inicialmente, elige los centroides iniciales, y el método más básico es elegir k muestras del conjunto de datos X. Después de la inicialización, *K-Means* asigna cada muestra a su centroide más cercano, como paso siguiente crea centroides más cercanos al tomar el valor medio de todas las muestras asignadas a cada *centroide* anterior. Se calcula la diferencia entre el centroide antiguo y el nuevo y el algoritmo repite estos dos últimos pasos hasta que este valor sea inferior a un umbral. En otras palabras, se repite hasta que los centroides no se mueven significativamente, de esta manera quedan agrupados los conjuntos según la cantidad de clusters que se haya introducido.

Este algoritmo se detiene cuando ha logrado agrupar en la cantidad de clúster que se han determinado a todos los elementos [65].

El algoritmo k-means es un algoritmo para asignar un número específico de centros, k, para representar la agrupación de N puntos ($k \leq N$). Estos puntos se ajustan iterativamente para que cada punto se asigne a un cluster, y el centroide de cada cluster es la media de sus puntos asignados. En general, la técnica k-means producirá exactamente k diferentes grupos de la mayor distinción posible [75].

Ventajas del algoritmo:

- Se adapta bien a un gran número de muestras.
- Este algoritmo admite pesos de muestra.
- Se puede introducir un parámetro específico a este algoritmo para que se ejecute en paralelo, llamado *njobs*.

El algoritmo Kmeans es un algoritmo iterativo que intenta dividir el conjunto de datos en K subgrupos distintos no superpuestos (grupos) predefinidos donde cada punto de datos pertenece a un solo grupo . Intenta hacer que los puntos de datos dentro del clúster sean lo más similares posible, al tiempo que mantiene los clústeres lo más diferentes (lo más lejos posible). Asigna puntos de datos a un grupo de modo que la suma de la distancia al cuadrado entre los puntos de datos y el centroide del grupo (media aritmética de todos los puntos de datos que pertenecen a ese grupo) es mínima. Cuanta menos variación tengamos dentro de los grupos, más homogéneos (similares) serán los puntos de datos dentro del mismo grupo.

La forma en que funciona el algoritmo kmeans es la siguiente:

1. Especificar número de agrupaciones K .
2. Inicializar los centroides primero barajando el conjunto de datos y luego seleccionando aleatoriamente K puntos de datos para los centroides sin reemplazo.
3. Sigue iterando hasta que no haya cambios en los centroides, es decir, la asignación de puntos de datos a grupos no está cambiando.

- El algoritmo calcula la suma de la distancia al cuadrado entre los puntos de datos y todos los centroides.
- Asignar cada punto de datos al grupo más cercano (centroide).
- Calcular los centroides para los grupos tomando el promedio de todos los puntos de datos que pertenecen a cada grupo.

El enfoque que kmeans sigue para resolver el problema se llama Expectativa-Maximización.El E-step es asignar los puntos de datos al clúster más cercano. El paso M está calculando el centroide de cada grupo.

La expresión matemática mediante la que queda expresada este algoritmo está dada por:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \mid\mid x_i - \mu_k \mid\mid^2 \tag{2.4}$$

donde $w_{ik} = 1$ para el punto de datos si x_i pertenece al clúster k; de lo contrario, $w_{ik} = 0$. μ_k es el centroide del grupo de x_i.

2.7.6 Otros Métodos de Aprendizaje

Aprendizaje por refuerzo (*Reinforcement learning*): estos algoritmos definen modelos y funciones enfocadas en maximizar una medida de "recompensas", basadas en "acciones", y al ambiente en el que el agente inteligente se desempeñará. Este algoritmo es el más apegado a la psicología conductista de los humanos, ya que es un modelo de acción-recompensa, que busca que el algoritmo se ajuste a la mejor "recompensa" dada por el ambiente y sus acciones por tomar.

2.8 Método aprendizaje de conjuntos

El método de aprendizaje de conjuntos (*Method Ensemble Learning*), es un algoritmo de aprendizaje que construye un conjunto de clasificadores básicos y luego clasifica nuevos ejemplos tomando un voto de sus predicciones y constituye una de las principales direcciones actuales en la comunidad de aprendizaje automatizado. El marco básico del aprendizaje de conjunto se ilustra en la Fig 2.8. En el conjunto, el clasificador 1 a través de N clasificadores se entrena primero usando los ejemplos de entrenamiento. Luego, para cada ejemplo, la salida pronosticada Oi de cada uno de estos clasificadores se combina para producir la salida O del conjunto. El aprendizaje de conjuntos utiliza modelos de clasificación como el algoritmo *Bagging* y el algoritmo *Boosting* que son dos de los algoritmos de conjunto más populares [66] para mejorar la predicción.

De acuerdo con el trabajo de Anifowose de 2015 [13], el método de aprendizaje de conjuntos tiene un desempeño superior al método convencional de aprendizaje de técnicas individuales cuando se aplica a problemas de clasificación y regresión. Este método es una técnica avanzada en la tecnología de aprendizaje supervisado. El paradigma que aplica este método de aprendizaje de conjuntos es una imitación del aprendizaje social humano [40], y se apoya sobre una metodología que es ideal para manejar ambos casos de demasiada información y muy poca información [13]. No obstante, uno de los problemas del método de aprendizaje de conjuntos *Ensembles* es que la interpretación es mucho más difícil, así como la preocupación de que su mayor complejidad conduzca a un sobreajuste, es decir, inexactitud en los nuevos datos. [101].

A continuación, la siguiente imagen ilustra de forma esquemática el método de aprendizaje de conjuntos.

Algoritmo *Bagging*

Uno de los métodos de conjunto o *Ensemble* para tareas de regresión es el *Bootstrap*

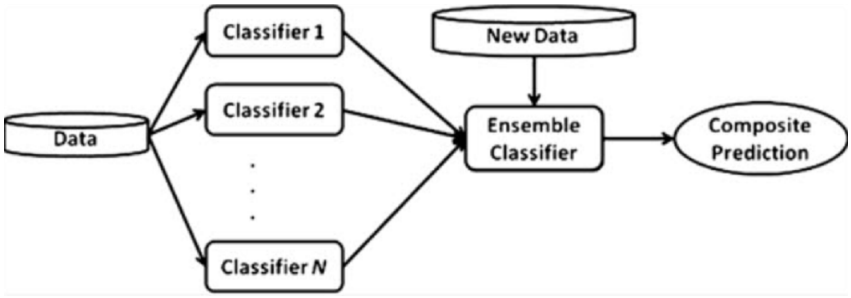


Figura 2.7: Vista esquemática del método ensemble [28].

Aggregate o *Bagging* [12]. El método *Bagging* para problemas de predicción funciona dando a la contribución de cada aprendiz base en el modelo de conjunto un peso igual. Para mejorar la varianza del modelo, el *Bagging* entrena a cada modelo en el *Ensemble* usando un subconjunto que se extrajo al azar del conjunto de entrenamiento con reemplazo [38]. Los resultados de los aprendices base se promedian sobre todos los aprendices base para obtener el resultado general del modelo de conjunto o *Ensemble*.

El proceso de construcción de cada clasificador base o aprendiz base es independiente el uno del otro. La precisión del algoritmo *bagging* mejorará enormemente si *bootstrap* puede inducir diferencias significativas en los clasificadores base construidos [66]. El método de conjunto convencional para tareas de regresión es el *Bagging* [13].

A continuación, en la siguiente figura se representa de forma esquemática el método *bagging* aplicando árboles de decisión.

En la imagen anterior se muestra un Ranodm Forest basado en el método *Bagging* (*Bootstrap + Aggregating*). En la etapa 1 se utiliza un método *Bootstrap* en M subconjuntos del dataset original de entrenamiento. En la etapa 2 se construyen M árboles de decisión independientes para el entrenamiento del modelo usando covariables de entrada. Para cada árbol de decisión se tiene la predicción de confianza. En la etapa 3 se obtener predicciones de cada árbol *boostrap* sobre M réplicas. En la etapa 4 se decide el resultado final por el promedio o mayoría de los votos.

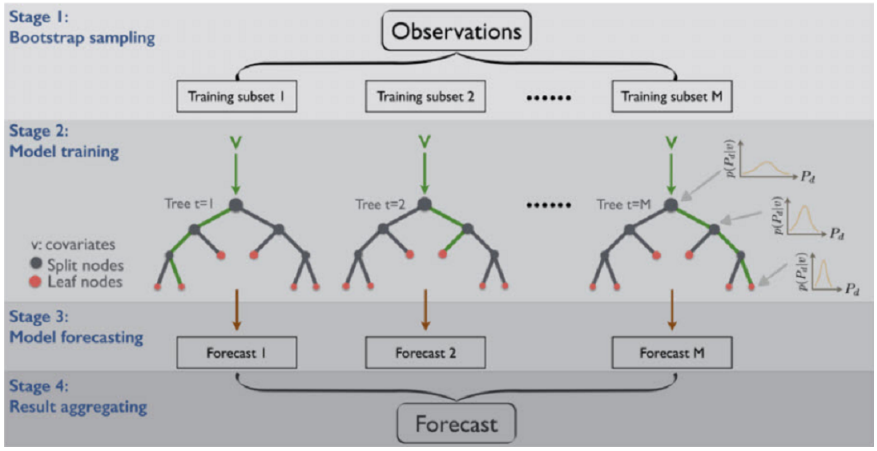


Figura 2.8: Vista esquemática del funcionamiento del método de aprendizaje *bagging* [51].

Algoritmo *Boosting*

Este algoritmo fue introducido por Y. Freund y R. Shapire en 1990 [42] y se implementó en la técnica Adaptive Boosting (*Adaboost*) para problemas de clasificación y agrupamiento [13].

La técnica de *AdaBoost* [42] fue el primer enfoque aplicable de *Boosting* y ha sido designado como uno de los diez mejores algoritmos de minería de datos [115].

Esta técnica, utiliza todo el conjunto de datos para entrenar cada clasificador en serie, pero después de cada ronda, se centra más en instancias difíciles, con el objetivo de clasificar correctamente los ejemplos en la próxima iteración que se clasificaron incorrectamente durante la iteración actual. Por lo tanto, le da más atención a los ejemplos que son más difíciles de clasificar, la cantidad de atención se mide por un peso, que inicialmente es igual para todas las instancias. Después de cada iteración, aumentan los pesos de las instancias mal clasificadas; por el contrario, los pesos de las instancias clasificadas correctamente disminuyen. Además, se asigna otro peso a cada clasificador individual en función de su precisión general que luego se utiliza en la fase de prueba; se da más confianza a los clasificadores más precisos. Finalmente, cuando se presenta una nueva instancia, cada clasificador otorga un voto ponderado

y la etiqueta de clase se selecciona por mayoría [44].

Sirve para mejorar el rendimiento de un algoritmo de aprendizaje débil. Hay muchas variedades de algoritmos de refuerzo y *AdaBoost*. En el trabajo de Shi et al. M1 se elige como el método de refuerzo. El método *Boosting* adaptativamente vuelve a ponderar el conjunto de entrenamiento de una manera basada en una tasa de error del anterior clasificador base. Inicialmente, los pesos son uniformes para todas las muestras de entrenamiento. Durante el procedimiento de refuerzo, el algoritmo de refuerzo mejora su comportamiento en reflejo de los últimos fallos que comete y la secuencia de la construcción de clasificadores base se detiene si la tasa de error de un clasificador base es mayor que 0.5 o igual a 0 [66].

2.9 Técnica de Escalarización utilizada

En el trabajo con los datos en el campo de la ciencia de los datos, en muchas ocasiones los datos no están todos en la misma escala, lo que puede afectar la precisión de los algoritmos que sean aplicados, ya sea k-means, regresión logística, redes neuronales, entre otros, por lo que se hace necesario aplicar métodos de escalarización a los datos. El escalado de los datos es el proceso de aumentar o disminuir la magnitud de acuerdo con una relación fija, en otras palabras, es cambiar el tamaño pero no la forma de los datos [105]. Es importante realizar la normalización del conjunto de datos utilizando técnicas de escalado de características para obtener una buena precisión en los resultados [105].

Antes de aplicar los algoritmos, previamente los datos han sido escalados (*Feature Scaling*) usando la librería *Standard Scaler* [88]. Este método transforma el conjunto de datos de modo que el valor medio de la distribución resultante sea cero y la desviación estándar sea uno. El valor transformado se obtiene restando el valor medio del valor original y dividiendo por la desviación estándar. La expresión dada a continuación se usa para la transformación [105].

$$Z = \frac{X - \mu}{\sigma} \tag{2.5}$$

donde:

X es el valor original

μ es la media

σ es la desviación standard

2.10 Selección de la métrica de distancia

Con el objetivo de contar con una métrica que indique la similitud entre objetos se escogió la métrica Euclidiana.

La métrica Euclidiana está definida por la siguiente expresión, como lo refleja el texto Unsupervised Classification [21]:

$$d_E(P,Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{2.6}$$

donde d_E es la distancia entre dos puntos de coordenadas $(x_1, y_1), (x_2, y_2)$

De forma genérica la expresión de la distancia euclidiana está determinada por:

$$d_{(i,j)} = \sqrt{\sum_{k=1}^{n} \left(x_{(ik)} - x_{(jk)}\right)^2} \tag{2.7}$$

2.11 Métricas utilizadas para evaluar el desempeño de los agrupamientos

Para evaluar los resultados de los algoritmos de agrupamiento DBSCAN y K-Means se consideraron 3 métricas, el Coeficiente de Silhouette, el coeficiente Ho-

mogeneity y el coeficiente Completeness (Integridad).

Coeficiente de Silhouette

El Coeficiente de Silhouette o coeficiente de silueta, como también se le llama en varios textos, es del tipo de métricas internas, que se basa únicamente en la información de los datos. Evalúa qué tan buena es la estructura del clustering sin necesidad de información ajena al propio algoritmo y su resultado, es decir cómo de bueno o de homogéneo es cada clúster formado.

Para evaluar conjuntos de clústeres de forma independiente, se utiliza el coeficiente de silueta. El coeficiente de silueta es una métrica de verificación de conglomerados ampliamente utilizada que determina qué tan similar es un punto a su propio conglomerado en comparación con otros con un valor entre uno negativo y uno. Un coeficiente de silueta de cero generalmente representa grupos potencialmente superpuestos, un valor positivo indica que una muestra está mejor representada en su propio grupo que cualquier otro, y un valor negativo representa lo contrario. Se toma una muestra s, se deja que la distancia entre s y un grupo C sea el promedio de distancias entre s y todos los puntos $p \in C$ [37].

El Coeficiente de Silhouette, está definido por la expresión:

$$SC(s) = \frac{B(s) - A(s)}{max(A(s), B(s))} \tag{2.8}$$

Donde:
A(s) es la distancia media dentro del grupo de todos los puntos a s
B(s) es la distancia entre s y el grupo no asociado más cercano.

El valor devuelto por A(s) también se puede interpretar como cuánto pertenece una muestra a su grupo (valores más pequeños significa mejor ajuste). A diferencia de las otras métricas de verificación, el coeficiente de silueta no requiere un conjunto

de clústeres de referencia y, en cambio, solo depende de los valores de distancia entre las propias muestras. Definimos Silhouette Score como el coeficiente de silueta promedio de todas las muestras, esto proporciona una medida general de calidad [37].

El valor de SC puede variar entre -1 y 1, siendo:

- 1 : buen agrupamiento o que el objeto s ha sido bien clasificado (en A)
- 0 : indiferente o que el objeto s se encuentra intermedio entre dos grupos (A y B)
- -1 : mal agrupamiento o que el objeto s ha sido mal clasificado (más cerca de B que de A)

Tabla de referencia del coeficiente Silhouette [104].

Tabla 2.1: Tabla de referencia del coeficiente Silhouette.

Rango de valores	**Interpretación**
0.71 -1.00	Se ha encontrado estructura fuerte
0.51 -0.70	La estructura es razonable
0.26 -0.51	La estructura es débil y podría ser artificial
≤ 0.25	No se ha encontrado una estructura sustancial

Homogeneity (h): en el sentido del criterio de homogeneidad el agrupamiento debe asignar solo aquellos puntos de datos que son miembros de una sola clase a un solo grupo, es decir, la distribución de la clase dentro de cada grupo debe adaptarse a una sola clase, es decir, entropía cero. Se determina qué tan cerca está una agrupación dada de este ideal examinando la entropía condicional de la distribución de clases dada la agrupación propuesta. En el caso perfectamente homogéneo, este valor, H(C | K), es 0.

En el trabajo de Rosenberg et al. del 2007 [94], se plantea la expresión que define a esta métrica, dada por:

$$h = \begin{cases} 1 & if \quad H(C,K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & else \end{cases} \tag{2.9}$$

donde

$$H(C|K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|}\frac{a_{ck}}{N} log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}} \tag{2.10}$$

$$H(C) = -\sum_{c=1}^{|C|}\frac{\sum_{k=1}^{|K|} a_{ck}}{n} log \frac{\sum_{k=1}^{|K|} a_{ck}}{n} \tag{2.11}$$

Completeness (Integridad)

La integridad es simétrica a la homogeneidad. Para satisfacer los criterios de integridad, una agrupación debe asignar todos los puntos de datos que son miembros de una sola clase a un solo grupo. Para evaluar la integridad, se examina la distribución de las asignaciones de grupos dentro de cada clase. En una solución de agrupación perfectamente completa, cada una de estas distribuciones estará completamente sesgada en un solo grupo. Podemos evaluar este grado de inclinación calculando la entropía condicional de la distribución de grupo propuesta dada la clase de datapoints componentes, H(K | C), como queda reflejado en el trabajo de Rosenberg et al. del 2007 [94]. La expresión que define a esta métrica, es:

$$c = \begin{cases} 1 & if \quad H(K,C) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & else \end{cases} \tag{2.12}$$

donde

$$H(K|C) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{a_{ck}}{N} log \frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}} \tag{2.13}$$

$$H(K) = -\sum_{k=1}^{|K|} \frac{\sum_{c=1}^{|C|} a_{ck}}{n} log \frac{\sum_{c=1}^{|C|} a_{ck}}{n} \tag{2.14}$$

2.12 Métricas utilizadas para evaluar a los clasificadores entrenados

Exactitud (*Accuracy*): es una medida de la exactitud de un clasificador, es decir es una medida de la corrección lograda en la predicción positiva, esto no es más que, de las observaciones predecidas como positivas, cuando son realmente positivas. Es utilizada para evaluar la capacidad de generalización de los clasificadores. A través de la exactitud, el clasificador entrenado se mide en base a la corrección total que se refiere al total de instancias que el clasificador entrenado predice correctamente cuando se prueba con los datos invisibles [55], se enfoca en la razón de prediciones correctas sobre el número total de instancias evaluadas. Está dada por la expresión [55]:

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{2.15}$$

Precisión (p): es utilizada para medir los patrones positivos que se predicen correctamente del total de patrones predichos en una clase positiva. Está dada por la expresión [55]:

$$p = \frac{tp}{tp + fp} \tag{2.16}$$

Recall (r): (sensibilidad): es una medida de la integridad de un clasificador, mide las observaciones reales que se etiquetan (predice) correctamente, es decir, cuántas observaciones de clase positiva se etiquetan correctamente, se usa para medir

la fracción de patrones positivos que se clasifican correctamente. Está dada por la expresión [55]:

$$r = \frac{tp}{tp + tn} \tag{2.17}$$

F-measure (medida F) (FM): Es un promedio ponderado entre la precisión y el Recall. Esta métrica representa la media armónica entre los valores de recuperación y precisión. Está dada por la expresión [55] :

$$FM = \frac{2 * p * r}{p + r} \tag{2.18}$$

Curva ROC:

De entre las técnicas más adecuadas para evaluar a los clasificadores en problemas de clases desbalanceadas se encuentra la curva ROC (Receiver Operating Characteristic) que es una herramienta para visualizar, organizar, y seleccionar clasificadores basados en su intercambio entre beneficios (TP) y costo (FP) [45]. Esta curva resume todas las matrices de confusión de los umbrales producidos.

La curva ROC (Receiver Operator Characteristic) es una medida de rendimiento para problemas de clasificación en varios ajustes de umbrales. ROC es una curva de probabilidad y AUC representa el grado o medida de separabilidad. Indica cuánto el modelo es capaz de distinguir entre clases. Cuanto mayor sea el AUC, mejor será el modelo para predecir 0s como 0s y 1s como 1s.

En este tipo de gráfico el eje “y”, muestra la Tasa positiva verdadera. La Tasa positiva verdadera es la tasa verdadera de positivos que se calcula mediante la expresión:

$$TP_{rate} = \frac{TP}{TP + FN} \tag{2.19}$$

Tasa de verdaderos positivos (TP_{rate})

Verdaderos positivos (TP)

Falsos Negativos (FN)

Esta tasa aporta información sobre que proporción de los registros fueron correctamente clasificados, el eje "x"muestra la Tasa de falsos positivos.

$$TN_{rate} = \frac{TN}{TN + FP} \tag{2.20}$$

Tasa de negativos positivos (TN_{rate})

Verdaderos Negativos (TN)

Falsos positivos (FP)

Estos falsos positivos, son aquellas muestras que fueron clasificadas de manera incorrecta. Cuando se establece un límite ajustado al clasificar a los casos positivos verdaderos y el valor obtenido de la métrica ROC es igual a 1, significa que el clasificador ha logrado clasificar a todas las muestras positivas de manera correcta, de forma análoga sucede con los casos falsos negativos.

AUC (Area Under the Curve): establece una comparación entre varios ROC, y permite comparar de este modo varios clasificadores, siendo aquel modelo de mayor valor el más indicado para clasificar.

2.13 Técnicas utilizadas para balancear los datos

Para afrontar el problema de las clases desbalanceadas existen varios métodos de remuestreo, según Burnaev del 2015 [29]:

- técnica SMOTE
- técnica de sobremuestreo (*Over-sampling technique*)
- técnica de submuestreo (*Random Under-sampling technique*)

Los métodos de remuestreo están diseñados para agregar o eliminar ejemplos del conjunto de datos de entrenamiento para cambiar la distribución de la clase.

Una vez que las distribuciones de clase están más equilibradas, el conjunto de algoritmos de clasificación de aprendizaje automático estándar puede ajustarse con éxito en los conjuntos de datos transformados.

Técnica de Submuestreo

La técnica del submuestreo equilibra el conjunto de datos al reducir el tamaño de la clase abundante o mayoritaria. Este método se utiliza cuando la cantidad de datos es suficiente. Al mantener todas las muestras en la clase minoritaria y seleccionar aleatoriamente un número igual de muestras en la clase mayoritaria, se puede recuperar un nuevo conjunto de datos equilibrado para un modelado adicional.

Según Haibo et al. del 2013 [50], el submuestreo aleatorio elimina los casos de clases mayoritaria de los datos de entrenamiento, lo cual reduce el número de observaciones de clase de la mayoría para equilibrar el conjunto de datos. Este método es uno de los mejores cuando la data en conjunto es grande ya que reduce el número de muestras de entrenamiento mejorando de tiempo de ejecución y almacenamiento.

Técnica de Sobremuestreo

Por el contrario, el sobremuestreo intenta equilibrar el conjunto de datos aumentando el tamaño de las muestras raras o minoritarias. En lugar de deshacerse de abundantes muestras, se generan nuevas muestras raras utilizando, por ejemplo, repetición, bootstrapping o SMOTE (Synthetic Minority Over-Sampling Technique).

Al igual que el submuestreo, este método también se puede dividir en dos tipos:

sobremuestreo aleatorio e informativo.

Teniendo en cuenta que no hay una ventaja absoluta de un método de remuestreo sobre otro. La aplicación de estos dos métodos depende del caso de uso al que se aplica y del conjunto de datos en sí. Una combinación de sobremuestreo y submuestreo a menudo también es exitosa.

En este trabajo se utilizarán las técnicas de sobremuestreo y submuestreo sin reemplazo. Los métodos de muestreo tienen como objetivo equilibrar la clase 1 de PayFlag, realizando esto solo en los datos de entrenamiento para el caso de los algoritmos de tipo supervisado.

2.14 Técnica para afrontar datos faltantes

Dentro de las situaciones que pueden producirse en el ámbito del análisis y procesamiento de datos, una de ellas es la ausencia de datos en las observaciones. Estos datos faltantes pueden involucrar desde algunas de las variables, hasta la totalidad de los datos de algunos de los registros seleccionados, siendo inevitable en diferentes estudios de investigación, independientemente de su diseño metodológico [35].

En el trabajo de Viada et al. del 2016 [109], han quedado descritos los estudios de autores, formulando diferentes tipos de técnicas de imputación.

Dentro de las técnicas determinísticas se encuentra:

a) Imputación de la media o moda: si la variable es cuantitativa se reemplaza el o los datos con el promedio, mientras que para variables cualitativas se reemplaza con la moda.

Este es uno de los métodos clásicos para el tratamiento de valores faltantes o perdidos. La sustitución o reemplazo por la media, reemplaza los valores faltantes

en una variable con el valor medio de los valores observados. Los valores faltantes imputados dependen de una y solo una variable, la media entre sujetos para esa variable basada en los datos disponibles. La sustitución de la media preserva la media de una distribución de variables; sin embargo, la sustitución de la media distorsiona típicamente otras características de una distribución de variables [73].

A partir de la existencia de valores faltantes en este trabajo, se determina realizar el tratamiento de estos valores mediante el método de sustitución / imputación por la media de la variable.

2.15 Método para determinar el valor apropiado del parámetro k

Existen varios métodos para determinar el valor apropiado del parámetro k, entre ellos el método del codo, como queda expresado en el trabajo de Kodinariya de 2013 [62].

El método del codo es un método que analiza el porcentaje de varianza explicado como una función del número de grupos. Este método existe con la idea de que uno debe elegir un número de grupos para que agregar otro grupo no proporcione un modelado mucho mejor de los datos. El porcentaje de varianza explicado por los grupos se representa gráficamente frente al número de grupos [26].

Los primeros grupos agregarán mucha información, pero en algún momento la ganancia marginal disminuirá drásticamente y dará un ángulo en el gráfico. En este punto, se elige la "k" correcta, es decir, el número de grupos, de ahí el "criterio de codo". La idea es que comience con K = 2 y siga incrementándolo en cada paso en 1, calculando sus grupos y el costo que conlleva la capacitación. En algún valor de K, el costo cae dramáticamente, y después de eso alcanza una meseta cuando lo aumenta aún más. Este es el valor K que se desea. La razón es que después de esto,

aumenta el número de clústeres, pero el nuevo clúster está muy cerca de algunos de los existentes como queda reflejado en el trabajo de Bholowalia et al. de 2014 [26].

En este trabajo se utiliza el método del codo (*Elbow Method*) para determinar el valor del parámetro "k".

2.16 Método para afrontar valores atípicos

Los valores atípicos son observaciones que se desvían significativamente de la mayoría de las observaciones. Pueden ser generados por un mecanismo diferente que corresponde a los datos normales y pueden deberse al ruido del sensor, las alteraciones del proceso, la degradación del instrumento y / o los errores relacionados con el ser humano. Es inútil hacer un análisis basado en datos cuando los datos están contaminados con valores atípicos porque los valores atípicos pueden conducir a una especificación errónea del modelo, estimación de parámetros sesgados y resultados de análisis incorrectos [70].

En este trabajo se utiliza el método de Tukey (**Tukeys Method**) para remover los valores atípicos. Se obtiene calculando la media (percentil 50) y el rango intercuartil (*Interquartile range*, IQR), es decir, el primer y el tercer cuartil) para escalar cada característica de forma independiente. Este método es más robusto que los valores atípicos, ya que la mediana y el IQR no se ven afectados tanto como la media y la varianza si algunos puntos (eventualmente solo uno) están lejos del centro [61].

2.17 Análisis de componentes principales(PCA)

Análisis de componentes principales (Principal Components Anlaysis, PCA) es un método para reducir la dimensión de los datos y determinar aquellas característi-

cas que más aportan para explicar la variabilidad de todo el conjunto.

Dada una matriz de datos con p variables y n muestras, los datos se centran primero en las medias de cada variable. Esto asegurará que la nube de datos se centre en el origen de nuestros componentes principales, pero no afecta las relaciones espaciales de los datos ni las variaciones a lo largo de nuestras variables. Los primeros componentes principales (Y1) están dados por la combinación lineal de las variables X1, X2, ..., Xp, [52].

Aunque estas relaciones pueden parecer obvias, cuando se trata con muchas variables, este proceso permite evaluar mucho más rápidamente cualquier relación entre variables. Para conjuntos de datos con muchas variables, la varianza de algunos ejes puede ser grande, mientras que otros pueden ser pequeños, de modo que pueden ignorarse. Esto se conoce como reducir la dimensionalidad de un conjunto de datos, de modo que uno podría comenzar con treinta variables originales, pero podría terminar con solo dos o tres ejes significativos. El nombre formal para este enfoque de rotación de datos de modo que cada eje sucesivo muestre una disminución de la varianza se conoce como Análisis de componentes principales, o PCA. PCA produce combinaciones lineales de las variables originales para generar los ejes, también conocidos como componentes principales o PC (principal components).

De acuerdo al trabajo de Wold et al. de 1987 [112], al aplicar el método PCA y analizar una matriz de datos se pueden tener diferentes objetivos entre los que están:

- selección de variables (variables selection)
- reducción de los datos (data reduction)
- detección de anomalías (outlier detection)
- clasificación (classification)

Para proceder a la aplicación del método de PCA, se siguen los siguientes pasos

[103]:

Paso 1: Obtener los datos

Paso 2: Sustraer la media

Para que PCA funcione correctamente, debe restar la media de cada una de las dimensiones de datos. La media restada es el promedio en cada dimensión. Por lo tanto, todos los x valores tienen $\overline{x}$ (la media de los x valores de todos los puntos de datos) restados, y todos los valores y tienen $\overline{y}$ restados de ellos. Esto produce un conjunto de datos cuya media es cero [103].

Paso 3: La covarianza siempre se mide entre 2 dimensiones. Si se tiene un conjunto de datos con más de 2 dimensiones, se puede calcular más de una medición de covarianza. Por ejemplo, a partir de un conjunto de datos tridimensionales (dimensiones x, y, z), se puede calcular $cov(x, y)$, $cov(x, z)$, $cov(y, z)$. De hecho, para un conjunto de datos tridimensionales, se puede calcular $\frac{n!}{(n-2)!*2}$ valores de covarianza diferentes [103].

Paso 4: Calcular los vectores propios y los valores propios de la matriz de covarianza

Como la matriz de covarianza es cuadrada, se puede calcular los vectores propios y los valores propios de esta matriz. Estos son valores importantes, ya que brindan información útil sobre los datos [103].

Paso 5: Elegir los componentes y formar un vector de características

Aquí es donde entra en juego la noción de compresión de datos y dimensionalidad reducida.

En general, una vez que se encuentran los vectores propios de la matriz de covarianza, el siguiente paso es ordenarlos por valor propio, de mayor a menor. Esto le da los componentes en orden de importancia. En este pueden ser descartados los componentes de menor importancia, con la consiguiente pérdida de información,

pero si los valores propios son pequeños, no la pérdida de información no es significativa. Si son omitidos algunos componentes, el conjunto de datos final tendrá menos dimensiones que el original, es decir si originalmente se tienen n dimensiones en los datos, por lo que se calculan n vectores propios y valores propios, y luego se elige solo los primeros p vectores propios, entonces el conjunto de datos final solo tiene p dimensiones.

A continuación, se necesita formar un vector de características, que no es más que una matriz de vectores. Esto se construye tomando los vectores propios que desea mantener de la lista de vectores propios, y formando una matriz con estos vectores propios en las columnas [103].

Paso 6: Derivando el nuevo conjunto de datos

Este es el paso final en el método PCA. Una vez que se han elegido los componentes (vectores propios) que se desea mantener en los datos y se ha formado un vector de características, se toma la transposición del vector y se multiplica a la izquierda del conjunto de datos original transpuesto.

Datos Finales = Vector-de-características-de-fila X Ajuste-de-datos-de-fila

donde Vector de características de fila, es la matriz con los vectores propios en las columnas transpuestas para que los vectores propios estén ahora en las filas, con el vector propio más significativo en la parte superior, y Ajuste de datos de fila, son la información transpuesta ajustada a la media. Los Datos Finales son el conjunto de datos final, con elementos de datos en columnas y dimensiones a lo largo de filas.

2.18 Revisión Bibliográfica

A continuación, se describen de forma general, los pasos de la metodología que se emplea para llevar a cabo el análisis bibliográfico:

1-Definir los temas y subtemas relacionados al trabajo de investigación.

2-Buscar entre 50 y 60 artículos más citados relacionados con el tema y según un conjunto de palabras clave. Ordenar en forma descendente según número de citas. Seleccionar los 20 más citados o los que tengan 50 citas o más.

3-En el total de artículos, identificar a los 20 autores más citados (aplicar un procedimiento similar al anterior para seleccionarlos) de los 50 o 60 artículos seleccionados, anotar los autores y coautores, y sumar las citas de los artículos en los que aparezcan. Ordenar por número de citas descendentes y seleccionar los 20 primeros.

4-Clasificar los artículos y autores según: subtemas, fecha de publicación considerándolos de la siguiente manera:

-clásico: antes del 2000

-contemporáneo: del 2000 al 2015

-recientes: del 2015 a la fecha actual.

5-Posteriormente se crea un libro de MS Excel, que contenga dos tablas en hojas separadas, una para los 20 artículos más citados y otra para los 20 autores más citados. La tabla de artículos contiene las columnas siguientes: año, autores, título, revista (o editorial en caso de libro, o institución en caso de tesis), palabras clave del artículo, subtema, número de citas. La tabla de autores contiene las columnas siguientes: nombre completo, número de citas de artículos clásicos, número de citas de artículos contemporáneos y número de citas de artículos recientes, número de artículos clásicos, número de artículos contemporáneos y número de artículos recientes, número de subtemas.

6-Repetir los pasos 2-4 restringiendo ahora el dominio a las listas de bibliografías que aparecen en los 20 artículos. Actualizar en consecuencia las tablas de Excel creadas en el paso 5. Como resultado se seleccionan los 20 artículos más cita-

dos y los 20 autores más citados.

7-Para cada uno de los artículos crear una ficha bibliográfica que contenga:

-Título.

-Autores.

-Revista/editorial, año de publicación.

-Subtema.

-Resumen.

-Conclusiones.

-Novedad científica.

-Crítica personal resaltando con relación al tema o subtema que falta por tratar, o se trata de forma insuficiente o poco rigurosa, o el alcance es limitado, etc...

Estos artículos constituyen un núcleo básico para continuar profundizando en la bibliografía relacionada con el tema sujeto a investigación.

En la Tabla 2.1 se muestra la cantidad de artículos más citados de acuerdo a la clasificación presentada anteriormente.

Tabla 2.2: Tabla de artículos más citados.

Clásicos	**Contemporáneos**	**Recientes**
6	30	9

En la Tabla 2.2 se presenta la cantidad de autores más citados de acuerdo a la clasificación presentada anteriormente.

Tabla 2.3: Tabla de autores más citados.

Clásicos	**Contemporáneos**	**Recientes**
5	30	9

2.18.1 Clásicos

En esta sección se abordan los artículos que se encontraron hasta el año 2000, entre los cuales se encuentra el trabajo de Holtz et al. de 1998 [53], en donde se propone una metodología de caracterización de reservas a partir de las propiedades de la permeabilidad relativa y la presión capilar. Uno de los puntos importantes de esta metodología propuesta, que consta de 4 pasos iterativos, es la identificación de la correspondencia entre la arquitectura de la reserva y las tendencias de unidades de flujo para luego establecer cual paquete o unidad genética actúan como compartimientos o unidades de flujo [53].

Otro de los trabajos es el de Mohaghegh de 1996 [79], en donde se utiliza una estructura de red neuronal artificial, para la caracterización de reservas de petróleo heterogéneas a partir de diferentes propiedades geológicas como la porosidad, la permeabilidad y la saturación de fluidos. En este trabajo se pone de manifiesto que se puede predecir con precisión este tipo de propiedad mediante una ANN. Esta técnica tiene varias ventajas como queda reflejado en el trabajo de Mohaghegh del 2000 [80] y el otro trabajo igualmente de Mohaghegh [78] donde quedan reflejadas las ventajas de las técnicas de ANN por encima de las técnicas convencionales existentes hasta el momento.

Otro de los trabajos de la literatura de este período es el de Ahmed et al. de 1997 [4] en el que se propone un modelo de red neuronal combinando datos de pozos, geología y datos de superficies sísmicas para predecir parámetros importantes del yacimiento. El modelo de este trabajo predice parámetros como la porosidad y la permeabilidad haciendo uso de datos sísmicos 3-D. Este tipo de metodología provee

una importante retroalimentación para el proceso de modelado de yacimientos. De igual modo, Balch et al. de 1999 [20] proponen el uso de la técnica de redes neuronales en una nueva metodología a fin de predecir propiedades del yacimiento en dos zonas de Nuevo México, EUA (Estados Unidos de América). Dado que utilizar todos los atributos es computacionalmente infactible utilizan un algoritmo basado en rankeo difuso *fuzzy-ranking*para seleccionar los parámetros más adecuados para realizar la predicción. En esta propuesta fueron utilizados regresiones no lineales. También aparece el modelo basado en una red neuronal para estimar la permeabilidad de Huang et al. de 1996 [57], en el que utilizan *back-propagation* ANN (BP-ANN) para modelar las relaciones entre la posición espacial, los registros de pozos y la permeabilidad.

En el trabajo de Gharbi et al. de 1997 [48] se propone un modelo basado en redes neuronales para estimar datos de campo como las propiedades de presión, volumen y temperatura (PVT), por la importancia que significan estos parámetros para determinar el desempeño del yacimiento.

Por otra parte, queda reflejado en el trabajo de Aminzadeh de 1999 [7], que la precisión de la clasificación de ANN se mejora drásticamente al transformar el espacio de características de entrada de la ANN en un nuevo espacio de entrada dimensionalmente más pequeño.

El año 2000 contó con varios trabajos basados en redes neuronales como el de Trappe [106], en donde se propone un modelo para estimar la porosidad a partir de datos sísmicos basado en redes neuronales aprovechando las bondades de esta técnica que es capaz de manejar bien la ambigüedad y el ruido de los datos fuente. Este modelo estima a partir de la clasificación multi-atributo de datos sísmicos en 3D las propiedades del yacimiento. De igual modo, Jamialahmadi et al. [59] propone un modelo basado en redes neuronales para determinar la relación entre varias propiedades de la roca como la permeabilidad, la porosidad y la profundidad, y en el trabajo de Yeten et al. [117] se propone un modelo basado en el uso de redes neuronales artificiales (ANN) para generar datos de permeabilidad a intervalos sin núcleo

de registros de porosidad a partir de datos conocidos del pozo. Dentro de este mismo período aparece otro trabajo en donde se propone una metodología basada en redes neuronales para caracterizar reservas fracturadas a partir de datos de la porosidad de la roca, del autor Sahimi [97]. El modelo de red neuronal es capaz de analizar los patrones de la red fracturada para estimar la distribución de la permeabilidad y construir correlaciones precisas. En el trabajo de Ouenes [86], también se propone una metodología que utiliza una red neuronal difusa para evaluar el efecto jerárquico de cada controlador geológico en las fracturas, de modo que los geólogos o ingenieros de yacimientos podrán identificar local y globalmente los factores geológicos clave que afectan las fracturas y contribuir así a la caracterización del yacimiento.

2.18.2 Contemporáneos

En esta sección se abordan los artículos que se encontraron en el período comprendido desde el año 2001 al 2015, entre los que se encuentra el trabajo de Nikravesh et al. del 2001 [75], en donde se aborda la importancia de la integración de las técnicas de análisis de datos (*Data Analytics*) como son: algoritmos genéticos, lógica difusa, redes neuronales, sistemas expertos, razonamiento probabilístico y técnicas de procesamiento en paralelo y las metodologías emergentes para la caracterización y exploración inteligente de yacimientos, lo cual redunda en una mejora de la capacidad para descubrir nuevas reservas potenciales. En general *Data Analytics* son el conjunto de herramientas robustas, manejables, eficientes y más económicas de implementar. En tanto, en otro trabajo del mismo autor, pero del 2003 [84] se propone una metodología integrada para identificar la relación no lineal y el mapeo entre datos sísmicos tridimensionales y registro de producción. Este trabajo involucra técnicas convencionales y a su vez las técnicas de inteligencia artificial de redes neuronales, lógica difusa, razonamiento probabilístico y computación genética, sin embargo algunas técnicas presentan limitaciones como los algoritmos genéticos (*Genetic Algorithms*, GA), autores como Huang et al. [56] realizan un estudio de la predicción

de la permeabilidad del yacimiento utilizando algoritmos genéticos y a pesar de que el estudio mostró que las redes entrenadas en GA (modelo neural-genético) dieron errores consistentemente más pequeños en comparación con las redes entrenadas por el algoritmo de descenso de gradiente convencional (propagación hacia atrás), sin embargo, las GA fueron comparativamente lentas en convergencia.

Por otra parte, en el trabajo del 2004 de nikravesh [83], se presenta una revisión de la caracterización de yacimientos donde se plantea que juega un papel crucial en el manejo moderno de yacimientos. Esta ayuda a tomar decisiones acertadas sobre los yacimientos y mejora el valor de los activos de las compañías de petróleo y gas. Maximiza la integración de datos y conocimientos multidisciplinarios y mejora la fiabilidad de las predicciones del yacimiento. El producto final es un modelo de depósito con tolerancia realista a la imprecisión y la incertidumbre. Análisis de datos (*Data Analytics*) (término con el que nombra al consorcio de herramientas de metodologías computacionales), pretende explotar tal tolerancia para resolver problemas prácticos. En la caracterización de yacimientos, estas técnicas inteligentes pueden usarse para el análisis de la incertidumbre, evaluación de riesgos, fusión de datos y minería de datos que son aplicables a la extracción de características de atributos sísmicos, registros de pozos, mapeo de reservorios e ingeniería. El objetivo principal es integrar datos blandos como datos geológicos con datos duros como sísmica 3D y datos de producción para construir un reservorio y modelo estratigráfico. Mientras que algunas metodologías individuales (especialmente la neurocomputación) han ganado mucha popularidad durante los últimos años, el verdadero beneficio de la (*Data Analytics*) radica en la integración de sus técnicas con las metodologías emergentes constituidas en vez de usarlas en forma aislada [83]. Además, en este trabajo se presenta un caso de estudio en el que analizan los clusters con 3 técnicas diferentes: *k-means*, red neuronal y *fuzzy c-means* y como resultado de la agrupación, estos clusters se superpusieron a los datos reconstruidos del registro de producción y se determinaron las ubicaciones óptimas para perforar nuevos pozos.

Un poco más adelante en el año 2005 aparecen varios trabajos como el de Wong

[113] donde se presenta una investigación basada en la técnica de máquina de soporte vectorial y la Redes neuronales de retropropagación (*Backpropagation Neural Networks, BPNN*) para la caracterización de un yacimiento partir de datos geológicos, obteniéndose mejor precisión por parte de la técnica de SVM, aunque una forma de obtener una mejora en la precisión sería integrar otras formas de preprocesamiento y posprocesamiento en el modelo BPNN, sin embargo, la compensación es, un aumento en el tiempo y la complejidad computacional con esta técnica. Otro trabajo de este período basado en la técnica de redes neuronales y lógica difusa para caracterizar reservorios fracturados naturalmente y determinar la relación compleja entre el índice de fractura y algunos factores geológicos y geomecánicos (facies, porosidad, permeabilidad, grosor del lecho, proximidad a fallas, pendientes y curvaturas de la estructura) con el fin de obtener un mapa de intensidad de fractura queda reflejado en el trabajo de Ouaheda et al. [1]. De igual modo, en este mismo año aparece el trabajo de Jong-Se Lim [69] donde se sugiere aplicar una estrategia que se apoya en la lógica difusa y una red neuronal para estimar propiedades de yacimientos a partir de registros de pozos. En este trabajo se realiza el análisis de curva difusa basado en la lógica difusa que se utiliza para seleccionar los mejores registros de pozos relacionados con datos de porosidad y permeabilidad del núcleo, tomando en cuenta que estas son las dos propiedades fundamentales de la roca que se relacionan con la cantidad de fluido contenido en un yacimiento y su capacidad de fluir cuando se somete a gradientes de presión aplicados. Estas propiedades tienen un impacto significativo en las operaciones de los campos petroleros y la gestión de los yacimientos. Sin embargo, la estimación de la porosidad y la permeabilidad de los registros de pozos convencionales en formación heterogénea constituyen un problema difícil y complejo de resolver mediante métodos estadísticos convencionales.

De entre las técnicas que pueden ser empleadas para predecir la permeabilidad se encuentran los árboles de clasificación, esta exploratoria para facilitar la determinación de la importancia relativa de los diferentes registros de pozos durante la clasificación de datos de acuerdo con el trabajo de Pérez del 2005 [89]. En este, se

efectúa un acercamiento al método basado en árboles proponiendo un enfoque que explique los registros de pozos faltantes durante las predicciones de la permeabilidad. Este último es uno de los parámetros más importantes en la caracterización de pozos de petróleo o yacimientos, según Aminzadeh del 2005 [8], pues controla el caudal y direccional movimiento de diferentes fluidos de fase (a saber, gas, agua y petróleo) a través de las formaciones del reservorio. Esto unido a la naturaleza compleja de los yacimientos con hidrocarburos y considerables heterogeneidades en la formación de rocas a través de las cuales se propagan las ondas sísmicas o fluidos, y dada la extremadamente escasa naturaleza de los datos disponibles, junto con mediciones directas y precisas limitadas (registros de pozos, caudales y muestras de núcleos) hace el modelado y trabajo de validación aún más difícil. Adicionalmente diferentes incertidumbres, errores de medición y aproximaciones de las propiedades medias hacen las ecuaciones teóricas menos realistas. Todo esto anterior hace a la tarea de caracterización de reservorios de petróleo a partir de datos geológicos una tarea compleja [8].

Existen otros ejemplos de la aplicación de la inteligencia artificial caso como lo desarrollado por Li et al. del 2010 [68] que emplean la técnica de árboles de decisión para predecir la producción del yacimiento a partir de datos geológicos.

Alcanzando el año 2011, llega el trabajo de Vaferi et al. [108] proponiendo un modelo de red neuronal multi-capa aplicado al reconocimiento de modelos de reserva a partir de datos geológicos. En el trabajo de Fatai et al. [39] se pone de manifiesto los métodos de lógica difusa Tipo-2 en conjunto con modelos de inteligencia computacional híbrida y SVM para la caracterización de yacimientos, específicamente en la predicción de dos importantes propiedades: la porosidad y la permeabilidad, a partir de datos geológicos de pozos. En este trabajo se pone de manifiesto claramente que no hay una técnica única y perfecta que funcione bien en todos los casos, en todas las situaciones; de ahí la necesidad de modelos híbridos que combinará las mejores características de cada técnica en un solo paquete, aumentando la confianza en la predicción de varias propiedades de yacimientos de petróleo. Una ves más se plantea

que la caracterización de yacimientos de petróleo es un proceso para describir cuantitativamente varias propiedades de yacimientos en la variabilidad espacial utilizando los datos de campo disponibles, desempeña un papel crucial en la gestión moderna del yacimiento, para así contribuir a tomar decisiones acertadas sobre el yacimiento y mejorar la confiabilidad de las predicciones del yacimiento. En este trabajo (de Fatai et al.), se ponen de manifiesto limitaciones de estas técnicas, para el caso de la Lógica difusa Tipo-2 no obtiene un buen rendimiento cuando el número de datos de entrenamiento es pequeño, pero funciona mejor cuando el número de prototipos de entrenamiento es grande. Esta debilidad se complementa con la capacidad de la técnica de la SVM para manejar pequeños conjuntos de datos. Esta última cuenta con fortalezas que radican principalmente en su relativa facilidad de entrenamiento, la ausencia de óptimos locales a diferencia del caso de ANN, escala relativamente bien a datos de alta dimensión, tiene la capacidad de controlar explícitamente la compensación entre la complejidad y el error, y tiene la capacidad de manejar datos no tradicionales como árboles como entrada al sistema, en lugar de vectores de características. Sin embargo, es débil en el sentido de que necesita una función de núcleo "buena" [39].

Por otro lado, existen varias variables que constituyen los datos de entrada a una red neuronal artificial como pueden ser la porosidad, el volumen de arcilla, rayos gamma, densidad, como se refleja en el trabajo de Anifowose et al. de 2011 [15], para predecir la permeabilidad. Otras características a predecir de los yacimientos de petróleo y gas, son la profundidad, temperatura, presión, volumen, estructura y sello, mecanismo de manejo y porosidad.

Más adelante en el año 2012 se sugiere un modelo basado en métodos ANN, unido a la realización de un análisis de atributos múltiples y datos de registros de pozos para determinar la alteración y la heterogeneidad de las litofacies en uno de los campos de petróleo estructural-estratigráfico, como queda reflejado en [91] de Raeesi et al., no obstante la técnica de ANN presenta limitaciones ya que el entrenamiento excesivo o la memorización de la red neuronal podría ser un inconveniente.

Una vez que la red memoriza un conjunto de datos, será incapaz de generalización. Se ajustará al conjunto de datos de entrenamiento con bastante precisión, pero sufre en generalización. La memorización y el sobreentrenamiento son aplicables a aquellas redes que usan un proceso iterativo durante el entrenamiento. Para evitar el sobre entrenamiento o la memorización, una práctica común es detener el proceso de entrenamiento de vez en cuando y aplicar la red al conjunto de datos de calibración. Dado que la salida del conjunto de datos de calibración no se presenta a la red, uno puede evaluar las capacidades de generalización de la red según cuán bien predice la salida del conjunto de calibración, y en el trabajo de Ferraretti et al. [41] de este mismo año, se combinan varias técnicas de aprendizaje no supervisado y supervisado, proponiendo una configuración en cascada, aplicando primero el algoritmo agrupamiento jerarquizado (*hierarchical clustering*) con el objeto de encontrar una estructura escondida en los datos obteniéndose un conjunto de clústers, de estos, los mejores son seleccionados y entregados como conjunto de entrenamiento al algoritmo supervisado de clasificación para caracterizar yacimientos a partir de datos geológicos. Uno de los mayores beneficios con esta propuesta, es que se logran manejar grandes cantidades datos simultánemente, y la aplicación del algoritmo agrupamiento jerarquizado, ahorra mucho tiempo a los geólogos.

Para el año 2013, aparece el trabajo de Fatai et al. [40] en donde se propone un modelo de conjunto (*Ensemble*) de Redes Neuronales Artificiales (ANN) que incorpora varias opiniones de expertos sobre el número óptimo de neuronas ocultas en la predicción de las propiedades de los yacimientos de petróleo, planteando que los modelos de conjunto funcionan mejor que el rendimiento promedio de los alumnos (*learners*) base individuales y que la técnica de (*Ensemble*) tiene un gran potencial en las tareas de caracterización de yacimientos de petróleo. Por otra parte, en el trabajo de Anifowose et al. [12] del mismo año, se propone un modelo de conjunto (*Ensemble*) de Redes Neuronales Artificiales (ANN) para predecir propiedades de los yacimientos. En este trabajo, se emplea esta técnica híbrida debido a la alta dimensionalidad y heterogeneidad de los datos de los sensores que se utilizan en los

pozos, viéndose la capacidad de las técnicas convencionales de IA limitada ya que no pueden manejar más de una hipótesis a la vez. El método de aprendizaje conjunto (*Ensemble*), tiene la capacidad de combinar varias hipótesis para desarrollar una solución de conjunto única a un problema. Seguidamente, en el trabajo de Anifowose et al. del mismo período [14] se propone un modelo de aprendizaje híbrido utilizando diferentes versiones de modelos híbridos adaptativos del sistema de inferencia Neuro-Fuzzy para predecir propiedades del yacimiento basado en datos geológicos. Se combinan una red funcional-SVM (FN-SVM) y una red funcional -Lógica difusa (FN-T2FL), esta útlima para extraer los atributos más relevantes de los datos de entrada. El primero es más prometedor ya que combina dos técnicas existentes que tienen un rendimiento muy cercano y son bien conocidas por su estabilidad computacional y procesamiento rápido. El entrenamiento y la prueba del componente SVM del modelo híbrido con las mejores variables reducidas dimensionalmente de los datos de entrada permitieron obtener como resultado un mejor rendimiento con coeficientes de correlación más altos, errores cuadrados medios de raíz más bajos y menos tiempo de ejecución que el modelo SVM tradicional. El modelo de FN-SVM mostró superioridad por encima de los demás modelos utilizados. En otra investigación llevada a cabo de Ahmadi et al. del 2013 [2] se propone un modelo basado en redes neuronales combinadas con algoritmo genético híbrido y optimización de Enjambre de partículas para predecir la permeabilidad del yacimiento.

En el trabajo de Li et al del 2013 [67], se utiliza una red neuronal para predecir la producción de un yacimiento a partir de datos geológicos.

2.18.3 Estado del Arte

En esta sección se realiza una revisión de los trabajos de los últimos 5 años sobre la caracterización de pozos de petróleo a partir de datos geológicos que recoge los trabajos de Fatai Adesina, Jared Schuetter, Yunxin Xie y Chenyang Zhu, entre otros.

En el trabajo propuesto por Fatai Adesina et al. del 2015 [13], se propone un modelo de conjunto de generalización apilado de SVM que incorpora diferentes opiniones de expertos sobre los valores óptimos de este parámetro en la predicción de porosidad y permeabilidad de los yacimientos de petróleo utilizando conjuntos de datos de diversas formaciones geológicas. Este método supera a las demás técnicas con el mayor coeficiente de correlación y la menor media, y el error absoluto. El estudio indicó que existe un gran potencial para el aprendizaje conjunto en la caracterización de yacimientos de petróleo para mejorar la precisión de las predicciones de propiedades de yacimientos para exploraciones más exitosas y una mayor producción de recursos de petróleo. Los resultados también confirmaron que los modelos de conjunto funcionan mejor que la implementación de SVM convencional. Otro de los trabajos de este mismo año es el de Schuetter et al. [100], que se enfoca en la construcción de modelos predictivos robustos y en desarrollar reglas de decisión que ayuden a identificar los factores que separan los buenos pozos de los que tienen un desempeño deficiente, con el objetivo de optimizar la producción de yacimientos no convencionales. En este trabajo se emplean métodos como los Bosques aleatorios (*Random Forests*), SVM, Máquina de aumento de gradiente (*Gradient Boosting Machine*), entre otros. Además en el trabajo de Akande et al. del 2015 [5] se propone un estudio basado en las ANN y la SVM para caracterizar yacimientos a partir de datos geológicos de pozos, utilizando pequeños datasets, tomando en cuenta que la SVM arroja mejores resultados sobre grupos de datos pequeños, no obstante esta técnica se ve limitada por el hecho de que la elección de ϵ es primordial y crítica para el rendimiento de SVM, ya que controla la complejidad del modelo donde el pequeño valor del parámetro conduce a un modelo ajustado y viceversa. Este trabajo demuestra que la ANN supera a la SVM en el caso de un conjunto de datos grande, mientras que la técnica de SVM funciona mejor en un escenario de conjuntos de datos pequeños [5].

Ya en el de 2016 se presenta el trabajo de Ani et al. [11] que se basa en el modelado de incertidumbre del yacimiento considerando las distintas variantes de

los algoritmos de inteligencia artificial más allá de los métodos convencionales.

Un año más tarde, en el 2017 se encuentra el trabajo de Anifowose [17], planteando que las ANN sufren de limitaciones como no tener un marco genérico para diseñar apropiadamente la red para una tarea específica, y la necesidad de determinar el número de capas y neuronas ocultas de la arquitectura de red es determinada mediante ensayo y error, además esta técnica de ANN requiere gran número de parámetros para ajustar bien la estructura de la red. En un trabajo de los mismos autores, Anifowose et al. en el mismo año [16], se presentan diversos ejemplos exitosos de la aplicación de la técnica aprendizaje de conjuntos (*Ensemble Learning*) en sistemas híbridos para modelar y caracterizar propiedades de yacimientos de petróleo, en específico predecir propiedades tales como la porosidad, permeabilidad, PVT, entre otras. Debido a que el problema de caracterización de los yacimientos de petróleo es tan complejo y lleno de incertidumbres que la existencia de diversas opiniones de expertos que conducen a diversas hipótesis necesita una solución más cooperativa y sólida. Dicha solución debería ser capaz de incorporar e integrar las diversidades existentes de opiniones de expertos para resolver estos problemas complejos. Se plantea que con la técnica de aprendizaje de conjuntos, es posible mejorar aún más el rendimiento de las técnicas de Inteligencia Artificial existentes.

En el trabajo de Oloso et al. de 2017 [85] se propone una forma de hacer mediante un método híbrido a partir de una red funcional híbrida predecir las propiedades PVT (Presión-Volumen-Temperatura) de una reserva. También en otro trabajo del mismo año 2017, Gulyani et al. [49] pone de manifiesto una implementación del método de aprendizaje de conjuntos de tipo (*Ensemble Bagging*) para predecir el factor de viscosidad de petróleo muerto.

En un período más reciente, de los trabajos revisados se encuentra el de Xie et al. del 2018 [116] donde se evalúan 5 tipos de métodos de aprendizaje automatizado, a saber, Algoritmos Naive Bayes (*Naïve Bayes*), SVM, ANN, Bosques aleatorios y mejora del árbol de gradiente (*Gradient Tree Boosting*), para la identificación de la

litología de formación utilizando datos de campo.

A continuación, una tabla con los cinco artículos más citados de los últimos 5 años.

Otro trabajo sobre el que se basa la presente investigación es el que tuvo lugar como resultado de una colaboración a través de un proyecto de investigación de la UANL a cargo del Dr. Arturo Berrones empleando técnicas de minería de datos. De ese proyecto se obtuvo como uno de sus resultados, el proyecto de tesis bajo el nombre Modelos para la evaluación de pozos de petróleo a partir de sus características geológicas [22]. En esta tesis se plantea que los datos tienen una complejidad importante, además en este trabajo se realizó un estudio descriptivo de los datos. Esta tiene una gran cantidad de datos a diferencia de los datos que se han encontrado geológicamente, con más de un millón de registros. El presente trabajo de investigación le da continuidad al trabajo de tesis antes mencionado.

En general a partir de los trabajos existen, se considera que existen limitaciones para realizar una caracterización de pozos de petróleo cuando los datos no están homogéneos, las clases están desbalanceadas y existe una gran cantidad de registros de pozos, por lo que se considera la aplicación de métodos de aprendizaje de conjuntos (*Ensemble*).

2.19 Conclusiones

La clasificación y caracterización de pozos de petróleo a partir de datos geológicos es un problema complejo, sobre todo cuando los datos están incompletos.

La intención de este trabajo es aportar una metodología de apoyo a la decisión, por lo que se incluye la teoría básica de la decisión multicriterio. La clasificación de pozos de petróleo a partir de datos geológicos para determinar la existencia o no de petróleo implica el uso de múltiples parámetros, lo que convierte a este problema en

un problema complejo.

Acorde a la literatura revisada sobre la caracterización de pozos de petróleo aplicando técnicas de aprendizaje automatizado, se considera que existe una amplia investigación abordada en su mayoría a través de las técnicas de redes neuronales, máquina de soporte vectorial, lógica difusa o combinaciones de estas técnicas. Sin embargo, se encontraron muy pocas referencias que reflejen la utilización de la técnica de bosques aleatorio de aprendizaje automatizado.

Tabla 2.4: Los cinco artículos más citados de los últimos 5 años.

Artículo	**Año**	**Autor**	**Citas**
Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machine	2015	Fatai A. Anifowose, J Labadin, A. Abdulraheem	43
Data Analytics for Production Optimization in Unconventional Reservoirs	2015	Jared Schuetter, Srikanta Mishra, Ming Zhong, and Randy LaFollette	35
Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances	2018	Yunxin Xie, Chenyang Zhu, Wen Zhou, Zhongdong Li, Xuan Liu, Mei Tu	31
Ensemble Machine Learning: An Untapped Modeling Paradigm for Petroleum Reservoir Characterization	2017	Fatai Adesina Anifowose, Jane Labadin, Abdulazeez Abdulraheem	24
Hybrid intelligent systems in petroleum reservoir characterization and modeling: the journey so far and the challenges ahead	2017	Fatai Adesina Anifowose, Jane Labadin, Abdulazeez Abdulraheem	17

Capítulo 3

Descripción del problema

3.1 Introduccion

En este capítulo se describe el problema de este trabajo de forma detallada, así como se explican las distintas variables que se toman en cuenta para la caracterización de los pozos de petróleo y además se describen los retos y costos de la actividad de exploración de petróleo.

3.2 Planteamiento del problema

El problema de caracterización de yacimientos de petróleo está vinculado al análisis e interpretación de grandes conjuntos de datos de entrada no etiquetados, datos incompletos y estructura de datos subyacente que se encuentra oculta o no está claramente definida. En la rama de la geología del petróleo, la comprensión y caracterización de los yacimientos necesita de la integración de diferentes datos del subsuelo para crear modelos de yacimientos confiables, que contribuyan a la labor de los geólogos de estimar y predecir varias propiedades de yacimientos para su uso a gran escala y a la determinación de la calidad y posible volumen de petróleo existente en un yacimiento. La gran cantidad de datos para cada pozo y la presencia

de diferentes pozos para analizar simultáneamente hacen que esta tarea sea compleja y lenta. En el contexto anterior, el desarrollo de métodos de caracterización confiables es de primordial importancia para ayudar al geólogo y reducir la subjetividad de la interpretación de los datos que contribuya a la toma de decisiones.

Actualmente la clasificación de los datos de pozos de petróleo es realizada por humanos, con base en la interpretación de los datos obtenidos de las lecturas de diversos sensores. Sin embargo, esto presenta varios problemas, entre ellos: la interpretación humana, la cual es subjetiva y depende de quién interprete los datos, también es un proceso que tiene implicaciones financieras importantes puesto que determina las inversiones en los pozos de petróleo. Por otra parte, disponer de herramientas automatizadas que validen la interpretación humana es altamente deseable, pero esto supone un gran reto ya que primero estas herramientas deben ser capaces de realizar interpretaciones con calidad similar a los humanos.

En resumen, el problema abordado en esta investigación puede formularse de la siguiente manera: ¿cómo determinar la influencia de la geología del terreno en la existencia o no de petróleo en una localización geográfica a partir de datos geológicos heterogéneos con la aplicación de aprendizaje automatizado?

3.3 Atributos que caracterizan a los pozos

El proceso de caracterización de pozos de petróleo se apoya en la relación entre los propiedades o atributos que reflejan las propiedades físicas de la roca y del sistema roca-fluido. En algunos casos para determinar el valor de algunos atributos, se necesita previamente obtener el valor de otros, como es el caso de los atributos: permeabilidad y en otro caso el *net pay*. Existen, además, un conjunto de atributos de los cuales se determina su valor a través de las lecturas de los sensores, entre los que se encuentran: los rayos gamma, la porosidad, la densidad, la saturación del agua, la resistividad somera y profunda, la porosidad de neutron, el caliper [13], y

otros atributos son el pay flag, la porosidad efectiva, la profundidad (*depth*), presión, volumen y la temperatura.

De las propiedades anteriores, la porosidad, la permeabilidad y el PayFlag son las propiedades más importantes para los ingenieros del petróleo ya que conjuntamente sirven como indicadores claves de la calidad del yacimiento.

A continuación, se describen varios de los atributos geológicos presentes en la literatura, utilizados para realizar caracterización de pozos:

1. **porosidad (p)(%)**: es la característica física más conocida de un reservorio o depósito. Esta propiedad determina los volúmenes de petróleo que pueden estar almacenados y la determinación de su valor es la base para definir los procesos de recuperación. Se define como la fracción del volumen total de la roca no ocupada por el esqueleto mineral de la misma. Es el porcentaje del espacio total de la formación que puede estar ocupado por los hidrocarburos [19]. En definitiva, es la capacidad de acumulación de la roca.

   La porosidad se determina por: medición directa realizada durante la perforación del pozo a través de testigos coronas, y también por medición indirecta en el momento de perfilar el pozo [19]. La unidad de medida de este atributo es en porciento.

   Existen varios tipos de porosidad, entre las que se encuentran: la porosidad efectiva, que se define por la existencia de poros continuos que están interconectados entre sí, la porosidad No efectiva (PHIE), que está presente cuando los poros en la roca están discontinuos y aislados.

   Cuanto mayor sea el porcentaje de poros en una muestra de roca, mayor será su capacidad para retener hidrocarburos, agua y gas [13]. Para este atributo lo deseable es que el pozo posea la mayor porosidad posible, pues cuanto mayor sea el valor de este atributo más fácil resultará en la extracción del hidrocarburo.

La expresión para el cálculo de la porosidad está dada por:

$$p = 100 * \frac{Vol. \quad de \quad espacios \quad porosos \quad en \quad la \quad roca}{Vol. \quad total \quad de \quad la \quad roca} \tag{3.1}$$

2. **permeabilidad (k)**: es la conductividad de la roca a los fluidos, o bien, es la capacidad de la roca de permitir el movimiento de los fluidos a través de la red de poros intercomunicados bajo cierta fuerza impulsora, guardando relación directa con la porosidad efectiva [19]. La unidad de medida de este atributo es en Darcy (D) o milidarcy (md). Se considera que la permeabilidad es baja cuando es menor que 1 md, es moderada cuando se encuentra en el rango de 1 md y 1 Darcy, mientras que la alta permeabilidad es mayor que 1 Darcy [27]. Para este atributo lo deseable es que el pozo posea la mayor permeabilidad posible. En general, en la reservas de pétróleo, el valor de la permeabilidad es menor de 1 darci. En general, un reservorio de petróleo que tiene una permeabilidad muy baja (en los dígitos pequeños bajos de milidarcies) puede no verse como un buen candidato para una producción sustancial durante un largo período [99].

3. **saturación de agua (Sw):** representa el espacio poroso de la formación o roca que puede estar ocupado por los fluidos: petróleo, agua y gas. El contenido de cada uno de estos fluidos en el espacio poroso representa la saturación. Depende de la resistividad, del volumen de Lutita, exponente de segmentación y de saturación, y de la porosidad total y efectiva. La forma de hallar la saturación (S) es mediante la expresión: $S_0 + Sg + Sx = 100$

   Una expresión clásica para obtener el valor de la saturación de agua está dada por la ecuación de Archie, que está expresada en función de la porosidad (S_w), la resistividad el agua (R_w) y la verdadera resistividad de formación (R_t).

$$S_w = \left(\frac{aR_w}{\phi^m R_t} \right)^{1/n} \tag{3.2}$$

   donde a, m, y n son constantes determinadas experimentalmente. Los parámetro pueden variar, pero con frecuencia están en el rango de a = 1, m = 2, n =

2. El valor del atributo saturación de agua está dado en porciento (%).

4. **pay flag**: este atributo establece la existencia o no de petróleo, y toma valor de 1 o 0. En algunos trabajos como el de Ramsay et al. [92], el atributo *pay flag* se toma como indicador de un proxy para una restricción espacial de facies geológicas en el modelo FCP y se emplea para distinguir las zonas de esquisto económicas de las no económicas. Este atributo toma valores de 0 o 1, cuando toma valor 0 significa que no se encontró petróleo en esa profundidad y si toma valor 1 es que sí se encontró petróleo en esa profundidad. Para cada pozo se tienen los valores de *pay flag*, los cuales tienen valor 1 para determinados valores de profundidad. Estos valores pueden variar para cada pozo, en función de las características geológicas de los mismos [22].

5. **rayos gamma (Gr)**: este atributo es utilizado como indicador de la calidad del yacimiento. En varios estudios, se busca predecir la respuesta de rayos gamma a partir de atributos sísmicos, de modo de establecer la relación entre este atributo y los atributos sísmicos.

6. **porosidad efectiva (PHIE):** se calcula a partir del volumen de lutita

7. **porosidad efectiva de neutrón (NPHIE):** se calcula a partir del volumen de lutita

8. **volumen de capa de arcilla (VCL):** uno de los parámetros clave que se utilizan para corregir los cálculos de la porosidad y la saturación de agua, debido a los efectos que produce la arcilla unida al agua sobre los valores de estas propiedades [33]. Para determinar el volumen de arcilla de una formación de rocas a una profundidad determinada pueden ser utilizadas fórmulas matemáticas que utilizan como entrada uno (indicadores simples) o dos(indicadores dobles) registros de pozo. Entre los registros de pozo que son indicadores simples de arcilla se encuentran los registros de Rayos Gamma, Neutrón y Resistividad [33]. La unidad de medida de este atributo es en porciento (%).

9. **volumen de agua Bruta (BVW, Bulk Volumn Water):** que se define como el producto de la saturación de agua y la porosidad, es uno de los parámetros importantes en yacimientos. Se emplea como indicador de homogeneidad de hidrocarburos y yacimientos para definir el estado de la roca, ya sea homogéneo o no, determinar la permeabilidad del yacimiento e indicar efecto de suavidad [72].

10. **net pay:** esta variable se calcula a partir de [22]:

    - Porosidad efectiva (PHIE)
    - Saturación del agua (SW)
    - Volumen de arcilla (VCL)
    - Permeabilidad (K) aceptable 0.01 (unidad: Millidarcy (mD)).

    La suma de estos parámetros se conoce como *net pay*, si algunas partes del pozo cumplen esto, se coloca una bandera verde limón en el carril de porosidad, no obstante obtener el valor de esta variable es complicado debido al número de factores como la dirección de la perforación, pues un pozo perforado direccionalmente necesita que las profundidades medidas logarítmicamente se conviertan en un espesor vertical verdadero, lo que explica la desviación del pozo además de la inmersión y el impacto de la formación. Otro factor es una inadecuada porosidad y permeabilidad, pues debe haber suficiente porosidad efectiva con permeabilidad de la formación o yacimiento para que las tasas de producción de petróleo o gas, o ambas, estén en niveles comercialmente viables.

    En general valores del atributo net pay relativamente grande conduce a una mejor productividad [99]. La unidad de medida de este atributo es feet (ft) y comunmente se llama espesor de net pay.

3.4 Comportamiento de algunos de los atributos

En el caso de estudio de este trabajo, para todos los pozos se tienen los valores de la profundidad, el volumen de capa de arcilla (VCL), la porosidad Efectiva (PHIE), saturación del agua (SW), BVW, permeabilidad (KTIX) y el PayFlag. En algunos casos se tienen los datos de los valores de la permeabilidad, y en otro caso el *net pay*, así como los atributos rayos gamma, la porosidad, la densidad, la saturación del agua, la resistividad somera y profunda, la porosidad de neutron y el caliper, entre otros.

Para el caso de este trabajo la profundidad está dada en metros, y todas las características que fueron medidas en cada pozo están registradas siguiendo la secuencia escalonada determinada por la profundidad [22].

A continuación, algunas imágenes sobre el comportamiento de algunos atributos para el caso de estudio de este trabajo.

La siguiente figura, muestra el comportamiento del atributo porosidad efectiva (ver Figura 3.1).

En la imagen anterior se aprecia el comportamiento del atributo porosidad efectiva (PHIE) que alcanza un valor mínimo de 0.0001 y un máximo 0.1925. Para estos rangos de valores de este atributo, se considera que este pozo presenta una baja porosidad.

Otro ejemplo del comportamiento de los atributos, en este caso la permeabilidad (ver Figura 3.2).

A continuación, se presenta la matriz de correlación que ilustra la relación entre cada par de variables para el caso del Pozo 433.

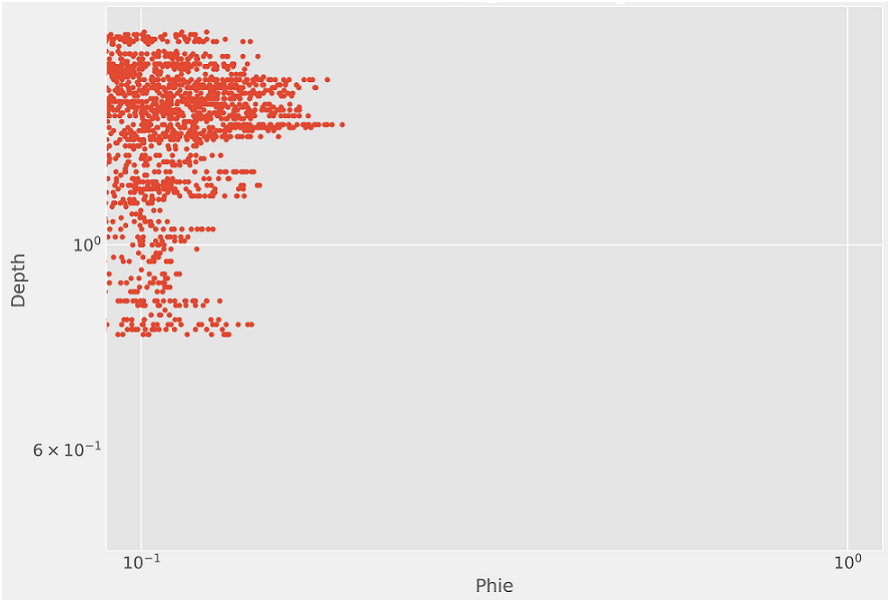


Figura 3.1: Comportamiento del atributo PHIE

Fuente: Elaboración propia

En la imagen anterior sobre la matriz de correlación se observa que de las correlaciones mostradas, las más altas con valor de 1, aparecen entre:

- las variables DTds y GRds
- las variables DTds y RHOBds
- las variables PHIE y las variables GRds, RHOBds y DTds
- las variables SW y las variables GRds, RHOBds y DTds
- las variables BVW y las variables GRds, RHOBds y DTds
- las variables DPHI Calc y las variables GRds, RHOBds y DTds
- las variables KTIX y las variables GRds, RHOBds y DTds

La mínima correlación que para todos los casos tiene un valor de -0.94, es entre:

- las variables NPHI y CALI

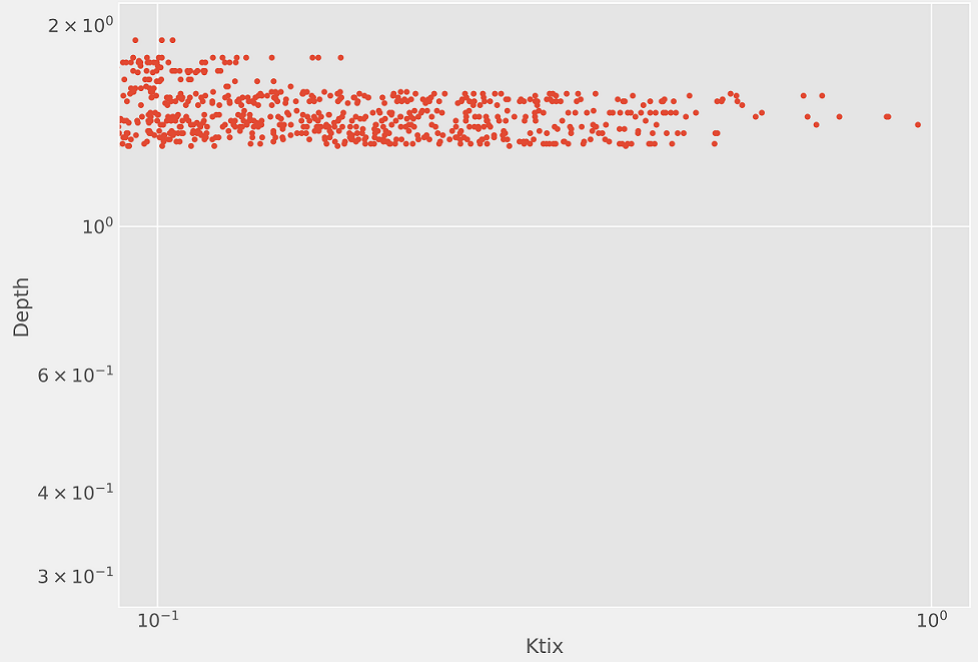


Figura 3.2: Comportamiento del atributo KTIX

Fuente: Elaboración propia

- las variables NPHI y RHOBds
- las variables NPHI y DTds
- las variables PHIE y NPHI
- las variables SW y NPHI
- las variables BVW y NPHI
- las variables DPHI Calc y NPHI
- las variables KTIX y NPHI

3.5 Sobre el caso de estudio

En el caso de estudio existe un desbalance de clases entre el 93 y el 99.5 % en promedio, de los registros en todos los datasets tienen la variable categórica *pay flag*

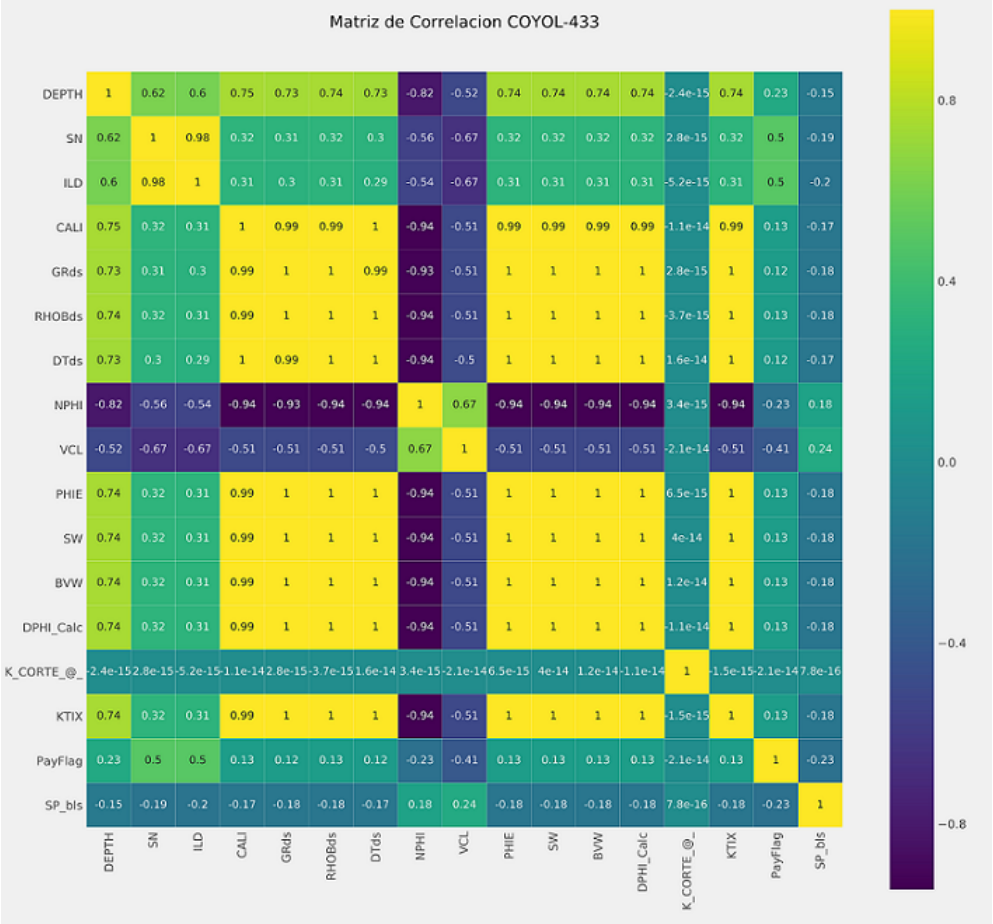


Figura 3.3: Matriz de correlación entre todos los parámetros del pozos 433

Fuente: Elaboración propia

con valor 0, y entre el 7 al 0.5 % aproximadamente de los registros, pertenecen a la clase donde el *pay flag* toma valor 1.

El desbalance de las clases se aprecia en la figura a continuación (ver Figura 3.4):

En la figura anterior se aprecia que no existe una homogeneidad en la cantidad de variables geológicas para cada pozo, esto hace complejo el dataset para aplicar los métodos de aprendizaje automatizado, en algunos pozos llegan a contar con 112 variables y en otros pozos solo 17 variables.

A continuación, se presenta la cantidad de registros con el *pay flag* con valor

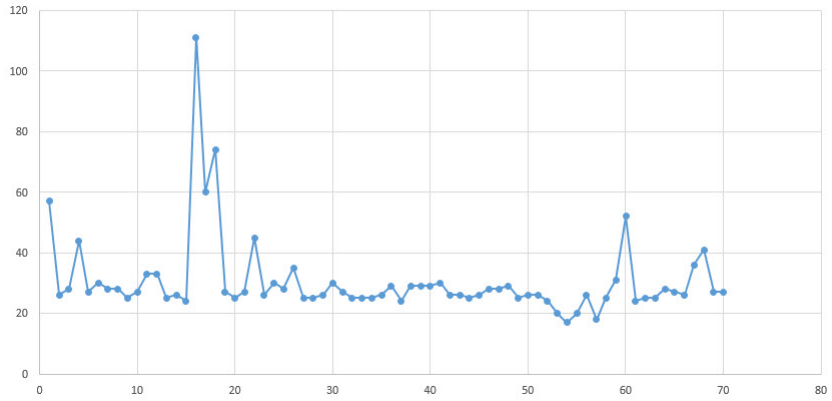


Figura 3.4: Comportamiento de la cantidad de variables por pozo

0, denotado por el color azul de las barras verticales, así como la cantidad de registros con valor 1, denotado por el color naranja de las barras verticales, en donde cada barra corresponde a un pozo. Notar la heterogeneidad en general que existe, evidenciando el desbalance de las clases (ver Figura 3.5).

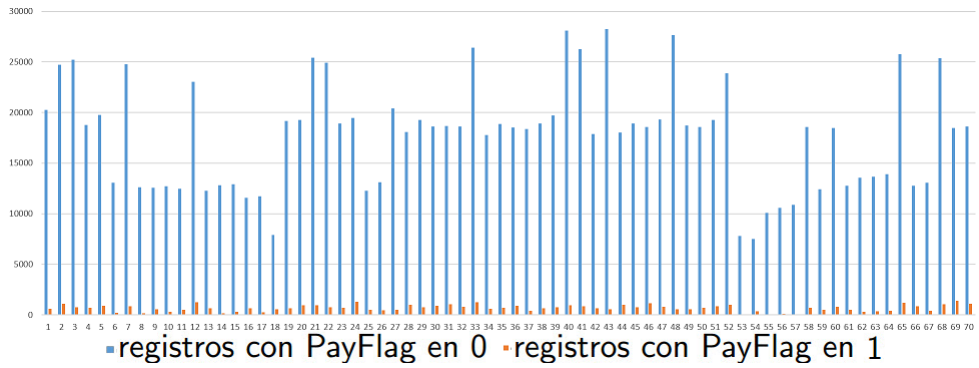


Figura 3.5: Cantidad de registros con pay flag en 0 y en 1 por pozo

Fuente: Elaboración propia

Tomando en cuenta las características de los datos, se han aplicado los algoritmos de aprendizaje automatizado del grupo No supervisados. Dentro de los algoritmos de este grupo, se encuentran los algoritmos de agrupamiento DBSCAN y *K-Means*.

3.6 Retos

Dentro de los retos más importantes en este trabajo está dado por la estructura de los datos con los que se cuenta. A continuación, se reflejan los aspectos más importantes de la estructura de los datos:

- En cada pozo se pueden presentar estructura-características diferentes y por lo tanto las mediciones pueden tener variables diferentes.
- El data set global que contiene los datos de todos los pozos alcanza más de un millón de registros.
- El data set global, es complejo por su estructura (con una cantidad diferente de variables).
- Los ejemplos en los cuales no hay petróleo son más del 90 % en todos los pozos.
- Existen muchos datos faltantes, en promedio por encima del 40 % de N/As (Valores faltantes).

En la siguiente figura 3.6, del caso de estudio de este trabajo se observa que existe una marcada presencia de datos faltantes denotada por el color claro en la imagen, esto es uno de los aspectos que convierte en un problema difícil la caraterización para resolver por los algoritmos del estado del arte.

La siguiente figura 3.6 refleja la heterogeneidad de los datos.

En la figura anterior el color rojo oscuro denota los valores observados. Notar que predomina el color claro en toda la matriz de datos.

En base al atributo *pay flag* se quiere realizar una caracterización general de la existencia o no de petróleo. Se desea establecer cómo influye la geología del terreno en el resultado del payflag, de modo que sirva para establecer bajo qué condiciones se puede encontrar petróleo cuando se haga una perforación. Es importante destacar en

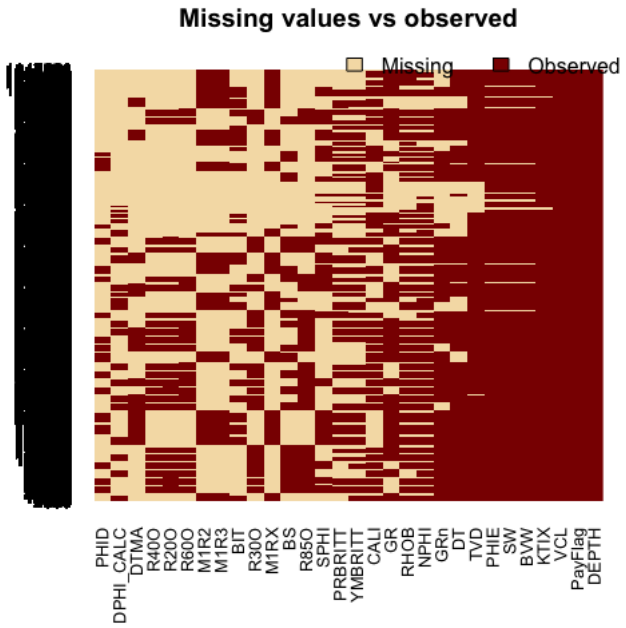


Figura 3.6: Gráfico que refleja la cantidad de registros faltantes

Fuente: Elaboración propia

cuanto a la profundidad que hay pozos que están más cerca de la superficie y pozos que están más alejados. Otra característica a destacar de los datos de los pozos es que para cada metro de profundidad se realizan varias mediciones [22].

3.7 Costos

Las investigaciones y análisis de los expertos en la etapa de exploración son costosas y dependen de la complejidad del estudio, por ejemplo, para el caso de estudios sísmicos típicamente varía entre $ 10000 (simple, marino) a $ 40000 (complejo, tierra) por kilómetro cuadrado de adquisición en 3D, y $ 5000 - $ 15000 por kilómetro cuadrado de procesamiento. Por otra parte, los pozos costa afuera resultan extremadamente costosos (en el Mar del norte el costo de los pozos típicamente está en el orden de los $ 20 millones de dólares), mientras que la perforación en tierra es mucho más económica [58].

La perforación de un pozo implica una inversión importante, que va desde unos pocos millones de dólares para un pozo en tierra, hasta cifras de alrededor de $ 100 millones de dólares más para un pozo de exploración en aguas profundas [58].

Figura 3.7: Comportamiento de los costos en un proyecto típico de exploración [58].

En la figura 3.5 se muestra como aumentan los costos durante la etapa de exploración de pozos de petróleo. En la figura, en el eje Y se encuentran los costos (cifras en millones de dólares) y en el eje X el período que comprende esta fase que puede variar de 2 a 4 años en un proyecto [58].

3.8 Conclusiones

En este capítulo queda determinado el problema en cuestión de forma precisa con vistas a realizar la caracterización de los pozos a partir de datos geológicos, así como también quedan explicados los atributos más importantes que se toman en cuenta para ello.

Capítulo 4

Metodología

4.1 Introduccion

En este capítulo se describe de forma detallada la metodología propuesta y cada una de sus fases.

La metodología que se ha concebido, que es la propuesta en este trabajo, se ha desarrollado teniendo en cuenta otras metodologías, en este caso la Metodología General de Apoyo a la Decisión Multicriterio **(MGAD)** que fue propuesta por el investigador Simon, H. (1960) [102] y la metodología fundamental para la ciencia de datos de IBM **(MFCD)**propuesta por Rollins, J. [93]. Además, se han tomado en cuenta los aspectos obtenidos a partir de la revisión de la literatura realizada.

Las fases proceden de forma secuencial, con bucles de retroalimentación a medida que el usuario vuelve a una etapa anterior antes de avanzar nuevamente.

En las figuras 4.1 y 4.2 se muestran las fases de la metodología propuesta y como se combinan las fases de MGAD y MFCD.

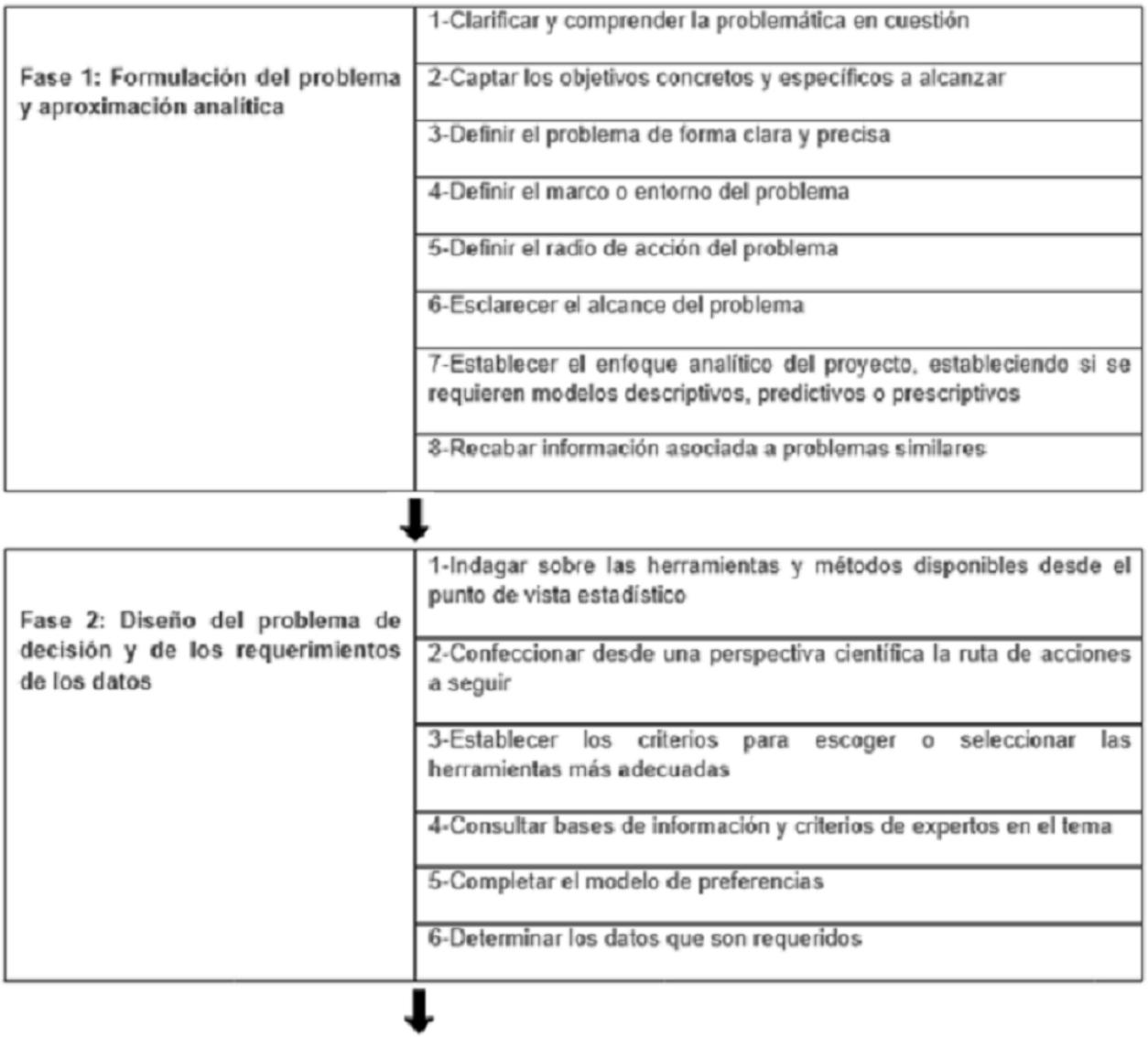


Figura 4.1: Fases 1 y 2 de la metodología propuesta

Fuente: Elaboración propia

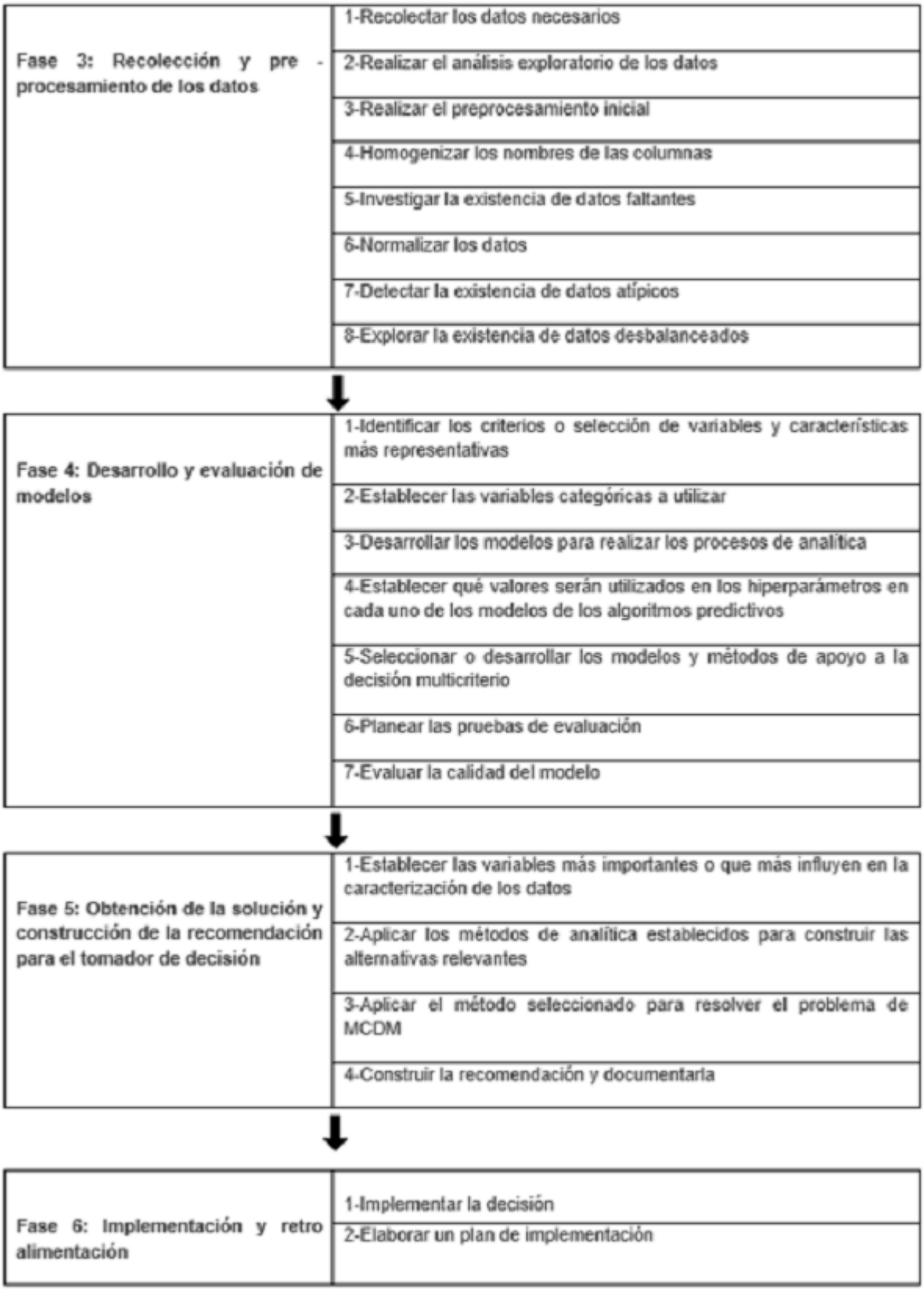


Figura 4.2: Fases 3, 4, 5 y 6 de la metodología propuesta

Fuente: Elaboración propia

4.2 Metodología propuesta

4.2.1 **Fase 1: Formulación del problema y aproximación analítica.**

El propósito de esta fase es propiciar un entendimiento del negocio en el que se origina el problema de decisión, así como formular el problema identificando sus elementos fundamentales desde ambas perspectivas (*Multicriteria Decision Making*, MCDM y Ciencia de datos, CD). Y finalmente desarrollar una aproximación analítica para identificar el enfoque que se seguirá para desarrollar el proyecto desde la perspectiva de la ciencia de datos.

Al finalizar esta fase se tendrá: (i) entendimiento del negocio, (ii) del problema, (iii) la formulación del problema y sus elementos fundamentales y (iv) se define el enfoque analítico a seguir para desarrollar la solución al problema desde la perspectiva de la ciencia de datos.

1. -**MFCD, MGAD**: Clarificar y comprender la problemática en cuestión, a través de la búsqueda y adquisición de información necesaria y conveniente que tributen a la resolución de la misma.
2. -**MGAD**: Captar los objetivos concretos y específicos a alcanzar mediante entrevistas a los actores del negocio, usuarios, jefes o entes involucrados.
3. -**MFCD, MGAD**: Definir el problema de forma clara y precisa.
4. -**MFCD, MGAD**: Definir el marco o entorno del problema.
5. -**MFCD, MGAD**: Definir el radio de acción del problema.
6. -**MFCD, MGAD**: Esclarecer el alcance del problema.

7. -**MFCD**: Establecer el enfoque analítico del proyecto, estableciendo si se requieren modelos descriptivos, predictivos o prescriptivos.

8. -**MFCD**: Recabar información asociada a problemas similares y buscar puntos en común que permitan esclarecer la problemática.

4.2.2 Fase 2: Diseño del problema de decisión y de los requerimientos de los datos

El propósito de esta fase es, una vez definido el problema y el enfoque analítico a emplear en la fase anterior, identificar los requerimientos de datos y las fuentes para iniciar la recolección de mismos. Estos pasos obedecen a los requerimientos del problema planteados por ambas metodologías.

Al finalizar esta fase se tendrá: (i) Requerimientos de datos para satisfacer ambas perspectivas (MCDM y CD), (ii) Fuentes de datos requeridos identificadas, (iii) Diseño de las alternativas del problema, así como de los criterios o atributos y otros elementos del problema de decisión que se aborda.

1. -**MGAD**: Indagar sobre las herramientas y métodos disponibles desde el punto de vista estadístico, de complejidad, rapidez de implementación que permitan escoger aquellas que sean las idóneas para la problemática ya definida anteriormente.

2. -**MGAD**: Confeccionar desde una perspectiva científica la ruta de acciones a seguir.

3. -**MGAD**: Establecer los criterios para escoger o seleccionar las herramientas más adecuadas.

4. -**MGAD**: Consultar bases de información y criterios de expertos en el tema.

5. -**MGAD**: Completar el modelo de preferencias, que se refiere a plasmar las preferencias del tomador de decisiones, establecidas como relaciones entre las distintas variables.

6. -**MFCD**: Una vez que se tienen las herramientas para análisis, determinar los datos que son requeridos en cuanto a sus características, volumen, para aplicar las herramientas anteriormente escogidas.

4.2.3 Fase 3: recolección y pre-procesamiento de los datos

El propósito de esta fase es recolectar los datos necesarios para realizar el proyecto según los requerimientos establecidos en la fase anterior. Realizar un análisis exploratorio de los mismos, para familiarizarse con estos. Así como pre-procesar los datos para garantizar la calidad requerida para que los modelos arrojen resultados confiables y de calidad.

Al terminar esta fase se tendrán: (i) Conjuntos de datos disponibles para el análisis, (ii) Conocimiento básico sobre los datos, su estructura y complejidad, así como sus propiedades principales. Lo que sirve para realizar la planeación del pre-procesamiento, el desarrollo de modelos y el análisis para establecer las conclusiones acorde a los objetivos planteados, (iii) Los datos validados, estandarizados o transformados según se requiera listos para ser usados en los modelos que se desarrollaran en las fases siguientes.

1. -**MFCD**: -Llevar a cabo una recolección de los datos pertinentes y captar las características de los mismos (estructura, volumen, interpretación, etc).

2. -**MFCD**:-Realizar el **análisis exploratorio de los datos (Exploratory Data Analysis, EDA)** como parte de la preparación de los datos, es decir conocer cómo se distribuyen los datos, saber si existen anomalías, comprobar rangos, entre otros, y para ello se visualizan las correlaciones entre los datos

que reflejan el comportamiento de las variables en cada dataset y se identifican todas las variables comunes entre todos los pozos para luego integrar todos los registros en un mismo dataset.

3. -**MFCD**:-Realizar el pre-procesamiento inicial que incluye el filtrado y limpieza de los datos. Esta actividad abarca:

4. -**MFCD**: Homogenizar los nombres de las columnas.

5. -**MFCD**:-Investigar la existencia de datos faltantes (*missing values*) (aquellos registros con valor NaN) y en caso de ser necesario, investigar los métodos existentes para afrontar esta situación de acuerdo a la literatura y aplicar el método de imputación de datos, ya sea si los mismos es posible calcularlos a partir de otros datos o se pueden rellenar a partir del valor promedio de la columna a la que pertenecen, analizar la posibilidad de reducir el dataset a las columnas o variables más relevantes que constituyen las características presentes en todos los datasets, o aplicar otro método apropiado.

6. -**MFCD**:-Normalizar los datos, lo que permite realizar una adecuada comparación entre diferentes características para asegurar que tienen el mismo impacto. Esto se refiere a que el valor de las variables esté en el mismo rango de valores y por lo tanto tengan una influencia intrínseca similar en el modelo analítico.

7. -**MFCD**:-Detectar la existencia de datos atípicos (*outliers*), y en caso de existir, investigar los métodos existentes para afrontar esta situación de acuerdo a la literatura, y luego aplicar el método apropiado para el tratamiento de estos valores.

8. -**MFCD**:-Explorar la existencia de datos desbalanceados, así como el impacto de esta situación en la calidad de la solución y de ser requerido, investigar las técnicas para balancear datos existentes de acuerdo a la literatura y aplicar la técnica apropiada.

4.2.4 Fase 4: Desarrollo y evaluación de modelos

El propósito de esta fase es planificar y desarrollar evaluaciones de los modelos para garantizar la calidad y eficiencia de los mismos al aplicarlos a los datos del proyecto.

Al finalizar esta fase se tendrá (i) implementación de los modelos desarrollados. (ii) Modelo de preferencias del tomador de decisiones (iii) Modelos y métodos para resolver los problemas de decisión establecidos. (iv) Modelos validados que arrojan resultados confiables y en tiempos aceptables.

1. -**MFCD**: Proceder a la identificación de los criterios o selección de variables y características más representativas mediante la técnica del Análisis de Componentes Principales (Principal Component Analysis, PCA), tomando en cuenta la varianza de los datos, lo cual permite reducir la dimensión de los mismos.

2. -**MFCD**: Establecer las variables categóricas que serán utilizadas.

3. -**MFCD**:-Desarrollar los modelos para realizar los procesos de analítica acorde con los objetivos definidos anteriormente.

4. -**MFCD**: Establecer qué valores serán utilizados en los hiperparámetros en cada uno de los modelos de los algoritmos predictivos de aprendizaje automático que serán utilizados para ajustar el modelo.

5. - **MGAD**: Seleccionar o desarrollar los modelos y métodos de apoyo a la decisión multicriterio, que se emplearan para darle solución al problema de decisión multicriterio formulado y la generación de la recomendación.

6. -**MFCD, MGAD**: Planeación de las pruebas de evaluación.

7. -**MFCD, MGAD**: Evaluar la calidad del modelo, esto se refiere a realizar una evaluación y análisis de los resultados a partir de las métricas de calidad y desempeño seleccionadas previamente de forma tal de obtener un modelo

con una alta precisión y sensibilidad, y contrastar los resultados para determinar la validez de los mismos. Estos resultados obtenidos por los modelos son consultados con los especialistas que toman las decisiones.

4.2.5 Fase 5: Obtención de la solución y construcción de la recomendación para el tomador de decisión

El propósito de esta fase es obtener la solución para el problema formulado en el proyecto, acorde a los objetivos planteados. Empleando los modelos seleccionados o desarrollados y generar una recomendación para el tomador de decisiones acorde a la aproximación analítica planificada y las preferencias del tomador de decisiones establecidas.

Al finalizar esta fase se tendrá: (i) La solución al problema planteado, (ii) recomendación de decisión para el tomador de decisiones desde la perspectiva de MCDM.

1. -**MFCD**: Establecer las variables más importantes o que más influyen en la caracterización de los datos de los pozos de petróleo.

   En este paso quedan esclarecidos por parte de los especialistas cuales son aquellas variables más importantes o que más aportan en la explicación de la variabilidad de los datos. Se toman en cuenta las variables que son las más importantes de acuerdo a la literatura consultada, y además las variables que se cuentan en los datos para realizar la caracterización.

2. - **MFCD**: Aplicar los métodos de analítica establecidos para construir las alternativas relevantes para el problema de decisión multicriterio.

3. -**MGAD**: Aplicar el método seleccionado para resolver el problema de MCDM.

4. -**MGAD**: Construir la recomendación y documentarla, lo cual permite mejorar a posteriori el modelo haciéndolo escalable y más robusto mediante mejoras.

4.2.6 Fase 6: Implementación y retroalimentación

El propósito de esta fase es implementar con éxito la decisión (considerando la recomendación construida en base a la solución del problema desarrollada). Para ello hay que hacer un plan de implementación que contemple, entre otros: la socialización de la implementación de la decisión con todos los actores involucrados, explicar a todos clara y detalladamente las consecuencias para la empresa y sus empleados, capacitar (de ser necesario) a aquellos actores que tengan que hacer cambios en sus actividades o rutinas, re-asignar o asignar recursos, etc. A la vez se planifica la retroalimentación para realizar ajustes en el periodo de implementación para que una vez concluido este la decisión tomada tenga impacto positivo y duradero en la organización.

Al terminar esta fase se tendrá (i) el problema resuelto, con la decisión implementada y creando un impacto positivo en la organización. (ii) Un plan de retroalimentación en un horizonte acordado para cualquier actualización de la situación o adecuación requeridas.

El alcance de este trabajo no abarca esta fase de implementación y retroalimentación.

4.3 Conclusiones

En este capítulo ha quedado descrita de forma explícita la metodología propuesta, detallando cada una de sus fases.

Capítulo 5

Experimentación y resultados

Este capítulo se describe acorde a las fases de la metodología propuesta. Además, se describen las especificaciones del software y hardware utilizado.

5.1 Fase 1: Formulación del problema y aproximación analítica

El contenido que corresponde a esta fase está plasmado en el capítulo 1 y 2 de este trabajo. En esta primera fase aspectos relacionados con la comprensión de la problemática, definición del problema, objetivos, marco teórico, radio de acción y alcance del problema han quedado definidos en los capítulos 1 y 2 del presente trabajo, así como la descripción de problemas similares se encuentra en el capítulo 2.

5.2 Fase 2: Diseño del problema de decisión y de los requerimientos de los datos

El diseño del problema de decisión y de los requerimientos de los datos están reflejados en el capítulo 3 de este trabajo.

5.3 Fase 3: Recolección y preprocesamiento de los datos

De los datos recolectados se utilizaron 3 datasets correspondientes a los pozos 433, 2654, 6076 para aplicar los algoritmos seleccionados.

La selección de estos dataset se realizó tomando en cuenta el criterio del dataset con la menor cantidad y con la mayor cantidad de registros, así como el dataset con una cantidad media de registros.

Tabla 5.1: Cantidad de registros para cada dataset seleccionado.

Pozo	**Total de registros**	**Total de columnas**
433	7876	19
2654	19458	28
6076	26416	41

Fuente: Elaboración propia

Los registros en los 3 datasets fueron ajustados de acuerdo a la variable Depth, en el rango de 600 a 1500, debido a que se encontró en ese rango la mayor concentración de valores de PayFlag (PF) en 1, como queda reflejado en la figura 5.1

A continuación, se muestra el ajuste a los 3 datasets escogidos.

La tabla 5.2 anterior muestra los 3 pozos seleccionados luego del ajuste to-

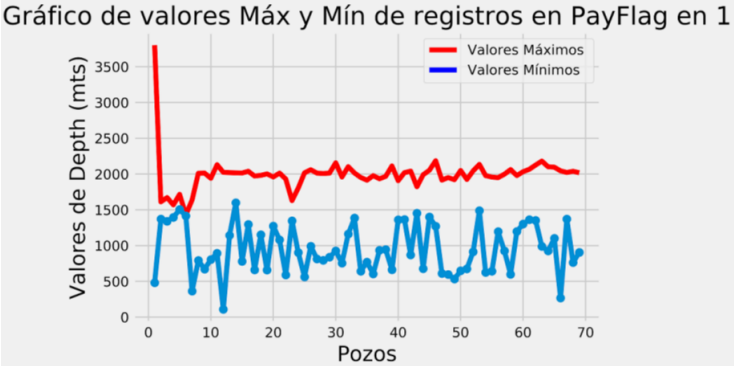


Figura 5.1: Valores máximos y mínimos de Depth para cada pozo

Fuente: Elaboración propia

Tabla 5.2: Datasets ajustados

Pozo	**Registros antes**	**Registros después**	**Reducción (%)**	**PFs antes**	**PFs después**
433	7876	5906	25.07	366	249
2654	19458	9002	53.74	922	144
6076	26416	11812	55.29	1058	303

Fuente: Elaboración propia

mando en cuenta la Depth en el rango de 600 a 1500, por ser el rango en el que predominan los registros que tienen el PayFlag en 1. La columna **PFs antes**, significa los registros con PayFlag en 1 antes de hacer la reducción, y **PFs después**, significa los registros con PayFlag en 1 una vez realizada la reducción.

Posteriormente, se realizó el filtrado y la limpieza de los datos, como parte del pre-procesamiento inicial que incluyó: homogenizar los nombres de las columnas, identificar los valores perdidos, detectar la existencia de datos atípicos, explorar la existencia de datos desbalanceados, normalizar los datos, entre otros. En el caso de los valores perdidos, estos valores faltantes están dados por el dominio del conocimiento del problema, por lo que los registros con el valor -999.25 indica que la medición no es válida a esa profundidad lo que equivale a registros con valor NaN.

A continuación, se presentan los resultados del análisis de los valores faltantes.

En las figuras 5.2, 5.3 y 5.4, aparecen representadas todas las características de cada uno de los pozos, siendo las barras de color oscuro que van desde el valor 0.0 a 1.0 (escala ubicada en el lado izquierdo de las gráficas) aquellas columnas o características en donde la información está completa y por otro lado en aquellas características donde faltan datos las barras de color oscuro son más cortas. En el lado derecho de las gráficas los valores representan la cantidad de registros que va desde 0 a 5906 para el caso del pozo 433, etc. De igual modo, en la tabla 5.3 queda expresado en porciento las cantidades faltantes en cada pozo.

Tabla 5.3: Datos faltantes asociados a los 3 pozos seleccionados.

IDPOZO	**Total de registros**	**Cantidad columnas afectadas**	**% datos faltantes**
433	5906	8	22.1
2654	9002	3	77.8
6076	11812	9	66.7

Fuente: Elaboración propia

A continuación, las matrices de los datos faltantes para cada pozo.

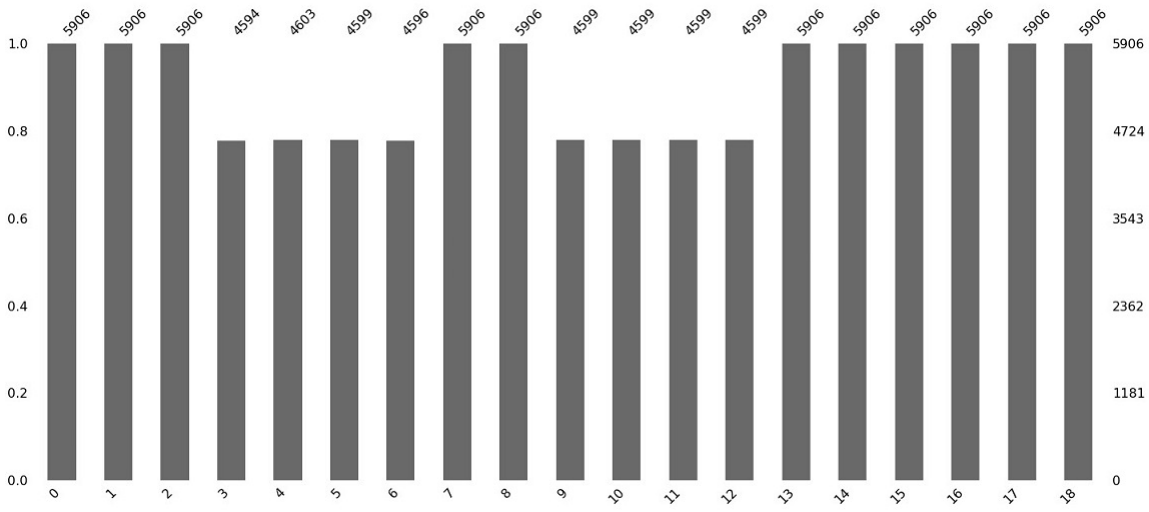


Figura 5.2: Datos faltantes Pozo 433

Fuente: Elaboración propia

De las figuras 5.2, 5.3, 5.4, como se ha comentado antes las barras de color oscuro denotan los datos existentes para cada columna o característica, siendo las barras de color oscuro las características en las que existen datos faltantes, cada barra corresponde a una característica. En el caso del pozo 433, existen 8 columnas afectadas y en cada una de ellas falta el 22.1 % de los datos, es decir existen 1307 registros faltantes en cada columna, para el caso del pozo 2654, existen 3 columnas afectadas y en cada una de ellas falta el 77.8 % lo que equivale a 7000 registros faltantes en cada columna y para el pozo 6076, existen 9 columnas afectadas, en 8 de ellas falta el 66.7 % de los datos, es decir existen 7873 registros faltantes en cada columna, y en la columna DTCO-VPVS el 99.9 % de los registros posee una medición errónea.

A continuación, precisiones de los datos faltantes:

En el caso del pozo 433, todos los registros de las columnas afectadas CALI, GRDS, RHOBDS, DTDS, PHIE, SW, BVW, DPHI-CALC corresponden a la clase mayoritaria donde el PayFlag vale 0.

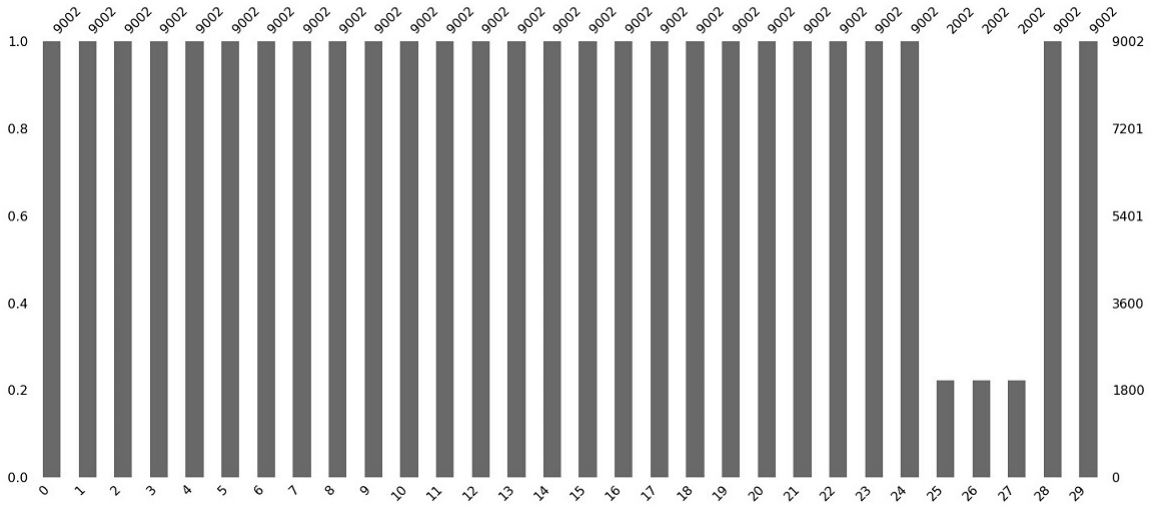


Figura 5.3: Datos faltantes Pozo 2654

Fuente: Elaboración propia

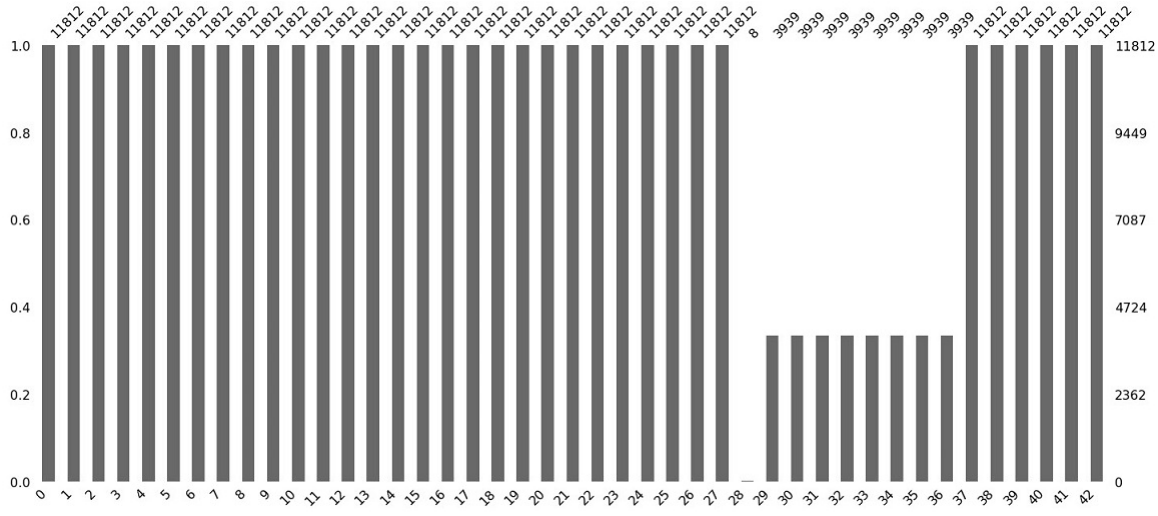


Figura 5.4: Datos faltantes Pozo 6076

Fuente: Elaboración propia

En el caso del pozo 2654, en las columnas afectadas YOUNG, PR y BRITT en las que faltan datos, de los 144 registros con PayFlag en 1, 10 registros corresponden a datos faltantes en esas mismas 3 columnas, y los otros 134 registros son de mediciones válidas.

En el caso del pozo 6076, la columna donde falta el 99.9 %, fue eliminada, y en las otras 8 (DTCO-VPVS, GAS, VEL-PERF, LUTITA, LIMOLITA, ARENISCA, MUDSTONE, BENTONITA y MARGA), todos los registros corresponden a la clase mayoritaria donde el PayFlag vale 0.

En el pozo 6076 las columnas K-CORTE, LIMOLITA, MUDSTONE, BENTONITA, MARGA fueron eliminadas por quedar con valor 0 en toda su columna.

Se aplicó en las variables de naturaleza cuantitativa el método de imputación con el promedio.

A continuación, el diagrama de cajas como resultado del análisis exploratorio (EDA) para los 3 pozos.

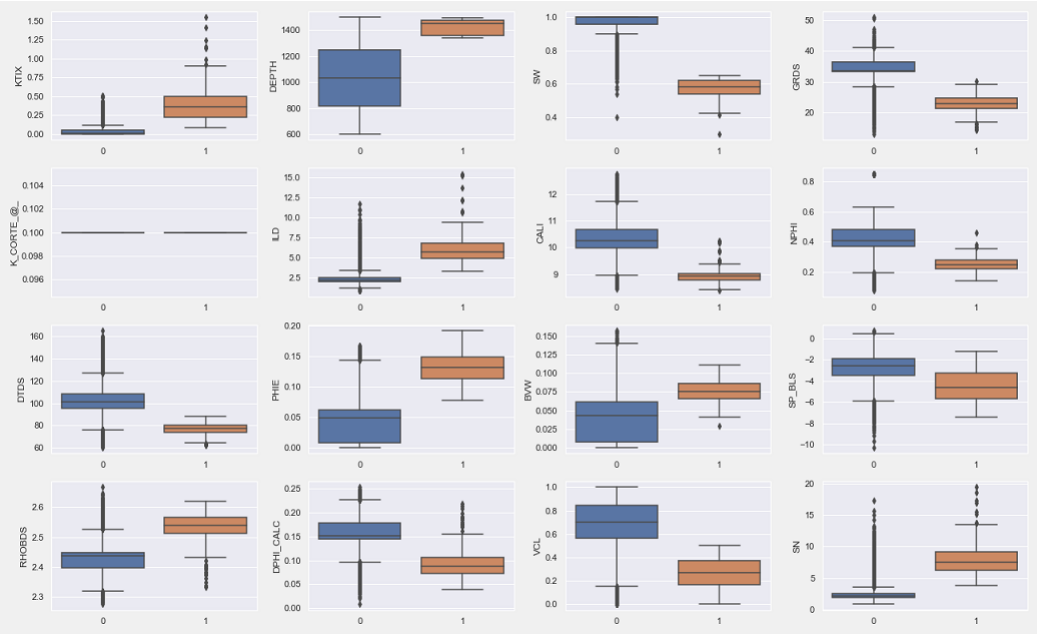


Figura 5.5: Pozo 433

Fuente: Elaboración propia

A continuación, el EDA para el caso del pozo 2654.

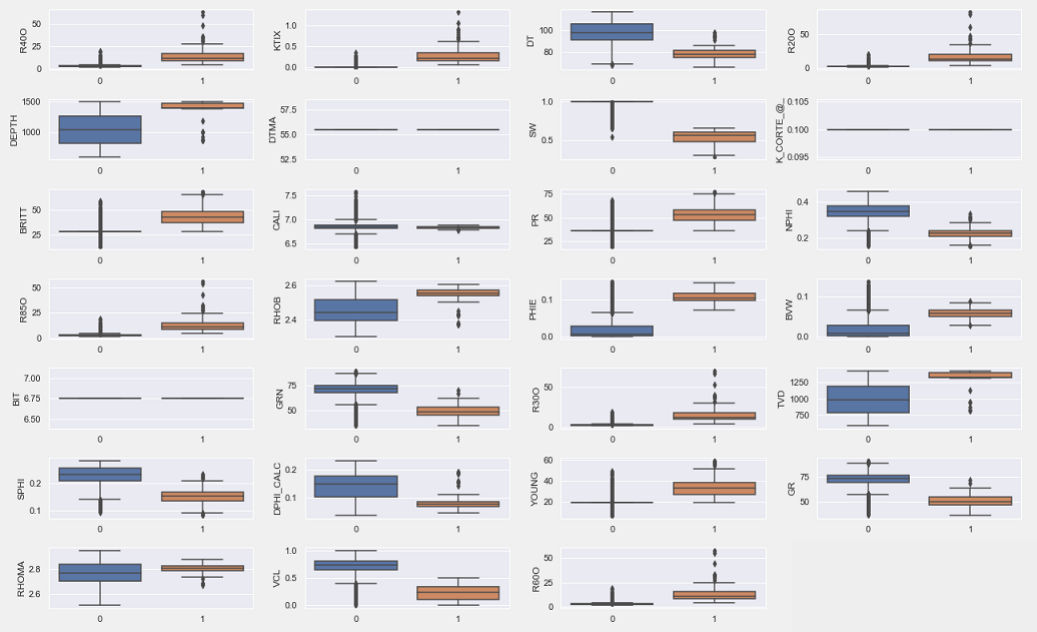


Figura 5.6: Pozo 2654

Fuente: Elaboración propia

A continuación, el EDA para el caso del pozo 6076.

En las figuras 5.5, 5.6 y 5.7 se observa a través de los diagramas de cajas que hay ciertas variables que presentan valores atípicos (*outliers*). En el caso del pozo 6076 principalmente en las variables KTIX, M1R1, VPVS, M1R2, M1R3, M1RX, POISDIN, M1R6, aparecen los valores atípicos. En el caso del pozo 2654 aparecen los valores atípicos principalmente en las variables CALI, PR, R300, YOUNG. En el caso del pozo 433 aparecen los valores atípicos principalmente en las variables BRITT, CALI, PR, PHIE, BVW. Para evitar la influencia de estos valores atípicos se trataron estos por el método de Tukey.

Los valores atípicos pueden sesgar una distribución de probabilidad y dificultar el escalado de datos mediante la estandarización, ya que la media calculada y la desviación estándar se verán sesgadas por la presencia de valores atípicos. Estos valores atípicos pueden afectar a los algoritmos como Regresión logística, entre otros

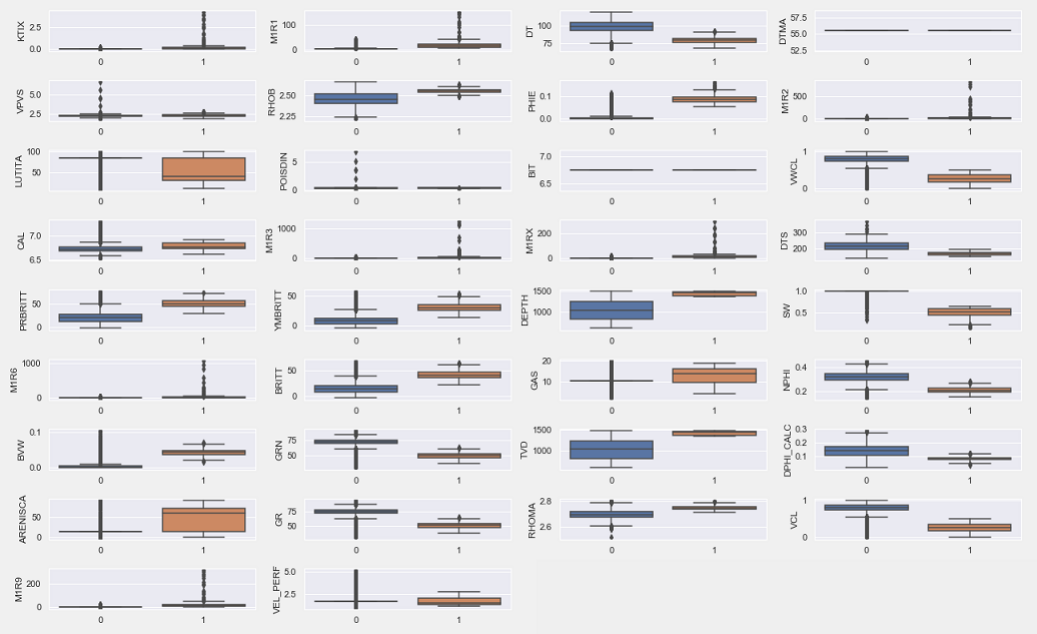


Figura 5.7: Pozo 6076

Fuente: Elaboración propia

algoritmos.

A continuación, se presenta la fig 5.8 que muestra los valores atípicos en el pozo 433 antes del tratamiento de *outliers*.

Fig 5.9 Valores atípicos en el pozo 433 después de tratamiento de *outliers*.

En las 2 figuras 5.8 y 5.9 se aprecia el antes y después de tratar los valores atípicos, en donde se aprecia una reducción de estos valores. Se procede de igual forma en los pozos 2654 y 6076.

Datos atípicos para el pozo 6076 en las figuras 5.12 y 5.13

Al término de esta parte ya los datos han quedado con todos los datos atípicos tratados en los 3 pozos, de modo que se pasa a balancear los mismos.

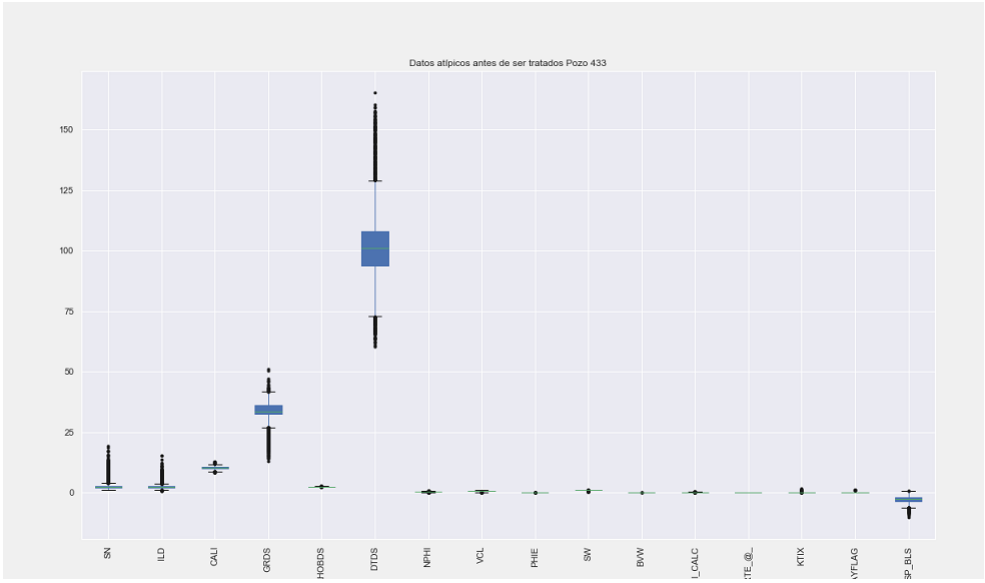


Figura 5.8: Pozo 433 antes

Fuente: Elaboración propia

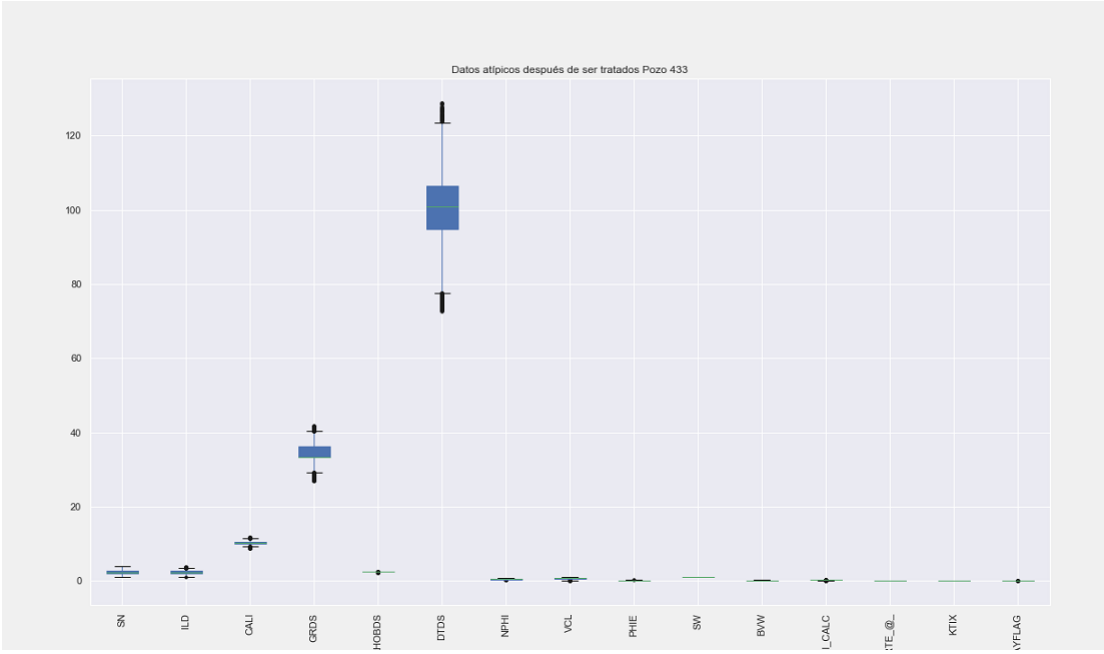


Figura 5.9: Pozo 433 después

Fuente: Elaboración propia

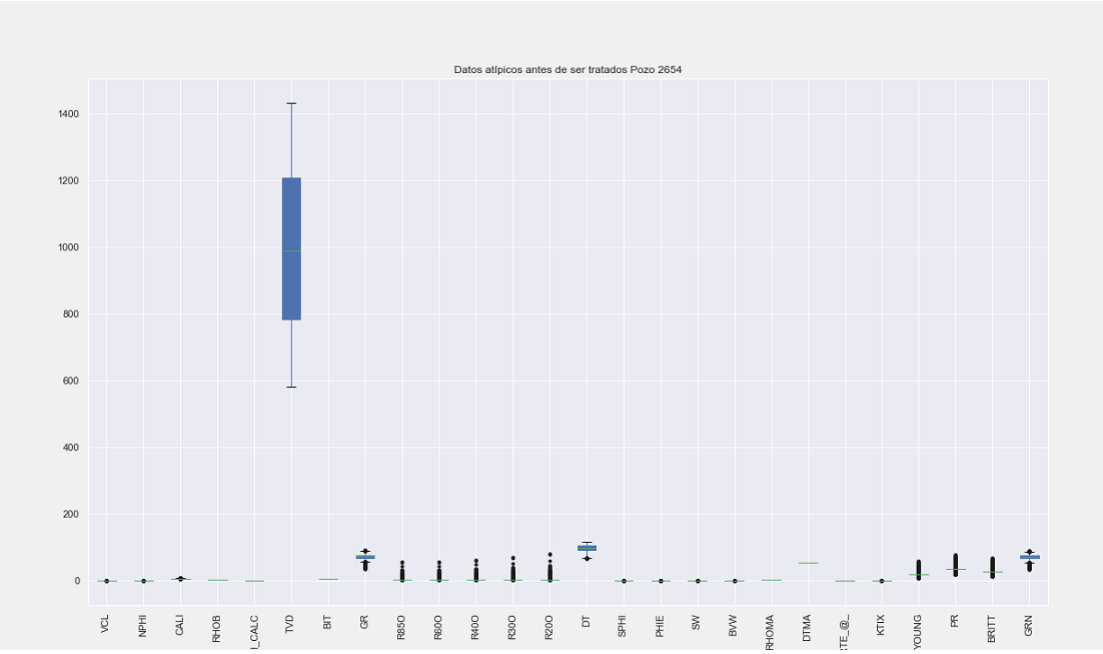


Figura 5.10: Pozo 2654 antes

Fuente: Elaboración propia

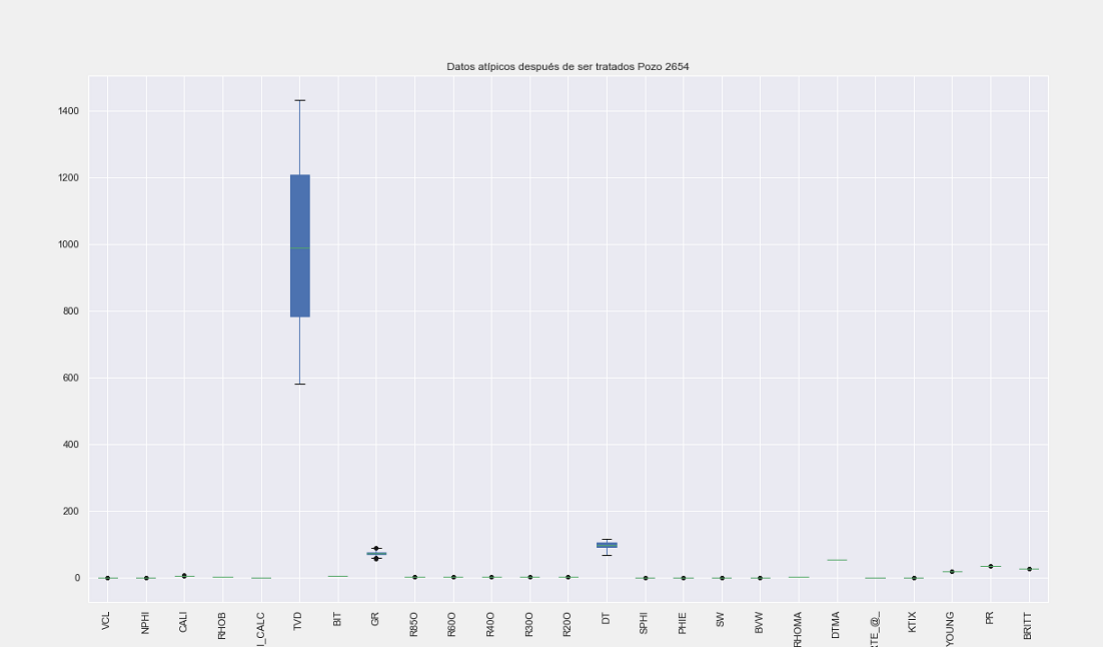


Figura 5.11: Pozo 2654 después

Fuente: Elaboración propia

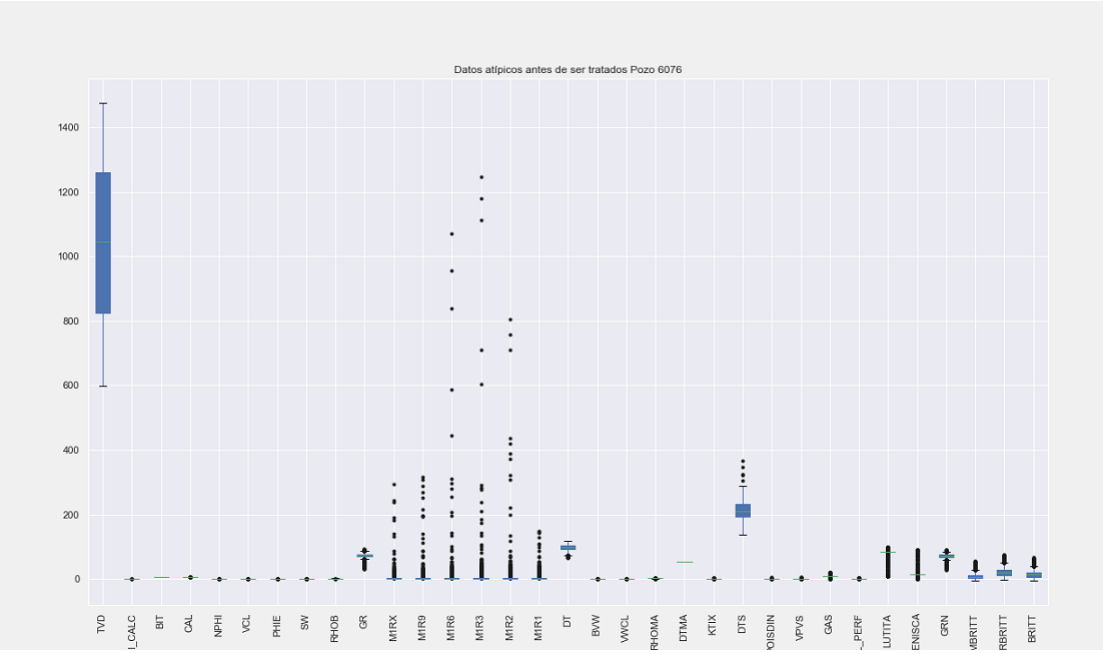


Figura 5.12: Pozo 6076 antes

Fuente: Elaboración propia

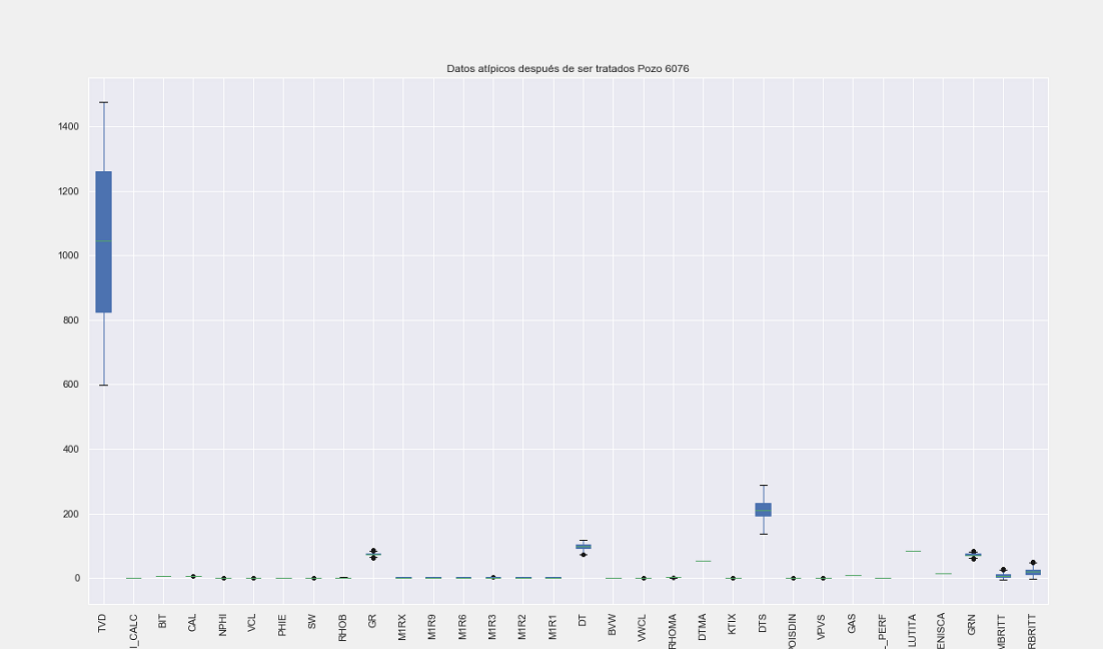


Figura 5.13: Pozo 6076 después

Fuente: Elaboración propia

5.3.1 Balanceo de las clases

Luego se balancearon los datos de los 3 pozos como se aprecia en las figuras 5.14, 5.15 y 5.16

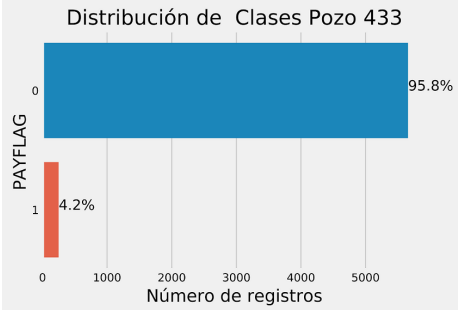


Figura 5.14: Pozo 433.

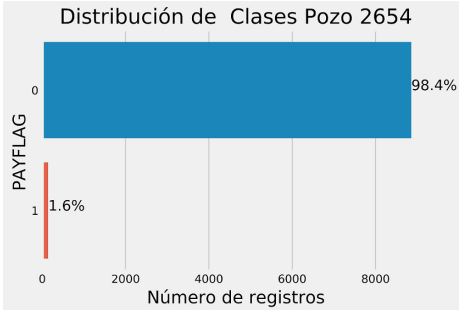


Figura 5.15: Pozo 2654. Fuente: Elaboración propia

Para el caso del pozo 6076 los valores atípicos es como se muestra.

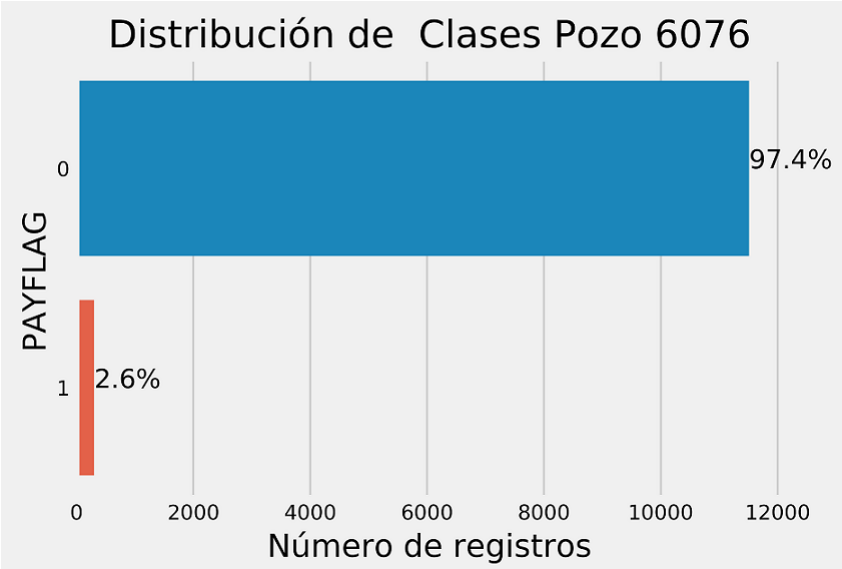


Figura 5.16: Pozo 6076

Fuente: Elaboración propia

A continuación de forma resumida la distribución de las clases en los 3 pozos en la tabla 5.4.

Distribución de la variable PayFlag según las clases 0 y 1

Tabla 5.4: Distribución de la variable PayFlag según las clases 0 y 1.

IDPOZO	**Frecuencia Clase 1**	**Frecuencia Clase 0**	**% Clase 0**	**% Clase 1**
433	249	5656	95.8	4.21
2654	144	8857	98.4	1.59
6076	303	11508	97.4	2.56

Fuente: Elaboración propia

En las figuras 5.14, 5.15, 5.16 y en la tabla 5.4 se puede apreciar la distribución de las clases en los pozos seleccionados. Se puede observar que existe una marcada diferencia entre las clases 1 y 0 en los 3 pozos, lo cual queda reflejado por el valor del porcentaje, pues para el pozo 433 la clase 1 representa el 4.21 %, para el pozo 2654 la clase 1 representa el 1.59 %, y para el pozo 6076 la clase 1 representa el 2.56 %. Esto es lo que denota la presencia de los datos desbalanceados, lo cual puede conllevar a un grado de sobre ajuste en los algoritmos de Regresión Logística y AdaBoost, y además se ven afectados los indicadores de precisión, por tal motivo se precisa aplicar uno o varios métodos de balanceo de clases.

A continuación, se muestran en las tablas 5.5, 5.6 y 5.7 los datos antes y después de aplicar los métodos de balanceo.

Tabla 5.5: Datos del pozo 433 antes y después del muestreo para equilibrarlos.

Métodos de muestreo	**Clase 0**	**Clase 1**	**Total**
Sin muestreo	4518	206	4724
Sobre-muestreo (down)	4518	4518	9036
Sub-muestreo (up)	206	206	412

Fuente: Elaboración propia

A continuación, de manera gráfica en la figuras 5.17, 5.18 y 5.19 para el pozo

Tabla 5.6: Datos del pozo 2654 antes y después del muestreo para equilibrarlos.

Métodos de muestreo	**Clase 0**	**Clase 1**	**Total**
Sin muestreo	7088	112	7200
Sobre-muestreo (down)	7088	7088	14176
Sub-muestreo (up)	112	112	224

Fuente: Elaboración propia

Tabla 5.7: Datos del pozo 6076 antes y después del muestreo para equilibrarlos.

Métodos de muestreo	**Clase 0**	**Clase 1**	**Total**
Sin muestreo	9199	249	9448
Sobre-muestreo (down)	9199	9199	18398
Sub-muestreo (up)	249	249	498

Fuente: Elaboración propia

433 el resultado de los muestreos.

A continuación, de manera gráfica en la figuras 5.20, 5.21 y 5.22 para el pozo 2654 el resultado de los muestreos:

Los muestreos obtenidos para el pozo 2654.

A continuación, de manera gráfica en la figuras 5.23, 5.24 y 5.25 para el pozo 6076 el resultado de los muestreos:

Los muestreos obtenidos para el pozo 6076.

En las gráficas de la 5.17 a la 5.25 y las tablas de la 5.5 a la 5.7 queda reflejado el balanceo de los datos tanto por la técnica de Sub-muestreo como por el Sobre-muestreo. Con el sobre-muestreo aumentó el tamaño del dataset de entrenamiento para cada pozo, de forma tal de equilibrar o igualar la clase minoritaria, en este caso la clase 1 (registros con PayFlag en 1), es decir seleccionó registros de la clase 1 hasta igualar la cantidad de la clase 0, quedando el dataset en 9036, para el pozo

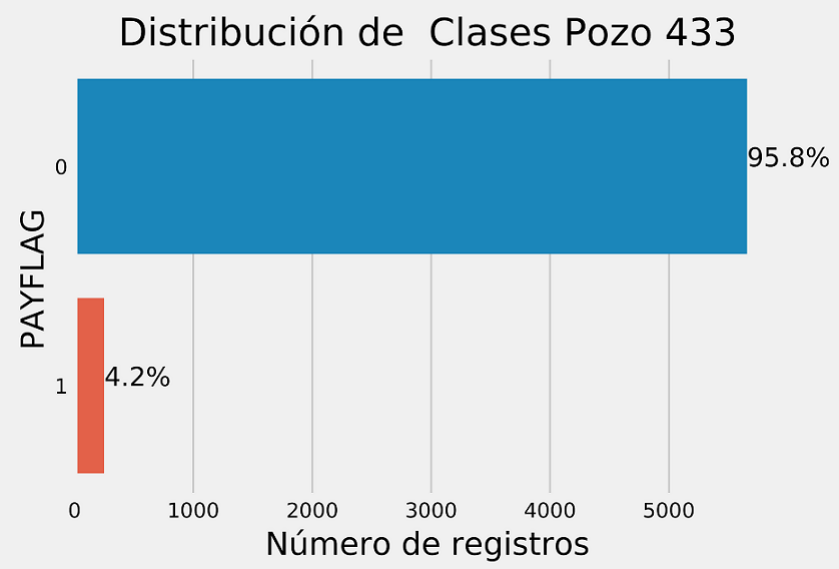


Figura 5.17: Distribución de clases Pozo 433 Sin muestreo

Fuente: Elaboración propia

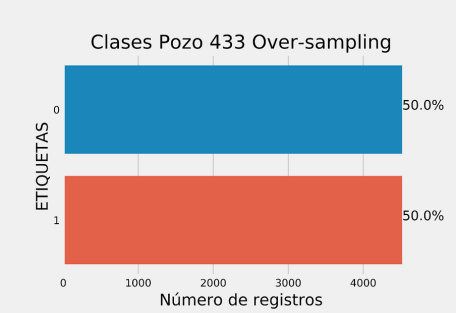


Figura 5.18: Distribución de clases Pozo 433 Sobre-muestreo.

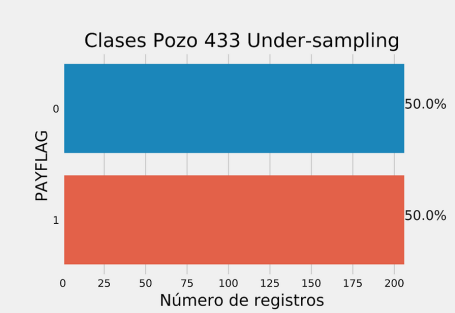


Figura 5.19: Distribución de clases Pozo 433 Sub-muestreo. Fuente: Elaboración propia

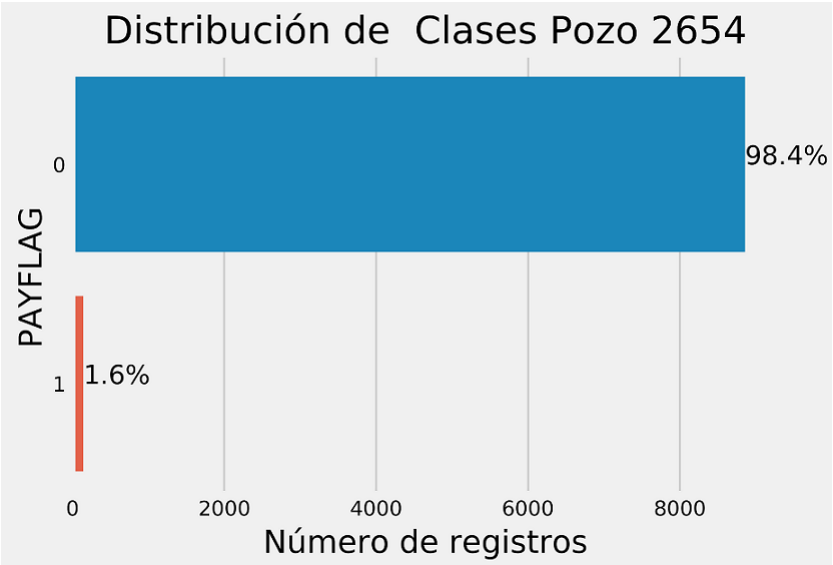


Figura 5.20: Distribución de clases Pozo 2654 Sin muestreo

Fuente: Elaboración propia

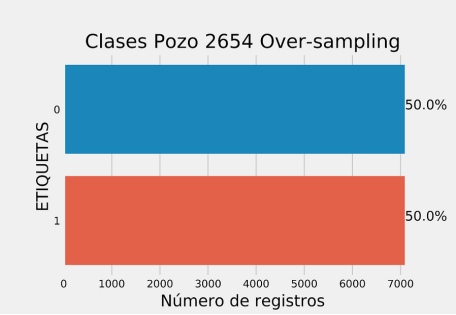


Figura 5.21: Distribución de clases Pozo 2654 Sobre-muestreo.

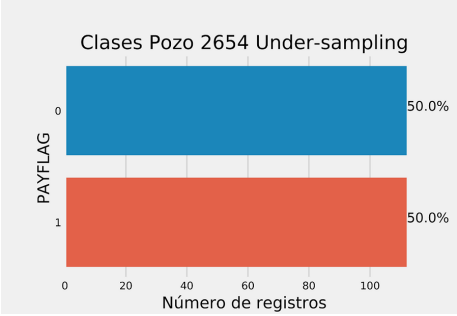


Figura 5.22: Distribución de clases Pozo 2654 Sub-muestreo. Fuente: Elaboración propia

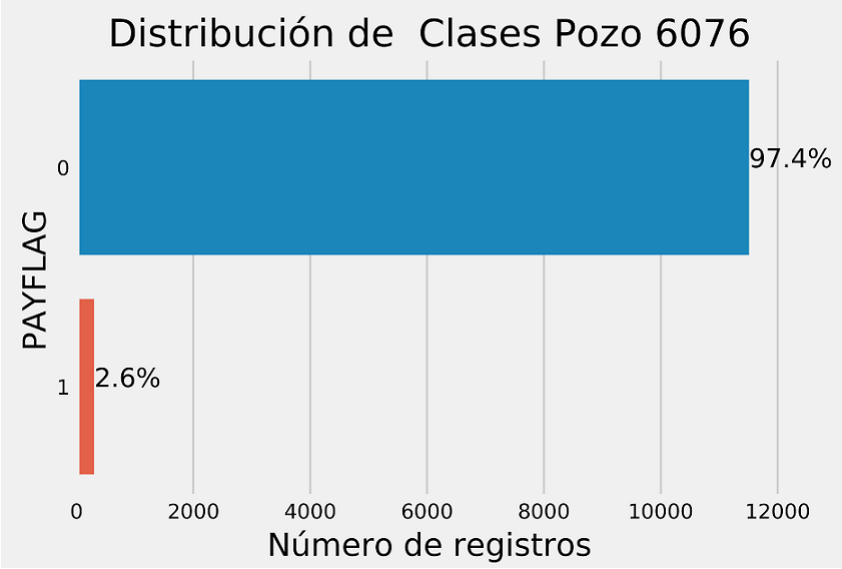


Figura 5.23: Distribución de clases Pozo 6076 Sin muestreo

Fuente: Elaboración propia

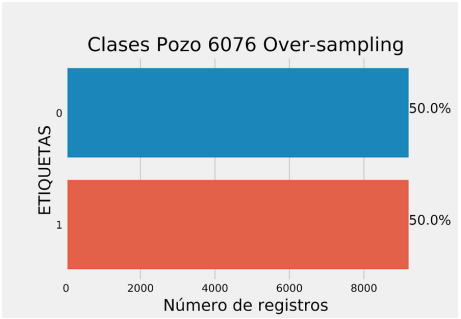


Figura 5.24: Distribución de clases Pozo 6076 Sobre-muestreo.

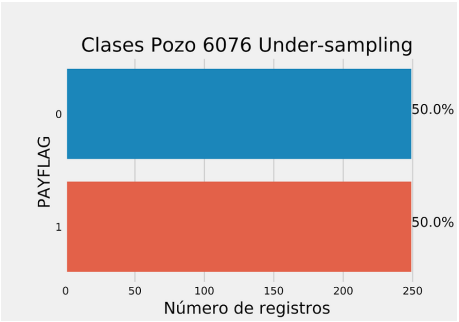


Figura 5.25: Distribución de clases Pozo 6076 Sub-muestreo. Fuente: Elaboración propia

433, 14176, para el pozo 2654, y 18398 para el pozo 6076, respectivamente.

Con el sub-muestreo sin reemplazo disminuyó el tamaño del dataset de entrenamiento para cada pozo, de forma tal que se equilibra o iguala la clase minoritaria o menos representada, en este caso la clase 1 (registros con PayFlag en 1), es decir se reducen los datos de la categoría más mayoritaria hasta igualar el tamaño de la categoría más baja, quedando el dataset en: 412 registros, para el pozo 433, 224 registros para el pozo 2654 y 498 para el pozo 6076, respectivamente.

Cabe mencionar que con la técnica del sobre-muestreo, se genera una cantidad registros repetidos dentro del dataset lo cual puede derivar en sobre entrenamiento de los algoritmos aplicados, en el caso del sub-muestreo, se pierde información significativa de la muestra lo cual puede influir en un detrimento de la precisión de los algoritmos.

5.4 Fase 4: Desarrollo y evaluación de modelos

A continuación, resultados de aplicar el algoritmo DBSCAN

Se consideraron 3 valores para el parámetro eps y 2 valores para el parámetro de muestra mínima (*min samples*) antes de aplicar este algoritmo. En el caso del eps se trabajó con los valores: 1.2, 1.5, 1.7, y para el parámetro muestra mínima los valores fueron: 20 y 50

De la tabla anterior se aprecia que los valores del pozo 433, en las combinaciones donde el parámetro toma valor de 1.5 y 1.7, el coeficiente de Silouette alcanza los valores más altos, tomando en cuenta que están en el rango de 0.5 a 0.7, lo cual corresponde a un clúster de estructura fuerte. En estos 4 casos del eps 1.5 y 1.7, del pozo 433, el DBSCAN identificó 2 clústers de etiquetas 0 y 1. Para los casos de los otros dos pozos el coeficiente de la métrica, no brinda un resultado favorable.

Tabla 5.8: Resultados completos del DBSCAN

IDPOZO	Coeficiente Silouette					
	eps(1.2)-20	eps(1.2)-50	eps(1.5)-20	eps(1.5)-50	eps(1.7)-20	eps(1.7)-50
433	0.480	0.445	**0.597**	**0.526**	**0.590**	**0.591**
2654	0.175	-	0.049	0.144	0.149	0.239
6076	-0.225	-	0.152	-	0.334	-0.037

Fuente: Elaboración propia

A continuación, se hayan las coincidencias para cada uno de los 4 casos antes mencionados en conjunto con los registros del dataset original del pozo 433, buscando los registros donde coinciden en el PayFlag en 1 y están en el clústert 1.

Tabla 5.9: Coincidencias para cada muestra seleccionada del pozo 433.

Muestra	**Coincidencias**	**Diferencias**
(eps(1.5)-20)	0	498
(eps(1.5)-50)	0	498
(eps(1.7)-20)	0	498
(eps(1.7)-50)	0	498

Fuente: Elaboración propia

A partir de la tabla anterior se puede concluir que ninguno de los registros catalogados en el dataset original con PayFlag en 1, coincide con los agrupados en el clúster 1 por el DBSCAN, en ninguna de las 4 corridas ejecutadas.

5.4.1 Resultados de aplicar el K-Means

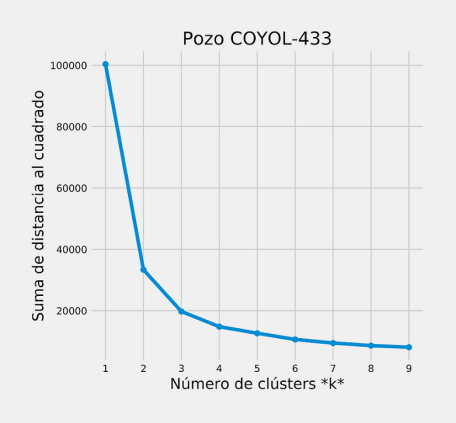
Para los 3 pozos que fueron seleccionados (433, 2654 y 6076) se aplicó el algoritmo K-Means, y se consideraron 3 valores en el parámetro cantidad de clústers (k), estos son 2, 3 y 4 clústers.

De igual modo a como se hizo para el algoritmo DBSCAN, se filtraron los registros por la variable Depth, en el rango de 600 a 1500 mts y sobre los datasets obtenidos se aplicó el K-Means. Además, antes de ejecutar el algoritmo se balancearon igualmente todos los datos en cada pozo.

Se utilizó el método del codo (*Elbow Method*) para determinar el mejor valor del parámetro (k). Este método es utilizado para este fin como queda expresado en el trabajo de Kodinariya de 2013 [62].

No obstante, los valores de k seleccionados anteriormente se aplicó el método del codo para determinar en cada uno de los 3 pozos, cuál es el valor adecuado del parámetro (k).

A continuación, se presentan las figuras 5.26, 5.27 y 5.28 del método del codo para cada pozo.



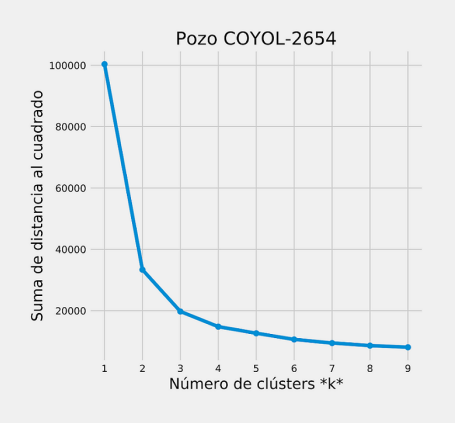


Figura 5.26: Pozo 433. Figura 5.27: Pozo 2654. Fuente: Elaboración propia

Para el caso del pozo 6076 el resultado es como se muestra en 5.28.

De las gráficas anteriores se observa que los valores más apropiados de k son 2 y 3 para los 3 pozos, en el caso del pozo 6076 además se considera el valor de 4.

Debido al interés en esta problemática basada en la caracterización de pozos de petróleo, se consideraron dos clases, la clase 0 y la clase 1, a partir de los valores

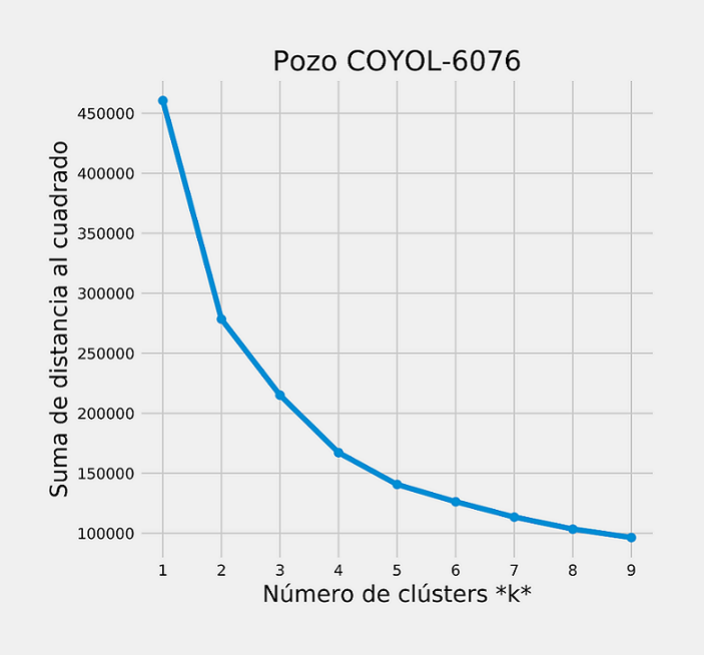


Figura 5.28: Pozo 6076

Fuente: Elaboración propia

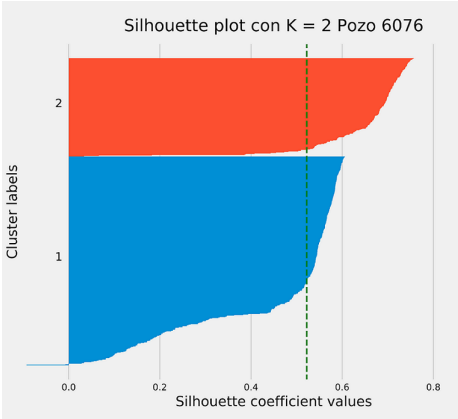
que toma la variable PayFlag que toma valor 0 o 1, en dependencia de si existe o no petróleo en esa profundidad.

A continuación, los resultados gráficos de la métrica Silhouette para 2, 3 y 4 clústers para el pozo 433 (figuras 5.32, 5.33 y 5.34) y 6076 (figuras 5.29, 5.30 y 5.31).

En la tabla 5.10 se muestra todos valores del coeficiente de Sihouette obtenido para todos los pozos.

Tabla de resultados de las métricas del K-Means.

De los gráficos anteriores y la tabla 5.7, se puede apreciar que el valor más alto del coeficiente de Silhouette corresponde al pozo 433, para el caso de $k = 2$, lo cual representa una estructura de clúster fuerte. En la figura 5.11 y 5.14, se puede apreciar que en el clúster 1 (color azul) contiene la mayor concentración de registros, esto está denotado por el grosor del color azul en la dirección del eje y. Para el caso del pozo 2654, se obtiene un valor de 0.520 para $k = 2$, y en el caso del pozo 6076



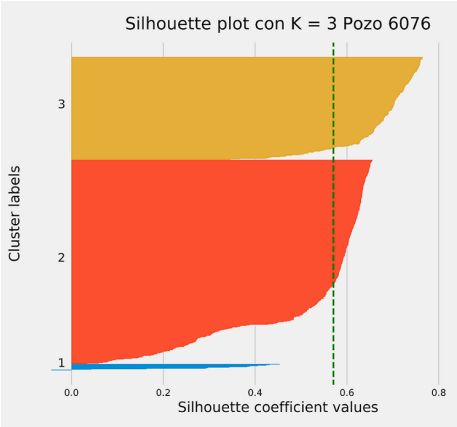


Figura 5.29: Pozo 6076. Figura 5.30: Pozo 6076. Fuente: Elaboración propia

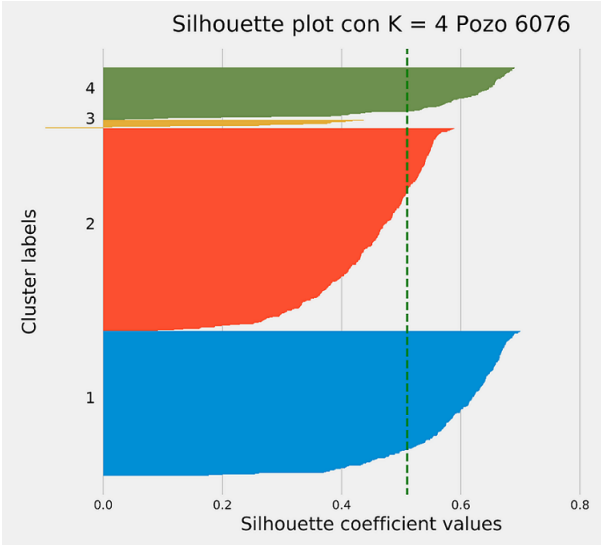


Figura 5.31: Pozo 6076

Fuente: Elaboración propia

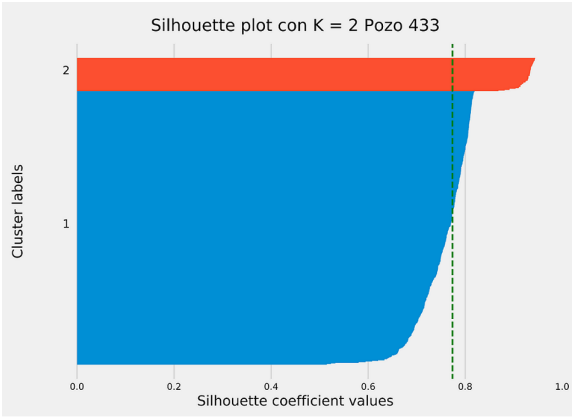


Figura 5.32: Pozo 433.

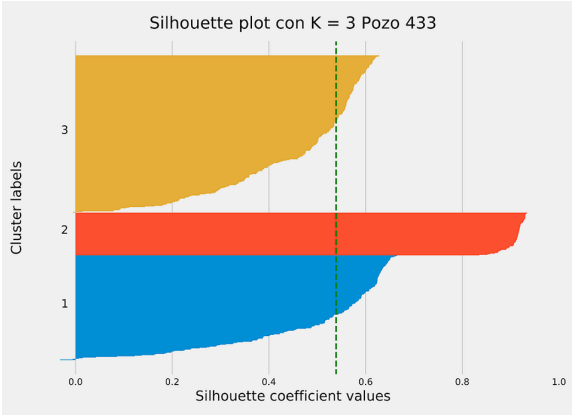


Figura 5.33: Pozo 433.

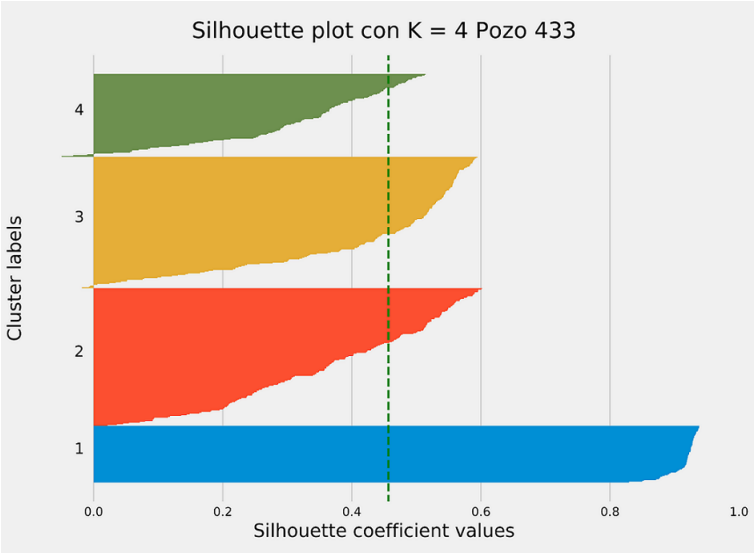


Figura 5.34: Pozo 433

Fuente: Elaboración propia

Tabla 5.10: Resultados completos del K-Means

IDPOZO	Coeficiente Silouette			Homogeneity	Completeness
	K = 2	K=3	K=4		
433	**0.773**	0.533	0.449	1	1
2654	0.520	0.411	0.442	1	1
6076	0.530	0.575	0.505	1	1

Fuente: Elaboración propia

en todas las combinaciones de clústers se obtienen registros de clúster fuertes.

A continuación, para los registros sombreados en negro en la tabla 5.11 se hallan las coincidencias, obteniéndose:

Tabla 5.11: Coincidencias para cada muestra seleccionada del K-Means.

Muestra	**Coincidencias**	**Diferencias**
433 ($k = 2$)	0	51
433 ($k = 3$)	0	59
2654 ($k = 2$)	135	9
6076 ($k = 2$)	299	114
6076 ($k = 3$)	0	214
6076 ($k = 4$)	0	206

Fuente: Elaboración propia

De la tabla 5.11 se puede observar que para el pozo 2654 en la combinación $k = 2$, se obtienen coincidencias para los registros que se encuentran en el clúster 1 y que además en el dataset original poseen el PayFlag en 1, a un rango de Depth de entre 600 a 1500. Además, en el pozo 6076 en la combinación $k = 2$, también se obtienen coincidencias para los registros que se encuentran en el clúster 1 y que además en el dataset original poseen el PayFlag en 1, a un rango de Depth de entre 600 a 1500.

Regresión Logística (RL) Binaria

Para el caso de la RL binaria, se dividen los datasets de los pozos seleccionados (433, 2654 y 6076) que previamente se filtraron por la Depth entre 600 a 1500, en dos partes: una parte para entrenar el modelo y la otra parte para probar el modelo, y luego se aplica el balanceo de clases a la parte de entrenamiento y luego se aplica el algoritmo.

Para evaluar el resultado obtenido del algoritmo RL binaria fueron seleccionadas las siguientes métricas:

- métrica Exactitud (Accuracy)
- métrica F1-score
- métrica Área dentro de la Curva Característica Operativa (ROC AUC)
- Curva ROC (Receiver Operating Characteristic Curve)

5.4.2 Datos de entrenamiento y prueba

Para aplicar los algoritmos se adoptó la proporción de 70 % de los datos destinados a entrenar los algoritmos de RL Binaria y Adaboost Ensemble y 30 % para probar esos algoritmos en cada uno de los 3 pozos seleccionados.

La forma en la que fueron divididos los 3 pozos es la que se muestra continuación.

En las 3 tablas 5.12, 5.13 y 5.14 ha quedado establecida la distribución de los datos para aplicar los algoritmos de aprendizaje supervisado que han sido seleccionados.

Tabla 5.12: Distribución de la variable PayFlag según las clases 0 y 1 en Pozo 433

Datos	**Clase 0**	**Clase 1**	**Total**	**Porcentaje**
Datos de entrenamiento	4518	206	4724	70 %
Datos de prueba	1127	54	1181	30 %

Fuente: Elaboración propia

Tabla 5.13: Distribución de la variable PayFlag según las clases 0 y 1 en Pozo 2654

Datos	**Clase 0**	**Clase 1**	**Total**	**Porcentaje**
Datos de entrenamiento	7088	112	7200	70 %
Datos de prueba	1764	37	1801	30 %

Fuente: Elaboración propia

Tabla 5.14: Distribución de la variable PayFlag según las clases 0 y 1 en Pozo 6076

Datos	**Clase 0**	**Clase 1**	**Total**	**Porcentaje**
Datos de entrenamiento	9199	249	9448	70 %
Datos de prueba	2304	59	2363	30 %

Fuente: Elaboración propia

5.4.3 Resultados del algoritmo de RL Binaria

A continuación, matrices de confusión para el pozo 433 en las tablas 5.15, 5.16 y 5.17

Tabla 5.15: Resultados con datos originales.

	Clase 0	**Clase 1**	**Total**
Clase 0	1679	19	1698
Clase 1	42	32	74
Total	1721	51	1772

Tabla 5.16: Sobre muestreo

	Clase 0	**Clase 1**	**Total**
Clase 0	1501	216	1717
Clase 1	80	1597	1677
Total	1581	1813	3394

Tabla 5.17: Sub muestreo

	Clase 0	**Clase 1**	**Total**
Clase 0	61	7	68
Clase 1	3	79	82
Total	64	86	150

A continuación, en la tabla 5.18 el resultado, antes y después del Sub-muestreo y Sobre-muestreo del pozo 433.

Tabla 5.18: Métricas de la RL Binaria del pozo 433 antes y después.

Medidas	**Sin ajuste**	**Sobre-muestreo**	**Sub-muestreo**
Exactitud (*Accuracy*)	0.9655	0.9127	0.9333
Precisión	0.6274	0.8808	0.9186
Recall	0.4324	0.9522	0.9634
F measure	0.512	0.9151	0.9404
AUC ROC	0.9797	0.9510	0.9652

Fuente: Elaboración propia

En las tablas 5.18, 5.22, 5.26 han quedado descritos los resultados del algoritmo RL Binaria aplicado. Fueron seleccionados la corridas que generaron en las métricas

Tabla 5.19: Resultados con datos originales Pozo 2654.

	Clase 0	**Clase 1**	**Total**
Clase 0	2649	4	2653
Clase 1	8	40	48
Total	2657	44	2701

Tabla 5.20: Sobre muestreo

	Clase 0	**Clase 1**	**Total**
Clase 0	2645	0	2645
Clase 1	35	2635	1677
Total	2680	2635	5315

Tabla 5.21: Sub muestreo

	Clase 0	**Clase 1**	**Total**
Clase 0	36	8	44
Clase 1	4	39	43
Total	40	47	78

el mejor valor y con estas se hallaron las coincidencias entre aquellos registros y los registros en el dataset original de los pozos, para determinar en cuáles registros coincide con PayFlag igual a 1.

A continuación, en la figura 5.35 las métricas Exactitud, Precisión, Recall, F measure y AUC el pozo 433 y 2654.

A continuación, en la figura 5.36 las métricas Exactitud, Precisión, Recall, F measure y AUC el pozo 6076.

Tabla 5.22: Aplicación de la RL binaria al pozo 2654 antes y después.

Medidas	**Sin ajuste**	**Sobre-muestreo**	**Sub-muestreo**
Exactitud (*Accuracy*)	0.9955	0.9934	0.8620
Precisión	0.9090	1.0	0.8297
Recall	0.8333	0.9868	0.9069
F measure	0.9545	0.9934	0.8666
AUC ROC	0.9869	0.9985	0.8895

Fuente: Elaboración propia

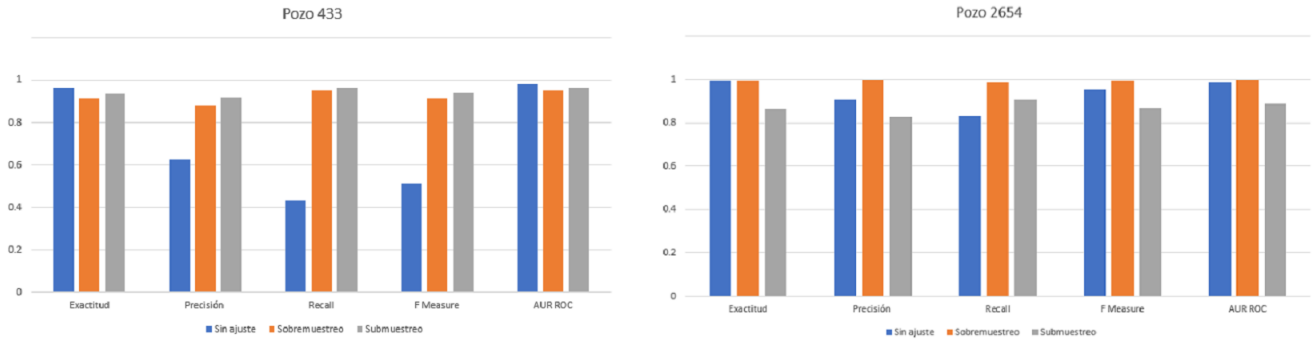


Figura 5.35: Métricas del Pozo 433 y 2654 en datos sin ajuste, sobre y sub muestreo

Fuente: Elaboración propia

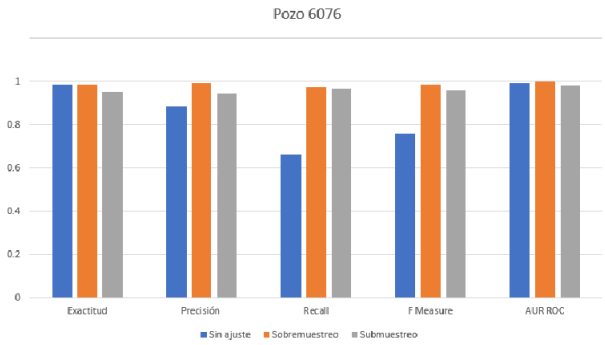


Figura 5.36: Métricas del Pozo 6076 en datos sin ajuste, sobre y sub muestreo

Fuente: Elaboración propia

Tabla 5.23: Resultados con datos originales Pozo 6076.

	Clase 0	**Clase 1**	**Total**
Clase 0	3418	10	3428
Clase 1	39	77	116
Total	3457	87	3544

Tabla 5.24: Sobre muestreo

	Clase 0	**Clase 1**	**Total**
Clase 0	3403	20	3423
Clase 1	80	3402	3482
Total	3483	3422	6905

Tabla 5.25: Sub muestreo

	Clase 0	**Clase 1**	**Total**
Clase 0	82	5	87
Clase 1	3	92	95
Total	85	97	182

A continuación, en la figura 5.37 el resultado de la métrica AUC aplicada en la RL para el Pozo 433 tanto para sobre-muestreo y sub-muestreo:

A continuación, en la figura 5.38 el resultado de la métrica AUC aplicada en la RL para el Pozo 2654 tanto para sobre-muestreo y sub-muestreo:

A continuación, en la figura 5.39 el resultado de la métrica AUC aplicada en la RL para el Pozo 6076 tanto para sobre-muestreo y sub-muestreo:

Tabla 5.26: Aplicación de la RL binaria al pozo 6076 antes y después.

Medidas	**Sin ajuste**	**Sobre-muestreo**	**Sub-muestreo**
Exactitud (*Accuracy*)	0.9861	0.9855	0.9560
Precisión	0.8850	0.9941	0.9484
Recall	0.6637	0.9770	0.9684
F measure	0.7586	0.9855	0.9583
AUC ROC	0.9958	0.9991	0.9822

Fuente: Elaboración propia

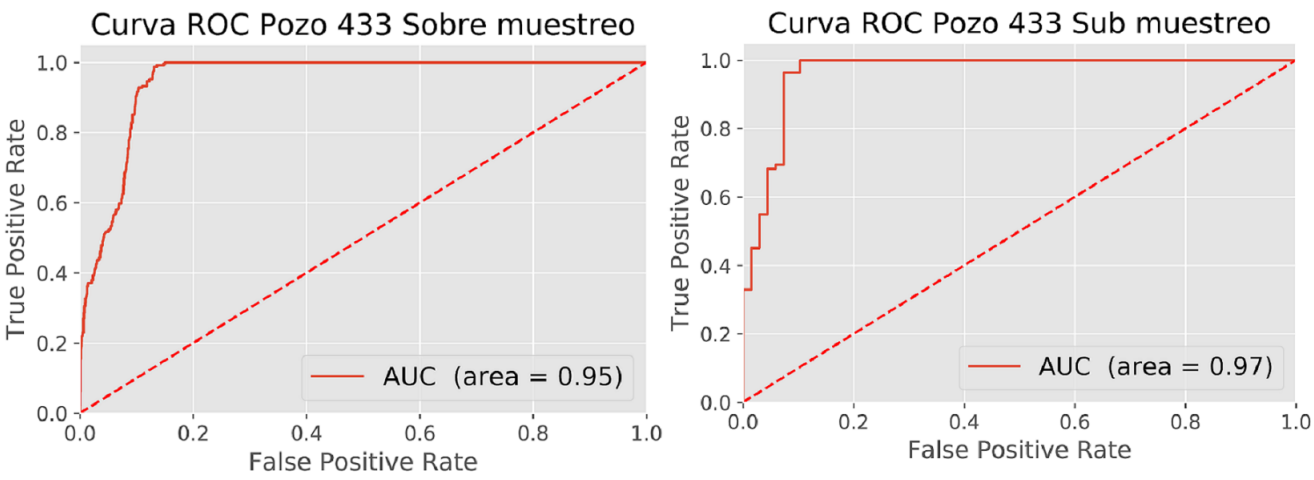


Figura 5.37: Métrica AUC al Pozo 433 Sobre y Sub muestreo

Fuente: Elaboración propia

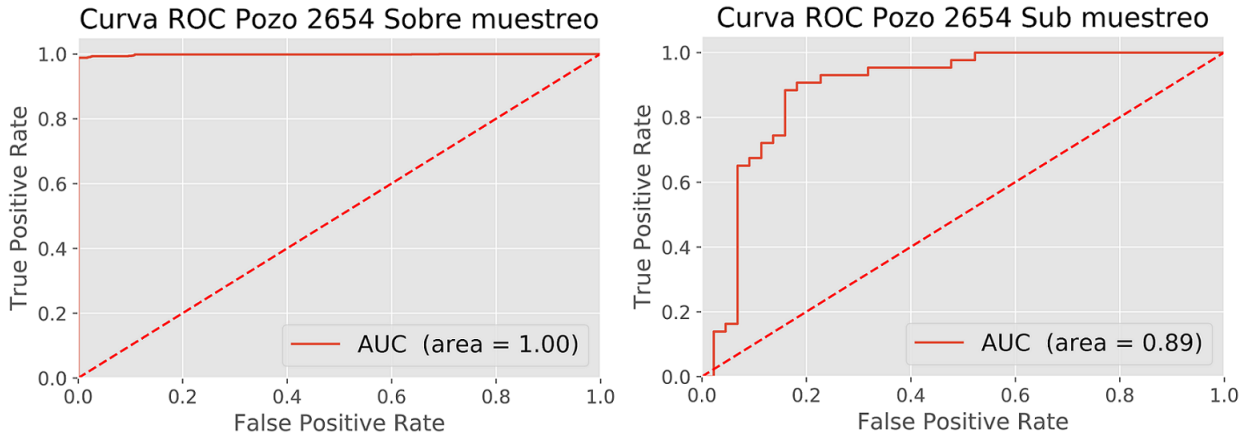


Figura 5.38: Métrica AUC al Pozo 2654 Sobre y Sub muestreo

Fuente: Elaboración propia

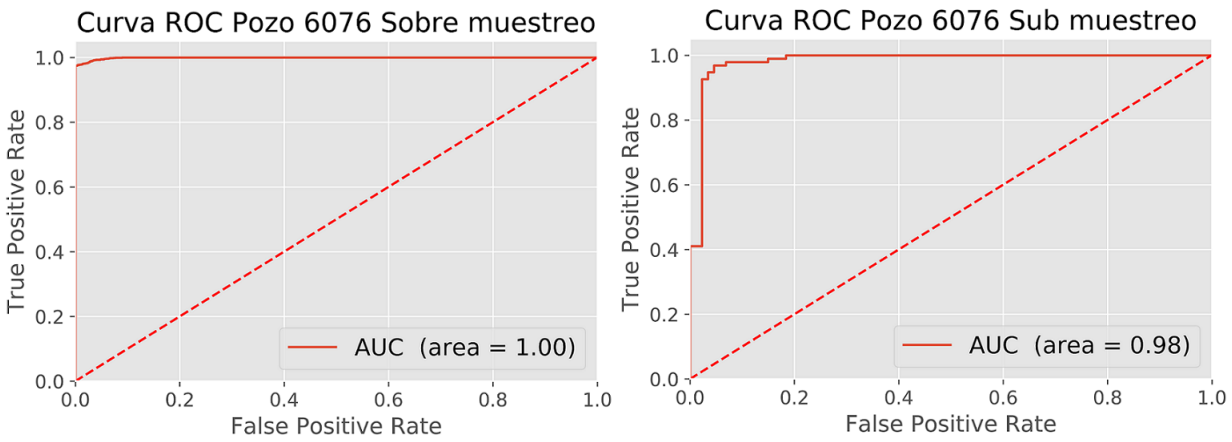


Figura 5.39: Métrica AUC al Pozo 6076 Sobre y Sub muestreo

Fuente: Elaboración propia

Las figuras 5.37, 5.38 y 5.39 anteriores muestran el valor del AUC (área bajo la curva) para la RL en cada método de muestreo con valores altos. Se observa que existe poca diferencia entre los métodos, excepto en el caso del valor del AUC para el Pozo 2654 submuestreo (0.86). Cada punto en el gráfico de la curva ROC, significa el rendimiento del algoritmo de clasificación. En estos gráficos lo deseable es contar con la mayor área posible, pues cuanto mayor sea el área, mayor es la precisión.

A continuación, la figura 5.40 muestra los resultados de la métrica AUC en forma conjunta (sobre-muestreo y sub-muestreo) aplicada en la RL para el Pozo 433 y 2654.

La siguiente, es la figura 5.41 muestra los resultados de la métrica AUC en forma conjunta (sobre-muestreo y sub-muestreo) aplicada en la RL para el Pozo 6076.

En las figuras 5.40 y 5.41 se muestran las curvas de cada pozo en conjunto. Se puede observar que en los pozos 2654 y 6076 con los métodos de sobre muestreo se obtienen los valores más altos de precisión con 0.99.

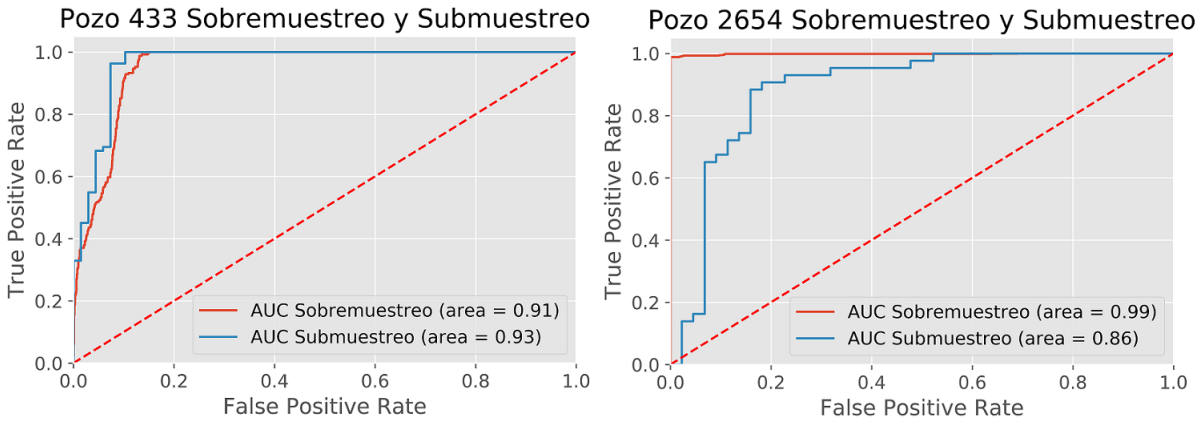


Figura 5.40: Métrica AUC en conjunto del Pozo 433 y 2654 de la RL

Fuente: Elaboración propia

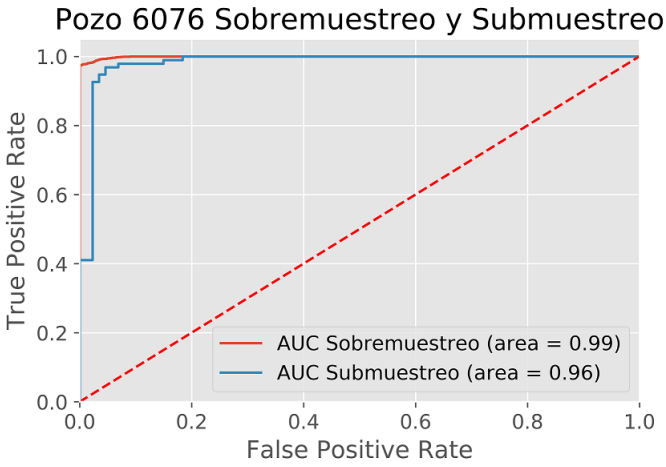


Figura 5.41: Métrica AUC en conjunto del Pozo 6076 de la RL

Fuente: Elaboración propia

5.4.4 Coincidencias

Tabla 5.27: Resultados de las coincidencias de la RL Binaria y los datos originales

IDPOZO	RL			
	Sobre-muestreo	Sub-muestreo	% Sobre-muestreo	% Sub-muestreo
433	231	77	92.77	30.92
2654	144	39	**100**	26.89
6076	295	91	97.35	30.03

Fuente: Elaboración propia

En la tabla 5.27 se observa que en el caso del pozo 2654 sobre-muestreo de las 144 registros con PayFlag en 1 en los datos originales la coincidencia es de un 100 %.

5.4.5 Resultados del AdaBoost

La técnica de aprendizaje de conjuntos (*AdaBoost*) se aplicó a los datasets de los pozos seleccionados como se muestra en las tablas 5.28, 5.29 y 5.30 a continuación.

A continuación, las figuras 5.42, 5.43 y 5.44 de los pozos 433, 2654 y 6076 respectivamente con los resultados de la métrica AUC (sobre-muestreo y sub-muestreo) aplicada en el método Adaboost.

En las tablas 5.28, 5.29, y 5.30 se aprecia que para el dataset 433 (sobre muestreo y sub-muestreo), así como en el pozo 2654 sobre-muestreo y en el pozo 6076 (sobre-muestreo y sub-muestreo) presentaron los mejores resultados, no obstante que para el pozo 2654 y 6076 sin ajuste se obtuvo un valor de 1.000, igualmente, que corresponde al dataset desbalanceado y sin la imputación de datos.

Luego de aplicar los algoritmos anteriores se procede a buscar las coincidencias

Tabla 5.28: Resultado de las métricas del clasificador Adaboost al pozo 433 antes y después.

Medidas	**Sin ajuste**	**Sobre-muestreo**	**Sub-muestreo**
Exactitud (*Accuracy*)	0.998	0.998	0.982
Precisión	0.991	0.998	0.983
Recall	0.980	0.998	0.982
F measure	0.989	0.998	0.982
AUC ROC	1.000	**1.000**	**1.000**

Fuente: Elaboración propia

Tabla 5.29: Resultado de las métricas del clasificador Adaboost al pozo 2654 antes y después.

Medidas	**Sin ajuste**	**Sobre-muestreo**	**Sub-muestreo**
Exactitud (*Accuracy*)	1.000	0.997	0.976
Precisión	0.998	0.997	0.977
Recall	0.997	0.998	0.980
F measure	1.000	0.997	0.979
AUC ROC	1.000	**1.000**	0.99

Fuente: Elaboración propia

Tabla 5.30: Resultado de las métricas del clasificador Adaboost al pozo 6076 antes y después.

Medidas	**Sin ajuste**	**Sobre-muestreo**	**Sub-muestreo**
Exactitud (*Accuracy*)	1.000	0.995	0.964
Precisión	1.000	0.995	0.967
Recall	1.000	0.995	0.964
F measure	1.000	0.995	0.965
AUC ROC	1.000	**1.000**	**1.000**

Fuente: Elaboración propia

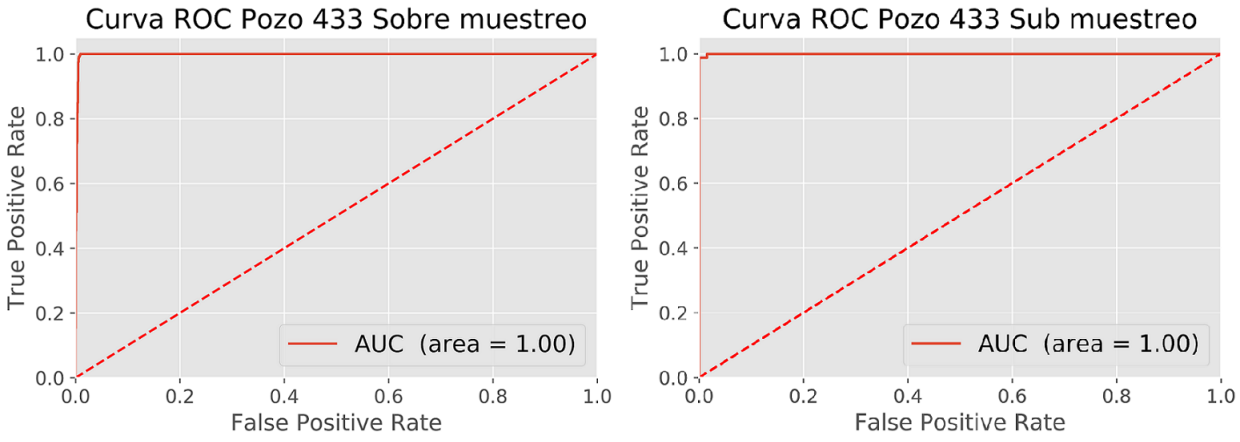


Figura 5.42: Métrica AUC en conjunto del Pozo 433

Fuente: Elaboración propia

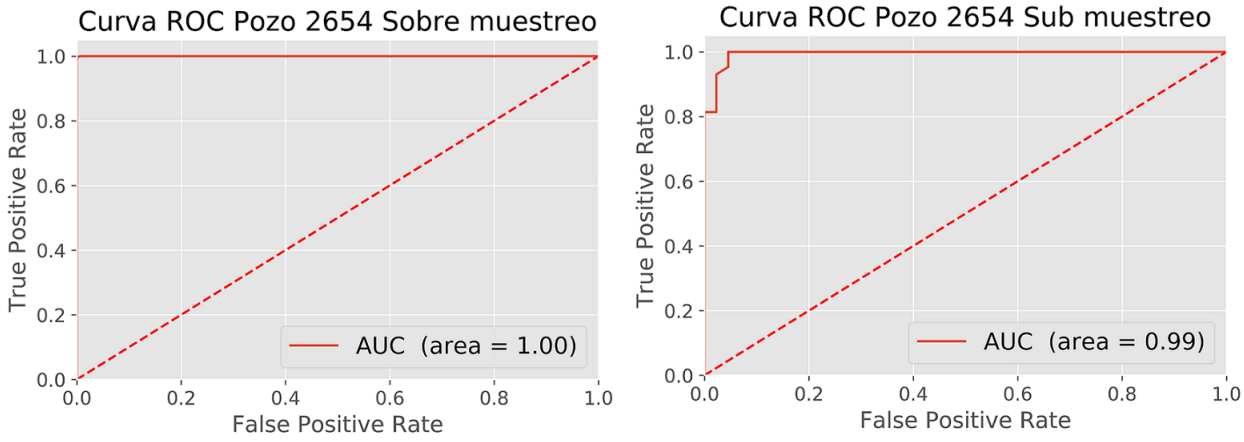


Figura 5.43: Métrica AUC en conjunto del Pozo 2654

Fuente: Elaboración propia

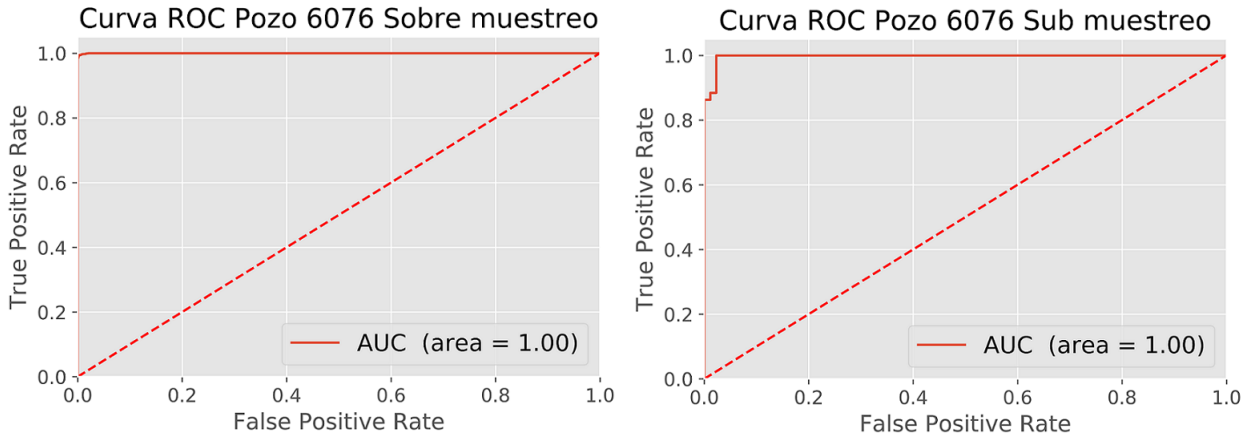


Figura 5.44: Métrica AUC en conjunto del Pozo 6076

Fuente: Elaboración propia

de los registros en los datasets originales.

A continuación, en la tabla 5.31 los resultados de las coincidencias con los datos originales.

5.4.6 Coincidencias

Tabla 5.31: Resultados de las coincidencias del Adaboost y los datos originales

IDPOZO	Adaboost			
	Sobre-muestreo	Sub-muestreo	% Sobre-muestreo	% Sub-muestreo
433	245	77	98.39	30.92
2654	144	43	**100**	29.65
6076	294	94	97.02	31.02

Fuente: Elaboración propia

En las tablas 5.27 y 5.31, de RL Binaria y Adaboost respectivamente se aprecian las coincidencias de los algoritmos aplicados, encontrándose la mayor coincidencia en el caso del pozo 2654 sobre-muestreo con un 100 % de coincidencias tanto para la RL Binaria como para el algoritmo Adaboost.

5.5 Fase 5: Obtención de la solución y construcción de la recomendación para el tomador de decisión

5.5.1 Resultados del PCA en la RL Binaria y Adaboost

A continuación, se aplicó el método de Análisis de Componentes Principales (Principal Component Analysis, PCA) a cada uno de los datasets para determinar las características más importantes en cada uno de los pozos.

Tabla 5.32: Resultados de aplicar PCA en las coincidencias del Adaboost en el Pozo 433

PC	Pozo 433 Adaboost Sobre muestreo														
	DEPTH	SN	ILD	CALI	GRDS	RHOBDS	DTDS	NPHI	VCL	PHIE	SW	BVW	DPHI-CALC	KCORTE	KTIX
PC-1	0.4492	0.0405	0.0893	0.2944	-0.0297	-0.3493	0.2327	0.1891	0.1421	-0.1268	0	-0.3163	0.3493	0	0.1395
PC-2	-0.3251	0.2161	0.3694	-0.1434	-0.5227	-0.2131	-0.1392	0.1692	0.2026	-0.2647	0	-0.2147	0.2131	0	0.0644
PC	Pozo 433 Adaboost Sub muestreo														
	DEPTH	SN	ILD	CALI	GRDS	RHOBDS	DTDS	NPHI	VCL	PHIE	SW	BVW	DPHI-CALC	KCORTE	KTIX
PC-1	0.4702	0.0551	0.0401	0.2919	0.0733	-0.3441	0.2765	0.1783	0.0053	-0.1088	0	1.32E-23	-0.2679	0	0.1406
PC-2	-0.2913	0.2601	0.3598	0.509	-0.4573	-0.2095	-0.1061	0.2375	0.3271	-0.3429	0	2.12E-22	-0.2331	0	-0.0176

Fuente: Elaboración propia

Tabla 5.33: Resultados de aplicar PCA en las coincidencias de la RL en el Pozo 433

PC	Pozo 433 RL Sobre muestreo														
	DEPTH	SN	ILD	CALI	GRDS	RHOBDS	DTDS	NPHI	VCL	PHIE	SW	BVW	DPHI-CALC	KCORTE	KTIX
PC-1	0.4696	-5.55E-17	0.0972	0.2929	0.0304	-0.3413	0.2377	0.1830	0.1383	-0.1168	0	-0.2951	0.3413	0	0.1403
PC-2	0.0345	-2.08E-16	0.0127	0.0086	-0.2320	0.4211	0.0305	0.1122	0.4875	-0.4336	0	-0.3414	-0.4210	0	0.1848
PC	Pozo 433 RL Sub muestreo														
	DEPTH	SN	ILD	CALI	GRDS	RHOBDS	DTDS	NPHI	VCL	PHIE	SW	BVW	DPHI-CALC	KCORTE	KTIX
PC-1	0.1628	0.3429	0.3424	-0.1356	-0.3711	0.2955	-0.3659	-0.3366	-0.3272	0.1362	0.0628	0.1421	-0.2955	0	0.1027
PC-2	-0.3375	-0.1778	-0.1661	0.1425	-0.0475	-0.0834	0.0465	0.1291	-0.3129	0.5065	0.0602	0.4667	0.0834	0	0.4286

Fuente: Elaboración propia

Tabla 5.34: Resultados de aplicar PCA en las coincidencias del Adaboost en el Pozo 2654

PC	Pozo 2654 Adaboost Sobre muestreo																						
	DEPTH	VCL	NPHI	CALI	RHOB	DPHI	TVD	BIT	GR	R850	R60O	R30O	R20O	DT	SPHI	PHIE	SW	RHOMA	KCORTE	KTIX	YOUNG	PR	BRITT
PC-1	-0.3885	-0.2414	-0.1044	0.0844	-0.3897	0.3896	-0.3884	0	-0.2334	0	0	0	0.2205	0.2808	0.0660	0	0	-0.2473	0	0	0	0	0
PC-2	-0.0532	0.2776	0.5503	0.1235	-0.0799	0.0798	-0.0481	0	0.2857	-1.06E-22	-6.62E-24	0	0.1965	-0.0832	0.3768	0	0	-0.4710	0	0	0	0	0
PC	Pozo 2654 Adaboost Sub muestreo																						
	DEPTH	VCL	NPHI	CALI	RHOB	DPHI	TVD	BIT	GR	R850	R60O	R30O	R20O	DT	SPHI	PHIE	SW	RHOMA	KCORTE	KTIX	YOUNG	PR	BRITT
PC-1	-0.4064	-0.2359	-0.1297	0.0806	-0.4052	0.4052	-0.4068	0	-0.2003	0	0	0	0	0.2840	0.1563	0	0	-0.1703	0	0	0	0	0
PC-2	0.1136	-0.2815	-0.5702	-0.1478	0.0691	-0.0689	0.1096	0	-0.3577	0	0	0	0	0.0390	-0.2315	0	0	0.5534	0	0	0	0	0

Fuente: Elaboración propia

Tabla 5.35: Resultados de aplicar PCA en las coincidencias de la RL en el Pozo 2654

PC	Pozo 2654 RL Sobre muestreo																						
	DEPTH	VCL	NPHI	CALI	RHOB	DPHI	TVD	BIT	GR	R850	R60O	R30O	R20O	DT	SPHI	PHIE	SW	RHOMA	KCORTE	KTIX	YOUNG	PR	BRITT
PC-1	-0.3878	-0.2408	-0.1043	0.0859	-0.3886	0.3886	-0.3877	3.31E-24	-0.2325	0	2.02E-28	0	0.2217	0.2804	0.0660	0	0	-0.2463	0	0	0	0	0
PC-2	-0.0527	0.2782	0.5501	0.1254	-0.0791	0.0790	-0.0477	-2.71E-20	0.2865	0	2.12E-22	0	0.1978	-0.0833	0.3766	0	0	-0.4699	0	0	0	0	0
PC	Pozo 2654 RL Sub muestreo																						
	DEPTH	VCL	NPHI	CALI	RHOB	DPHI	TVD	BIT	GR	R850	R60O	R30O	R20O	DT	SPHI	PHIE	SW	RHOMA	KCORTE	KTIX	YOUNG	PR	BRITT
PC-1	-0.3926	-0.1361	-0.2286	0.0193	-0.3946	0.3946	-0.3937	-1.32E-23	-0.2420	0	-8.08E-28	0	0	0.3779	0.0706	0	0	-0.0852	0	0	0	0	0
PC-2	-0.2124	0.4221	0.4287	0.1276	-0.1672	0.1670	-0.2085	0	0.2971	0	0	0	0	-0.2296	0.0818	0	0	-0.5276	0	0	0	0	0

Fuente: Elaboración propia

Tabla 5.36: Resultados de aplicar PCA en las coincidencias del Adaboost en el Pozo 6076

PC	Pozo 6076 Adaboost Sobre muestreo																								
	DEPTH	TVD	DPHI-CALC	BIT	CALC	NPHI	SW	RHOB	GR	M1RX	M1R9	M1R6	M1R3	M1R2	M1R1	DT	BVW	VWCL	RHOMA	DTMA	KTIX	DTS	POISDIN	VPVS	GRN
PC-1	-0.4069	-0.4071	-0.2091	4.34E-19	-0.3173	-0.0455	1.69E-21	0.2091	-0.0642	2.07E-25	0	0	1.58E-30	0	0	0.0796	0	0	0.0075	0	0	-0.1894	-0.3611	0.3668	-0.0642
PC-2	-0.0606	-0.0605	0.4547	-5.55E-17	0.0307	-0.3998	-5.55E-17	-0.4546	-0.2414	0	1.69E-21	0	-3.31E-24	0	0	0.0037	0	0	-0.1457	0	0	0.0510	-0.0748	-0.0828	-0.2414
PC	Pozo 6076 Adaboost Sub muestreo																								
	DEPTH	TVD	DPHI-CALC	BIT	CALC	NPHI	SW	RHOB	GR	M1RX	M1R9	M1R6	M1R3	M1R2	M1R1	DT	BVW	VWCL	RHOMA	DTMA	KTIX	DTS	POISDIN	VPVS	GRN
PPC-1	-0.4022	-0.4024	-0.1949	2.60E-18	-0.3262	-0.0368	0	0.1949	-0.0791	2.07E-25	6.46E-27	0	0	0	0	0.0703	0	0	-0.0024	0	0	-0.2101	-0.3625	-0.374	-0.0791
PC-2	-0.0587	-0.0588	0.4199	1.11E-16	-0.0064	-0.4017	-1.73E-18	-0.4198	-0.2482	-1.69E-21	-1.06E-22	0	0	0	0	-0.0915	0	0	-0.1355	0	0	0.1552	-0.0191	-0.0236	-0.2482

Fuente: Elaboración propia

Tabla 5.37: Resultados de aplicar PCA en las coincidencias de la RL en el Pozo 6076

PC	Pozo 6076 RL Sobre muestreo																								
	DEPTH	TVD	DPHI-CALC	BIT	CALC	NPHI	SW	RHOB	GR	M1RX	M1R9	M1R6	M1R3	M1R2	M1R1	DT	BVW	VWCL	RHOMA	DTMA	KTIX	DTS	POISDIN	VPVS	GRN
PC-1	-0.4069	-0.4071	-0.2103	4.34E-19	-0.3176	-0.0435	0	0.2104	-0.0642	0	-3.23E-27	-1.01E-28	0	0	0	0.0802	0	0	0.0084	0	0	-0.1897	-0.3601	-0.3659	-0.0642
PC-2	-0.0612	-0.0611	0.4546	-3.33E-16	0.0319	-0.4010	0	-0.4546	-0.2402	-5.42E-20	-1.69E-21	0	0	0	0	0.0027	0	0	-0.1480	0	0	0.4499	-0.0791	-0.0869	-0.2402
PC	Pozo 6076 RL Sub muestreo																								
	DEPTH	TVD	DPHI-CALC	BIT	CALC	NPHI	SW	RHOB	GR	M1RX	M1R9	M1R6	M1R3	M1R2	M1R1	DT	BVW	VWCL	RHOMA	DTMA	KTIX	DTS	POISDIN	VPVS	GRN
PPC-1	-0.4033	-0.4034	-0.2135	0	-0.3240	-0.0136	4.24E-22	0.2136	0	1.29E-26	2.02E-28	3.16E-30	0	0	0	0.0704	0	0	0.0136	0	0	-0.2192	-0.3629	-0.3750	0
PC-2	-0.1078	-0.1078	-0.4344	-5.55E-17	-0.0064	0.4234	2.17E-19	0.4343	0	-1.06E-22	3.31E-24	0	0	0	0	0.1055	0	0	0.1027	0	0	-0.1419	0.0541	0.0661	0

Fuente: Elaboración propia

Como ha quedado mostrado en las tablas 5.32 a la 5.37, las variables o características que más contribuyen a la explicación de gran parte de la variabilidad en los datos son:

1. DEPTH(*)
2. DPHI-CALC
3. RHOBDS
4. VCL(*)
5. SN
6. PHIE(*)
7. BVW(*)
8. KTIX(*)
9. DPHI
10. NPHI
11. RHOMA
12. VPVS
13. PayFlag(*)

De estas características las que están marcadas con (*) significa que son aquellas que coinciden con la literatura que son las más importantes para realizar caracterizaciones de pozos.

5.5.2 Resumen comparativo de análisis de los resultados de los algoritmos empleados

A continuación, se analizan las similitudes y diferencias de los registros de los clústers generados entre los algoritmos RL Binaria y Adaboost sobre la base de aquellos registros en los que coinciden con los datos originales con PayFlag en 1.

Tabla 5.38: Resultados de las similitudes y diferencias entre RL Binaria y Adaboost

IDPOZO	RL Binaria y Adaboost	
	Similitudes	Diferencias
433 Sobre-muestreo	232	3
433 Sub-muestreo	76	0
2654 Sobre-muestreo	42	101
2654 Sub-muestreo	38	4
6076 Sobre-muestreo	294	0
6076 Sub-muestreo	90	3

Fuente: Elaboración propia

A partir de las observaciones de la tabla 5.38, se puede apreciar que en gran parte de los registros tanto de RL Binaria como de Adaboost en los 3 pozos existe similitud en los registros clasificados de los que corresponden al PayFlag en 1 y que coinciden con los datos originales, excepto en el caso de la muestra del pozo 2654 Sobre muestreo, en donde la mayoría de los registros son diferentes.

5.6 Especificaciones de software

A continuación, se presenta una descripción detallada del lenguaje de programación utilizado.

5.6.1 Lenguaje de programación Python

Para aplicar los algoritmos, se utilizó el lenguaje de programación de distribución libre llamado *Python*. Este lenguaje es interpretado e interactivo orientado a objeto. Provee un alto nivel de estructura de datos como listas y arreglos asociados (llamados diccionarios), tipado dinámico, clases, excepciones, módulos, administración de memoria, etc. Tiene una sintaxis elegante y además, es potente y de uso general. Fue diseñado en 1990 por Guido van Rossum [98]. Tiene una base de documentación sólida, además la comunidad que contribuye con código de programación en este lenguaje es bastante grande, lo que ayuda a que sea fácil de entender para otros expertos y programadores.

La razón de elegir *Python* sobre otros lenguajes de programación que existen está enmarcada en varios aspectos, a saber:

- Python es un lenguaje de programación que tiene un alto nivel de abstracción a nivel de máquina, lo cual quiere decir que su código se modela alrededor de sus clases y estas pueden ser mucho más sencillas de remodelar en caso de que el proyecto que se está realizando cambie a ser más grande de cuando se inició.
- Por otra parte en las áreas de desarrollo web, robótica, inteligencia artificial ha tomado un papel importante dentro de estos lo cual ha llevado a este lenguaje a ampliar su entorno, dándole robustez a sus funciones.
- Por otra parte, la idea detrás de la creación de este lenguaje de programación de código abierto fue la de asemejar lo mayor posible a la sintaxis humana. Esto, lo convierte en uno de los lenguajes más sencillos de aprender, al mismo tiempo que asegura que otras personas puedan tener un mayor entendimiento del código, lo cual lo hace versátil para compartirlo con la comunidad científica.

5.7 Especificaciones de Hardware

Para realizar los experimentos descritos, se utilizó una estación de trabajo, marca DELL serie 5000, con las siguientes características:

- *Processor*: Core i7-8550U 3.2 Ghz
- *System Memory*: 16 GB
- Hard Drive: 2 Tb SATA
- Sistema Operativo: Windows 10

5.8 Conclusiones

En este capítulo ha quedado descrito toda la experimentación y los resultados de los algoritmos aplicados.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1 Conclusiones

A continuación, se describen las conclusiones importantes en este trabajo:

1. Se ha desarrollado una nueva metodología para la caracterización de pozos de petróleo a partir de datos geológicos, tomando en cuenta la Metodología General de Apoyo a la Decisión Multicriterio y la metodología fundamental para la ciencia de datos propuesta por IBM, lo cual facilita y contribuye a obtener resultados razonables que pueden aportar valor en proyectos de prospección de pozos de petróleo a partir de datos de tipo geológicos.

2. Se obtuvieron los resultados de la clasificación realizada a partir de los algoritmos de clasificación aplicados obteniendo aquellas variables o características más importantes en los datos que se tienen.

3. Ha quedado determinado que entre las profundidades de 500 a 2300 mts se encuentra la mayor concentración de registros con PayFlag en 1 en la mayoría de los casos.

4. En las variables relevantes, varias coinciden con la literatura; que son DEPTH, PHIE, KTIX, además del PayFlag que logran explicar entre el 60 - 66 % de la

variabilidad de los datos.

5. Los resultados confirman que con los métodos aplicados, se puede clasificar el PayFlag tomando como base las variables que se tienen.

6. Se confirma que se puede hacer una reducción de la cantidad de variables.

7. Se obtienen coincidencias en cuanto a la calidad de las predicciones en la clase de PayFlag 1.

8. Se confirma que los modelos no supervisados son buenos, pero que los modelos supervisados brindan un mejor resultado.

9. Se confirma que en el tratamiento de los datos desbalanceados, con la técnica Sobre-muestreo se obtuvieron mejores resultados que con la técnica Sub-muestreo.

6.2 Recomendaciones y Trabajo Futuro

A continuación, se describen las recomendaciones y trabajo futuro propuestos en este trabajo:

1. Se propone la aplicación de otras técnicas de clasificación más sofisticadas y los resultados que se obtengan contrastarlos con los obtenidos en este trabajo.

2. Se propone la utilización de otros métodos para determinar el valor del hiperparámetro K en el algoritmo KMeans.

3. Se sugiere refinar los parámetros de los algoritmos aplicados, en este caso el algoritmo de RL Binaria y Adaboost.

4. Utilizar otros datasets de pozos de petróleo con características de desbalance de clases y establecer coincidencias.

5. Aplicar técnicas de Rough Sets (Conjuntos Aproximados) o Fuzzy Rough Sets (Conjuntos Aproximados difusos) para hacer caracterización por reglas sobre los datos.

Bibliografía

[1] Djebbar Tiab Amine Mazouzi Abdelkader Kouider, El Ouahed. Application of artificial intelligence to characterize naturally fractured zones in hassi messaoud oil field, algeria. *Journal of Petroleum Science and Engineering*, 49:122–141, 2005.

[2] Mohammad Ali Ahmadi, Sohrab Zendehboudi, Ali Lohi, Ali Elkamel, and Ioannis Chatzis. Reservoir permeability prediction by neural networks combined with hybrid genetic algorithm and particle swarm optimization. *Geophysical Prospecting*, 61(3):582–598, 2013.

[3] Mohiuddin Ahmed and Al-Sakib Khan Pathan. *Data Analytics: Concepts, Techniques, and Applications*. CRC Press, 2018.

[4] T Ahmed, CA Link, KW Porter, CJ Wideman, P Himmer, J Braun, et al. Application of neural network parameter prediction in reservoir characterization and simulation-a case history: the rabbit hills field. In *Latin American and Caribbean Petroleum Engineering Conference*. Society of Petroleum Engineers, 1997.

[5] Kabiru O Akande, Sunday O Olatunji, Taoreed O Owolabi, AbdulAzeez AbdulRaheem, et al. Comparative analysis of feature selection-based machine learning techniques in reservoir characterization. In *SPE Saudi Arabia Section Annual Technical Symposium and Exhibition*. Society of Petroleum Engineers, 2015.

[6] K Aminian and S Ameri. Application of artificial neural networks for reservoir characterization with limited data. *Journal of Petroleum Science and Engineering*, 49(3-4):212–222, 2005.

[7] F Aminzadeh, Jacob Barhen, CW Glover, and NB Toomarian. Estimation of reservoir parameter using a hybrid neural network. *Journal of Petroleum Science and Engineering*, 24(1):49–56, 1999.

[8] Fred Aminzadeh. Applications of ai and soft computing for challenging problems in the oil industry. *Journal of Petroleum Science and Engineering*, 47:5–14, 2005.

[9] Fred Aminzadeh and Shankar Chatterjee. Applications of clustering in exploration seismology. *Geoexploration*, 23(1):147–159, 1984.

[10] et al Andrew McCallum. A machine learning approach to buildingdomain specific search engines. 1999.

[11] Maureen Ani, Gbenga Oluyemi, Andrei Petrovski, Sina Rezaei-Gomari, et al. Reservoir uncertainty analysis: The trends from probability to algorithms and machine learning. In *SPE Intelligent Energy International Conference and Exhibition*. Society of Petroleum Engineers, 2016.

[12] Fatai Anifowose, Jane Labadin, and Abdulazeez Abdulraheem. Predicting petroleum reservoir properties from downhole sensor data using an ensemble model of neural networks. In *Proceedings of Workshop on Machine Learning for Sensory Data Analysis*, pages 27–34, 2013.

[13] Fatai Anifowose, Jane Labadin, and Abdulazeez Abdulraheem. Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines. *Applied Soft Computing*, 26:483–496, 2015.

[14] Fatai A Anifowose, Jane Labadin, and Abdulazeez Abdulraheem. Prediction of petroleum reservoir properties using different versions of adaptive neuro-fuzzy

inference system hybrid models. *Int. J. Comput. Inf. Syst. Ind. Manage. Appl*, 5:413–426, 2013.

[15] Fatai Adesina Anifowose. Artificial intelligence application in reservoir characterization and modeling whitening the black box. *SPE Saudi Arabia section Young Professionals Technical Symposium*, 34:1834–1843, 2011.

[16] Fatai Adesina Anifowose, Jane Labadin, and Abdulazeez Abdulraheem. Ensemble machine learning: An untapped modeling paradigm for petroleum reservoir characterization. *Journal of Petroleum Science and Engineering*, 151:480–487, 2017.

[17] Fatai Adesina Anifowose, Jane Labadin, and Abdulazeez Abdulraheem. Hybrid intelligent systems in petroleum reservoir characterization and modeling: the journey so far and the challenges ahead. *Journal of Petroleum Exploration and Production Technology*, 7(1):251–263, 2017.

[18] Nancy Maribel Arratia Martínez. *Metodología de apoyo a la decisión en la selección de carteras de proyectos con beneficios o impactos de carácter social*. PhD thesis, Universidad Autónoma de Nuevo León, 2012.

[19] Marcelo Artigas. Exploración y producción de petróleo: reservorios, perforación y terminación de pozos. page 36. Cámara argentina del libro, 2010.

[20] RS Balch, BS Stubbs, WW Weiss, S Wo, et al. Using artificial intelligence to corellate multiple seismic attributes to reservoir properties. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers, 1999.

[21] Sanghamitra Bandyopadhyay and Sriparna Saha. *Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications*. Springer Science & Business Media, 2012.

[22] Adrián Chao Bataller. Modelos para la evaluación de pozos de petróleo a partir de sus características geológicas. Master's thesis, Universidad Autónoma de Nuevo León, 2018.

[23] A. Annie Portia Benson Edwin Raj S. Analysis on credit card fraud detection methods. 2011.

[24] Pavel Berkhin. Survey of clustering data mining techniques. accrue software. *Inc. TR, San Jose, USA*, 2002.

[25] Sergio A Berumen and Francisco Llamazares Redondo. La utilidad de los métodos de decisión multicriterio (como el ahp) en un entorno de competitividad creciente. *Cuadernos de administración*, 20(34), 2007.

[26] Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.

[27] Kirill Y Bogatchev. *Developing a tight gas sand advisor for completion and stimulation in tight gas reservoirs worldwide*. PhD thesis, Texas A & M University, 2008.

[28] Max Bramer. *Principles of data mining*, volume 180. Springer, 2007.

[29] Evgeny Burnaev, Pavel Erofeev, and Artem Papanov. Influence of resampling on accuracy of imbalanced classification. In *Eighth International Conference on Machine Vision (ICMV 2015)*, volume 9875, page 987521. International Society for Optics and Photonics, 2015.

[30] RJ Chappaz. Application of the fuzzy sets theory to the interpretation of seismic. In *Abstract of the 47th SEG meeting, Calgary, Paper R-9*, 1977.

[31] Rohit Choudhry and Kumkum Garg. A hybrid machine learning system for stock market forecasting. 2008.

[32] Alternativa Económica Consultores. Evaluación de proyectos de desarrollo de reservas de hidrocarburos. 14 a Reunión Académica Mexicana de Profesionistas en Evaluación Socioeconómica de Proyectos, 2014.

[33] ER Crain. Petrophysical handbook. *Alberta: Spectrum*, 2000.

[34] Vasant Dhar. *Data Science and Prediction*. Communications of the ACM, second edition, 2013.

[35] Pablo Durán. Los datos perdidos en estudios de investigación¿ son realmente datos perdidos? *Archivos argentinos de pediatría*, 103(6):566–568, 2005.

[36] Jeffrey Erman, Martin Arlitt, and Anirban Mahanti. Traffic classification using clustering algorithms. In *Proceedings of the 2006 SIGCOMM workshop on Mining network data*, pages 281–286, 2006.

[37] Houtan Faridi, Srivathsan Srinivasagopalan, and Rakesh Verma. Performance evaluation of features and clustering algorithms for malware. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 13–22. IEEE, 2018.

[38] Abdulazeez Abdulraheemc Fatai Adesina Anifowose, Jane Labadin. Ensemble machine learning: An un tapped modeling paradigm for petroleum reservoir characterization. 2017.

[39] Abdulazeez Abdulraheem Fatai Anifowose. Fuzzy logic-driven and svm-driven hybrid computational intelligence models applied to oil and gas reservoir characterization. 2011.

[40] Abdulazeez Abdulraheem Fatai Anifowose, Jane Labadin. Ensemble learning model for petroleum reservoir characterization: a case of feed-forward back-propagation neural networks. 2013.

[41] Denis Ferraretti, Giacomo Gamberoni, and Evelina Lamma. Unsupervised and supervised learning in cascade for petroleum geology. *Expert Systems with Applications*, 39(10):9504–9514, 2012.

[42] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[43] Said Gaci and Olga Hachay. *Oil and Gas Exploration: Methods and Application*, volume 72. John Wiley & Sons, 2017.

[44] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011.

[45] Vicente García, José Salvador Sánchez, and Ramón Alberto Mollineda. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21, 2012.

[46] Miguel Garre, Juan José Cuadrado, Miguel A Sicilia, Daniel Rodríguez, and Ricardo Rejas. Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software, 2007.

[47] Edgar René Rangel German. Ior-eor: Una oportunidad histórica para méxico. 2015.

[48] RB Gharbi, Adel M Elsharkawy, et al. Neural network model for estimating the pvt properties of middle east crude oils. In *Middle East Oil Show and Conference*. Society of Petroleum Engineers, 1997.

[49] Bharat B Gulyani, BG Prakash Kumar, and Arshia Fathima. Bagging ensemble model for prediction of dead oil viscosity. *International Journal of Chemical Engineering and Applications*, 8(2):102, 2017.

[50] Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

[51] Xiaogang He, Nathaniel W Chaney, Marc Schleiss, and Justin Sheffield. Spatial downscaling of precipitation using adaptable random forests. *Water resources research*, 52(10):8217–8237, 2016.

[52] Steven M Holland. Principal components analysis (pca). *Department of Geology, University of Georgia, Athens, GA*, pages 30602–2501, 2008.

[53] Mark H Holtz, Douglas S Hamilton, et al. Reservoir characterization methodology to identify reserve growth potential. In *International Petroleum Conference and Exhibition of Mexico*. Society of Petroleum Engineers, 1998.

[54] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[55] Mohammad Hossin and MN Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

[56] YT Huang, PM Wong, and TD Gedeon. Prediction of reservoir permeability using genetic algorithms. *AI applications (USA)*, 1998.

[57] Zehui Huang, John Shimeld, Mark Williamson, and John Katsube. Permeability prediction with artificial neural network modeling in the venture gas field, offshore eastern canada. *Geophysics*, 61(2):422–436, 1996.

[58] Frank Jahn, Mark Cook, and Mark Graham. *Hydrocarbon exploration and production*. Elsevier, 2008.

[59] M Jamialahmadi and FG Javadpour. Relationship of permeability, porosity and depth using an artificial neural network. *Journal of Petroleum Science and Engineering*, 26(1-4):235–239, 2000.

[60] Michael Jordan and Tom M. Mitchell. Machine learning: Trends, perspectives, and prospects. 2015.

[61] K Senthamarai Kannan, K Manoj, and S Arumugam. Labeling methods for identifying outliers. *International Journal of Statistics and Systems*, 10(2):231–238, 2015.

[62] Trupti M Kodinariya and Prashant R Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[63] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. 2001.

[64] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[65] Brett Lantz. *Machine learning with R*. Packt Publishing Ltd, 2013.

[66] Ximming Ma Lei Xi Lei Shi, Mei Weng. Rough set based decision tree ensemble algorithm for text classification, 2010.

[67] X Li, CW Chan, and HH Nguyen. Application of the neural decision tree approach for prediction of petroleum production. *Journal of Petroleum science and engineering*, 104:11–16, 2013.

[68] Xiongmin Li and Christine W Chan. Application of an enhanced decision tree learning approach for prediction of petroleum production. *Engineering Applications of Artificial Intelligence*, 23(1):102–109, 2010.

[69] Jong-Se Lim. Reservoir properties determination using fuzzy logic and neural networks from well data in offshore korea. *Journal of Petroleum Science and Engineering*, 49(3-4):182–192, 2005.

[70] Hancong Liu, Sirish Shah, and Wei Jiang. On-line outlier detection and data cleaning. *Computers & chemical engineering*, 28(9):1635–1647, 2004.

[71] Xuewei Liu, Ping Xue, and Yanda Li. Neural network method for tracing seismic events. In *SEG Technical Program Expanded Abstracts 1989*, pages 716–718. Society of Exploration Geophysicists, 1989.

[72] Walid M Mabrouk. Bvw as an indicator for hydrocarbon and reservoir homogeneity. *Journal of Petroleum Science and Engineering*, 49(1-2):57–62, 2005.

[73] Ms R Malarvizhi and Antony Selvadoss Thanamani. K-nearest neighbor in missing data imputation. *International Journal of Engineering Research and Development*, 5(1):5–7, 2012.

[74] Rúsbel Domínguez Domínguez Martínez. *Aplicación De Ciencia De Datos Para La Creación De Software Predictivo De Morbilidad Materna En México.* PhD thesis, Universidad de Montemorelos, 2017.

[75] Fred Aminzadeh Masoud Nikravesh. Past, present and future intelligent reservoir characterization trends. *Journal of Petroleum Science and Engineering*, 31:67–79, 2001.

[76] Ana Isabel Caballero Merino, Angel M Gento Municio, and Alfonso Redondo Castán. Toma de decisiones multicriterio, con incertidumbre, en el ámbito de los recursos humanos. In *V Congreso de Ingeniería de Organización*, 2003.

[77] A Mirzaei-Paiaman and S Salavati. The application of artificial neural networks for the prediction of oil production flow rate. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 34(19):1834–1843, 2012.

[78] S Mohaghegh and S Ameri. Artificial neural network as a valuable tool for petroleum engineers. *Paper SPE*, 29220, 1995.

[79] Shahab Mohaghegh, Reza Arefi, Sam Ameri, Khashayar Aminiand, and Roy Nutter. Petroleum reservoir characterization with the aid of artificial neural networks. *Journal of Petroleum Science and Engineering*, 16(4):263–274, 1996.

[80] Shahab Mohaghegh et al. Virtual-intelligence applications in petroleum engineering: Part 1—artificial neural networks. *Society of Petroleum Engineers*, 52:64–73, 2000.

[81] Shahab D Mohaghegh et al. Recent developments in application of artificial intelligence in petroleum engineering. *Journal of Petroleum Technology*, 57(04):86–91, 2005.

[82] Mohamed B.Bader-El-Den James M.Buicka Munirudeen A.Oloso, Mohamed G.Hassan. Hybrid functional networks for oil reservoir pvt characterisation. 2017.

[83] Masoud Nikravesh. Soft computing-based computational intelligent for reservoir characterization. *Expert Systems with Applications*, 26:19–38, 2004.

[84] Masoud Nikravesh and Mahnaz Hassibi. Intelligent reservoir characterization (iresc). In *IEEE International Conference on Industrial Informatics, 2003. INDIN 2003. Proceedings.*, pages 369–373. IEEE, 2003.

[85] Munirudeen A Oloso, Mohamed G Hassan, Mohamed B Bader-El-Den, and James M Buick. Hybrid functional networks for oil reservoir pvt characterisation. *Expert Systems with Applications*, 87:363–369, 2017.

[86] Ahmed Ouenes. Practical application of fuzzy logic and neural networks to fractured reservoir characterization. *Computers & Geosciences*, 26(8):953–962, 2000.

[87] Norman R Paterson and Colin V Reeves. Applications of gravity and magnetic surveys; the state-of-the-art in 1985. *Geophysics*, 50(12):2558–2594, 1985.

[88] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[89] Hector H Perez, Akhil Datta-Gupta, Srikanta Mishra, et al. The role of electrofacies, lithofacies, and hydraulic flow units in permeability predictions from well logs: a comparative analysis using classification trees. *SPE Reservoir Evaluation & Engineering*, 8(02):143–155, 2005.

[90] Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* .ºReilly Media, Inc.", 2013.

[91] Morteza Raeesi, Ali Moradzadeh, Faramarz Doulati Ardejani, and Mashallah Rahimi. Classification and identification of hydrocarbon reservoir lithofacies and their heterogeneity using seismic attributes, logs data and artificial neural networks. *Journal of Petroleum Science and engineering*, 82:151–165, 2012.

[92] Travis Ramsay, Jeffrey Yarus, et al. Petrofacies determination in unconventional reservoirs driven by a simulation-to-seismic process. In *EUROPEC 2015*. Society of Petroleum Engineers, 2015.

[93] JB Rollins. Foundational methodology for data science. *Domino Data Lab, Inc., Whitepaper*, 2015.

[94] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.

[95] Bernard Roy. *Multicriteria methodology for decision aiding*, volume 12. Springer Science & Business Media, 2013.

[96] Irina Samoylova Rustam Alimkhanov. Application of data mining tools for analysis and prediction of hydraulic fracturing efficiency for the bv8 reservoir of the povkh oil field. 2014.

[97] Muhammad Sahimi. Fractal-wavelet neural-network approach to characterization and upscaling of fractured reservoirs. *Computers & Geosciences*, 26(8):877–905, 2000.

[98] Michel F Sanner et al. Python: a programming language for software integration and development. *J Mol Graph Model*, 17(1):57–61, 1999.

[99] Abdus Satter, Ghulam M Iqbal, and James L Buchwalter. *Practical enhanced reservoir engineering: assisted with simulation software*. Pennwell Books, 2008.

[100] Jared Schuetter, Srikanta Mishra*, Ming Zhong, and Randy LaFollette. Data analytics for production optimization in unconventional reservoirs. In *Unconventional Resources Technology Conference, San Antonio, Texas, 20-22 July 2015*, pages 249–269. Society of Exploration Geophysicists, American Association of Petroleum ..., 2015.

[101] Giovanni Seni and John F Elder. Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis lectures on data mining and knowledge discovery*, 2(1):1–126, 2010.

[102] H Simon. The new science of management decision 3rd revised edition (1960) prentice-hall. *Englewood Cliffs, NJ*, 1977.

[103] Lindsay I Smith. A tutorial on principal components analysis. Technical report, 2002.

[104] Anja Struyf, Mia Hubert, Peter Rousseeuw, et al. Clustering in an object-oriented environment. *Journal of Statistical Software*, 1(4):1–30, 1997.

[105] DK Thara, BG PremaSudha, and Fan Xiong. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. *Pattern Recognition Letters*, 128:544–550, 2019.

[106] H Trappe and C Hellmich. Using neural networks to predict porosity thickness from 3d seismic data. *First break*, 18(9), 2000.

[107] Martyn Unsworth. New developments in conventional hydrocarbon exploration with electromagnetic methods. *CSEG Recorder*, 30(4):34–38, 2005.

[108] Behzad Vaferi, Reza Eslamloueyan, and Shahab Ayatollahi. Automatic recognition of oil reservoir models from well testing data by using multi-layer perceptron networks. *Journal of Petroleum Science and Engineering*, 77(3-4):254–262, 2011.

[109] Carmen Viada, Carlos Bouza, Javier Ballesteros, Martha Fors, Mayteé Robaina, and Rolando Uranga. Revisión sistemática de los métodos de imputación de datos faltantes.

[110] Begoña Vitoriano. Teoría de la decisión: decisión con incertidumbre, decisión multicriterio y teoría de juegos. *Universidad Complutense de Madrid*, 107, 2007.

[111] Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*, 63(8):826–833, 2010.

[112] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[113] Kok Wai Wong, Chun Che Fung, Yew Soon Ong, and Tamas D Gedeon. Reservoir characterization using support vector machines. *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, 2:354–359, 2005. Accessed: 2019-04-19.

[114] Patrick Wong, Fred Aminzadeh, and Masoud Nikravesh. *Soft computing for reservoir characterization and modeling*, volume 80. Physica, 2013.

[115] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[116] Yunxin Xie, Chenyang Zhu, Wen Zhou, Zhongdong Li, Xuan Liu, and Mei Tu. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *Journal of Petroleum Science and Engineering*, 160:182–193, 2018.

[117] B Yeten and F Gümrah. The use of fractal geostatistics and artificial neural networks for carbonate reservoir characterization. *Transport in porous media*, 41(2):173–195, 2000.

[118] Shuo Liang Lin Yu-dong Cai. Support vector machines for predicting rrna-, rna-, and dna-binding proteins from amino acid sequence. 2003.

[119] Paulina Alejandra Ávila Torres. Un enfoque integrado multicriterio para la planificación de las frecuencias de paso y las tablas de tiempo de una empresa de transporte urbano. Master's thesis, Universidad Autónoma de Nuevo León, 2012.

Resumen autobiográfico

Daniel Chong Sánchez

Candidato para obtener el grado de
Maestría en Ciencias de la Ingeniería
con Orientación en Sistemas

Universidad Autónoma de Nuevo León
Facultad de Ingeniería Mecánica y Eléctrica

Tesis:
Descubrimiento de conocimiento en pozos de petróleo basado en datos geológicos

Nacido el 22 de septiembre de 1985, en La Habana, Cuba. Hijo de Maricela Sánchez Lorenzo y Pedro David Chong Castro, graduado de Ingeniería en Informática, en el centro de altos estudios: "Ciudad Universitaria José Antonio Echeverría". Como profesional se ha desempeñado en el área de informática en el hospital municipal "Freyre de Andrade", ubicado en La Habana, Cuba.